Multimedia Systems (2023) 29:3847–3861
https://doi.org/10.1007/s00530-023-01165-z

SPECIAL ISSUE PAPER

LET-Net: locally enhanced transformer network for medical image segmentation

Na Ta[1,2,3] · Haipeng Chen[1,3] · Xianzhu Liu[4] · Nuo Jin[5]

Received: 13 April 2023 / Accepted: 12 August 2023 / Published online: 5 September 2023

Abstract
Medical image segmentation has attracted increasing attention due to its practical clinical requirements. However, the prevalence of small targets still poses great challenges for accurate segmentation. In this paper, we propose a novel locally enhanced transformer network (LET-Net) that combines the strengths of transformer and convolution to address this issue. LET-Net utilizes a pyramid vision transformer as its encoder and is further equipped with two novel modules to learn more powerful feature representation. Specifically, we design a feature-aligned local enhancement module, which encourages discriminative local feature learning on the condition of adjacent-level feature alignment. Moreover, to effectively recover high-resolution spatial information, we apply a newly designed progressive local-induced decoder. This decoder contains three cascaded local reconstruction and refinement modules that dynamically guide the upsampling of high-level features by their adaptive reconstruction kernels and further enhance feature representation through a split-attention mechanism. Additionally, to address the severe pixel imbalance for small targets, we design a mutual information loss that maximizes task-relevant information while eliminating task-irrelevant noises. Experimental results demonstrate that our LET-Net provides more effective support for small target segmentation and achieves state-of-the-art performance in polyp and breast lesion segmentation tasks.

Keywords Medical image segmentation · Feature alignment · Local-induced decoder · Mutual information · Transformer

1 Introduction

Multimodal medical image segmentation aims to accurately identify and annotate regions of interest from images produced by various medical devices, such as segmenting polyps from colonoscopy images [1], breast lesions from ultrasound images [2], and focal cortical dysplasia lesions from magnetic resonance images [3]. It has been an essential procedure for computer-aided diagnosis [4], which assists clinicians in making accurate diagnoses, planning surgical procedures, and proposing treatment strategies. Hence, the development of automatic, accurate, and robust medical image segmentation methods is of great value to clinical practice.

However, medical image segmentation still encounters some challenges, one of which is the prevalence of small lesions. Figure 1 illustrates small lesion samples and size distribution histograms for several different benchmarks, where the ratio of lesion area to whole image is significantly concentrated in a smaller range, with proportions in descending order: 0−0.1 first, 0.1−0.2 s. Specifically, a vast majority of polyps and breast lesions occupy only a small proportion of the entire medical image. Meanwhile, some small lesions, e.g., early stage polyps, exhibit an inconspicuous appearance. These small targets inevitably pose great difficulties for accurate segmentation for several reasons. First, small targets are prone to being lost during repeated

✉ Nuo Jin
jinnuo0412@163.com

1 College of Computer Science and Technology, Jilin University, Changchun 130012, China

2 College of Computer, Hulunbuir University, Hulunbuir 021008, China

3 Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

4 National and Local Joint Engineering Research Center of Space Optoelectronics Technology, Changchun University of Science and Technology, Changchun 130022, China

5 Southampton Business School, University of Southampton, Southampton SO17 1BJ, United Kingdom

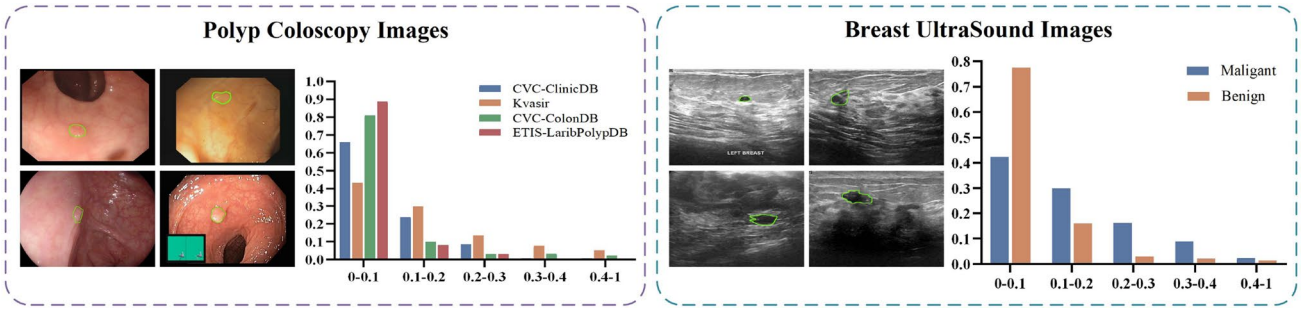

Fig. 1 An illustration of small lesion samples and size distributions for different medical image datasets, including polyp coloscopy images and breast ultrasound images. Ground truth for each image is represented by a green line. In a histogram, the horizontal axis represents the proportion of the entire image occupied by the lesion area, while the vertical axis indicates the proportion of samples with a particular lesion size relative to the total sample

downsampling operations and are hard to recover. Second, there is a significant class imbalance problem in the number of pixels between the foreground and background, leading to a biased network and suboptimal performance. Whereas, the ability of computer-aided diagnosis to identify small objects is highly desired, as early detection and diagnosis of small lesions are crucial for successful cancer prevention and treatment.

Nowadays, the development of medical image segmentation has greatly advanced due to the efficient feature extraction ability of convolutional neural networks (CNNs) [5–7]. Modern CNN-based methods typically utilize a U-shaped encoder–decoder structure, where the encoder extracts semantic information and the decoder restores resolution to facilitate segmentation. Additionally, skip connections are employed to compensate for detailed information. Some advanced U-shaped works focus on the following studies, which include designing novel encoding blocks [8–10] to enhance feature representation ability, adopting attention mechanisms to further recalibrate features [11, 12], extracting and fusing multi-scale reasonable context information to improve accuracy [13–15], and so on. Despite their promising performance, these methods share a common flaw, i.e., lacking global contexts essential for better recognition of target objects.

Due to their superior ability to model global contexts, Transformer-based architectures have become popular in segmentation tasks while achieving promising performance. Recent works [16–18] utilize vision transformers (ViT) as a backbone to incorporate global information. Despite their good performance, ViT produces single-scale low-resolution features and has a very high computational cost, which hampers their performance in dense prediction. In contrast to ViT, pyramid vision transformer (PVT) [19] inherits the advantages of both CNN and Transformer and produces hierarchical multi-scale features that are more favorable for segmentation. Unfortunately, Transformer-based methods destroy part of local features when modeling global contexts, which may result in imprecise predictions for small objects.

In the field of small target segmentation, a couple of approaches have been devised to improve the sensitivity of small objects. They overcome the segmentation difficulties brought by small objects from multiple aspects, such as exploiting the complementarity between low-level spatial details and high-level semantics [20], multi-scale feature learning [21, 22], and augmenting spatial dimension strategies [23–25]. Although their skip connections can compensate for detail loss to some extent and even eliminate somewhat irrelevant noises by extra equipping with attention mechanisms, these methods are still insufficient, as some local contexts may be overwhelmed by dominant semantics due to feature misalignment issues. In addition, another important factor that has been overlooked is how to effectively restore spatial information of downsampled features. Most methods adopt common upsampling operations, such as nearest-neighbor interpolation and bilinear interpolation, which may still lack local spatial awareness to handle small object positions. As a result, they are not compatible with the recovery of target objects and produce suboptimal segmentation performance.

In this paper, we propose a novel locally enhanced transformer network (LET-Net) for medical image segmentation. By leveraging the merits of both Transformer and CNN, our LET-Net can accurately segment small objects and precisely sharpen local details. First, the PVT-based encoder produces hierarchical multi-scale features where low-level features tend to retain local details, while high-level features provide strong global representations. Second, to further emphasize detailed local contexts, we propose a feature-aligned local enhancement (FLE) module, which can learn discriminative local cues from adjacent-level features on the condition of feature alignment and then utilize the local enhancement block equipped with local receptive fields to further recalibrate features. Third, we design a progressive

local-induced decoder that contains cascaded local reconstruction and refinement (LRR) modules to achieve effective spatial recovery of high-level features under the adaptive guidance of reconstruction kernels and optimization of a split-attention mechanism. Moreover, to alleviate the class imbalance between foreground and background, we design a mutual loss based on an information-theoretic objective, which can impose task-relevant restrictions while reducing task-irrelevant noises.

The contributions of this paper mainly include:

(1) We put forward a novel LET-Net, which combines the strengths of Transformer and CNN for accurate medical image segmentation.
(2) We propose two novel modules, FLE and LRR, to enhance the sensitivity of small objects. FLE can extract discriminative local cues under the alignment of adjacent-level features, while LRR enables effective spatial recovery by guiding upsampling of high-level features via its adaptive reconstruction kernels and recalibrating features through a split-attention mechanism.
(3) To mitigate the class imbalance caused by small targets, we design a mutual information loss, which enables our model to extract task-relevant information while reducing task-irrelevant noises.
(4) By evaluating our LET-Net in challenging colorectal polyp segmentation and ultrasound breast segmentation, we demonstrate its state-of-the-art segmentation ability and strong generalization capability.

2 Related work

2.1 Medical image segmentation

With the great development of deep learning, especially convolutional neural networks (CNNs), various CNN-based methods, such as U-Net [7], have significantly improved the performance of medical image segmentation. These approaches possess the popular U-shaped encoder–decoder structure. To further assist precise segmentation, a battery of innovative improvements based on encoder–decoder architecture has emerged [26–30]. One direction is to design a new module for enhancing the encoder or decoder ability. For instance, Dai et al. [26] designed Ms RED network, which, respectively, employs a multi-scale residual encoding fusion module (MsR-EFM) and a multi-scale residual decoding module (MsR-DFM) in the encoder and decoder stages to improve skin lesion segmentation. In the work [27], a selective receptive filed module (SRFM) was designed to obtain suitable sizes of receptive fields, thereby boosting breast mass segmentation. Another direction is optimizing skip connection to facilitate the recovery of spatial information. UNeXt [28] proposed an encoder–decoder structure involving convolutional stages and tokenized MLP stages, achieving better segmentation performance while also improving the inference speed. However, these methods directly fuse unaligned features from different levels, which may hamper accuracy, especially for small objects. In this paper, we propose a powerful feature-aligned local enhancement module, which ensures that feature maps at adjacent levels can be well aligned and then explore substantial local cues to optimally enhance the discriminative details.

2.2 Feature alignment

Feature alignment has drawn much attention and is now an active research topic in computer vision. Numerous researchers have devoted considerable effort to addressing this challenge [6, 31–37]. For instance, SegNet [6] utilized max-pooling indices computed in the encoder to perform an upsampling operation in the corresponding decoder stage. Mazzini et al. [32] proposed a guided upsampling module (GUM) that generates learnable guided offsets to enhance the upsampling operation. IndexNet [33] built a novel index-guided encoder–decoder structure in which pooling and upsampling operators are guided by self-learned indices. AlignSeg [34] learned 2D transformation offsets by a simple learnable interpolation strategy to alleviate feature misalignment. Huang et al. [35] designed an FaPN framework consisting of feature alignment and feature selection modules, achieving substantial and consistent performance improvements on dense prediction tasks. SFNet [31] presented a flow alignment module that effectively broadcasts high-level semantic features to high-resolution detail features by its semantic flow. Our method shares a similar aspect with the work [31], in which efficient spatial alignment is achieved by learning offsets. However, unlike these methods, we further enhance discriminative representations by subtraction under the premise of aligning low-resolution and high-resolution features, which facilitates excavating imperceptible local cues related to small objects.

2.3 Attention mechanism

Attention-based algorithms have been developed to assist in segmentation. In general, attention mechanisms can be categorized into channel attention, spatial attention, and self-attention according to different focus perspectives. Inspired by the success of SENet [38], various networks [39–41] have incorporated the squeeze-and-excitation (SE) module to recalibrate features by modeling channel relationships, thereby improving segmentation performance. K. Wang et al. [42] proposed a dual attention network (DANet), which

combines position spatial attention and channel attention modules to capture rich contexts.

Additionally, Transformer networks based on self-attention have been popular in medical image segmentation [43–48]. For instance, TransUnet [43] inserted Transformer layers between CNN-based encoder and decoder stages to model global contexts, achieving excellent performances in multi-organ and cardiac segmentation. Wu et al. [48] proposed FAT-Net with a dual encoder that is, respectively, based on CNNs and Transformers for skin lesion segmentation. However, the loss of local contexts may still hinder the prediction accuracy of Transformer-based methods. In this paper, we propose a feature-aligned local enhancement module and progressive local-induced decoder, which, respectively, emphasize local information and adaptively recover spatial information to improve predictions.

3 Method

Figure 2 illustrates our proposed LET-Net, which combines Transformer and CNN architectures to achieve accurate segmentation. In the encoder stage, we utilize a pre-trained pyramid vision transformer (PVT) [19] as the backbone to extract hierarchical multi-scale features. Then, three feature-aligned local enhancement (FLE) modules are inserted in the skip connections to enhance discriminative local features. Afterward, we employ a novel progressive local-induced decoder composed of cascaded local reconstruction and refinement (LRR) modules to effectively recover spatial resolution and produce the final segmentation maps. In what follows, we elaborate on the key components of our model.

3.1 PVT-based encoder

Although CNN-based methods have achieved great success in medical image segmentation, they have general limitations in modeling global contexts. In contrast, pyramid vision transformer (PVT) [19] inherits the advantages of both Transformer and CNN while proving to be more effective for segmentation. Thus, we choose PVT as the backbone to obtain global receptive fields and learn effective multi-scale features.

As shown in Fig. 2, the PVT-based encoder has four stages with a similar architecture. Each stage contains a patch embedding layer and multiple Transformer layers. Benefiting from its progressive shrinking pyramid and spatial-reduction attention strategy, the PVT-based encoder can produce multi-scale feature maps with fewer memory costs. Specifically, given an input image $X \in \mathbb{R}^{H\times W\times 3}$, it produces features $\{\mathbf{E}_i | 1 \le i \le 4\}$, in which $\mathbf{E}_i \in \mathbb{R}^{H/2^{i+1}\times W/2^{i+1}\times C_i}$. Therefore, we obtain high-resolution detail features and low-resolution semantic features, which are beneficial for segmentation.

3.2 Feature-aligned local enhancement module

The powerful global receptive field of PVT-based encoder makes it challenging for our model to adequately capture critical local details. Although low-level features can provide some local context, directly transmitting them to the decoder via a simple skip connection is problematic, as this may introduce a large amount of irrelevant background information. As a solution, leveraging high-level features is an effective way, but one significant issue, i.e., feature

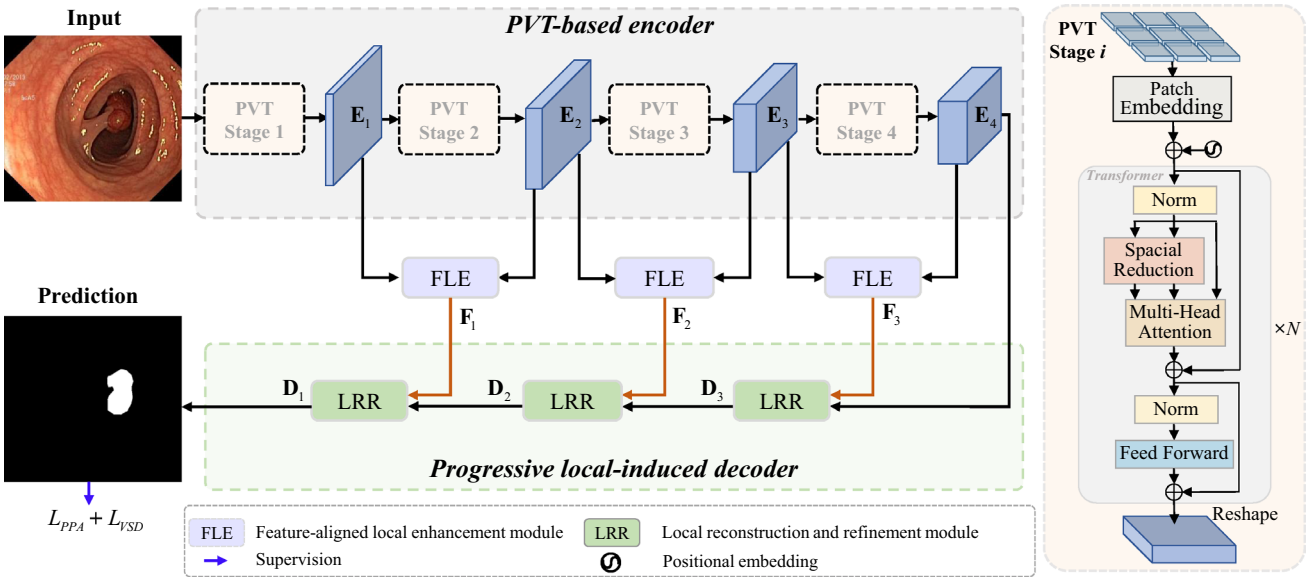

Fig. 2 The pipeline of our proposed LET-Net

alignment, should be fully considered in this procedure to prevent local contexts from being overshadowed by global contexts. To this end, we propose a feature-aligned local enhancement (FLE) module, in which informative detailed features are effectively captured under the premise of feature alignment, producing discriminative representation. The internal structure of FLE is illustrated in Fig. 3, and it consists of two steps: feature-aligned discriminative learning and local enhancement.

Feature-aligned discriminative learning Due to the information gap between semantics and resolution, feature representation is still suboptimal when directly upsampling high-level feature maps to guide low-level features. To obtain strong feature representations, more attention and effort should be given to position offset between low-level and high-level features. Inspired by previous work [31], we propose a feature-aligned discriminative learning (FDL) block that aligns adjacent-level features and further excavates discriminative features, leading to high sensitivity to small objects. Within FDL, two 1×1 convolution layers are first employed to compress adjacent-level features (i.e., $\mathbf{E}_i$ and $\mathbf{E}_{i-1}$) into the same channel depth. Then, a semantic flow field is calculated by a 3×3 convolution operation, as described in Eq. 1

$$\Delta_{i-1} = f_{3\times3}\left(f_{1\times1}\left(\mathbf{E}_{i-1}\right) \copyright \mathrm{U}\left(f_{1\times1}\left(\mathbf{E}_i\right)\right)\right), \tag{1}$$

where $f_{s\times s}(\cdot)$ indicates $s \times s$ convolution layer followed by batch normalization and a ReLU activation function, while © and U$(\cdot)$, respectively, represent concatenation and upsampling operation. Next, according to learned semantic flow Δ_{i-1}, we obtain a feature-aligned high-resolution feature $\tilde{\mathbf{E}}_i$ with semantic cues, Mathematically

$$\tilde{\mathbf{E}}_i = Warp\left(f_{1\times1}\left(\mathbf{E}_i\right), \Delta_{i-1}\right), \tag{2}$$

where $Warp(\cdot)$ indicates the mapping function, $\mathbf{E}_i$ is a C_i dimensional feature map defined on the spatial grid Ω_i of the specific size $\left(H/2^{i+1}, W/2^{i+1}\right)$. Schematically as shown in Fig. 4, the warp procedure consists of two steps. In the

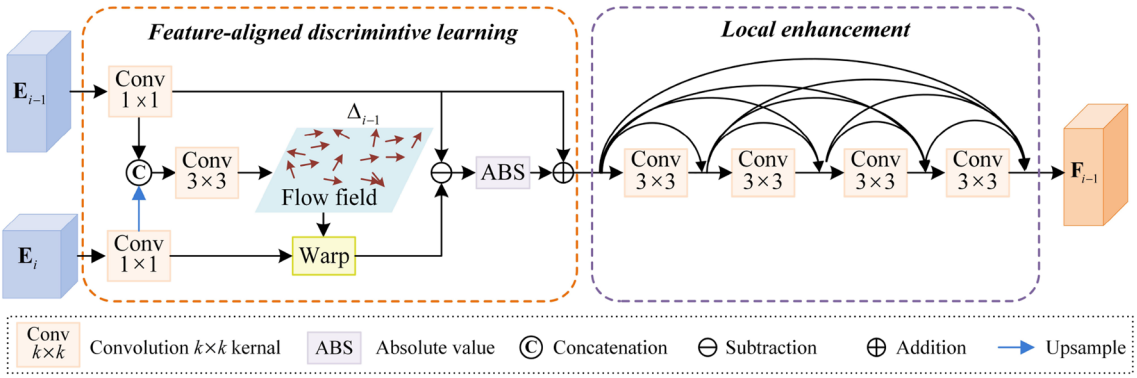

Fig. 3 The architecture of feature-aligned local enhancement module, which performs two steps: First, feature-aligned discriminative learning initially produces a flow field to align adjacent features and then constructs discriminative representation using subtraction and a residual connection. Second, local enhancement with a dense connection structure is adopted to highlight local details

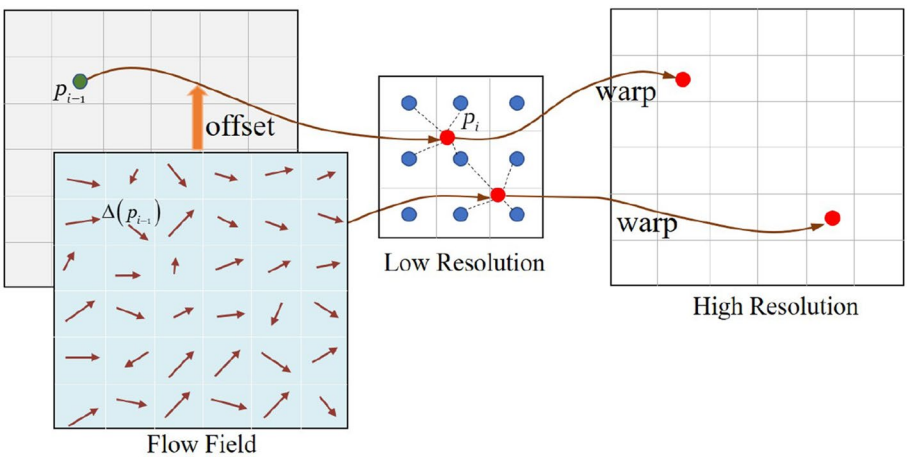

Fig. 4 An illustration of the warp procedure

first step, each point p_{i-1} on the spatial grid Ω_{i-1} is mapped to p_i on low-resolution feature, which is formulated by Eq. 3

$$p_i = \frac{p_{i-1} + \Delta_{i-1}(p_{i-1})}{2}. \tag{3}$$

It is worth mentioning that due to the resolution gap between the flow field and features (see Fig. 4), Eq. 3 contains a halved operation to reduce the resolution. In the second step, we adopt the differentiable bilinear sampling mechanism [49] to approximate the final feature $\tilde{\mathbf{E}}_i$ by linearly interpolating the scores of four neighboring points (top-right, top-left, bottom-right, and bottom-left) of p_i.

After that, to enhance the discriminative local context representation, we further utilize subtraction, absolute value, and residual learning procedures. Conclusively, the final optimized feature $\hat{\mathbf{E}}_{i-1}$ can be expressed as follows:

$$\hat{\mathbf{E}}_{i-1} = \mid \mathbf{E}_{i-1} - \tilde{\mathbf{E}}_i \mid + \mathbf{E}_{i-1}. \tag{4}$$

Local enhancement In the PVT-based encoder, attention is established between each patch, allowing information to be blended from all other patches, even if their correlation is not high. Meanwhile, since small targets only occupy a portion of the entire image, the global interaction in transformer architecture cannot fully meet the requirements of small target segmentation where more detailed local contexts are needed. Considering that the convolution operation with a fixed receptive field can blend the features of each patch's neighboring patches, we construct a local enhancement (LE) block to increase the weights associated with adjacent patches to the center patch using convolution, thereby emphasizing the local features of each patch.

As shown in Fig. 3, LE has a convolution-based structure and consists of four stages. Each stage includes a 3×3 convolutional layer followed by batch normalization and a ReLu activation layer (denoted as $f_{3\times3}(\cdot)$). Additionally, dense connections are added to encourage feature reuse and strengthen local feature propagation. As a result, the feature map obtained by LE contains rich local contexts. Let x_0 denote the initial input, and the outputs of i^{th} stage within LE can be formulated as follows:

$$x_i = \begin{cases} f_{3\times3}(x_0), & i = 1, \\ f_{3\times3}([x_0, \cdots, x_{i-1}]), & 2 \le i \le 4, \end{cases} \tag{5}$$

where [ ] represents the concatenation operation. In summary, LE utilizes the local receptive field of the convolution operation and dense connections to achieve local enhancement.

3.3 Progressive local-induced decoder

Efficient recovery of spatial information is critical in medical image segmentation, especially for small objects. Inspired by previous works [50, 51], we propose a progressive local-induced decoder to adaptively restore feature resolution and detailed information. As shown in Fig. 2, the decoder consists of three cascaded local reconstruction and refinement (LRR) modules. The internal structure of LRR is illustrated in Fig. 5, where two steps are performed: local-induced reconstruction (LR) and split-attention-based refinement (SAR).

Local-induced reconstruction LR aims to transfer the spatial detail information from low-level features into high-level features, thereby facilitating accurate spatial recovery of high-level features. As shown in Fig. 5, LR first produces a reconstruction kernel $\kappa \in \mathbb{R}^{k^2 \times H_{i-1} \times W_{i-1}}$ based on low-level feature $\mathbf{F}_{i-1}$ and high-level feature $\mathbf{D}_i$, in which k indicates the neighborhood size for reconstructing local features. The procedure of generating the reconstruction kernel κ can be expressed as follows:

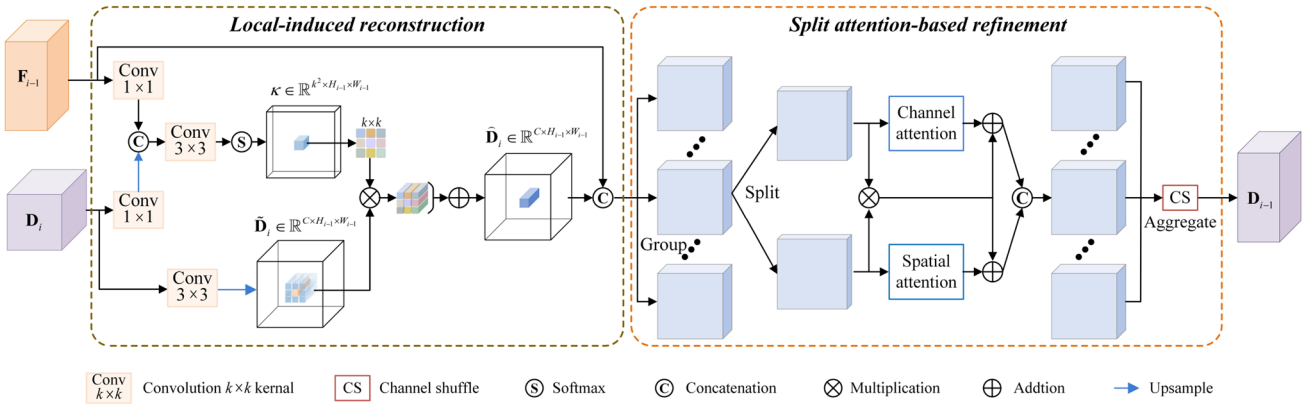

Fig. 5 The structure of local reconstruction and refinement module. It contains two blocks: local-induced reconstruction and split-attention-based refinement

$$\kappa = Soft\left(f_{3\times3}\left(\mathrm{U}(f_{1\times1}\left(\mathbf{D}_i\right))\copyright f_{1\times1}\left(\mathbf{F}_{i-1}\right)\right)\right), \tag{6}$$

where $f_{s\times s}(\cdot)$ represents an $s \times s$ convolution layer followed by batch normalization and a ReLU activation function. $\mathrm{U}(\cdot)$, $\copyright$, and $Soft(\cdot)$, respectively, indicate upsampling, concatenation, and Softmax activation operations. Meanwhile, another 3×3 convolution and upsampling operation are applied on $\mathbf{D}_i$ to obtain $\tilde{\mathbf{D}}_i$ with the same resolution size as $\mathbf{F}_{i-1}$. Mathematically, $\tilde{\mathbf{D}}_i = \mathrm{U}\left(f_{3\times3}\left(\mathbf{D}_i\right)\right)$. Note that, $\mathbf{D}_4 = \mathbf{E}_4$ here. Next, we optimize pixel $\tilde{\mathbf{D}}_i[u, v]$ under the guidance of reconstruction kernel $\kappa_{[u,v]} \in \mathbb{R}^{k\times k}$, producing refined local feature $\hat{\mathbf{D}}_i[u, v]$. This can be written as Eq. 7, where $r = \lfloor k/2 \rfloor$

$$\hat{\mathbf{D}}_i[u, v] = \sum_{m=-r}^{r}\sum_{n=-r}^{r} \kappa_{[u,v]}[m, n] \times \tilde{\mathbf{D}}_i[u + m, v + n]. \tag{7}$$

Subsequently, $\hat{\mathbf{D}}_i$ and $\mathbf{F}_{i-1}$ are concatenated together and then passed through two convolutional layers to produce an optimized feature. Conclusively, LR overcomes the limitations of traditional upsampling operations in precisely recovering pixel-wise prediction, since it takes full advantage of low-level features to adaptively predict reconstruction kernel and then effectively combines semantic contexts with spatial information toward accurate spatial recovery. This can strengthen the recognition of small objects.

Split-attention-based refinement To enhance feature representation, we implement an SAR block in which grouped sub-features are further split and fed into two parallel branches to capture channel dependencies and pixel-level pairwise relationships through two types of attention mechanisms. As shown in Fig. 5, SAR is composed of two basic components: a spatial attention block and a channel attention block. Given an input feature map M, SAR first divides it along the channel dimension to produce $M = \left\{M_1, M_2, \cdots, M_G\right\}$. For each M_i, valuable responses are specified by attention mechanisms. Specifically, M_i is split into two features, denoted as M_i^1 and M_i^2, which are separately fed into the channel attention block and spatial attention block to reconstruct features. This allows our model to focus on "what" and "where" are valuable through these two blocks.

In channel attention block, global average pooling (denoted as $GAP(\cdot)$) is performed to produce channel-wise statistics, which can be formulated as

$$S = GAP(M_i^1) = \frac{1}{H \times W}\sum_{m=1}^{H}\sum_{n=1}^{W} M_i^1(m, n). \tag{8}$$

Then, channel-wise dependencies are captured according to the guidance of a compact feature, which is generated by a Sigmoid function (i.e., $Sig(\cdot)$). Mathematically

$$\tilde{M}_i^1 = \mathrm{Sig}\left(W_1 \times S + b_1\right) \times M_i^1, \tag{9}$$

in which parameters W_1 and b_1 are used for scaling and shifting S.

In spatial attention block, spatial-wise statistics are calculated using Group Norm (GN) [52] on M_i^2. The pixel-wise representation is then strengthened by another compact feature calculated by two parameters W_2 and b_2 and a Sigmoid function. This process can be expressed as

$$\tilde{M}_i^2 = \mathrm{Sig}\left(W_2 \times GN\left(M_i^2\right) + b_2\right) \times M_i^2. \tag{10}$$

Next, $\tilde{M}_i^1$ and $\tilde{M}_i^2$ are optimized by an additional consistency embedding path and then concatenated. This procedure is represented as

$$\tilde{M}_i = \left(\tilde{M}_i^1 + M_i^1 \times M_i^2\right)\copyright\left(\tilde{M}_i^2 + M_i^1 \times M_i^2\right). \tag{11}$$

After aggregating all sub-features, a channel shuffle [53] is performed to facilitate cross-group information exchange along the channel dimension.

3.4 Mutual information loss

As stated in the previous study [54], training models with only pixel-wise loss may limit segmentation performance, especially resulting in prediction errors for small objects. This is due to class imbalance between foreground and background, such that task-relevant information is overwhelmed by irrelevant noise. Therefore, to facilitate preserving task-relevant information, we explore novel supervision at the feature level to further assist accurate segmentation. Let X and Y denote the input medical image and its corresponding ground truth, respectively. Z represents the deep feature extracted from input X.

Mutual information (MI) Mutual information is a fundamental quantity that measures the amount of information shared between two random variables. Mathematically, the statistical dependency of Y and Z can be quantified by MI, which is expressed as

$$I(Y;Z) = \mathbb{E}_{p(Y,Z)}\left[\log\frac{p(Y, Z)}{p(Y)p(Z)}\right], \tag{12}$$

where $p(Y, Z)$ is the probability distribution between Z and Y, while $p(Z)$ and $p(Y)$ are their marginals.

Mutual Information Loss Our primary objective is to maximize the amount of task-relevant information about Y in the latent feature Z while reducing irrelevant information. This is achieved by two mutual information terms [55, 56]. Formally

$$\mathrm{IB}(Y, Z) = \mathrm{Max}\ I(Z;Y) - I(Z;X). \tag{13}$$

Owing to the notorious difficulty of the conditional MI computations, these terms are estimated by existing MI estimators [56, 57]. In detail, the first term is accomplished through

the use of Pixel Position Aware (PPA) loss [57] (L_{PPA}). Since PPA loss assigns different weights to different positions, it can better explore task-relevant structure information and give more attention to important details. The second term is estimated by Variational Self-Distillation (VSD) [56] (L_{VSD}) that uses KL-divergence to compress Z and remove irrelevant noises, thereby addressing the effect of imbalances in the number of foreground and background pixels caused by small targets. Thus, our total loss can be expressed as

$$L_{total} = L_{\text{PPA}} + L_{\text{VSD}}. \tag{14}$$

4 Experiments

4.1 Experimental setup

4.1.1 Implementation details

We implement our experiments based on the hardware environment with NVIDIA GeForce RTX 3090. The AdamW algorithm is chosen to optimize our model's parameters, and the initial learning rate is set to 1e-4. During training, a multi-scale training strategy is employed, in which input images are reshaped according to a ratio of [0.75, 1, 1.25]. The total number of epochs and batch size are set to 200 and 16, respectively. In the pre-processing step, all images and corresponding ground truths are resized to 352×352 in our experiments.

4.1.2 Datasets

To verify the capability of our proposed model, we evaluate LET-Net in two medical image segmentation tasks. For polyp segmentation, we utilize five public benchmarks: CVC-ClinicDB [62], Kvasir [63], CVC-ColonDB [64], ETIS-LaribPolypDB [65], and CVC-300 [66]. To ensure a fair comparison, we follow the work [59] and divide large-scale CVC-ClinicDB and Kvasir datasets into training, validation, and testing datasets in a ratio of [8:1:1], while the remaining three datasets are used only for testing to evaluate the model's generalization abilities. For breast lesion segmentation task, we choose the public breast ultrasound dataset (BUSIS) [67] to assess the effectiveness of our LET-Net. This dataset includes 133 normal cases, 437 benign cases, and 210 malignant cases. We follow the same settings as work [2] to separately conduct experiments on benign and malignant samples.

4.1.3 Evaluation metrics

As done in recent related work of polyp segmentation [20], we employ both mean Dice (mDice) and mean IoU (mIoU) to quantitatively evaluate the performance of our model and other state-of-the-art methods on polyp benchmarks. For breast lesion segmentation, we adopt four widely used metrics, including Accuracy, Jaccard index, Precision, and Dice to validate the segmentation performance in our study. Theoretically, high scores for all metrics indicate better results.

4.2 Experimental results

To investigate the effectiveness of our proposed method, we validate LET-Net in two applications: polyp segmentation from coloscopy images and breast lesion segmentation from ultrasound images.

4.2.1 Polyp segmentation

Quantitative comparison To demonstrate the effectiveness of our LET-Net, we compare it to several state-of-the-art methods on five polyp benchmarks. Table 1 summarizes the quantitative experimental results in detail. From it, we can see that our LET-Net outperforms the other methods on all datasets. Concretely, on the seen CVC-ClinicDB dataset, it achieves significantly higher mDice and mIoU scores (94.5% and 89.9%, respectively). On Kvasir dataset, our method exceeds SANet [20] and BLE-Net [61] by 2.2% and 2.1% mDice improvements, respectively. The underlying reason for their limited performance is that these two methods follow a pure CNN architecture, which lacks global long-range dependencies. By contrast, our method captures global contexts by its PVT-based encoder, and further excavates valuable local information using FLE module, demonstrating superior segmentation ability. Most importantly, our LET-Net still exhibits excellent generalization capabilities when applied to unseen datasets (i.e., CVC-ColonDB, ETIS-LaribPolypDB, and CVC-300). Specifically, LET-Net gets ahead of the CNN-based SOTA CaraNet [22] by 2.2% and 2.8% in terms of mDice and mIoU on CVC-ColonDB. Compared with other Transformer-based approaches, our LET-Net also presents excellent segmentation and generalization abilities. Concretely, on ETIS-LaribPolypDB dataset, we can observe that LET-Net achieves 4.7% and 4.2% higher mDice than SETR-PUP [18] and TransUnet [43], respectively. This performance improvement can be attributed to two factors. One is that our proposed FLE module compensates for the loss of local details in the Transformer architecture. The other is that the LRR module effectively recovers spatial information.

Visual Comparison To further evaluate the proposed LET-Net intuitively, we visualize some segmentation maps produced by our model and other methods in Fig. 6. It is apparent that our LET-Net can not only clearly highlight polyp regions but also identify small polyps more accurately than other counterparts. This is mainly because our method effectively leverages and combines global and local

Table 1 Comparisons between different method in polyp segmentation task. The best results are highlighted in bold

Method	Seen dataset				Unseen dataset					
	CVC-ClinicDB		Kvasir		CVC-ColonDB		ETIS- Larib-PolypDB		CVC-300	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
FCN [5]	0.825	0.747	0.775	0.686	0.578	0.481	0.379	0.313	0.660	0.558
U-Net [7]	0.842	0.775	0.818	0.746	0.512	0.444	0.398	0.335	0.710	0.627
UNet++ [30]	0.846	0.774	0.821	0.743	0.599	0.499	0.456	0.375	0.707	0.624
AttentionU-Net [11]	0.809	0.744	0.782	0.694	0.614	0.524	0.440	0.360	0.686	0.580
DCRNet [58]	0.896	0.844	0.886	0.825	0.704	0.631	0.556	0.496	0.856	0.788
SegNet [8]	0.915	0.857	0.878	0.814	0.647	0.570	0.612	0.529	0.841	0.773
SFA [1]	0.700	0.607	0.723	0.611	0.469	0.347	0.297	0.217	0.467	0.329
PraNet [59]	0.899	0.849	0.898	0.840	0.709	0.640	0.628	0.567	0.871	0.797
ACSNet [39]	0.912	0.858	0.907	0.850	0.709	0.643	0.609	0.537	0.862	0.784
EU-Net [60]	0.902	0.846	0.908	0.854	0.756	0.681	0.687	0.609	0.837	0.765
SANet [20]	0.916	0.859	0.904	0.847	0.753	0.670	0.750	0.654	0.888	0.815
BLE-Net [61]	0.926	0.878	0.905	0.854	0.731	0.658	0.673	0.594	0.879	0.805
CaraNet [22]	0.936	0.887	0.918	0.865	0.773	0.689	0.747	0.672	0.903	0.838
SETR-PUP [18]	0.934	0.885	0.911	0.854	0.773	0.690	0.726	0.646	0.889	0.814
TransUnet [43]	0.935	0.887	0.913	0.857	0.781	0.699	0.731	0.660	0.893	0.824
LET-Net(Ours)	**0.945**	**0.899**	**0.926**	**0.876**	**0.795**	**0.717**	**0.773**	**0.698**	**0.907**	**0.839**

contexts. In addition, we introduce mutual information loss as an assistant to learning task-relevant representation. Furthermore, we find that our LET-Net successfully deals with other challenging cases, including cluttered backgrounds (Fig. 6 (b),(c), (g), (i)) and low contrast (Fig. 6 (a),(h)). For example, as illustrated in Fig. 6 (b),(i), ACSNet [39] and PraNet [59] misidentify background tissues as polyps, but our LET-Net overcomes this drawback. Due to combining the strengths of CNN and Transformer, our LET-Net produces good segmentation performance in these scenarios. Overall, our model achieves leading performance.

4.2.2 Breast lesion segmentation

Quantitative comparison To further evaluate the effectiveness of our method, we conduct extensive experiments in breast lesion segmentation and perform a comparative analysis with ten segmentation approaches. Table 2 presents the detailed quantitative comparison among different methods on BUSIS dataset. Obviously, our LET-Net exhibits excellent performance in both benign and malignant lesion segmentation. In benign lesion segmentation, LET-Net achieves 97.7% Accuracy, 74% Jaccard, 83.5% Precision, and 81.5% Dice. Compared with other competitors, LET-Net significantly outperforms them by a large margin. In detail, it, respectively, excels C-Net [2], CPF-Net [29], and PraNet [59] by 1.6%, 4.1%, and 4.9% in terms of Jaccard. Meanwhile, in malignant lesion segmentation, we obtain an Accuracy score of 93% and a Dice score of 72.7%, respectively, demonstrating the superiority of our LET-Net over other methods. In particular, LET-Net presents a significant improvement of 1.8% in Jaccard and 2.8% in Dice compared with C-Net [2]. The reason behind this is that although C-Net constructs a bidirectional attention guidance network to capture both global and local features, long-range dependencies are not fully modeled due to the limitations of convolution.

Visual comparison: To intuitively demonstrate the performance of our model, we present segmentation results of different methods in Fig. 7. We observe that other methods often produce segmentation maps with incomplete lesion structures or false positives, while our prediction maps are superior to others. This is mainly due to our FLE's ability to facilitate discriminative local feature learning and the effectiveness of our proposed LRR module for spatial reconstruction. In addition, it is worth noting that our LET-Net performs well in handling various shapes [Fig. 7(a)–(h)] and low-contrast images [Fig. 7(d)(h)], which can be attributed to the powerful and robust feature learning ability of LET-Net.

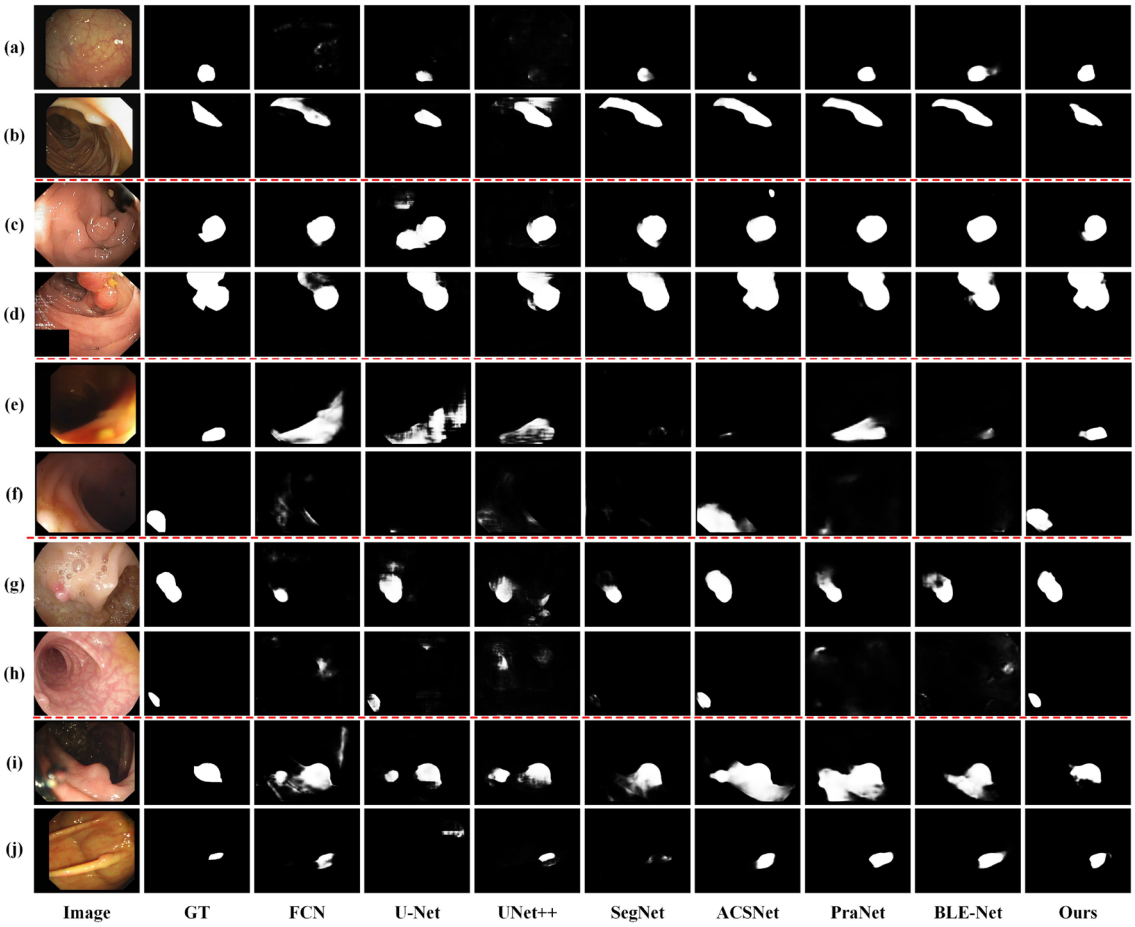

Fig. 6 Visualization results of our LET-Net and several other methods on five polyp datasets. From top to down, the images are from CVC-ClinicDB, Kvasir, CVC-ColonDB, ETIS-LaribPolypDB, and CVC-300, which are separated by red dashed lines

Table 2 Comparison with different state-of-the-art methods on BUSIS dataset

Method	Benign lesion				Malignant lesion			
	Accuracy	Jaccard	Precision	Dice	Accuracy	Jaccard	Precision	Dice
U-Net [7]	0.966	0.615	0.750	0.705	0.901	0.511	0.650	0.635
STAN [21]	0.969	0.643	0.744	0.723	0.910	0.511	0.647	0.626
AttentionU-Net [11]	0.969	0.650	0.752	0.733	0.912	0.511	0.616	0.630
Abraham et al. [68]	0.969	0.667	0.767	0.748	0.915	0.541	0.675	0.658
UNet++ [30]	0.971	0.683	0.759	0.756	0.915	0.540	0.655	0.655
UNet3+ [69]	0.971	0.676	0.756	0.751	0.916	0.548	0.658	0.662
SegNet [8]	0.972	0.679	0.770	0.755	0.922	0.549	0.638	0.659
PraNet [59]	0.972	0.691	0.799	0.763	0.925	0.582	0.763	0.698
CPF-Net [29]	0.973	0.699	0.801	0.766	0.927	0.605	0.755	0.716
C-Net [2]	0.975	0.724	0.827	0.794	0.926	0.597	0.757	0.699
LET-Net(Ours)	**0.977**	**0.740**	**0.835**	**0.815**	**0.930**	**0.615**	**0.772**	**0.727**

4.3 Ablation study

In this section, we conduct a series of ablation studies to verify the effectiveness of each critical component in our proposed LET-Net, including FLE, LRR, and mutual information loss.

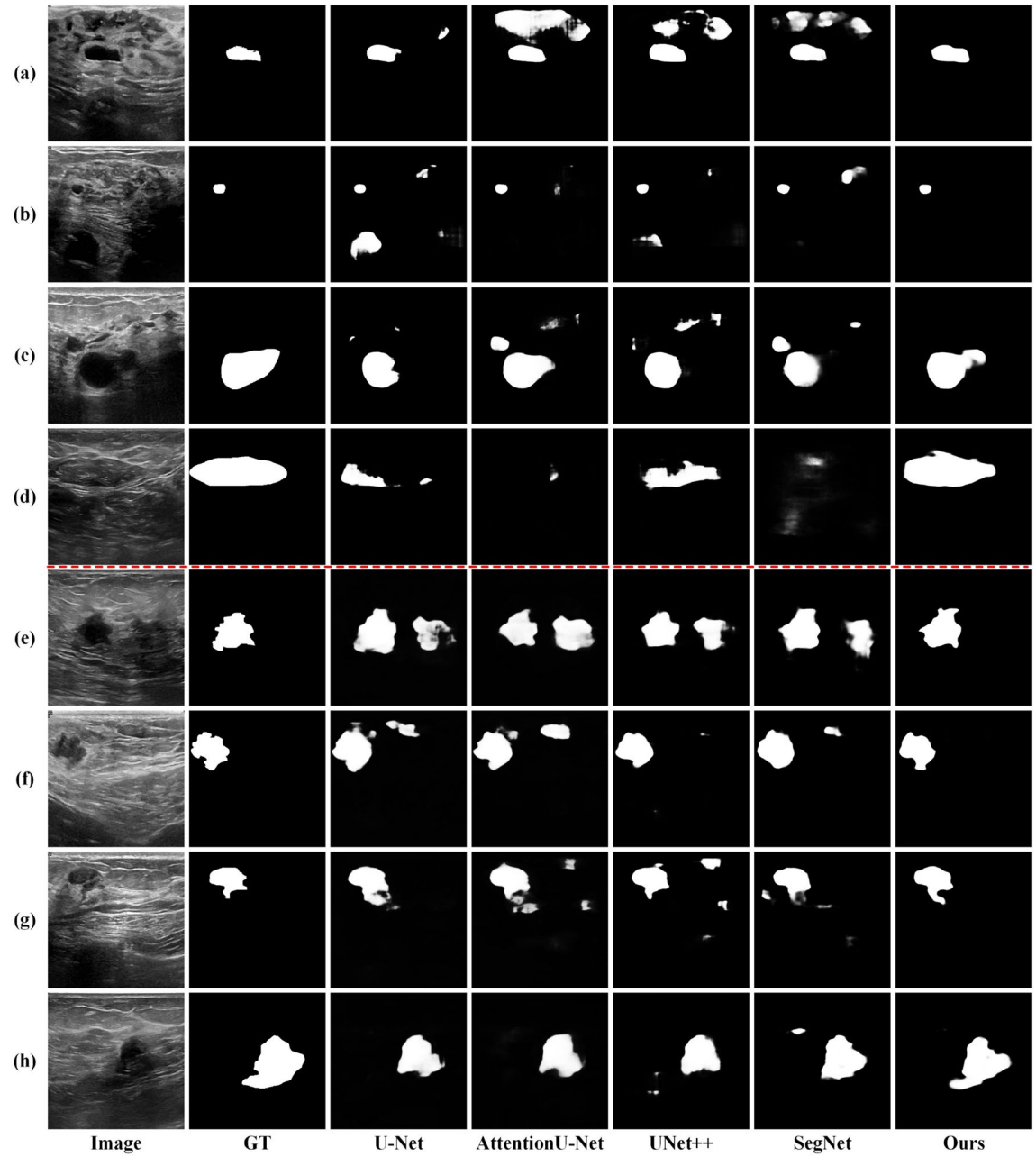

Fig. 7 Visual comparison among different methods in breast lesion segmentation, where the segmentation results of benign and malignant lesions are separated by a red dashed line

4.3.1 Impact of FLE and LRR modules

To validate the effectiveness of FLE and LRR modules, we remove them individually from our full net, resulting in two variants, namely w/o FLE and w/o LRR. As shown in Table 3, the variant without FLE (w/o FLE) achieves a 93.6% mDice score on CVC-ClinicDB dataset. When we apply the FLE module, the mDice score increases to 94.5%. Moreover, it boosts mDice by 1.6%, 2.2%, and 2.3% on CVC-ColonDB, ETIS-LaribPolypDB, and CVC-300 datasets, respectively. These results indicate that our FLE module effectively supports accurate segmentation due to its ability to learn discriminative local features under the feature alignment condition. Furthermore, when comparing the second and third lines of Table 3, it can be seen that LRR module is also conducive to segmentation, with performance gains of 1.6% and 1.7% in terms of mDice and mIoU on Kvasir dataset. The main reason is

Table 3 Ablation analysis w.r.t the effectiveness of FLE and LRR modules. The best results are shown in bold

Method	Seen dataset				Unseen dataset					
	CVC-ClinicDB		Kvasir		CVC-ColonDB		ETIS- Larib-PolypDB		CVC-300	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
w/o FLE	0.936	0.887	0.918	0.871	0.779	0.698	0.751	0.674	0.884	0.816
w/o LRR	0.940	0.894	0.910	0.859	0.790	0.711	0.759	0.681	0.890	0.821
LET-Net	**0.945**	**0.899**	**0.926**	**0.876**	**0.795**	**0.717**	**0.773**	**0.698**	**0.907**	**0.839**

Table 4 Ablation analysis of mutual information loss. The best results are shown in bold

Loss setting	Seen dataset				Unseen dataset					
	CVC-ClinicDB		Kvasir		CVC-ColonDB		ETIS- Larib-PolypDB		CVC-300	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
w/o L_{PPA}	0.937	0.888	0.917	0.864	0.782	0.697	0.737	0.663	0.885	0.812
w/o L_{VSD}	0.940	0.892	0.923	0.872	0.785	0.702	0.762	0.688	0.895	0.826
w/o L_{PPA} & L_{VSD}	0.934	0.882	0.914	0.861	0.772	0.692	0.716	0.648	0.879	0.807
LET-Net	**0.945**	**0.899**	**0.926**	**0.876**	**0.795**	**0.717**	**0.773**	**0.698**	**0.907**	**0.839**

that LRR module is capable of effective spatial recovery via its dynamic reconstruction kernels and split-attention mechanism, thereby facilitating segmentation.

4.3.2 Effectiveness of mutual information loss

To validate the effectiveness and necessity of our mutual information loss, we retrain our proposed LET-Net with different loss settings. Specifically, we denote three variants, i.e., w/o L_{PPA}, w/o L_{VSD}, and w/o L_{PPA} & L_{VSD}, each of which removes the corresponding loss item. Note that we apply conventional binary-cross entropy loss to supervise our model when removing L_{PPA}. Table 4 reports the quantitative evaluation. Comparing the first and fourth lines in Table 4, we can observe that our model performs poorly without PPA loss supervision, obtaining a 1.1% lower mIoU on CVC-ClinicDB dataset. Also, a similar dropping situation occurs with the variant w/o L_{VSD}. Specifically, our model has witnessed performance degradation without L_{VSD}, decreasing mIoU by 1.5%, 1%, and 1.3%, respectively, on CVC-ColonDB, ETIS-LaribPolypDB, and CVC-300 datasets. This confirms that each term in our total loss is effective for segmentation. The reasons can be summarized as: first, in contrast to binary-cross entropy loss, PPA loss can guide our model to pay more attention to local details by synthesizing local structure information of a pixel, resulting in superior performance. Second, L_{VSD} assists task-relevant feature learning, thereby improving the sensitivity of small objects. In addition, it can be seen that our method outperforms w/o L_{PPA} & L_{VSD} by a large margin, achieving 2.3% mDice and 2.5% mIoU performance gains with the help of our mutual information loss on CVC-ColonDB dataset. In summary, our experimental results fully demonstrate that mutual information loss is beneficial for LET-Net.

5 Conclusion

In this work, we propose a novel locally enhanced transformer network for accurate medical image segmentation. Our model adopts a PVT-based encoder to extract global contexts and utilizes a feature-aligned local enhancement module to highlight detailed local contexts while effectively recovering high-resolution spatial information by its progressive local-induced decoder. In addition, we design a mutual information loss to encourage our LET-Net to learn powerful representations from the task-relevant perspective. LET-Net is validated in polyp and breast lesion segmentation and achieves state-of-the-art performance, especially demonstrating its ability for small target segmentation. In future work, we aim to apply our proposed LET-Net to other medical image segmentation tasks with different modalities or anatomies, thereby developing our model to be more robust.

Acknowledgements This research is supported by the National Natural Science Foundation of China (62276112), the National Natural Science Foundation of China Regional Joint Fund of NSFC (U19A2057), Jilin Province Science and Technology Development Plan Key R &D Project (20230201088GX), and Collaborative Innovation Project of Anhui Universities (GXXT-2022-044).

References

1. Fang, Y., Chen, C., Yuan, Y., Tong, R.K.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 302–310 (2019). https://doi.org/10.1007/978-3-030-32239-7_34
2. Chen, G., Dai, Y., Zhang, J.: C-net: Cascaded convolutional neural network with global guidance and refinement residuals for breast ultrasound images segmentation. Comput. Methods. Programs Biomed. **225**, 107086 (2022)
3. Thomas, E., Pawan, S., Kumar, S., Horo, A., Niyas, S., Vinayagamani, S., Kesavadas, C., Rajan, J.: Multi-res-attention unet: a cnn model for the segmentation of focal cortical dysplasia lesions from magnetic resonance images. IEEE J. Biomed. Health Informat. **25**(5), 1724–1734 (2020)
4. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: A survey. IET Image Process. **16**(5), 1243–1267 (2022). https://doi.org/10.1049/ipr2.12419
5. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015). https://doi.org/10.1109/CVPR.2015.7298965
6. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017). https://doi.org/10.1109/TPAMI.2016.2644615
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28
8. Lou, A., Guan, S., Loew, M.: Cfpnet-m: A light-weight encoder-decoder based network for multimodal biomedical image real-time segmentation. Comput. Biol. Med. **154**, 106579 (2023)
9. Xie, X., Pan, X., Zhang, W., An, J.: A context hierarchical integrated network for medical image segmentation. Comput. Elect. Eng. **101**, 108029 (2022). https://doi.org/10.1016/j.compeleceng.2022.108029
10. Wang, R., Ji, C., Zhang, Y., Li, Y.: Focus, fusion, and rectify: Context-aware learning for covid-19 lung infection segmentation. IEEE Trans. Neural Netw. Learn. Syst. **33**(1), 12–24 (2021)
11. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
12. Cheng, J., Tian, S., Yu, L., Lu, H., Lv, X.: Fully convolutional attention network for biomedical image segmentation. Artif. Intell. Med. **107**, 101899 (2020)
13. Wang, X., Jiang, X., Ding, H., Liu, J.: Bi-directional dermoscopic feature learning and multi-scale consistent decision fusion for skin lesion segmentation. IEEE Trans. Image Processing **29**, 3039–3051 (2019)
14. Wang, X., Li, Z., Huang, Y., Jiao, Y.: Multimodal medical image segmentation using multi-scale context-aware network. Neurocomputing **486**, 135–146 (2022). https://doi.org/10.1016/j.neucom.2021.11.017
15. Liang, X., Li, N., Zhang, Z., Xiong, J., Zhou, S., Xie, Y.: Incorporating the hybrid deformable model for improving the performance of abdominal ct segmentation via multi-scale feature fusion network. Med. Image Anal. **73**, 102156 (2021)
16. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12179–12188 (2021)
17. Li, Y., Wang, Z., Yin, L., Zhu, Z., Qi, G., Liu, Y.: X-net: a dual encoding–decoding method in medical image segmentation. The Visual Computer, pp. 1–11 (2021)
18. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., 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, pp. 6881–6890 (2021)
19. Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
20. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 699–708 (2021). Springer
21. Shareef, B., Xian, M., Vakanski, A.: Stan: Small tumor-aware network for breast ultrasound image segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2020)
22. Lou, A., Guan, S., Ko, H., Loew, M.H.: Caranet: context axial reverse attention network for segmentation of small medical objects. In: Medical Imaging 2022: Image Processing, vol. 12032, pp. 81–92 (2022)
23. Valanarasu, J.M.J., Sindagi, V.A., Hacihaliloglu, I., Patel, V.M.: Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 363–373 (2020). Springer
24. Pang, Y., Zhao, X., Xiang, T.-Z., Zhang, L., Lu, H.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2160–2170 (2022)
25. Jia, Q., Yao, S., Liu, Y., Fan, X., Liu, R., Luo, Z.: Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4713–4722 (2022)
26. Dai, D., Dong, C., Xu, S., Yan, Q., Li, Z., Zhang, C., Luo, N.: Ms red: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. Med. Image Anal. **75**, 102293 (2022)
27. Xu, C., Qi, Y., Wang, Y., Lou, M., Pi, J., Ma, Y.: Arf-net: An adaptive receptive field network for breast mass segmentation in whole

mammograms and ultrasound images. Biomed. Signal Process Control **71**, 103178 (2022)
28. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. arXiv preprint arXiv:2203.04967 (2022)
29. Feng, S., Zhao, H., Shi, F., Cheng, X., Wang, M., Ma, Y., Xiang, D., Zhu, W., Chen, X.: Cpfnet: Context pyramid fusion network for medical image segmentation. IEEE Trans. Med. Imaging **39**(10), 3008–3018 (2020). https://doi.org/10.1109/TMI.2020.2983721
30. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging **39**(6), 1856–1867 (2019)
31. Li, X., You, A., Zhu, Z., Zhao, H., Yang, M., Yang, K., Tan, S., Tong, Y.: Semantic flow for fast and accurate scene parsing. In: European Conference on Computer Vision, pp. 775–793 (2020)
32. Mazzini, D.: Guided upsampling network for real-time semantic segmentation. arXiv preprint arXiv:1807.07466 (2018)
33. Lu, H., Dai, Y., Shen, C., Xu, S.: Indices matter: Learning to index for deep image matting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3266–3275 (2019)
34. Huang, Z., Wei, Y., Wang, X., Liu, W., Huang, T.S., Shi, H.: Alignseg: Feature-aligned segmentation networks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(1), 550–557 (2021)
35. Huang, S., Lu, Z., Cheng, R., He, C.: Fapn: Feature-aligned pyramid network for dense image prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 864–873 (2021)
36. Wu, J., Pan, Z., Lei, B., Hu, Y.: Fsanet: Feature-and-spatial-aligned network for tiny object detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. **60**, 1–17 (2022)
37. Hu, H., Chen, Y., Xu, J., Borse, S., Cai, H., Porikli, F., Wang, X.: Learning implicit feature alignment function for semantic segmentation. In: European Conference on Computer Vision, pp. 487–505 (2022)
38. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
39. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 253–262 (2020). https://doi.org/10.1007/978-3-030-59725-2_25
40. Tomar, N.K., Jha, D., Riegler, M.A., Johansen, H.D., Johansen, D., Rittscher, J., Halvorsen, P., Ali, S.: Fanet: A feedback attention network for improved biomedical image segmentation. IEEE Trans. Neural Netw. Learn, Syst (2022)
41. Shen, Y., Jia, X., Meng, M.Q.-H.: Hrenet: A hard region enhancement network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 559–568 (2021)
42. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019)
43. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
44. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 14–24 (2021)
45. He, X., Tan, E.-L., Bi, H., Zhang, X., Zhao, S., Lei, B.: Fully transformer network for skin lesion analysis. Med. Image Anal. **77**, 102357 (2022)
46. Yuan, F., Zhang, Z., Fang, Z.: An effective cnn and transformer complementary network for medical image segmentation. Pattern Recogn **136**, 109228 (2023)
47. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J., Merhof, D.: Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6202–6212 (2023)
48. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: Fat-net: Feature adaptive transformers for automated skin lesion segmentation. Med. Image Anal. **76**, 102327 (2022)
49. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Adv. Neural Info. Processing Syst **28**, (2015)
50. Song, J., Chen, X., Zhu, Q., Shi, F., Xiang, D., Chen, Z., Fan, Y., Pan, L., Zhu, W.: Global and local feature reconstruction for medical image segmentation. IEEE Trans. Med. Imaging (2022)
51. Zhang, Q.-L., Yang, Y.-B.: Sa-net: Shuffle attention for deep convolutional neural networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2235–2239 (2021)
52. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
53. Ma, N., Zhang, X., Zheng, H.-T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
54. Zhao, S., Wang, Y., Yang, Z., Cai, D.: Region mutual information loss for semantic segmentation. Adv. Neural Info. Processing Syst. **32**, (2019)
55. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. arXiv preprint arXiv:1612.00410 (2016)
56. Tian, X., Zhang, Z., Lin, S., Qu, Y., Xie, Y., Ma, L.: Farewell to mutual information: Variational distillation for cross-modal person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1522–1531 (2021)
57. Wei, J., Wang, S., Huang, Q.: F^3net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12321–12328 (2020)
58. Yin, Z., Liang, K., Ma, Z., Guo, J.: Duplex contextual relation network for polyp segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2022). https://doi.org/10.1109/ISBI52829.2022.9761402
59. Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 263–273 (2020). https://doi.org/10.1007/978-3-030-59725-2_26
60. Patel, K., Bur, A.M., Wang, G.: Enhanced u-net: A feature enhancement network for polyp segmentation. In: 2021 18th Conference on Robots and Vision (CRV), pp. 181–188 (2021). https://doi.org/10.1109/CRV52889.2021.00032
61. Ta, N., Chen, H., Lyu, Y., Wu, T.: Ble-net: boundary learning and enhancement network for polyp segmentation. Multimed. Syst. 1–14 (2022)
62. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp

highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. **43**, 99–111 (2015). https://doi.org/10.1016/j.compmedimag.2015.02.007
63. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., Lange, T.d., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling, pp. 451–462 (2020)
64. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging **35**(2), 630–644 (2016). https://doi.org/10.1109/TMI.2015.2487997
65. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. **9**(2), 283–293 (2014). https://doi.org/10.1007/s11548-013-0926-3
66. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.C.: A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthc. Eng. (2017)
67. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
68. Abraham, N., Khan, N.M.: A novel focal tversky loss function with improved attention u-net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 683–687 (2019)
69. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1055–1059 (2020)