



## Article

# Automatic Medical Image Segmentation with Vision Transformer

Jie Zhang <sup>1,2,\*</sup> , Fan Li <sup>1</sup>, Xin Zhang <sup>1</sup>, Huaijun Wang <sup>1</sup> and Xinhong Hei <sup>1</sup> 

<sup>1</sup> School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China; lf53mail@gmail.com (F.L.); xinzhang246745@foxmail.com (X.Z.); wanghuaijun@xaut.edu.cn (H.W.); heixinhong@xaut.edu.cn (X.H.)

<sup>2</sup> Provincial Key Laboratory of Network Computing and Security Technology, Xi'an University of Technology, Xi'an 710048, China

\* Correspondence: jiezhong1984@xaut.edu.cn

**Featured Application:** This work investigates the application of transformers for general medical image segmentation tasks. The key future directions involve developing unified AI-based system for integrated diagnosis, interactive segmentation to enable human.

**Abstract:** Automatic image segmentation is vital for the computer-aided determination of treatment directions, particularly in terms of labelling lesions or infected areas. However, the manual labelling of disease regions is inconsistent and a time-consuming assignment. Meanwhile, radiologists' comments are exceedingly subjective, regularly impacted by personal clinical encounters. To address these issues, we proposed a transformer learning strategy to automatically recognize infected areas in medical images. We firstly utilize a parallel partial decoder to aggregate high-level features and then generate a global feature map. Explicit edge attention and implicit reverse attention are applied to demonstrate boundaries and enhance their expression. Additionally, to alleviate the need for extensive labeled data, we propose a segmentation network combining propagation and transformer architectures that requires only a small amount of labeled data while leveraging fundamentally unlabeled images. The attention mechanisms are integrated within convolutional networks, keeping their global structures intact. Standalone transformers connected straightforwardly and receiving image patches can also achieve impressive segmentation performance. Our network enhanced the learning ability and attained a higher quality execution. We conducted a variety of ablation studies to demonstrate the adequacy of each modelling component. Experiments conducted across various medical imaging modalities illustrate that our model beats the most popular segmentation models. The comprehensive results also show that our transformer architecture surpasses established frameworks in accuracy while better preserving the natural variations in anatomy. Both quantitatively and qualitatively, our model achieves a higher overlap with ground truth segmentations and improved boundary adhesion.

**Keywords:** automatic segmentation; transformer; medical images



**Citation:** Zhang, J.; Li, F.; Zhang, X.; Wang, H.; Hei, X. Automatic Medical Image Segmentation with Vision Transformer. *Appl. Sci.* **2024**, *14*, 2741. <https://doi.org/10.3390/app14072741>

Academic Editor: Christos Bouras

Received: 30 January 2024

Revised: 26 February 2024

Accepted: 29 February 2024

Published: 25 March 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Medical image segmentation using automated techniques is vital for many clinical tasks, including discovering new biomarkers, monitoring disease progression, and facilitating computer-assisted determination and treatment direction [1], particularly in the labeling of lesions or infected areas during public health crises. The recent advances in deep neural network structures have led to major enhancements in the performance of medical image segmentation models. However, medical image segmentation tasks exhibit immense diversity, as different imaging modalities produce images with distinct characteristics and appearances while the presentation of diseases also varies considerably. This heterogeneity makes the direct application of even a successful architecture such as U-Net [2], a convolutional network for biomedical image segmentation, to a new task unlikely to achieve optimal results. Since the manual annotation of infected areas is a

laborious and time-consuming task, and radiologists' annotations are highly subjective and prone to personal bias informed by past clinical encounters, it is an urgent requirement to explore automated and robust segmentation techniques for medical images [3].

Segmenting organs and lesions in computed tomography (CT) slides provides critical clinical data for disease identification and quantification [4]. Recently, many algorithms using feature extraction and classification techniques have been developed for lung nodule segmentation, achieving promising performance. Nevertheless, visual similarities between nodules and the background make isolating them challenging. To address these challenges, researchers have developed deep learning approaches to learn more discriminative visual representations for segmentation [5–9]. For instance, from heterogeneous CT slices, the authors of [5] introduced a centrally focused convolutional neural network (CNN) to isolate lung nodules. Meanwhile, in generative adversarial networks, GAN-synthesized information is utilized to enhance the discriminative training model for lung pathology segmentation. A dual deep network that incorporates multiple residual streams at varying resolutions [9] was designed to isolate lung tumors in CT images [8]. In summary, deep learning shows promise for improving the segmentation of lung nodules and pathology markers in CT scans via learning robust features to overcome the difficulty of differentiating nodules from image backgrounds.

Artificial intelligence technologies have been widely applied to combat COVID-19 (coronavirus disease 2019), with many deep learning systems having been proposed for detecting infected patients through radiological imaging. However, labeled data are still limited. In this context, semi-supervised models can identify target regions from other anomalies given minimal labeling, making them well-suited for COVID-19 assessments. Additionally, transformer learning methods constitute another promising approach for situations with limited data [10,11]. Transformers excel at modeling global contexts and, with large-scale pre-training, demonstrate superior transfer learning ability on downstream tasks. This has been evidenced by their success in natural language processing (NLP) and machine translation tasks [12,13]. Recently, transformers have achieved or surpassed cutting-edge performance on various image recognition challenges [6,14]. In summary, semi-supervised learning and transformers are two effective artificial intelligence (AI) techniques which can address the lack of labeled medical imaging data, capitalizing on their abilities to learn from limited supervision and leverage pre-training. Their adoption could enable accurate screening from radiological images despite data constraints.

Detecting infected regions and affected areas in medical images remains challenging due to the variance in morphological characteristics like texture, size, and positioning across different medical image modalities. Additional difficulties arise from subtle visual differences between normal and infected tissues, as well as low-intensity contrast. Moreover, collecting extensive labeled infection data is arduous, as acquiring precise pixel-level annotations to train deep models is expensive and time-consuming [15]. Infection detection in medical images is difficult due to infections' heterogeneous appearances, their visual similarity to normal tissue, and the lack of abundant annotated data. Overcoming these obstacles is key to improving automated infection detection from medical images.

Here, we develop a novel medical segmentation network that establishes self-attention mechanisms for sequence-to-sequence prediction. In order to compensate for transformers' need to sacrifice fine-grained feature resolution, we employ a hybrid CNN–Transformer structure. This combines the advantages of high-resolution spatial details from CNN representations with the overall context modeling capacities of transformers. Transformers are better at modeling long-range dependencies in images, more flexible at handling varying input sizes, and they provide greater model interpretability, offering a promising alternative to CNNs for medical image segmentation. We integrate transformers in medical segmentation, providing both locally precision segmentation and global contextual modeling for accurate analysis.

Inspired by parallel structure designs, the self-attentive features from transformers are up-sampled and combined with CNN features of high-resolution skipped from the

encoding path. This fusion strategy enables the precise delineation of boundaries while preserving the transformers' self-attention mechanism advantages for medical segmentation. Empirical experimental results suggest our transformer framework better leverages self-attention compared to existing self-attention CNN-based models. Simultaneously, incorporating more intensive low-level feature details generally improves segmentation accuracy. Extensive experiments across various medical image modalities demonstrate our method superiority over compared approaches. Our parallel integration of transformer global context modeling with CNN local detail extraction, coupled with multi-level feature fusion, allows transformers to significantly enhance the performance of medical image segmentation tasks. In summary, our framework seamlessly combines the strengths of transformers and CNNs, facilitating accurate and robust segmentation of medical images by effectively capturing both global and local contextual information.

In summary, our main contributions in this work are as follows:

- We develop the attempt to investigate the integration of transformer architectures into general medical image segmentation tasks, demonstrating their applicability beyond NLP.
- We introduce a novel transformer-based network tailored for infection segmentation, designed to mitigate the challenges posed by limited labeled data availability. Our system exhibits superior learning capabilities through random propagation and self-supervised pre-training on unlabeled data.
- Our proposed framework achieves high-performance segmentation results in a self-supervision manner, highlighting the significant potential of pre-training transformers on large unlabeled medical image datasets to learn rich representations that can be effectively transferred to the segmentation task.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of related work, including existing segmentation methods and relevant techniques. Section 3 details our conceptual framework and the methodologies employed in our automatic segmentation network. Section 4 reports the extensive experimental results and provides in-depth analysis. Finally, Section 5 concludes and discusses future research directions.

## 2. Related Works

In this section, we review related studies that are most relevant to our work, encompassing the following areas: medical image segmentation tasks, approaches that combine self-attention mechanisms with CNNs, and the application of different transformer architectures in computer vision.

### 2.1. Medical Segmentation Studies

Medical segmentation tasks pose several challenges, including the scarcity of manual annotations and high memory requirements for processing high resolution image scans. In contrast with natural images frameworks like DeepLab [16] and PSPNet [17], U-Net [2]-based architectures like VNet [18], U-Net++ [19], HDenseU-Net [20], and nnU-Net [21] demonstrate superior performance in preserving fine anatomical details while being more memory efficient.

Attention mechanisms have gained popularity for U-Nets in medical image segmentation, enabling models to better capture informative features and spatial dependencies [22]. Recurrent attention in RAAGR2-Net [23] focuses on relevant regions during segmentation. Multi-scale attention gates in M-Unet [24] extract local and global context across scales. PsLSNet [25] leverages position-sensitive self-attention to model intricate spatial relationships. Additionally, MSAU-Net [26] attends to features across multiple scales simultaneously, while SegR-Net [27] reweights features spatially and channel-wise via attention. U-NetPlus [28] incorporates attention gates and dense skip connections for enhanced information flow. Moreover, U-Det [29] jointly performs detection and segmentation by integrating attention. Overall, these various attention mechanisms improve U-Nets for

medical images by enabling more effective feature representation and spatial dependency modeling, leading to gains in segmentation accuracy.

Traditional semi-supervised medical image segmentation methods primarily rely on handcrafted features to contrive models, including the prior-based [30] and clustering-based models [31]. Performance often depends heavily on the quality of the feature representation. Adversarial learning is also leveraged to exploit the unlabeled data. For instance, with uncertainty guidance [32], the mean teacher model [33] is extended for left atrium segmentation in a semi-supervised way. Li et al. [34] introduced a shape-aware strategy leveraging unlabeled data and geometric constraints. However, our method synergizes local precision and global modeling via a simple yet effective CNN–transformer fusion tailored for medical images. By combining complementary strengths, it achieves accurate analysis without hand-tuning features or complex adversarial objectives. Unlabeled data improve the consistency regularization.

## 2.2. Combining Self-Attention Mechanisms into CNNs

Numerous studies have explored the incorporation of CNNs and self-attention mechanisms, aiming to model global interactions between pixel-level features. For example, a non-local operator was designed, which could be inserted into multiple convolution layers [35]. Based on an encoder–decoder parallel structure, additive attention gate modules [36] have been proposed to inject self-attention into skip-connections.

Unlike the existing approaches, our method employs transformers to embed global self-attention. In our framework, transformers enable the capture of long-range dependencies and global context that complement CNNs’ local feature extraction capabilities, providing a novel and effective way of incorporating self-attention for medical image segmentation tasks.

## 2.3. Transformers in Vision

Originally developed for machine translation [37], transformer architectures have achieved state-of-the-art results on numerous NLP tasks [38]. More recently, transformer models have been adapted to computer vision tasks [39], such as image classification [40], saliency detection [41], and semantic segmentation [42], demonstrating their strong potential in this domain.

To apply transformers to image data, the authors of [43] employed self-attention locally instead of globally on image regions. In [44], the authors used approximations for scalable self-attention computation. The vision transformer (ViT) [45] directly applies global self-attention transformers on full-sized images, matching the performance of CNN on the ImageNet classification task for large-scale visual recognition. Our medical segmentation framework builds upon the ViT architecture. Transformers show significant potential for medical image analysis, especially for segmentation tasks, due to their better handling of long-range dependencies, flexibility in input size, opportunities for transfer learning, improved global context modeling, enhanced interpretability, and efficient training mechanisms. Subsequently, our work represents an initial effort to integrate transformers and their benefits of global context modeling into the domain of medical image segmentation.

## 3. Method

Given an input image  $x \in \mathbb{R}^{H \times W \times C}$  with spatial dimensions  $H \times W \times C$  channels, our objective was to predict the corresponding  $H \times W$  label map through pixel-wise segmentation. The common approach trains a CNN, such as U-Net, to directly encode the input image into high-level feature representations and then decode these representations back to full spatial resolution.

In contrast to the current approaches, our framework incorporated self-attention mechanisms via transformer encoders. We first explain how transformers can directly encode features from decomposed image patches (Section 3.1). Unlike CNN encoders, our transformer encoder incorporated self-attention to model global contexts and long-range

dependencies in the image. This provides complementary information to aid medical image segmentation. Then, we elaborate on the overall framework.

Rather than completely replacing CNNs, we selectively applied transformers to leverage both local precision from convolutions and global reasoning from self-attention. This hybrid architecture balances strengths by dividing responsibilities—the transformer focuses solely on encoding semantics, while the CNN decodes for spatial acuity.

In summary, our key insight is to combine complementary components in a tailored encoder–decoder design for medical images. This moves beyond the existing methods by fusing advancements from both computer vision and NLP to best suit the problem domain of medical image segmentation.

### 3.1. Transformers Encoder

Following the vision transformer (ViT) [45], we first tokenized the input image  $x$  by reshaping it into a sequence of flattened 2D patches,  $x_p^i \in \mathbb{R}^{p^2 \cdot C} \mid i = 1, \dots, N$ , and every image patch was  $P \times P$  size. The patches number  $N = \frac{H \times W}{P^2}$  was as the input sequence length.

We employed a trainable linear projection layer to map these image patches  $x_p$  into a  $D$ -dimensional latent space. To incorporate spatial information, we added positional embeddings to the patch embeddings before passing them through the projection layer:

$$z_0 = [x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{pos} \quad (1)$$

Here,  $E \in \mathbb{R}^{(p^2 \cdot C) \times D}$  is the linear projection of patch embedding, and  $E \in \mathbb{R}^{(p^2 \cdot C) \times D}$  is the position embedding. The input image is divided into patches, which are then projected into an embedding space. Learned positional encodings are provided through learned embedding vectors. This tokenization and embedding process allows the transformer to model interactions between image patches for global context modeling. In other words, the image is divided into local patches and each patch is mapped to a vector space that retains spatial relationships. This tokenization presents the transformer with visual concepts and their arrangements, priming it to model semantic and positional interactions. The output patch embeddings summarized identity and layout to inform the downstream decoding.

The transformer encoder contains  $L$  identical layers, with each layer containing a multi-layer perceptron (MLP) block and a multi-head self-attention (MSA) block (Equations (2) and (3)). The MSA block first projects the embeddings into query, key, and value vectors. It then calculates attention based on the dot-product over all tokens using these projections. This models the global relationships in the image. The MLP then processes each token individually with non-linear transformations to encode semantic concepts. Layer normalization and residual connections are employed around each block. The  $\ell$  – layer output can be obtained by:

$$z'_\ell = \text{MSA}(\text{LN}(z_{\ell-1})) + z_{\ell-1} \quad (2)$$

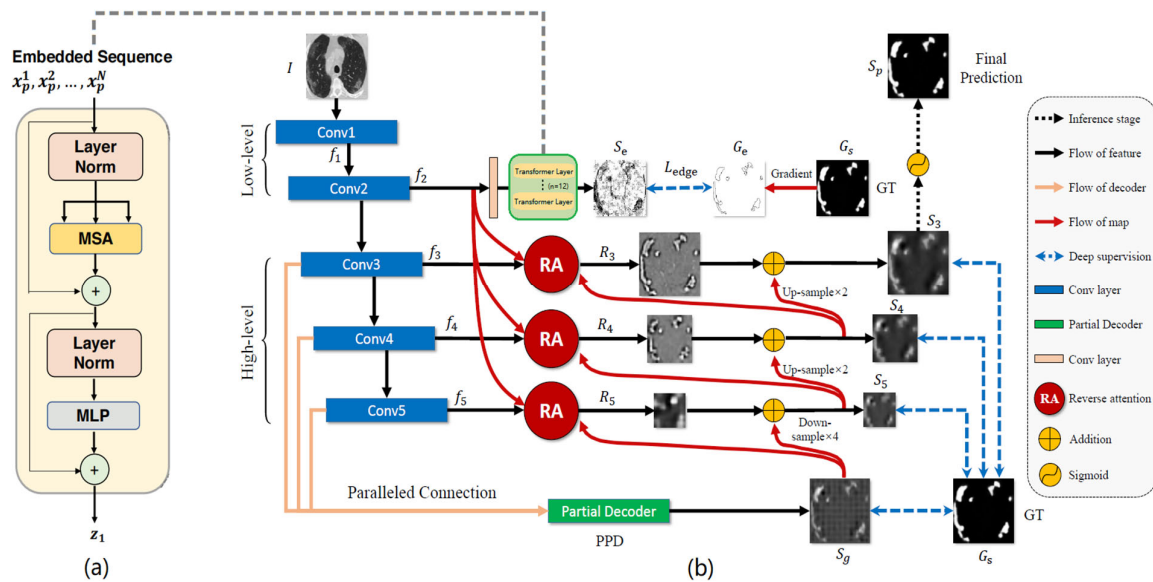
$$z_\ell = \text{MLP}(\text{LN}(z'_\ell)) + z'_\ell \quad (3)$$

Here,  $\text{LN}(\cdot)$  represents applying layer normalization, and  $z_L$  is the transformer output encoding the image features after  $L$  layers. The stacked MSA and MLP blocks enable complex reasoning through multiple rounds of global self-attentive processing and local non-linear transformations.

Figure 1a illustrates the structure of the transformer layers. There are 12 transformer layers in the second convolutional block of the CNN encoder, and each transformer encoder contains multiple identical layers stacked on top of each other, with each layer containing two sublayers, a multi-head self-attention (MSA) layer and a multi-layer perceptron (MLP) layer. Since we mapped the vectorized patches into a latent  $D$ -dimensional embedding space using a trainable linear projection, the number of hidden dimensions and attention



heads can adjust based on the input image dimensions, reflecting the transformer’s capability to accommodate images of varying sizes—a crucial and practical feature in clinical settings. The self-attention mechanism models the global interactions between image patches, while the MLP provides non-linear transformations to encode semantic concepts. Stacking these layers allows modeling of the global context and long-range dependencies present in the input image.



**Figure 1.** The framework overview. (a) The schematic of transformer layer; (b) structure of our proposed framework.

### 3.2. Infection Segmentation Network

#### 3.2.1. Framework Overview

The overview of our framework’s architecture is shown in Figure 1. Firstly, the input medical images are passed through two convolutional layers to obtain high-resolution feature representations with low-level semantic. An edge attention is then applied to explicitly enhance the representation of region boundaries.

The obtained low-level features  $f_l$  are passed through three convolutional layers to extract high-level features  $f_h$ . These high-level features serve two purposes. First, they are aggregated in parallel by a partial prediction decoder (PPD)  $f_h$  to generate a coarse global infection localization map  $S_g$ . Second,  $f_h$  combined with  $S_g$  are passed through a cascade of reverse attention (RA) modules guided by the localization map  $S_g$ . Notably, each RA module guides the subsequent RA module in the cascade. Finally, the output of the last RA module passes through a sigmoid activation function to predict the infection regions.

Our network extracted multi-level features, utilized self-attention to model global context, and hierarchically refined localization via a cascaded decoder design for precise infection segmentation. The cascaded decoder mimics the workflow of radiologists, providing context and direction to focus the model on areas of interest. This hierarchical strategy tailors the model capacity to efficiently improve segmentation performance.

#### 3.2.2. Edge Attention Module

As shown in prior work, edge information provides useful constraints to guide feature extraction for segmentation tasks [46–48]. Since low-level feature representations (e.g.,  $f_l$  in our model) contain several edge features at a moderate resolution, we fed these low-level feature representations  $f_l$  to a newly introduced edge attention (EA) module. This module learns an explicit edge-aware feature representation. Meanwhile,  $f_l$  passes through a convolutional layer with a single filter to produce an edge map prediction.

The edge attention module improves medical image segmentation in the following way. It first identifies edges and boundaries between structures in the input image using gradient information. These detected edges provide vital cues for delineating infection regions from surrounding healthy tissue. The predicted edge maps are then used to refine the convolutional feature representations extracted from the image. By element-wise multiplying the edge maps with the convolutional filters, the features are enhanced at semantically meaningful edges, incorporating critical localization cues. The refined edge-aware features help guide the network's attention to focus on the most relevant regions for accurate segmentation. The edge-guided spatial attention provides fine-grained localization signals to help distinguish infection areas from other tissues. In essence, the edge attention module incorporates explicit edge knowledge to enhance the convolutional features. By refining features and steering attention based on edges, the module enables the more precise delineation of infection regions, thus improving medical image segmentation performance. The edge-guided spatial attention is crucial for accurate localization of infection regions.

We subsequently assessed the disparity between the edge map prediction from the EA module and the ground-truth (GT) edge map  $G_e$  derived from the annotated segmentation tasks. This was performed by computing the binary cross entropy (BCE) loss function:

$$\mathcal{L}_{edge} = - \sum_{x=1}^w \sum_{y=1}^h [G_e \log(S_e) + (1 - G_e) \log(1 - S_e)] \quad (4)$$

Here,  $(x, y)$  indexes the spatial coordinates within the predicted edge map  $S_e$  and the ground-truth edge map  $G_e$ . The ground-truth edge map  $G_e$  is derived by computing the gradient of the annotated segmentation mask. The  $w$  and  $h$  are the corresponding width and height of the edge maps, respectively. The binary cross-entropy loss compares the predicted and ground-truth edge maps at each pixel location, providing localized supervision for predicting salient boundaries. By explicitly guiding the model to focus on edges, it learns to delineate challenging vague transitions between infection regions and healthy tissues.

We exploited low-level edge information by passing the convolutional features through the edge attention module. By optimizing the edge map predictions against the ground-truth edge maps using the BCE loss, it guides the network to obtain edge-aware feature representations that are beneficial for the segmentation task. The BCE loss compares the predicted and ground-truth edge maps pixel-by-pixel, guiding the network to learn edge-aware features that can effectively capture and represent the boundary information present in the input image.

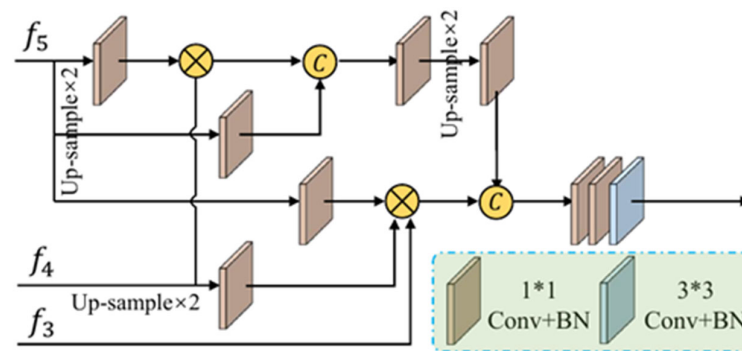
### 3.2.3. Parallel Partial Decoder

Many established medical segmentation frameworks utilize both high-level and low-level representations extracted by the encoder to segment organs/lesions [2,19,49–51]. However, as pointed out in [47], compared to high-level feature representations, low-level representations require substantially increased computational resources due to their larger spatial resolutions, while contributing minimally to performance improvements.

We introduced a parallel partial decoder (PPD), which is beneficial for medical image segmentation by allowing efficient multi-scale feature utilization, information fusion, and balancing precision with efficiency while being robust to scale variations. The parallel branches in the PPD allow for the concurrent decoding of features from multiple scales of the encoder. This incorporates both fine local details from early layers and high-level global context from later layers. The fusion of multi-scale information enables a more comprehensive understanding of image structures, improving segmentation accuracy. The decoder can selectively aggregate multi-scale features to balance precision and efficiency. To prioritize precision for small structures, fine-grained features are emphasized. For scenarios where efficiency is prioritized over some precision sacrifice, coarse-scale features are utilized instead. This configurability allows catering the trade-off to diverse segmentation scenarios with varying precision needs and resource constraints. In essence, the PPD facilitates the

flexible utilization of multi-scale encoder representations. By fusing features from layers of varying granularity, it balances segmentation precision and efficiency in an application-specific manner. The configurability enables optimized trade-offs for different precision requirements and computational budgets.

This motivated a more efficient utilization of multi-scale encodings in our framework. Rather than naively fusing layers, we strategically applied low-level features to refine boundaries via our edge attention module. The high-level features focus solely on generating an initial coarse prediction to localize regions of interest, as shown in Figure 2.



**Figure 2.** Paralleled partial decoder to obtain global map.

Our PPD aggregation and cascaded reverse attention streams then allow low-level cues to be exploited in a targeted manner. By deliberating directing model capacity based on scale utility, we balance precision and efficiency. Experiments validated that our structured approach prevents wasteful redundancy while improving contour accuracy. These optimizations help address deployment constraints concerning inference speed, memory footprint, and power consumption.

Specifically, given an input medical image, we extracted two low-level features  $\{f_i, i = 1, 2\}$  and three high-level features  $\{f_i, i = 3, 4, 5\}$  using Res2Net convolutional blocks [52]. We then aggregated the high-level feature maps through the PPD  $p_d(\cdot)$  [53] to produce a coarse global localization map  $S_g = p_d\{f_3, f_4, f_5\}$ .

### 3.2.4. Reverse Attention Module

Clinicians typically segment regions of infection in a two-step process—coarsely localizing regions of infection, and then accurately annotating local tissue structures. We mirrored this workflow with two components. First, the PPD acts as a coarse locator, generating a global localization map  $S_g$  indicating the rough locations of an infection region without structural details. Next, a progressive reverse attention (RA) framework acts as the refined labeler to discriminatively delineate infections by removing estimated regions from high-level feature representations [53,54]. Rather than simply aggregating all levels [54], we dynamically acquired reverse attention across three parallel high-level feature maps. This progressively exploited complementary regions and details from the deeper layers. This was achieved by eliminating elements of the current estimations from the up-sampled high-level features.

The resulting RA features  $R_i$  are obtained through fusing high-level feature maps  $\{f_i, i = 3, 4, 5\}$  and edge-aware features  $e_{att} = f_2$  from the edge attention module using the reverse attention weights  $A_i$ :

$$R_i = C(f_i, Dow(e_{att})) \odot A_i \quad (5)$$

where  $dow(\cdot)$  is the down-sampling operation;  $C(\cdot)$  is the concatenation operation. Our network mimics the clinician workflow with coarse localization followed by fine labeling of structural details. The global guidance provided by the localization map and the progres-

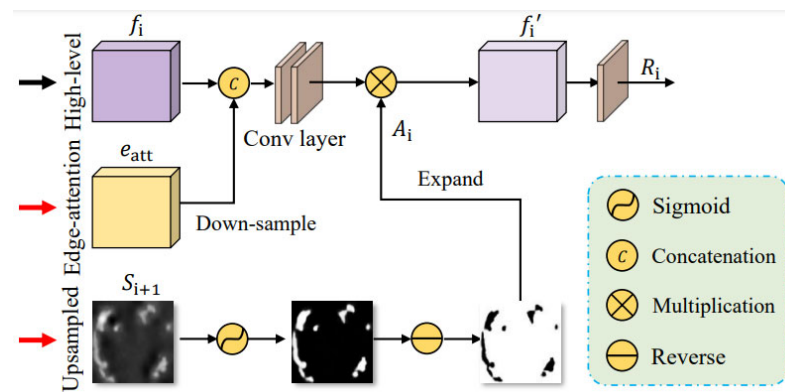


sive reverse attention mechanism allow the module to selectively zoom in on informative regions.

The RA weight  $A_i$  is crucial for salient object detection [46], with the definition as:

$$A_i = \varepsilon(\ominus(\sigma(P(S_{i+1})))) \quad (6)$$

where  $P(\cdot)$  is the up-sampling operation;  $\sigma(\cdot)$  represents the sigmoid activation function; and  $\ominus(\cdot)$  denotes the reversed subtraction operation between the input and a matrix with each element as 1. The reversed subtraction  $\varepsilon$  subtracts the localization map from a matrix of ones to reverse the activation in Equation (5). Figure 3 shows the details of this procedure.



**Figure 3.** Utilizing reverse attention module to learning edge features implicitly.

This reverse attention mechanism essentially reverses the background to foreground patterns learned by the CNN, helping to disentangle infection regions from normal tissue. The  $A_i$  weights then filter out irrelevant regions while passing through discriminative cues that are progressively refined across stages. Together with the edge guidance from the edge attention module, this multi-level erasing and fusion process elegantly captures subtle characteristics that are difficult for standalone networks to model, moving towards an expert-level understanding.

Importantly, the erasing strategy, which is guided by the RA module, progressively refines the initial coarse prediction map into a more accurate final segmentation prediction. The reverse attention weights are learned to erase the prior coarse estimations and focus on complementary infection details. By iteratively refining the localization in this manner, the initial coarse prediction is transformed into a precise final segmentation map.

The reverse attention module progressively refines coarse segmentation predictions in a hierarchical manner to enhance accuracy and capture finer details. It disentangles infection regions from normal tissues by using attention mechanisms to selectively focus on amplifying features of infection regions while suppressing irrelevant features of normal tissues. This highlights infection boundaries and contrasts between infected and healthy areas.

The module employs an iterative erasing strategy that gradually removes less relevant features from the coarse predictions to emphasize salient infection-related features. This erasing strategy aligns with the intuition of selective attention, where the network iteratively attends to and refines the most infection-relevant features while filtering out distracting or less informative features.

Through this progressive selective attention and erosion of non-essential features, the reverse attention module disentangles infections from complex surroundings and refines the segmentation map for improved accuracy and localization.

### 3.2.5. Loss Function

We employed a weighted intersection over union (IoU) loss and a symmetric binary cross-entropy (BCE) loss for supervision during training. The weighted IoU loss allows

higher importance to be assigned to certain classes or regions, helping the model prioritize clinically significant areas and improving segmentation accuracy in critical regions. It addresses class imbalance and differences in contextual importance. The asymmetric BCE loss allows customization based on clinical implications. It mitigates class imbalance issues and makes the model more robust to misclassifications. The weighted IoU guides the focus on salient areas, while the asymmetric BCE improves robustness. This makes the model more effective at delineating anomalies from normal tissues.

Together, these losses align the model with clinical objectives by emphasizing critical regions, balancing class importance, and penalizing asymmetric errors tailored to medical imaging challenges. By combining losses tailored for medical images, the model becomes better equipped to handle complexities like class imbalance while improving performance in clinically valuable regions.

As Equation (4) mentioned, we take the edge loss function  $\mathcal{L}_{edge}$  for supervision. The full loss function  $\mathcal{L}_{edge}$  comprises a weighted intersection over union loss (IoU)  $\mathcal{L}_{IoU}^w$ , combined with a weighted BCE loss  $\mathcal{L}_{BCE}^w$  for segmentation supervision,

$$\mathcal{L}_{seg} = \mathcal{L}_{IoU}^w + \mathcal{L}_{BCE}^w \quad (7)$$

The proposed loss function consists of a weighted IoU and an asymmetric BCE loss. The global and local losses provide effective supervision at the image and pixel levels, respectively, for more accurate segmentation. Different from the standard IoU loss, the weighted IoU loss emphasizes hard pixels to highlight their importance, while the asymmetric BCE assigns greater weight to challenging mislabeled hard pixels, instead of treating all pixels equally. The formulation of these losses draws from successful applications in other domains like salient object detection [55,56]. Correntropy-induced losses [57,58] could also be explored here for improved robustness.

Additionally, we adopt the global map  $S_g$  as well as three side-outputs (i.e.,  $S_3$ ,  $S_4$ , and  $S_5$ ) with deep supervision by comparing their up-sampled versions ( $\mathcal{L}_3^{up}$ ) to the ground truth segmentation mask during training. This provides direct optimization guidance at multiple encoder stages. The total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{seg}(G_s, S_g^{up}) + \mathcal{L}_{edge} + \sum_{i=3}^{i=5} \mathcal{L}_{seg}(G_s, S_i^{up}) \quad (8)$$

where  $G_s$  is the labeled mask, and  $\mathcal{L}_{edge}$  denotes cross-entropy loss for edge prediction. This benefits gradient flow across features and prevents overfitting.

#### 4. Experiments and Discussions

In our experiments, standard augmentations like random rotation and flipping are applied. We utilized a 12 layer Vision Transformer (ViT) [45] architecture for the pure transformer encoder model, with an input resolution of  $8 \times 8$  and a patch size of 16, unless specified. Four  $2 \times$  up-sampling was used in the cascaded up-sampler (CUP) for full resolution. The models were trained for 20k iterations using the Adam optimizer with a learning rate of 0.0001 and a batch size of 6. The training was performed on an Nvidia RTX2080Ti GPU (Graphics Processing Unit, TechPowerUp, Yorkshire, NY, USA).

We demonstrated our framework on different medical image datasets in distinct modalities, CT and skin image dataset. The CT segmentation dataset contains 100 axial CT scans from COVID patients, with radiologist segmentations of lung infections, and the HAM10000 dataset [59], Human Against Machine, contains 10,000 training images, a large collection of dermatoscopic images from multi-sources showing common pigmented skin lesions, including the ISIC (International Skin Imaging Collaboration) 2016 and 2018 skin lesion datasets. The COVID CT dataset contains 349 CT scans from 216 patient cases, with extracted metadata like age, gender, location, medical history, scan time, COVID severity, and radiology reports. It has more COVID positive cases than other datasets, providing greater image diversity.

The HAM10000 dataset includes ISIC 2016 (900 images) and ISIC 2018 (1000 images) of dermoscopic images from different populations and modalities, covering important diagnostic categories of pigmented skin lesions.

The datasets were split into training (70%), validation (20%), and testing (10%) sets in a 7:2:1 ratio. All images were resized to  $512 \times 512$  pixels.

The combined datasets include CT scans across age groups, as well as chest and eye images, reflecting a rigorous and diverse collection for model training and evaluation. Utilizing these diverse data spanning COVID CT, dermatology cases, and varied anatomies/modalities aimed to develop robust, generalizable segmentation models for clinical use.

These datasets contain expert annotations and clinical severity ratings, using standard training schemes and hardware. This benchmarks performance on realistic data.

#### 4.1. Experimental Setting

Thorough validation is essential for developing trust in automatic segmentation models prior to clinical use. Comprehensive validation provides confidence that these models can generalize to new data while maintaining reliable performance across diverse clinical scenarios. Key validation techniques include quantitative metrics to evaluate segmentation accuracy, qualitative visual analysis for assessing clinical utility, comparisons against baseline methods, and incorporation of feedback from clinicians. Ultimately, rigorous multifaceted validation combining metrics, visual analysis, comparisons, and clinician input enables a holistic understanding of model capabilities, limitations, and clinical relevance.

##### 4.1.1. Baselines

For the experiments on infection region segmentation, the proposed network is compared with traditional medical segmentation models such as U-Net [2], Attention U-Net [36], and Inf-Net [60]. This benchmarks the proposed method against widely used medical segmentation architectures, including a standard U-Net baseline and prior attention-based models.

##### 4.1.2. Evaluation Metrics

According to [61], we utilized several common evaluation metrics: the Dice Similarity Coefficient, specificity, and precision. We also included three object detection metrics, structure measure [62], mean absolute error, and enhanced-alignment measure [63].

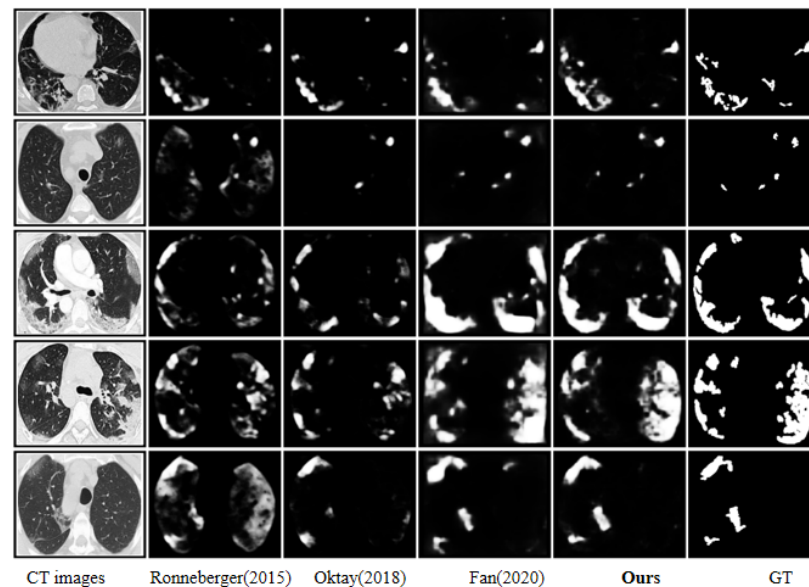
These metrics compare the prediction map with Sigmoid activation to the ground-truth  $G$  of object-level annotations. The metrics measure similarity/dissimilarity between the ground truth and predicted segmentation. We quantified segmentation performance using popular medical metrics along with additional detection-based metrics. These evaluate alignment, error, and topological similarity between predictions and ground truth annotations.

#### 4.2. CT Segmentation Dataset Results

In this experiment, we compiled a CT infection segmentation dataset, leveraging unlabeled images to augment the limited labeled training data. Our labeled data  $D_{Labeled}$  come from the CT segmentation dataset [13], which has 45 CT images randomly chosen for training, 5 for validation, and 50 for testing.

The infection segmentation results in Figure 4 show that our network markedly outperforms the baseline algorithms. Specifically, it yields segmentation maps that closely match the ground truth with significantly fewer incorrectly segmented tissue regions. Additionally, U-Net results in inadequate segmentation with many mis-segmented areas. Our success is attributable to a strategy of coarse-to-fine segmenting. First, a parallel partial decoder localizes infection regions roughly, then the segmentation is refined by multiple edge attention modules. This mimics radiologists' workflow of segmenting infections from CT images, and consequently achieves a promising performance. Furthermore, Figure 4

confirms the superiority of our semi-supervised approach over baselines. Our model produces segmentation with more accurate boundaries, while baselines result in relatively fuzzy boundaries, particularly in regions of subtle infection. The additional unlabeled data and multi-scale edge attention allow our model to capture fine-grained infection patterns and boundaries.



**Figure 4.** CT infection segmentation results for visual comparison: Ronneberger [2], Oktay [36], Fan [60].

As shown in Table 1, our proposed model has specificity 0.01 higher than U-Net and 0.03 than Attention U-Net. It also has a lower MAE than the baselines. This improvement is likely due to our model's implicit mechanism of reverse attention and unambiguous models of edge attention modules providing strong feature extraction abilities.

**Table 1.** Quantitative results of infection region on CT dataset.

Methods	Dice	Sen	Spec	$S_{\alpha}$	$E_{\phi}^{mean}$	MAE
U-Net [2]	0.574	0.561	0.949	0.706	0.744	0.105
Atte-U-Net [36]	0.582	0.508	0.971	0.684	0.762	0.096
Inf-Net [60]	0.647	0.709	0.910	0.737	0.827	0.103
Ours	0.695	0.721	0.939	0.778	0.854	0.083

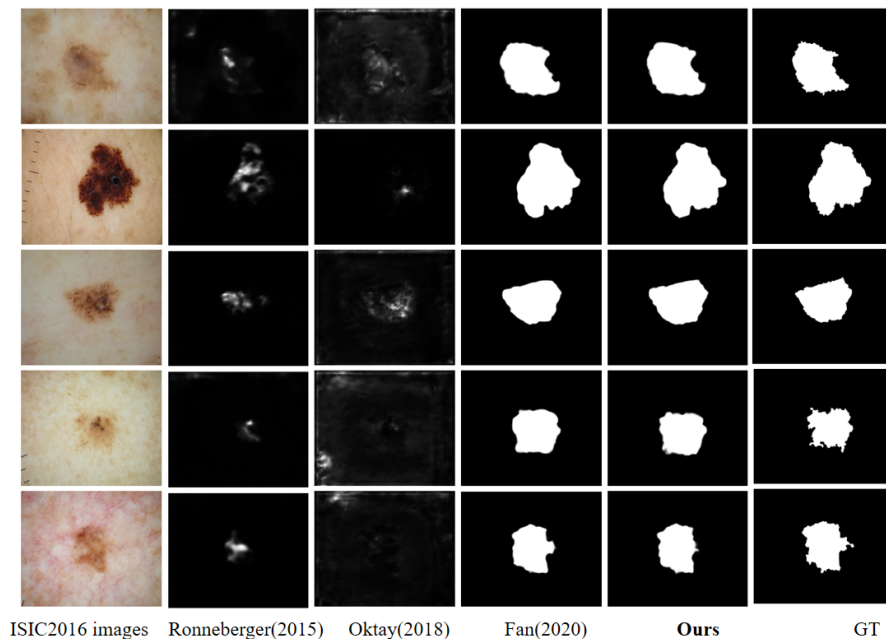
Overall, our model outperforms the baselines on most metrics. The additional unlabeled data and refined attention mechanisms allow our model to better distinguish infection regions.

#### 4.3. Skin Lesions Dataset Results

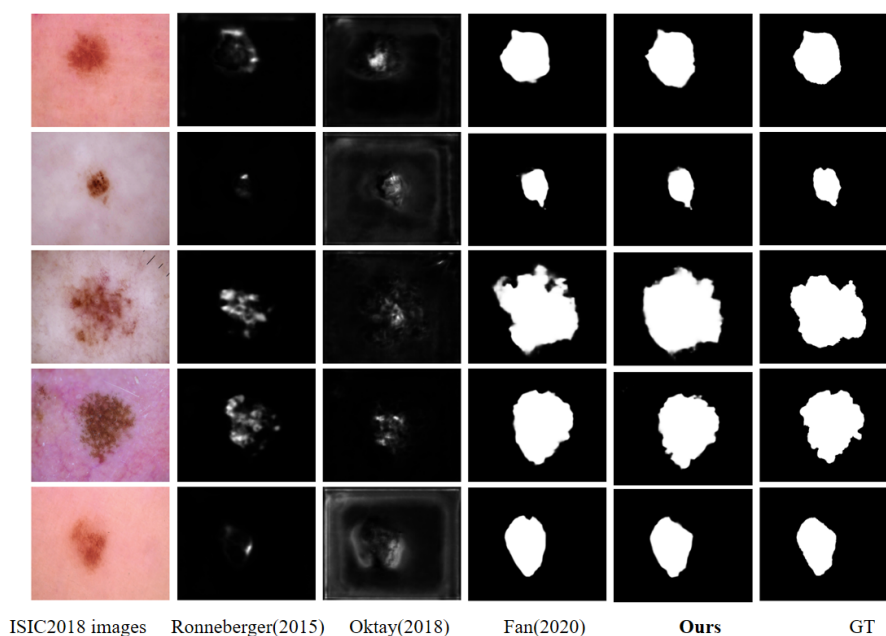
To validate the effectiveness of the algorithm across different medical image modalities, we utilized the ISIC2016 and ISIC2018 skin lesion image datasets. The 2016 dataset contained 900 images, which we divided into 600 for training, 200 for testing, and 100 for validation. The 2018 dataset had 5000 images, which we split into 3500 training images, 1000 testing images, and 500 validation images.

As shown in Figures 5 and 6, our network demonstrates superior lesion segmentation compared to the baseline algorithm. The segmentation maps from our network align closely with the ground truth data and contain significantly fewer erroneous segmented regions than the baselines. In contrast, U-Net and Attention U-Net yield very unsatisfactory

segmentation results on this task. While Inf-Net delivers a comparable performance to our network, our approach still produces slightly better delineation of edge details in the infection regions. Overall, these results highlight the effectiveness of our network for accurate lesion segmentation in the ISIC2016 dataset.



**Figure 5.** ISIC2016 infection segmentation results for visual comparison: Ronneberger [2], Oktay [36], Fan [60].



**Figure 6.** ISIC2018 infection segmentation results for visual comparison: Ronneberger [2], Oktay [36], Fan [60].

The results in Tables 2 and 3 demonstrate that our proposed model achieves the best performance across all evaluation metrics compared to the baseline algorithms. Our model also outperforms Inf-Net on multiple indicators, which can be attributed to the addition of the transformer module in our architecture.



**Table 2.** Quantitative results of infection region on ISIC2016 dataset.

Methods	Dice	Sen	Spec	$S_\alpha$	$E_\phi^{mean}$	MAE
U-Net [2]	0.155	0.107	0.908	0.408	0.351	0.275
Atte-U-Net [36]	0.058	0.064	0.666	0.368	0.294	0.309
Inf-Net [60]	0.899	0.900	0.970	0.880	0.916	0.054
Ours	0.905	0.905	0.972	0.884	0.918	0.052

**Table 3.** Quantitative results of infection region on ISIC2018 dataset.

Methods	Dice	Sen	Spec	$S_\alpha$	$E_\phi^{mean}$	MAE
U-Net [2]	0.125	0.093	0.879	0.407	0.343	0.251
Atte-U-Net [36]	0.086	0.084	0.826	0.395	0.319	0.276
Inf-Net [60]	0.905	0.946	0.935	0.899	0.924	0.054
Ours	0.915	0.944	0.942	0.906	0.934	0.048

In summary, our model consistently surpasses the baseline algorithms on most metrics for all the datasets. The use of additional unlabeled data and refined attention mechanisms enables our model to more accurately identify infection regions. The quantitative results validate the superior infection segmentation capability of our proposed model across different imaging modalities.

#### 4.4. Ablation Study

In this section, we employ several experiments to confirm the performance of every key element of our framework, including the PPD, RA, EA, and TRANS modules; in Table 4, we set Backbone as Bb. The experiments validated the effectiveness of the PPD and other core components of the proposed model and Table 4 shows the results.

**Table 4.** Ablation studies of proposed model.

Methods	Dice	Sen	Spec	$S_\alpha$	$E_\phi^{mean}$	MAE
(NO.1) Bb	0.442	0.570	0.825	0.651	0.569	0.207
(NO.2) Bb + EA	0.541	0.665	0.807	0.673	0.659	0.205
(NO.3) Bb + PPD	0.669	0.744	0.880	0.720	0.810	0.125
(NO.4) Bb + RA	0.625	0.826	0.809	0.668	0.736	0.177
(NO.5) Bb + EA + RA	0.672	0.754	0.882	0.738	0.804	0.122
(NO.6) Bb + PPD + RA	0.655	0.690	0.927	0.761	0.812	0.098
(NO.7) Bb + PPD + RA + EA	0.647	0.790	0.910	0.737	0.827	0.103
(NO.8) Bb + PPD + RA + EA+TRANS	0.695	0.721	0.939	0.778	0.854	0.083

**Effectiveness of PPD.** To determine the contribution of the PPD, we compared backbone only No. 1 and backbone+ PPD (No. 3) in Table 4. The PPD clearly improves performance, demonstrating its necessity.

**Effectiveness of RA.** Comparing backbone only (No. 1) and backbone + RA (No. 4) shows RA boosts metrics like Dice, sensitivity, and MAE. This indicates RA helps accurately distinguish infected regions.

**Effectiveness of PPD and RA.** No. 6 combines PPD and RA. As Table 4 shows, No. 6 outperforms other settings on most metrics. This shows PPD and RA are central components underlying our strong performance.

**Effectiveness of EA.** Adding EA (No. 2 vs. No. 1, No. 5 vs. No. 4, No. 7 vs. No. 6) consistently improves segmentation, demonstrating its contribution.

**Effectiveness of TRANS.** Finally, adding the TRANS module (No. 8 vs. No. 7) boosts our model across indicators, validating its role in improving segmentation.

In summary, our ablation studies demonstrate the necessity and complementary value of the PPD, RA, EA, and TRANS components proposed. Together, they enable our model's cutting-edge infection segmentation performance.

#### 4.5. Time Costs

We validated the effectiveness of our experiments using an Nvidia GTX 2080Ti GPU and an Intel Xeon Silver 4210 CPU (Intel, Santa Clara, CA, USA). Due to setting the size of all three datasets to  $512 \times 512$ , the average training and testing time of different models and segmentation methods on the three datasets remained basically consistent, as shown in Table 5. The time is reported in seconds per image.

**Table 5.** Average training and test time comparison.

Methods	CT/ISIC2016/ISIC2018	
	Train	Test
U-Net [2]	0.216	8.307
Atte-U-Net [36]	0.278	15.875
Inf-Net [60]	8.17	45.66
Ours	8.86	65.75

Although our method does not match the computational efficiency of Inf-Net for training and testing times, we achieved a better image segmentation performance at the cost of increased computation time. This trade-off between accuracy and speed is acceptable and expected, as our approach prioritizes segmentation quality over pure efficiency on the hardware used. The validation results demonstrate the effectiveness of our method, albeit with some sacrifice in speed compared to highly optimized networks like Inf-Net.

#### 4.6. Potential Limitations

Image quality variability can impact the performance of segmentation models. Models trained on high-quality data, especially in the selected public dataset in our work may not generalize to real-world clinical images of varying quality.

Biases in the training data arising from limited or labeling errors can propagate biases in model predictions, leading to disparities across patient populations. Careful curation for representativeness and techniques like augmentation and adversarial training can help mitigate such biases.

Models trained on data from specific clinical settings may not generalize well to new clinical settings with different data distributions. Domain adaptation and transfer learning approaches can potentially improve generalization across domains.

In summary, key limitations around image variability, data bias, and generalization across clinical settings need to be addressed through robust algorithms, careful data curation to ensure representativeness, evaluation on diverse data, and techniques like augmentation, adversarial training, and transfer learning. Addressing these challenges will enable the reliable deployment of automatic segmentation models in clinical practice.

## 5. Conclusions

Transformers are structured with inherent powerful self-attention mechanisms that enable them to model long-range dependencies effectively. This work investigates the application of transformers for general medical image segmentation tasks. To fully utilize the capabilities of transformers, we encoded robust global context by training image features as sequences. Additionally, we incorporated low-level CNN features through a hybrid design. Our approach provides an alternative to the prevailing FCN-based methods for segmentation tasks across different medical image modalities. Our transformer-based model outperforms various popular segmentation methods, demonstrating that transformers can move beyond their success in NLP to effectively capture visual relationships in medical

images. The promising results highlight the potential of transformers in medical imaging applications. Rather than relying on the local self-attention mechanisms in CNNs, pure transformer encoders can better model global dependencies, which is advantageous for medical image segmentation. Overall, our model advances the field of medical image segmentation by demonstrating the applicability and advantages of transformers for encoding domain knowledge in this task.

The key future directions involve developing unified AI-based system for integrated diagnosis, interactive segmentation to enable human–AI collaboration, multimodal fusion for leveraging complementary information from different imaging modalities, multi-task learning for joint optimization of multiple clinical tasks, clinical decision support systems, and the seamless integration of these AI solutions into clinical workflows and mobile health applications.

**Author Contributions:** Conceptualization, J.Z.; methodology, J.Z. and X.Z.; software, F.L.; validation, X.Z.; resources, H.W.; data curation, X.Z.; writing—original draft preparation, J.Z.; funding acquisition, X.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by NSFC Grant No. 61702409.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. This data can be found here: <https://github.com/UCSD-AI4H/COVID-CT>, <https://challenge.isic-archive.com/data/#2016> and <https://challenge.isic-archive.com/data/#2018> (accessed on 29 January 2024).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

- Shen, D.; Wu, G.; Suk, H.-I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **2017**, *19*, 221–248. [CrossRef] [PubMed]
- Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional net works for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
- Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–13 June 2015*; pp. 3431–3440.
- Gordaliza, P.M.; Mu, A.; Abella, M.; Desco, M.; Sharpe, S.; Vaquero, J.J. Unsupervised CT lung image segmentation of a mycobacterium tuberculosis infection model. *Sci. Rep.* **2018**, *8*, 9802. [CrossRef] [PubMed]
- Jin, D.; Xu, Z.; Tang, Y.; Harrison, A.P.; Mollura, D.J. CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 732–740.
- Ye, Z.; Zhang, Y.; Wang, Y.; Huang, Z.; Song, B. Chest CT manifestations of new coronavirus disease 2019 (COVID-19): A pictorial review. *Eur. Radiol.* **2020**, *30*, 4381–4389. [CrossRef] [PubMed]
- Jiang, J.; Hu, Y.C.; Liu, C.J.; Halpenny, D.; Hellmann, M.D.; Deasy, J.O.; Mageras, G.; Veeraraghavan, H. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images. *IEEE Trans. Med. Imaging* **2018**, *38*, 134–144. [CrossRef]
- Yu, L.; Cheng, J.Z.; Dou, Q.; Yang, X.; Chen, H.; Qin, J.; Heng, P.A. Automatic 3D cardiovascular MR segmentation with densely connected volumetric convnets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 287–295.
- Wu, Y.H.; Gao, S.H.; Mei, J.; Xu, J.; Fan, D.P.; Zhang, R.G.; Cheng, M.M. JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 3113–3126. [CrossRef]
- Shin, H.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* **2016**, *35*, 1285–1298. [CrossRef] [PubMed]
- Cheplygina, V.; de Bruijne, M.; Pluim, J.P. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* **2019**, *54*, 280–296. [CrossRef] [PubMed]
- Ng, M.Y.; Lee, E.Y.; Yang, J.; Yang, F.; Li, X.; Wang, H.; Lui, M.M.S.; Lo, C.S.Y.; Leung, B.; Khong, P.L.; et al. Imaging profile of the COVID-19 infection: Radiologic findings and literature review. *Radiol. Cardiothorac. Imaging* **2020**, *2*, e200034. [CrossRef]

13. Yang, X.; He, X.; Zhao, J.; Zhang, Y.; Zhang, S.; Xie, P. COVID-CT-dataset: A CT scan dataset about COVID-19. *arXiv* **2020**, arXiv:2003.13865.
14. Cohen, J.P.; Morrison, P.; Dao, L. COVID-19 image data collection. *arXiv* **2020**, arXiv:2003.11597.
15. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.-W.; Heng, P.-A. H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674. [[CrossRef](#)]
16. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018; pp. 801–818.
17. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
18. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
19. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Edesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [[CrossRef](#)]
20. Li, Q.; Song, H.; Zhang, W.; Fan, J.; Ai, D.; Lin, Y.; Yang, J. CC-DenseUNet: Densely connected U-Net with criss-cross attention for liver and tumor segmentation in CT volumes. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 9–12 December 2021.
21. Isensee, F.; Jäger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. Automated design of deep learning methods for biomedical image segmentation. *arXiv* **2019**, arXiv:1904.08128.
22. Rehman, M.U.; Ryu, J.; Nizami, I.F.; Chong, K.T. RAAGR2-Net: A brain tumor segmentation network using parallel processing of multiple spatial frames. *Comput. Biol. Med.* **2023**, *152*, 106426. [[CrossRef](#)] [[PubMed](#)]
23. Soni, A.; Koner, R.; Villuri, V.G.K. M-unet: Modified u-net segmentation framework with satellite imagery. In Proceedings of the Global AI Congress 2019; Springer: Singapore, 2020; pp. 47–59.
24. Dash, M.; Londhe, N.D.; Ghosh, S.; Semwal, A.; Sonawane, R.S. PsLSNet: Automated psoriasis skin lesion segmentation using modified U-Net-based fully convolutional network. *Biomed. Signal Process. Control* **2019**, *52*, 226–237. [[CrossRef](#)]
25. Chattopadhyay, S.; Basak, H. Multi-scale attention u-net (msaunet): A modified u-net architecture for scene segmentation. *arXiv* **2020**, arXiv:2009.06911.
26. Ryu, J.; Rehman, M.U.; Nizami, I.F.; Chong, K.T. SegR-Net: A deep learning framework with multi-scale feature fusion for robust retinal vessel segmentation. *Comput. Biol. Med.* **2023**, *163*, 107132. [[CrossRef](#)] [[PubMed](#)]
27. Hasan, S.M.K.; Linte, C.A. U-NetPlus: A modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 7205–7211.
28. Keetha, N.V.; Annavarapu, C.S.R. U-Det: A modified U-Net architecture with bidirectional feature network for lung nodule segmentation. *arXiv* **2020**, arXiv:2003.09293.
29. Pun, N.S.; Agarwal, S. Modality specific U-Net variants for biomedical image segmentation: A survey. *Artif. Intell. Rev.* **2022**, *55*, 5845–5889. [[CrossRef](#)] [[PubMed](#)]
30. You, X.; Peng, Q.; Yuan, Y.; Cheung, Y.-M.; Lei, J. Segmentation of retinal blood vessels using the radial projection and semi-supervised approach. *Pattern Recognit.* **2011**, *44*, 2314–2324. [[CrossRef](#)]
31. Portela, N.M.; Cavalcanti, G.D.; Ren, T.I. Semisupervised clustering for MR brain image segmentation. *Expert Syst. Appl.* **2014**, *41*, 1492–1497. [[CrossRef](#)]
32. Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; Heng, P.-A. Uncertainty-aware self-ensembling model for semisupervised 3D left atrium segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 605–613.
33. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the NeurIPS, Long Beach, CA, USA, 4–9 December 2017; pp. 1195–1204.
34. Li, S.; Zhang, C.; He, X. Shape-aware Semisupervised 3D Semantic Segmentation for Medical Images. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 552–561.
35. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
36. Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; Rueckert, D. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **2019**, *53*, 197–207. [[CrossRef](#)] [[PubMed](#)]
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
38. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
39. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *arXiv* **2021**, arXiv:2101.01169. [[CrossRef](#)]



40. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
41. Liu, N.; Zhang, N.; Wan, K.; Shao, L.; Han, J. Visual saliency transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
42. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
43. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4055–4064.
44. Child, R.; Gray, S.; Radford, A.; Sutskever, I. Generating long sequences with sparse transformers. *arXiv* **2019**, arXiv:1904.10509.
45. Jiao, J.; Tang, Y.M.; Lin, K.Y.; Gao, Y.; Ma, J.; Wang, Y.; Zheng, W.S. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE Trans. Multimed.* **2023**, *25*, 8906–8919. [[CrossRef](#)]
46. Zhao, J.-X.; Liu, J.-J.; Fan, D.-P.; Cao, Y.; Yang, J.; Cheng, M.-M. EGNet: Edge guidance network for salient object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8779–8788.
47. Wu, Z.; Su, L.; Huang, Q. Stacked cross refinement network for edge-aware salient object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7264–7273.
48. Zhang, Z.; Fu, H.; Dai, H.; Shen, J.; Pang, Y.; Shao, L. ET-Net: A generic edge-attention guidance network for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 442–450.
49. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. CE-Net: Context Encoder Network for 2D Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2281–2292. [[CrossRef](#)] [[PubMed](#)]
50. Zhang, S.; Fu, H.; Yan, Y.; Zhang, Y.; Wu, Q.; Yang, M.; Tan, M.; Xu, Y. Attention Guided Network for Retinal Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 797–805.
51. Chakraborty, S.; Kalyani, M. An overview of biomedical image analysis from the deep learning perspective. In *Research Anthology on Improving Medical Imaging Techniques for Analysis and Intervention*; IGI Global: Harrisburg, PA, USA, 2023; pp. 43–59.
52. Gao, S.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P.H. Res2Net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [[CrossRef](#)]
53. Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; Yan, S. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1568–1576.
54. Chen, S.; Tan, X.; Wang, B.; Hu, X. Reverse attention for salient object detection. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 234–250.
55. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. BASNet: Boundary-aware salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7479–7489.
56. Wei, J.; Wang, S.; Huang, Q. F3Net: Fusion, feedback and focus for salient object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
57. Chen, L.; Qu, H.; Zhao, J.; Chen, B.; Principe, J.C. Efficient and robust deep learning with correntropy-induced loss function. *Neural Comput. Appl.* **2016**, *27*, 1019–1031. [[CrossRef](#)]
58. Liangjun, C.; Honeine, P.; Hua, Q.; Jihong, Z.; Xia, S. Correntropybased robust multilayer extreme learning machines. *Pattern Recognit.* **2018**, *84*, 357–370. [[CrossRef](#)]
59. Tschandl, P.; Rosendahl, C.; Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **2018**, *5*, 180161. [[CrossRef](#)] [[PubMed](#)]
60. Fan, D.-P.; Zhou, T.; Ji, G.-P.; Zhou, Y.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images. *IEEE Trans. Med. Imaging* **2020**, *39*, 2626–2637. [[CrossRef](#)]
61. Shan, F.; Gao, Y.; Wang, J.; Shi, W.; Shi, N.; Han, M.; Xue, Z.; Shen, D.; Shi, Y. Lung infection quantification of COVID-19 in CT images with deep learning. *arXiv* **2020**, arXiv:2003.04655.
62. Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; Borji, A. Structuremeasure: A new way to evaluate foreground maps. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4548–4557.
63. Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; Borji, A. Enhanced-alignment measure for binary foreground map evaluation. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 698–704.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.