C2FTrans: Coarse-to-Fine Transformers for Medical Image Segmentation

Xian Lin, Li Yu, Kwang-Ting Cheng, and Zengqiang Yan

***Abstract*—Convolutional neural networks (CNN), the most prevailing architecture for deep-learning based medical image analysis, are still functionally limited by their intrinsic inductive biases and inadequate receptive fields. Transformer, born to address this issue, has drawn explosive attention in natural language processing and computer vision due to its remarkable ability in capturing long-range dependency. However, most recent transformer-based methods for medical image segmentation directly apply vanilla transformers as an auxiliary module in CNN-based methods, resulting in severe detail loss due to the rigid patch partitioning scheme in transformers. To address this problem, we propose C2FTrans, a novel multi-scale architecture that formulates medical image segmentation as a coarse-to-fine procedure. C2FTrans mainly consists of a cross-scale global transformer (CGT) which addresses local contextual similarity in CNN and a boundary-aware local transformer (BLT) which overcomes boundary uncertainty brought by rigid patch partitioning in transformers. Specifically, CGT builds global dependency across three different small-scale feature maps to obtain rich global semantic features with an acceptable computational cost, while BLT captures mid-range dependency by adaptively generating windows around boundaries under the guidance of entropy to reduce computational complexity and minimize detail loss based on large-scale feature maps. Extensive experimental results on three public datasets demonstrate the superior performance of C2FTrans against state-of-the-art CNN-based and transformer-based methods with fewer parameters and lower FLOPs. We believe the design of C2FTrans would further inspire future work on developing efficient and lightweight transformers for medical image segmentation. The source code of this paper is publicly available at https://github.com/xianlin7/C2FTrans.**

***Index Terms*—Transformer, Coarse-to-Fine, Multi-scale, Entropy, Medical Image Segmentation**

I. Introduction

MEDICAL image segmentation, a vital technique for computer-assisted diagnosis, can be formulated as a coarse-to-fine process: existence and rough area determination, boundary identification, and boundary refinement [1], which largely relies on multi-level features from deep to shallow. Through progressively merging multi-level features with skip connections, U-shaped convolutional neural networks (CNN) have dominated the field of medical image segmentation with superior performance, such as U-Net [2], Res-Unet [3], Attention U-Net [4], and UNet++ [5]. However, these CNN-based models still suffer from limited receptive fields.

Corresponding author: Zengqiang Yan

X. Lin, L. Yu, and Z. Yan are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: z_yan@hust.edu.cn).

K. -T. Cheng is with the School of Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong.

Transformers [6]–[10], born to capture long-term interaction, have recently attracted extensive attention in medical image segmentation [25]–[29]. Chen *et al.* [11] first explored the potential of transformers in medical image segmentation by introducing transformers to the last encoder layer of the vanilla U-Net framework. Then, a series of frameworks have been proposed focusing on combining transformer- and CNN-blocks in the encoder part for better feature modeling, such as MedT [12], TransFuse [13], TransBTS [14], UNETR [15], and PMTrans [1]. However, directly coupling vanilla transformers with the encoder to build global dependency usually results in daunting computational complexity, making it bleak for deploying transformers on high-resolution feature maps. To make transformers more computationally efficient on high-resolution feature maps, CoTr [16] introduced deformable self-attention to make full use of just a small set of key points. Swin-Unet [17] divided feature maps into a series of shifted windows and performed dependency establishment within each window separately. UTNet [18] proposed an efficient self-attention mechanism to reduce computational complexity by projecting K and V into low-dimension. Similarly, GT U-Net [19] introduced the grouping structure and the bottleneck structure for lightweight deployment. While effective in complexity reduction, the above frameworks suffer from vital local feature loss and insufficient robustness for multi-scale object segmentation due to rigid patch partition.

Recently, several approaches have been proposed to preserve the locality of transformer-based architectures. DS-TransUNet [20] adopted dual-scale Swin transformers to extract features under different window divisions, based on which to restore the window internal structure destroyed by single-scale window division. TransAttUnet [21] introduced multi-scale skip connections among decoder layers to minimize information loss. MISSFormer [22] redesigned the feed-forward network with depth-wise convolution to supply locality. MT-UNet [23] proposed a local-global Gaussian-weighted self-attention by alternating self-attention with local and global receptive fields. Despite the effectiveness in increasing the locality of transformers, the partitioning schemes of these approaches still are rigid, struggling to segment objects with irregular shapes and varying sizes.

In this paper, we propose a novel multi-scale architecture with coarse-to-fine transformers, named C2FTrans, to deal with the challenging scale-variant medical image segmentation tasks. Specifically, a U-shaped CNN backbone, consisting of three different scales of branch heads and one synthetic head, is developed to extract multi-scale features for segmentation. A cross-scale global transformer (CGT) is embedded

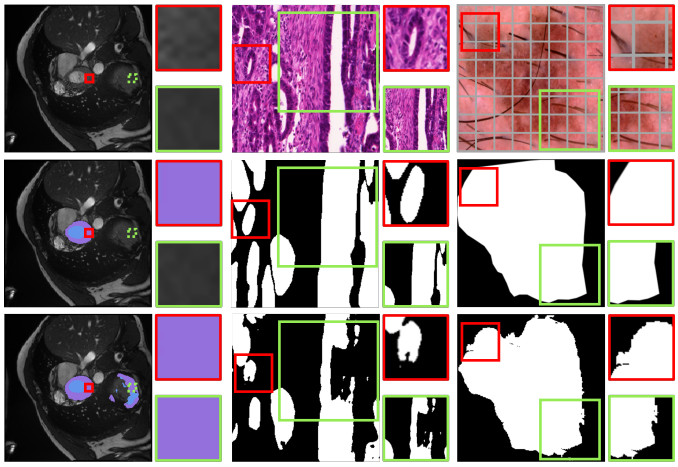
in the smallest-scale branch head to locate the rough area of objects by capturing global dependency among the three deepest layers of features, while a boundary-aware local transformer (BLT) is deployed in the middle-scale branch head to extract adequate features for boundary identification by adaptively generating windows around object boundaries and performing self-attention within each window. Then, a simple CNN module is utilized in the largest-scale branch head for boundary refinement. Predictions of the three branch heads are fused for final segmentation. Qualitative and quantitative comparison results with the state-of-the-art methods on three public datasets, including the ISIC 2018 dataset [37], [38], the ACDC dataset [39], and the GlaS dataset [40], demonstrate the effectiveness of C2FTrans for both 2D and 3D medical image segmentation. More importantly, the proposed 2D framework C2FTrans can achieve better segmentation performance on 3D medical images with much lower computational complexity than existing 3D transformer-based approaches. The contributions can be summarized as follows:

- C2FTrans, a multi-scale segmentation framework based on coarse-to-fine transformers for segmenting medical objects with varying shapes and sizes.
- A cross-scale global transformer (CGT) to extract global-range dependency with lower computational complexity.
- A boundary-aware local transformer (BLT) to capture mid-range dependency for accurate boundary segmentation. To our best knowledge, this is the first work to study transformers with dynamic and adaptive window partitioning for medical image segmentation.
- Superior segmentation performance, compared to the state-of-the-art CNN-based and transformer-based approaches on three public datasets covering 2D and 3D medical imaging data.

II. Problem Analysis

Organs, tissues, lesions, and other objects in medical image segmentation can have varying sizes and irregular shapes, segmenting which often suffers from extensive misclassifications and poor boundary detection, due to the following reasons:

1) **Inadequate receptive field in CNN:** In medical image segmentation, both deep and shallow features are crucial since the former corresponds to large receptive fields for global semantic capturing while the latter corresponds to small receptive fields for detail localization. Unfortunately, existing CNN-based methods are subject to limited receptive fields, resulting in inaccurate segmentation when dealing with local contextual similarity and large-scale variation. As shown in Fig. 1(a), the green patch suffers from serious misclassification due to its similar local contextual information with the target region (*i.e.* the red patch). Meanwhile, in Fig. 1(b), the small-scale instance (*i.e.* the red patch) requires rich local features for separation while the large-scale object (*i.e.* the green patch) requires global features for segmentation.
2) **Rigid partitioning scheme in transformers:** The inadequate receptive field issue can be perfectly addressed by transformers [30]–[34]. However, as shown in Fig. 1(c), though the completeness of the large-scale skin lesion is well preserved by TransUNet [11], the corresponding boundaries are inaccurate due to its rigid patch partitioning scheme where each image is simply divided through an evenly tiled grid (*i.e.*, the gray grid shown in Fig. 1(c)). In such a partitioning scheme, detailed information around the split lines would be seriously destroyed. It may not be a big issue for the interior regions of objects but is fatal for the fine boundary regions.



(a) ACDC (b) GlaS (c) ISIC 2018

Fig. 1: Challenges in medical image segmentation. From left to right: the raw images, the manual annotations, and the segmentation results of cardiac, gland, and skin lesion from the ACDC [39], GlaS [40], and ISIC 2018 [37], [38] datasets respectively produced by U-Net [4] or TransUNet [11].

Based on the above analysis, developing transformers with adequate receptive fields and adaptive patch partitioning has the potential of achieving better segmentation performance, especially for objects with varying shapes and sizes.

III. Method

A. Overview

The overall architecture of C2FTrans is depicted in Fig. 2, consisting of a U-shaped backbone for initial multi-scale feature exaction, three branch heads based on coarse-to-fine transformers to utilize multi-scale features, and a synthetic head to produce the final segmentation results. In C2FTrans, the segmentation task is modeled as a coarse-to-fine procedure corresponding to the three-scale features. As analyzed in Section II, inadequate receptive fields in CNN are largely responsible for segmentation errors of large-scale objects and regions with local contextual similarity. Thus, we build a cross-scale global transformer (CGT) to identify the existence of each class and their rough locations by capturing the holistic characteristics. Considering the inaccurate boundaries of the rough regions due to rigid patch partitioning in CGT, a boundary-aware local transformer (BLT) is deployed in the second branch head for precise boundary localization. Following BLT, a simple convolution operation is applied in

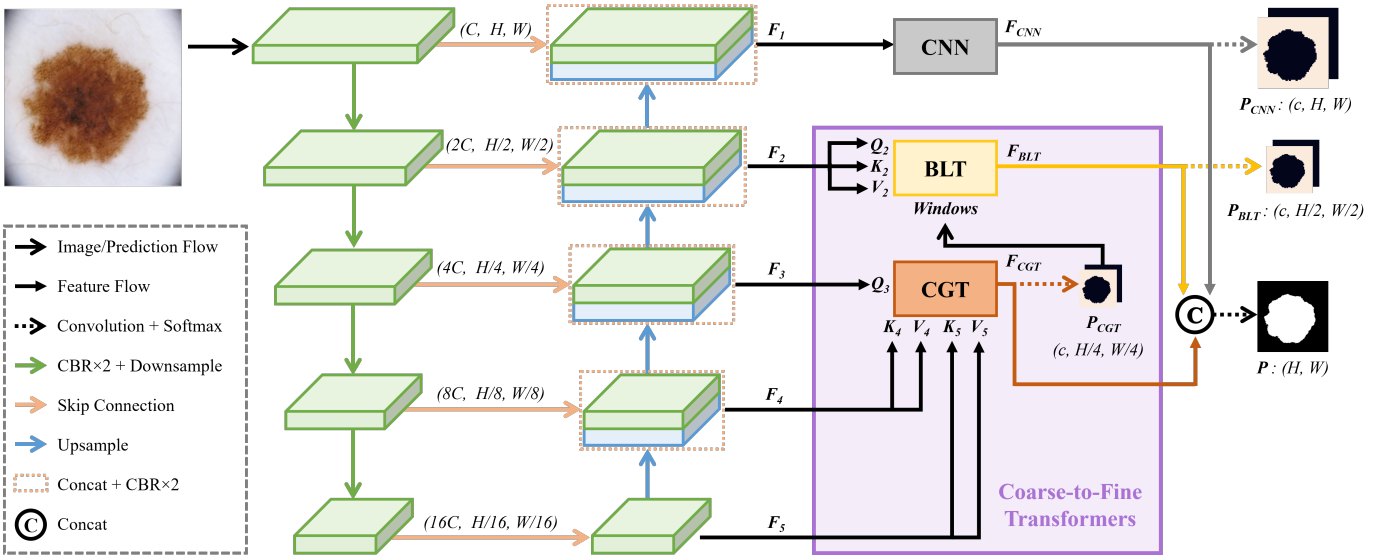


Fig. 2: Overview of the proposed multi-scale medical image segmentation framework based on coarse-to-fine transformers (C2FTrans), where CGT and BLT represent the cross-scale global transformer and the boundary-aware local transformer respectively.

the first branch head to refine the boundaries as the small receptive fields of convolution are better for capturing details. Finally, the results corresponding to rough area localization (*i.e.* CGT), boundary identification (*i.e.* BLT), and boundary refinement (*i.e.* CNN) are combined for final segmentation. In the following, we detail each component of the proposed coarse-to-fine transformers.

B. Cross-Scale Global Transformer

CGT consists of two cross-scale attention modules and one feed-forward network (FFN) as illustrated in Fig. 3, aiming to build global dependency among the three small-scale feature maps generated by the backbone, *i.e.*, $F_3 \in \mathbb{R}^{4C\times\frac{H}{4}\times\frac{W}{4}}$, $F_4 \in \mathbb{R}^{8C\times\frac{H}{8}\times\frac{W}{8}}$, and $F_5 \in \mathbb{R}^{16C\times\frac{H}{16}\times\frac{W}{16}}$ where C is the channel count of the first branch head and (H, W) is the resolution of the input image. In the cross-scale attention module, F_3 with the maximum resolution is projected into query $Q_{3,i} \in \mathbb{R}^{\frac{HW}{16}\times d}, i \in \{4, 5\}$ where d is the dimension of the transformer module, while F_4 and F_5 are projected into two groups of $\{K_i, V_i\} \in \mathbb{R}^{\frac{HW}{2^{2(i-1)}}\times d}$. Then, cross-scale attention can be described as:

$$\mathcal{F}^i_{ca}(Q_{3,i}, K_i, V_i) = softmax(\frac{Q_{3,i}K_i^T}{\sqrt{d}})V_i. \quad (1)$$

In contrast to the vanilla self-attention which generates Q, K, and V from the same input, the proposed cross-scale attention derives K and V from the other two smaller-scale features. In this way, computational complexity can be reduced by a factor of 2^2 or 4^2 since the sequence lengths of K and V are shorter. More importantly, diverse features are included for dependency establishment, since different groups of K and V correspond to different scales of receptive fields and larger receptive fields would contain richer semantic information.

Before feeding to FFN, the two sets of cross-scale attention are combined and refined according to the following

$$F_{ca} = (\mathcal{F}^4_{ca,1}ⓒ\cdots ⓒ\mathcal{F}^4_{ca,g}ⓒ\mathcal{F}^5_{ca,1}\cdots ⓒ\mathcal{F}^5_{ca,g})\cdot W_{ca}⊕F_3, \quad (2)$$

where g is the predefined number of transformer heads in CGT, $W_{ca} \in \mathbb{R}^{2gd\times d}$ is a learnable projection matrix for combination, ⓒ is the concatenation operation, and ⊕ represents residual connection via element-wise addition.

FFN after cross-scale attention is further processed to obtain the final output of CGT according to

$$F_{CGT} = (max(0, F_{ca}W_{c1} + b_{c1})\cdot W_{c2} + b_{c2})⊕F_{ca}, \quad (3)$$

where $W_{c1} \in \mathbb{R}^{d\times 4d}$ and $W_{c2} \in \mathbb{R}^{4d\times d}$ are the learnable projection matrices, and b_{c1}, $b_{c2} \in \mathbb{R}$ are the offsets.

C. Boundary-Aware Local Transformer

BLT contains three stages: dynamic boundary-aware window generation for adaptively boundary localization, boundary-around self-attention for boundary identification, and FFN for feature distillation. As analyzed in Section II, existing transformers adopt the rigid window/patch partitioning scheme, which severely destroys the vital details around boundaries. Therefore, the core of BLT is to perform local self-attention within boundary-aware windows.

Boundary-aware windows are generated under the guidance of entropy calculation. Firstly, evenly and densely tiled windows over the feature map $F_2 \in \mathbb{R}^{2C\times\frac{H}{2}\times\frac{W}{2}}$ are obtained as the initial window set $\{w\}$ containing all possible window positions. Before calculating the probability of each window being boundary-around, the entropy of each position (m, n) in the probability map $P_{CGT} \in \mathbb{R}^{C\times\frac{H}{4}\times\frac{W}{4}}$ produced by CGT is calculated by

$$\mathcal{C}_p(m, n) = -\frac{1}{\log_2 L}\sum_{l=1}^{L} P_{CGT}(l, m, n)\log_2 P_{CGT}(l, m, n), \quad (4)$$

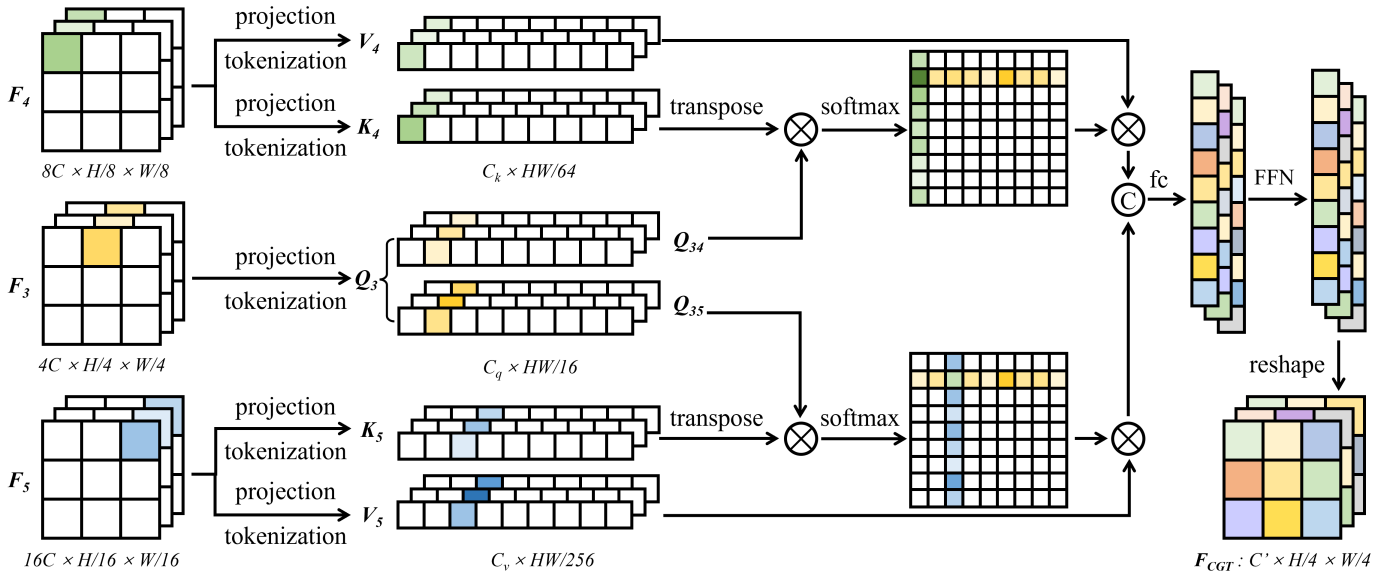


Fig. 3: Overview of the cross-scale global transformer (CGT). The residual structures of self-attention and the feed-forward neural networks (FFN) are omitted for brevity.

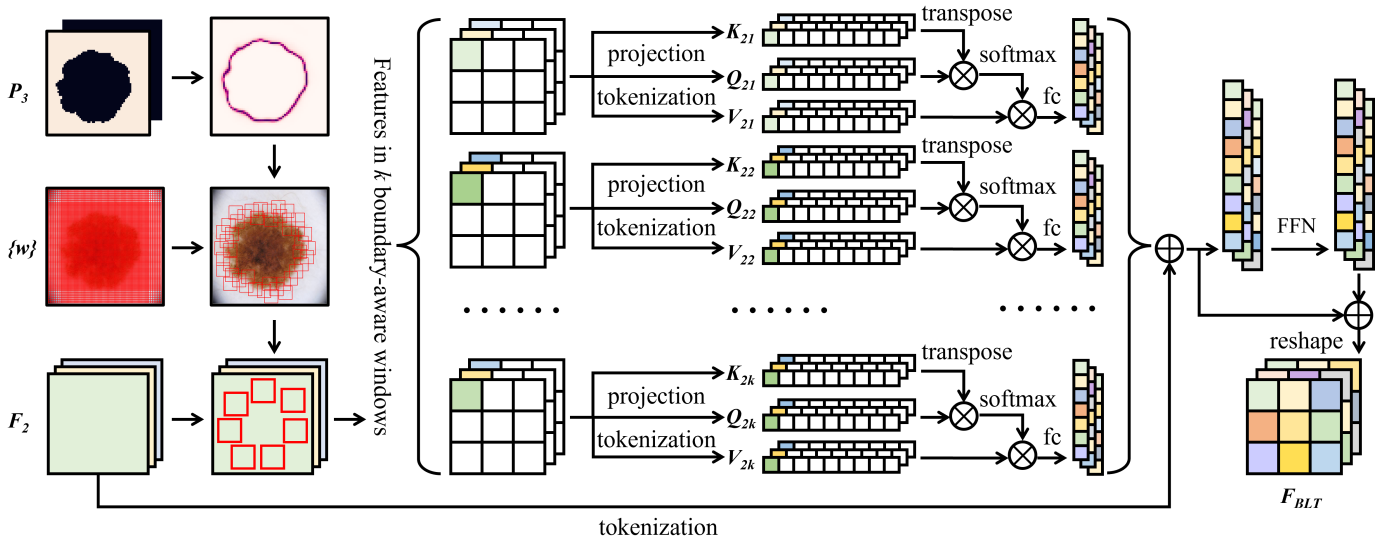


Fig. 4: Overview of the boundary-aware local transformer (BLT).

where L is the number of classes. Then, the score of each window is defined as:

$$\mathcal{C}_w(x,y) = \frac{1}{\lfloor \frac{h}{2} \rfloor \lfloor \frac{w}{2} \rfloor} \sum_{m=0}^{\lfloor \frac{h}{2} \rfloor} \sum_{n=0}^{\lfloor \frac{w}{2} \rfloor} \mathcal{C}_p(\lfloor \frac{x}{2} \rfloor + m, \lfloor \frac{y}{2} \rfloor + n), \quad (5)$$

where (x, y) is the upper-left coordinate of the window and (h, w) is the size of the window. As is known, entropy is to measure the uncertainty degree of information, namely positions with higher entropy scores are more likely to be real boundaries. Therefore, scoring through entropy can effectively localize the boundary windows. After non-maximum suppression [35], a boundary-aware window set $\{w^*\}$ and the corresponding feature set $\{f^*\}$ can be obtained by removing redundant windows and aligning $\{w^*\}$ with F_2 [36]. Then, performing multi-head self-attention within each f_j^* can extract specific features for the corresponding window according to

$$\begin{aligned} \mathcal{F}_{sa}(f_j^*) &= softmax(\frac{f_j^* E_q f_j^* E_k}{\sqrt{d}}) f_j^* E_v \\ F_{sa} &= (\mathcal{F}_{sa,1} \textcircled{c} \cdots \textcircled{c} \mathcal{F}_{sa,g}) \cdot W_{sa} \oplus F_2, \end{aligned} \quad (6)$$

where $E_q \in \mathbb{R}^{2C \times d}$, $E_k \in \mathbb{R}^{2C \times d}$, $E_v \in \mathbb{R}^{2C \times d}$, and $W_{sa} \in \mathbb{R}^{gd \times d}$ are learnable projection matrices. Finally, the output of BLT can be described as:

$$F_{BLT} = (max(0, F_{sa} W_{s1} + b_{s1}) \cdot W_{s2} + b_{s2}) \oplus F_{sa}, \quad (7)$$

where $W_{s1} \in \mathbb{R}^{d \times 4d}$ and $W_{s2} \in \mathbb{R}^{4d \times d}$ are learnable projection matrices and b_{s1}, $b_{s2} \in \mathbb{R}$ are the offsets in FFN. It should be noticed that only regions around boundaries will

go through the above boundary-around local multi-head self-attention and feature updating. As a result, its computational complexity is

$$\Omega_{BLT} = k(6dhwC + 2d(hw)^2 + d^2hw), \tag{8}$$

where k is the maximum number of windows. In our experiments, we adopt $h = \frac{H}{32}$, $w = \frac{W}{32}$, and set k as $\alpha\frac{HW}{hw}, \alpha \in (0, 1)$. Then, Eq. 8 is rewritten as

$$\begin{aligned}\Omega_{BLT} &= \alpha\left(6dHWC + \frac{2}{32\times 32}d(HW)^2 + d^2HW\right)\\ &\approx \mathrm{O}\left(\frac{\alpha}{16\times 32}d(HW)^2\right).\end{aligned} \tag{9}$$

Compared to directly applying Vision Transformer (ViT) [7] onto F_2 whose computational complexity is

$$\begin{aligned}\Omega_{ViT} &= \frac{3}{8}dHWC + \frac{1}{8}d(HW)^2 + \frac{1}{16}d^2HW\\ &\approx \mathrm{O}\left(\frac{1}{8}d(HW)^2\right),\end{aligned} \tag{10}$$

the computational complexity of BLT is at most $\frac{1}{64}$ of ViT .

D. Multi-Scale Soft Supervision

For medical image segmentation using multi-scale features, inspired by [50], we introduce soft supervision to alleviate boundary uncertainty brought by down-sampling or up-sampling. Concretely, the ground truth mask $M \in \mathbb{R}^{H\times W}$ is resized to multiple scales for supervision (*i.e.*, $G_{BLT} \in \mathbb{R}^{L\times\frac{H}{2}\times\frac{W}{2}}$ for training BLT and $G_{CGT} \in \mathbb{R}^{L\times\frac{H}{4}\times\frac{W}{4}}$ for training CGT) while preserving the original boundary distributions. Given class l and the down-sampling scale s, the resized ground-true map $G_{l,s}$ is constructed as

$$G_{l,s}(i,j) = \frac{1}{2^{2(s-1)}}\sum_{(m,n)\in O_{i,j}} |M(m,n) == l|, \tag{11}$$

where $O_{i,j}$ represents the down-sampling block in M corresponding to (i, j) in $G_{l,s}$ defined as $M((i-1)2^{s-1}+1 : i2^{s-1}, (j-1)2^{s-1}+1 : j2^{s-1})$. In this way, the probability value of (i, j) in $G_{l,s}$ is determined by the frequency of class l in the down-sampling block. In this way, the boundary distributions in $G_{l,s}$ would approach to those in the original mask M. Given the down-sampling scale s, a set of resized ground-true maps are constructed accordingly as $G_s = \{G_{l,s}|l = 1, 2, ..., L\}$ where L is the total number of classes. Then, the loss between $P_s = \{P_{l,s}|c = 1, 2, ..., L\}$ and G_s of the same scale is defined as

$$\mathcal{L}_s = \sum_{l=1}^{L}\sum_{i=1}^{\frac{H}{2^{s-1}}}\sum_{j=1}^{\frac{W}{2^{s-1}}} |G_{l,s}(i,j) - P_{l,s}(i,j)| \tag{12}$$

After setting $s = 1, 2, 3$ to define the loss functions $\mathcal{L}_{CNN}$, $\mathcal{L}_{BLT}$, and $\mathcal{L}_{CGT}$ for CNN, BLT, and CGT respectively, their features are further fused to generate the final predictions which are penalized by both cross-entropy loss and dice loss denoted as $\mathcal{L}_C$. The overall loss function $\mathcal{L}$ for multi-scale soft supervision of C2FTrans is

$$\mathcal{L} = \beta_1\mathcal{L}_{CNN} + \beta_2\mathcal{L}_{BLT} + \beta_3\mathcal{L}_{CGT} + \beta_4\mathcal{L}_C \tag{13}$$

where β_1, β_2, β_3, and β_4 are balancing hyper-parameters set as 0.2, 0.1, 0.1, and 0.6 respectively in our experiments.

IV. Evaluation

In this section, we conduct extensive experiments comparing C2FTrans against various state-of-the-art CNN-based and transformer-based approaches on publicly-available datasets.

A. Datasets

Three datasets covering 2D and 3D medical image data are adopted for evaluation, including the ISIC 2018 dataset [37], [38], the GlaS dataset [39], and the ACDC dataset [40].

1) ISIC 2018: The dataset contains 2596 images with well-annotated labels. Following the setting in [41], these images are resized into 256×256 and split into 1815 images for training, 259 images for validation, and 520 images for testing.

2) ACDC: The dataset contains 150 magnetic resonance imaging (MRI) 3D cases collected from different patients. Following the setting in [11], [42], only 100 cases are utilized for evaluation, and the ratio of 7:1:2 is adopted to split these cases into training, validation, and testing sets. Slices are resized to the same resolution 256×256.

3) GlaS: The dataset consists of 165 microscopic images of hematoxylin and eosin-stained slides and is officially split into training and testing sets. Images are resized to the same resolution 256×256 for training.

B. Implementation Details

All the networks were trained using an Adam optimizer with an initial learning rate of 0.001 and a batch size of 4. All methods were implemented within the PyTorch framework and trained on NVIDIA 3090ti GPUs for 400 rounds. In the last 50 rounds, the parameters of the backbone are frozen, and only the parameters of the branch heads and the synthetic head are updated with a learning rate of 10^{-4}. Random rotation, random scaling, cropping, contrast adjustment, and gamma augmentation are applied for data augmentation.

C. Evaluation on ISIC 2018

1) Learning Frameworks for Comparison: Both CNN-based and transformer-based/-hybrid state-of-the-art architectures have been included for comparison on the ISIC 2018 dataset. CNN-based architectures include U-Net [2], AttU-Net [4], ResUnet++ [43], CPFNet [44], DAGAN [45], and CKDNet [46], while transformer-based approaches include SETR [10], MedT [12], and FAT-Net [41]. Among these methods, DAGAN, CKDNet, and FAT-Net are specifically designed for skin legion segmentation, SETR is developed for semantic segmentation, and the rest approaches are the most recent medical image segmentation works. MedT and SETR are re-implemented according to the released source codes.

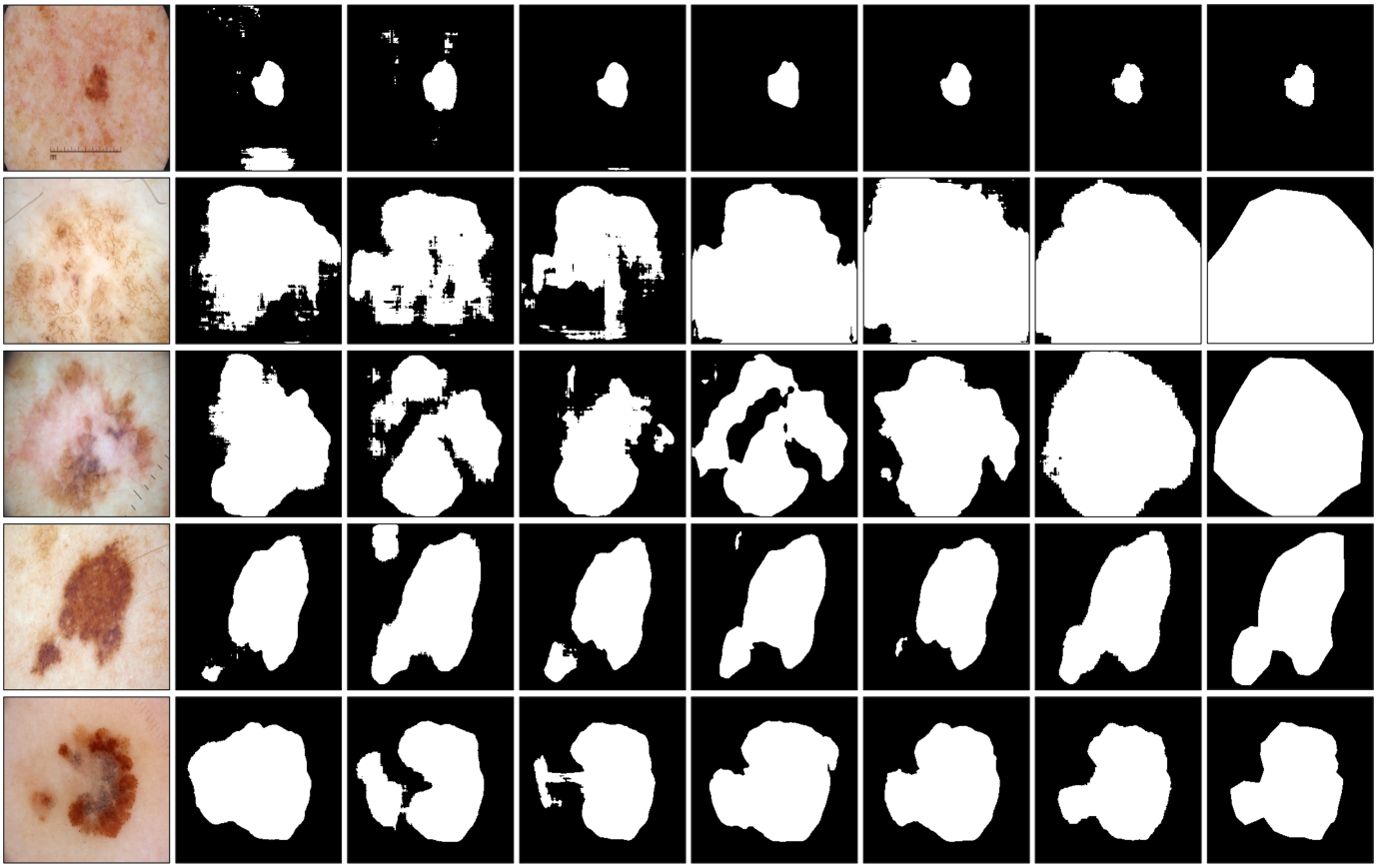


Fig. 5: Qualitative comparison results on the ISIC 2018 dataset. From left to right: the raw images, the segmentation maps produced by U-Net [2], ResUnet++ [43], AttU-Net [4], SETR [10], MedT [12], and C2FTrans respectively, and the manual annotations.

TABLE I: Comparison results on the ISIC dataset.

Model Type	Method	Year	Dice (%)	IoU (%)	ACC (%)	SE (%)	SP (%)	Params (M)	GFLOPs
CNN	U-Net [2]	2015	85.45	77.33	94.04	88.00	96.77	40	89
	AttU-Net [4]	2019	85.66	77.64	93.76	86.00	98.26	45	84
	ResUNet++ [43]	2019	85.36	77.21	93.82	87.35	97.21	87	95
	CPFNet [44]	2020	87.69	79.88	94.96	89.53	96.55	43	16
	DAGAN [45]	2020	88.07	81.13	93.24	90.72	95.88	54	62
	CKDNet [46]	2021	87.79	80.41	94.92	90.55	97.01	51	44
Transformer	SETR-PUP* [10]	2021	88.03	80.53	95.51	91.51	96.52	40	39
	MedT [12]	2021	89.28	82.48	95.83	88.16	**98.00**	1.6	4.4
	FAT-Net [41]	2022	89.03	82.02	95.78	91.00	96.99	30	23
	C2FTrans	2022	**90.76**	**84.64**	**96.76**	**91.22**	97.74	**1.2**	**4.1**

2) Quantitative Results: Quantitative comparison results of different methods on the ISIC 2018 dataset are depicted in Table I. Compared to the CNN-based methods, well-designed transformer-based methods achieve better segmentation performance with fewer parameters and lower FLOPs, among which FAT-Net achieves the best performance. Compare to FAT-Net, C2FTrans achieves consistent performance improvements by an average increase of 1.73%, 2.62%, 0.98%, 0.22%, and 0.75% in Dice, IoU, ACC, SE, and SP respectively. Though MedT owns the best SP results in Table I, the overall performance is much worse than that of C2FTrans. Furthermore, C2FTrans is the most lightweight framework with the fewest parameters and the lowest FLOPs, indicating the effectiveness and efficiency of the proposed coarse-to-fine transformers.

3) Qualitative Results: Exemplar qualitative results produced by different approaches on several challenging cases of the ISIC 2018 dataset are provided in Fig. 5. Given local contextual similarity as shown in the first row of Fig. 5, CNN-based methods can hardly distinguish skin lesions from background resulting in extensive false negatives, while transformer-based methods can recognize skin lesions correctly. Given the large-scale skin lesions as shown in the sec-

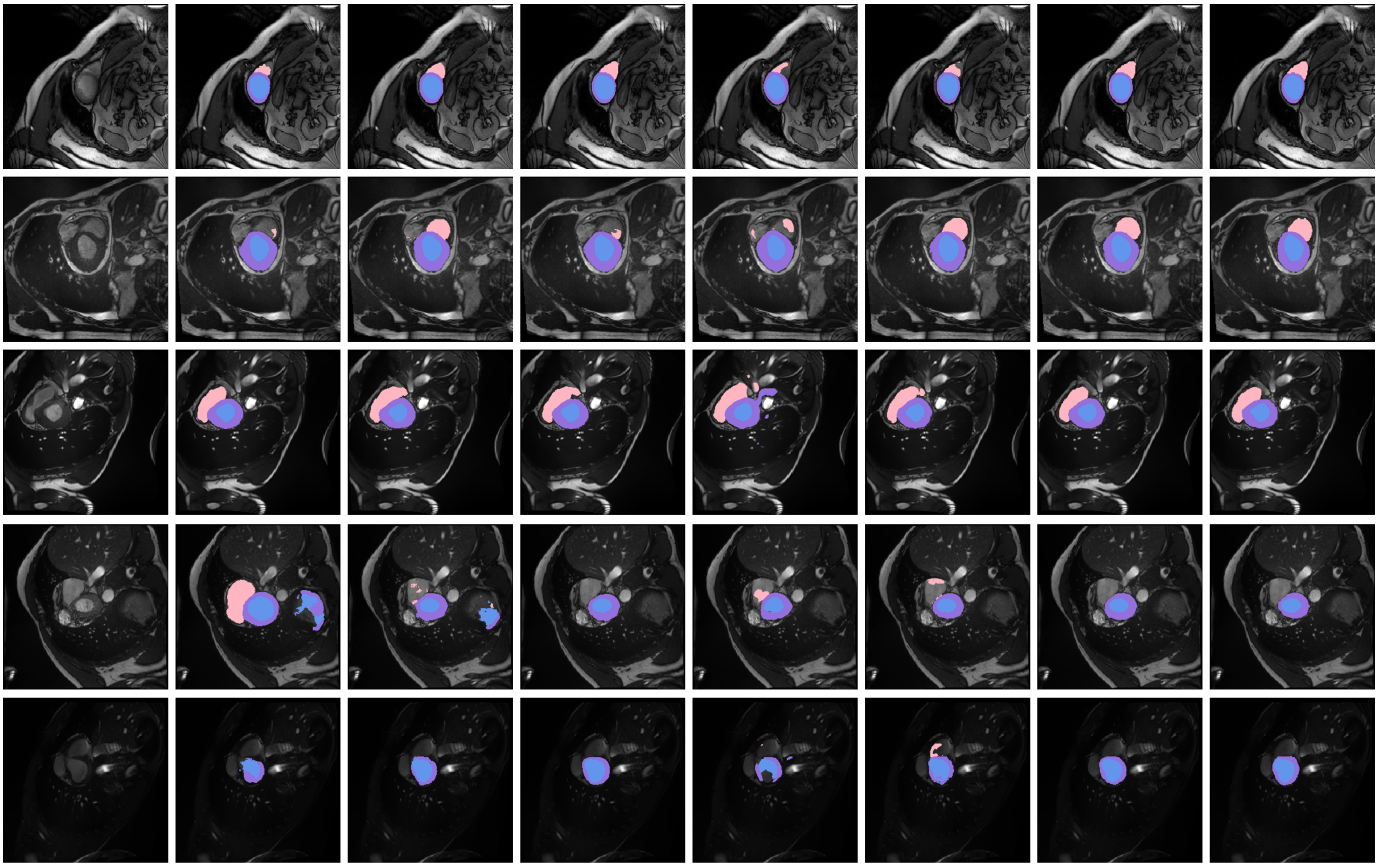


Fig. 6: Qualitative comparison results on the ACDC dataset. From left to right: the raw images, the segmentation maps produced by U-Net [2], ResUnet++ [43], AttU-Net [4], SETR [10], MedT [12], and C2FTrans respectively, and the manual annotations. The pink, purple, and blue regions denote right ventricle, myocardium, and left ventricle respectively.

TABLE II: Comparison results on the ACDC dataset. RV, Myo, and LV represent right ventricle, myocardium, and left ventricle.

Model Type	Method	Year	Dice (%)				Params (M)	GFLOPs
			Avg.	RV	Myo	LV		
CNN	R50 U-Net [2]	2015	87.55	87.10	80.63	94.52	-	-
	R50 AttUNet [4]	2019	86.75	87.58	79.20	93.47	-	-
3D Transformer	nnFormer [42]	2021	92.06	90.94	89.58	95.65	159	158
	UNETR [15]	2022	88.61	85.29	86.52	94.02	92	86
	D-Former [49]	2022	92.29	91.33	**89.60**	**95.93**	44	54
2D Transformer	SETR-PUP* [10]	2021	83.87	82.29	78.54	90.79	40	39
	MISSFormer [22]	2021	87.90	86.36	85.75	91.59	-	-
	TransUNet [11]	2021	89.71	88.86	84.54	95.73	105	25
	Swin-Unet [17]	2021	90.00	88.55	85.62	95.83	41	12
	MedT [12]	2021	91.44	91.43	87.72	95.16	1.6	4.4
	C2FTrans	2022	**92.83**	**93.17**	89.55	95.76	**1.2**	**4.1**

ond and the third rows of Fig. 5, segmentation maps generated by transformer-based methods are more complete than those produced by the CNN-based methods, validating the analysis in Section II, *i.e.*, CNN-based methods would suffer from local contextual similarity and large-scale object segmentation due to inadequate receptive fields. Compared to the state-of-the-art transformer-based methods, C2FTrans effectively reduces false negatives with better boundary preservation, benefiting from both CGT and BLT respectively.

D. Evaluation on ACDC

1) Learning Frameworks for Comparison: For evaluation on the 3D ACDC dataset, CNN-based methods including U-Net [2] and AttUNet [4], 2D transformer-based methods including SETR [10], MISSFormer [22], TransUNet [11], Swin-Unet [17], and MedT [12], and 3D transformer-based methods including UNETR [15], nnFormer [42], and D-Former [49] are used for comparison.

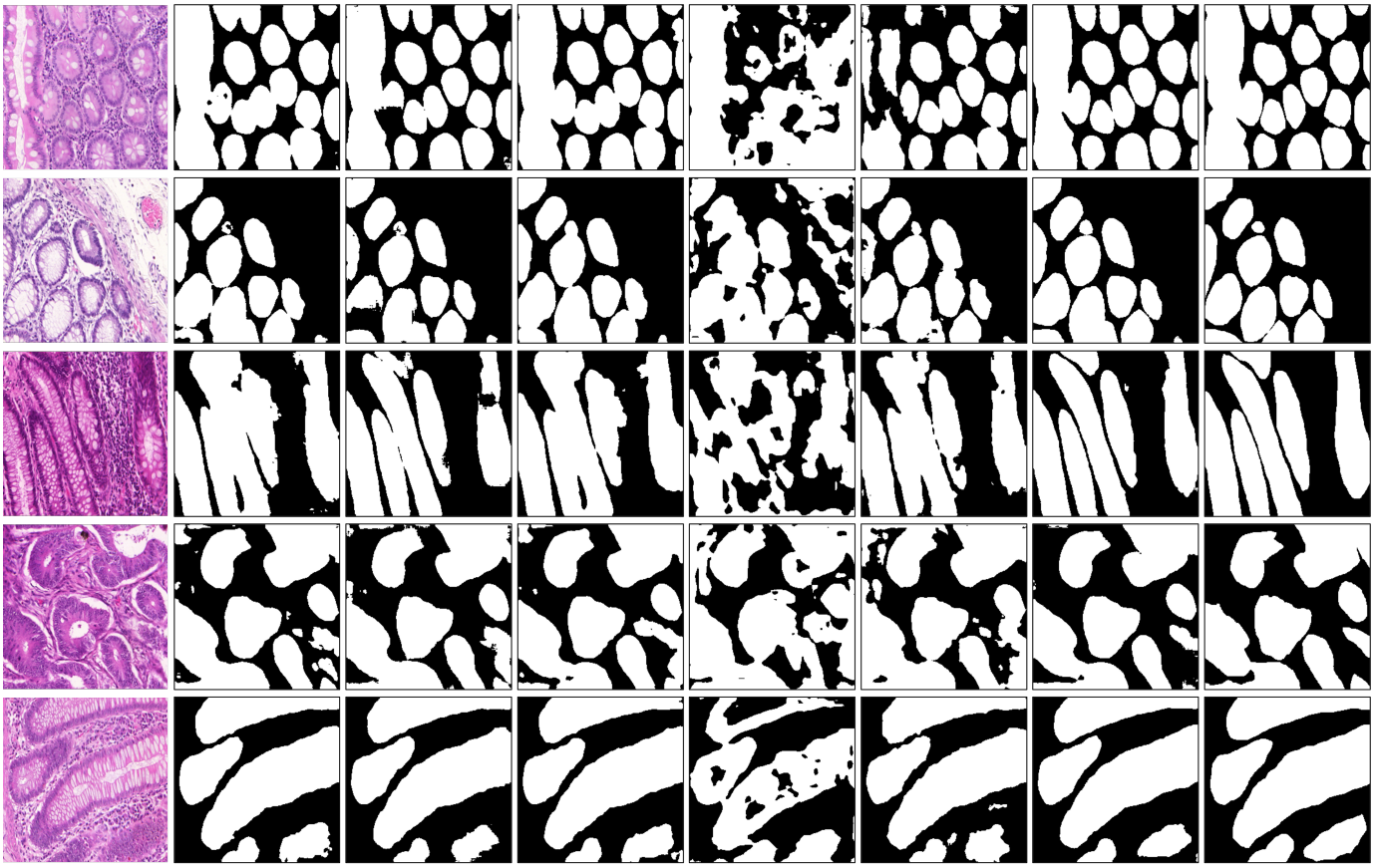


Fig. 7: Qualitative comparison results on the GlaS dataset. From left to right: the raw images, the segmentation maps produced by U-Net [2], ResUnet++ [43], AttU-Net [4], SETR [10], MedT [12], and C2FTrans respectively, and the manual annotations.

TABLE III: Comparison results on the GlaS dataset.

Model Type	Method	Year	Dice (%)	IoU (%)
CNN	U-Net [2]	2015	79.76	67.63
	Seg-Net [47]	2017	78.61	65.96
	AttUNet [4]	2019	81.59	70.06
	KiU-Net [48]	2020	83.25	72.78
Transformer	SETR-PUP [10]	2021	77.87	64.80
	MedT [12]	2021	81.02	69.61
	DS-TransUNet [20]	2021	87.19	78.45
	TransAttUNet [21]	2021	89.11	81.13
	C2FTrans	2022	**91.35**	**84.65**

2) Quantitative Results: According to the quantitative results in Table II, transformer-based methods can achieve better segmentation performance on the ACDC dataset, especially the 3D transformer-based methods with the help of inter-slice spatial information. Specifically, D-Former has the best performance with an average dice of 92.29%. Compared to the state-of-the-art 3D transformers, though C2FTrans is based on 2D slices, it achieves considerable performance improvements on the ACDC dataset, outperforming D-Former by an average increase of 0.54% in Dice. It should be noted that the proposed 2D framework C2FTrans is much more lightweight than D-Former, which is more feasible in clinical scenarios.

3) Qualitative Results: Qualitative segmentation results of different methods on the ACDC dataset are presented in Fig. 6. Suffering from local contextual similarity, CNN-based methods produce more false positives. Comparatively, transformer-based methods produce extensive false negatives with distorted boundaries due to rigid partitioning. Thanks to both CGT and BLT, C2FTrans achieves the best segmentation performance on all slices in Fig. 6, especially on boundary consistency with the manual annotations.

E. Evaluation on GlaS

1) Learning Frameworks for Comparison: Both CNN-based and transformer-based/-hybrid frameworks are included for evaluation. CNN-based architectures include U-Net [2], Seg-Net [47], AttU-Net [4], and KiU-Net [48], while transformer-based approaches include SETR [10], MedT [12], DS-TransUNet [20], and TransAttUNet [21]. Among these methods, KiU-Net, DS-TransUNet, and TransAttUNet represent the state-of-the-art performance on the GlaS dataset.

2) Quantitative Results: Quantitative comparison results of different methods are provided in Table III. Compared to skin lesion segmentation, gland instances can have various sizes and irregular shapes, most of which are small-scale. As a result, transformer-based methods may not always outperform the CNN-based methods. According to Table III,

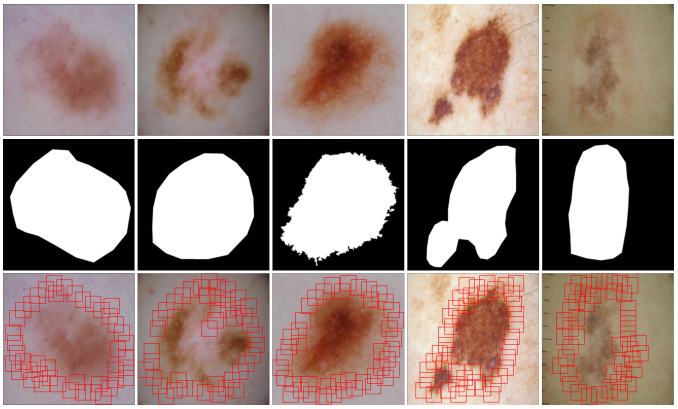


Fig. 8: Visualization of boundary-around window generation by the proposed BLT module.

TABLE IV: Ablation study on different component combinations of C2FTrans on the ISIC 2018 dataset.

Method	Dice (%)	IoU (%)	ACC (%)	SE (%)	SP (%)
backbone	88.41	81.09	95.53	89.46	97.51
+ CGT	89.60	82.84	96.07	90.85	97.48
+ BLT	89.73	83.30	96.36	**91.55**	97.47
C2FTrans	**90.76**	**84.64**	**96.76**	91.22	**97.74**

TABLE V: Ablation study on different window sizes of the BLT module on the ISIC 2018 dataset.

window size	Dice (%)	IoU (%)	ACC (%)	SE (%)	SP (%)
8×8	89.66	83.05	96.24	90.35	**97.87**
12×12	89.59	82.88	96.32	**92.07**	97.11
16×16	**89.73**	**83.30**	**96.36**	91.55	97.47
20×20	89.55	82.97	96.30	90.99	97.63

DS-TransUNet and TransAttUnet achieve inevitable performance improvements by introducing locality to transformers in addition to long-range dependency with larger receptive fields. In general, C2FTrans achieves the best segmentation performance in both Dice (*i.e.* 91.35%) and IoU (*i.e.* 84.65%), outperforming the most competitive TransAttUNet by 2.24% and 3.52% respectively.

3) Qualitative Results: Qualitative comparison results on the GlaS dataset are shown in Fig. 7. Different from the results in Fig. 5, transformer-based methods own many more false negatives. SETR, extracting features via a vanilla global transformer, fails to segment most instances. By introducing locality, MedT achieves much better performance but can hardly identify boundaries to separate instances. The poor performance of SETR and MedT demonstrates the limitation of the rigid window/patch partitioning strategy on boundary preservation. In contrast, C2FTrans, especially the boundary-aware local transformer, can better segment the instances with irregular shapes. Compared to CNN-based methods, C2FTrans achieves better boundary preservation and instance separation.

V. Discussion

To further verify the effectiveness of C2FTrans, ablation studies are conducted in the following.

A. Effectiveness of Each Component in C2FTrans

We evaluate the effectiveness of CGT and BLT by separately introducing them to the backbone network (*i.e.* multi-scale UNet) for medical image segmentation on the ISIC 2018 dataset. Quantitative comparison results summarized in Table IV indicate that coupling either component of the coarse-to-fine transformers is helpful for performance improvement, which is consistent with the analysis in Section II. Specifically, by capturing global dependency based on cross-scale features, CGT improves the overall segmentation performance by an average increase of 1.19 %, 1.75%, 0.54%, and 1.21% in Dice, IoU, ACC, and SE respectively. Comparatively, BLT achieves greater performance improvements, achieving an average increase of 1.32 %, 2.21%, 0.83%, and 2.09% in Dice, IoU, ACC, and SE respectively. This can be explained by the fact that boundary uncertainty is more common and challenging in medical image segmentation, and thus addressing it would be relatively more beneficial. To better evaluate the effectiveness of BLT, we further plot the generated boundary-around windows as shown in Fig. 8. BLT can accurately localize the boundary regions and build mid-range dependency for boundary identification, which in turn alleviates boundary uncertainty for performance improvement. When jointly using CGT and BLT, C2FTrans can better handle both local contextual similarity and boundary uncertainty, leading to maximum performance improvements.

B. Effect of Window Size in BLT

To analyze the effect of the window size (h, w) in BLT, we evaluate the performance of the baseline multi-scale UNet coupled with BLT under different window sizes $(h, w) = (8, 8)$, $(12, 12)$, $(16, 16)$, and $(20, 20)$. Quantitative comparison results are provided in Table V. With smaller window sizes, more windows would be regarded as boundary windows, resulting in redundancy and poor mid-range dependency establishment. Comparatively, using larger window sizes can better capture mid-range dependency within each window for boundary identification but may fail to preserve complete boundaries. Therefore, the selection of (h, w) is task-specific. According to the comparison results in Table V, setting (h, w) as $(16, 16)$ achieves the best BLT performance on the ISIC 2018 dataset. It should be noted that the performance gaps under various (h, w) settings are quite limited, demonstrating the robustness of BLT in boundary preservation.

VI. Conclusion

In this paper, we propose C2FTrans for medical image segmentation by reformulating medical image segmentation as a coarse-to-fine process, coarse region identification, boundary localization, and boundary refinement. Specifically, a cross-scale global transformer is developed to capture long-range dependency for coarse segmentation, which can effectively address local contextual similarity and improve the segmentation performance of large-scale objects/instances. Then, a

boundary-aware local transformer is introduced to alleviate boundary uncertainty by adaptively generating boundary-aware windows and performing mid-range dependency establishment within each window. Boundary refinement is accomplished through a simple convolution to utilize small-size receptive fields and local inductive bias for detail extraction. Experimental results on three widely-used datasets demonstrate C2FTrans' greater effectiveness and better generalization ability compared to both of the state-of-the-art CNN-based and transformer-based approaches for medical image segmentation. Furthermore, C2FTrans achieves consistent performance improvements across 2D and 3D medical image data with lower computational complexity, which is another attractive feature of the proposed approach.

REFERENCES

[1] Z. Zhang, B. Sun, and W. Zhang, "Pyramid medical transformer for medical image segmentation," 2021, *arXiv:2104.14702*.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234-241.

[3] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted ResUNet for high-quality retina vessel segmentation," in *Proc. ITME*, 2018, pp. 327-331.

[4] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197-207, 2019.

[5] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "A nested U-Net architecture for medical image segmentation," 2018, *arXiv:1807.10165*.

[6] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998-6008.

[7] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[8] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[9] K. He *et al.*, "Masked autoencoders are scalable vision learners," 2021, *arXiv:2111.06377*.

[10] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. CVPR*, 2021, pp. 6881-6890.

[11] J. Chen *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[12] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," 2021, *arXiv:2102.10662*.

[13] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," 2021, *arXiv:2102.08005*.

[14] W. Wang *et al.*, "TransBTS: Multimodal brain tumor segmentation using transformer," 2021, *arXiv:2103.04430*.

[15] A. Hatamizadeh *et al.*, "UNETR: Transformers for 3D medical image segmentation," in *Proc. WACV*, 2022, pp. 574-584.

[16] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," 2021, *arXiv:2103.03024*.

[17] H. Cao *et al.*, "Swin-UNet: UNet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.

[18] Y. Gao, M. Zhou, and D. N. Metaxas, "UTNet: A hybrid transformer architecture for medical image segmentation," in *Proc. MICCAI*, 2021, pp. 61-71.

[19] Y. Li *et al.*, "GT U-Net: A U-Net like group transformer network for tooth root segmentation," in *Proc. MLMI*, 2021, pp. 386-395.

[20] A. Lin, B. Chen, J. Xu, Z. Zhang, and G. Lu, "DS-TransUNet: Dual swin transformer U-Net for medical image segmentation," 2021, *arXiv:2106.06716*.

[21] B. Chen, Y. Liu, Z. Zhang, G. Lu, and D. Zhang, "TransAttUNet: Multi-level attention-guided U-Net with transformer for medical image segmentation," 2021, *arXiv:2107.05274*.

[23] H. Wang *et al.*, "Mixed transformer U-Net for medical image segmentation," 2021, *arXiv:2111.04734*.

[22] X. Huang, Z. Deng, D. Li, and X. Yuan, "MISSFormer: An effective medical image segmentation transformer," 2021, *arXiv:2109.07162*.

[24] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203-211, 2021.

[25] S. Li *et al.*, "Medical image segmentation using squeeze-and-expansion transformers," 2021, *arXiv:2105.09511*.

[26] Y. Li, W. Cai, Y. Gao, and X. Hu, "More than encoder: Introducing transformer decoder to upsample," 2021, *arXiv:2106.10637*.

[27] D. Karimi, S. Vasylechko, and A. Gholipour, "Convolution-free medical image segmentation using transformers," 2021, *arXiv:2102.13645*.

[28] Y. Chang, H. Menghan, Z. Guangtao, and Z. Xiao-Ping, "TransClaw U-Net: Claw U-Net with transformers for medical image segmentation," 2021, *arXiv:2107.05188*.

[29] G. Xu, X. Wu, X. Zhang, and X. He, "LeViT-UNet: Make faster encoders with transformer for medical image segmentation," 2021, *arXiv:2107.08623*.

[30] S. Tuli, I. Dasgupta, E. Grant, and T. L. Grifiths, "Are convolutional neural networks or transformers more like human vision," 2021, *arXiv:2105.07197*.

[31] B. Yun *et al.*, "SpecTr: Spectral transformer for hyperspectral pathology image segmentation," 2021, *arXiv:2103.03604*.

[32] K. He *et al.*, "Transformers in medical image analysis: A review," 2022, *arXiv:2202.12165*.

[33] Y. Gao, M. Zhuo, D. Liu, and D. Metaxas, "A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks," 2022, *arXiv:2203.00131*.

[34] A. Reza, H. Moein, W. Yuli, and M. Dorit, "Contextual attention network: Transformer meets U-Net," 2022, *arXiv:2203.01932*.

[35] R. Rothe, M. Guillaumin, and L. V. Gool, "Non-maximum suppression for object detection by passing messages between windows," in *Proc. ACCV*, 2014, pp. 290-306.

[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. CVPR*, 2017, pp. 2961-2969.

[37] N. Codella *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.

[38] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1-9, 2018.

[39] K. Sirinukunwattana *et al.*, "Gland segmentation in colon histology images: The GlaS challenge contest," *Med. Image Anal.*, vol. 35, pp. 489-502, 2017.

[40] O. Bernard *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514-2525, 2018.

[41] H. Huisi *et al.*, "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.*, vol. 76, pp. 102327, 2022.

[42] H. Zhou *et al.*, "nnFormer: Interleaved transformer for volumetric segmentation," 2021, *arXiv:2109.03201*.

[43] D. Jha *et al.*, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. ISMM*, 2019, pp. 225-2255.

[44] S. Feng *et al.*, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008-3018, 2020.

[45] B. Lei *et al.*, "Skin lesion segmentation via generative adversarial networks with dual discriminators," *Med. Image Anal.*, vol. 64, pp. 101716, 2020.

[46] Q. Jin, H. Cui, C. Sun, Z. Meng, and R. San, "Cascade knowledge diffusion network for skin lesion diagnosis and segmentation," *Appl. Soft Comput.*, vol. 99, pp. 106881, 2021.

[47] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481-2495, 2017.

[48] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Vishal, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," in *Proc. MICCAI*, 2020, pp. 363-373.

[49] Y. Wu *et al.*, "D-Former: A U-shaped dilated transformer for 3D medical image segmentation," 2022, *arXiv:2201.00462*.

[50] C. Gros, A. Lemay, and L. Cohen-Adad, "SoftSeg: Advantages of soft versus binary training for image segmentation," *Med. Image Anal.*, vol. 71, pp. 102038, 2021.