XBound-Former: Toward Cross-scale Boundary Modeling in Transformers

Jiacheng Wang, Fei Chen, Yuxi Ma, Liansheng Wang, Zhaodong Fei, Jianwei Shuai, Xiangdong Tang, Qichao Zhou, Jing Qin

***Abstract*— Skin lesion segmentation from dermoscopy images is of great significance in the quantitative analysis of skin cancers, which is yet challenging even for dermatologists due to the inherent issues, i.e., considerable size, shape and color variation, and ambiguous boundaries. Recent vision transformers have shown promising performance in handling the variation through global context modeling. Still, they have not thoroughly solved the problem of ambiguous boundaries as they ignore the complementary usage of the boundary knowledge and global contexts. In this paper, we propose a novel cross-scale boundary-aware transformer, XBound-Former, to simultaneously address the variation and boundary problems of skin lesion segmentation. XBound-Former is a purely attention-based network and catches boundary knowledge via three specially designed learners. First, we propose an implicit boundary learner (im-Bound) to constrain the network attention on the points with noticeable boundary variation, enhancing the local context modeling while maintaining the global context. Second, we propose an explicit boundary learner (ex-Bound) to extract the boundary knowledge at multiple scales and convert it into embeddings explicitly. Third, based on the learned multi-scale boundary embeddings, we propose a cross-scale boundary learner (X-Bound) to simultaneously address the problem of ambiguous and multi-scale boundaries by using learned boundary embedding from one scale to guide the boundary-aware attention on the other scales. We evaluate the model on two skin lesion datasets and one polyp lesion dataset, where our model consistently outperforms other convolution- and transformer-based models, especially on the boundary-wise metrics. All resources could be found in `https://github.com/jcwang123/xboundformer`.**

***Index Terms*— Skin lesion segmentation, transformer, boundary modeling, cross-scale.**

Jiacheng Wang is with the Department of Computer Science at School of Informatics, Xiamen University, Xiamen 361005, China. He finished this work during the intern of Manteia Technologies Co.,Ltd. (e-mail:jiachengw@stu.xmu.edu.cn)

Fei Chen, Liansheng Wang are with the Department of Computer Science at School of Informatics, Xiamen University, Xiamen 361005, China. (e-mail:feichen@stu.xmu.edu.cn, lswang@xmu.edu.cn)

Yuxi Ma is with the Xiamen University, Xiamen 361005, China. (e-mail:mayuxi1@stu.xmu.edu.cn)

Zhaodong Fei is with the Department of Radiotherapy, Fujian Cancer Hospital, Fuzhou 350000, China (e-mail:feizhaodong@yeah.net)

Jianwei Shuai is with the Oujiang Laboratory (Zhejiang Lab for Regenerative Medicine, Vision and Brain Health), and Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou, Zhejiang 325001, China. (jianweishuai@xmu.edu.cn)

Xiangdong Tang is with the Sleep Medicine Center, Mental Health Center, Department of Respiratory and Critical Care Medicine, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu 610041, China. (2372564613@qq.com)

Qichao Zhou is with Manteia Technologies Co.,Ltd. (e-mail:zhouqc@manteiatech.com)

Jing Qin is with Center for Smart Health, School of Nursing, The Hong Kong Polytechnic University. (e-mail:harry.qin@polyu.edu.hk)

Liansheng Wang and Qichao Zhou are the corresponding authors.

I. Introduction

Melanoma is one of the most rapidly increasing cancers over the world, consistently leading to about 100,000 new cases and 7000 deaths per year [1], [2]. Segmenting skin lesions from dermoscopy images is critical in the diagnosis and treatment planning, which is usually tedious, time-consuming, and error-prone for human beings. In this regard, automated segmentation methods are highly demanded in clinical practice to improve clinical workflow in terms of accuracy and efficiency. It remains a very challenging task because (1) skin lesions have large size, shape and color variance (see Fig. 1 (a-b)), (2) the hair will partially cover the lesions destroying the local context (see Fig. 1 (c-d)), (3) sometimes, the contrast between lesions to normal skin is relatively low, resulting in ambiguous boundaries (see Fig. 1 (e-h)).

Many efforts have been dedicated to overcoming these challenges. Hand-crafted features are adopted in the early years, which are usually not stable and robust, leading to poor segmentation performance when facing lesions with large variations [3]. To solve this problem, deep learning models based on convolutional neural networks (CNN) have been proposed and achieved remarkable performance gains [4], [5]. However, due to the lack of global context modeling, these models are still insufficient to counteract the large variation of skin lesion segmentation. To enlarge the receptive fields, researchers propose various approaches inspired by the advancement of residual convolution [6], recurrent design [7], and dilated convolution [8], [9]. Lee *et al.* [10] extensively incorporate the dilated attention module with boundary prior so that the network predicts boundary key-points maps to guide the attention module.

Nevertheless, the receptive field of convolution is inevitably limited and the length of recurrent design can not be large. Therefore, these solutions are still incapable of effectively capturing sufficient global context to deal with the challenges mentioned above. Recently, vision transformers have been proposed to regard an image as a sequence of patches and aggregate features in a global manner by self-attention mechanisms [11]–[13]. It is also verified that transformers

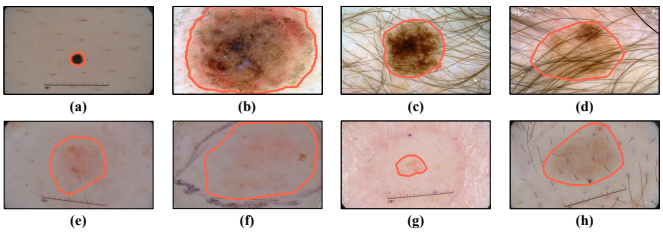

Fig. 1. The challenges of automatic skin lesion segmentation from dermoscopy images: (a)-(b) large skin lesion variations in size, shape, and color, (c)-(d) partial occlusion by hair, and (e)-(h) ambiguous boundaries.

can be used to handle medical image segmentation tasks, i.e., TransUNet [14] and TransFuse [15]. In the field of skin lesion segmentation, studies improve the transformer-based networks with boundary information [16], [17], while they have not thoroughly explored the potential usefulness of boundary information and global context in a multi-scale manner. Furthermore, these transformers still contain convolutional modules that may decrease the performance thanks to the inductive bias.

In this paper, we propose a novel cross-scale boundary-aware transformer (**XBound-Former**) to ably handle the problems mentioned above by holistically leveraging the advancement of boundary-wise prior knowledge and self-attention mechanism. This method is inspired by the intuition that human beings perceive lesions in vision, i.e., considering global context to coarsely locate lesion areas and paying particular attention to the ambiguous area to specify the exact boundary. Concretely, we enhance the boundary modeling ability of the transformer-based network via three key learners: implicit boundary learner (im-Bound), explicit boundary learner (ex-Bound), and cross-scale boundary learner (X-Bound).

- **Im-Bound** is recommended to explore local contexts for accurate boundary modeling implicitly. As the points with large boundary variation contribute more to the segmentation result than other boundary points, we constrain the network attention on such boundary key points. It enhances the local boundary modeling while maintaining the global context.
- **Ex-Bound** is proposed to explicitly extract the boundary knowledge as multiple embeddings where each embedding represents the boundary knowledge at a unique scale. They are used to further enhance the local boundary modeling and boost the cross-scale communication.
- **X-Bound** is suggested as a cross-scale attention mechanism for simultaneously addressing the problems of ambiguous boundaries and size variation. Acting like human beings that determine the accurate boundaries by zooming in and zooming out, we use the learned boundary embedding at one scale to guide the boundary-aware attention at the other scales to enhance the cross-scale knowledge communication.

We evaluate our model on two skin lesion datasets, ISIS-2016&PH2 and ISIC-2018, following the standard experimental setup [10], [16], [17]. To evaluate the generalization, we perform an extensive experiment on the polyp lesion which has closed characteristics. Our model has achieved superior performance in all experiments compared to state-of-the-art CNN-based and transformer-based models, indicating the advanced power in addressing object segmentation with ambiguous boundaries, especially for skin lesion segmentation.

II. Related Work

A. Skin Lesion Segmentation

In the early years, traditional methods apply various hand-crafted features to learn lesion segmentation that are not robust and stable. It leads to poor segmentation performance when facing large lesions with large variations [3]. Later, a fully convolutional network (FCN) [18] brings the deep learning model to skin lesion segmentation and achieves a much better result. Several improved networks following its direction are proposed to solve the imbalance between foreground and background pixels [4], multi-scale feature representations [6], and limited receptive fields [7]. With the widespread use of the attention-based mechanism, channel and spatial attention-based methods are applied to enhance the lesion modeling [19], [20]. The performance indeed reaches a higher score, but skin lesions' ambiguous boundaries are still hard to recognize. To address this issue, [21] propose adaptive dual attention modules to let the network focus on lesion boundaries while it fails to cope with blurry boundaries owing to poor use of boundary-aware prior knowledge. More recently, seeing the excellent success achieved by vision transformers, several studies employ transformer-based networks in the field of skin lesion segmentation [16], [17]. It works to solve the problem of large lesion variation by capturing the global context. However, they are still unable to handle the problem of ambiguous boundaries, especially the ones with size variation. Instead, our proposed **XBound-Former** exploits multi-scale boundary information through the advanced self-attention blocks and utilizes the boundary-aware prior knowledge to supervise the transformer training. Thus it can outperform the State-of-the-Arts and the latest vision transformers.

B. Vision Transformers

Transformer, as a standard model in natural language process [22], has made great progress in the field of computer vision recently. The first vision transformer, ViT [23], proposes to split an image into a certain number of patches and utilize self-attention blocks to embed the features, achieving competitive performance in image classification tasks compared to the latest convolution-based neural networks. Later work [24] introduces a series of strategies to increase the training efficiency and improve the accuracy on small datasets. Although the transformers are originally proposed to explore global dependency, recent studies find that the transformers also need local communication [13], [25], [26], which can be achieved through the local window shift or pyramid architecture [13], especially for the tasks requiring dense representations [25], [27], [28]. As for medical image segmentation, the effectiveness of vision transformers is verified by TransUNet [14] and TransFuse [15]. In the field of skin lesion segmentation, vision transformers also boost the performance to reach new higher scores [16], [17]. Despite their success, these models have

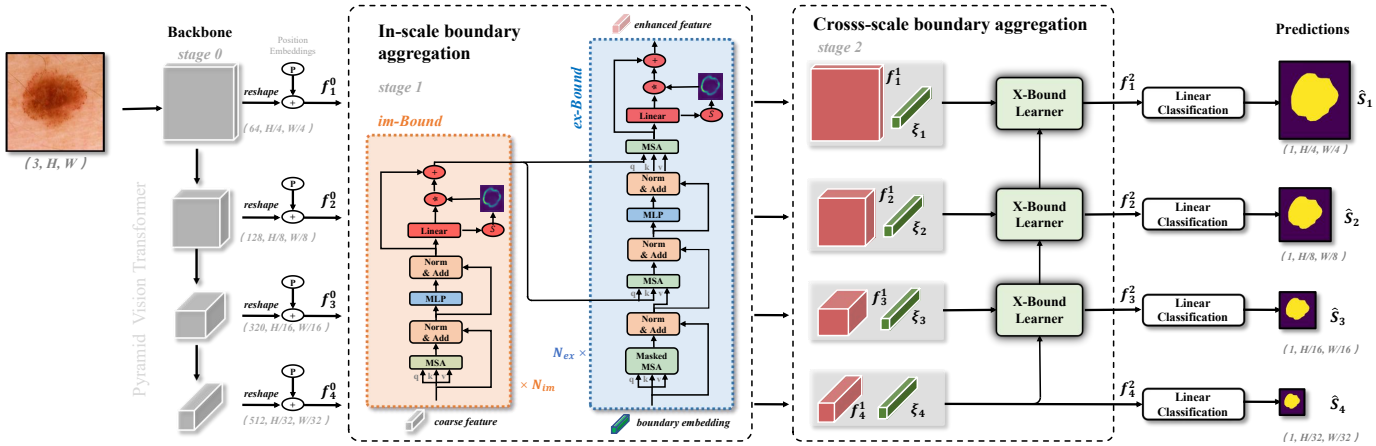

Fig. 2. An overview of the proposed cross-scale boundary-aware transformer (XBound-Former). Briefly, it enhances the boundary modeling of transformers via in- and cross-scale boundary aggregation. Given coarsely extracted features ($\{f_l^0\}_{l=1}^4$, gray cuboids), (1) the in-scale boundary aggregation implicitly (im-Bound) and explicitly (ex-Bound) explores boundary knowledge, enhancing features ($\{f_l^1\}_{l=1}^4$, red cuboids) and obtaining boundary embeddings ($\{\xi\}_{l=1}^4$, green cuboids) at each scale, and (2) the cross-scale boundary aggregation further exploits boundary information by fusing learned boundary embeddings from different scales and yielding final enhanced features ($\{f_l^2\}_{l=1}^4$). After that, several classification heads are used to predict the segmentation maps. Note that the MSA takes the query/key/value vectors (from left to right) as the input. Since the MSA in ex-Bound learners utilizes boundary embeddings to refine the features while the MSA in im-Bound learners refines the features through the self-attention, their inputs are different.

not considered the complementary knowledge of boundary knowledge and global context in a multi-scale manner, which may help segment the extremely challenging lesions. **XBound-Former** aims to mitigate this issue through cross-scale boundary learners and, besides, builds a pure attention-based network instead of the fusion of transformer and convolution to prevent the inductive bias.

C. Boundary-aware Prior Knowledge

The accurate recognition of ambiguous boundaries is one of the most tricky problems in medical image segmentation. There are plenty of works to address this issue by taking full advantage of the boundary-aware prior knowledge. The earliest works propose to modify the loss function to give boundary-aware supervision for network optimization, i.e., HD loss [29], Boundary loss [30], etc. Later, multi-task learning is applied in this direction where manually designed tasks are used to provide extra supervision on the boundaries [31], [32]. Apart from the boundary-aware supervision, several networks propose to utilize spatial attention mechanisms to enhance the representation of boundaries [21]. By contrast, we not only introduce the boundary-aware prior knowledge into vision transformers but also present a novel key-patch map generator that can select the most ambiguous points among the boundaries and convert them to the key-patch map to give supervision to the transformers.

III. Method

An overview of the cross-scale boundary-aware transformer (**XBound-Former**) is presented in Fig. 2, where we show the details about how to leverage boundary prior knowledge and global dependency across different scales. It first utilizes a pyramid vision transformer [13] to coarsely extract the features of an input dermoscopy skin image. As a pyramid feature extractor, the backbone yield features at four different scales, $\{f_l^0\}_{l=1}^4$, where l denotes the layer number. Here, f_1^0 denotes the lowest feature with the largest scale and f_4^0 denotes the deepest feature with the smallest scale, as $f_l^0 \in \mathbb{R}^{C^l \times \frac{H}{2^{l+1}} \times \frac{W}{2^{l+1}}}$, where H, W is the size of input images and C^l is the channel number. Each feature will be enhanced through the in-scale and cross-scale boundary aggregation to strengthen the boundary representation. Finally, several linear classification heads are used to predict the segmentation maps ($\{\hat{S}\}_{l=1}^4$), and the map with the largest size ($\hat{S}_1$) will be upsampled to the original size by bi-linear interpolation.

A. In-scale Boundary Modeling

As an attention-based mechanism, transformers treat each image as a sequence of patches and explore the global dependency to represent them. The global view is precisely helpful for the vision tasks, while recent studies have shown that they also require local context modeling in the dense-level vision tasks [13], [25]. For the segmentation tasks, especially for skin lesions with ambiguous boundaries, global dependency can help locate coarse boundary but lacks local contexts to segment accurate boundaries. Therefore, we propose to fuse boundary information in the transformers to explore the local context of boundaries. It is achieved by using a sequence of N_{im} implicit boundary learners (im-Bound) and N_{ex} explicit boundary learners (ex-Bound) to refine the feature at each scale as $\{f_l^1\}_{l=1}^4$, where $f_l^1 \in \mathbb{R}^{C^l \times \frac{H}{2^{l+1}} \times \frac{W}{2^{l+1}}}$. The process is denoted as,

$$f^1, \xi \longleftarrow \mathcal{F}_{ex\text{-}Bound}^{1\ldots N_{ex}}\left(\mathcal{F}_{im\text{-}Bound}^{1\ldots N_{im}}\left(f^0\right)\right), \tag{1}$$

where we simplify the notation l. As the in-scale boundary modeling module takes the sequential features instead of 2-D maps as inputs and outputs, we re-define the inputted features as $z_l \in \mathbb{R}^{C^l \times \frac{H*W}{4^{l+1}}}$. They are the encoded features after sequentialization and are added with position embeddings [33].

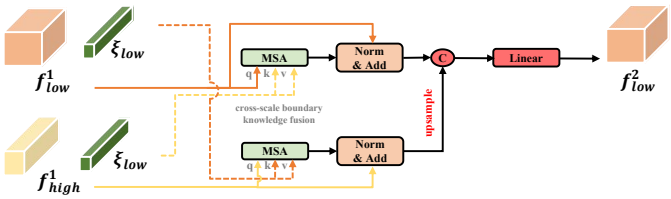

Fig. 3. Details of the cross-scale boundary learner (x-Bound). It utilizes the boundary embedding at one scale to guide the multi-head self-attention (MSA) module at another scale. The enhanced features are combined and projected to keep the same dimension with the feature at a low level (f^1_{low}) via a linear projection. Note that the dashed lines denote the key/value branches for boundary embeddings and the orange/yellow colors denote features at two scales.

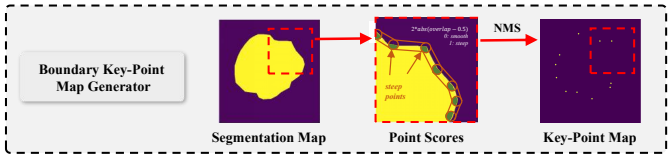

Fig. 4. The pipeline of the boundary key-point map generation. It aims to transform the ground-truth segmentation map into the key-point map for supervising boundary learners.

1) Implicitly Boundary-wise Attention: The **im-Bound** aims to constrain the model's attention on the points with large boundary variation as they contribute more to the final segmentation result. With this inspiration, we propose to utilize the self-attention module to find such points in the manner of predicting boundary key-point map. The map is used for the feature refinement and offering a boundary-aware constraint. Specifically, it contains N_{im} cascaded blocks in total. Assumed that at the i-th block, given the inputted feature as z^{i-1}, where $z^0 \leftarrow z$, we firstly feed it into a sequence of multi-head self-attention (MSA) and multi-layer perception (MLP) to gather the global dependency for coarsely locating the boundaries [23]. After each part, there is a Layer Normalization with residual short connection for a stable training process [22]. We denote this intermediate feature as,

$$\rho^i = \mathcal{F}_{MSA}(z^{i-1}) \oplus \mathcal{F}_{MLP}(\mathcal{F}_{MSA}(z^{i-1})), \tag{2}$$

where $\oplus$ denotes the element-wise addition and $\mathcal{F}_{MSA}(query, key, value)$ denotes the MSA operation. As the self-attention modules embed *query*, *key* and *value* together from z^{i-1}, we simplify the equation. Additionally, the LayerNorm operation is also simplified to save space. Then, a linear predictor with Sigmoid activation is utilized to classify each patch whether it is the point with large boundary variation, supervised by the boundary key-point map pre-produced by our boundary key-point map generation algorithm (see Sec. III-C). We denote the predicted key-point map as $\hat{M}^i$ so that we could obtain the enhanced feature as,

$$z^i = \rho^i \ \oplus \ (\ \rho^i \otimes \hat{M}^i\ ), \tag{3}$$

where $\oplus$ denotes the element-wise multiplication. After N_{im} cascaded blocks, the resulted feature $z^{N_{im}}$ will be sent to ex-Bound for further refinement.

2) Learn Explicit Boundary Embedding: The **ex-Bound** is proposed to embed boundary information into a set of feature vectors explicitly $\{\xi_l\}_{l=1}^4$, where each embedding contains the high-level boundary semantics at a unique scale. This learner is different from the **im-Bound** regarding the implementation, as well as the motivation that it not only refines the features but also provides the explicit expression for subsequent cross-scale communication. To achieve this goal, we treat the boundary key points as query objects and employ a transformer decoder [27], [34] to learn the boundary embeddings. The decoder contains a sequence of the Masked MSA module, MSA module, and the MLP module, each after which there is a LayerNorm layer and the short connection [27]. It is noteworthy that the Masked MSA is equal to the MSA module here since the boundary embeddings have a fixed length. Thanks to the global context modeling, it refines the inputted randomly initialized vector into the boundary embedding that contains abundant boundary knowledge. After that, we send the feature and boundary embedding into the MSA module and the boundary key-point prediction part for the consideration of refining features and, of more importance, obtaining a preciser boundary embedding.

We repeat the ex-Bound N_{ex} times to guarantee the adequate boundary learning. For the j-th block, it takes feature $z^{N_{im}+j-1}$ and current embedding ξ^{j-1} as input and output the aggregated feature $z^{N_{im}+j}$, the embedding ξ_j, and predicted key-point map $\hat{M}^{N_{im}+j}$. After N_{ex} blocks, the resulted feature $z^{N_{im}+N_{ex}}$ is reshaped as f^1 and sent to the cross-scale boundary aggregation along with the learned boundary embedding.

B. Attention-based Cross-scale Boundary Fusion

Automatic skin lesion segmentation suffers from the significant variance in lesion size and ambiguous boundaries. We take the first attempt to address these two issues simultaneously through the attention-based mechanism, our cross-scale boundary learners (**X-Bound**). It is inspired by the human beings that determine the accurate boundaries by zooming in and zooming out boundaries and combining multi-perspective information across different scales to make the final decision.

Generally, we visualize the details in Fig. 3 where features and boundary embeddings at low scale ($f^1_{low} \in \mathbb{R}^{C\times H_{low}\times W_{low}}, \xi_{low} \in \mathbb{R}^{1\times C}$) and high scale ($f^1_{high} \in \mathbb{R}^{C\times H_{high}\times W_{high}}, \xi_{high} \in \mathbb{R}^{1\times C}$) are inputted and the enhanced feature at low scale (f^2_{low}) is outputted. (H_{low}, W_{low}) denotes the size larger than (H_{high}, W_{high}). Theoretically, the boundary embedding at a lower scale focuses on more local details and the boundary embedding at a larger scale focuses more on the high-level semantics. Thus, utilizing the embedding at one scale to attentively refine the features at another scale provides complementary boundary knowledge.

In detail, we compare ξ_{high} to each point in the lower feature f^1_{low} and compute the distance matrix, which is then used to transfer boundary knowledge in ξ_{high} to each point in the feature f^1_{low}. It means that the intermediate features can be calculated as:

$$\gamma_{low} = f^1_{low} \oplus \mathcal{F}_{MSA}\ (f^1_{low}\ ,\ \xi_{high},\ \xi_{high}), \tag{4}$$

$$\gamma_{high} = f^1_{high} \oplus \mathcal{F}_{MSA}\ (f^1_{high}\ ,\ \xi_{low}\ ,\ \xi_{low}), \tag{5}$$

where $\mathcal{F}_{MSA}$ is the multi-head attention module used in Equation 2. After that, the intermediate features are concatenated after the up-sample operation of f_{high}^1, which is fed into a linear projection head to reduce the feature dimension and refine the fusion. The resulted feature is denoted as f_{low}^2.

Totally, except the deepest feature f_4^1, we perform the cross-scale boundary learning on $\{f_l^1\}_{l=1}^3$ to obtain $\{f_l^2\}_{l=1}^3$ and $\{f_4^2\}$ is straightly set as $\{f_4^1\}$. For the consideration of multi-scale model learning, we feed each feature into a linear classification head to predict the segmentation maps $\{\hat{S}_l\}_{l=1}^4$.

C. Boundary Key-point Generation Algorithm

As the boundary learners do not naturally know which points can best represent the ambiguous boundaries, we propose a novel generation algorithm to pre-produce a ground-truth key-point map supervising the boundary point map prediction and boundary embedding learning, as shown in Fig. 4. The first step is to calculate all points on the boundary using a conventional contour detection algorithm [35]. After that we could obtain a set of coordinates of the boundary points. Then, as points with larger boundary deviation should be paid more attention to than those with smoother deviation, we propose filtering the points by scoring the deviation. For each point in this set, we draw a circle of radius r and calculate the proportion p of the lesion area in this circle region, where the larger or smaller p indicates that the boundary is not smooth in this circle region. Hence, we score each point as $|p - 0.5|$ to representation its deviation. To find the most valuable points, non-maximum suppression is performed in which the points with larger p than neighbor k points are selected. Specifically, given the sorted boundary point list, we denote the neighborhoods of each point as the k points before the point and the k points after the point, totally resulting in $2k$ points. Note that points at the beginning and the end of the list are connected. If one point's score is larger than that the neighbour $2k$ points' scores, it will be saved in the list, otherwise it will be removed from the list. Next, selected points' 2D locations are mapped into the binary key-point map M, where points at the selected location are set to one and others are set to zero. By minimizing the error between M and $\hat{M}$, the supervision helps the boundary learners focus on the ambiguous boundary regions and helps the boundary embeddings learn correct boundary knowledge.

D. Objective Function

We design a joint objective to train the entire network, including the lesion segmentation loss L_{Seg} for predicted segmentation maps and the key-point map loss L_{Map} for predicted boundary key-point maps, as

$$L_{Seg} = \frac{1}{4}\sum_{l=1}^{4}\phi_{Dice}(\hat{S}_l, S_l), \tag{6}$$

$$L_{Map} = \frac{1}{N_{im}+N_{ex}}\sum_{l=1}^{4}\phi_{CE}(\hat{M}_l, M_l), \tag{7}$$

$$L_{Total} = L_{Seg} + \lambda L_{Map}, \tag{8}$$

where ϕ_{DICE}, ϕ_{CE} denote Dice loss function and Cross Entropy function and $\{S_l\}_{l=1}^4, \{M_l\}_{l=1}^4$ are the ground-truth segmentation and boundary key-point maps pre-produced. λ is the weight to balance the two objectives. Moreover, we computed the averaged loss of segmentation maps and boundary maps for better controlling the weight and intuitively indicating the performance. The detailed calculation is described as,

$$\begin{aligned}\phi_{Dice}(\hat{S}, S) &= 1 \ - \ 2 * |S| * |\hat{S}| \ / \ (|S| + |\hat{S}|),\\ \phi_{CE}(\hat{M}, M) &= -\hat{M}\log(M) - (1-\hat{M})\log(1-M).\end{aligned} \tag{9}$$

For deeply multi-scale supervision, given the original segmentation label, $S \in \mathbb{R}^{1\times H\times W}$, we repeat the down-sample operation with different rates to obtain the set of ground-truth segmentation maps $\{S_l\}_{l=1}^4$, where $S_l \in \mathbb{R}^{1\times \frac{H}{2^{l+1}}\times \frac{W}{2^{l+1}}}$. For the key-point maps, we also repeat the down-sample operation and obtain $\{M_l\}_{l=1}^4$ where $M_l \in \mathbb{R}^{1\times \frac{H}{2^{l+1}}\times \frac{W}{2^{l+1}}}$.

IV. Experiments

A. Dataset

Following the classical experimental setting in the previous studies [10], we evaluate our model on two skin lesions segmentation datasets, **ISIC-2016&PH^2** and **ISIC-2018**. To further evaluate the model generalization, we evaluate it on the polyp lesion segmentation using five public polyp image datasets, named **Polyp-seg**.

- The **ISIC-2016&PH^2** contains samples from two centers to evaluate the accuracy and generalization ability of skin lesion segmentation. One is the ISIC-2016 dataset that contains a total number of 900 samples for training and 379 samples for validation. The other one is the PH^2 dataset [36], containing 200 samples in total. Here, we use samples in the ISIC-2016 dataset for model learning through the official *train-validation* split and test the model on the 200 samples from the PH^2 dataset.
- The **ISIC-2018** dataset was also collected by ISIC in 2018, which contains 2594 images and labels. The resolution of each image varies from 720×540 to 6708×4439. As the public test set has not been released, we perform a 5-fold cross-validation for a fair comparison.
- The **Polyp-seg** dataset is collected following the most popular setting [37], which contains five public datasets: Kvasir-SEG [38], ClinicDB [39], ColonDB [40], Endoscene [41], and ETIS [42]. The Kvasir-SEG and ClinicDB contain 612 and 1000 samples, respectively, of which 548 and 900 samples are used for training and the rest samples are used for testing. To evaluate the generalization ability, samples from the rest three datasets are also used for testing.

B. Evaluation Metrics

We employ four widely-used metrics to quantitatively evaluate the skin lesion segmentation performances, including $Dice$ coefficient, IoU score, Average symmetric surface distance ($ASSD$), and Hausdorff distance of boundaries (95^{th} percentile; $HD95$). Generally, a better segmentation performance

TABLE I

COMPARISON OF SKIN LESION SEGMENTATION WITH DIFFERENT APPROACHES ON THE ISIC-2016&PH^2 DATASET. WE REPORT THE AVERAGED SCORES OF THE ISIC-2016 VALIDATION SET AND THE SCORES OF THE PH^2 TEST SET.

Method	*validation*-ISIC-2016 [43]				*test*-PH^2 [36]			
	IoU↑	Dice↑	ASSD↓	HD95↓	IoU↑	Dice↑	ASSD↓	HD95↓
U-Net [44]	80.25	87.81	15.51	45.88	73.91	83.66	21.50	60.12
U-Net++ [45]	81.84	88.93	15.01	44.83	81.26	88.99	15.97	46.66
Polar Res-UNet++ [46]	69.21	80.25	20.35	46.24	75.23	85.36	18.55	42.96
Attention U-Net [47]	79.70	87.43	16.41	48.78	69.52	80.52	26.73	74.51
DeepLabV3+ [48]	85.62	91.76	9.85	26.66	82.03	89.56	14.93	37.81
CE-Net [49]	84.39	90.74	11.77	31.01	83.48	90.44	13.48	33.97
CA-Net [20]	80.73	88.10	15.67	44.98	75.18	84.66	21.06	64.53
TransFuse [15]	86.19	92.03	10.04	30.33	82.32	89.75	15.00	39.98
TransUNet [14]	84.89	91.26	10.63	28.51	83.99	90.96	12.65	33.30
XBound-Former (Ours)	**87.69**	**93.08**	**8.21**	**21.83**	**85.38**	**91.80**	**10.72**	**26.00**

TABLE II

COMPARISON OF SKIN LESION SEGMENTATION WITH DIFFERENT APPROACHES WITH 5-FOLD CROSS-VALIDATION ON ISIC-2018 DATASET. WE PRESENT THE AVERAGED RESULT AND THE STANDARD ERROR OF ALL FOLDS.

Method	Overall				Fold-1				Fold-2				Fold-3				Fold-4				Fold-5			
	IoU↑	Dice↑	ASSD↓	HD95↓	IoU↑	Dice↑	ASSD↓	HD95↓	IoU↑	Dice↑	ASSD↓	HD95↓	IoU↑	Dice↑	ASSD↓	HD95↓	IoU↑	Dice↑	ASSD↓	HD95↓	IoU↑	Dice↑	ASSD↓	HD95↓
U-Net [44]	75.40	83.53	17.74	53.98	76.50	84.64	16.66	51.35	77.84	85.44	15.43	46.79	73.90	82.32	19.76	60.87	74.38	82.54	18.36	55.77	74.40	82.72	18.50	55.11
U-Net++ [45]	77.92	85.77	16.13	49.90	79.22	86.96	13.83	42.00	79.28	86.76	14.72	44.99	76.30	84.46	19.24	60.40	77.07	84.92	15.70	48.20	77.72	85.74	17.14	53.93
Polar Res-UNet++ [46]	60.50	72.61	21.06	47.55	61.48	73.45	19.51	43.13	61.30	73.18	20.84	47.27	60.32	72.60	22.12	50.63	59.15	71.37	21.59	48.54	60.28	72.45	21.27	48.17
Attention U-Net [47]	75.94	83.88	17.24	52.56	78.01	85.93	15.13	48.86	77.31	85.05	16.04	50.51	73.78	81.98	19.16	58.00	75.58	83.43	17.15	51.89	75.01	82.99	18.73	53.56
DeepLabV3+ [48]	82.49	89.38	10.75	28.64	84.33	90.79	8.57	23.38	84.05	90.44	9.81	27.30	82.01	89.18	10.90	29.28	82.57	89.42	10.50	27.63	79.49	87.10	14.00	35.64
CE-Net [49]	82.86	89.62	10.59	28.69	84.10	90.70	8.83	23.77	83.75	90.16	10.38	28.46	81.45	88.60	12.12	32.80	82.81	89.62	10.34	28.20	82.18	89.02	11.26	30.24
CA-Net [20]	78.94	86.56	13.52	39.90	80.56	87.76	11.70	33.61	79.53	86.94	13.21	37.75	77.88	85.85	14.09	42.77	78.55	86.32	14.06	44.04	78.20	85.94	14.54	41.31
TransFuse [15]	83.59	90.13	10.21	28.33	85.05	91.23	8.80	25.61	84.45	90.74	10.05	28.52	82.52	89.46	11.07	30.10	83.29	89.81	10.01	27.46	82.66	89.42	11.13	29.97
TransUNet [14]	82.61	89.50	10.88	29.05	83.53	90.10	9.72	25.31	83.90	90.41	10.00	27.77	81.67	88.84	11.86	31.15	82.37	89.29	10.77	29.32	81.58	88.84	12.06	31.71
XBound-Former (Ours)	**84.51**	**90.89**	**8.61**	**22.47**	**85.31**	**91.47**	**7.67**	**20.05**	**85.22**	**91.33**	**8.48**	**22.29**	**84.12**	**90.64**	**8.92**	**23.31**	**84.00**	**90.43**	**8.87**	**23.14**	**83.88**	**90.56**	**9.13**	**23.56**

TABLE III

COMPARISON OF POLYP LESION SEGMENTATION. WE REPORT THE OVERALL $\boldsymbol{IoU}$ AND $\boldsymbol{F_\beta^w}$ SCORES AND THE SCORES OF EACH DATASET.

Method	Overall		Kvasir-SEG		ClinicDB		ColonDB		ETIS		Endoscene	
	IoU↑	F_β^w ↑	IoU↑	F_β^w ↑	IoU↑	F_β^w ↑	IoU↑	F_β^w ↑	IoU↑	F_β^w ↑	IoU↑	F_β^w ↑
ACSNet [50]	65.07	69.08	83.80	88.20	82.60	87.30	63.10	68.40	49.60	50.60	78.80	83.00
PraNet [37]	67.52	72.41	84.00	88.50	84.90	89.60	64.00	69.90	56.70	60.00	79.70	84.30
TGANet [51]	64.16	71.40	81.60	87.79	81.09	86.71	60.08	67.94	52.33	59.36	82.18	89.47
MSEG [52]	66.68	70.75	83.90	88.50	86.40	90.70	64.90	69.70	50.90	53.00	80.40	85.20
DCRNet [53]	70.00	74.95	82.50	86.80	84.40	89.00	66.60	72.40	63.00	67.10	78.70	82.50
EU-Net [54]	70.41	74.55	85.40	89.30	84.60	89.10	68.10	73.00	60.90	63.60	76.50	80.50
SANet [55]	71.38	76.09	84.70	89.20	85.90	90.90	67.00	72.60	65.40	68.50	81.50	85.90
Polyp-PVT [56]	76.00	81.60	86.40	**91.10**	88.90	93.60	72.70	79.50	70.60	75.00	83.30	88.40
XBound-Former (Ours)	**77.50**	**84.30**	**87.10**	89.70	**91.10**	**94.40**	**73.20**	**81.40**	**75.10**	**82.00**	**83.60**	**90.10**

shall have higher area-based metrics ($Dice, IoU$) and lower boundary-based metrics ($ASSD, HD95$).

The area-based similarity of predicted segmentation map $\hat{S}$ and the ground-truth S are computed as:

$$\begin{aligned} \psi_{Dice}(\hat{S}, S) &= 2 * \frac{|\hat{S} * S|}{|\hat{S}| + |S|}, \\ \psi_{IoU}(\hat{S}, S) &= \frac{|\hat{S} * S|}{|\hat{S}| + |S| - |\hat{S} * S|}. \end{aligned} \tag{10}$$

To better evaluate the segmentation performance of boundaries, we employ another two boundary-based metrics, as

$$\begin{aligned} \psi_{ASSD}(\hat{S}, S) &= \frac{\sum_{a \in P_b} d\left(a, G_b\right) + \sum_{b \in G_b} d\left(b, P_b\right)}{|P_b| + |G_b|}, \\ \psi_{HD95}(\hat{S}, S) &= \max\left\{h(P_b, G_b), h(G_b, P_b)\right\}, \end{aligned} \tag{11}$$

where P_b and G_b denote the predicted boundary points and the ground-truth boundary points in the $\hat{S}$ and S, and $d(\cdot)$ denotes the minimum Euclidean distance function. Moreover, $h(P_b, G_b) = \mathop{\hat{\max}}_{a \in P_b} \left\{ \min_{b \in G_b} \|a - b\| \right\}$ denotes the one-way hausdorff distance from P_b to G_b, and $\hat{\max}\{\cdot\}$ refers to the calculation of the 95^{th} percentile of the distances.

As for the polyp segmentation, we adopt the same metrics as the latest work, Polyp-PVT [56], including the area-based metric, IoU, and the boundary-based metric, F_β^w.

C. Implementation Details

All methods are implemented on the Pytorch with a single NVIDIA Geforce GTX 3090 GPU with a memory of 24 GB. We empirically resize all images to (512×512) considering the computation efficiency and don't keep the original ratio as it will not break the lesion appearance [10], [16], [17], [57]. A series of data augmentations are implemented to

increase the data diversity, including vertical flip, horizontal flip, and random scale change (limited 0.9-1.1). Each mini-batch includes eight images, and the AdamW [58] optimizer with an initial learning rate of 0.0003 is used to optimize the parameters. We train the network for 200 epochs and save the model parameters with the best performance during validation. We adopt the pyramid vision transformer, PVTv2 [13], as the backbone and pre-train it on the ImageNet dataset. As for the hype-parameters, we set N_{im} and N_{ex} to 2 by default and discuss it in Section IV-F.3. In the boundary key-point generation algorithm, considering the image size and lesion size, we set $r = 2$ and $k = 30$ by default.

D. Comparisons with state-of-the-art Methods

1) Quantitative results for skin lesion segmentation: We compare our model to several popular segmentation models, including the CNN-based models, U-Net [44], U-Net++ [45], Polar Res-UNet++ [46], Attention U-Net [47], DeepLabV3+ [48], CE-Net [49], CA-Net [20], and the transformer-based models, TransFuse [15] and TransUNet [14]. All compared models are trained under the same experimental setting as our model. For the models using their special training manners, i.e., transformed data [46], we re-implement them using their official resources and follow the best hyper-parameters claimed in their manuscripts.

For the ISIC-2016&PH^2 dataset, it is found that our model has achieved the best performance on whatever the validation set or the test set. Since the samples from the PH^2 dataset are unseen during the model learning, our superior performance indicates the satisfactory generalization ability, which is owing to the learning of boundaries that are the general features among different distributions. In comparison to us, TransFuse generalize poorly to the test set and TranUNet has poor segmentation accuracy on the validation set. Furthermore, it is seen that our model has obviously lower $ASSD$ (-1.83 and -1.93) and $HD95$ (-6.68 and -7.30), demonstrating the promising advantage in handling boundary segmentation.

To extensively evaluate the models, we perform the 5-fold cross-validation in the ISIC-2018 dataset and show the evaluated scores of each fold as well as the overall scores in Table II. The results illustrate that our model achieves the highest IoU score and the shortest ASSD distance on all sets. In addition to this, although the improvement on the IoU score is not as large as that on the ISIC-2016&PH^2 dataset, the ASSD score has decreased a lot compared to the other models. It means that our model has superior accuracy in reducing the false positives away from the boundaries and detecting the ambiguous boundaries that are ignored by other models.

2) Visualized Comparison for Skin Lesion Segmentation: We visualize the predictions of some representative images in Fig. 5, including the lesions with hair occlusion, various sizes, and ambiguous boundaries. The first row shows that our model can detect the lesion covered by the hair with the largest accuracy. The second and third rows prove that our model consistently yields stable and the best prediction on the smallest or largest lesions. For all rows, particularly the last two rows where lesions show an extremely close appearance to neighbor tissues, our model is still able to give accurate segmentation.

3) Evaluation for Model Generalization: In order to further study the generalization of our proposed model and show our potentials in other similar targets with ambiguous boundaries, we conduct experiments on the polyp image segmentation and compare our model to the most popular models in this field, including ACSNet [50], PraNet [37], TGANet [51], MSEG [52], DCRNet [53], EUNet [54], SANet [55] and Polyp-PVT [56]. Since we follow the experiment setting in Polyp-PVT [56] and the results of compared methods are all presented in it, we straightly show the publicly presented scores. In difference, as TGANet is not included in Polyp-PVT, we re-implement TGANet using its official resources and adopt the best hyper-parameters claimed in its manuscript.

We show the compared results in Table III, where the overall scores and the scores of each dataset are presented. We highlight the best score in bold, and it is found that our model nearly achieves the best scores on all metrics. For overall performance, compared to the latest model, Polyp-PVT, which has also used PVTv2 as the backbone, our model yields obvious performance improvement, i.e., 1.5% on the IoU score and 2.7% on the F_β^w score. As F_β^w demonstrates the ability of accurate boundary segmentation, the result indicates that our boundary learners are genuinely able to enhance the determination of boundary points. The results on each dataset also support the conclusion, especially for the ETIS dataset. Samples from the ETIS dataset are more challenging to segment, leading to relatively poorer performance in all experiments. On such a difficult sampler, our model has a 4.5% improvement on the IoU score and 7.0% improvement on the F_β^w score, indicating its superior ability to handle challenging boundaries.

E. Analytical Ablation Study

We conduct extensive ablation experiments on the ISIC-2016&PH^2 and ISIC-2018 datasets to demonstrate the effectiveness of the three bound learners in our proposed method. We analyze the performance of the validation and test sets for the ISIC-2016&PH^2 dataset and discuss the results of all folds for the ISIC-2018 dataset. For the baseline comparison, we remove the learners of XBound-Former and maintain the same linear prediction and up-sampling fusion as U-Net. Then, we add the im-Bound learners, ex-Bound learners, and X-Bound learners step by step and obtain three models that are the imBound-Former, exBound-Former, and XBound-Former.

1) Quantitative Analysis: The results of the ablation experiment are shown in Fig. 6(a) using bar plots, and the evaluated IoU scores are highlighted by red scores. Compared to the baseline model, imBound-Former has gained a 0.76% improvement on the ISIC-2016 validation set, 1.12% improvement on the PH^2 test set, 0.57% improvement on the ISIC-2018 dataset, verifying that the implicit boundary modeling and attention truly benefit the segmentation accuracy and generalization. In addition, exBound-Former gains further improvements on the ISIC-2016 validation set (0.78%), the PH^2 test set (0.24%), and the ISIC-2018 dataset (0.53%).

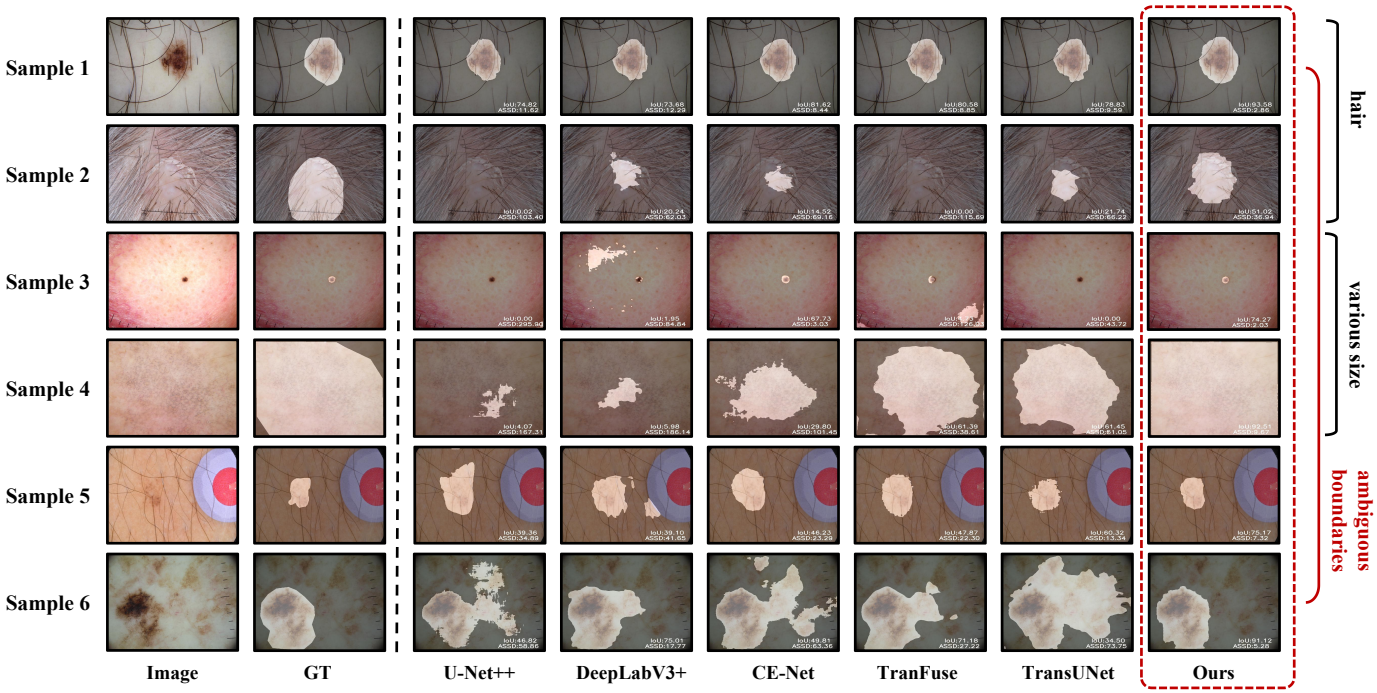

Fig. 5. Visual comparison on some representative and challenging images (Samples 1,3,5 are from the ISIC-2016&PH^2 dataset and the rest are from the ISIC-2018 dataset). It includes the tricky lesions caused by hair occlusion, size variance, and especially the ambiguous boundaries. We show the specific IoU and ASSD scores in each visualized imag e at the right-bottom corner.

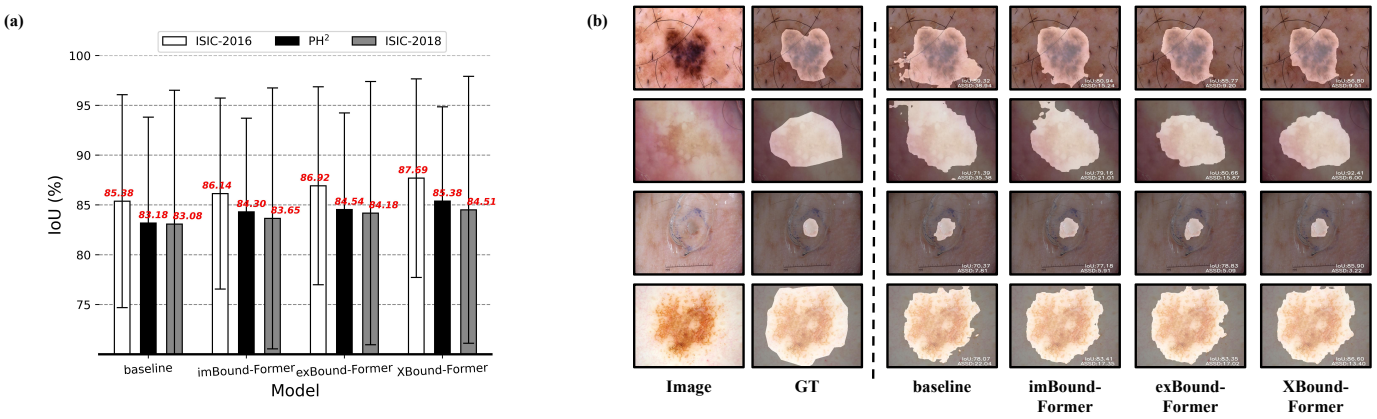

Fig. 6. Ablation analysis of the boundary learners through adding the im-Bound learners, ex-Bound learners, and X-Bound learners step by step to the baseline model. (a) Quantitative analysis of the ISIC-2016 validation set, PH^2 test set, and ISIC-2018 dataset. The specific IoU score on each set is shown upper the plotted bar. (b) Visual comparison of some representative images. The first two rows are selected from the PH^2 test set and the last two rows are from the ISIC-2018 dataset. We show the specific IoU and ASSD scores at the right-bottom corner of the visualized image.

Since the ex-Bound learnders majorly aim to learn explicit embeddings for boundary knowledge which are used for the cross-scale boundary learning, the improvement is slight yet not important. The complete version, XBound-Former, shows obvious and consistent improvements on all sets, verifying the usefulness of our attention-based cross-scale boundary fusion.

2) Visual Comparison on Lesion Boundaries: We also visually analyze the effectiveness of each component in Fig. 6, including two samples from the PH^2 test set (the first two rows) and two samples from the ISIC-2018 dataset (the last two rows). As it shows, the baseline model lacks sufficient ability to address lesions with ambiguous boundaries as there are a lot of false positives. This issue has decreased significantly in the predictions of imBound-Former and exBound-Former, while the determination is still not accurate enough. By combining the multi-scale boundary knowledge, XBound-Former achieves the best performance on the small lesion (the third row) and the large lesions (the second and last rows).

F. Detailed Analysis of Bound Learners

1) Boundary Supervision: As shown in Equation 6, we utilize the factor (λ) to balance the segmentation map loss and boundary key-point map loss. The smaller λ may fail to provide strong enough supervision, while the larger λ may sometimes bring the noise to the model learning. Hence, we have a discussion about how it affects the final segmentation performance. The results are shown in Fig. 7, where λ is set to

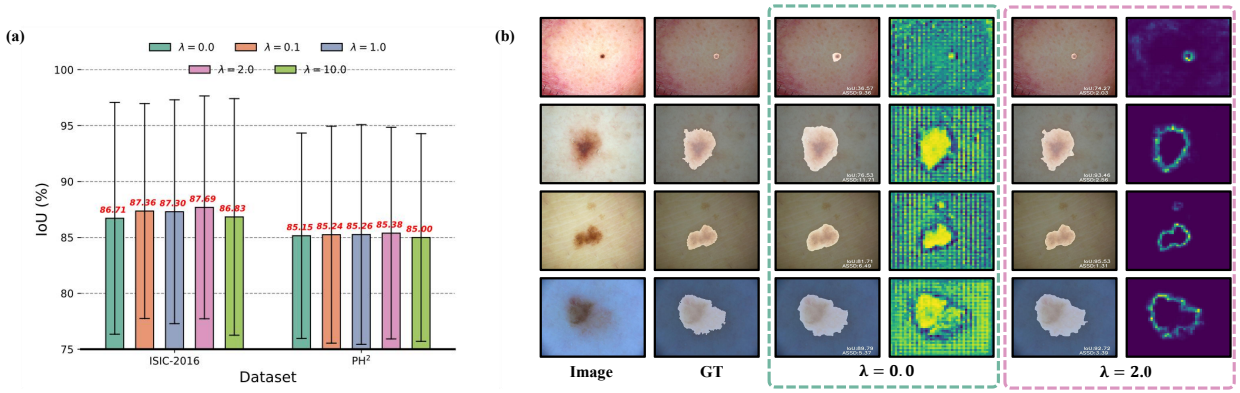

Fig. 7. Analysis of how the boundary key-point supervises segmentation learning. (a) Evaluated scores with different controlling weight, λ. (b) Visual comparison of using key-point supervision or not.

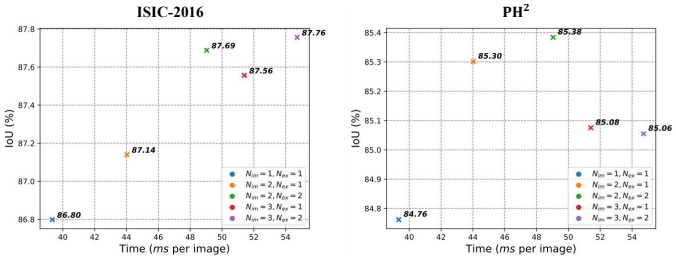

Fig. 8. Inference speed analysis for models with different N_{im}, N_{ex} on the ISIC-2016 validation set (left) and PH² test set (right). We measure the speed by calculating the inference time per image.

TABLE IV
THE IOU SCORES OF USING NMS OPERATOR TO FILTER THE BOUNDARY POINTS OR NOT. WE SHOW THE AVERAGED SCORES AND STANDARD ERRORS ON THE ISIC-2016 VALIDATION SET AND PH² TEST SET.

Method	ISIC-2016	PH^2
w/o NMS	87.19 ± 10.37	85.06 ± 10.04
with NMS	87.69 ± 9.97	85.38 ± 9.46

$\{0, 0.1, 1, 2, 10\}$ and all models adopt the same architecture as XBound-Former. As the plot shows in Fig. 7(a), the evaluated scores increase on both sets when enlarging the λ from 0.0 to 2.0. However, they decrease when the λ reaches 10.0. It verifies the assumption that the small λ limits the improvement and the large one will harm the segmentation training. We additionally visualize the predicted segmentation map along with the point map in Fig. 7(b). As it shows, the model without boundary supervision is still able to predict coarse lesion regions for spatial attention while it lacks the ability to recognize the most challenging regions of the boundaries. In comparison, our predicted point map concentrates on the ambiguous boundaries so that it can boost the challenging lesion's segmentation.

2) Non-Maximum Suppression: In the boundary key-point map generation algorithm, the Non-Maximum Suppression aims to filter the boundary points and find the most valuable points on the most ambiguous regions. To study the effectiveness of filtering points, we remove the NMS operator in our XBoundFormer and show the results in Table IV. As the statistics show, using NMS to filter the most ambiguous boundary points can improve the IoU scores on both sets. The results support the assumption that models should focus their attention on the most ambiguous regions.

3) Statistics of the Efficiency: We set N_{im}, N_{ex} to control the number of im-Bound and ex-Bound learners. Enlarging them leads to more computation, while few learners may not be able to learn the correct boundary knowledge. Fig. 8 shows the evaluated IoU scores and inference time of the models with different N_{im}, N_{ex}. For the validation set, the evaluated IoU score increase obviously with more boundary learners, and the score changes a few when $N_{im} = 2, N_{ex} = 2$. The IoU score also increases with increasing $N_{im} = 1, N_{ex} = 1$ to $N_{im} = 2, N_{ex} = 2$ but it also drops with $N_{im} = 3$. The underlying reason may be that more learners bring larger hardness to model optimization. Considering both the efficiency, accuracy and generalization ability, we take $N_{im} = 2, N_{ex} = 2$ as our final setting.

V. DISCUSSION

Skin lesion segmentation plays a vital role in the quantitative analysis of skin cancers, i.e., lesion size and shape analysis. Existing studies adopt attention-based networks to catch global context, and boundary-aware supervision is proved to be effective for object segmentation in other fields. In this work, we exploit the complementary advantage of global context and boundary knowledge at multi-scale, proposing a cross-scale boundary-aware transformer, XBound-Former, for precise segmentation of skin lesions with ambiguous boundaries. The main contribution is our three boundary learners to explore in-scale and cross-scale boundary knowledge. The experiment is conducted on two skin lesion datasets and an external polyp lesion dataset. The results have shown that our model has the best segmentation performance, especially in the determination of challenging boundaries. The generalization ability on unseen images and different tasks has also been verified.

In the medical field, targets usually have ambiguous boundaries that are hard to determine, even for human beings. The challenges majorly come from the limitation of imaging techniques and would be solved in the future by the new

evolution of advanced imaging techniques. However, in the current community, how to segment these challenging objects has huge significance for the diagnosis, quality control, and treatment planning of patients. Therefore, we thoroughly investigate and aim to solve the challenges in the skin lesion segmentation and preliminarily discuss the potential users on the other targets with similar characteristics.

How to fuse boundary information into the segmentation tasks is one of the most well-known topics in object segmentation. It can be achieved through designing boundary-aware loss objectives like HD loss. Recent studies show that it is more effective to transfer the boundary loss as boundary key-point map loss. In addition to the supervision, the predicted boundary key-point map can also be used as the spatial attention map. Following this direction, we propose XBound-Former, which takes the complementary usage of the attention-based network and boundary supervision. Based on this theory, we further explore the potential help in exploring cross-scale boundary knowledge. All our proposals are proved to be effective in our ablation experiment and the detailed discussion.

Our model still has some limitations that will further improve the segmentation if broken. First, in some extremely challenging images, the boundary key points are still unable to detect clearly. The false point detection may bring harmful guidance to the branch of lesion segmentation. Although they have the complementary advantage in most cases, we should consider the potential harm in some noisy cases. Second, boundary key-point detection is a different task that requires unique representations compared to lesion segmentation. In future work, utilizing different models for the two branches instead of sharing the same architecture may be helpful to guarantee the accuracy of the two branches.

VI. Conclusion

In this paper, we present a novel cross-scale boundary-aware transformer (XBound-Former) to handle the large lesion variance and ambiguous boundaries in skin lesion segmentation, by holistically perceiving the advantage of boundary-wise prior knowledge and long-range dependency modeling. Based on the pyramid features extracted by transformers, we propose three boundary learners (im-Bound, ex-Bound, X-Bound) to explore the in-scale and cross-scale boundary knowledge to enhance the boundary segmentation accuracy. We perform comparison experiments on two skin lesion datasets where the results clearly verify the advantage of our method. The detailed ablation study proves that each boundary learner contributes to the performance boost and the learners can be fused to further improve the accuracy. The extensive experiments conducted on the polyp segmentation also indicate our potentials in the similar targets with ambiguous boundaries. However, it is also found that our model still fails on some extremely low-contrast lesions, which may be solved by fusing a deep learning-based model and low-level feature extractor in future work.

VII. Acknowledgement

This work is supported by the Ministry of Science and Technology of the People's Republic of China under grant No. 2021ZD0201900 and 2021ZD0201904.

References

[1] P. Mathur, K. Sathishkumar, M. Chaturvedi, P. Das, K. L. Sudarshan, S. Santhappan, V. Nallasamy, A. John, S. Narasimhan, F. S. Roselind *et al.*, "Cancer statistics, 2020: report from national cancer registry programme, india," *JCO Global Oncology*, vol. 6, pp. 1063–1075, 2020.

[2] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021," *CA: a cancer journal for clinicians*, vol. 71, no. 1, pp. 7–33, 2021.

[3] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3d brain image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1744–1757, 2010.

[4] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE transactions on medical imaging*, vol. 36, no. 9, pp. 1876–1886, 2017.

[5] L. Yu, H. Chen, Q. Dou, J. Qin, and P. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Transactions on Medical Imaging*, vol. 36, no. 4, pp. 994–1004, 2017.

[6] H. Li, X. He, F. Zhou, Z. Yu, D. Ni, S. Chen, T. Wang, and B. Lei, "Dense deconvolutional network for skin lesion segmentation," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 527–537, 2018.

[7] M. Attia, M. Hossny, S. Nahavandi, and A. Yazdabadi, "Skin melanoma segmentation using recurrent and convolutional neural networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 292–296.

[8] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472–480.

[9] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2016.

[10] H. J. Lee, J. U. Kim, S. Lee, H. G. Kim, and Y. M. Ro, "Structure boundary preserving segmentation for medical image with ambiguous boundary," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4816–4825.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[12] T. Prangemeier, C. Reich, and H. Koeppl, "Attention-based transformers for instance segmentation of cells in microstructures," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 700–707.

[13] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvtv2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 1–10, 2022.

[14] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," 2021.

[15] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *Medical Image Computing and Computer Assisted*, ser. Lecture Notes in Computer Science, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds., vol. 12901. Springer, 2021, pp. 14–24.

[16] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, and J. Qin, "Boundary-aware transformers for skin lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 206–216.

[17] W. Cao, G. Yuan, Q. Liu, C. Peng, J. Xie, X. Yang, X. Ni, and J. Zheng, "Icl-net: Global and local inter-pixel correlations learning network for skin lesion segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[19] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 251–266.

[20] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 2, pp. 699–711, 2020.

[21] H. Wu, J. Pan, Z. Li, Z. Wen, and J. Qin, "Automated skin lesion segmentation via an adaptive dual attention module," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 357–370, 2020.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[26] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *arXiv preprint arXiv:2101.11986*, 2021.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[28] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[29] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on medical imaging*, vol. 39, no. 2, pp. 499–513, 2019.

[30] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International conference on medical imaging with deep learning*. PMLR, 2019, pp. 285–296.

[31] S. Wang, K. He, D. Nie, S. Zhou, Y. Gao, and D. Shen, "Ct male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation," *Medical image analysis*, vol. 54, pp. 168–178, 2019.

[32] Y. Meng, H. Zhang, Y. Zhao, X. Yang, Y. Qiao, I. J. MacCormick, X. Huang, and Y. Zheng, "Graph-based region and boundary aggregation for biomedical image segmentation," *IEEE transactions on medical imaging*, 2021.

[33] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1243–1252.

[34] T. Prangemeier, C. Reich, and H. Koeppl, "Attention-based transformers for instance segmentation of cells in microstructures," 2020.

[35] S. Suzuki *et al.*, "Topological structural analysis of digitized binary images by border following," *Computer vision, graphics, and image processing*, vol. 30, no. 1, pp. 32–46, 1985.

[36] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," in *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2013, pp. 5437–5440.

[37] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and 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.

[38] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.

[39] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[40] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 630–644, 2015.

[41] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of healthcare engineering*, vol. 2017, 2017.

[42] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International journal of computer assisted radiology and surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[43] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1605.01397*, 2016.

[44] O. Ronneberger, P. Fischer, and 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.

[45] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3–11.

[46] M. Benčević, I. Galić, M. Habijan, and D. Babin, "Training on polar image transformations improves biomedical image segmentation," *IEEE Access*, vol. 9, pp. 133 365–133 375, 2021.

[47] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[48] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[49] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[50] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu, "Adaptive context selection for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 253–262.

[51] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "Tganet: Text-guided attention for improved polyp segmentation," in *MICCAI*, 2022.

[52] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, "Hardnet-mseg: a simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps," *arXiv preprint arXiv:2101.07172*, 2021.

[53] Z. Yin, K. Liang, Z. Ma, and J. Guo, "Duplex contextual relation network for polyp segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[54] K. Patel, A. M. Bur, and G. Wang, "Enhanced u-net: A feature enhancement network for polyp segmentation," in *2021 18th Conference on Robots and Vision (CRV)*. IEEE, 2021, pp. 181–188.

[55] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 699–708.

[56] D. Bo, W. Wenhai, L. Jinpeng, and F. Deng-Ping, "Polyp-pvt: Polyp segmentation with pyramidvision transformers," *arXiv preprint arXiv:2108.06932v3*, 2021.

[57] J. Qi, M. Le, C. Li, and P. Zhou, "Global and local information based deep network for skin lesion segmentation," *arXiv preprint arXiv:1703.05467*, 2017.

[58] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," 2018.