# Multi-dimension Transformer with Attention-based Filtering for Medical Image Segmentation

Wentao Wang<sup>a</sup>, Xi Xiao<sup>a</sup>, Mingjie Liu<sup>b,\*</sup>, Qing Tian<sup>a</sup>, Xuanyao Huang<sup>b</sup>, Qizhen Lan<sup>a</sup>, Swalpa Kumar Roy<sup>c</sup> and Tianyang Wang<sup>a</sup>

<sup>a</sup>University of Alabama at Birmingham
<sup>b</sup>Chongqing University of Posts and Telecommunications
<sup>c</sup>Alipurduar Government Engineering and Management College

The accurate segmentation of medical images is crucial for diagnosing and treating diseases. Recent studies demonstrate that vision transformer-based methods have significantly improved performance in medical image segmentation, primarily due to their superior ability to establish global relationships among features and adaptability to various inputs. However, these methods struggle with the low signal-to-noise ratio inherent to medical images. Additionally, the effective utilization of channel and spatial information, which are essential for medical image segmentation, is limited by the representation capacity of self-attention. To address these challenges, we propose a multi-dimension transformer with attentionbased filtering (MDT-AF), which redesigns the patch embedding and self-attention mechanism for medical image segmentation. MDT-AF incorporates an attention-based feature filtering mechanism into the patch embedding blocks and employs a coarse-to-fine process to mitigate the impact of low signal-to-noise ratio. To better capture complex structures in medical images, MDT-AF extends the selfattention mechanism to incorporate spatial and channel dimensions, enriching feature representation. Moreover, we introduce an interaction mechanism to improve the feature aggregation between spatial and channel dimensions. Experimental results on three public medical image segmentation benchmarks show that MDT-AF achieves state-of-the-art (SOTA) performance.

### 1 Introduction

Segmentation plays a pivotal role in medical imaging analysis by identifying and outlining regions of interest within medical images [35, 22]. However, relying on experts for diagnosis is time-consuming and prone to observer bias by clinical experience [11, 12]. Therefore, the automated segmentation technique is an excellent assistance tool. Convolutional neural networks (CNNs) have become the standard in computer vision and are widely adopted for medical image segmentation [47]. UNet [32] and its variants are widely used in medical image segmentation, such as Attention UNet [29], UNet++[50], and UNet 3+[17]. However, due to the limitations of the convolution operation, CNN-based methods struggle to model long-range dependencies effectively, which can result in limited performance in medical image segmentation [48]. Firstly, medical images often necessitate the modeling of global information to achieve reliable segmentation, as the shape and size of the region of interest

can vary significantly. Secondly, CNN-based methods lack the flexibility to accommodate inputs with diverse characteristics, thereby limiting their ability to adapt to varying content within input images. Recently, vision transformers [13] have achieved remarkable success in various vision tasks [27, 46, 37], demonstrating competitive performance compared to other CNN-based methods. Given their capability for long-range information interaction and dynamic feature encoding [26], researchers have explored the use of transformers in medical image segmentation [35]. TransUNet [4] represents the pioneering effort to integrate Transformers into medical image segmentation, utilizing a hybrid structure of CNNs and Transformers. Subsequently, a pure Transformer encoder-decoder architecture named SwinUNet [3] was introduced, demonstrating strong performance in this domain.

While Transformers perform well in modeling long-range dependencies, they still face drawbacks and limitations when applied to medical images. i) these methods struggle to handle the inherently low signal-to-noise ratio of medical images [14, 21], which can significantly hinder feature learning and discrimination. ii) their multidimension representation capability is still limited, particularly in capturing channel and spatial information, which are crucial for medical image segmentation. Therefore, there remains significant room for improvement in Transformer-based methods.

To address the above limitations, we propose MDT-AF, a novel transformer variant designed for robust and precise medical image segmentation. MDT-AF comprises two primary components: Patch embedding with attention-based filtering and self-attention that extends across spatial and channel dimensions. Specifically, to address the first problem, we redesigned the patch embedding mechanism of the vision transformer by incorporating an attention-based filtering mechanism parallel to the patch embedding module. This filtering mechanism aims to refine coarse features and reduce noise, thereby enhancing the model's ability to focus on relevant signals and adapt dynamically to varying noise levels across different image regions. As a result, this mechanism improves the quality of extracted features and enhances the model's accuracy in identifying and delineating critical areas in medical images. To tackle the second challenge, we introduced multi-dimension transformer blocks, which expand self-attention to spatial and channel dimensions and perform feature interaction and aggregation within blocks. Spatial self-attention enriches the spatial expression of each feature map, while channel selfattention facilitates global information exchange between features.

 $<sup>^{\</sup>ast}$  Corresponding Author. Email: liumj@cqupt.edu.cn

As a result, this mechanism improves the richness of feature representation and enhances the model's ability to model medical images with various shapes and scenes effectively.

In contrast to existing approaches, our method offers the following three main contributions:

- 1) We propose a novel transformer variant for stable and precise medical image segmentation, which can generate rich feature representation while minimizing redundant information.
- 2) We introduce an attention-based filtering mechanism. This mechanism filters the coarse features and noise acquired from patch embedding, which can enable the construction of more refined feature representations. Moreover, the self-attention mechanism is expanded to spatial and channel dimensions and incorporates feature interaction and aggregation within blocks to establish comprehensive feature representation.
- 3) Experimental evaluations conducted on three public medical image segmentation benchmarks across Lung, Skin, and Polyp demonstrate the effectiveness and superiority of our method compared to the state-of-the-art methods.

#### 2 Related Works

#### 2.1 CNN-based segmentation methods

CNNs have been primarily used for medical image segmentation in recent years due to their powerful feature representation capabilities [34]. U-shaped architectures play a significant role and are widely used in medical image segmentation due to their encoderdecoder architecture [25]. UNet, an encoder-decoder architecture featuring multiscale skip connections, has demonstrated state-of-theart performance across various medical image segmentation tasks. Subsequently, several variants of UNet have been developed, including UNet++, Attention UNet, R2U-Net [1], and UNet3+. UNet++ utilized the advantage of dense skip connections to link its nested encoder-decoder subnetworks. The Attention UNet introduced an attention mechanism to refine the output features of the encoder before merging these features with the corresponding decoder features at each resolution level. R2U-Net is an extension of standard U-Net, which combines recurrent neural networks and residual skip connections. UNet3+ extends the concept of full-scale skip connections by incorporating intra-connections between the decoder blocks. Inspired by these architectures, specific methods have been tailored for particular tasks, such as polyp segmentation [45, 30], skin lesion segmentation [36, 33], etc. Although these methods have improved the abilities of context modeling to some extent, CNN's limited receptive fields strand their performance.

# 2.2 Transformer-based segmentation methods

Recently, vision transformer-based methods have achieved state-of-the-art performance in various vision tasks. Given their success, numerous studies have explored the application of transformers in medical image segmentation. Compared to CNNs, transformers can capture long-range dependencies through sequence modeling and multi-head self-attention, leading to improved segmentation performance. TransUNet [4] integrates UNet with Transformers to learn both local and global pixel relations. TransFuse [49] combines Transformers and CNNs in parallel to capture global dependencies and low-level spatial details in a more efficient and shallower manner. MedT [38] leverages both local and global information by employing two branches: A gated axial Transformer to explore global information, and a CNN to learn local information. SwinUNet [3] in-

troduces a pure U-shaped transformer architecture using Swin transformer blocks, where all convolutional layers in the U-Net are replaced by Swin transformer blocks. DS-TransUnet [24] further extends SwinUNet by introducing a fusion module designed to model long-range dependencies between features of different scales. MISS-former [18] implements an encoder-decoder architecture using enhanced transformer blocks and introduces a convolutional layer within the Transformer to enhance its ability to capture local information. Hiformer [16] proposes a combination of a Swin Transformer module and a CNN-based encoder to generate two multi-scale feature representations, which are integrated via a Double-Level Fusion module.

These transformer-based methods had several shortcomings despite their encouraging results. 1) They overlook important aspects of medical image segmentation, such as the representation and aggregation of features across spatial and channel dimensions. 2) The poor signal-to-noise ratio in medical images makes these methods challenging to use and would negatively impact segmentation performance. In this study, we redesign the patch embedding and selfattention mechanism, where the modified patch embedding process provides an attention-based filtering mechanism. By generating a coarse-to-fine process, a finer representation of the features is obtained. Furthermore, we expand self-attention across channel and spatial dimensions, enabling comprehensive representation and aggregation of cross-dimension information within blocks. Finally, a novel transformer variant, termed MDT-AF, is proposed with superior performance, which can provide rich feature representation and redundant information filtering for medical image segmentation.

#### 3 Methods

In this section, we first present the overview of our proposed MDT-AF. Then, the details of the proposed patch embedding with attention-based filtering and multi-dimension transformer block will be given.

#### 3.1 Overall Architecture

The overall architecture of the proposed MDT-AF mainly comprises three modules: Attention-based patch embedding blocks, multidimension transformer blocks, and MLP layers, as illustrated in Fig 1. Specifically, given an input image, it is initially processed through the patch embedding block. This block comprises two parallel branches: One is the overlap patch embedding, utilized for extracting initial features, while another parallel branch is the attentionbased filtering mechanism, which generates attention weights to filter the output of the first branch. Subsequently, we obtain more refined features after this filtering process, which are further passed into the transformer block. Within the transformer block, self-attention is extended across spatial and channel dimensions, enabling the model to establish and aggregate features across multi-dimension. The input image undergoes processing through four levels of patch embedding blocks and transformer blocks. The processed features are fed into the MLP decoders, which integrate features from four levels to predict semantic segmentation masks.

## 3.2 Patch Embedding with Attention-based Filtering

Medical images generally confront challenges related to low signalto-noise ratio, which directly result in the blurring of critical features such as boundaries, thus complicating the pixel-level classification. Patch Embedding is a crucial component of transformer-based

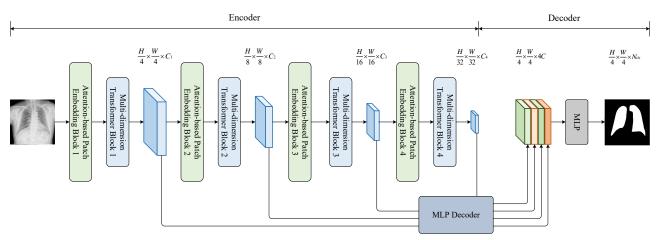


Figure 1. The proposed MDT-AF framework comprises three primary modules: 1) Attention-based patch embedding blocks that generate patch tokens while concurrently producing attention weights. These weights are instrumental in filtering out coarse features and noise. 2) Multi-dimension transformer blocks extend self-attention across spatial and channel-wise dimensions to build and aggregate a comprehensive feature representation. 3) MLP decoders fuse these multi-level features to accurately predict the semantic segmentation mask. Where  $C_1$  is set to 64,  $C_2$  is set to 128,  $C_3$  is set to 320, and  $C_4$  is set to 512.

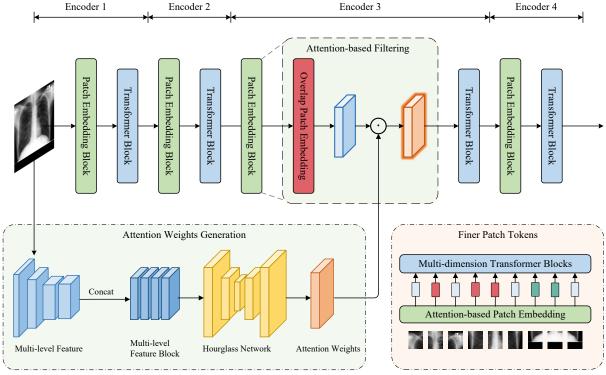


Figure 2. The proposed Patch Embedding with Attention-based Filtering consists of two parallel branches: 1) The Overlap Patch Embedding branch processes an input feature map to extract coarse features. 2) Simultaneously, a parallel branch generates corresponding attention weights. These weights are applied to the coarse features, filtering out noise and refining the feature representation from coarse to fine. Notably, this approach is consistently employed in

the patch embedding block of each encoder stage to generate patch tokens. methods, which plays a vital role in preprocessing input images and generating embeddings. These embeddings are then processed into tokens utilized by subsequent transformer blocks for feature extraction and classification tasks. Improving Patch Embedding is essential for several reasons: i) It will provide a more robust initial feature representation, crucial for handling noisy inputs in medical images; ii) It will preserve more detailed spatial information, which is vital for identifying subtle anatomical structures in blurred images; iii) It will improve the model's efficiency by ensuring that higher-quality embeddings lead to more effective subsequent processing; iv) Optimized Patch Embeddings will be able to enhance the robustness of

the model against variations in image quality, a common issue in different clinical environments.

Motivated by [44, 43, 42, 40], we propose a patch embedding with an attention-based filtering mechanism. Specifically, we incorporate an attention-based filtering branch parallel to the patch embedding module to filter the coarse features and noise, as is shown in Fig 2. This mechanism aims to refine the feature representation by selectively enhancing relevant features and suppressing noise, thereby improving the segmentation accuracy in medical images. Next, we will describe the details below.

1) Overlap Patch Embedding: The initial design of patch embed-

ding in Vision Transformers (ViT) [13] involved combining nonoverlapping images or feature patches. However, this method can not effectively preserve local continuity around the patches. In this study, we employ overlapping patch embedding to better capture features [41]. In our experiments, we applied two configurations to facilitate overlapping patch embedding: K = 7, S = 4, P = 3, and K = 3, S = 2, P = 1. The coarse features obtained from overlap patch embedding are recorded as  $F_1$ .

2) Attention-based Filtering: Given an input X, we first use a  $1 \times 1$  convolution operation to reshape the channels of the input to 40, noted as  $X_r$ . Then, a convolution with  $3 \times 3$  kernel size and 40 groups is applied to the  $X_r$ , noted as  $X_{gwc}$  Next, the first 8 channels of the grouped features are further processed by a  $3 \times 3$  convolution with a dilation of 1 (Level 1). The next 16 channels are processed by a  $3 \times 3$  convolution with a dilation of 2 (Level 2). The last 16 channels are processed by a  $3 \times 3$  convolution with a dilation of 3 (Level 3). The process is defined as:

$$V_{l1} = \text{DilatedConv}_{l1}(X_{\text{gwc}}[:,:8]), \tag{1}$$

$$V_{l2} = \text{DilatedConv}_{l2}(X_{\text{gwc}}[:, 8:24]), \tag{2}$$

$$V_{l3} = \text{DilatedConv}_{l3}(X_{\text{gwc}}[:, 24:40])$$
 (3)

where  $DilatedConv_{li}$  corresponds to the dilated convolution operation for level i with different dilation rates and group settings. The outputs of these three levels are concatenated along the channel dimension, resulting in a combined feature map that includes multiscale feature information, which can be formulated as:

$$V_{\text{mpmv}} = \text{Concat}(V_{l1}, V_{l2}, V_{l3}) \tag{4}$$

Subsequently, we apply two convolutions and a 2D hourglass network [15] to regularize  $V_{\rm mpmv}$ . Following this, another convolutional layer compresses the channels to 1, deriving the attention weights denoted as A. With the obtained attention weights, we utilize them to eliminate redundant information in the coarse feature. The output feature  $X_{out}$  at channel i is computed as:

$$X_{out}(i) = A \odot F_1, \tag{5}$$

where  $\odot$  indicates the element-wise product, and the attention weights A are applied to all channels of the coarse feature.

#### 3.3 Multi-dimension Transformer Block

A rich representation of channel and spatial dimensions is essential in medical image analysis. Self-attention can not model spatial and channel information well, as it lacks spatial expression of each feature map and ignores the relationship among channels. Inspired by [23, 7, 8, 6], we introduce the Multi-dimension Transformer block during the feature extraction, mainly consisting of efficient self-attention attention, spatial dimension self-attention, and channel dimension self-attention, as shown in Fig 3. Specifically, spatial and channel dimension self-attentions are designed to model spatial and channel information. Considering that self-attention primarily captures global features, we integrated a convolutional branch parallel to the self-attention module to introduce locality into the Transformer architecture. And an interaction mechanism is introduced to adaptively re-weight features from the spatial or channel dimensions, allowing for better fusion of features from the two branches.

1) Efficient Self-attention (ESA): Self-attention [39] calculates the attention matrix using queries(Q), keys(K), and values(V) to

be a sequence of  $H \times W$  feature vectors with dimensions D, where H and W is the height and width of the image, respectively. The formula is as follows:

$$V' = Softmax \left(\frac{Q^T K}{\sqrt{D}}\right) V^T. \tag{6}$$

the operation has  $O(n^2)$  complexity. While for the efficient self-attention [41], a modified equation reduces computation burden:

$$\hat{K} = \text{Reshape}(\frac{N}{R}, C \cdot R)(K)$$

$$K = \text{Linear}(C \cdot R, C)(\hat{K}),$$
(7)

where K is the sequence to be reshaped, Reshape  $\left(\frac{N}{R}, C \cdot R\right)(K)$  reshapes K to  $\frac{N}{R} \times (C \cdot R)$ , and Linear  $(C_{in}, C_{out})(\cdot)$  is a linear layer transforming an input tensor of dimension  $C_{in}$  to an output tensor of dimension  $C_{out}$ . Consequently, the new K has dimensions  $\frac{N}{R} \times C$ , resulting in reduced complexity of  $O\left(\frac{N^2}{R}\right)$ . In this case, given an input  $X_{\rm in}$ , we get the output from Efficient Self-attention (ESA) as follows:

$$Y_E = ESA(X_{\rm in}) + X_{\rm in} \tag{8}$$

where  $Y_E$  represents the output from the ESA operation.

2) Spatial Self-attention (SSA): Spatial-wise expression is a benefit for accurate boundary delineation and shape representation. To enrich the spatial expression of each feature map, we expand self-attention to the spatial dimension, as shown in Fig 3. Given a feature map  $f \in \mathbb{R}^{H \times W \times C}$ , we first derive the query (Q), key (K), and value (V) through linear projection:

$$Q = fW_{Q}, \quad K = fW_{K}, \quad V = fW_{V} \tag{9}$$

where  $W_Q, W_K, W_V \in \mathbb{R}^{C \times C}$  are the projection matrices without bias terms. We then partition Q, K, and V into non-overlapping windows and flatten each window, noted as  $Q_{\rm sp}, K_{\rm sp}$ , and  $V_{\rm sp}$ , respectively. Next, we distribute them into h attention heads:  $Q_{\rm sp} = [Q_{\rm sp}^1, \ldots, Q_{\rm sp}^h], K_{\rm sp} = [K_{\rm sp}^1, \ldots, K_{\rm sp}^h],$  and  $V_{\rm sp} = [V_{\rm sp}^1, \ldots, V_{\rm sp}^h]$ . Each head's output is then computed as follows:

$$Y_{\text{sp}}^{i} = Softmax \left( \frac{Q_{\text{sp}}^{i} (K_{\text{sp}}^{i})^{T}}{\sqrt{d}} + D \right) \cdot V_{\text{sp}}^{i}$$
 (10)

where D signifies the relative position encoding. By reshaping and concatenating the outputs  $Y_{\rm sp}^i$  from all heads, we acquire the feature map  $Y_{\rm sp}$ :

$$Y_{Sp} = \operatorname{concat}(Y_{\operatorname{sn}}^1, \dots, Y_{\operatorname{sn}}^h) W_{\operatorname{merge}}$$
 (11)

where  $W_{\rm merge}$  is the linear projection to fuse all features. To better encode spatial location information and inject inductive bias, we adopt depthwise convolution as a parral branch to extract local spatial features. Specifically, given a feature map f, we adopt the convolutional block to model local spatial features, which are shown as follows:

$$Y_{local} = Dw\text{-}Conv(f) \tag{12}$$

 $Y_{local}$  is the local features, which contain 2D spatial location information, making it possible to encode position information. Given two features  $X_1$  and  $X_2 \in \mathbb{R}^{H \times W \times C}$ , we apply two interaction mechanisms to better fuse and aggregate these features:

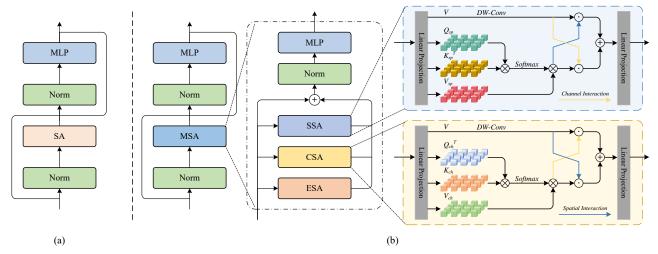


Figure 3. The proposed Multi-dimension Transformer Block (b) expands self-attention to spatial and channel dimensions. Unlike transformer block (a), it incorporates feature interaction and aggregation within blocks, with spatial self-attention capturing contextual information across image positions and channel self-attention analyzing feature channel relationships to highlight significant features. And a convolution branch is parallel with self-attention to add locality. "SA" denotes self-attention, "ESA" signifies efficient self-attention, "SSA" stands for spatial self-attention, and "CSA" represents channel self-attention.

$$I_s(X_1, X_2) = X_1 \odot \sigma(W_{\text{sp2}} \cdot GELU(W_{\text{sp1}} \cdot X_2)),$$
  

$$I_c(X_1, X_2) = X_1 \odot \sigma(W_{\text{ch2}} \cdot GELU(W_{\text{ch1}} \cdot GAP(X_2)))$$
(13)

where  $\sigma$  indicates sigmoid activation. GAP refers to the global average pooling layer. W indicates the weight of the point-wise convolution for downscaling or upscaling channel dimensions. The ratios are set to  $W_{\rm spl}=r_1\downarrow$ ,  $W_{\rm sp2}=\frac{C}{r_1}\downarrow$  and  $W_{\rm chl}=r_2\downarrow$ ,  $W_{\rm ch2}=r_2\uparrow$ . In this case, given an input  $X_{\rm in}$ , we get the output from Spatial Selfattention (SSA) as follows:

$$Y_S = (I_c(Y_{Sp}, Y_{local}) + I_s(Y_{local}, Y_{Sp})) W_{merge} + X_{in}$$
 (14)

where  $W_{\rm merge}$  indicates the integrating projection matrix,  $Y_S$  represents the output from the SSA operation.

3) Channel Self-attention (CSA): Capturing intricate channel-wise relationships is crucial for medical image segmentation with different shapes and scales. Thus, we expand self-attention to the channel dimension, enhancing feature representation by focusing on inter-channel relationships, which complements token-wise self-attention in Transformer. Given an input f, linear projections are utilized to generate channel queries  $(Q_{\rm ch})$ , keys  $(K_{\rm ch})$ , and values  $(V_{\rm ch})$ , all are reshaped to  $\mathbb{R}^{HW \times C}$ . Following a similar process above, the projection vectors are split into h heads. The channel-wise attention for the i-th head is computed as:

$$Y_{\text{ch}}^{i} = V_{\text{ch}}^{i} \cdot \operatorname{softmax} \left( \frac{(Q_{\text{ch}}^{i})^{T} K_{\text{ch}}^{i}}{\alpha} \right),$$

$$Y_{Ch} = \operatorname{concat}(Y_{\text{ch}}^{1}, \dots, Y_{\text{ch}}^{h}) W_{\text{merge}}$$
(15)

where  $Y_{\rm ch}^i$  represents the output of the i-th head, enhancing channel-specific features.  $\alpha$  is a learnable scaling factor that modulates the attention mechanism's sensitivity. The feature map  $Y_{Ch}$  is obtained by concatenating and reshaping all  $Y_{\rm ch}^i$  outputs. Similarly, in this case, given an input  $X_{in}$ , we get the output from Channel Self-attention (CSA) as follows:

$$Y_C = (I_s(Y_{Ch}, Y_{local}) + I_c(Y_{local}, Y_{Ch})) W_{merge} + X_{in}$$
 (16)

The output of the multi-dimension Transformer block can be formulated as:

$$Y = MLP(Y_E + \lambda_1 Y_S + \lambda_2 Y_C) \tag{17}$$

where MLP denoted the multilayer perceptron,  $\lambda_1$  is set to 0.6, and  $\lambda_2$  is set to 0.4. Finally, combining spatial and channel dimension information with initial self-attention facilitates comprehensive feature representation for medical image segmentation.

## 4 Experimental setup

## 4.1 Datasets

To comprehensively validate the effectiveness of MDT-AF, we use three widely used medical image datasets, each acquired with different imaging devices and capturing distinct subjects. The metrics we employ for the evaluation include Accuracy (Acc) and Dice Score (DSC).

- *1) Lung X-ray Dataset*: The dataset was acquired jointly by Shenzhen Hospital, China [2], and the tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA [19]. It comprises 704 chest X-ray (CXR) images and corresponding label masks. We conducted 5-fold cross-validation and reported the average results across these folds.
- 2) Skin Lesion Dataset: This study presents a dermoscopic image analysis benchmark challenge aimed at automated skin cancer diagnosis. We utilized a fusion dataset comprising images from ISIC 2017 [10], ISIC 2018 [9], and the PH2 dataset [28]. Specifically, the dataset used for this study consists of 2794 samples for training and an additional 600 samples for testing purposes.
- 3) Kvasir-SEG Dataset: The Polyps Dataset (Kvasir-SEG) is a publicly available dataset that includes 1000 polyp images along with their corresponding segmentation masks [20]. In this study, we conducted 5-fold cross-validation and reported the average results across these validation folds.

# 4.2 Implementation details

All experiments were conducted using PyTorch and trained on a single Nvidia A100 GPU. Input medical images were resized to

**Table 1.** Comparison with state-of-the-art methods was conducted on the Lung X-ray, Skin Lesion, and Kvasir-SEG datasets. The evaluation highlighted the best-performing methods with bold for the highest dice score (DSC) and accuracy (Acc). The second-best results were marked in red, while the third-best results were indicated in blue.

Lung X-ray

Skin Lesion

Kvasir-SEG

Model	Lung X-ray		Skin Lesion		Kvasir-SEG	
	Acc(%)↑	DSC(%)↑	Acc(%)↑	DSC(%)↑	Acc(%)↑	DSC(%)↑
UNet [32] [MICCAI'15]	96.67	94.27	93.86	83.54	95.62	87.96
AttnUNet [29] [MIDL'18]	98.11	95.97	93.74	83.27	95.49	87.14
DeepLabv3+ [5] [ECCV'18]	97.54	94.51	96.65	91.03	96.45	86.35
UNet++ [50] [TMI'20]	96.69	94.33	93.59	82.96	95.51	87.01
UNet3+ [17] [ICASSP'20]	97.99	95.63	95.18	88.52	95.19	83.95
TransUNet [4] [ArXiv'21]	97.58	95.25	93.32	82.25	96.40	89.21
MedT [38] [MICCAI'21]	96.07	91.91	92.18	81.83	88.96	45.97
SwinUNet [3] [ECCVW'22]	95.88	90.67	93.46	84.01	93.31	70.71
Missformer [18] [TMI'22]	97.86	95.49	93.93	84.50	92.98	71.56
Hiformer [16] [WACV'23]	97.79	94.73	95.57	88.59	94.51	80.51
MERIT [31] [MIDL'23]	89.85	73.20	94.47	79.73	93.94	76.85
MDT-AF [Ours]	98.60	97.17	97.77	94.25	98.12	93.38

512×512 for consistency in comparison. We employed the AdamW optimizer with a momentum of 0.9 and weight decay of 1e-2 to train our model for 100 epochs with a batch size of 8. A cosine learning rate scheduler was utilized during training, with maximum and minimum learning rates set to 1e-4 and 1e-6, respectively. For the loss function, we only utilized CrossEntropy loss, defined as shown in Formula 18:

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} \left[ y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]$$
 (18)

where N represents the number of samples,  $y_i$  denotes the label of the  $i^{th}$  sample, and  $\hat{y}_i$  indicates the predicted probability of the  $i^{th}$  sample.

# 5 RESULTS

## 5.1 Comparison with state-of-the-arts

1) Results of Lung X-ray Dataset: The comparison of our MDT-AF with state-of-the-art (SOTA) methods on the Lung X-ray Dataset is presented in Table 1. Our MDT-AF achieves 98.60% accuracy (Acc) and 97.17% Dice Score (DSC) on this dataset. MDT-AF demonstrates significant performance improvements over CNNbased methods, i.e., AttnUNet [29] (+0.49% Acc and +1.2% DSC) and UNet3+ [17] (+0.61% Acc and +1.54% DSC). Compared to transformer-based methods, MDT-AF remains competitive, surpassing four widely recognized models: TransUNet [4] (+1.02% Acc and +1.92% DSC), SwinUNet [3] (+2.72% Acc and +6.5% DSC), Missformer [18] (+0.74% Acc and +1.68% DSC), and Hiformer [16] (+0.81% Acc and +2.44% DSC). Qualitative results of some methods are depicted in Fig 4. We can observe that MDT-AF accurately segments delicate and complex structures with more precise boundaries, showing robustness against complicated backgrounds. In contrast, many baseline methods like TransUNet and Hiformer struggle to locate regions of interest precisely and exhibit misclassified pixels.

2) Results of Skin Lesion Dataset: To further demonstrate the generalization capability of our MDT-AF, we evaluated it on the Skin Lesion Dataset, and the experimental results are summarized in Table 1. Our MDT-AF performs better than state-of-the-art methods, achieving 97.77% accuracy (Acc) and 94.25% Dice Score (DSC). Figure 4 further compares skin lesion segmentation visual results, illustrating that our method captures finer structures and produces more precise contours. As depicted in Figure 4, our method outperforms hybrid methods like Hiformer, particularly in boundary areas.

**Table 2.** Ablation studies were conducted on each component using the Lung X-ray, Skin Lesion, and Kvasir-SEG datasets. The components evaluated include Patch Embedding (PE), Patch Embedding with Attention-based Filtering (AF), Transformer Blocks with Efficient Self-attention (ESA), and Transformer Blocks with Multi-dimension Self-attention (MSA). The best re-

sults are highlighted in bold.			l in bold.	DSC (%) ↑				
PE	AF	ESA	MSA	Lung X-ray	Skin Lesion	Kvasir-SEG		
$\checkmark$		✓		96.84 (-0.33)	94.16 (-0.09)	92.27 (-1.11)		
	$\checkmark$	$\checkmark$		97.12 (-0.05)	94.20 (-0.05)	92.95 (-0.43)		
$\checkmark$			$\checkmark$	97.12 (-0.05)	94.19 (-0.06)	92.60 (-0.78)		
	$\checkmark$		$\checkmark$	97.17	94.25	93.38		

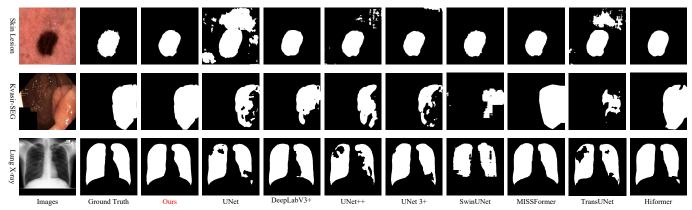
Additionally, MDT-AF demonstrates greater robustness to noisy elements than purely transformer-based methods such as SwinUNet. This performance results from the superiority of our attention-based patch embedding and transformer blocks with multi-dimension self-attention.

3) Results of Kvasir-SEG Dataset: In Table 1, we present results from the Kvasir-SEG dataset, where the MDT-AF architecture consistently outperforms state-of-the-art methods. Additionally, Figure 4 showcases segmentation outputs of the proposed MDT-AF, highlighting how our predictions closely align with the provided ground truth. MDT-AF's key advantage lies in its ability to achieve multidimension information representation and aggregation. Moreover, it effectively suppresses background noise through the attention-based filtering mechanism in the patch embedding process. In contrast, other methods easily encounter issues like boundary-blurring and misclassification when processing endoscopic images.

#### 5.2 Ablation study

Extensive ablation studies were conducted on the Lung X-ray, Skin Lesion, and Kvasir-SEG datasets in order to verify the effectiveness of the attention-based filtering mechanism and transformer with multi-dimension self-attention. The Dice score (DSC) was selected as the default evaluation metric, and quantitative results are reported in Table 2.

1) Effectiveness of the Attention-based Filtering: In this experiment, we tested two variants of MDT-AF: a) PE+ESA: This variant removes the attention-based filtering procedure, relying solely on patch embedding to generate all patch tokens. b) AF+ESA: This variant uses a coarse-to-fine approach to generate finer patch tokens by incorporating attention-based filtering to guide the patch embedding process. The quantitative results in Table 2 demonstrate that



**Figure 4.** The visual comparison results are presented for the Lung X-ray, Skin Lesion, and Kvasir-SEG datasets. The images in the first row depict the results of skin lesion segmentation, the second row shows the outcomes of polyp segmentation, and the final row displays the results of lung segmentation.

both the model equipped with the attention-based filtering mechanism and the overall MDT-AF model outperform the baseline method across several datasets. Specifically, improvements over the baseline are observed in the Lung X-ray (+0.05% and +0.33%), Skin Lesion (+0.05% and +0.09%), and Kvasir-SEG datasets (+0.43% and +1.11%).

2) Strength of the Multi-dimension Self-attention: The multi-dimension self-attention module was removed in order to create a baseline model (PE+ESA). The models that use transformer blocks with multi-dimension self-attention (PE+MSA and MDT-AF) show significant improvements over the baseline. Specifically, improvements were observed on the Lung X-ray (+0.05% and +0.33%), Skin Lesion (+0.06% and +0.09%), and Kvasir-SEG datasets (+0.78% and +1.11%). The model can learn and aggregate richer feature representation across spatial and channel by integrating multi-dimension self-attention mechanism, which improves overall performance and feature discriminative ability.

#### 6 Conclusion

In this paper, we introduce the MDT-AF, a novel transformer variant (MDT-AF) customized for precise and robust medical image segmentation. This model combines patch embedding with an attentionbased filtering mechanism to provide a coarse-to-fine process, where coarse features with noise will be filtered for a finer feature representation. Low signal-to-noise ratio is a significant challenge in medical image segmentation, and this design demonstrated its ability to solve it. Additionally, richer feature representation can be captured via the modified self-attention mechanism, which effectively constructs and aggregates multi-dimension information across the spatial and channel. It provides advantages in processing the complex structures of medical images across various scales. The effectiveness and competitiveness of the MDT-AF have been proved by its superior performance, which is higher than the state-of-the-art methods on three publicly available medical image segmentation benchmarks without any advanced training strategies. In the future, we will explore the extension of MDT-AF to other downstream medical image analysis tasks.

## Acknowledgements

## References

- M. Z. Alom et al. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. arXiv preprint arXiv: 1802.06955, 2018.
- [2] S. Candemir et al. Lung Segmentation in Chest Radiographs Using Anatomical Atlases With Nonrigid Registration. *IEEE Transactions on Medical Imaging (TMI)*, 33(2):577–590, 2014.
- [3] H. Cao et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision – ECCV 2022 Workshops*, pages 205–218, 2023.
- [4] J. Chen et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv: 2102.04306, 2021.
- [5] L.-C. Chen et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [6] S. Chen et al. Sparse Sampling Transformer with Uncertainty-Driven Ranking for Unified Removal of Raindrops and Rain Streaks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 13106–13117, 2023.
- [7] S. Chen et al. Dual-former: Hybrid self-attention transformer for efficient image restoration. *Digital Signal Processing*, 149:104485, 2024.
- [8] Z. Chen et al. Dual aggregation transformer for image super-resolution. In Proceedings of the IEEE/CVF international conference on computer vision, pages 12312–12321, 2023.
- [9] N. Codella et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (isic), 2019.
- [10] N. C. F. Codella et al. Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (isbi), Hosted by the International Skin Imaging collaboration (isic), 2018.
- [11] P.-H. Conze et al. Current and Emerging Trends in Medical Image Segmentation With Deep Learning. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 7(6):545–569, 2023.
- [12] A. J. DeGrave et al. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. *Nature Biomedical Engineering*, pages 1–13, 2023.
- [13] A. Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv: 2010.11929, 2021.
- [14] S. Gerl et al. A Distance-Based Loss for Smooth and Continuous Skin Layer Segmentation in Optoacoustic Images. In *Proceedings of Medical Image Computing and Computer Assisted Intervention*, pages 309–319, 2020.
- [15] X. Guo et al. Group-Wise Correlation Stereo Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
- [16] M. Heidari et al. Hiformer: Hierarchical Multi-Scale Representations Using Transformers for Medical Image Segmentation. In *Proceedings* of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 6202–6212, 2023.
- [17] H. Huang et al. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1055– 1059, 2020.
- [18] X. Huang et al. MISSFormer: An Effective Transformer for 2D Medical Image Segmentation. *IEEE Transactions on Medical Imaging (TMI)*, 42 (5):1484–1494, 2023.
- [19] S. Jaeger et al. Automatic Tuberculosis Screening Using Chest Radiographs. *IEEE Transactions on Medical Imaging (TMI)*, 33(2):233–245, 2014.
- [20] D. Jha et al. Kvasir-SEG: A Segmented Polyp Dataset. In MultiMedia Modeling, pages 451–462, 2020.
- [21] A. Kaur et al. A complete review on image denoising techniques for medical images. *Neural Processing Letters*, 55(6):7807–7850, 2023.
- [22] J. Li et al. A Systematic Collection of Medical Image Datasets for Deep Learning. ACM Computing Surveys, 56(5):51, 2023.
- [23] X. Li et al. Dlgsanet: lightweight dynamic local and global selfattention networks for image super-resolution. In *Proceedings of* the IEEE/CVF International Conference on Computer Vision, pages 12792–12801, 2023.
- [24] A. Lin et al. Ds-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation. *IEEE Transactions on Instrumentation and Measurement (TIM)*, 71:1–15, 2022.
- [25] L. Liu et al. A survey on U-shaped networks in medical image segmentations. *Neurocomputing*, 409:244–258, 2020.

- [26] Y. Liu et al. A Survey of Visual Transformers. IEEE Transactions on Neural Networks and Learning Systems, pages 1–21, 2023.
- [27] D. Marin et al. Token Pooling in Vision Transformers for Image Classification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 12–21, January 2023.
- plications of Computer Vision (WACV), pages 12–21, January 2023.

  [28] T. Mendonça et al. PH2 A dermoscopic image database for research and benchmarking. In Proceedings of 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 5437–5440, 2013.
- [29] O. Oktay et al. Attention U-Net: Learning Where to Look for the Pancreas. CoRR, abs/1804.03999, 2018. URL http://arxiv.org/abs/1804.03999.
- [30] K. Patel et al. Enhanced U-Net: A Feature Enhancement Network for Polyp Segmentation. In *Proceedings of 2021 18th Conference on Robots and Vision (CRV)*, pages 181–188, 2021.
- [31] M. M. Rahman et al. Multi-scale Hierarchical Vision Transformer with Cascaded Attention Decoding for Medical Image Segmentation. In Proceedings of Medical Imaging with Deep Learning (MIDL), volume 227, pages 1526–1544, 2024.
- [32] O. Ronneberger et al. U-net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), pages 234–241, 2015.
- [33] J. Ruan et al. Ege-UNet: An Efficient Group Enhanced UNet for Skin Lesion Segmentation. In Proceedings of Medical Image Computing and Computer Assisted Intervention, pages 481–490, 2023.
- [34] D. Sarvamangala et al. Convolutional neural networks in medical image understanding: a survey. *Evolutionary intelligence*, 15(1):1–22, 2022.
- [35] F. Shamshad et al. Transformers in medical imaging: A survey. Medical Image Analysis, 88:102802, 2023.
- [36] Y. Sun et al. Msca-Net: Multi-scale contextual attention network for skin lesion segmentation. *Pattern Recognition*, 139:109524, 2023.
- [37] C. Tang et al. Learning Spatial-Frequency Transformer for Visual Object Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):5102–5116, 2023.
- [38] J. M. J. Valanarasu et al. Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 36– 46, 2021.
- [39] A. Vaswani et al. Attention is All you Need. In Proceedings of Advances in Neural Information Processing Systems (NeurIPS), volume 30, 2017.
- [40] X. Wang et al. Selective-Stereo: Adaptive Frequency Information Selection for Stereo Matching, 2024.
- [41] E. Xie et al. Segformer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 12077–12090, 2021.
- [42] G. Xu et al. Attention Concatenation Volume for Accurate and Efficient Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12981–12990, 2022.
- [43] G. Xu et al. Cgi-Stereo: Accurate and Real-Time Stereo Matching via Context and Geometry Interaction. arXiv preprint arXiv:2301.02789, 2023.
- [44] G. Xu et al. Memory-Efficient Optical Flow via Radius-Distribution Orthogonal Cost Volume, 2023.
- [45] M. Yeung et al. Focus U-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy. Computers in Biology and Medicine, 137:104815, 2021.
- [46] D. Yu et al. Dstrans: Dual-Stream Transformer for Hyperspectral Image Restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3739–3749, January 2023.
- [47] H. Yu et al. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neuro-computing*, 444:92–110, 2021.
- [48] F. Yuan et al. An effective CNN and Transformer complementary network for medical image segmentation. *Pattern Recognition*, 136: 109228, 2023.
- [49] Y. Zhang et al. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI), pages 14–24, 2021.
- [50] Z. Zhou et al. Unet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging (TMI)*, 39(6):1856–1867, 2020.