



OPEN ACCESS

EDITED BY

Zetao Chen,
Tianjin University, China

REVIEWED BY

Wencai Du,
University of Saint Joseph, Macao SAR, China
Jia Guo,
Hosei University, Japan

*CORRESPONDENCE

Teeradaj Racharak,
✉ racharak@jaist.ac.jp
Kai Li,
✉ kaili@fudan.edu.cn
Guanqun Sun,
✉ sun.guanqun@hmc.edu.cn

RECEIVED 30 September 2024
ACCEPTED 12 December 2024
PUBLISHED 24 December 2024

CITATION

Pan Y, Xin J, Yang T, Li S, Nguyen L-M,
Racharak T, Li K and Sun G (2024) A mutual
inclusion mechanism for precise boundary
segmentation in medical images.
Front. Bioeng. Biotechnol. 12:1504249.
doi: 10.3389/fbioe.2024.1504249

COPYRIGHT

© 2024 Pan, Xin, Yang, Li, Nguyen, Racharak, Li
and Sun. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

A mutual inclusion mechanism for precise boundary segmentation in medical images

Yizhi Pan¹, Junyi Xin^{1,2}, Tianhua Yang¹, Siqi Li³, Le-Minh Nguyen³,
Teeradaj Racharak^{3*}, Kai Li^{1,2*} and Guanqun Sun^{1,2,3*}

¹School of Information Engineering, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, China, ²Zhejiang Provincial Engineering Research Center for Brain Cognition, Disease and Digital Medical Devices, Hangzhou Medical College, Hangzhou, China, ³School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan

Introduction: Accurate image segmentation is crucial in medical imaging for quantifying diseases, assessing prognosis, and evaluating treatment outcomes. However, existing methods often fall short in integrating global and local features in a meaningful way, failing to give sufficient attention to abnormal regions and boundary details in medical images. These limitations hinder the effectiveness of segmentation techniques in clinical settings. To address these issues, we propose a novel deep learning-based approach, MIPC-Net, designed for precise boundary segmentation in medical images.

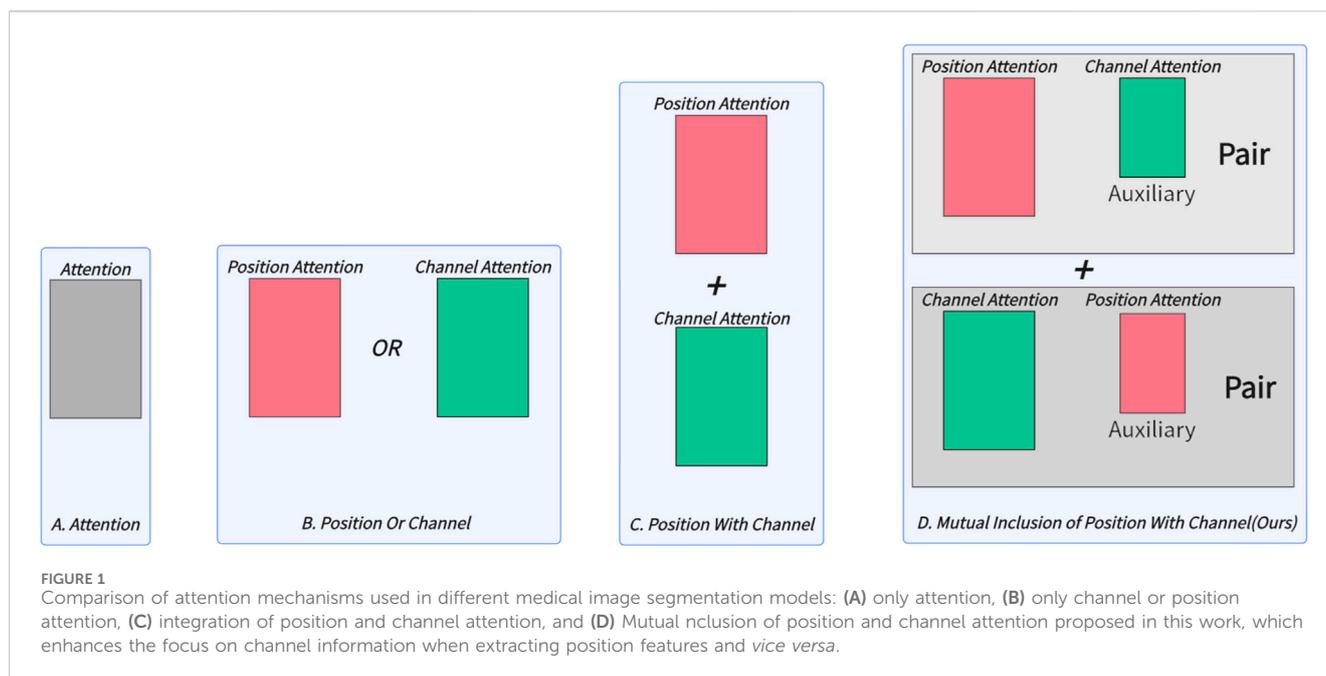
Methods: Our approach, inspired by radiologists' working patterns, introduces two distinct modules: 1. Mutual Inclusion of Position and Channel Attention (MIPC) Module: To improve boundary segmentation precision, we present the MIPC module. This module enhances the focus on channel information while extracting position features and vice versa, effectively enhancing the segmentation of boundaries in medical images. 2. Skip-Residue Module: To optimize the restoration of medical images, we introduce Skip-Residue, a global residual connection. This module improves the integration of the encoder and decoder by filtering out irrelevant information and recovering the most crucial information lost during the feature extraction process.

Results: We evaluate the performance of MIPC-Net on three publicly accessible datasets: Synapse, ISIC2018-Task, and Segpc. The evaluation uses metrics such as the Dice coefficient (DSC) and Hausdorff Distance (HD). Our ablation study confirms that each module contributes to the overall improvement of segmentation quality. Notably, with the integration of both modules, our model outperforms state-of-the-art methods across all metrics. Specifically, MIPC-Net achieves a 2.23 mm reduction in Hausdorff Distance on the Synapse dataset, highlighting the model's enhanced capability for precise image boundary segmentation.

Conclusion: The introduction of the novel MIPC and Skip-Residue modules significantly improves feature extraction accuracy, leading to better boundary recognition in medical image segmentation tasks. Our approach demonstrates substantial improvements over existing methods, as evidenced by the results on benchmark datasets.

KEYWORDS

U-Net, medical image segmentation, mutual inclusion, transformer, deep learning



1 Introduction

Medical image segmentation plays a pivotal role in quantifying diseases, assessing prognosis, and evaluating treatment outcomes. It describes crucial observations in images, such as the degree, size, and location of lesions. However, manual segmentation by skilled professionals is both time-consuming and tedious (Sun et al., 2024). Therefore, with the advance of deep learning technologies, automatic medical image segmentation has attracted growing research interest.

Existing medical image segmentation methods usually follow the practice of combining Convolutional Neural Networks (CNNs) with Vision Transformer modules under the U-Net structure (Ronneberger et al., 2015; Long et al., 2015; Dosovitskiy et al., 2020). For example, various U-Net variants have been proposed to improve medical image segmentation performance. ResUNet (Diakogiannis et al., 2020), Unet++ (Zhou et al., 2018), and Unet3++ (Huang et al., 2020) introduced residual connections and complex skip connections, while Attention-Unet (Oktay et al., 2018) integrated attention mechanisms into the U-Net architecture. TransUNet (Chen et al., 2021) and Swin-Unet (Cao et al., 2022) incorporated Transformer and Swin-Transformer (Liu et al., 2021) modules, respectively, to capture global information. However, medical image segmentation differs from generic image segmentation tasks. In medical image segmentation, data is characterized by small sample sizes and the need for precise boundary delineation. Unlike generic image segmentation models, which are required to cover all details of the image, medical image segmentation demands special attention to abnormal regions and boundary details in organ or pathological images. Therefore, local image features need to be combined with global features. To this end, attention mechanisms focusing on both channel and position information need to be introduced into the research.

In recent research, there has been a trend towards incorporating both channel and position attention mechanisms into models. SA-UNet (Guo et al., 2021) and AA-TransUNet (Yang and Mehrkanoon, 2022) incorporated spatial and channel attention, respectively, but lack comprehensive utilization of image features. TransUNet++ (Jamali et al., 2023) and DS-TransUNet (Lin et al., 2022) integrated Transformers into skip connections but have limitations in overall architecture and feature integration. DA-TransUNet (Sun et al., 2023) merges position and channel attention but merely adapts a block from road segmentation, lacking tailored feature extraction for medical images. In TransUNet (Chen et al., 2024), a versatile framework is proposed that allows the integration of the self-attention mechanism at multiple stages of the model, while still focusing on exploring the Transformer mechanism. In MVRM (Zuo et al., 2024), MambaBlock is used to enhance feature extraction; however, the improvement of the model's boundary segmentation capability has not been considered. These methods achieve improved performance over previous medical image segmentation models. However, they focus primarily on the overall segmentation overlap rather than specifically enhancing the boundary details of the segmentation results. Moreover, when extracting features from the perspective of channel and position, these models only focus on repeated feature extraction, potentially disrupting the original information without considering how to restore the boundary details of the image.

Inspired by radiologists' working patterns, this paper proposes a simple and effective mutual inclusion mechanism for medical image segmentation. Instead of simply stacking Transformer-related modules, we introduce the Mutual Inclusion of Position and Channel Attention (MIPC) module, which enhances the focus on channel information when extracting position features and vice versa. Figure 1 illustrates the superiority of our proposed mutual inclusion of position and channel attention compared to existing attention mechanisms. We propose two pairs of channel and position combinations, each pair emphasizing either channel or

position information while mutually including the other. This approach mimics the radiologists' working patterns, where mutual inclusion is practiced with varying emphasis. The experimental results demonstrate that this method effectively improves the model's ability to accurately segment image boundary. Furthermore, we focus on the restoration of medical images by proposing the Global-Skip-Connections. This connection introduces a Dual Attention mechanism to filter out invalid information while utilizing a Skip-Residue to restore the most effective information lost during the feature extraction process.

We evaluate our proposed methods on three publicly accessible datasets: the Synapse dataset (Landman et al., 2015), the ISIC2018-Task dataset (Codella et al., 2019; Tschandl et al., 2018), and the Segpc dataset (Gupta et al., 2021). In addition to the Dice coefficient (DSC) metrics, which deal with class imbalance problems, we adopt the Hausdorff Distance (HD) to analyze the quality of the segmentation results, as it is particularly convincing in evaluating boundary region segmentations. The results show that the proposed method achieves state-of-the-art performance on both DSC and HD metrics. Notably, there was a 2.23 mm reduction over competing models in the HD metric on the benchmark Synapse dataset, strongly evidencing our model's enhanced capability for precise image boundary segmentation. This finding also indicates that medical image segmentation benefits from the mutual inclusion mechanism of position and channel attention.

The main contributions are as follows:

- 1) This paper proposes a novel model, MIPC-Net, which incorporates a Mutual Inclusion attention mechanism for position and channel information. This approach further enhances the precision of boundary segmentation in medical images.
- 2) This paper introduces the Skip-Residue, a global residual connection that improves image restoration by enhancing the integration of the encoder and decoder.
- 3) Experiments demonstrate that the proposed components achieve consistent performance improvements. Furthermore, our model achieves state-of-the-art performance across all metrics on the public Synapse (Landman et al., 2015), ISIC2018-Task (Codella et al., 2019; Tschandl et al., 2018), and Segpc (Gupta et al., 2021) datasets.
- 4) The rest of this article is organized as follows. Section 2 reviews the related works of automatic medical image segmentation, and the description of our proposed MIPC-Net is given in Section 3. Next, the comprehensive experiments and visualization analyses are conducted in Section 4. Finally, Section 5 makes a conclusion of the whole work.

2 Related work

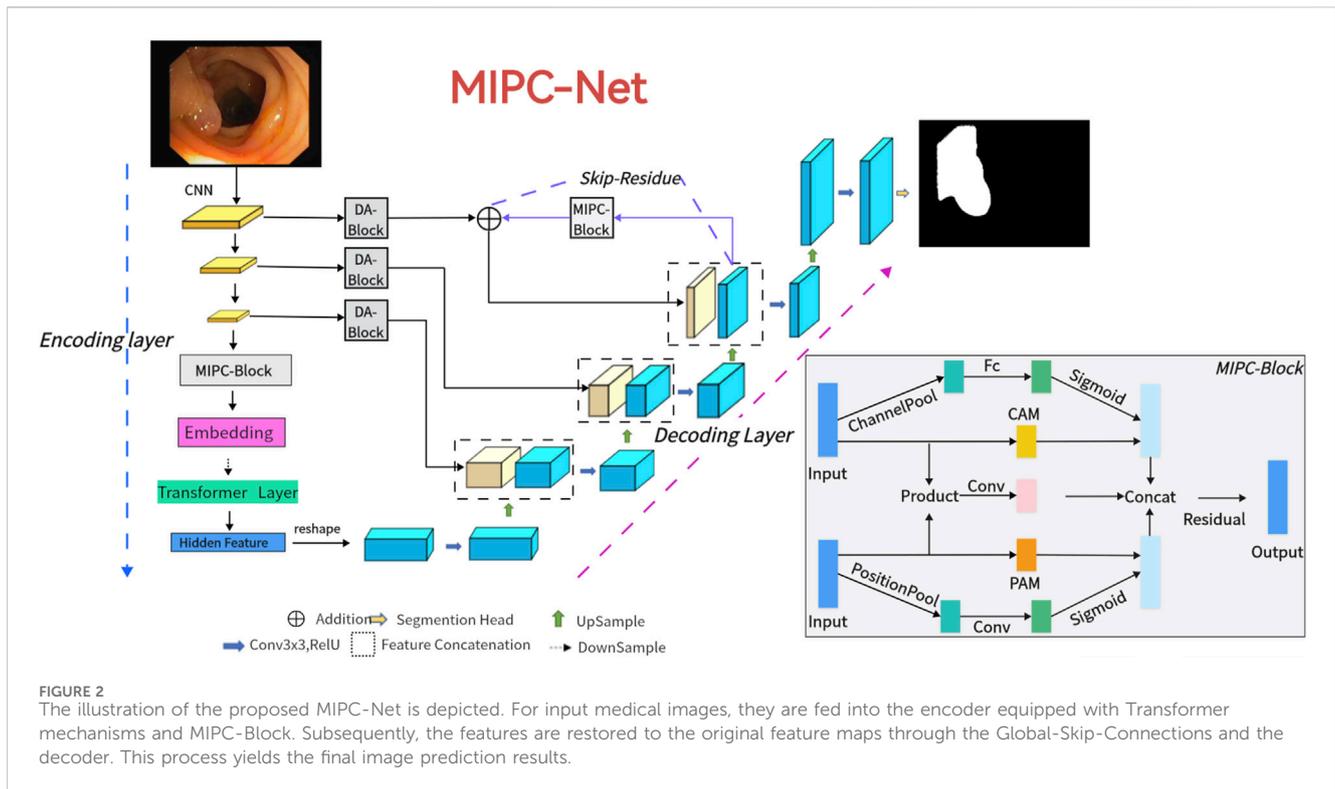
2.1 Model integration of U-structure

Research on U-Net architecture optimization has made significant strides in recent years. Proper utilization of residual learning and skip connections can enhance a model's learning capacity when constructing deep neural networks. The original

U-Net (Ronneberger et al., 2015) introduced skip connections to improve feature fusion, but the potential of these connections was not fully exploited. Subsequent works have aimed to address this limitation. UNet++ (Zhou et al., 2018) incorporated a densely connected network architecture to enrich skip connections and improve model performance, but it did not explore the integration of other optimization mechanisms. Building upon this, UNet3++ (Huang et al., 2020) introduced hierarchical skip connections to further enhance the model's feature extraction capability, focusing on enriching skip connections without specifically optimizing feature transmission during the process. Several works have sought to refine skip connections by integrating attention mechanisms and Transformer components. DAREsUNet (Shi et al., 2020a) incorporated residual modules and Dual Attention (DA) Blocks, but only optimized the first-layer skip connection. DS-TransUNet (Lin et al., 2022) merged Transformer mechanisms into the skip connections, but did not fully consider the overall model structure. Similarly, IB-TransUNet (Li et al., 2023) integrated a multi-resolution fusion mechanism into skip connections without a holistic view of the model architecture. A recent work, DA-TransUNet (Sun et al., 2023), optimized skip connections using image feature positions and channels, but the integration of these components into the overall model was insufficient, leaving room for further improvement. While these works have made valuable contributions, there is still a need for a more comprehensive approach that optimizes skip connections and enhances the overall integration of the model components. In this paper, we propose a novel architecture that not only optimizes the skip connections at multiple levels but also strengthens the overall integration of the model components. Our approach leverages the strengths of residual learning, attention mechanisms, and Transformer modules to capture rich contextual information and enhance feature fusion.

2.2 The utilization of attention modules

The attention mechanism has become a crucial component in enhancing model performance by enabling focus on target features. Since its introduction in the Bahdanau Attention paper (Bahdanau et al., 2014) in 2014 for machine translation, the field of attention mechanisms has witnessed continuous advancements and iterations. In 2015, the introduction of attention mechanisms for image generation significantly enhanced the quality of the produced images (Gregor et al., 2015), while the application of visual attention mechanisms to image description generation sparked substantial interest in the image captioning domain (Xu et al., 2015). The same year, the introduction of various attention mechanism variants, such as global attention and local attention, marked a significant advancement (Luong et al., 2015). The evolution of attention mechanisms was further propelled forward in 2017 with the proposal of sub-attention mechanisms (Vaswani et al., 2017). In 2019, the pioneering introduction of dual attention mechanisms was marked by the employment of dual attention modules for scene segmentation, integrating both spatial and channel attention mechanisms (Fu et al., 2019). The modular DAN (Dual Attention Network) framework, combining visual and textual attention, achieved significant outcomes in visual



question-answering (VQA) tasks (Nam et al., 2017). The introduction of the Dual Attention Matching (DAM) module enhanced high-level event information modeling over extended video durations, complemented by a global cross-check mechanism for precise localization of visible and audible events in videos (Wu et al., 2019). Furthermore, the application of dual attention mechanisms in medical image segmentation has shown promising results, but the strategies for optimizing feature extraction through position and channel attention mechanisms require further investigation (Shi et al., 2020b). Despite the significant progress in attention mechanisms, there remain limitations in their application to medical image segmentation tasks. Mutual Inclusion has been explored in other fields, but its application in the fusion of attention modules has not been previously investigated (Hosseinzadeh and Wang, 2021; Zhang et al., 2015). In this paper, we propose a novel Mutually Inclusion of Position and Channel (MIPC) Block, which aims to enhance the segmentation performance of the model by mutually including position and channel attention modules and incorporating the concept of residue. Our approach seeks to leverage the complementary nature of position and channel information, enabling the model to capture more comprehensive and discriminative features for medical image segmentation.

3 Methods

In the following section, we introduce the MIPC-Net architecture, as depicted in Figure 2. We begin by providing an overview of the overall structure. Subsequently, we present its key components in the following sequence: Mutual Inclusion of Position

and Channel (Section 3.2), the encoder (Section 3.3), Global-Skip-Connections (Section 3.4), and the decoder (Section 3.5).

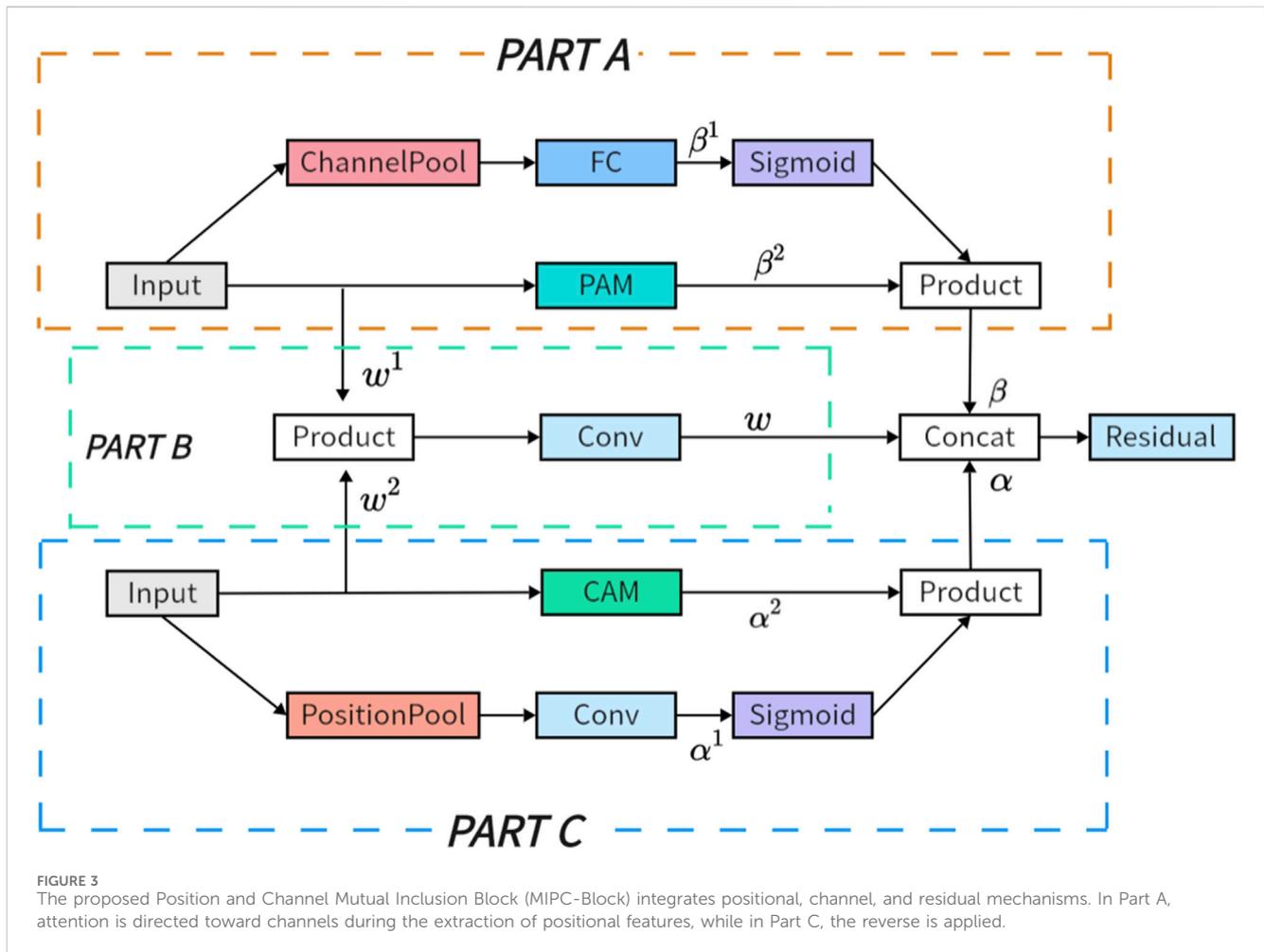
3.1 Overview of MIPC-Net

Figure 2 illustrates the detailed configuration of our MIPC-Net model, which is a medical image segmentation model capable of capturing image-specific channel and position information and incorporates improved skip connections.

Our model consists of three main components: the encoder, the decoder, and the Global-Skip-Connections. Notably, the encoder integrates traditional convolutional neural network (CNN) and Transformer mechanisms, while using MIPC-Block to enhance the encoding capability (Section 3.3). The decoder relies on deconvolution to restore the features to the original feature map size (Section 3.5). Global-Skip-Connections employ DA-Block to purify the features of skip connection transmission. Furthermore, they use the Global-Skip to further enhance the integrity of the encoder and decoder (Section 3.4). MIPC-Net, comprised of three integral components, exhibits superior performance in image segmentation.

Given the constraints highlighted by traditional models, it is evident that while the conventional U-Net architecture excels in capturing image features, it lacks effective methods for preserving and extracting global features.

On the other hand, Transformers exhibit remarkable proficiency in preserving and extracting global features through self-attention mechanisms (Chen et al., 2021). However, they are inherently limited to unidirectional positional attention, overlooking the utilization of image-special position and channel. To address



these limitations, we have integrated the Mutual Inclusion of Position and Channel Block (MIPC-Block) and leveraged Global-Skip-Connections to enhance the integrity of the encoder and decoder, thereby improving medical image segmentation performance.

In medical image segmentation tasks, current models usually use attention mechanisms to enhance the segmentation capabilities of the model. For example, TransUNet uses ViT, and Swin-Unet uses Swin-Transformer. These approaches fail to adapt attention mechanisms to the specific features of the image, hence unable to extract deep image-related information. To solve this problem, our proposed MIPC-Block enhances the segmentation capabilities of the model by leveraging image-specific features related to position and channel. It effectively combines these two features in a mutually inclusive manner to extract deeper image-related features, achieving subdivided extraction of image features and more fully mining features.

As illustrated in Figure 3, the MIPC-Block architecture seamlessly integrates image-specific channel and positional features, enriched by the application of residual concepts. The amalgamation of channel and positional features empowers the MIPC-Block with profound insights into the image, surpassing the capabilities of conventional attention modules.

The MIPC-Block architecture consists of three parts: PART A, PART B, and PART C. PART A and PART C serve as crucial feature extraction modules, ingeniously integrating both position and channel information of image features. The tight coupling of positional and channel information further enhances the feature extraction capability of the module. In Part A, our module undergoes a channel-wise average pooling layer (ChannelPool) to compress the feature map. Subsequently, it passes through fully connected layers to learn the correlations between different channels in the features. Following this, a sigmoid function is applied to constrain the values between 0 and 1, yielding channel correlations. Multiplying these correlations with the features obtained through the Position Attention Module (PAM) results in information where the position is the primary focus and channels act as auxiliary. Conversely, in Part C, features are first subjected to MaxPool and AvgPool operations (PositionPool) along the spatial dimensions. The resulting features from these two pooling operations are concatenated, and through fully connected layers, correlations between different spatial dimensions in the features are learned. Similar to Part A, a sigmoid function constrains the values between 0 and 1. Multiplying these spatial correlations with the features obtained through the Channel Attention Module (CAM) produces information where channels are the main focus, and spatial dimensions serve as auxiliary. Part B employs a residual

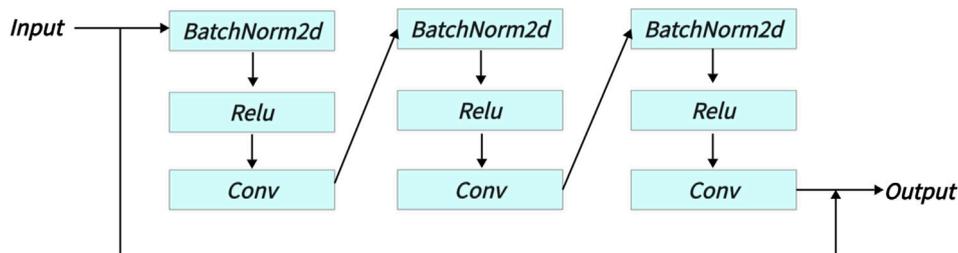


FIGURE 4
The specific structure of the last Residual module in MIPC-Block.

approach to minimize the loss of valuable original information introduced by the convolution and attention modules.

MIPC-Net offers several key advantages in medical image segmentation. Firstly, it integrates position and channel information through the novel Mutual Inclusion of Position and Channel (MIPC) block, enhancing feature extraction by leveraging mutually inclusive attention mechanisms. This approach enables deeper extraction of image-specific features, thereby improving segmentation accuracy. Additionally, the model employs Global-Skip-Connections and DA-Blocks to purify and enhance the integrity of feature transmission between the encoder and decoder. Compared to traditional models such as U-Net, MIPC-Net excels in capturing and preserving global features, overcoming the limitations of conventional architectures. Finally, by adapting to the unique characteristics of medical images, MIPC-Net effectively extracts image-specific information, enhancing its segmentation capability, particularly in boundary segmentation.

3.2 Mutual inclusion of position and channel

Part A (Position-Dominant Extraction with Channel): As illustrated in Figure 3, the extraction of channel information from the input features is facilitated by ChannelPool. Subsequently, a series of fully connected layers is employed to capture inter-channel correlations, yielding Equation 1. Concurrently, another set of input features undergoes processing by the Position Attention Module (PAM) to extract position information features, resulting in Equation 2. Following sigmoid processing of β^1 , it is multiplied element by element with β^2 to obtain Equation 3. In contrast to Part C, where channel-wise modulation is utilized for distributing feature maps from the spatial module, this process generates feature maps with spatial and channel emphasis.

$$\beta^1 = FC(\text{ChannelPool}(\text{Input})), \quad (1)$$

$$\beta^2 = \text{PAM}(\text{Input}), \quad (2)$$

$$\beta = \text{Sigmoid}(\beta^1) \cdot \beta^2, \quad (3)$$

PART B (Residual Part): As shown in the figure, Part A and Part B inputs undergo a convolutional operation to obtain Equation 4 and Equation 5, respectively. Subsequently, the two are element-wise multiplied and then passed through another convolutional layer to

yield Equation 6. It extracts and refines features from both inputs, thereby refining the original features.

$$\omega^1 = \text{Conv}(\text{PartA's Input}), \quad (4)$$

$$\omega^2 = \text{Conv}(\text{PartC's Input}), \quad (5)$$

$$\omega = \text{Conv}(\omega^1 \cdot \omega^2), \quad (6)$$

PART C (Channel-Dominant Extraction with Position): As shown in Figure 3, the input features undergo PositionPool along the spatial dimension to effectively extract spatial information while eliminating noise and irrelevant details in the image. Subsequently, the feature maps are further processed by convolution to capture spatial correlations, resulting in Equation 7. Simultaneously, another set of input features is processed by the Channel Attention Module (CAM) to extract channel features, denoted as Equation 8. The channel attention module is employed to extract detailed channel features from the image. After sigmoid processing of α^1 , it is element-wise multiplied by α^2 to obtain the output Equation 9. Unlike Part A, where the feature maps extracted by the spatial module are weighted by the channel attention module, effectively integrating image-specific spatial and channel features, generating feature maps with channel emphasis and spatial emphasis.

$$\alpha^1 = \text{Conv}(\text{PositionPool}(\text{Input})), \quad (7)$$

$$\alpha^2 = \text{CAM}(\text{Input}), \quad (8)$$

$$\alpha = \text{Sigmoid}(\alpha^1) \cdot \alpha^2, \quad (9)$$

Finally, the outputs of Parts A, B, and C are summed along the channel dimension and then passed through a residual network (see Figure 4) to obtain the Equation 10.

$$\text{Output} = \text{Residual}(\alpha + \beta + \omega), \quad (10)$$

The Mutual Inclusion of Position and Channel block (MIPC-Block) mutually includes the image features' position and channel, capturing deeper features associated with image features compared to standard attention modules.

3.3 Encoder

As shown in Figure 2, the encoder consists of four key components: convolution blocks, MIPC-Block, an embedding layer, and transformer layers.

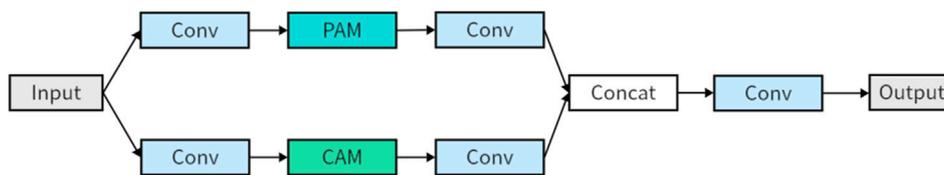


FIGURE 5
Architecture of dual attention block (DA-Block).

It is particularly significant that the MIPC-Block is introduced just before the transformer layers. The purpose is to subject the convolutional features to specialized image processing, enhancing the transformer's feature extraction capabilities with respect to the image's content. The Transformer architecture excels at capturing global information. Integrating the MIPC-Block enhances its ability to maintain and extract global features specifically from images, enriching the Transformer's image processing capabilities. This approach effectively combines image-specific channel and positional features with global features.

It begins with three U-Net convolutional blocks. Each block consists of a series of convolutions, normalization, and activation, designed to progressively refine input features, halve their size, and double their dimensions, thereby achieving efficient feature extraction. The MIPC-Block then purifies these features, emphasizing image-specific details for deeper analysis. An embedding layer adjusts feature dimensions for transformer layers, which address CNN limitations by capturing global information. Finally, the transformer's output is recombined and directed through skip connections to the decoder, ensuring comprehensive information retention and enhancing segmentation performance in a streamlined process.

By incorporating convolutional neural networks, transformer architecture, and Mutual Inclusion of Position and Channel, the encoder configuration ultimately attains robust feature extraction capabilities, resulting in synergistic strength.

3.4 Global-Skip-Connections

Within the framework of the U-shaped encoder-decoder architecture, skip connections are utilized to alleviate semantic discrepancies between encoder and decoder components. However, the optimization of skip connections remains an area in need of improvement. Primarily, there exist challenges such as loss of feature fidelity during transmission and insufficient overall integrity between the encoder and decoder. To address these issues, we employed two strategies: purifying the features transmitted via skip connections and augmenting skip connections with global information. These approaches facilitate the decoder in accurately restoring the original feature map, thereby significantly enhancing the model's segmentation capabilities. Here we call the entire skip connection part Global Mutual Inclusion Of Position With Channel-Skip-Connections (Global-Skip-Connections). It is divided into two parts: DA-Skip-Connections and Skip-Residue.

3.4.1 DA-Skip-Connections

Analogous to the conventional U-structured models (Ronneberger et al., 2015; Shi et al., 2020a), our approach utilizes traditional skip connections to diminish the semantic disparity between the encoder and decoder. We have incorporated Dual Attention Blocks (DA-Blocks) within all three skip connections to further narrow this gap, as illustrated in Figure 5. This enhancement stems from our observation that features conveyed through skip connections frequently harbor redundancies, which DA-Blocks are adept at filtering out, thereby refining the feature transmission process.

Integrating Dual Attention Blocks (DA-Blocks) into skip connections empowers the model to meticulously refine features relayed from the encoder, through the lens of image-specific positional and channel-based considerations. This process facilitates the extraction of more pertinent information while minimizing redundancy. Such an enhancement bolsters the model's robustness and significantly reduces the likelihood of overfitting, thereby contributing to superior performance and enhanced generalization capabilities.

3.4.2 Skip-Residue

Our approach differs from other U-Net models through the fine-tuning of decoder features and their strategic integration into the skip connections, as illustrated in Figure 2. The purple lines in the figure represent our added Skip-Residue, a skip connection that combines our custom MIPC-Block with the concept of capturing global information. This approach is motivated by the realization that, although encoder features are extensively leveraged via skip connections, decoder features often remain underexploited. By purifying decoder features before their integration into skip connections—thereby enhancing the restoration process of the original feature map—we facilitate a more profound utilization of decoder features.

Purifying features within the decoder, after three stages of upsampling, using Mutual Inclusion of Position and Channel (MIPC-Blocks) — oriented specifically towards image-relevant channels and positions — significantly elevates the quality of information. Subsequent transmission of these enhanced features to the skip connections, followed by their integration into the decoder, ensures the comprehensive utilization of decoder features. Incorporating the Skip-Residue module only in the top-level skip connection, rather than in every layer or other skip connections, allows for the effective utilization of the model's overall framework. This approach maximizes the extraction of valuable information while avoiding the potential negative effects of overemphasizing image features. This methodology effectively

minimizes redundancy between encoder and decoder, enriches feature depth, mitigates overfitting risks, and augments the model's image segmentation and generalization capabilities.

3.5 Decoder

As depicted in Figure 2, the diagram's right section represents the decoder. The decoder's fundamental task is to leverage features sourced from the encoder and those transmitted via skip connections. Through processes including upsampling, it endeavors to accurately reconstruct the original feature map.

The decoder architecture is structured around three pivotal elements: feature fusion, the segmentation head, and a series of three upsampling convolution blocks. Initially, feature fusion operates by amalgamating feature maps received through skip connections with current feature maps, thereby equipping the decoder to accurately reconstitute the original feature map. Subsequently, the segmentation head undertakes the task of adjusting the final output feature map back to its original dimensions. The final element comprises three upsampling convolution blocks, methodically increasing the size of the input feature map at each stage to adeptly reinstate the image's resolution.

Owing to the synergistic operation of these three components, the decoder showcases formidable decoding prowess. It adeptly harnesses features conveyed via skip connections as well as those derived from intermediate layers, enabling a proficient reconstruction of the original feature map.

4 Experiment

4.1 Datasets

In the Dataset section of our paper, we chose to conduct experiments on two distinct datasets: Synapse (Landman et al., 2015), ISIC 2018 (Codella et al., 2019; Tschandl et al., 2018) and Segpc (Gupta et al., 2021) for the following reasons:

Firstly, the Synapse dataset is among the most frequently utilized benchmark datasets in medical image segmentation, featuring segmentation tasks for eight different organs. This variety not only challenges but also demonstrates the generalization capabilities of our model across diverse anatomical structures.

Secondly, the selection encompasses both a 3D multi-class segmentation challenge (Synapse) and a 2D single-class segmentation task (ISIC 2018; Segpc). This combination allows us to evaluate our model's segmentation abilities from different perspectives, effectively showcasing its versatility and robustness in handling both complex three-dimensional data and simpler two-dimensional images.

This strategic choice of datasets underscores our commitment to validating the model's performance across a range of segmentation tasks, highlighting its potential for widespread application in medical image analysis.

4.1.1 Synapse

Under Institutional Review Board (IRB) supervision, 50 abdomen CT scans were randomly selected from a

combination of an ongoing colorectal cancer chemotherapy trial, and a retrospective ventral hernia study. After data processing, the Synapse dataset consists of 30^{- Δ CT} scan images of eight abdominal organs (Landman et al., 2015). Including left kidney, right kidney, aorta, spleen, gallbladder, liver, pancreas and stomach, A total of 3779 axial contrast-enhanced abdominal clinical CT images were obtained. In-plane resolution varies from 0.54 × 0.54 mm² to 0.98 × 0.98 mm², while slice thickness ranges from 2.5 to 5.0 mm.

4.1.2 ISIC-2018-task

ISIC-2018-dataset used in the 2018 ISIC Challenge addresses the challenges of skin diseases (Codella et al., 2019; Tschandl et al., 2018). It comprises a total of 2,512 images, with a file format of JPG. Where the ground truth data of the mask image is generated by several techniques and has been reviewed and curated by a specialized dermatologist. The images of lesions were obtained using various dermatoscopic techniques from different anatomical sites (excluding mucous membranes and nails). These images are sourced from historical samples of patients undergoing skin cancer screening at multiple institutions. Each lesion image contains only a primary lesion.

4.1.3 Segpc

This challenge targets robust segmentation of cells and is the first stage in building such tools for plasma cell cancers known as multiple myeloma (MM), a blood cancer. Provides images of stained colors normalized. The Segpc dataset (Gupta et al., 2021) contains a total of 298 images. Images are derived from microscope and camera shots.

4.2 Implementation settings

4.2.1 Baselines

In order to innovate in the field of medical image segmentation, we conducted benchmark testing of our proposed model against a series of well-regarded baselines, including U-Net, UNet++, Residual U-Net, Att-UNet, TransUNet, and MultiResUNet. U-Net has been a foundational model in the medical image segmentation domain (Ronneberger et al., 2015). UNet++ enriches the skip connections (Zhou et al., 2018). Residual U-Net integrates a single residual module into the U-Net model (Diakogiannis et al., 2020), while MultiResUNet incorporates multiple residual modules (Ibtehaz and Rahman, 2020). Att-UNet utilizes attention mechanisms to improve the weight of feature maps (Oktay et al., 2018). Finally, TransUNet integrates the Transformer architecture, establishing a new benchmark in segmentation accuracy (Chen et al., 2021). Through comprehensive comparisons with these renowned baselines, our objective is to highlight the unique advantages and wide-ranging potential applications of our proposed model. Additionally, we benchmarked our model against advanced models. UCTransNet allocates attention modules in the traditional U-Net model for skip connections (Wang et al., 2022a), while MISSFormer moves attention module allocation into a Transformer module-based U-shaped structure (Huang et al., 2022). TransNorm integrates Transformer modules into the encoder and skips standard U-Net connections (Azad et al., 2022). A novel

Transformer module was designed, and a model named MT-UNet was constructed with it (Wang et al., 2022b). Swin-UNet further enhances segmentation by extensively applying Swin-transformer modules (Cao et al., 2022). DA-TransUNet enhances model segmentation capabilities by using image feature location contracts (Sun et al., 2023). Through extensive comparisons with current state-of-the-art solutions, we aim to showcase its outstanding segmentation performance.

4.2.2 Implementation details

We implemented MIPC-Net using the PyTorch framework and trained it on a single NVIDIA RTX 3090 GPU (Paszke et al., 2019). The Transformer module we use employs the pre-trained model “R50-ViT”. The input resolution and patch size set to 224×224 and 16, respectively. We trained the model using the SGD optimizer, setting the learning rate to 0.01, momentum of 0.9, and weight decay of $1e-4$. The default batch size was set to 24. The loss function employed for dataset is defined as follows:

$$\text{Loss} = \frac{1}{2} \times \text{Cross-Entropy Loss} + \frac{1}{2} \times \text{DiceLoss}$$

4.2.3 Model evaluation

When evaluating the performance of MIPC-Net, we utilize a comprehensive set of metrics, including Dice Coefficient (DSC), and Hausdorff Distance (HD). These metrics are industry standards for computer vision and medical image segmentation and allow a multi-faceted assessment of a model’s accuracy, precision, and robustness.

AC(Accuracy): Accuracy is a widely used metric that assesses the overall correctness of a model’s predictions. It calculates the proportion of correctly predicted samples over the total number of samples. Accuracy gives a general idea of how well the model is performing across all classes.

$$AC = \frac{TP + TN}{TP + TN + FP + FN}$$

PR (Precision): Precision focuses on the accuracy of the positive predictions made by the model. Precision is the ratio of correctly predicted positive observations to the total predicted positives. High precision indicates that the model is good at not misclassifying negative instances as positive.

$$PR = \frac{TP}{TP + FP}$$

SP (Specificity): Specificity measures the accuracy of negative predictions made by the model. Specificity is the ratio of correctly predicted negatives to the total predicted negatives. A high specificity suggests that the model is effective at correctly identifying true negatives.

$$SP = \frac{TN}{TN + FP}$$

In summary, Accuracy provides an overall view of model performance, Precision emphasizes positive predictions’ accuracy, and Specificity assesses the accuracy of negative predictions.

The Dice coefficient (also known as Sørensen-Dice coefficient, F1-score, DSC) is a measure of model performance in image segmentation tasks and is particularly useful for dealing with

class imbalance problems. It measures the degree of overlap between prediction results and ground-truth segmentation results, and is particularly effective when dealing with object segmentation with unclear boundaries. The Dice coefficient is commonly used in image segmentation tasks as a measure of the accuracy of the model in the target area.

$$Dice(P, T) = \frac{|P_1 \cap T_1|}{|P_1| + |T_1|} \Leftrightarrow \text{Dice} = \frac{2|T \cap P|}{|F| + |P|}$$

Hausdorff distance (HD) is a distance metric used to measure the similarity between two sets and is often used to evaluate the performance of models in image segmentation tasks. It is particularly useful in the field of medical image segmentation, where it can quantify the difference between predicted and true segmentations, and is particularly convincing in evaluating boundary region segmentations. The calculation of the Hausdorff distance captures the maximum difference between the true and predicted segmentation results.

$$H(A, B) = \max \left\{ \max_{a \in A} \min_{b \in B} \|a - b\|, \max_{b \in B} \min_{a \in A} \|b - a\| \right\}$$

We use Dice and HD in the Synapse dataset, use the AC, PR, SP, Dice in the ISIC-2018-Task and Segpc datasets.

4.3 Comparison to the state-of-the-art methods

Our MIPC-Net has achieved promising results on the Synapse, ISIC 2018-Task, and Segpc datasets, showcasing its versatility and effectiveness across a range of medical image segmentation tasks. On the Synapse dataset, MIPC-Net excels in complex multi-organ 3D segmentation, significantly improving segmentation accuracy and boundary delineation compared to state-of-the-art models. Similarly, on the ISIC 2018-Task dataset, our model effectively handles the challenges of skin lesion segmentation, outperforming existing transformer-based models in multiple metrics such as Accuracy, Precision, and Specificity. Additionally, MIPC-Net demonstrates its ability to tackle simpler binary lesion segmentation tasks on the Segpc dataset, where it consistently achieves superior performance in accurately separating overlapping cells and dealing with low contrast. These results highlight the robustness and generalizability of MIPC-Net, making it a powerful solution for both complex and simpler segmentation tasks in medical segmentation.

4.3.1 Synapse

To evaluate the performance of our proposed MIPC-Net model, we conducted extensive experiments on the widely-used Synapse multi-organ segmentation dataset (Landman et al., 2015). We compared MIPC-Net with 12 state-of-the-art (SOTA) methods, including both CNN-based and transformer-based approaches, such as U-Net (Ronneberger et al., 2015), Res-UNet (Diakogiannis et al., 2020), TransUNet (Chen et al., 2021), U-Net++ (Zhou et al., 2018), Att-UNet (Oktay et al., 2018), TransNorm (Azad et al., 2022), UCTransNet (Wang et al., 2022a), MultiResUNet (Ibtehaz and Rahman, 2020), Swin-UNet (Cao et al., 2022), MT-UNet (Wang et al., 2022b), and DA-

TABLE 1 The experimental results on the Synapse dataset include the average Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) for each organ, as well as the individual DSC for each organ.

Model	mDSC		mHD	DSC of a single organ							
	Year	DSC↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
U-Net (Ronneberger et al., 2015)	2015	76.85%	39.70	89.07	69.72	77.77	68.6	93.43	53.98	86.67	75.58
U-Net++ (Zhou et al., 2018)	2018	76.91%	36.93	88.19	68.89	81.76	75.27	93.01	58.20	83.44	70.52
Residual U-Net (Diakogiannis et al., 2020)	2018	76.95%	38.44	87.06	66.05	83.43	76.83	93.99	51.86	85.25	70.13
Att-Unet (Oktay et al., 2018)	2018	77.77%	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
MultiResUNet (Ibtehaz and Rahman, 2020)	2020	77.42%	36.84	87.73	65.67	82.08	70.43	93.49	60.09	85.23	75.66
TransUNet (Chen et al., 2021)	2021	77.48%	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
UCTransNet (Wang et al., 2022a)	2022	78.23%	26.75	84.25	64.65	82.35	77.65	94.36	58.18	84.74	79.66
TransNorm (Azad et al., 2022)	2022	78.40%	30.25	86.23	65.1	82.18	78.63	94.22	55.34	89.50	76.01
MT-UNet (Wang et al., 2022b)	2022	78.59%	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
swin-unet (Cao et al., 2022)	2022	79.13%	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
DA-TransUNet (Sun et al., 2023)	2023	79.80%	23.48	86.54	65.27	81.70	80.45	94.57	61.62	88.53	79.73
MIPC-Net(Ours)		80.00%	19.32	87.30	66.43	83.24	80.37	94.48	59.45	89.20	79.55

The bold values indicate the best performance among all the methods compared in each respective evaluation metric. For each row in a table, the bold number represents the method that achieves the highest score or lowest error on that particular metric, demonstrating its superior performance relative to the other approaches.

TransUNet (Sun et al., 2023). The experimental results are presented in Table 1.

As shown in Table 1, MIPC-Net achieves the highest average Dice Similarity Coefficient (DSC) of 80.00% and the lowest average Hausdorff Distance (HD) of 19.32 mm among all the compared methods. This demonstrates the superior performance of MIPC-Net in both overall segmentation accuracy and boundary delineation precision. Compared to the popular transformer-based model TransUNet (Chen et al., 2021), MIPC-Net significantly improves the DSC by 2.52% and reduces the HD by 12.37 mm, highlighting the effectiveness of our proposed mutual inclusion mechanism and global integration strategy.

Moreover, MIPC-Net consistently outperforms TransUNet in terms of DSC for all eight individual organs, with improvements ranging from

0.07% to 4.12%. Notably, MIPC-Net achieves substantial DSC improvements of 3.29%, 3.35%, 3.59%, 4.12%, and 3.93% for the gallbladder, right kidney, pancreas, spleen, and stomach, respectively. These organs are known to be particularly challenging to segment due to their variable shapes, sizes, and locations, as well as their low contrast with surrounding tissues. The significant performance gains achieved by MIPC-Net demonstrate its strong capability in handling these difficult cases and accurately delineating organ boundaries.

Figure 6 provides a visual comparison of the DSC and HD values achieved by MIPC-Net and several other advanced models on the Synapse dataset. It is evident that MIPC-Net achieves the highest DSC and the lowest HD among all the compared models, further confirming its state-of-the-art performance in multi-organ segmentation.

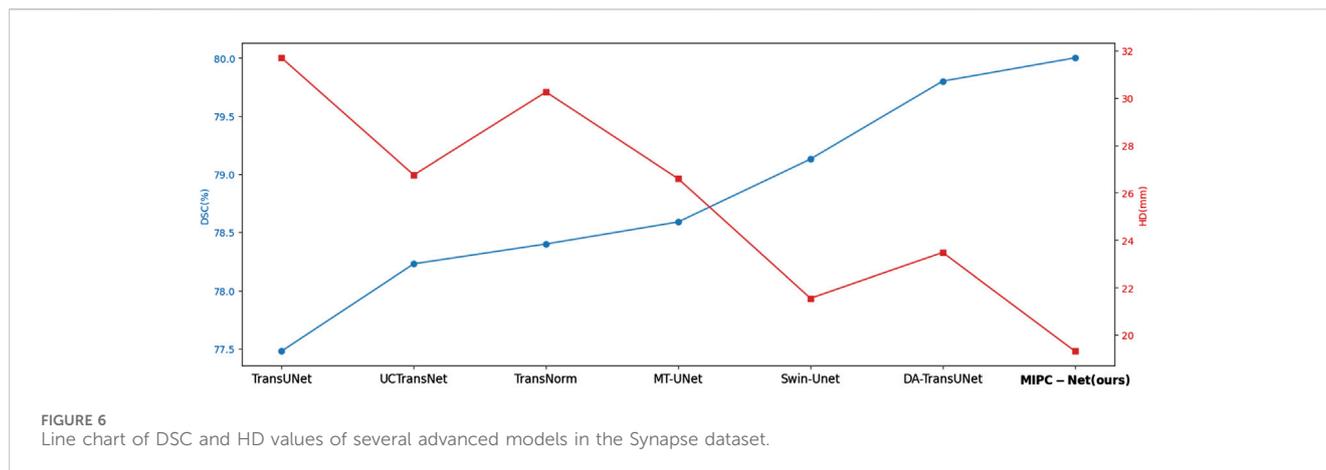


TABLE 2 The Hausdorff Distance (HD) for each organ in the Synapse dataset experimental results.

Model	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
TransUNet	14.94 mm	15.81 mm	59.92 mm	45.76 mm	37.86 mm	17.34 mm	43.33 mm	18.56 mm
swin-unet	8.64 mm	27.98 mm	41.83 mm	34.00 mm	22.17 mm	12.43 mm	9.90 mm	15.45 mm
DA-TransUNet	11.37 mm	27.93 mm	30.76 mm	48.93 mm	20.26 mm	12.29 mm	12.91 mm	23.37 mm
MIPC-Net(Ours)	8.63 mm	15.74 mm	41.65 mm	27.12 mm	22.33 mm	11.58 mm	12.09 mm	15.39 mm

The bold values indicate the best performance among all the methods compared in each respective evaluation metric. For each row in a table, the bold number represents the method that achieves the highest score or lowest error on that particular metric, demonstrating its superior performance relative to the other approaches.

To gain deeper insights into the boundary delineation performance of MIPC-Net, we also evaluated the HD metric for each individual organ, as shown in Table 2. MIPC-Net achieves the lowest HD for five out of eight organs, including the aorta, gallbladder, right kidney, pancreas, and stomach. Particularly, MIPC-Net significantly reduces the HD by 6.31 and 2.73 mm for the aorta compared to TransUNet and DA-TransUNet, respectively. These results highlight the superior boundary segmentation capability of MIPC-Net, which can be attributed to the effective integration of position and channel information through our proposed mutual inclusion mechanism.

It is worth noting that while MIPC-Net achieves state-of-the-art performance, its computational efficiency is comparable to that of TransUNet. The image segmentation time of MIPC-Net is 38.51 m, only slightly higher than TransUNet's 33.58 m. This indicates that the superior performance of MIPC-Net does not come at the cost of significantly increased computational overhead, making it a practical solution for real-world clinical applications.

Figure 7 presents a qualitative comparison of the segmentation results produced by TransUNet and MIPC-Net on the Synapse dataset. The regions highlighted by orange borders clearly demonstrate that MIPC-Net generates more accurate and precise segmentations compared to TransUNet, especially in challenging areas such as organ boundaries and small structures. The visual results further validate the effectiveness of our proposed approach in capturing fine-grained details and producing high-quality segmentation masks.

The experimental results clearly demonstrate that MIPC-Net outperforms existing models, in both segmentation accuracy and boundary delineation, particularly for challenging organs.

4.3.2 ISIC 2018-task dataset

To further validate the generalizability of MIPC-Net, we conducted experiments on the ISIC 2018 dataset (Codella et al., 2019; Tschandl et al., 2018) for skin lesion segmentation. This dataset presents unique challenges, such as varying lesion sizes, shapes, and color variations.

Table 3 compares MIPC-Net with several state-of-the-art models on the ISIC 2018 dataset. MIPC-Net achieves the highest Accuracy (AC) of 0.9560, Precision (PR) of 0.9279, and Specificity (SP) of 0.9831, demonstrating its superior performance in accurately segmenting skin lesions. Notably, MIPC-Net significantly outperforms the transformer-based model TransUNet, with improvements of 0.0108 in AC, 0.0453 in PR, 0.0178 in SP, and 0.0376 in Dice index. These improvements can be attributed to the effectiveness of our proposed mutual inclusion mechanism and global integration strategy in capturing both local and global contextual information.

Interestingly, while MIPC-Net achieves the highest AC, PR, and SP, its Dice index of 0.8875 is slightly lower than that of UCTransNet (0.8898). This suggests a potential trade-off between precision and recall, which could be further investigated in future work.

Figure 8 qualitatively compares the segmentation results of TransUNet and MIPC-Net on the ISIC 2018 dataset. MIPC-Net generates more precise and accurate segmentations, especially in

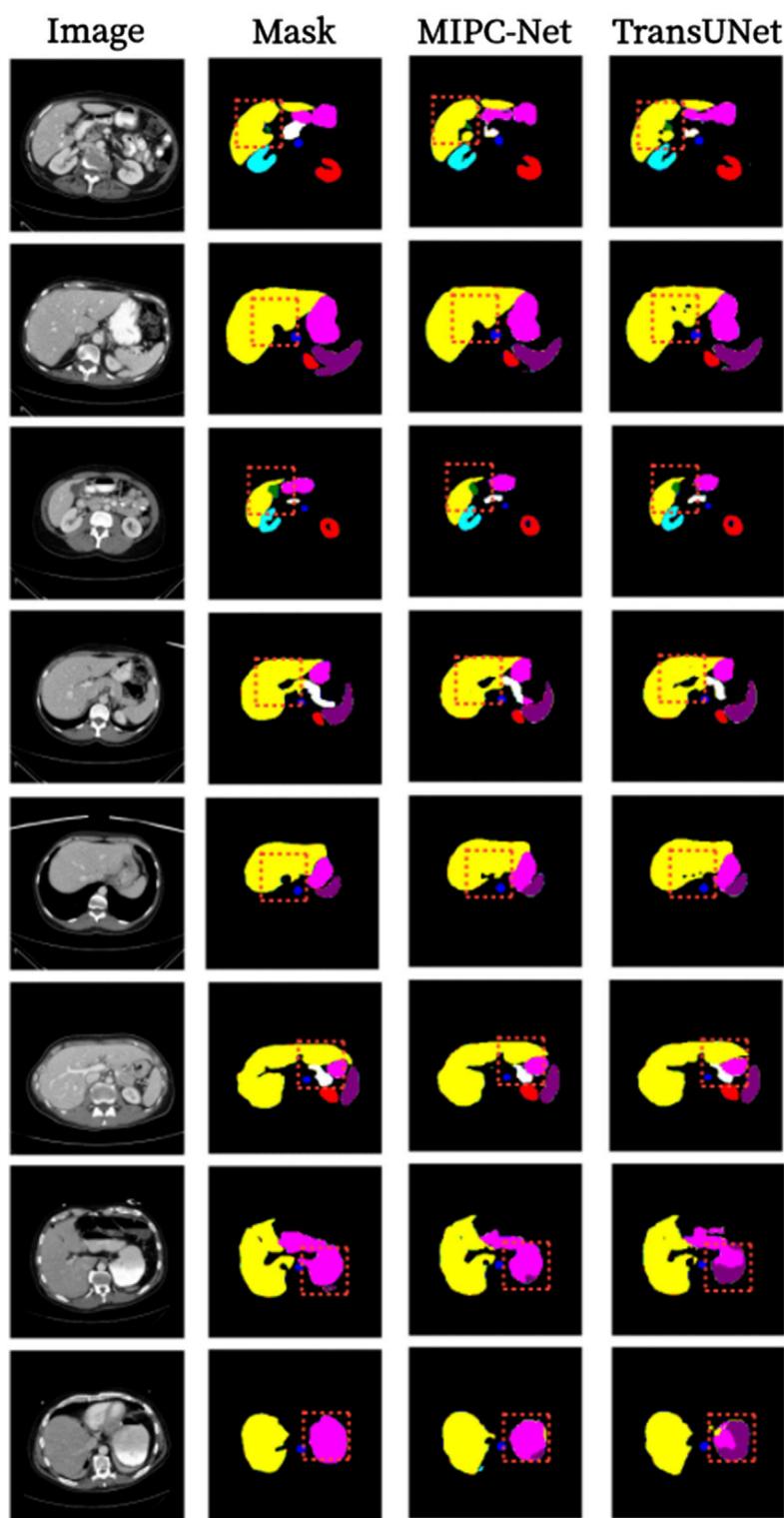


FIGURE 7
Segmentation results of TransUNet and MIPC-Net on the Synapse dataset.

challenging cases with irregular lesion boundaries and low contrast. The visual results further validate the superiority of our approach in capturing fine-grained details and producing high-quality segmentation masks for skin lesions.

4.3.3 Segpc dataset

We further assessed the performance of MIPC-Net on the Segpc dataset (Gupta et al., 2021) for cell segmentation in microscopy images. This dataset presents challenges such as overlapping cells,

TABLE 3 Experimental results on the ISIC2018-Task dataset.

Method	AC	PR	SP	Dice
U-Net (Ronneberger et al., 2015)	0.9446	0.8746	0.9671	0.8674
Att-UNet (Oktay et al., 2018)	0.9516	0.9075	0.9766	0.8820
U-Net++ (Zhou et al., 2018)	0.9517	0.9067	0.9764	0.8822
MultiResUNet (Ibtehaz and Rahman, 2020)	0.9473	0.8765	0.9704	0.8694
Residual U-Net (Diakogiannis et al., 2020)	0.9468	0.8753	0.9688	0.8689
TransUNet (Chen et al., 2021)	0.9452	0.8823	0.9653	0.8499
UCTransNet (Wang et al., 2022a)	0.9546	0.9100	0.9770	0.8898
MISSFormer (Huang et al., 2022)	0.9453	0.8964	0.9742	0.8657
MIPC-Net(ours)	0.9560	0.9279	0.9831	0.8875

The bold values indicate the best performance among all the methods compared in each respective evaluation metric. For each row in a table, the bold number represents the method that achieves the highest score or lowest error on that particular metric, demonstrating its superior performance relative to the other approaches.

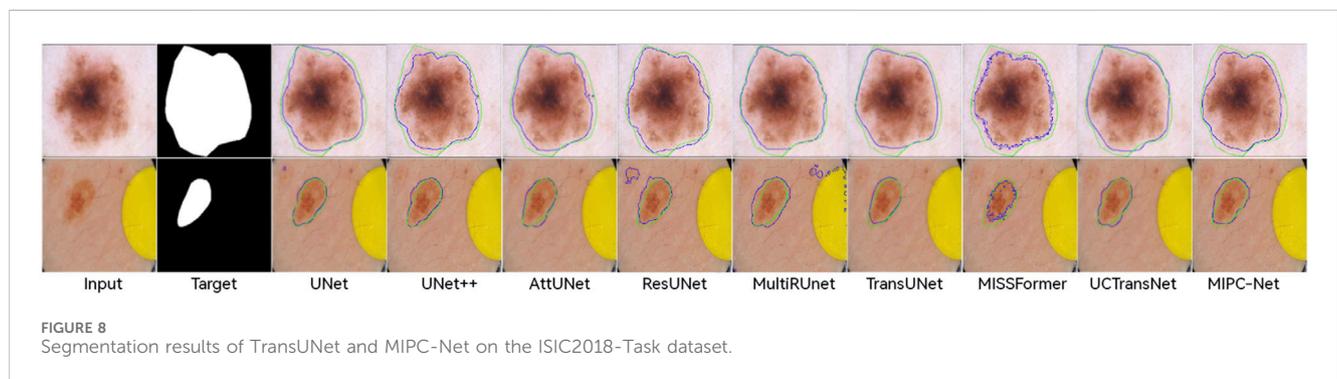


FIGURE 8 Segmentation results of TransUNet and MIPC-Net on the ISIC2018-Task dataset.

TABLE 4 Experimental results on the Segpc dataset.

Method	AC	PR	SP	Dice
Residual U-Net (Diakogiannis et al., 2020)	0.9733	0.8917	0.9871	0.8479
MultiResUNet (Ibtehaz and Rahman, 2020)	0.9753	0.8391	0.9834	0.8613
TransUNet (Chen et al., 2021)	0.9671	0.8598	0.9882	0.8005
MISSFormer (Huang et al., 2022)	0.9663	0.8152	0.9823	0.8082
DA-TransUNet (Sun et al., 2023)	0.9713	0.8789	0.9845	0.8366
MIPC-Net(ours)	0.9817	0.9079	0.9898	0.8675

The bold values indicate the best performance among all the methods compared in each respective evaluation metric. For each row in a table, the bold number represents the method that achieves the highest score or lowest error on that particular metric, demonstrating its superior performance relative to the other approaches.

variable cell sizes and shapes, and low contrast between cells and background.

Table 4 compares MIPC-Net with state-of-the-art models on the Segpc dataset. MIPC-Net consistently outperforms all compared methods, achieving the highest Accuracy (AC) of 0.9817, Precision (PR) of 0.9079, Specificity (SP) of 0.9898, and Dice index of 0.8675. Compared to TransUNet, MIPC-Net significantly improves performance across all metrics, with improvements of 0.0146 in

AC, 0.0481 in PR, 0.0016 in SP, and 0.067 in Dice index. These substantial improvements demonstrate the effectiveness of our approach in accurately separating overlapping cells and dealing with low contrast.

Notably, MIPC-Net achieves a significantly higher Dice index (0.8675) compared to all other methods, indicating a good balance between precision and recall when segmenting cells, which is crucial for accurate cell analysis and quantification.

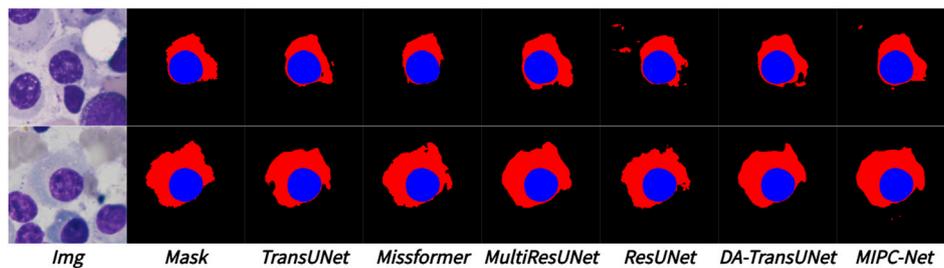


FIGURE 9 Segmentation results of TransUNet and MIPC-Net on the Segpc dataset.

TABLE 5 Effects of mutual inclusion of position and channel.

	Mutual inclusion	DSC \uparrow	HD \downarrow
PC-Net		79.09	23.34
MIPC-Net	√	80.00	19.32

The bold values indicate the best performance among all the methods compared in each respective evaluation metric. For each row in a table, the bold number represents the method that achieves the highest score or lowest error on that particular metric, demonstrating its superior performance relative to the other approaches.

Figure 9 visually compares the segmentation results of TransUNet and MIPC-Net on the Segpc dataset. MIPC-Net generates more accurate and precise segmentations, successfully separating individual cells and capturing their fine boundaries, even in dense cell clusters.

The strong performance of MIPC-Net on the ISIC 2018 and Segpc datasets, along with its state-of-the-art results on the Synapse dataset, highlights the versatility and generalizability of our approach across different medical image segmentation tasks and modalities.

4.4 Ablation study

To gain a deeper understanding of the effectiveness of the key components in our proposed MIPC-Net model, we conducted a comprehensive ablation study on the Synapse dataset. The study focused on three main aspects: the effects of mutual inclusion of position and channel, the impact of different configurations within the MIPC-Block, and the influence of the Skip-Residue in skip connections.

4.4.1 The effects of mutual inclusion of position and channel

As shown in Table 5, MIPC-Net, which incorporates the mutual inclusion mechanism, outperforms PC-Net by 0.91% in terms of DSC and achieves a reduction of 4.02 mm in HD. This improvement can be attributed to the effective integration of position and channel information through the mutual inclusion mechanism. By allowing the position and channel attention modules to interact and mutually guide each other, MIPC-Net is able to capture more comprehensive and discriminative features, leading to more accurate and precise segmentations. In contrast, simply using position and channel information independently, as in PC-Net, fails to fully exploit the potential synergies between these two types of information, resulting in suboptimal performance.

4.4.2 The effects of how to mix MIPC-Block internal mechanisms

Table 6 presents the results of different configurations within the MIPC-Block. The optimal configuration, where position attention (PAM) is used as the primary focus and channel attention (ChannelPool) as the auxiliary focus in Part A, and channel attention (CAM) is used as the primary focus and position attention (PositionPool) as the auxiliary focus in Part C, achieves the best performance with a DSC of 80.00% and an HD of 19.32 mm. This suggests that a balance between position and channel attention is crucial for achieving the best segmentation results. By employing different primary attention modules in Part A and Part C, the MIPC-Block is able to capture complementary information from both position and channel perspectives, leading to more comprehensive feature extraction. Furthermore, the results demonstrate that using PAM and CAM as the primary attention modules consistently outperforms using ChannelPool and PositionPool as the primary modules, indicating that the self-attention mechanisms employed in PAM and CAM are more

TABLE 6 Effects of how to mix MIPC-Block internal mechanisms.

	Part.A primary	Part.A auxiliary	Part.C primary	Part.A auxiliary	DSC \uparrow	HD \downarrow
MIPC-Net	PAM	ChannelPool	CAM	PositionPool	80.00	19.32
MIPC-Net	PAM	ChannelPool	PositionPool	CAM	78.87	21.55
MIPC-Net	ChannelPool	PAM	CAM	PositionPool	79.10	26.38
MIPC-Net	ChannelPool	PAM	PositionPool	CAM	79.11	24.27

The bold values indicate the best performance among all the methods compared in each respective evaluation metric. For each row in a table, the bold number represents the method that achieves the highest score or lowest error on that particular metric, demonstrating its superior performance relative to the other approaches.

TABLE 7 Effects of the Skip-Residue in skip connections.

	Skip-residue						
	1st	2nd	3rd	DA-Skip-Connections	Encoder with MIPC	DSC↑	HD↓
MIPC-Net				√	√	79.28	25.27
MIPC-Net	√			√	√	80.00	19.32
MIPC-Net		√		√	√	79.90	21.82
MIPC-Net			√	√	√	78.64	27.78
MIPC-Net	√	√	√	√	√	78.25	28.06
MIPC-Net						77.48	31.69

The bold values indicate the best performance among all the methods compared in each respective evaluation metric. For each row in a table, the bold number represents the method that achieves the highest score or lowest error on that particular metric, demonstrating its superior performance relative to the other approaches.

effective in capturing long-range dependencies and global contextual information.

4.4.3 The effect of the Skip-Residue in skip connections

Table 7 shows the impact of the Skip-Residue module on the overall performance of MIPC-Net. Adding the Skip-Residue module to the first skip connection layer alone achieves the best performance, with a DSC of 80.00% and an HD of 19.32mm, outperforming the baseline MIPC-Net without any Skip-Residue by 0.72% in terms of DSC and reducing the HD by 5.95 mm. This suggests that the Skip-Residue module is most effective when applied to the shallower skip connection layers, particularly the first layer, as it captures more low-level and spatial information crucial for accurate boundary delineation. The Skip-Residue module provides a direct path for the propagation of high-resolution spatial information from the encoder to the decoder, helping to preserve fine-grained details and improve localization accuracy. However, applying the Skip-Residue module to all skip connection layers leads to a significant performance drop, indicating that excessive use of the module can be counterproductive.

In conclusion, the ablation study demonstrates the importance of the mutual inclusion mechanism, the careful design of attention mechanisms within the MIPC-Block, and the strategic placement of the Skip-Residue module in skip connections. These components work together to capture comprehensive and discriminative features, leading to improved segmentation accuracy and precise boundary delineation in medical images.

4.5 Discussion

In this study, we found that the Mutual Inclusion of image-specific channels and positions can provide significant assistance for medical image segmentation tasks. The proposed MIPC-Block, based on the Mutual Inclusion mechanism, combined with Skip-Residue, further enhances the overall integration of the encoder and decoder. Our proposition has been validated through experiments on datasets, with the HD metric showing improvement to 2.23 mm compared to competing models on the Synapse dataset, demonstrating strong boundary segmentation capabilities.

Analyzing the ablation experiments validates the effectiveness of our proposed MIPC Block and Skip-Residue. Firstly, according to the experimental results presented in Tables 5, 6, we conclude that Mutual Inclusion of image feature positions and channels yields better performance compared to simple usage. Additionally, as demonstrated by the results in Table 7, the Skip-Residue module enhances the overall integrity of the encoder-decoder. We conclude that reducing the loss of effective features is of paramount importance when deeply exploring features.

Despite these advantages, our model has some limitations. Firstly, the introduction of MIPC-Block and DA-Blocks leads to an increase in computational complexity. This added cost may pose a barrier for real-time or resource-constrained applications. Furthermore, this approach combines feature positions and channels attention with the Vision Transformer in a parallel manner, without achieving deep integration between them, indicating potential areas for further research and enhancement. At the same time, exploring ways to enhance performance while reducing computational complexity is also an important direction to consider.

5 Conclusion

In conclusion, the proposed MIPC-Net represents a significant advancement in medical image segmentation, offering a powerful tool for precise boundary delineation. Inspired by radiologists' working patterns, our model integrates the Mutual Inclusion of Position and Channel Attention (MIPC) module and the Skip-Residue, a global residual connection, to effectively combine global and local features while focusing on abnormal regions and boundary details. The effectiveness of MIPC-Net is validated through extensive experiments on three publicly accessible datasets, outperforming state-of-the-art methods across all metrics and notably reducing the Hausdorff Distance by 2.23 mm on the Synapse dataset. The mutual inclusion mechanism and the Skip-Residue contribute to the model's superior performance by allowing for a more comprehensive utilization of image features and enhancing the restoration of medical images. The improved precision in boundary segmentation has the potential to significantly impact clinical practice, leading to more accurate diagnosis, treatment planning, and ultimately better patient care. Future work may focus on extending the application of MIPC-Net

to other medical imaging modalities, exploring its potential in tasks beyond segmentation, and incorporating domain-specific knowledge and multi-modal data to further enhance the model's performance and robustness.

In future work, we will further explore the integration of image-specific position and channel attention mechanisms with the self-attention mechanism of Transformers, aiming to enhance the model's ability to more effectively capture both local and global contextual information while improving the extraction of image-related features. Additionally, we will focus on optimizing the model's efficiency, striving to reduce computational complexity while enhancing its overall performance.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Landman et al., 2015; Codella et al., 2019; Tschandl et al., 2018; Gupta et al., 2021.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the patients/participants or patients/participants legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

YP: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing. JX: Conceptualization, Data curation, Investigation, Software, Supervision, Writing–review and editing. TY: Data curation, Methodology, Project administration, Supervision, Writing–review and editing. SL: Supervision, Visualization, Writing–review and editing. L-MN: Conceptualization, Data curation, Formal Analysis, Methodology, Validation, Writing–original draft, Writing - review and editing. TR: Conceptualization, Data curation, Methodology, Writing - original draft, Writing–review and editing. KL:

Funding acquisition, Resources, Writing–review and editing. GS: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Supervision, Validation, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was supported by multiple grants: The Soft Science Research Planning Project of Zhejiang Province (Grant No. 2024C35064) for the project “Study on Performance Evaluation and Optimization Path of Digital Aging Transformation Driven by User Experience”; the Online and Offline Hybrid First-Class Course “Computer Networks” of Zhejiang Province; the General Program of National Natural Science Foundation of China (Grant No. 82371484); and the Key Research and Development Program of Zhejiang Province (Grant Nos. 2021C03116 and 2022C03064); and Innovation and Entrepreneurship Training Program for College Students of Zhejiang Province (S202413023054).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Azad, R., Al-Antary, M. T., Heidari, M., and Merhof, D. (2022). Transnorm: transformer provides a strong spatial normalization mechanism for a deep segmentation model. *IEEE Access* 10, 108205–108215. doi:10.1109/access.2022.3211501
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv Prepr. arXiv:1409.0473*. doi:10.48550/arXiv.1409.0473
- Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2022). “Swin-UNET: unet-like pure transformer for medical image segmentation,” in *European conference on computer vision* (Springer), 205–218.
- Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*
- Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., et al. (2024). Transunet: rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* 97, 103280. doi:10.1016/j.media.2024.103280
- Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., et al. (2019). Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*
- Diakogiannis, F. I., Waldner, F., Caccetta, P., and Wu, C. (2020). Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogrammetry Remote Sens.* 162, 94–114. doi:10.1016/j.isprsjprs.2020.01.013
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*

- Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., et al. (2019). "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3146–3154.
- Gregor, K., Danihelka, I., Graves, A., Rezende, D., and Wierstra, D. (2015). "Draw: a recurrent neural network for image generation," in *International conference on machine learning* (Lille, France: PMLR), 1462–1471.
- Guo, C., Szemeyeni, M., Yi, Y., Wang, W., Chen, B., and Fan, C. (2021). "Sa-unet: spatial attention u-net for retinal vessel segmentation," in *2020 25th international conference on pattern recognition (ICPR)* (IEEE), 1236–1242.
- Gupta, A., Gupta, R., Gehlot, S., and Goswami, S. (2021). Segpc-2021: segmentation of multiple myeloma plasma cells in microscopic images. *IEEE Dataport* 1 (1). doi:10.1016/j.media.2022.102677
- Hosseinzadeh, M., and Wang, Y. (2021). "Image change captioning by learning from an auxiliary task," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2725–2734.
- Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., et al. (2020). "Unet 3+: a full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE), 1055–1059.
- Huang, X., Deng, Z., Li, D., Yuan, X., and Fu, Y. (2022). Missformer: an effective transformer for 2d medical image segmentation. *IEEE Trans. Med. Imaging* 42, 1484–1494. doi:10.1109/tmi.2022.3230943
- Ibtehaz, N., and Rahman, M. S. (2020). Multiresunet: rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural Netw.* 121, 74–87. doi:10.1016/j.neunet.2019.08.025
- Jamali, A., Roy, S. K., Li, J., and Ghamisi, P. (2023). Transu-net++: rethinking attention gated transu-net for deforestation mapping. *Int. J. Appl. Earth Observation Geoinformation* 120, 103332. doi:10.1016/j.jag.2023.103332
- Landman, B., Xu, Z., Igelsias, J. E., Styner, M., Langerak, T., and Klein, A. (2015). "Segmentation outside the cranial vault challenge," in *MICCAI: multi atlas labeling beyond cranial vault-workshop challenge*.
- Li, G., Jin, D., Yu, Q., and Qi, M. (2023). Ib-transunet: combining information bottleneck and transformer for medical image segmentation. *J. King Saud University-Computer Inf. Sci.* 35, 249–258. doi:10.1016/j.jksuci.2023.02.012
- Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., and Zhang, D. (2022). Ds-transunet: dual swin transformer u-net for medical image segmentation. *IEEE Trans. Instrum. Meas.* 71, 1–15. doi:10.1109/tim.2022.3178991
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.
- Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.
- Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv Prepr. arXiv:1508.04025*. doi:10.48550/arXiv.1508.04025
- Nam, H., Ha, J.-W., and Kim, J. (2017). "Dual attention networks for multimodal reasoning and matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 299–307.
- Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention u-net: learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: an imperative style, high-performance deep learning library. *Adv. neural Inf. Process. Syst.* 32. doi:10.48550/arXiv.1912.01703
- Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, october 5-9, 2015, proceedings, Part III 18* (Springer), 234–241.
- Shi, Z., Miao, C., Schoepf, U. J., Savage, R. H., Dargis, D. M., Pan, C., et al. (2020a). A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images. *Nat. Commun.* 11, 6090. doi:10.1038/s41467-020-19527-w
- Shi, Z., Miao, C., Schoepf, U. J., Savage, R. H., Dargis, D. M., Pan, C., et al. (2020b). A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images. *Nat. Commun.* 11, 6090. doi:10.1038/s41467-020-19527-w
- Sun, G., Pan, Y., Kong, W., Xu, Z., Ma, J., Racharak, T., et al. (2023). Da-transunet: integrating spatial and channel dual attention with transformer u-net for medical image segmentation. *arXiv Prepr. arXiv:2310.12570*. doi:10.3389/fbioe.2024.1398237
- Sun, G., Shu, H., Shao, F., Racharak, T., Kong, W., Pan, Y., et al. (2024). Fkd-med: privacy-aware, communication-optimized medical image segmentation via federated learning and model lightweighting through knowledge distillation. *IEEE Access* 12, 33687–33704. doi:10.1109/access.2024.3372394
- Tschandl, P., Rosendahl, C., and Kittler, H. (2018). The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. data* 5, 180161–180169. doi:10.1038/sdata.2018.161
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. neural Inf. Process. Syst.* 30. doi:10.48550/arXiv.1706.03762
- Wang, H., Cao, P., Wang, J., and Zaiane, O. R. (2022a). Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. *Proc. AAAI Conf. Artif. Intell.* 36, 2441–2449. doi:10.1609/aaai.v36i3.20144
- Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.-H., Chen, Y.-W., et al. (2022b). "Mixed transformer u-net for medical image segmentation," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE), 2390–2394.
- Wu, Y., Zhu, L., Yan, Y., and Yang, Y. (2019). "Dual attention matching for audio-visual event localization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 6292–6300.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., et al. (2015). "Show, attend and tell: neural image caption generation with visual attention," in *International conference on machine learning (PMLR)*, 2048–2057.
- Yang, Y., and Mehrkanoo, S. (2022). "Aa-transunet: attention augmented transunet for nowcasting tasks," in *2022 international joint conference on neural networks (IJCNN)* (IEEE), 01–08.
- Zhang, X., Zhao, J., Hao, J.-K., Zhao, X.-M., and Chen, L. (2015). Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic acids Res.* 43, e31. doi:10.1093/nar/gku1315
- Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). "Unet++: a nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, granada, Spain, september 20, 2018, proceedings 4* (Springer), 3–11.
- Zuo, Q., Shi, Z., Liu, B., Ping, N., Wang, J., Cheng, X., et al. (2024). Multi-resolution visual mamba with multi-directional selective mechanism for retinal disease detection. *Front. Cell Dev. Biol.* 12, 1484880. doi:10.3389/fcell.2024.1484880