Journal Pre-proof

Dual-scale enhanced and cross-generative consistency learning for semi-supervised medical image segmentation

Yunqi Gu, Tao Zhou, Yizhe Zhang, Yi Zhou, Kelei He, Chen Gong, Huazhu Fu

Please cite this article as: Y. Gu, T. Zhou, Y. Zhang et al., Dual-scale enhanced and cross-generative consistency learning for semi-supervised medical image segmentation, *Pattern Recognition* (2024), doi: https://doi.org/10.1016/j.patcog.2024.110962.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Dual-scale Enhanced and Cross-generative Consistency Learning for Semi-supervised Medical Image Segmentation

Yunqi Gu[a], Tao Zhou[a,*], Yizhe Zhang[a], Yi Zhou[b], Kelei He[c], Chen Gong[a], Huazhu Fu[d]

[a]*PCA Lab, the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*
[b]*School of Computer Science and Engineering, Southeast University, Nanjing 211189, China*
[c]*Medical School, Nanjing University, Nanjing 210023, China, and the National Institute of Healthcare Data Science at Nanjing University, Nanjing 210023, China*
[d]*Institute of High Performance Computing, A*STAR, Singapore*

Abstract

Medical image segmentation plays a crucial role in computer-aided diagnosis. However, existing methods heavily rely on fully supervised training, which requires a large amount of labeled data with time-consuming pixel-wise annotations. Moreover, accurately segmenting lesions poses challenges due to variations in shape, size, and location. To address these issues, we propose a novel Dual-scale Enhanced and Cross-generative consistency learning framework for semi-supervised medical image Segmentation (DEC-Seg). First, we propose a Cross-level Feature Aggregation (CFA) module that integrates cross-level adjacent layers to enhance the feature representation ability across different resolutions. To address scale variation, we present a scale-enhanced consistency constraint, which ensures consistency in the segmentation maps generated from the same input image at different scales. This constraint helps handle variations in lesion sizes and improves the robustness of the model. Furthermore, we propose a cross-generative consistency scheme, in which the original and perturbed images can be reconstructed using cross-segmentation maps. This consistency constraint allows us to mine effective feature representations and boost the segmentation performance. To further exploit the scale information, we propose a Dual-scale Complementary Fusion (DCF) module that integrates features from two scale-specific decoders operating at different scales to help produce more accurate segmentation maps. Extensive experimental results on multiple medical segmentation tasks (polyp, skin lesion, and brain glioma) demonstrate the effectiveness of our DEC-Seg against other state-of-the-art semi-supervised segmentation approaches. The implementation code will be released at https://github.com/taozh2017/DECSeg.

*Corresponding author: taozhou.ai@gmail.com (T. Zhou)

1. Introduction

Accurate segmentation of organs or lesions from medical images is a crucial task within the realm of medical image processing and analysis. The successful accomplishment of this task holds immense significance for disease diagnosis, organ localization, and lesion segmentation. In recent years, several deep learning models [1–4] have been developed and have shown promising performance in medical image segmentation. However, these models often rely on a fully supervised training strategy, which necessitates a substantial number of pixel-wise annotations. Unfortunately, the annotation process for medical images is both time-consuming and costly due to the need for specialized expertise. To address these challenges, researchers have turned to semi-supervised learning (SSL) as an effective approach to leverage unlabeled data and enhance the performance of segmentation models [5, 6]. By utilizing SSL techniques, models can exploit the abundance of available unlabeled data in medical imaging datasets, reducing their dependence on fully annotated data. This approach not only alleviates the burden of extensive annotation requirements but also has the potential to improve the model's accuracy in medical image segmentation tasks.

Recently, various semi-supervised medical image segmentation methods [7–9] based on deep learning have been developed. Among these methods, the consistent constraint is a widely used strategy, by ensuring that the perturbations on unlabeled data should not significantly vary their outputs or predictions. One of the most representative frameworks is the mean teacher (MT) [10], which designs a perturbation-based consistency loss between the teacher and student models on the unlabeled examples. Inspired by MT, several improved methods focus on designing different perturbation strategies to achieve SSL segmentation. For instance, an uncertainty-aware framework (UA-MT) [11] was proposed to make the student model gradually learn more reliable targets and eliminate unreliable predictions by exploiting the uncertainty information. Verma *et al.* [12] proposed an interpolation consistency training framework, which constrains the prediction at an interpolation of unlabeled points to be consistent with that of the predictions at those points. Chen *et al.* [13] developed a cross-pseudo supervision strategy, which initializes two identical networks with different weights and encourages high consistency between the predictions of two different networks with different weights for the same input. Luo *et al.* [14] designed an uncertainty rectified pyramid consistency (URPC) scheme to enable the segmentation model can

2

produce consistent predictions at different scales. Besides, CLCC [15] was proposed to fuse global and patch-wise information by contrastive learning and consistency constraint. SLCNet [16] was proposed to utilize shape information for the training by feeding pseudo-labels into the network simultaneously. Despite the achievement made, semi-supervised medical image segmentation

35 remains challenging due to variations in the shape, size, and location of lesions. Therefore, it is crucial to develop effective and robust semi-supervised learning strategies that leverage a large amount of unlabeled data to boost the medical image segmentation performance.

To this end, we propose a novel **D**ual-scale **E**nhanced and **C**ross-generative consistency learning framework for semi-supervised medical image **Seg**mentation (**DEC-Seg**), which fully ex-

40 ploits multi-scale information from the labeled and unlabeled data to enhance segmentation performance. Specifically, we first propose a Cross-level Feature Aggregation (CFA) module to integrate the adjacent layer features in the encoder, and the cross-level aggregated features are further incorporated into the scale-specific decoder. Then, we present a scale-enhanced consistency strategy to encourage the consistency of segmentation maps from the same inputs with

45 different scales. This consistency can guide our segmentation network to learn more powerful features for handling scale variations. Moreover, a Dual-scale Complementary Fusion (DCF) module is further proposed to filter the effective information from different scale features and fuse them to boost segmentation performance. Meanwhile, a cross-generative consistency is presented to constrain the original images and perturbed ones can be reconstructed by cross-segmentation

50 maps, which can effectively harness the knowledge from unlabeled data. Finally, we conduct the comparison experiments on three medical image segmentation tasks, and the results demonstrate the superiority of the proposed model over other state-of-the-art semi-supervised segmentation methods.

The main contributions of this paper are listed as follows:

55 • We propose a novel semi-supervised medical image segmentation framework, which leverages scale-enhanced consistency and cross-generative consistency to exploit the relationships between labeled and unlabeled data for boosting the segmentation model.

• A cross-level feature aggregation module is proposed to fuse the cross-level features, which can enhance the representation ability of features within different resolutions.

60 • We present a scale-enhanced consistency constraint to reduce the discrepancy between the predicted segmentation maps from different scale inputs. Besides, a cross-generative consistency is designed to constrain that the original and perturbed images can be reconstructed
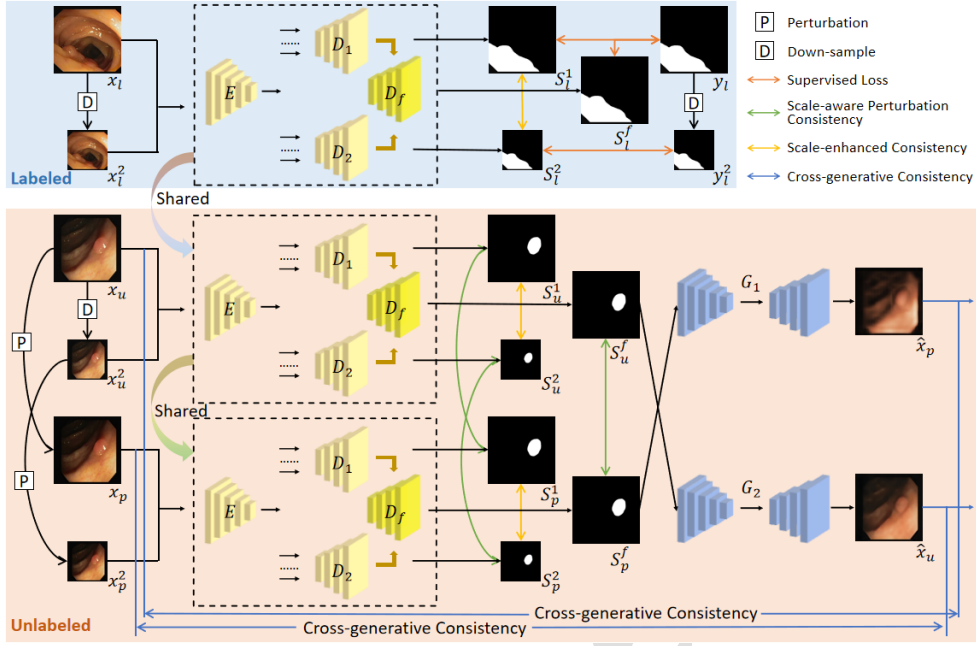
3

Figure 1: Architecture of our DEC-Seg framework. Labeled data are fed to the model to achieve the segmentation task under supervised learning. Unlabeled data and the corresponding perturbed versions are passed through the model, which is constrained under scale-enhanced consistency, scale-aware perturbation consistency, and cross-generative consistency. See Sec. 3 for details.

using cross-segmentation maps, which aims to mine effective feature representations and boost the segmentation performance.

- A dual-scale complementary fusion module is proposed to benifit the better prediction generation through the complementation and fusion of different scale information from the two scale-specific decoders.

The remainder of this paper is structured as follows. In Section 2, we briefly review some related works to our model, consisting of medical image segmentation, semi-supervised learning, and consistency learning. In Section 3, we give the overview framework and then provide the details of the proposed semi-supervised segmentation model. Then, we present the used datasets, experimental settings, comparison results, and ablation study in Section 4. Finally, we conclude this paper in Section 5.

4

2. Related Works

₇₅ *2.1. Medical Image Segmentation*

Recently, numerous studies have proposed different deep learning architectures and techniques to enhance the accuracy of medical image segmentation. The UNet framework, proposed by Ronneberger *et al.* [17], has significantly advanced the field of medical image segmentation. It incorporates skip connections between the contracting and expanding paths, enabling the

₈₀ network to retain fine-grained spatial information while also capturing high-level contextual features. UNet has demonstrated state-of-the-art performance across various segmentation tasks, including liver and tumor segmentation, cell segmentation, brain tumor segmentation, and polyp segmentation. To further improve the accuracy of UNet-based models, researchers have developed several variants, such as attention UNet [18], UNet++ [1], and ResUNet++ [19]. For

₈₅ example, Zhou *et al.* [1] presented UNet++ for medical image segmentation which designs a series of nested, dense skip pathways to reduce the semantic gap between the feature maps. Jha *et al.* [19] developed a ResUNet++ architecture, which takes advantage of residual blocks, squeeze and excitation blocks, atrous spatial pyramidal pooling, and attention blocks. Moreover, the cross-level and multi-level feature fusion technique is widely recognized for enhancing seg-

₉₀ mentation performance by integrating spatial information from low-level features with semantic information from high-level features. For instance, Li *et al.* [20] proposed a cross-level information processing module that aggregates and processes multi-scale features transmitted by different encoder layers to improve polyp segmentation. Zhou *et al.* [21] designed a cross-level feature fusion module to merge adjacent features from various levels, enabling the characterization of cross-level

₉₅ and multi-scale information to address scale variations in polyps. Additionally, CFU-Net [22] incorporates a multi-level attention module to exploit interactions of contextual information, decision information, and long-range dependencies, thereby enhancing feature propagation within skip connections.

2.2. Semi-Supervised Learning

₁₀₀ Existing SSL methods can be categorized into self-training, consistency learning, contrastive learning, and adversarial learning. Self-training [23, 24] methods use the predictions of the fully-supervised algorithm to act as the pseudo labels of the unlabeled data and retrain the model by mixing with the labeled data. Consistency learning methods [10, 13, 14, 25] enforce agreement between different views to improve the model's performance. By utilizing a consis-

₁₀₅ tency loss, these methods ensure that the model's predictions remain consistent across different

5

representations or augmentations of the same input. Contrastive learning methods [15, 26–29] obtain better representation learning by contrasting similar features against dissimilar features. Adversarial learning methods [30–32] employ a minimax game between a generator and a discriminator to train the model. By generating new and realistic samples, these methods enhance
110 the model's ability to generalize and produce accurate segmentation. Several SSL methods have been developed for the medical image segmentation task. For example, Li *et al.* [25] proposed a segmentation model with multiple auxiliary decoders and encouraged the consistency of the predictions made by the main decoder and the auxiliary decoders. Besides, contrastive learning [26, 33, 34] has been widely used in SSL medical image segmentation. Moreover, adversarial
115 learning [30, 31, 35] has also emerged to improve the robustness of the model. For example, Peiris *et al.* [30] equipped the network with a critic network to influence the segmentation network to produce the resemble prediction map as ground truth. Lei *et al.* [31] proposed double discriminators which are used to learn the prior relationship between labeled and unlabeled data. However, these semi-supervised methods do not deeply explore the exploitability and importance
120 of scale information for semi-supervised learning.

2.3. Consistency Learning

Consistency learning as one of the most popular methods in semi-supervised learning aims to force the network to learn potential knowledge from different predictions by applying consistency constraints, which may come from different views of the same input from the same network, or
125 may come from the same input from different networks. Currently, several methods [10, 13, 36–38] based on the mean teacher framework focus on designing different consistent and perturbation strategies to achieve SSL segmentation. For instance, Tarvainen *et al.* [10] proposed to generate two different augmented views of the same input and use the output of the teacher model to supervise the student model. Chen *et al.* [13] proposed cross-pseudo supervision by enforcing the
130 pseudo supervision between the predictions of two models with different parameter initialization. Verma *et al.* [36] regularized semi-supervised learning by encouraging consistent predictions at interpolations of different unlabeled points. Wu *et al.* [37] proposed three decoders and a different upsampling module in each decoder to amplify the perturbation and impose consistency across the three prediction maps. Zhong *et al.* [38] proposed a multi-attention tri-branch network
135 (MTNet) that each branch uses a different attention mechanism. Among these methods, the consistent constraint is a widely used strategy, aiming to ensure that perturbations applied to unlabeled data do not drastically alter their outputs.
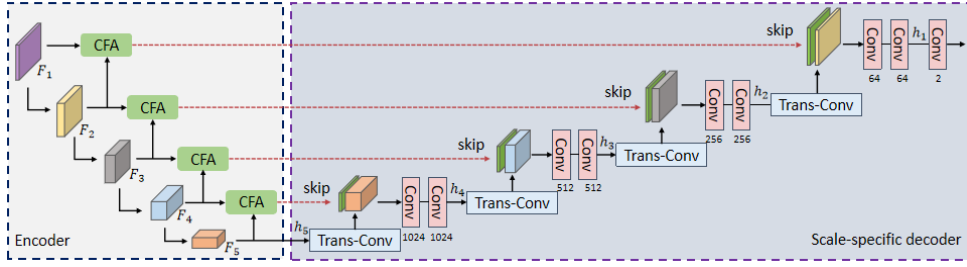
6

Figure 2: The detailed architectures of the encoder (*i.e.*, E in Fig. 1) and scale-specific decoder (*i.e.*, D_1 and D_2 in Fig. 1) networks. "Conv" denotes a sequential operation that consists of a 3×3 convolution, batch normalization, and a *ReLU* activation, and "Trans-Conv" denotes a transpose convolution.

3. Proposed Method

3.1. Overview

The proposed DEC-Seg framework, as depicted in Fig. 1, fully leverages multi-scale information from both labeled and unlabeled data for medical image segmentation. Our segmentation network comprises a shared encoder (E), two scale-specific decoders (D_1 processes features solely from the original scale while D_2 exclusively processes features from the downsampled scale), and a scale-fused decoder (D_f is used to fuse features from both scales). Additionally, two generative networks (G_1 and G_2) assist in the segmentation process, enhancing the feature mining ability of our segmentation network through a cross-generation strategy. Specifically, the original and downsampled images from the labeled data are fed into the model (*i.e.*, the encoder E and scale-specific decoders D_1 and D_2), and then the scale-enhanced consistency and supervised loss are calculated to constrain the network. Then, the unlabeled data and its perturbed versions are inputted into the model. Meanwhile, we conduct scale-enhanced consistency, scale-aware perturbation consistency, and cross-generative consistency for unlabeled data, to fully leverage unlabeled samples to improve the segmentation performance. Moreover, to make use of multi-scale information, we design the scale-fused decoder (*i.e.*, D_f) to generate the final segmentation maps. For convenience, we denote $X_l = \left\{x_l^1, x_l^2, \cdots, x_l^{n_l}\right\}$ and $Y_l = \left\{y_l^1, y_l^2, \cdots, y_l^{n_l}\right\}$ as the labeled dataset and the corresponding label set, respectively, where n_l is the number of the labeled images. We use $X_u = \left\{x_u^1, x_u^2, \cdots, x_u^{n_u}\right\}$ to denote the unlabeled dataset with n_u images, and we typically have $n_l \ll n_u$. Given inputs X, they are fed into the encoder to learn five levels of features, namely $\{F_i\}_{i=1}^5$.

7

3.2. Cross-level Feature Aggregation Module

To fully exploit cross-level information from various receptive fields and provide more effective supplementation from different convolutional layers, we propose the CFA module. This module fuses features from every two adjacent layers, enhancing the understanding of context and relationships within the cross-level features. The aggregated feature is then incorporated into our designed decoder, as depicted in Fig. 2. Specifically, as shown in Fig. 3, two adjacent features F_i and F_{i+1} first proceeded by a 1×1 convolution, and then the two features are concatenated (*i.e.*, F_{cat}) to obtain $F_{cat}^{'} = \mathcal{B}_{conv3 \times 3}(F_{cat})$, where $\mathcal{B}_{conv3 \times 3}(\cdot)$ is a sequential operation that consists of a 3×3 convolution, batch normalization, and a *ReLU* activation. To learn the cross-level attention-based enhanced feature, we conduct a global average pooling (GAP) on the cascaded feature $F_{cat}^{'}$ and utilize the point-wise convolution (PWC) [39] to capture channel interactions across different spatial positions. Therefore, we obtain the attention-based weights by

$$W = \sigma\big(\mathrm{Conv}_{pwc2}(\zeta(\mathrm{Conv}_{pwc1}(\mathcal{G}_{ave}(F_{cat}^{'})))) \big), \tag{1}$$

where the kernel sizes of Conv_{pwc1} and Conv_{pwc2} denote as $\frac{C}{r} \times C \times 1 \times 1$ and $C \times \frac{C}{r} \times 1 \times 1$ (C is channel size and r is a reduction ratio), respectively. Besides, $\zeta(\cdot)$ and $\sigma(\cdot)$ indicate *ReLU* and *Sigmoid* activation functions, respectively, and $\mathcal{G}_{ave}$ is a GAP operation. Next, an element-wise multiplication is utilized to enhance $F_{cat}^{'}$ with W, and then a residual structure is also adopted to fuse the enhanced feature and the original cascaded feature. Finally, we obtain the aggregated feature by

$$F_i^{\mathrm{CFA}} = \mathcal{B}_{conv3 \times 3}\big(F_{cat}^{'} \otimes W \oplus F_{cat}^{'}\big), \tag{2}$$

where $\oplus$ and $\otimes$ represent element-wise addition and multiplication, respectively. It is worth noting that our CFA module can explore contextual information from the diversity of resolutions to enhance the features' representation ability.

3.3. Scale-enhanced Consistency

The scale variation is still a great challenge for medical image segmentation. As discussed in [40], reducing the gap between the network outputs from different scale images can help the model learn scale-correlated features. Motivated by this observation, we propose a scale-enhanced consistency scheme to constrain the outputs of different scale images to be closed, in which the different scale features can be refined by each other. To increase the perturbation, we use two independent decoders for the single-scale prediction task, which are called scale-special decoders (*i.e.*, D_1 and D_2). Specifically, given unlabeled images X_u, X_u and its downsampled
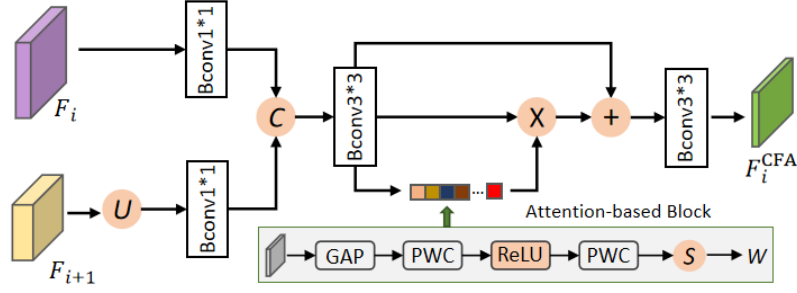
8

Figure 3: Illustrations of cross-level feature aggregation module. "GAP" denotes the global average pooling, "PWC" denotes the point-wise convolution, and "U" denotes the upsampling using bilinear interpolation.

versions $X_u^2 = Down(X_u)$ are fed to a weight-shared encoder to extract two sets of multi-level features, and then we can obtain the predicted maps S_u^1 and S_u^2, respectively. Specifically, the two segmentation maps can be obtained by $S_u^1 = D_1(E(X_u))$ and $S_u^2 = D_2(E(X_u^2))$, where $E(\cdot)$ is the encoder, and $Down(\cdot)$ denotes a $1/2$ downsampling operation. Furthermore, we employ the mutual supervision approach outlined in [13] to ensure the consistency of S_u^1 and S_u^2 with different sizes. Concretely, $\hat{Y}_u^1 = \arg\max_c(S_u^1)$ and $\hat{Y}_u^2 = \arg\max_c(S_u^2)$ are computed. Subsequently, $\hat{Y}_u^1$ is downsampled to the size of S_u^2 and utilized as a pseudo-label for S_u^2, while $\hat{Y}_u^2$ is upsampled to the size of S_u^1 and employed as a pseudo-label for S_u^1. Therefore, the scale-enhanced consistency loss for unlabeled data is defined by

$$\mathcal{L}_{SC}^u(S_u^1, S_u^2) = \mathcal{L}_{ce}(S_u^1, Up(\hat{Y}_u^2)) + \mathcal{L}_{ce}(S_u^2, Down(\hat{Y}_u^1)), \tag{3}$$

where $\mathcal{L}_{ce}$ denotes the widely used cross entropy loss and $Up(\cdot)$ denotes a $2\times$ upsampling operation. Similarly, for labeled images X_l, we can obtain the predicted segmentation maps S_l^1 and S_l^2 from X_l and its downsampled versions X_l^2 by $S_l^1 = D_1(E(X_l))$ and $S_l^2 = D_2(E(X_l^2))$, respectively. As a result, $\mathcal{L}_{SC}^l(S_l^1, S_l^2)$ can be obtained in the same way.

3.4. Dual-scale Complementary Fusion Module

The two independent decoders, as designed in Section 3.3, are specialized for processing information at two distinct scales. Throughout network training, these decoders acquire unique features based on the input image scale, with each emphasizing different aspects. Despite this distinction, the features from both scales pertain to the same unlabeled image, indicating the opportunity to combine and extract features from these two scale features further, potentially resulting in an enhanced segmentation map. Consequently, to fully leverage information from
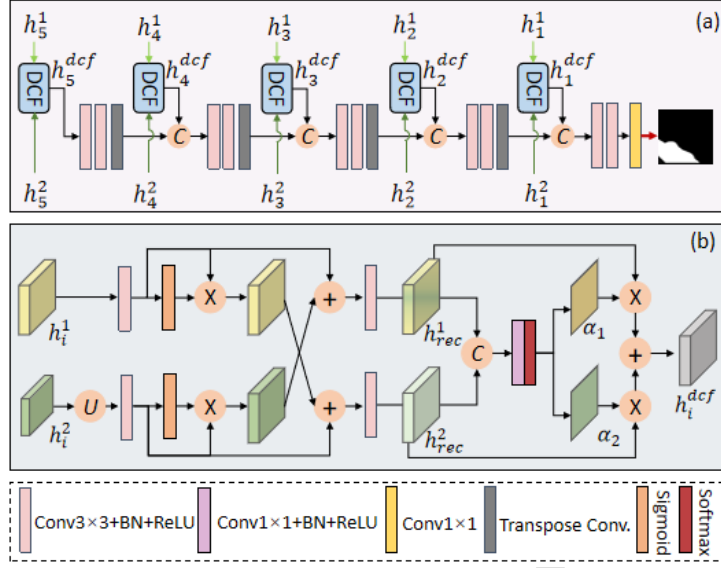
9

Figure 4: Illustration of (a) the architecture of the scale-fused decoder (*i.e.*, D_f in Fig. 1) and (b) the proposed dual-scale complementary fusion module. "U" denotes the upsampling using a bilinear interpolation.

different scales, in comparison to the scale-specific decoder, we also introduce the scale-fused decoder (D_f in Fig. 1) to produce more reliable predictions as the final segmentation maps.

As shown in Fig. 1, the features from D_1 and D_2 are fused and then fed into the scale-fused decoder. The feature flow of the scale-fused decoder is shown in Fig. 4(a), and the structure of the scale-fused decoder is basically the same as that of the scale-specific decoder. To achieve the fuse operation, we present a dual-scale complementary fusion (DCF) module to fuse the features from the two scale-specific decoders. This module replaces the features transmitted from the encoder through the skip operation. The proposed DCF module aims to fuse the multi-scale features and enable the features from the original scale and downsampled scale to complement each other (as shown in Fig. 4(b)). Taking the original scale feature as an example, we first conduct a 3×3 convolution on $Up(h_i^2)$ and Sigmoid activation function to obtain a scale-aware weight $W_2 = \sigma(\mathcal{B}_{conv3 \times 3}(Up(h_i^2)))$. Then, we multiply W_2 by $\mathcal{B}_{conv3 \times 3}(Up(h_i^2))$, and the resultant feature is the complementary information (*i.e.*, h_{com}^2) which is encouraged to learn the information required by the original scale. Finally, we add the complementary feature obtained at the downsampled scale to the original scale and then smooth the feature by a 3×3 convolution, thus we can obtain $h_{rec}^1 = \mathcal{B}_{conv3 \times 3}(\mathcal{B}_{conv3 \times 3}(h_i^1) + \mathcal{B}_{conv3 \times 3}(Up(h_i^2)) * W_2)$. Similarly,
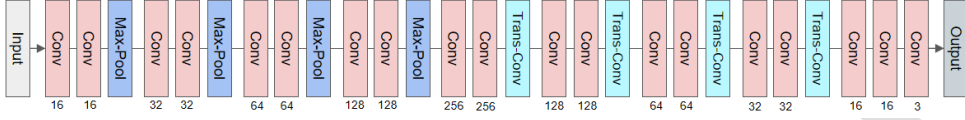
10

Figure 5: The detailed architecture of the generative network (*i.e.*, G_1 and G_2). "Conv" denotes a sequential operation that consists of a 3×3 convolution, batch normalization, and a *ReLU* activation, and "Trans-Conv" denotes a transpose convolution.

we can obtain the complementary-enhanced downsampled feature h^2_{rec}. Through a cross-scale complementary-enhanced process, we learn more rich multi-scale feature representations. Further, two features h^1_{rec} and h^2_{rec} are concatenated and then proceeded by a 1×1 convolution to two channels and a *Softmax* function to obtain two weight maps α_1 and α_2, where $\alpha_1 + \alpha_2 = 1$. Finally, we obtain the multi-scale fused feature by

$$h^{dcf}_i = h^1_{rec} \otimes \alpha_1 + h^2_{rec} \otimes \alpha_2. \tag{4}$$

Subsequently, the fused features h^{dcf}_i will concatenate with the features from the previous layer and continue to propagate forward. Specifically, h^{dcf}_5 is first passed through two convolutional layers and a transpose convolutional layer to have the same resolution with h^{dcf}_4. Then, the two features are concatenated and fed into the next two convolutional layers. Repeating the above process, we will obtain the final segmentation maps S^f. It is important to highlight that the scale-fused decoder plays a vital role in fully integrating multi-scale information, consequently enhancing the segmentation performance. Consequently, during the inference stage, the segmentation predictions obtained from the scale-fused decoder are taken as the final segmentation results.

3.5. Scale-aware Perturbation and Cross-generative Consistency

To constrain the predictions of unlabeled data on the scale-fused decoder, we follow previous works [25, 41, 42], and introduce scale-aware perturbation consistency to encourage the outputs between the original version and perturbed version to be closed. Specifically, we impose the perturbation consistency on different scale images, *i.e.*, $S^1_p = D_1(E(\mathcal{P}(X_u)))$, $S^2_p = D_2(E(\mathcal{P}(X^2_u)))$, and S^f_p, where a random color jitter operation is adopted as the perturbation operation (*i.e.*, $\mathcal{P}$). Thus, $\mathcal{L}^p_{SC}(S^1_p, S^2_p)$ is obtained and the scale-aware perturbation consistency scheme can be optimized by minimizing $\mathcal{L}_{SPC}$, which is defined by

$$\mathcal{L}_{SPC} = \mathcal{L}_{mse}(S^1_u, S^1_p) + \mathcal{L}_{mse}(S^2_u, S^2_p) + \mathcal{L}_{mse}(S^f_u, S^f_p). \tag{5}$$

11

Further, to utilize the rich information contained in different feature maps, we propose a perturbation-based cross-generative consistency constraint to enhance the feature mining ability of our segmentation network with information reconstruction and to correct the cognitive bias on a single input version with cross-generation. Therefore, we design two generative networks G_1 and G_2 (The detailed structure is shown in Fig. 5) for image reconstruction. For the generative network to accept richer information, produce better-reconstructed maps, and deliver more valuable feedback to the segmentation network, the logit predictions z_u^f and z_p^f obtained from the scale-fused decoder are fed into G_2 and G_1, respectively, and then we can obtain two reconstructed images $\hat{X}_p = G_1(z_p^f)$ and $\hat{X}_u = G_2(z_u^f)$. Finally, we form a cross-generative consistency loss ($\mathcal{L}_{CC}$), which is given by

$$\mathcal{L}_{CC} = \mathcal{L}_{mse}(\hat{X}_p, X_u) + \mathcal{L}_{mse}(\hat{X}_u, X_p). \tag{6}$$

It is worth noting that $\mathcal{L}_{CC}$ can enforce $\hat{X}_p$ and X_u to be consistent and $\hat{X}_u$ and X_p to be closed. Using the input image as a learning target provides unlabeled data with an authentic constraint containing structural and semantic information. This empowers the segmentation network to prioritize the structural and semantic details that are pertinent to category information during the generation of segmentation maps. The cross-generation process compels the network to investigate inconsistent predictions caused by color variations, thereby rectifying the network's cognitive biases toward a single input and ensuring the accuracy of predictions for both the original input and the perturbed input from different perspectives. Consequently, such cross-generation can further enhance the robustness of our model, enabling it to learn powerful features and effectively leverage the knowledge from unlabeled data to enhance the proposed segmentation model.

3.6. Overall Loss Function

In this study, the supervised loss for labeled data is the sum of the loss on the prediction maps of the designed three decoders and its corresponding ground truth, and it is formulated by

$$\mathcal{L}_S = \mathcal{L}_{sup}(S_l^1, Y_l) + \mathcal{L}_{sup}(S_l^2, Down(Y_l)) + \mathcal{L}_{sup}(S_l^f, Y_l), \tag{7}$$

where $\mathcal{L}_{sup} = (\mathcal{L}_{CE} + \mathcal{L}_{Dice})/2$, $\mathcal{L}_{CE}$ and $\mathcal{L}_{Dice}$ denote the cross-entropy (CE) loss and Dice loss, respectively. Finally, the total loss can be given as follows:

$$\mathcal{L}_{total} = \mathcal{L}_S + \mathcal{L}_{SPC} + \mathcal{L}_{SC} + \mathcal{L}_{CC}, \tag{8}$$

12

where $\mathcal{L}_{SC} = \mathcal{L}_{SC}^{l} + \mathcal{L}_{SC}^{u} + \mathcal{L}_{SC}^{p}$ is the sum of scale-enhanced consistency loss in labeled images,

200 unlabeled images, and perturbed images.

4. Experiments and Results

4.1. Datasets

Colonoscopy Datasets: We carry out the comparison experiments on four public datasets, *i.e.*, CVC-ColonDB [43], Kvasir [44], CVC-ClinicDB [45], and ETIS [46]. Following the setting
205 in [47], $1,450$ images are randomly selected from the two datasets (*i.e.*, CVC-ClinicDB and Kvasir) for the training set, and the remaining images from the two datasets and the other two datasets (*i.e.*, ETIS and CVC-ColonDB) to form the testing set. Besides, 10% (145 images) or 30% (435 images) from the training set are used as the labeled data, while the remaining images are adopted as the unlabeled data.

210 **Brain Tumor Dataset**: This dataset contains brain MR images together with manual FLAIR abnormality segmentation masks and it comes from 110 patients included in The Cancer Genome Atlas (TCGA) lower-grade glioma [48, 49]. In this study, we have $1,090$ images for the training set and 283 ones for the testing set. In addition, 10% and 30% images from the training set are adopted as the labeled data, respectively, while the remaining training images are taken
215 as unlabeled data.

Skin Lesion Dataset: The dataset comes from ISIC-2018 challenge [50] and is used for skin lesion segmentation. It includes $2,594$ dermoscopic images for training and $1,000$ images for testing. In the same way, 10% or 30% images from the training set are used as the labeled data, while the remaining images are formed as the unlabeled set.

220 ### 4.2. Implementation Details and Evaluation Metrics

Implementation Details: The proposed segmentation framework is implemented in PyTorch and trained on one NVIDIA GeForce RTX3090 GPU. The input images from the polyp segmentation dataset and skin lesion segmentation dataset are uniformly rescaled to 352×352, and the inputs of the brain MRI segmentation dataset are uniformly rescaled to 256×256. Our
225 model converges over $10,000$ iterations with a batch size of 6 including 3 labeled samples and 3 unlabeled samples, and uses an SGD optimizer with an initial learning rate set to $1e$-2 and a poly learning rate strategy to update the learning rate.

13

Table 1: Quantitative results on CVC-ClinicDB and Kvasir datasets with 10% or 30% labeled data.

Methods		10% labeled					30% labeled				
		mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	MAE ↓	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	MAE ↓
CVC-ClinicDB	MT [10]	0.747	0.662	0.741	0.830	0.028	0.772	0.689	0.763	0.852	0.028
	DAN [32]	0.755	0.674	0.733	0.842	0.027	0.775	0.696	0.772	0.848	0.027
	UA-MT [11]	0.749	0.676	0.742	0.839	0.027	0.790	0.715	0.787	0.853	0.025
	URPC [14]	0.769	0.696	0.760	0.854	0.021	0.777	0.709	0.785	0.853	0.026
	CLCC [15]	0.794	0.720	0.786	0.859	0.027	0.840	0.786	0.840	0.891	0.019
	SLC-Net [16]	0.752	0.689	0.735	0.844	0.027	0.835	0.774	0.830	0.888	0.020
	MC-Net+ [37]	0.767	0.702	0.751	0.856	0.025	0.845	0.783	0.844	0.891	0.020
	CDMA [38]	0.759	0.676	0.750	0.837	0.025	0.803	0.728	0.802	0.863	0.024
	SCP-Net [51]	0.776	0.703	0.758	0.854	0.024	0.839	0.786	0.837	0.888	0.020
	MCF [52]	0.779	0.718	0.767	0.860	0.021	0.823	0.766	0.819	0.882	0.019
	DEC-Seg (ours)	**0.836**	**0.774**	**0.834**	**0.886**	**0.019**	**0.859**	**0.804**	**0.860**	**0.900**	**0.015**
Kvasir	MT [10]	0.814	0.722	0.797	0.841	0.050	0.814	0.732	0.790	0.852	0.052
	DAN [32]	0.808	0.723	0.786	0.839	0.059	0.841	0.760	0.828	0.866	0.043
	UA-MT [11]	0.799	0.713	0.782	0.834	0.058	0.845	0.771	0.840	0.873	0.037
	URPC [14]	0.811	0.728	0.796	0.842	0.057	0.849	0.778	0.846	0.874	0.040
	CLCC [15]	0.806	0.724	0.796	0.844	0.052	0.864	0.804	0.853	0.892	0.034
	SLC-Net [16]	0.840	0.773	0.830	0.868	0.042	0.867	0.805	0.856	0.890	0.034
	MC-Net+ [37]	0.817	0.735	0.807	848	0.051	0.831	0.765	0.816	0.862	0.044
	CDMA [38]	0.786	0.695	0.769	0.826	0.056	0.839	0.758	0.826	0.862	0.044
	SCP-Net [51]	0.810	0.723	0.790	0.840	0.056	0.862	0.789	0.847	0.878	0.037
	MCF [52]	0.822	0.751	0.816	0.852	0.051	0.856	0.794	0.847	0.882	0.039
	DEC-Seg (ours)	**0.859**	**0.787**	**0.853**	**0.877**	**0.044**	**0.893**	**0.830**	**0.886**	**0.902**	**0.032**

Evaluation Metrics: To evaluate the effectiveness, we employ five commonly adopted metrics [19, 47], namely mean Dice (mDice), mean IoU (mIoU), F_β^w, S_α, and mean absolute error (MAE).

4.3. Comparison with State-of-the-arts

We compare the proposed DEC-Seg with ten state-of-the-art semi-supervised segmentation methods, *i.e.*, MT [10], DAN [32], UA-MT [11], URPC [14], CLCC [15], SLC-Net [16], MC-Net+ [37], CDMA [38], SCP-Net [51] and MCF [52]. For a fair comparison, we change the backbones of all compared methods to "Res2Net", and then train all compared methods without

14

Table 2: Quantitative results on two unseen datasets (CVC-ColonDB and ETIS) with 10% or 30% labeled data.

Methods		10% labeled					30% labeled				
		mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	MAE ↓	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	MAE ↓
CVC-ColonDB	MT [10]	0.589	0.497	0.581	0.743	0.048	0.600	0.510	0.597	0.749	0.045
	DAN [32]	0.620	0.517	0.604	0.756	0.046	0.633	0.546	0.630	0.767	0.045
	UA-MT [11]	0.553	0.471	0.551	0.727	0.051	0.648	0.563	0.635	0.771	0.046
	URPC [14]	0.556	0.480	0.549	0.732	0.047	0.598	0.519	0.601	0.754	0.045
	CLCC [15]	0.538	0.473	0.533	0.720	0.053	0.564	0.505	0.562	0.738	0.050
	SLC-Net [16]	0.595	0.525	0.582	0.755	0.044	0.653	0.582	0.644	0.783	0.041
	MC-Net+ [37]	0.562	0.486	0.543	0.725	0.055	0.589	0.518	0.578	0.745	0.050
	CDMA [38]	0.507	0.419	0.494	0.691	0.054	0.624	0.536	0.609	0.756	0.047
	SCP-Net [51]	0.577	0.495	0.559	0.739	0.049	0.694	0.615	0.683	0.802	0.039
	MCF [52]	0.566	0.498	0.559	0.742	**0.045**	0.621	0.550	0.614	0.768	0.042
	DEC-Seg (ours)	**0.648**	**0.565**	**0.630**	**0.771**	0.046	**0.721**	**0.640**	**0.709**	**0.814**	**0.035**
ETIS	MT [10]	0.356	0.288	0.337	0.636	0.035	0.495	0.420	0.462	0.706	0.039
	DAN [32]	0.437	0.358	0.397	0.666	0.046	0.485	0.418	0.472	0.708	0.027
	UA-MT [11]	0.459	0.393	0.441	0.693	0.029	0.541	0.473	0.523	0.738	0.024
	URPC [14]	0.420	0.356	0.406	0.677	0.028	0.559	0.490	0.549	0.749	0.021
	CLCC [15]	0.409	0.346	0.395	0.668	0.031	0.474	0.422	0.465	0.707	0.020
	SLC-Net [16]	0.431	0.374	0.409	0.682	0.030	0.566	0.503	0.543	0.753	0.024
	MC-Net+ [37]	0.439	0.369	0.403	0.684	0.035	0.535	0.470	0.505	0.735	0.026
	CDMA [38]	0.309	0.254	0.296	0.611	0.031	0.531	0.446	0.494	0.724	0.024
	SCP-Net [51]	0.394	0.323	0.355	0.656	0.045	0.571	0.500	0.543	0.747	0.023
	MCF [52]	0.425	0.367	0.407	0.679	**0.025**	0.487	0.425	0.469	0.707	0.022
	DEC-Seg (ours)	**0.592**	**0.511**	**0.558**	**0.758**	**0.025**	**0.634**	**0.564**	**0.608**	**0.789**	**0.019**

any data augmentation. All the compared methods undergo 10,000 iterations, similar to our approach, utilizing the same data and image dimensions. In the case of CLCC, we maintain its original setting of 320×320 due to the specificity of the method. Moreover, the optimizer, learning rate, and descent strategy all remain consistent with their original configurations across
240   the compared methods.

4.3.1. Experiments on Polyp Segmentation

Quantitative comparisons are reported in Table 1 and Table 2. To validate the learning ability of the two seen training datasets (CVC-ClinicDB and Kvasir), it can be seen from Table 1

15

Table 3: Quantitative results on Brain MRI dataset with 10% or 30% labeled data.

Methods	10% labeled					30% labeled				
	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	MAE ↓	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	MAE ↓
MT [10]	0.491	0.404	0.512	0.705	0.015	0.674	0.561	0.704	0.792	0.012
DAN [32]	0.534	0.429	0.564	0.719	0.015	0.703	0.582	0.737	0.803	0.011
UA-MT [11]	0.549	0.435	0.584	0.721	0.015	0.703	0.586	0.736	0.807	0.011
URPC [14]	0.547	0.448	0.571	0.729	0.014	0.653	0.548	0.677	0.785	0.012
CLCC [15]	0.595	0.505	0.615	0.760	**0.012**	0.728	0.615	0.754	0.824	0.010
SLC-Net [16]	0.483	0.400	0.506	0.704	0.015	0.702	0.591	0.714	0.807	0.010
MC-Net+ [37]	0.580	0.481	0.595	0.749	0.015	0.746	0.640	0.766	0.834	0.009
CDMA [38]	0.495	0.408	0.508	0.705	0.015	0.627	0.526	0.647	0.775	0.011
SCP-Net [51]	0.559	0.467	0.571	0.741	0.014	0.643	0.544	0.660	0.783	0.012
MCF [52]	0.589	0.491	0.623	0.750	0.013	0.704	0.612	0.722	0.819	0.010
DEC-Seg (ours)	**0.626**	**0.533**	**0.639**	**0.777**	0.012	**0.767**	**0.671**	**0.783**	**0.850**	**0.008**

Table 4: Quantitative results on ISIC-2018 dataset with 10% or 30% labeled data.

Methods	10% labeled					30% labeled				
	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	MAE ↓	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	MAE ↓
MT [10]	0.822	0.738	0.809	0.811	0.104	0.844	0.760	0.822	0.825	0.087
DAN [32]	0.831	0.744	0.817	0.815	0.097	0.842	0.757	0.825	0.824	0.091
UA-MT [11]	0.839	0.752	0.818	0.820	0.094	0.845	0.762	0.829	0.829	0.086
URPC [14]	0.847	0.763	0.829	0.831	0.089	0.853	0.772	0.835	0.880	0.082
CLCC [15]	0.842	0.756	0.815	0.820	0.088	0.845	0.762	0.816	0.821	0.092
SLC-Net [16]	0.843	0.754	0.818	0.822	0.091	0.846	0.752	0.811	0.821	0.085
MC-Net+ [37]	0.848	0.767	0.829	0.832	0.087	0.862	0.782	0.841	0.842	0.076
CDMA [38]	0.858	0.777	0.842	0.839	0.083	0.868	0.788	0.848	0.846	0.073
SCP-Net [51]	0.852	0.770	0.835	0.835	0.085	0.853	0.769	0.831	0.834	0.083
MCF [52]	0.856	0.773	0.837	0.836	0.083	0.861	0.778	0.839	0.840	0.077
DEC-Seg (ours)	**0.867**	**0.796**	**0.858**	**0.851**	**0.078**	**0.875**	**0.801**	**0.862**	**0.856**	**0.069**

that while most methods learn well, our method performs better than other compared methods, both on 10% and 30% of labeled data, which indicate that our DEC-Seg can fully leverage unlabeled data to improve the polyp segmentation performance. Additionally, we report the comparison results on two unseen datasets (ETIS, and CVC-ColonDB) from Table 2 to verify the generalization ability of our DEC-Seg. The CVC-ColonDB and ETIS datasets are difficult,
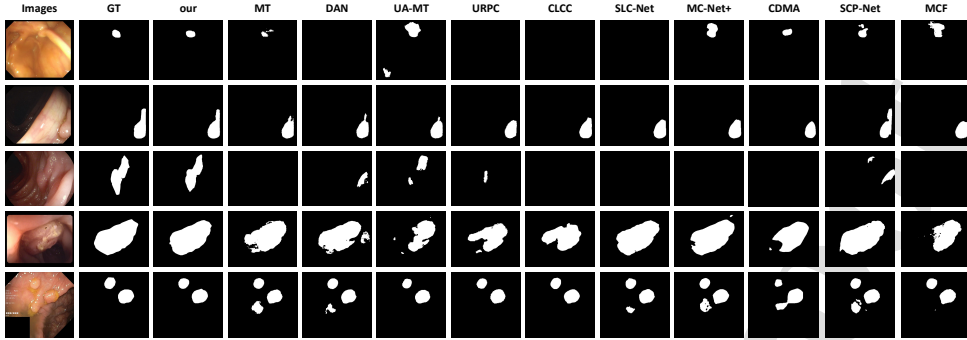
16

Figure 6: Qualitative results of our model and other ten compared semi-supervised segmentation methods on polyp segmentation datasets.

as polyps are with large variations in size. As reported in Table 2, comparing the SCP-Net
method on the CVC-ColonDB dataset using 30% labeled data, our method achieves 3.9% and
4.1% improvements on mDice and mIoU, respectively. Similarly, on the ETIS dataset, our
performance has a more significant superiority over other methods. The primary reason is that
our exploration and utilization of scale information enhance the model's robustness to scale
variations. This capability allows it to small polyps, while other methods cannot identify them
accurately. Overall, the results indicate that our DEC-Seg has a better generalization ability
and effectively utilizes unlabeled data to boost polyp segmentation.

Fig. 6 presents some of the visualization results of the test examples under 30% labeled
data. It can be seen that our method can accurately locate and segment polyps under different
challenging factors. For example, in the first row of Fig. 6, the polyps have very small sizes. It can
be seen that DAN, URPC, CLCC, and SLC-Net methods fail to locate polyps. Other methods
produce over-segmented fragments and confuse the edges of polyps, while our method accurately
and completely segments polyps. In the 3rd row, the polyps are visually embedded in their
surrounding mucosa, thus it is very difficult to accurately locate and segment these polyps. From
the results, our method performs better than other comparison methods to accurately segment
polyps. In the 4th row, the polyps have relatively large sizes, making it challenging to complete
locate the polyps. In this case, some methods (*e.g.*, UA-MT, URPC, CLCC, CDMA, and MCF)
only locate some fragments of polyps, while our method can obtain promising segmentation
results and produce fine details of the boundary. This is mainly because our method makes
use of multi-scale information and cross-generative consistency to learn more powerful feature
representations, which help boost the segmentation performance. In addition, we also show
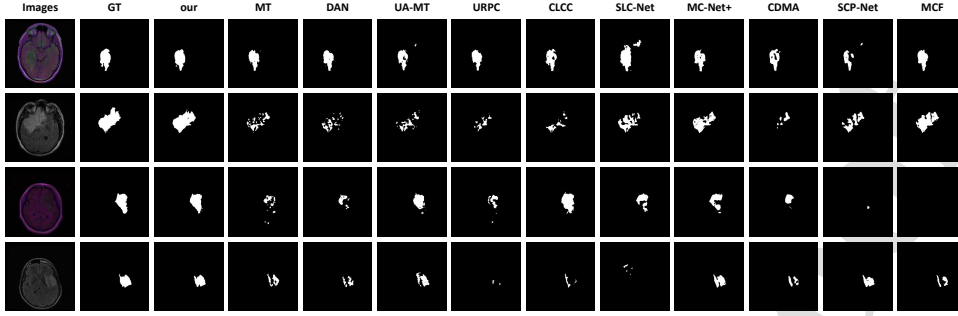
17

Figure 7: Qualitative results of our model and other ten compared semi-supervised segmentation methods on brain tumor segmentation dataset.
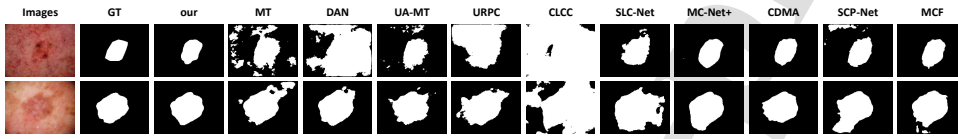

Figure 8: Qualitative results of our model and other ten compared semi-supervised segmentation methods on skin lesion segmentation dataset.

some segmentation results containing multiple polyps (see the 5th row). Compared with other methods, the segmentation of multiple polyps by our method is more complete and accurate. It can be also observed that our method can effectively locate and segment polyps under different challenging factors, such as scale variation, homogeneous regions, non-sharp boundaries, and multiple polyps.

4.3.2. Experiments on Brain Tumor Segmentation

As shown in Table 3, our DEC-Seg achieves significant improvement in five metrics, and compared to the second best, our method achieves 5.2% and 5.5% improvements in terms of mDice and mIou with 10% labeled data, and the corresponding improvements are 2.8% and 4.8% with 30% labeled data. Moreover, it can be seen in Fig. 7 that our method can segment the glioma completely and coherently. In the 2rd row, the segmentation maps of other methods are discrete, while our method produces more coherent segmented fragments. In addition, in the 4th row, some methods (*e.g.*, URPC, CLCC, and SLC-Net) do not find glioma regions, while our method accurately locates and segments glioma regions.

18

Table 5: Ablative results on the Kvasir and CVC-ColonDB datasets with 30% labeled data.

Settings	Kvasir				CVC-ColonDB			
	mDice	mIoU	F_β^w	S_α	mDice	mIoU	F_β^w	S_α
(No.1) Baseline	0.850	0.778	0.829	0.873	0.612	0.539	0.600	0.760
(No.2) Baseline + SC	0.873	0.812	0.865	0.894	0.637	0.572	0.632	0.773
(No.3) Baseline + SC + DCF	0.882	0.818	0.874	0.892	0.675	0.599	0.658	0.790
(No.4) Baseline + SC + DCF + CC	0.888	0.827	0.882	0.900	0.702	0.626	0.694	0.807
(No.5) Baseline + SC + DCF + CFA	0.888	0.824	0.878	0.898	0.698	0.621	0.688	0.804
(No.6) Baseline + SC + DCF + CC + CFA (ours)	**0.893**	**0.830**	**0.886**	**0.902**	**0.721**	**0.640**	**0.709**	**0.814**

²⁸⁵ *4.3.3. Experiments on Skin Lesion Segmentation*

Table 4 shows the performance on skin lesion segmentation dataset with 10% and 30% labeled ratios. DEC-Seg surpasses all state-of-the-arts. Our method achieves 1.3% and 3.0% improvements over MCF in terms of mDice and mIou with 10% labeled data, and the corresponding improvements are 1.6% and 3.0% with 30% labeled data. Besides, Fig. 8 shows the visual segmentation results in the skin lesion segmentation dataset. Compared with other methods, our DEC-Seg achieves better and more accurate segmentation in blurred boundaries.

4.4. Ablation Study

To verify the effectiveness of each key component in DEC-Seg, we conduct ablation studies with 30% labeled data on Kvasir and CVC-ColonDB datasets. The ablative results are shown ²⁹⁵ in Table 5, where "Baseline" denotes the semi-supervised framework with only scale-aware perturbation consistency.

Effectiveness of SC. To investigate the importance of scale-enhanced consistency (SC), we add the SC loss to encourage the consistency of predictions from the same inputs with different scales. From Table 5, we observe that No.2 (Baseline + SC) outperforms No.1 and obtains a ³⁰⁰ 4.1% improvement in mean Dice. This result indicates that scale-enhanced consistency is very helpful in improving polyp segmentation performance and is robust to scale variation.

Effectiveness of DCF. As shown in Table 5, No.3 (using the proposed DCF module) outperforms No.2 on two datasets. This indicates that the fusion of multi-scale features can further improve the segmentation performance. In addition, to further validate the effectiveness ³⁰⁵ of the dual-scale complement fusion strategy in the proposed DCF module, we construct a "Basic" strategy, which conducts a concatenation operation followed by two convolution layers to integrate the two features from different scales. The comparison results are shown in Table

19

Table 6: Ablative results of our DCF module and a basic fusion strategy on the Kvasir and CVC-ColonDB datasets. The labeled ratio is set to 30%.

Settings	Kvasir			CVC-ColonDB		
	mDice	mIoU	F_β^w	mDice	mIoU	F_β^w
Basic	0.885	0.823	0.878	0.710	0.633	0.703
DCF	**0.893**	**0.830**	**0.886**	**0.721**	**0.640**	**0.709**

6. From Table 6, it can be observed that our DCF performs better than the "Basic" strategy, indicating the effectiveness of the designed dual-scale complementary fusion module. Moreover, we visualize the segmentation results by using three different decoders, *i.e.*, the scale-specific and scale-fused decoders, and the comparison results are shown in Fig. 9. It can be observed that our method can not accurately locate the boundaries of the polyps when only using the original scale features or the downsampled scale features. However, we integrate the features from the two scales and then propagate them into the scale-fused decoder, which can produce more accurate segmentation maps (as shown in Fig. 9 (c)). This further confirms that integrating features from different scales enhances segmentation performance.

Effectiveness of CC. We further study the contributions of cross-generative consistency (CC). As shown in Table 5, it can be seen that No.4 improves the No.3 performance on the CVC-ColonDB dataset, as the mean Dice is improved from 0.675 to 0.702. Therefore, these improvements indicate that introducing cross-generative consistency loss can help accurately segment polyp tissues in the learning of details and textures.

Effectiveness of CFA. We then examine the significance of the proposed CFA module. To do this, we integrate the CFA into configurations No.3 and No.4 by aggregating the features of adjacent layers in the encoder before passing them to the decoder. As shown in Table 5, No.5 outperforms No.3, and No.6 shows a substantial improvement over No.4 on the CVC-ColonDB dataset, which has significant scale variation. This demonstrates that the CFA module effectively captures scale information and enhances segmentation performance.

4.5. Limitation and Future Work

The qualitative and quantitative evaluations confirm the effectiveness and superiority of our model. However, DEC-Seg encounters challenges in accurately segmenting polyps when dealing with ambiguous areas and unclear boundaries. Some failure cases generated by our model are
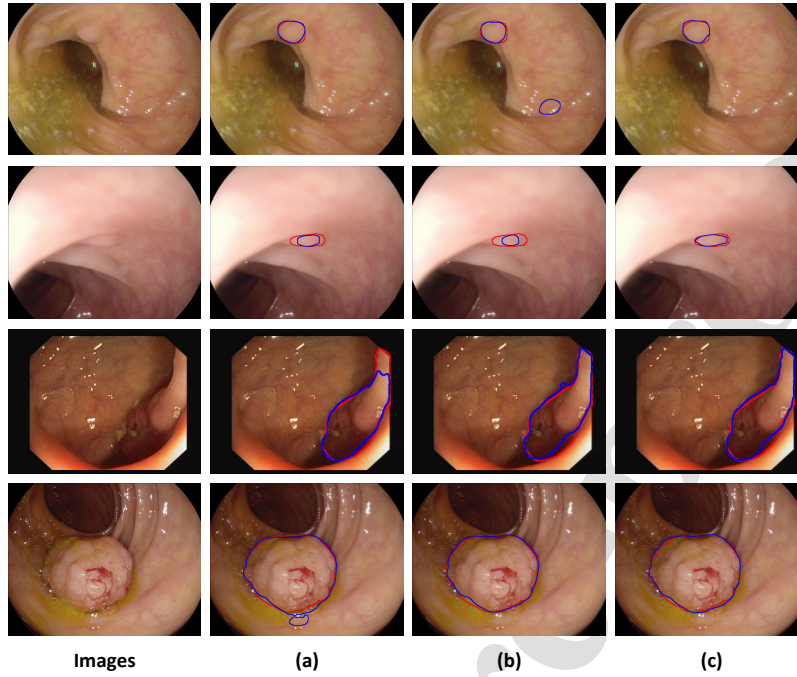
20

Figure 9: Visualization results of the three decoders. (a) predictions of the decoder (*i.e.*, D_1) using the original scale features; (b) predictions of the decoder (*i.e.*, D_2) using the downsampled scale features; (c) predictions of the scale-fused decoders (*i.e.*, D_f) using the multi-scale integrated features. The red and blue lines denote the ground truth and predictions, respectively.

depicted in Fig. 10. In the 1st and 2nd rows, while the prominent polyp regions are segmented accurately, there are inaccuracies in parts areas of the segmentation. In the 3rd and 4th rows, it
335   is evident that the polyp regions are notably small and masked within the background, featuring very unclear boundaries, which presents a significant challenge in accurately segmenting the polyp regions. It is apparent that under these circumstances, our DEC-Seg struggles to identify and segment the polyps. Therefore, dealing with ambiguous areas and segmenting polyps with unclear boundaries is desired to be investigated in future work. Additionally, unlike a simple
340   UNet framework, our model may have higher complexity due to incorporating three decoders and two generative networks. Furthermore, our model has only been validated on 2D segmentation tasks. Moving forward, we intend to address and overcome these limitations in our future research endeavors.
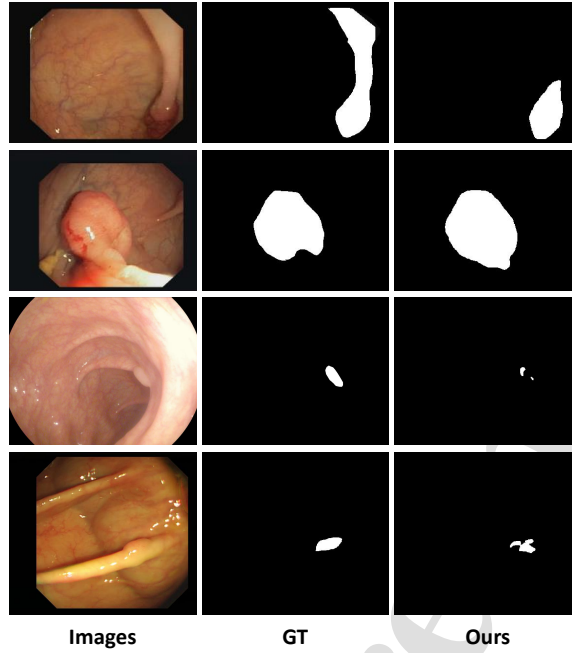
Figure 10: Some failure cases generated by our DEC-Seg.

5. Conclusion

345   In this paper, we have presented a novel semi-supervised learning framework (DEC-Seg) for medical image segmentation. The proposed cross-level feature aggregation module integrates the adjacent features from different resolutions, to enhance the features' representation ability. Then, a scale-enhanced consistency is proposed to handle scale variation and learn more scale-aware features. Meanwhile, we design the scale-fused decoders and a dual-scale complementary

350 fusion module to aggregate the features from the scale-specific decoders and produce the final segmentation maps. Moreover, multiple consistency strategies, *i.e.*, scale-aware perturbation consistency and cross-generative consistency, are presented to enhance the learning process and fully leverage unlabeled data to boost the segmentation performance. Experimental results on multiple datasets from three medical image segmentation tasks show that our DEC-Seg is superior

355 to state-of-the-art semi-supervised segmentation methods.

22

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.

Acknowledgment

References

[1] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Transactions on Medical Imaging 39 (6) (2019) 1856–1867.

[2] T. Zhou, Y. Zhou, G. Li, G. Chen, J. Shen, Uncertainty-aware hierarchical aggregation network for medical image segmentation, IEEE Transactions on Circuits and Systems for Video Technology 2024.

[3] G. Yue, W. Han, B. Jiang, T. Zhou, R. Cong, T. Wang, Boundary constraint network with cross layer feature integration for polyp segmentation, IEEE Journal of Biomedical and Health Informatics 26 (8) (2022) 4090–4099.

[4] T. Lei, H. Liu, Y. Wan, C. Li, Y. Xia, A. K. Nandi, Shape-guided dual consistency semi-supervised learning framework for 3-d medical image segmentation, IEEE Transactions on Radiation and Plasma Medical Sciences 7 (7) (2023) 719–731.

[5] K. Wang, B. Zhan, C. Zu, X. Wu, J. Zhou, L. Zhou, Y. Wang, Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning, Medical Image Analysis 79 (2022) 102447.

[6] Y. Meng, H. Zhang, Y. Zhao, D. Gao, B. Hamill, G. Patri, T. Peto, S. Madhusudhan, Y. Zheng, Dual consistency enabled weakly and semi-supervised optic disc and cup segmentation with dual adaptive graph convolutional networks, IEEE Transactions on Medical Imaging 42 (2) (2022) 416–429.

23

[7] X. Zhao, Z. Wu, S. Tan, D.-J. Fan, Z. Li, X. Wan, G. Li, Semi-supervised spatial temporal attention network for video polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 456–466.

385 [8] K. Han, L. Liu, Y. Song, Y. Liu, C. Qiu, Y. Tang, Q. Teng, Z. Liu, An effective semi-supervised approach for liver CT image segmentation, IEEE Journal of Biomedical and Health Informatics 26 (8) (2022) 3999–4007.

[9] L. Qiu, J. Cheng, H. Gao, W. Xiong, H. Ren, Federated semi-supervised learning for medical image segmentation via pseudo-label denoising, IEEE Journal of Biomedical and Health 390 Informatics 27 (10) (2023) 4672–4683.

[10] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Advances in Neural Information Processing Systems, 2017.

[11] L. Yu, S. Wang, X. Li, C.-W. Fu, P.-A. Heng, Uncertainty-aware self-ensembling model 395 for semi-supervised 3D left atrium segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 605–613.

[12] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, Y. Bengio, D. Lopez-Paz, Interpolation consistency training for semi-supervised learning, Neural Networks 145 (2022) 90–106.

[13] X. Chen, Y. Yuan, G. Zeng, J. Wang, Semi-supervised semantic segmentation with cross 400 pseudo supervision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 2613–2622.

[14] X. Luo, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, S. Zhang, Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency, in: International Conference on Medical Image 405 Computing and Computer-Assisted Intervention, Springer, 2021, pp. 318–329.

[15] X. Zhao, C. Fang, D.-J. Fan, X. Lin, F. Gao, G. Li, Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation, in: IEEE International Symposium on Biomedical Imaging, 2022, pp. 1–5.

[16] J. Liu, C. Desrosiers, Y. Zhou, Semi-supervised medical image segmentation using cross-410 model pseudo-supervision with shape awareness and local context constraints, in: Inter-

24

national Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 140–150.

[17] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[18] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. Mc-Donagh, N. Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999.

[19] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H. D. Johansen, Resunet++: An advanced architecture for medical image segmentation, in: IEEE International Symposium on Multimedia, 2019, pp. 225–2255.

[20] W. Li, Z. Huang, F. Li, Y. Zhao, H. Zhang, CIFG-Net: Cross-level information fusion and guidance network for polyp segmentation, Computers in Biology and Medicine 169 (2024) 107931.

[21] T. Zhou, Y. Zhou, K. He, C. Gong, J. Yang, H. Fu, D. Shen, Cross-level feature aggregation network for polyp segmentation, Pattern Recognition 140 (2023) 109555.

[22] H. Yin, Y. Shao, CFU-Net: A coarse-fine u-net with multi-level attention for medical image segmentation, IEEE Transactions on Instrumentation and Measurement 42 (2023) 5020412.

[23] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images, IEEE Transactions on Medical Imaging 39 (8) (2020) 2626–2637.

[24] S. Qiao, W. Shen, Z. Zhang, B. Wang, A. Yuille, Deep co-training for semi-supervised image recognition, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 135–152.

[25] Y. Li, L. Luo, H. Lin, H. Chen, P.-A. Heng, Dual-consistency semi-supervised learning with uncertainty quantification for COVID-19 lesion segmentation from CT images (2021) 199–209.

25

[26] H. Wu, Z. Wang, Y. Song, L. Yang, J. Qin, Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 11666–11675.

[27] J. Shi, T. Gong, C. Wang, C. Li, Semi-supervised pixel contrastive learning framework for tissue segmentation in histopathological image, IEEE Journal of Biomedical and Health Informatics 27 (1) (2022) 97–108.

[28] X. Zhao, Z. Qi, S. Wang, Q. Wang, X. Wu, Y. Mao, L. Zhang, RCPS: Rectified contrastive pseudo supervision for semi-supervised medical image segmentation, arXiv preprint arXiv:2301.05500.

[29] X. Hu, D. Zeng, X. Xu, Y. Shi, Semi-supervised contrastive learning for label-efficient medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 481–490.

[30] H. Peiris, Z. Chen, G. Egan, M. Harandi, Duo-segnet: adversarial dual-views for semi-supervised medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 428–438.

[31] T. Lei, D. Zhang, X. Du, X. Wang, Y. Wan, A. K. Nandi, Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network, IEEE Transactions on Medical Imaging 42 (5) (2022) 1265–1277.

[32] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, D. Z. Chen, Deep adversarial networks for biomedical image segmentation utilizing unannotated images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 408–416.

[33] J. Chen, C. Chen, W. Huang, J. Zhang, K. Debattista, J. Han, Dynamic contrastive learning guided by class confidence and confusion degree for medical image segmentation, Pattern Recognition 145 (2024) 109881.

[34] C. Tang, X. Zeng, L. Zhou, Q. Zhou, P. Wang, X. Wu, H. Ren, J. Zhou, Y. Wang, Semi-supervised medical image segmentation via hard positives oriented contrastive learning, Pattern Recognition 146 (2024) 110020.

26

[35] C. Chen, K. Zhou, Z. Wang, R. Xiao, Generative consistency for semi-supervised cerebrovascular segmentation from TOF-MRA, IEEE Transactions on Medical Imaging 42 (2) (2022) 346–353.

[36] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, D. Lopez-Paz, Interpolation consistency training for semi-supervised learning, Neural Networks 145 (2022) 90–106.

[37] Y. Wu, Z. Ge, D. Zhang, M. Xu, L. Zhang, Y. Xia, J. Cai, Mutual consistency learning for semi-supervised medical image segmentation, Medical Image Analysis 81 (2022) 102530.

[38] L. Zhong, X. Liao, S. Zhang, G. Wang, Semi-supervised pathological image segmentation via cross distillation of multiple attentions, in: MICCAI, Springer, 2023, pp. 570–579.

[39] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, K. Barnard, Attentional feature fusion, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision., 2021, pp. 3560–3569.

[40] Y. Wang, J. Zhang, M. Kan, S. Shan, X. Chen, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 12272–12281.

[41] L. Wang, J. Wang, L. Zhu, H. Fu, P. Li, G. Cheng, Z. Feng, S. Li, P.-A. Heng, Dual multiscale mean teacher network for semi-supervised infection segmentation in chest ct volume for covid-19, IEEE Transactions on Cybernetics 53 (10) (2022) 6363–6375.

[42] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, G. J. Brostow, Harmonic networks: Deep translation and rotation equivariance, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5028–5037.

[43] N. Tajbakhsh, S. R. Gurudu, J. Liang, Automated polyp detection in colonoscopy videos using shape and context information, IEEE Transactions on Medical Imaging 35 (2) (2015) 630–644.

[44] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: MultiMedia Modeling, Springer, 2020, pp. 451–462.

27

[45] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, Computerized Medical Imaging and Graphics 43 (2015) 99–111.

[46] J. Silva, A. Histace, O. Romain, X. Dray, B. Granado, Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer, International Journal of Computer Assisted Radiology and Surgery 9 (2) (2014) 283–293.

[47] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 263–273.

[48] M. Buda, A. Saha, M. A. Mazurowski, Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm, Computers in Biology and Medicine 109 (2019) 218–225.

[49] M. A. Mazurowski, K. Clark, N. M. Czarnek, P. Shamsesfandabadi, K. B. Peters, A. Saha, Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the cancer genome atlas data, Journal of Nuro-oncology 133 (2017) 27–35.

[50] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration, arXiv preprint arXiv:1902.03368.

[51] Z. Zhang, R. Ran, C. Tian, H. Zhou, X. Li, F. Yang, Z. Jiao, Self-aware and cross-sample prototypical learning for semi-supervised medical image segmentation, in: MICCAI, Springer, 2023, pp. 192–201.

[52] Y. Wang, B. Xiao, X. Bi, W. Li, X. Gao, Mcf: Mutual correction framework for semi-supervised medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15651–15660.

28

Dual-scale Enhanced and Cross-generative Consistency Learning for Semi-supervised Medical Image Segmentation

Yunqi Gu[a], Tao Zhou[a,*], Yizhe Zhang[a], Yi Zhou[b], Kelei He[c], Chen Gong[a], Huazhu Fu[d]

[a]PCA Lab, the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

[b]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China

[c]Medical School, Nanjing University, Nanjing 210023, China, and the National Institute of Healthcare Data

Science at Nanjing University, Nanjing 210023, China

[d]Institute of High Performance Computing, A*STAR, Singapore

1) We propose a novel semi-supervised medical image segmentation framework, leveraging scale-enhanced consistency and cross-generative consistency to boost the segmentation model.

2) A cross-level feature aggregation module (CFA) is proposed to fuse the cross-level features, which can enhance the representation ability of features within different resolutions.

3) We present a scale-enhanced consistency constraint and cross-generative consistency to boost the ability of feature representations.

4) A dual-scale complementary fusion (DCF) module is proposed to generate better predictions through the complementation and fusion of different scale information from the two scale-specific decoders.

5) Extensive experimental results show the effectiveness of the proposed method.

Yunqi Gu is currently pursuing the master's degree with School of Computer Science and Engineering, Nanjing University of Science and Technology, China. Her research is focus on the medical image analysis.

Tao Zhou is currently a Professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. He received the Ph.D. degree in Pattern Recognition and Intelligent Systems from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, in 2016. From 2016 to 2018, he was a Postdoctoral Fellow in the BRIC and IDEA lab, University of North Carolina at Chapel Hill. From 2018 to 2020, he was a Research Scientist at the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi. His research interests include machine learning, computer vision, and medical image analysis.

Yizhe Zhang is currently a Associate Professor in the School of Computer Science and Engineering at Nanjing University of Science and Technology. He received the Ph.D. degree in computer science and engineering at the University of Notre Dame, IN, USA. His research interests include medical (and biological) image analysis, machine learning and algorithm design.

Yi Zhou is currently an Associate Professor at Southeast University. He received a M.Sc. degree from the Department of Electronic and Electrical Engineering, University of Sheffield, U.K. in 2014, and a Ph.D. degree from the School of Computing Sciences, University of East Anglia, U.K. in 2018. He was a Research Scientist with IIAI. His research interests include computer vision, pattern recognition, machine learning, and medical imaging.

Kelei He received the Ph.D. degree in computer science and technology from Nanjing University, China. He is currently the assistant dean of National Institute of Healthcare Data Science at Nanjing University. He is also an assistant researcher of Medical School at Nanjing University, China. His research interests include medical image analysis, computer vision and deep learning.

Chen Gong received his B.E. degree from East China University of Science and Technology (ECUST) in 2010, and dual doctoral degree from Shanghai Jiao Tong University (SJTU) and University of Technology Sydney (UTS) in 2016 and 2017, respectively. Currently, he is a full professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests mainly include machine learning, data mining, and learning-based vision problems.

Huazhu Fu is currently a Senior Scientist at Inception Institute of Artificial Intelligence, UAE. He received a Ph.D. degree from Tianjin University in 2013 and was a Research Fellow at NTU for two years. From 2015 to 2018, he was a Research Scientist with I2R, A∗STAR, Singapore. His research interests include computer vision, machine learning, and AI in healthcare. He serves as an Associate Editor for IEEE TMI and IEEE JBHI, and also served as the Area Chair for MICCAI 2021 and Co-Chair for the OMIA Workshop.

Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

```



```