China Instrument and Control Society
Instrumentation
INSTRUMENTATION, Vol. 12, No. 2, June 2025 https://doi.org/10.15878/j.instr.202500297

Article

BiCLIP-nnFormer: A Virtual Multimodal Instrument for Efficient and Accurate Medical Image Segmentation

Wang Bo[1], Yue Yan[2], Mengyuan Xu[3], Yuqun Yang[1*], Xu Tang[4], Kechen Shu[1], Jingyang Ai[5], Zheng You[1]

[1] Institute of Medical Equipment Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[2] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[3] School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China

[4] School of Artificial Intelligence, Xidian University, Xi'an 710071, China

[5] New York University, New York, NY 10003, USA

[*] Corresponding author email: yqunyang@hust.edu.cn

Abstract: Image segmentation is attracting increasing attention in the field of medical image analysis. Since widespread utilization across various medical applications, ensuring and improving segmentation accuracy has become a crucial topic of research. With advances in deep learning, researchers have developed numerous methods that combine Transformers and convolutional neural networks (CNNs) to create highly accurate models for medical image segmentation. However, efforts to further enhance accuracy by developing larger and more complex models or training with more extensive datasets, significantly increase computational resource consumption. To address this problem, we propose BiCLIP-nnFormer (the prefix "Bi" refers to the use of two distinct CLIP models), a virtual multimodal instrument that leverages CLIP models to enhance the segmentation performance of a medical segmentation model nnFormer. Since two CLIP models (PMC-CLIP and CoCa-CLIP) are pre-trained on large datasets, they do not require additional training, thus conserving computation resources. These models are used offline to extract image and text embeddings from medical images. These embeddings are then processed by the proposed 3D CLIP adapter, which adapts the CLIP knowledge for segmentation tasks by fine-tuning. Finally, the adapted embeddings are fused with feature maps extracted from the nnFormer encoder for generating predicted masks. This process enriches the representation capabilities of the feature maps by integrating global multimodal information, leading to more precise segmentation predictions. We demonstrate the superiority of BiCLIP-nnFormer and the effectiveness of using CLIP models to enhance nnFormer through experiments on two public datasets, namely the Synapse multi-organ segmentation dataset (Synapse) and the Automatic Cardiac Diagnosis Challenge dataset (ACDC), as well as a self-annotated lung multi-category segmentation dataset (LMCS).

Keywords: medical image analysis; image segmentation; CLIP; feature fusion; deep learning

Citation: Wang Bo, Yue Yan, Mengyuan Xu, Yuqun Yang, Xu Tang, Kechen Shu, Jingyang Ai, Zheng You. "BiCLIP-nnFormer: A Virtual Multimodal Instrument for Efficient and Accurate Medical Image Segmentation." Instrumentation 12, no. 2 (June 2025). https://doi.org/10.15878/j.instr.202500297

1 Introduction

Image segmentation plays a crucial role in the field of medical image analysis, which has been used in various practical applications, such as bone structure diagnosis[1,2], tumor detection[3,4] and multi-organ

segmentation[5,6]. As shown in Fig. 1, for 3-dimensional (3D) medical images, its goal is to generate the visual voxel-wise segmentation maps for the regions of interest[7], which can help physicians conduct accurate diagnosis, treatment and disease monitoring[8]. Therefore, ensuring and improving segmentation accuracy has become a hot topic, attracting increasing attention from researchers[9].

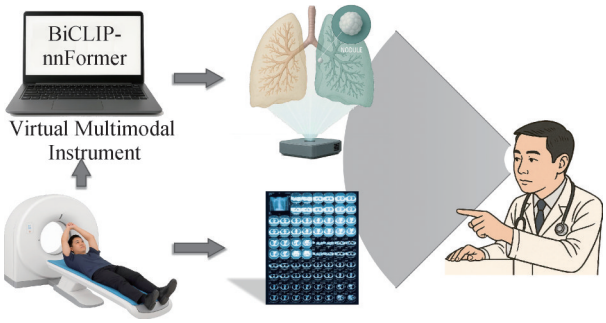

Fig.1 The workflow of 3D reconstruction for patient diagnosis. The process begins with a CT scan of the patient, generating a series of cross-sectional images. These images are processed by the BiCLIP-nnFormer model on a computer, which performs segmentation and 3D reconstruction of anatomical structures (such as lungs, vessels and nodule). The results are visualized and interpreted by clinicians to support accurate medical diagnosis and decision-making.

In recent years, deep learning[10] technology has achieved rapid development and demonstrated significant advantages in many fields, such as medical image analysis[11], natural language processing[12], and autonomous driving[13]. Thus, many deep learning-based methods have been proposed for medical image segmentation, achieving promising performance[14-17]. As a classical method, Unet[18] leverages a convolutional neural network (CNN) [19] to implement an encoder-decoder architecture with skip-connections. This structure allows it to extract hierarchical features from 3D medical images and decode them into precise voxel-wise segmentation results, leading to the proposal of many CNN-based UNet variants[16,20,21]. However, the limited receptive field of CNNs hinders further improvements in segmentation accuracy. To address this issue, researchers have introduced the Transformer[22] to capture long-term contextual information for analyzing medical images[23-25]. For instance, TransUNet[23] was developed to demonstrate the potential of the Transformer in the field of medical image segmentation. It relies on a CNN to extract main features from medical images for segmentation tasks, while the long-term contextual information provided by Transformer further enhances segmentation accuracy. Despite the impressive performance of these methods, the imbalanced use of CNNs and Transformers still influence their effectiveness negatively, as both techniques are quite capable. Therefore, Zhou et al. [17] introduced nnFormer for medical image segmentation, employing a hybrid stem that alternates between convolution and self-attention to effectively combine the strengths of both methods. In addition, to completely explore the potential of Transformer in medical image segmentation, several convolution-free methods have been introduced for medical image segmentation[26-28] that entirely eliminate the use of convolutions.

The aforementioned methods, through careful design and robust nonlinear fitting capabilities, achieve remarkable segmentation performance. However, they still often fall short in applications that require high precision. Moreover, given the high cost of computational resources, the continuous increase in model complexity and parameter scaling to enhance precision significantly complicates the training process, posing a significant challenge. Fortunately, the appearance of universal multimodal model[29,30], especially contrastive language-image pre-training (CLIP) [31,32], provides a potential solution for this dilemma. Under the training of large image-text dataset, CLIP model obtains the ability to perform well on a broad range of visual tasks with impressive accuracy, even in zero-shot scenarios[33]. Moreover, through innovative design, some CLIP variants[34,35] can perform the visual question answering (VQA) task, where they analyze an image and provide a corresponding text-based answer. For example, Yu et al. [35] introduced CoCa-CLIP, which features a visual encoder and a multimodal text decoder. Given an image, CoCa-CLIP can directly generate the corresponding text embedding, enabling the creation of text-based answers. Considering the slow progress of CLIP in medical image analysis due to data scarcity, PMC-CLIP[36] was proposed and trained on a large-scale biomedical image-text dataset. It achieves the outstanding performance in multiple downstream medical image analysis tasks.

Building on these advantages, integrating trained CLIP models with medical image segmentation frameworks presents a promising strategy for enhancing the performance of these models. This enhancement is driven by CLIP's pre-training on large-scale datasets, which provides it with robust generalization capabilities. Consequently, CLIP can accurately interpret a diverse range of images and provide corresponding global image or textual information. This information can help segmentation models better understand and interpret the context of medical images, leading to more accurate and effective segmentation.

Therefore, in this paper, we introduce a novel approach that combines PMC-CLIP and CoCa-CLIP with nnFormer for conducting medical image segmentation, named BiCLIP-nnFormer. Our proposed method outperforms nnFormer by integrating two CLIP models to gather additional global multimodal information (image and text). This information supplement leads to superior segmentation performance improvement. After obtaining 3D medical scans, BiCLIP-nnFormer first encodes them

into corresponding feature maps using the image encoder of nnFormer. Meanwhile, these 3D scans are input into PMC-CLIP and CoCa-CLIP to extract image and text embeddings, respectively. Then, a proposed 3D CLIP adapter processes these embeddings for fine-tuning. Finally, the feature maps are fused with the two embeddings from the adapter and fed into the image decoder of nnFormer for generating the predicted 3D mask. The three main contributions of this paper are as follows:

- To enhance the performance of nnFormer, we integrate PMC-CLIP and CoCa-CLIP with it to develop a novel virtual multimodal instrument for medical image segmentation, named BiCLIP-nnFormer. Compared to nnFormer alone, it achieves higher segmentation accuracy by fusing feature maps with image and text embeddings.
- We construct a self-annotated lung multi-category segmentation dataset, adding clinically significant labels including right lung, left lung, bronchus, nodule, pulmonary artery, and pulmonary vein. These annotations are crucial for tumor resection planning and serve as a valuable resource for future research.
- Extensive experiments are conducted to evaluate the effectiveness of BiCLIP-nnFormer on two public medical image segmentation datasets and a self-annotated dataset, including ablation study and comparison study.

The rest of this paper is recognized as follows. Section 2 introduces the related works of medical image segmentation and CLIP. The model details of BiCLIP-nnFormer will be explained in Section 3. In Section 4, the experimental results are shown for proving the effectiveness of BiCLIP-nnFormer. The conclusion of this paper is presented in Section 5.

2 Related Works

To provide a clear background for BiCLIP-nnFormer, the related works are categorized into three main sections: 1) CNN-based methods for medical image segmentation, 2) Transformer-based methods for medical image segmentation, and 3) CLIP models. In the first two sections, we review both CNN-based and Transformer-based segmentation methods. In the third section, we delve into various classic CLIP models and their variants for medical image analysis.

2.1 CNN-based Methods

Inspired by the classical UNet model[18], numerous CNN-based variants[20,37,38] have been developed and shown strong performance in medical image segmentation. 3D U-Net[39] adopts the classical U-Net architecture by replacing 2D operations with their 3D counterparts and incorporates efficient data augmentation during training. However, this operation requires replacing all 2D convolutions with 3D convolutions, which consumes large computational resources. To address this, H-DenseUNet[40] was proposed. It employs a 2D DenseUNet to capture intra-slice features and a 3D counterpart to hierarchically explore volumetric context information. To learn rich information from previous different layers, DenseRes-UNet[21] employs the dense blocks and residual connections, which effectively improve the representation capacity of features. Similarly, NucleiSegNet[41] also adopts a robust residual block for efficient extraction of high-level semantic features. Given the limitations of existing approaches that frequently require manual tuning, nnUNet[16] was developed as a self-configuring solution for various medical image segmentation tasks, covering preprocessing, network architecture, training, and post-processing. In addition to adopting UNet architecture, many self-designed models also achieve the promising segmentation results. For example, V-net[42] design a novel objective function based on dice coefficient, which can alleviate the negative influence caused by the number imbalance of foreground and background. Since traditional segmentation methods require complex inter-subject image registrations, Gibson et al.[43] proposed an automatic multi-organ segmentation method that uses deep learning to directly segment abdominal organs from CT images without registration, improving both accuracy and efficiency. In medical image segmentation, the objects of segmentation often exhibit significant variations in position, shape, and scale, presenting challenges for precise segmentation. Therefore, Gu et al.[37] proposed CA-Net that integrates various attention mechanisms within a CNN framework. It enhances medical image segmentation accuracy and interpretability by simultaneously recognizing the most crucial spatial positions, channels, and scales.

2.2 Transformer-based Methods

As a backbone capable of capturing long-term dependencies, Transformer attracts more and more attention in the field of medical image segmentation. Many methods involving Transformers appear and demonstrate more powerful performance compared to CNN-based methods. For instance, Chen et al.[23] introduced TransUNet, which initially explores the potential of Transformers in medical image segmentation. Since many methods employ deep networks and aggressive downsampling, issues like redundant parameters and loss of detail often arise. To alleviate this, Zhang et al.[25] proposed a parallel-in-branch structure called TransFuse, which combines CNNs and Transformers in parallel to simultaneously capture long-term dependencies and preserve spatial details. Due to the scarcity of labeled data in medical image analysis, efficiently training Transformers presents a challenge.

Therefore, a medical Transformer (MedT)[44] was proposed to improve training effectiveness. It leverages gated axial-attention to extend the existing architecture and introduces an additional control mechanism within the self-attention module. Moreover, self-supervised

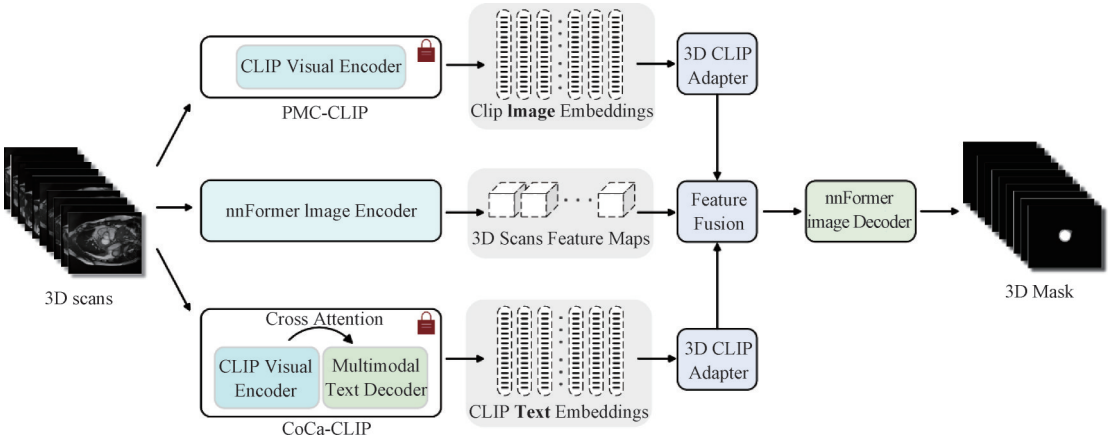

Fig.2 The framework of BiCLIP-nnFormer. After obtaining the 3D medical scans, they are inputted into PMC-CLIP, nnFormer image encoder and CoCa-CLIP to extract the image embeddings, 3D scans feature map and text embeedings, respectively. Then, feature maps are fused with embeddings from 3D CLIP adapter and inputted into nnFormer image decoder for generating predicted 3D mask.

learning (SSL) also is employed to address this issue. In the literature[45], LoGoNet utilizes a combination of masking and contrastive learning to implement SSL, significantly enhancing model performance in the absence of large labeled datasets. Although the patch mechanism in Transformers aids in capturing long-range relationships, it overlooks pixel-level structural information within single patch. Thus, Lin et al. [46] introduced the dual swin Transformer UNet (DS-TransUNet), which combines hierarchical swin Transformers to explore non-local dependencies and multiscale context information. Additionally, UNETR[24] was developed to reframe the 3D medical image segmentation task as a sequence-to-sequence prediction problem. Through capturing global multiscale information during sequence learning, it achieves enhanced segmentation accuracy and robustness.

2.3 CLIP Models

Through contrastive learning and training on large-scale datasets, CLIP[31] establishes a mapping relationship between images and text, which facilitates the creation of models that can understand the content of images and provide the corresponding textual descriptions. For example, CoCa-CLIP[35] can adopt multimodal text decoder to conduct VQA task. In addition to establishing relationships between images and text, grounded language-image pre-training (GLIP)[47], a variant of CLIP, extends its focus from whole images to specific objects. It successfully learns object-level, language-aware, and semantically rich representations. In order to conduct dense prediction task, e.g., segmentation, CLIPSeg[48] was presented, which can create one model for three segmentation tasks. Similarly, Wand et al. proposed a CLIP-driven referring image segmentation (CRIS) [49] method. It enhances feature consistency between the image and text modalities by propagating fine-grained semantic information from textual representations to pixel-level activations. Although CLIP and its variants play a crucial role in universal image analysis, they struggle to perform well in medical image analysis due to the domain gap. To address this issue, Zhang et al. [50] proposed PubMedCLIP, a version of CLIP fine-tuned for the medical domain. They chose the VQA task to demonstrate its effectiveness in medical image analysis. To enhance the utility of CLIP in medical contexts, PMC-CLIP[36] and FairCLIP[51] were introduced. These models are trained on large, proprietary medical datasets using the CLIP framework and have demonstrated significant performance improvements. Given the high costs of building large medical datasets for training, Zhang et al. [52] explored few-shot learning for the CLIP model in medical imaging, and introduced MediCLIP that employs SSL for efficient fine-tuning.

3 Materials and Methods

The framework of the proposed method BiCLIP-nnFormer is shown in Fig. 2, which consists of nnFormer, PMC-CLIP, CoCa-CLIP and corresponding 3D CLIP adapters. Here, nnFormer is divided into two components: the image encoder and the decoder. By integrating multimodal information from two CLIP models, the feature maps extracted by the encoder provide a comprehensive representation of 3D scans, resulting in more accurate segmentation results. Compared to the original CLIP adapter[53], the 3D CLIP adapter explores the intra-relationships among slices to more effectively represent the volumetric information of 3D medical scans.

3.1 3D CLIP Adapter

Before introducing the 3D CLIP adapter, let us explain how to generate CLIP embeddings. In BiCLIP-nnFormer, the input of PMC-CLIP and CoCa-CLIP is a 3D scan $\mathcal{X} \in \mathbb{R}^{H \times W \times D}$, where H, W and D represent the height, width and depth, respectively. However, these two

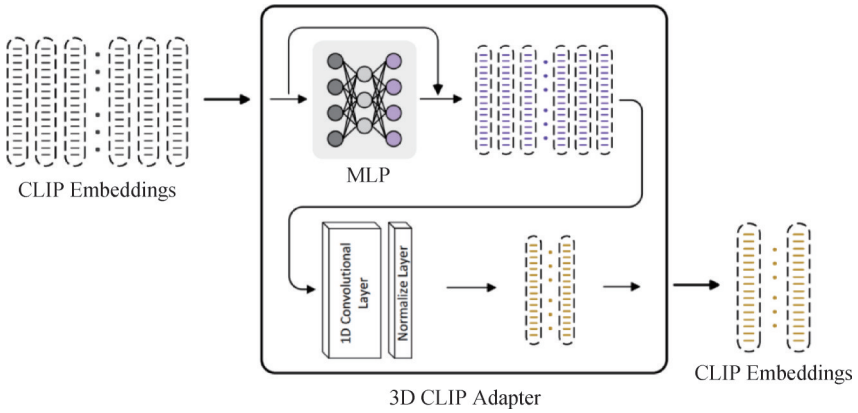

Fig.3 The structure of 3D CLIP adapter. First, the obtained CLIP embeddings are fed into the MLP with a residual connection, serving as the original CLIP adapter. Next, a 1D convolutional layer is employed across the embeddings to capture the volumetric information from the 3D medical scan. Finally, a normalization layer is used to ensure that the output remains standardized.

CLIP models only can process 2D image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$. Thus, we separate a scan to D slices along the depth dimension and replicate each slice three times, i.e.,

$$\mathcal{X} \rightarrow \{I_1, I_2, ..., I_D | I_j \in \mathbb{R}^{H\times W\times 3}\}. \quad (1)$$

Thus, $I_{j,j\in[1,D]}$ can be inputted into PMC-/CoCa-CLIP models to generate D image/text embeddings, each with a channel number C_e of 768. In summary, each CLIP model can extract embeddings $\in \mathbb{R}^{D\times 768}$ from $\mathcal{X}$. Then, two 3D CLIP adapters, identical in structure but differing in parameters, are applied to two groups of embeddings from two CLIP models. The goal of CLIP adapter is to adapt the knowledge from CLIP model to downstream tasks by fine-tuning it. However, original CLIP adapter[53] are designed for 2D images, which do not consider the depth information crucial for 3D scans. Therefore, we propose the 3D CLIP adapter for 3D medical segmentation tasks. This new adapter builds on the original design while also focusing on capturing the relationships between different embeddings to enhance the fine-tuning of CLIP for 3D segmentation tasks. Next, we will use the embeddings from PMC-CLIP as an example to introduce the 3D CLIP adapter. The process for the CoCa-CLIP embeddings is the same.

After obtaining the PMC-CLIP embeddings, they are firstly processed by a multilayer perceptron (MLP) with a residual connection, corresponding to the original CLIP adapter, as illustrated in the top half of Fig. 3. Here, the MLP is composed of two fully connected layers. Subsequently, we add a component consisting of a 1D convolutional layer and a normalization layer. The input channel number for the convolution is D, and its output channel number is aligned with the depth of feature maps from the nnFormer image encoder to facilitate the following feature fusion. Here, using 1D convolution help integrating information among embeddings to new ones. This not only more accurately capture the volumetric information but also adapt better to 3D spatial changes when affine transformations are used for data augmentation. It is important to note that the extraction of CLIP embeddings is performed offline. Consequently, these pre-extracted embeddings do not directly correspond to the slices of an augmented 3D medical scan. Therefore, it becomes crucial to use the 3D CLIP adapter to effectively explore volumetric information.

3.2 Feature Fusion

By fine-tuning the 3D CLIP adapter, the knowledge from two CLIP models can be effectively adapted to conduct the 3D medical image segmentation task. Therefore, by fusing the feature maps obtained from the nnFormer encoder with the embeddings of PMC-CLIP and CoCa-CLIP, the global representation of feature maps for 3D medical scans can be improved. To achieve this, a feature fusion strategy is proposed, as illustrated in Fig. 4. This strategy primarily involves the creation of CLIP features, concatenation of features, and the application of a global self-attention layer.

Since the shapes of 3D feature maps $\in \mathbb{R}^{H\times W\times D\times C}$ and CLIP embeddings $\in \mathbb{R}^{D\times C}$ are not consistent, the CLIP feature creation process (see the right side of Fig. 4) is designed to merge the embeddings from two CLIP models, and align their shapes with 3D feature maps. Here, C denotes the channel number. First, the embeddings $\in \mathbb{R}^{D\times C}$ are expanded to $\mathbb{R}^{1\times 1\times D\times C}$. Second, the first two dimensions 1×1 are repeated to $H\times W$ for realizing feature creation, resulting in CLIP feature maps $\in \mathbb{R}^{H\times W\times D\times C}$. Finally, the feature maps created from PMC-CLIP and CoCa-CLIP are combined at the pixel level to produce the final CLIP feature maps.

Next, the feature maps of 3D scans and those from the CLIP models are concatenated and fed into an MLP for feature fusion. Given the rich information provided by the two CLIP models, this can significantly aid in further enhancing features. Therefore, the fused feature maps are processed through two global self-attention layers. The use of a large receptive field allows these layers to highlight important information beneficial for the 3D segmentation task, while effectively suppressing redundant information. Note that the structure of the two global self-attention layers is identical to that of the last layer in the bottleneck module of nnFormer. Since they

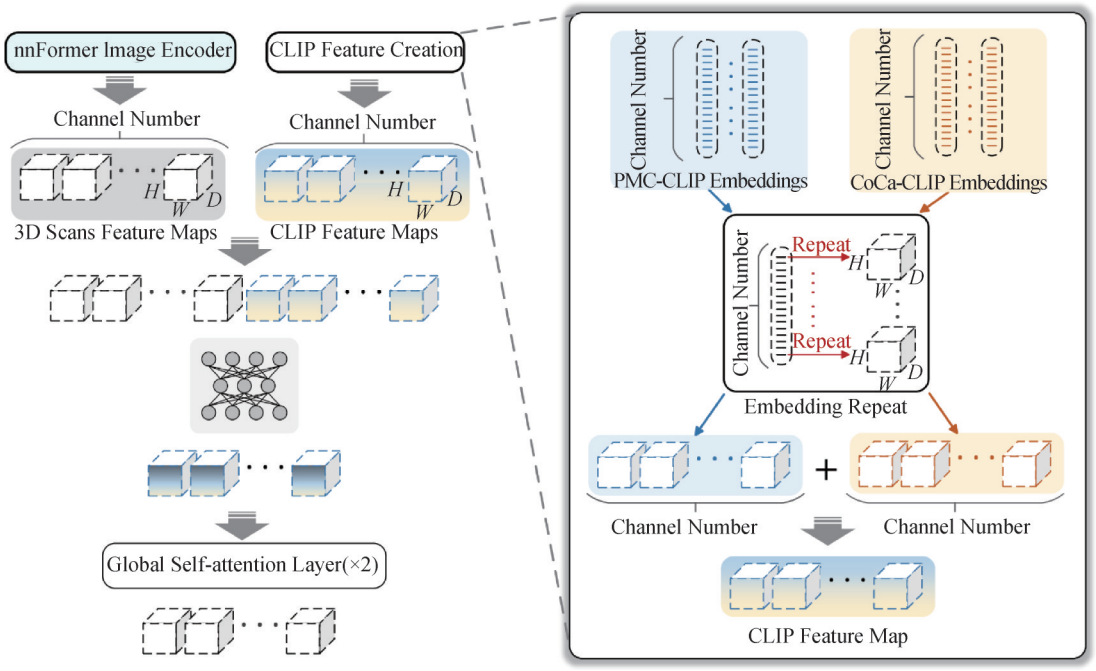

Fig.4 The diagram for describing how to fuse CLIP embedding and the feature maps of 3D scans. This process involves the creation of CLIP feature maps, concatenation of features, and the application of a global self-attention layer.

serve a similar function, utilizing a large receptive field to capture long-range dependencies, we load the same initial pre-training parameters into them, though they are not shared.

4 Experimental Results

4.1 Datasets

To evaluate the effectiveness of BiCLIP-nnFormer in utilizing CLIP to improve the performance of nnFormer, we select two public datasets: synapse multi-organ segmentation (referred to as Synapse) and automatic cardiac diagnosis challenge (ACDC), and a self-annotated lung multi-category segmentation (LMCS) dataset. The visual samples are shown in Fig. 5 and 6.

Synapse. It comprises 30 abdominal CT scans, divided into 18 for training and 12 for testing, as used in[23]. Originating from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge, it includes 3779 axial

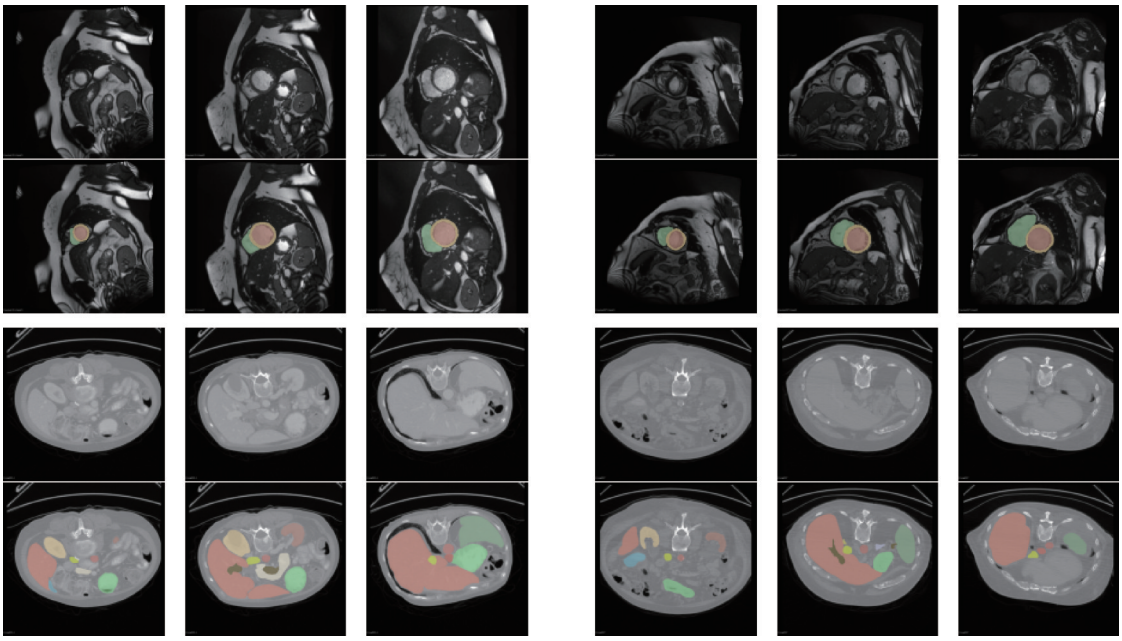

Fig.5 The samples shown include slices and their corresponding segmentation labels. The first and third rows represent the slices, while the second and fourth rows show the segmentation labels. The samples in the first two rows are selected from the ACDC dataset, and those in the last two rows are from the Synapse dataset, each set featuring two patients.

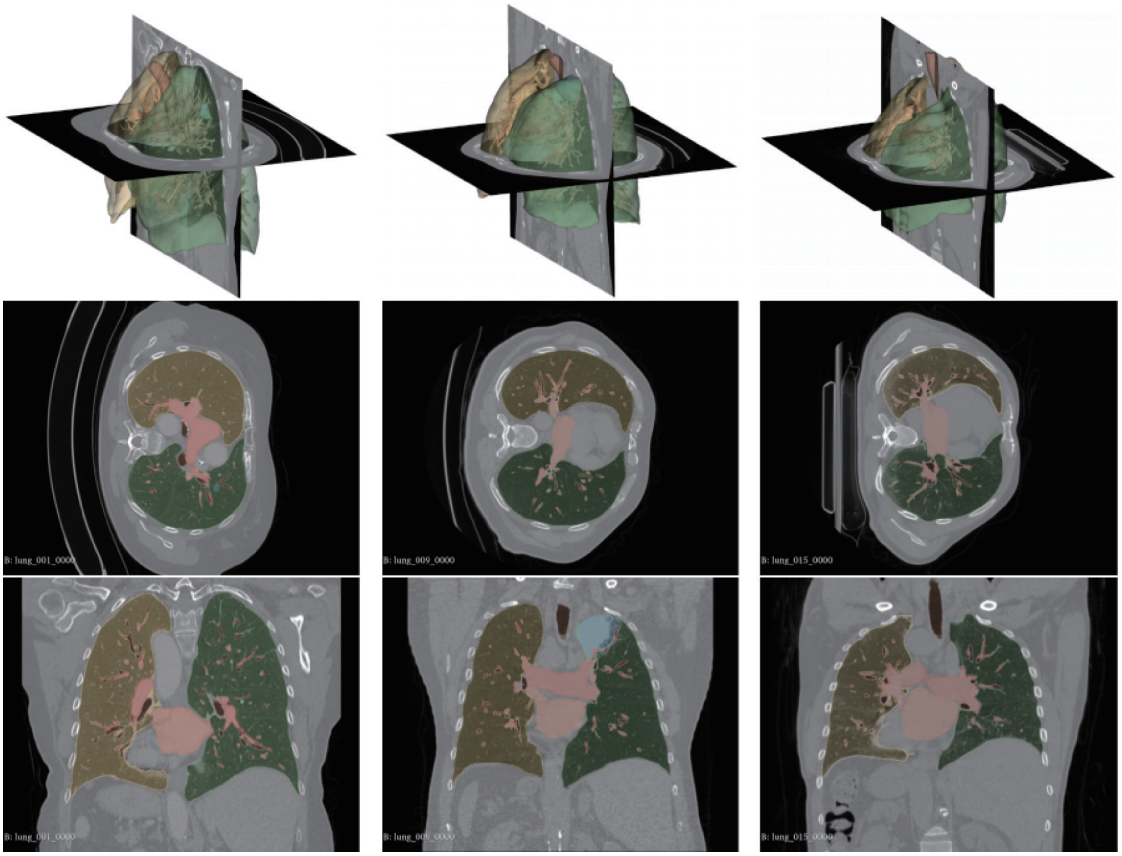

Fig.6 The samples shown include slices and their corresponding segmentation labels of LMCS dataset.

contrast-enhanced abdominal clinical CT slices. Each scan in the dataset contains between 85 and 198 slices, each slice with image shape of 512×512 pixels and a voxel spatial resolution of ([0.54 ~ 0.54] × [0.98 ~ 0.98] × [2.5 ~ 5.0]) mm^3. Performance is assessed on eight abdominal organs: aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas and stomach.

ACDC. The dataset comprises cardiac magnetic resonance imaging (MRI) images sourced from 100 different patients, divided into training, validation, and testing subsets containing 70, 10, and 20 patients, respectively. These images capture a series of short-axis slices that span from the base to the apex of the left ventricle, with each slice measuring between 5 and 8 *mm* in thickness. The spatial resolution of these short-axis images ranges from 0.83 to 1.75 mm^2/pixel. This dataset targets anatomical regions including the cavities of the right and left ventricles, as well as the myocardium of the left ventricle, and provides annotations for the left ventricle (LV), right ventricle (RV), and myocardium (MYO) for segmentation.

LMCS. In addition to using public datasets, we further extended the MSD Task06 Lung dataset by performing additional manual annotations on anatomical and pathological structures relevant to lung tumor resection. While the original dataset provides tumor annotations, we annotated total six categories that are crucial for surgicalplanning and intraoperative navigation: right lung, left lung, bronchus, pulmonary nodule, pulmonary artery, and pulmonary vein. These structures play vital roles in assessing tumor resectability, determining surgical margins, and avoiding critical vasculature during procedures. A total of 64 patient cases were annotated, with the number of slices per scan ranging from 211 to 636 (median: 271.0). The voxel spacing varies across scans: Z-axis (slice thickness) ranges from 0.62 mm to 1.25 mm (median: 1.24 mm), while the X and Y in-plane resolutions range from 0.60 mm to 0.98 mm (median: 0.80 mm). The dataset was randomly divided into training (38 cases), validation (13 cases), and test (13 cases) subsets.

4.2 Implementation Details

● *Experimental Settings.* Building on the original nnFormer, the proposed BiCLIP-nnFormer incorporates modifications on Python 3.9, Pytorch 1.10, and Ubuntu 22.04. All experiments are conducted on a high-performance computer equipped with four NVIDIA GeForce RTX 3090 GPUs, each with 24 GB of memory, and two E5-2667 v4 CPUs. Consistent with the original nnFormer settings, we utilize the SGD optimizer with an initial learning rate of 0.01, a weight decay of 3e-5, and a momentum of 0.99. Observing that the best-performing model typically emerges within 500 epochs, we have capped the maximum number of epochs for model optimization at 500. Model parameters are initialized using the provided pre-trained configurations of nnFormer. For the Synapse and ACDC datasets, the batch

sizes are set at 32 and 16, respectively. The other hyperparameters remain consistent with those used in nnFormer. The 3D CLIP adapter is mainly equipped with an MLP and a 1D convolutional layer. Specifically, the MLP is configured with dimensions of 768×192×1536 for the Synapse dataset and 768×192×768 for the ACDC dataset. Additionally, the convolutional layer features 64/16 output channels and 4/3 input channels for the Synapse/ACDC dataset.

● *Data augmentation.* During the training and validation process, all images are initially resampled to uniform spacing. To enhance the robustness of the models, a series of augmentations are systematically applied in the training stage. These include rotation, scaling, the addition of Gaussian noise, application of Gaussian blur, adjustments to brightness and contrast, simulation of low resolution, gamma corrections, and mirroring.

● *Assessment Criteria.* In this paper, we employ two metrics to quantitatively assess the performance of the BiCLIP-nnFormer and other methods: the dice similarity coefficient (DSC) and the 95% hausdorff distance (HD95). The DSC measures the extent of overlap between the segmentation labels and the predicted masks. A higher DSC indicates greater accuracy of the model. Its definition is as follows,

$$DSC(L,P)=2\times\frac{|L\cap P|}{|L|\cup|P|}=2\times\frac{L\cdot P}{L^2+P^2}, \tag{2}$$

where L and P denote the label and prediction. HD95 acts as a boundary-focused metric, specifically evaluating the 95th percentile of distances between the boundaries of predicted mask and the corresponding segmentation labels. It is calculated using the formula:

$$HD95(L,P)=max(D_{L\to P},D_{P\to L}), \tag{3}$$

where$D_{L\to P}$ represents the maximum 95th percentile distance from the label voxels to the prediction, and $D_{P\to L}$ indicates the maximum 95th percentile distance from the prediction voxels to the label.

4.3 Comparison Study

In this section, we compare the proposed BiCLIP-nnFormer with 12 other methods to demonstrate its effectiveness, including VIT+CUP[23], R50-VIT+CUP[23], TransUNet[23], SwinUNet[28], TransClaw UNet[54], LeVit-UNet-384s[55], MISSFormer[56], CoTr[57], UNETR[24], Swin UNETR[58], nnFormer[17] and SegFormer3D[59]. The results of LMCS, Synapse and ACDC datasets are listed in Table 1, 2 and 3, respectively, with the best results highlighted in bold.

Table 1 The comparative results of different methods on LMCS dataset. Here, "Pul." Denotes pulmonary.

Methods	Average	Lung (R)	Lung (L)	Bronchi	Nodules	Pul. Artery	Pul. Vein
R50-VIT+CUP[23]	60.83	94.84	94.70	48.51	22.23	50.95	53.75
VIT+CUP[23]	63.99	94.10	93.82	49.51	39.05	50.65	56.81
SwinUNet[28]	71.25	96.55	96.49	62.18	41.77	63.96	66.57
SegFormer3D[59]	74.93	93.99	93.54	63.73	59.44	68.32	70.55
TransUNet[23]	75.06	**96.91**	**96.72**	66.25	46.25	70.65	73.56
SwinUNETR[28]	75.30	93.13	92.78	69.68	52.13	70.65	73.45
UNETR[24]	76.37	94.70	94.28	71.96	54.41	70.10	72.76
nnFormer[17]	81.84	96.30	96.19	73.77	**66.00**	78.32	80.47
Ours	**83.64**	95.70	95.44	**79.53**	64.78	**82.27**	**84.11**

For the LMCS dataset, BiCLIP-nnFormer achieves the best overall performance with an average Dice score of 83.64%, outperforming all baseline methods. Compared to the strong baseline nnFormer, our method improves the average Dice by 1.80%, demonstrating the effectiveness of integrating CLIP-derived multimodal features. Across individual anatomical structures, BiCLIP-nnFormer consistently achieves the highest Dice scores on three target classes: bronchus (79.53%), pulmonary artery (82.27%), and pulmonary vein (84.11%), while maintaining competitive performance on the remaining classes. Particularly noteworthy is the significant improvement on smaller and more challenging structures such as the bronchus and pulmonary artery, with gains of 5.76% and 3.95%, respectively, over nnFormer, and up to 37.50% improvement over early Transformer baselines such as R50-ViT+CUP. These results validate the capacity of BiCLIP-nnFormer to leverage global semantic priors for better localization and boundary delineation, especially in anatomically complex or low-contrast regions. The consistent gains across public and private datasets confirm the generalizability and robustness of the proposed virtual multimodal instrument.

For the Synapse dataset, the proposed BiCLIP-nnFormer achieves state-of-the-art performance in two comprehensive metrics, DSC (86.80%) and HD95 (8.82). Compared to nnFormer, the DSC has improved by 0.22%, and there is a significant reduction of 1.81 in

Table 2 The comparative results of different methods on synapse dataset. Here, "Gall." and "Kid." denote gallbladder and kidney.

Methods	HD95 ↓	DSC ↑	Aorta	Gall.	Kid.(L)	Kid.(R)	Liver	Pancreas	Spleen	Stomach
VIT+CUP[23]	36.11	67.86	70.19	45.10	74.70	67.40	91.32	42.00	81.75	70.44
R50-VIT+CUP[23]	32.87	71.29	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet[23]	31.69	77.48	87.23	63.16	81.87	77.02	94.08	55.86	85.08	75.62
TransClaw UNet[54]	26.38	78.09	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
LeViT-UNet-384s[55]	16.84	78.53	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76
SwinUNet[28]	21.55	79.13	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
UNETR[24]	22.97	79.56	89.99	60.56	85.66	84.80	94.46	59.25	87.81	73.99
CoTr[57]	19.15	80.78	85.42	68.93	85.45	83.62	93.89	63.77	88.58	76.23
MISSFormer[56]	18.20	81.96	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
Swin UNETR[58]	14.78	83.51	90.75	66.72	86.51	85.88	95.33	70.07	**94.59**	78.20
nnFormer[17]	10.63	86.57	**92.04**	70.17	86.57	86.25	**96.84**	**83.35**	90.51	**86.83**
Ours	**8.82**	**86.80**	89.96	**72.44**	**87.74**	**86.89**	96.10	82.45	92.11	86.70

Table 3 The comparative results of different methods on ACDC dataset.

Methods	Average	Left Ventricle	Right Ventricle	Myocardium
ViT-CUP[23]	81.45	81.46	70.71	92.18
R50-ViT-CUP[23]	87.57	86.07	81.88	94.75
UNETR[24]	88.61	85.29	86.52	94.02
TransUNet[23]	89.71	88.86	84.54	95.73
SwinUNet[28]	90.00	88.55	85.62	95.83
LeViT-UNet-384s[55]	90.32	89.55	87.64	93.76
SegFormer3D[59]	90.96	88.50	88.86	95.53
nnFormer[17]	92.06	90.94	89.58	95.65
Ours	**92.49**	**91.66**	**89.68**	**96.13**

HD95. Furthermore, it is important to highlight that BiCLIP-nnFormer demonstrates a significant performance improvement in DSC and HD95 compared to other methods. For instance, improvements are noted as 3.29%/5.96 over Swin UNETR, 7.24%/14.15 over UNETR, 6.02%/10.33 over CoTr, 4.84%/9.38 over MISSFormer, and 8.27%/8.02 over LeVit-UNet-384s. These results not only show improved overlap between the predicted mask and the segmentation label but also better consistency in the prediction boundary with the segmentation label. Additionally, integrating two CLIP models has boosted nnFormer's performance across four specific organs: the gallbladder, left kidney, right kidney, and spleen. Notably, the performance on the gallbladder achieved the highest value, significantly outperforming other methods.

For ACDC dataset, BiCLIP-nnFormer obtains the highest results on all metrics, including "Average" (92.49%), "LV" (91.66%), "RV" (89.68%) and "MYO" (96.13%). Compared to nnFormer, improvements in "Average", "LV", "RV" and "MYO" are 0.43%, 0.72%, 0.10%, and 0.48%, respectively. The excellent results and clear improvements with nnFormer highlight how BiCLIP-nnFormer advances performance and proves that using CLIP models can effectively enhance segmentation models.

4.4 Ablation Study

BiCLIP-nnFormer consists of three components: nnFormer, PMC-CLIP, and CoCa-CLIP. To accurately analyze the contribution of each CLIP model, we have developed four network variations, from Net_1 to Net_4. Net_1 is the basic nnFormer. Net_2 integrates PMC-CLIP with nnFormer, and Net_3 incorporates CoCa-CLIP with nnFormer. Net_4, which combines both CLIP models, represents the complete BiCLIP-nnFormer setup.

We conduct ablation experiments on two datasets, ACDC and LMCS, with results summarized in Tables 4 and 5, respectively. On the ACDC dataset, both Net_2 and Net_3 gain improvements in the average Dice score compared to the baseline (Net_1), indicating that incorporating either CLIP model contributes positively to segmentation accuracy. Notably, Net_2 achieves better performance in left ventricle (LV) segmentation, while Net_3 performs slightly better in myocardium (MYO) segmentation. When both CLIP models are integrated in Net_4, the model achieves the best results across all metrics, confirming the complementary nature of PMC-CLIP and CoCa-CLIP.

To evaluate the impact of integrating two CLIP models on inference efficiency, we measured the average inference time per volume across four networks, with a batch size of 1 using a single NVIDIA 3090 GPU. As shown in the results, the inference time on the LMCS

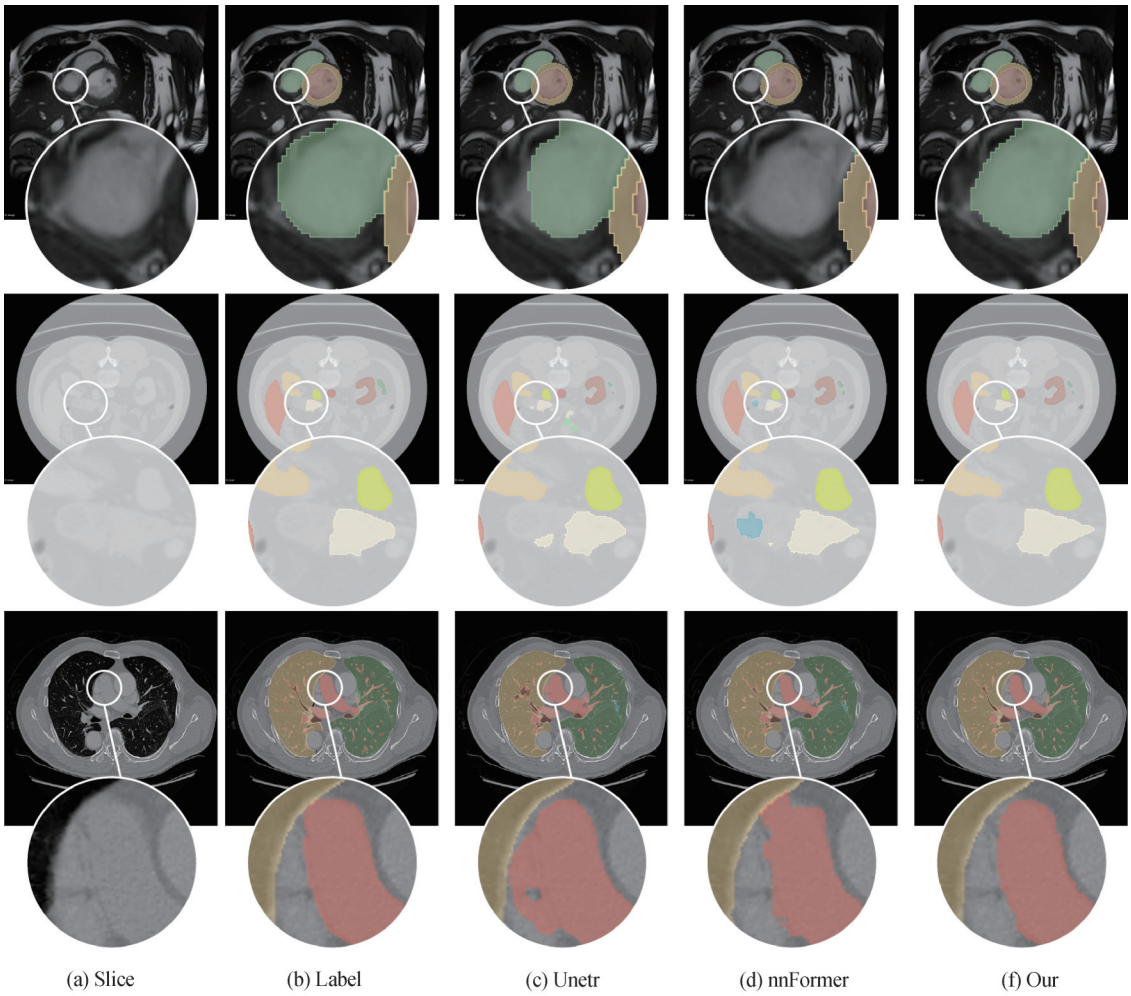

Fig.7 Visual results of three datasets. The first/second/third rows present scans from patients in the ACDC/Synapse/LMCS datasets, each showing both the original slice and its corresponding segmentation label. To better highlight the differences, a circular region is selected and magnified for clearer comparison.

Table 4 Ablation study of different networks on ACDC dataset.

Networks	nnFormer	PMC-CLIP	CoCa-CLIP	Average	LV	RV	MYO	Time
Net_1	√			92.06	90.94	89.58	95.65	2.24 s
Net_2	√	√		92.21	91.53	89.35	95.76	2.25 s
Net_3	√		√	92.24	91.33	89.53	95.88	2.62 s
Net_4	√	√	√	**92.49**	**91.66**	**89.68**	**96.13**	2.64 s

Table 5 Ablation study of different networks on LMCS dataset.

Net.	nnFormer	PMC-CLIP	CoCa-CLIP	Average	Lung (R)	Lung (L)	Bronchi	Nodules	Pul. Artery	Pul. Vein	Time
Net_1	√			81.84	**96.30**	**96.19**	73.77	**66.00**	78.32	80.47	77.86 s
Net_2	√	√		83.35	95.47	95.14	79.37	64.49	81.67	82.31	78.08 s
Net_3	√		√	83.04	94.79	95.44	79.47	64.91	81.69	82.77	81.10 s
Net_4	√	√	√	**83.64**	95.70	95.44	**79.53**	64.78	**82.27**	**84.11**	81.33 s

dataset is noticeably higher than that on the ACDC dataset, primarily due to the significantly larger image dimensions (height, width, and depth) in LMCS. This is a common phenomenon in medical image segmentation, where large image sizes often lead to longer inference times as a result of mirroring augmentation and sliding window strategies adopted by most existing methods. When incorporating the PMC-CLIP and CoCa-CLIP

models into nnFormer, the inference time on the ACDC dataset increases only slightly, by 0.01 s with PMC-CLIP and 0.38 s with CoCa-CLIP. The slightly higher cost in CoCa-CLIP is attributed to the additional step of generating text embeddings. A similar pattern is observed on the LMCS dataset. These results demonstrate that adding the two CLIP models leads to notable performance improvements with only an acceptable increase in inference time.

Similar trends are observed on the LMCS dataset. Compared to Net_1, both Net_2 and Net_3 significantly improve segmentation performance for smaller and anatomically complex structures such as the bronchi, pulmonary artery, and pulmonary vein. Net_4 again achieves the highest average Dice score (83.64%) and outperforms all other variants across three categories, including bronchi, pulmonary artery, and pulmonary vein. These results highlight that the multimodal information extracted from image and text representations in CLIP can enhance feature representation and improve segmentation accuracy, especially in challenging regions.

4.5 Visual Results

To qualitatively assess the segmentation performance, we present visual comparisons of BiCLIP-nnFormer against baseline models on the ACDC, Synapse, and LMCS datasets, as shown in Fig. 7. For each dataset, a representative slice from a patient scan is selected, with a circular region magnified to facilitate clearer visual comparison. In each case, the original slice, ground truth segmentation, and outputs from UNETR, nnFormer, and BiCLIP-nnFormer are provided for comparison.

As illustrated, the proposed BiCLIP-nnFormer generates predictions that are more consistent with the ground truth in terms of organ boundaries, spatial continuity, and anatomical shape. In the first row (ACDC), the right ventricle segmentation by UNETR and nnFormer appears incomplete or under-segmented, while BiCLIP-nnFormer produces a more accurate and complete boundary. In the middle row (Synapse), our model shows improved delineation of organs such as the gallbladder, capturing their contours more faithfully than competing methods. In the final row (LMCS), notable improvements are seen in the segmentation of lung vessels and bronchi, where BiCLIP-nnFormer generates finer, more coherent structures and cleaner edges. These visual comparisons confirm that incorporating CLIP-derived multimodal information enables BiCLIP-nnFormer to achieve more robust and anatomically accurate segmentation results, particularly for small or low-contrast structures.

5 Conclusion

In this paper, we introduce BiCLIP-nnFormer, a novel CLIP-based virtual multimodal instrument for medical image segmentation. To enhance the performance of nnFormer, we integrate two pretrained CLIP models, PMC-CLIP and CoCa-CLIP, which provide complementary global semantic information through image and text embeddings. By leveraging this multimodal knowledge, our model significantly improves segmentation accuracy, particularly in anatomically complex regions. To effectively adapt CLIP for 3D medical imaging, we design a 3D CLIP adapter that aligns embedding dimensions and enables fusion with volumetric features extracted by the nnFormer encoder. The fusion process incorporates both concatenation and global self-attention mechanisms to enhance representation learning. To evaluate the effectiveness of our approach, we conduct experiments on two public datasets, Synapse and ACDC, as well as a self-annotated lung multi-category segmentation (LMCS) dataset. The LMCS dataset includes additional annotations for critical structures such as the bronchi, pulmonary artery and pulmonary vein, which are highly relevant to lung tumor diagnosis and resection planning. Extensive experiments demonstrate that BiCLIP-nnFormer achieves superior performance across all datasets, highlighting the benefits of integrating multimodal CLIP features.

Author Contribution:

The authors' contributions are as follows: Bo Wang was responsible for methodology development, experimental implementation, and funding acquisition. He also contributed to project coordination and participated in the writing of the original draft and visualization. Yue Yan and Mengyuan Xu focused on the execution of experiments and contributed to writing the experimental sections of the manuscript. Yuqun Yang provided overall supervision and conceptual guidance. He was also responsible for reviewing and editing the manuscript, contributed to methodology design, and secured research funding. Xu Tang and Kechen Shu were in charge of data collection and analysis, and played a key role in reviewing and editing the manuscript. Jingyang Ai contributed to algorithm optimization and experimental validation, and also assisted in data analysis and manuscript revision. Zheng You supported the project through supervision, conceptual input, and resource provision.

Acknowledgments:

The authors acknowledge the assistance of ChatGPT and the DALL·E 3 model (OpenAI, 2025) in generating part of the graphical content in Fig. 1 for illustrative purposes.

Foundation Information:

This research was funded by the National Natural Science Foundation of China (Grant No. 6240072655), the Hubei Provincial Key Research and Development Program (Grant No. 2023BCB151), the Wuhan Natural

Science Foundation Exploration Program (Chenguang Program, Grant No. 2024040801020202), and the Natural Science Foundation of Hubei Province of China (Grant No. 2025AFB148).

Data Availability:

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest:

The authors declare no competing interests.

Dates:

Received 20 May 2025; Accepted 11 July 2025; Published online 14 July 2025

References

[1] G. Zeng and G. Zheng. 3D tiled convolution for effective segmentation of volumetric medical images. *Medical Image Computing and Computer Assisted Intervention-MICCAI 2019* , Springer, pp. 146-154, **2019**.

[2] J. C. G. Sánchez, M. Magnusson, M. Sandborget al. *Segmentation of bones in medical dual-energy computed tomography volumes using the* 3D U-Net. Physica Medica, 69:241, **2020**.

[3] W. Chen, B. Liu, S. Penget al. S3D-UNet: Separable 3D U-Net for brain tumor segmentation. *BrainLes* **2018**, MICCAI Workshop, Springer, pp. 358-368, **2019**.

[4] B. H. Menze, A. Jakab, S. Baueret al. *The multimodal brain tumor image segmentation benchmark (BraTS). IEEE Transactions on Medical Imaging* , 34:**1993**, **2014**.

[5] M. P. Heinrich, O. Oktay, N. Bouteldja. Obelisk-Net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions. *Medical Image Analysis* , 54: 1, 2019.

[6] H. Kakeya, T. Okada, Y. Oshiro. 3D U-JAPA-Net: Mixture of convolutional networks for abdominal multi-organ CT segmentation. *MICCAI 2018* , Springer, pp. 426-433, **2018**.

[7] R. Wang, T. Lei, R. Cuiet al. *Medical image segmentation using deep learning* : A survey. IET Image Processing, 16: 1243, **2022**.

[8] A. Norouzi, M. S. M. Rahim, A. Altameemet al. Medical image segmentation methods, algorithms, and applications. *IETE Technical Review* , 31:199, **2014**.

[9] K. Ramesh, G.K. Kumar, K. Swapnaet al. *A review of medical image segmentation algorithms* . EAI Endorsed Transactions on Pervasive Health and Technology, **7**:e6, 2021.

[10] S. K. Zhou, H. Greenspan, D. Shen. Deep Learning for Medical Image Analysis. Academic Press, **2023**.

[11] A. Singh, S. Sengupta, V. Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging* , 6:52, **2020**.

[12] Y. Shao, Z. Geng, Y. Liuet al. *CPT* : A pre-trained unbalanced transformer for both Chinese language understanding and generation. Science China Information Sciences, 67: 152102, **2024**.

[13] K. Muhammad, A. Ullah, J. Lloretet al. *Deep learning for safe autonomous driving* : Current challenges and future directions. IEEE Transactions on Intelligent Transportation Systems, 22:4316, **2020**.

[14] L. Chen, P. Bentley, K. Moriet al. *DRI-Net for medical image segmentation. IEEE Transactions on Medical Imaging* , 37: 2453, **2018**.

[15] Y. Yang, Y. Zhu, Y. Zhanget al. *FSVS-Net* : A few-shot semi-supervised vessel segmentation network for multiple organs based on feature distillation and bidirectional weighted fusion. Information Fusion, 103281, **2025**.

[16] F. Isensee, P. F. Jaeger, S. A. Kohlet al. *nnU-Net* : A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18:203, **2021**.

[17] H. -Y. Zhou, J. Guo, Y. Zhanget al. nnFormer: Volumetric medical image segmentation via a 3*D transformer* . IEEE Transactions on Image Processing, **2023**.

[18] O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional networks for biomedical image segmentation. *MICCAI 2015* , Springer, pp. 234-241, **2015**.

[19] Z. Li, F. Liu, W. Yanget al. *A survey of convolutional neural networks* : analysis, applications, and prospects. IEEE Transactions on Neural Networks and Learning Systems, 33: 6999, **2021**.

[20] S. Cai, Y. Tian, H. Luiet al. *Dense-UNet* : A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. Quantitative Imaging in Medicine and Surgery, 10:1275, **2020**.

[21] I. Kiran, B. Raza, A. Ijazet al. *DenseRes-UNet* : Segmentation of overlapped/clustered nuclei from multi-organ histopathology images. Computers in Biology and Medicine, 143:105267, 2022.

[22] K. Han, Y. Wang, H. Chenet al. *A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence* , 45:87, **2022**.

[23] J. Chen, Y. Lu, Q. Yuet al. *TransUNet* : Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, **2021**.

[24] A. Hatamizadeh, Y. Tang, V. Nathet al. UNETR: Transformers for 3*D medical image segmentation* . WACV, pp. 574-584, **2022**.

[25] Y. Zhang, H. Liu, Q. Hu. TransFuse: Fusing transformers and CNNs for medical image segmentation. *MICCAI* **2021** , Springer, pp. 14-24, **2021**.

[26] D. Karimi, S. D. Vasylechko, A. Gholipour. Convolution-free medical image segmentation using transformers. *MICCAI 2021* , Springer, pp. 78-88, **2021**.

[27] R. Azad, M. Heidari, M. Shariatniaet al. *TransDeepLab* : Convolution-free transformer-based DeepLab V3+ for medical image segmentation. PRIME Workshop, Springer, pp. 91-102, **2022**.

[28] H. Cao, Y. Wang, J. Chenet al. *Swin-UNet* : UNet-like pure transformer for medical image segmentation. ECCV, Springer, pp. 205-218, **2022**.

[29] K. Bayoudh, R. Knani, F. Hamdaouiet al. *A survey on deep multimodal learning for computer vision. The Visual Computer* , 38:2939, **2022**.

[30] P. Xu, X. Zhu, D. A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , **2023**.

[31] A. Radford, J.W. Kim, C. Hallacyet al. *Learning transferable visual models from natural language supervision. ICML* , PMLR, pp. 8748-8763, **2021**.

[32] Z. Zhao, Y. Liu, H. Wuet al. *CLIP in medical imaging* : A comprehensive survey. arXiv preprint arXiv: 2312.07353, **2023**.

[33] Z. Zhou, Y. Lei, B. Zhanget al. *ZegCLIP* : Towards adapting CLIP for zero-shot semantic segmentation. CVPR, pp. 11175-11185, **2023**.

[34] H. Song, L. Dong, W.-N. Zhanget al. *CLIP models are few-shot learners* : Empirical studies on VQA and visual entailment. arXiv preprint arXiv:2203.07190, **2022**.

[35] J. Yu, Z. Wang, V. Vasudevanet al. *CoCa* : Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917, **2022**.

[36] W. Lin, Z. Zhao, X. Zhanget al. *PMC-CLIP* : Contrastive language-image pre-training using biomedical documents. MICCAI, Springer, pp. 525-536, **2023**.

[37] R. Gu, G. Wang, T. Songet al. *CA-Net* : Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Transactions on Medical Imaging, 40:699, **2020**.

[38] A.M. Shaker, M. Maaz, H. Rasheedet al. UNETR++: Delving into efficient and accurate 3*D medical image segmentation* . IEEE Transactions on Medical Imaging, **2024**.

[39] Ö. Çiçek, A. Abdulkadir, S. S. Lienkampet al. 3U-NetD: Learning dense volumetric segmentation from sparse annotation. *MICCAI* **2016**, Springer, pp. 424-432, **2016**.

[40] X. Li, H. Chen, X. Qiet al. *H-DenseUNet* : Hybrid densely connected UNet for liver and tumor segmentation. IEEE Transactions on Medical Imaging, 37:2663, **2018**.

[41] S. Lal, D. Das, K. Alabhyaet al. *NucleiSegNet* : Robust deep learning architecture for the nuclei segmentation of liver cancer histopathology images. Computers in Biology and Medicine, 128:104075, **2021**.

[42] F. Milletari, N. Navab, S. -A. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *3DV* , IEEE, pp. 565-571, **2016**.

[43] E. Gibson, F. Giganti, Y. Huet al. *Automatic multi-organ segmentation on abdominal CT with dense V-networks. IEEE Transactions on Medical Imaging* , 37:1822, **2018**.

[44] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglet al. *Medical Transformer* : Gated axial-attention for medical image segmentation. MICCAI **2021**, Springer, pp. 36-46, **2021**.

[45] A.K. Monsefi, P. Karisani, M. Zhouet al. Masked LogoNet: Fast and accurate 3*D image analysis for medical domain* . arXiv preprint arXiv:2402.06190, **2024**.

[46] A. Lin, B. Chen, J. Xuet al. *DS-TransUNet* : Dual Swin Transformer UNet for medical image segmentation. IEEE Transactions on Instrumentation and Measurement, 71: 1, **2022**.

[47] L.H. Li, P. Zhang, H. Zhanget al. *Grounded language-image pre-training. CVPR* , pp. 10965-10975, **2022**.

[48] T. Lüddecke, A. Ecker. Image segmentation using text and image prompts. *CVPR* , pp. 7086-7096, **2022**.

[49] Z. Wang, Y. Lu, Q. Liet al. *CRIS* : CLIP-driven referring image segmentation. CVPR, pp. 11686-11695, **2022**.

[50] S. Eslami, G. de Melo, C. Meinel. Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint* arXiv: 2112.13906, **2021**.

[51] Y. Luo, M. Shi, M. O. Khanet al. *FairCLIP* : Harnessing fairness in vision-language learning. CVPR, pp. 12289-12301, **2024**.

[52] X. Zhang, M. Xu, D. Qiuet al. *MediCLIP* : Adapting CLIP for few-shot medical image anomaly detection. arXiv preprint arXiv:2405.11315, **2024**.

[53] P. Gao, S. Geng, R. Zhanget al. *CLIP-Adapter* : Better vision-language models with feature adapters. International Journal of Computer Vision, 132:581, **2024**.

[54] C. Yao, M. Hu, Q. Liet al. *TransClaw U-Net* : CLAW U-Net with transformers for medical image segmentation. ICICSP, IEEE, pp. 280-284, **2022**.

[55] G. Xu, X. Zhang, X. Heet al. *LeViT-UNet* : Make faster encoders with transformer for medical image segmentation. PRCV, Springer, pp. 42-53, **2023**.

[56] X. Huang, Z. Deng, D. Liet al. *MissFormer* : An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162, **2021**.

[57] Y. Xie, J. Zhang, C. Shenet al. CoTr: Efficiently bridging CNN and transformer for 3*D medical image segmentation* . MICCAI 2021, Springer, pp. 171-180, **2021**.

[58] A. Hatamizadeh, V. Nath, Y. Tanget al. *Swin UNETR* : Swin transformers for semantic segmentation of brain tumors in MRI images. MICCAI BrainLesion Workshop, Springer, pp. 272-284, **2021**.

[59] S. Perera, P. Navard, A. Yilmaz. SegFormer3D: An efficient transformer for 3D medical image segmentation. *CVPR* , pp. 4981-4988, **2024**.