Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A novel LVPA-UNet network for target volume automatic delineation: An MRI case study of nasopharyngeal carcinoma

Yu Zhang [a,*,1], Hao-Ran Xu [a], Jun-Hao Wen [a], Yu-Jun Hu [b,c], Yin-Liang Diao [a], Jun-Liang Chen [a], Yun-Fei Xia [b,d,**]

[a] *College of Electronic Engineering (College of Artificial Intelligence), South China Agricultural University, Guangzhou, 510642, China*
[b] *State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-Sen University Cancer Center, Guangzhou, China*
[c] *Department of Radiology, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Guangzhou, China*
[d] *Department of Radiation Oncology, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Guangzhou, China*

ARTICLE INFO

Keywords:
Nasopharyngeal carcinoma
Image segmentation
Deep learning
Convolutional network
Layer volume Attention-UNet

ABSTRACT

Accurate delineation of Gross Tumor Volume (GTV) is crucial for radiotherapy. Deep learning-driven GTV segmentation technologies excel in rapidly and accurately delineating GTV, providing a basis for radiologists in formulating radiation plans. The existing 2D and 3D segmentation models of GTV based on deep learning are limited by the loss of spatial features and anisotropy respectively, and are both affected by the variability of tumor characteristics, blurred boundaries, and background interference. All these factors seriously affect the segmentation performance. To address the above issues, a Layer-Volume Parallel Attention (LVPA)-UNet model based on 2D-3D architecture has been proposed in this study, in which three strategies are introduced. Firstly, 2D and 3D workflows are introduced in the LVPA-UNet. They work in parallel and can guide each other. Both the fine features of each slice of 2D MRI and the 3D anatomical structure and spatial features of the tumor can be extracted by them. Secondly, parallel multi-branch depth-wise strip convolutions adapt the model to tumors of varying shapes and sizes within slices and volumetric spaces, and achieve refined processing of blurred boundaries. Lastly, a Layer-Channel Attention mechanism is proposed to adaptively adjust the weights of slices and channels according to their different tumor information, and then to highlight slices and channels with tumor. The experiments by LVPA-UNet on 1010 nasopharyngeal carcinoma (NPC) MRI datasets from three centers show a DSC of 0.7907, precision of 0.7929, recall of 0.8025, and HD95 of 1.8702 mm, outperforming eight typical models. Compared to the baseline model, it improves DSC by 2.14 %, precision by 2.96 %, and recall by 1.01 %, while reducing HD95 by 0.5434 mm. Consequently, while ensuring the efficiency of segmentation through deep learning, LVPA-UNet is able to provide superior GTV delineation results for radiotherapy and offer technical support for precision medicine.

[*] Corresponding author.

[**] Corresponding author. State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-Sen University Cancer Center, Guangzhou, China.

E-mail addresses: zhangyu@scau.edu.cn (Y. Zhang), xiayf@sysucc.org.cn (Y.-F. Xia).

[1] Lead contact.

https://doi.org/10.1016/j.heliyon.2024.e30763
Received 8 February 2024; Received in revised form 8 April 2024; Accepted 3 May 2024
Available online 4 May 2024

1. Introduction

Intensity-modulated radiation Therapy (IMRT) employs high-energy rays or particle beams to target and eradicate cancer cells while avoiding the surrounding normal tissues to minimize damage [1]. Therefore, identifying and delineating Gross Tumor Volume (GTV) rapidly and accurately is a prerequisite for a good tumor radiotherapy plan formulation, effective dose delivery, and efficient and precise treatment. When the delineated GTV is smaller or larger than its actual size, under-treatment of tumor cells or damage of normal cells may occur by the corresponding biased radiation plan [2]. Currently, GTV delineation in MRIs is a predominantly manual, slice-by-slice process conducted by radiation oncologists, which is time-consuming and several hours may be needed for one complex case. It is difficult to achieve a consistent boundary judgment among radiation oncologists of varying expertise levels [3], which brings certain obstacles to the accurate formulation of the IMRT plan. In this situation, there is an urgent need for intelligent GTV segmentation methods, where machines replace humans to achieve precise and consistent delineation of GTV.

In recent years, deep learning technology has achieved significant breakthroughs and applications across various fields, including industry [4,5], agriculture [6], psychology [7,8], and medicine [9,10]. Especially in the medical field, deep learning technology provides intelligent and precise solutions for GTV segmentation. This technology is capable of learning the features of the image in depth, automatically identifying the boundary between tumor and normal tissue, and thus accurately locating and segmenting the GTV, which provides strong support for the formulation of the IMRT plan.

Existing deep learning-based GTV segmentation methods can be categorized into three types: The first approach deconstructs 3D images into a series of slices and employs 2D neural networks for precise segmentation by recognizing and differentiating minor structures and lesions. Subsequently, the segmented 2D slices are reassembled into cohesive 3D images [11–13]. However, due to the lack of spatial correlation between slices in the 2D network segmentation process, this might lead to an incoherent understanding of the 3D anatomical structure and loss of critical 3D spatial feature information, potentially causing inconsistencies in the predicted tumor areas across different slices. The second method employs 3D neural networks to segment 3D images directly [14–24], which is effective in maintaining continuity between slices and overall spatial correlation, providing a comprehensive perception of volumetric spatial features. However, this method might lean towards learning incorrect volumetric structural information due to the resolution difference between the x, y, and z axes in MRIs (image anisotropy), overlooking important details within slices and causing distortion in segmentation results [25,26]. The third approach utilizes hybrid 2D-3D networks, combining the capabilities of 2D networks to focus on fine details within slices with the advantages of 3D networks in understanding the overall spatial structures [26–28], effectively solving the issue of spatial feature loss encountered with the use of only the 2D networks, as well as the issue of the 3D networks in dealing with the anisotropy of the images [29].

Nonetheless, hybrid 2D-3D networks still share common issues with 2D and 3D networks, including: (a) Tumors exhibit significant variabilities in shapes and sizes, and the differences in gray level with surrounding tissues are not distinct, resulting in their boundaries appearing blurred in images. If the receptive fields of the networks are not wide enough, they may not be able to accurately capture the complete shape, structure, and boundary information of the tumors, and especially perform poorly when dealing with tumor scenes with slim edges or blurred boundaries. (b) In MRIs, tumors occupy a small spatial proportion, and the networks, while extracting large amounts of irrelevant background information, might overshadow critical semantic features of tumors, affecting the ability to differentiate details between tumors and normal tissues. In such cases, attention mechanisms that target specific regions become the key to addressing this issue [30,31].

To address the issues of the existing methods, this study proposes the Layer-Volume Parallel Attention (LVPA)-UNet network, which fully harnesses the strengths of hybrid 2D-3D networks. Built upon the classic UNet architecture, LVPA-UNet incorporates an LVPA module at each stage of the encoder, employing three strategies to accurately capture key tumor characteristics. Firstly, this study extends the multi-scale convolutional attention (MSCA) module [32] and proposes an integrated 2D-3D parallel workflow strategy that combines parallel Layer MSCA (L-MSCA) and Volume MSCA (V-MSCA). This parallel strategy facilitates collaboration between feature extraction processes across different dimensions, effectively extracting the anatomical structure information of tumors within slices and their volumetric spaces. It addresses issues of segmentation inconsistency between slices and morphological distortion, caused by the loss of spatial features and anisotropy, respectively. Secondly, for the modular design of L-MSCA and V-MSCA, 2D and 3D multi-branch depth-wise strip convolutions are respectively implemented. This design significantly expands the model's receptive field, enabling it to adapt to variations in tumor shape and size within slices and volumetric spaces, and enhances its ability to identify and process blurred boundaries. Thirdly, leveraging the concept of the dual attention mechanism [31,33], the Layer-Channel Attention module is proposed to focus on slices and channels that are closely related to tumors, while effectively suppressing background noise and interfering signals from lesion-free tissues, thus enhancing the detailed differentiation between tumors and normal tissues. The contributions of this study are as follows.

(a) Based on the hybrid 2D-3D architecture, this study proposes an LVPA-UNet to realize GTV segmentation, which can assist radiation therapists in formulating more accurate radiation dose distribution plans.
(b) For each stage of the encoder, this paper introduces three innovative strategies: 2D-3D parallel workflows, parallel multi-branch depth-wise strip convolutions, and Layer-Channel Attention—to address the issues that 2D, 3D, or hybrid 2D-3D networks face, including spatial feature loss, anisotropy, variable tumor characteristics, blurred boundaries, and background interference in MRIs. By integrating these three strategies, a significant improvement in the overall performance of the network is achieved.
(c) In a performance comparison conducted on 1010 stage II NPC T1-weighted MRIs, LVPA-UNet outperforms eight typical models, showcasing its superior effectiveness in GTV segmentation.

2. Methodology

The purpose of this study is to develop and evaluate a novel model for predicting GTV in MRIs. A dataset comprising 1010 cases of stage II NPC from three research institutions has been gathered to train, validate, and test the proposed model LVPA-UNet, which is an end-to-end deep learning network designed specifically for GTV segmentation. The LVPA-UNet's performance improvement is confirmed by comparing it with typical models and through ablation experiments.

2.1. The overall architecture of LVPA-UNet

Fig. 1 shows the network architecture of the proposed LVPA-UNet. The input to LVPA-UNet is a 3D T1-weighted MRI, denoted as $X = \mathbb{R}^{C \times D \times H \times W}$, where X comprises D slices of 2D spatial resolution $H \times W$. Given that the input modality is solely T1-weighted, the channel number C is 1. Based on the UNet architecture, LVPA-UNet primarily comprises an encoder and a decoder. The encoder comprises multiple stages, each leveraging the LVPA module to generate high-level information through the integration of three advanced mechanisms. Specifically, these encoder enhancements include parallel 2D and 3D workflows to improve both 2D slice details and 3D spatial information, L-MSCA and V-MSCA [32] composed of parallel multi-branch depth-wise strip convolutions to expand the receptive field with a heightened focus on boundary processing and variable tumors, and Layer-Channel Attention module to increase attention on tumor-related slices and channels. Within the decoder module, each stage employs a skip connection to fuse information from corresponding stages in the encoder and decoder, progressively generating the segmentation map. The final high-resolution segmentation result is formed by combining maps from adjacent stages.

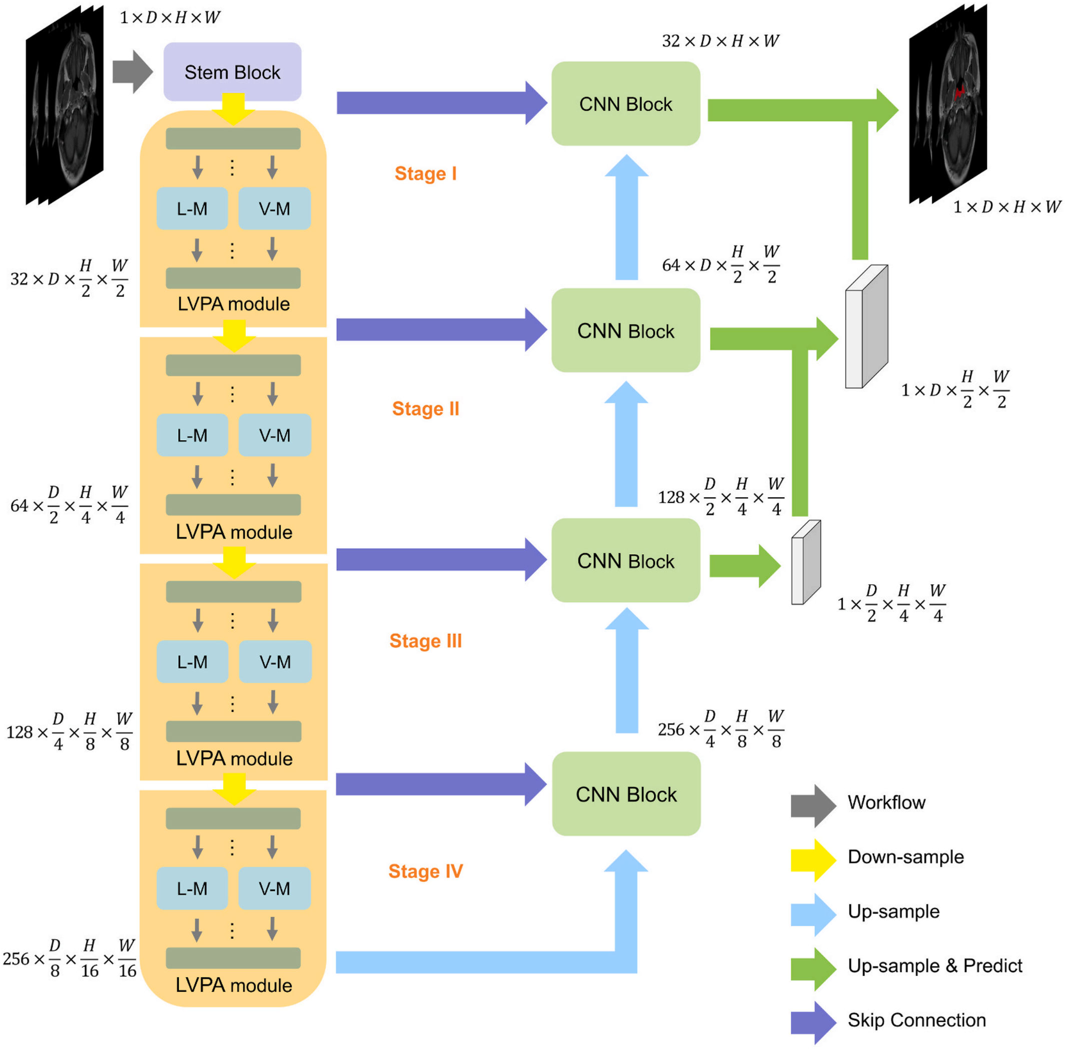
$1 \times D \times H \times W$
Stem Block
$32 \times D \times H \times W$
CNN Block
$1 \times D \times H \times W$
Stage I
L-M
V-M
$32 \times D \times \frac{H}{2} \times \frac{W}{2}$
LVPA module
$64 \times D \times \frac{H}{2} \times \frac{W}{2}$
CNN Block
$1 \times D \times \frac{H}{2} \times \frac{W}{2}$
Stage II
L-M
V-M
$64 \times \frac{D}{2} \times \frac{H}{4} \times \frac{W}{4}$
LVPA module
$128 \times \frac{D}{2} \times \frac{H}{4} \times \frac{W}{4}$
CNN Block
$1 \times \frac{D}{2} \times \frac{H}{4} \times \frac{W}{4}$
Stage III
L-M
V-M
$128 \times \frac{D}{4} \times \frac{H}{8} \times \frac{W}{8}$
LVPA module
$256 \times \frac{D}{4} \times \frac{H}{8} \times \frac{W}{8}$
CNN Block
Stage IV
L-M
V-M
$256 \times \frac{D}{8} \times \frac{H}{16} \times \frac{W}{16}$
LVPA module
Workflow
Down-sample
Up-sample
Up-sample & Predict
Skip Connection

Fig. 1. The overall architecture of the proposed LVPA-UNet.
Abbreviation: CNN = convolution neural network; L-M = L-MSCA module; V-M = V-MSCA module; LVPA = Layer-Volume Parallel Attention.

2.2. Network encoder

The encoder of LVPA-UNet consists of four stages, each encompassing an LVPA module that is designed to integrate information from 2D slices with 3D volumetric space and adaptively perceive highly correlated slices and channels with tumors. In constructing the encoder, specific considerations are given to the feature dimensions at each stage: (a) The Stem block [32,34] is incorporated with the intent to preserve detailed features at the original resolution as much as possible. Instead of simply halving D, H, and W, they are tailored to be consistent with the size of the input MRI. (b) Considering that the dataset in this study suffers from anisotropy, with D averaging of only 33.01 layers, it's imperative to minimize the loss of depth-related feature information due to a substantial reduction in D. Hence, in stage I, D is maintained to be consistent with the input image. (c) From stage II to stage IV, before the extraction of features by the LVPA module in each stage, overlapping patch embedding operations are configured to reduce resolution while extracting more enriched semantic features.

2.3. The architecture of the LVPA module

As depicted in Fig. 2, the LVPA module is a composite built from a stack of parallel L-MSCA, V-MSCA, Layer-Channel Attention, and auxiliary modules.

2.3.1. The overview of L (V)-MSCA module

At the core of the LVPA module, inspired by Ref. [32], an innovative approach is taken to handle 3D images. This marked the inception of Layer MSCA (L-MSCA) and Volume MSCA (V-MSCA). As illustrated in Fig. 2(A and B), both L-MSCA and V-MSCA adopt similar architectural designs, encompassing four key components: (a) a 5 × 5 × 5 depth-wise convolution serving as the conduit for local information aggregation; (b) multi-branch depth-wise strip convolutions engineered to capture semantic features of tumor at multiple scales and different shapes, thus expanding the receptive field; (c) a 1 × 1 × 1 convolution to fuse information across different channel; and (d) a subsequent application of the output from the aforementioned stages as attention weights, through multiplication with the L (V)-MSCA module's input, effectuating a re-weighting of the input features for both L-MSCA and V-MSCA.

As shown in Fig. 2(C), the introduction of L-MSCA and V-MSCA integrates two key strategies of this study. The parallel approach

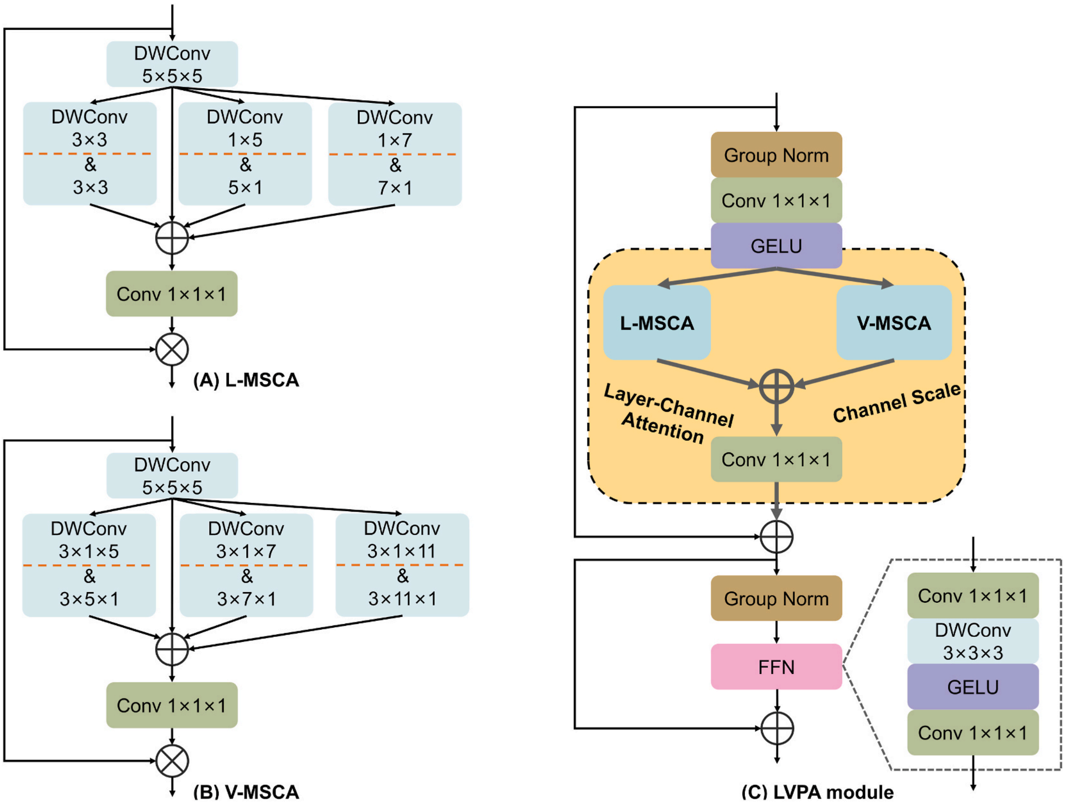

Fig. 2. The architecture of the proposed LVPA module in detail. (A) denotes the L-MSCA, (B) denotes the V-MSCA, and (C) denotes the L-MSCA and V-MSCA are parallel in the yellow region of the LVPA module and then subsequently fused by Conv 1 × 1 × 1.
Abbreviation: Conv = convolution; DWConv = depth-wise convolution; FFN = feedforward neural network; GELU = gaussian error linear unit. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

fosters cooperation between 2D and 3D feature extraction processes, effectively extracting anatomical information within slices and their volumetric spaces, overcoming the loss of spatial features and the distortion caused by image anisotropy; the design of parallel multi-branch depth-wise strip convolutions expands the model's receptive field, adapting to the variability of tumor characteristics and enhancing the capability to handle ambiguous boundaries. Detailed descriptions of the L(V)-MSCA design are provided in sections 2.3.2 and 2.3.3.

2.3.2. L-MSCA module

In the pursuit of robust segmentation from MRI data, one must account for the inherent anisotropy within the data. For instance, it is not uncommon to observe a glaring presence of a tumor in one slice and its complete absence in an adjacent one. This discontinuity poses a conundrum: an excessive reliance on 3D modules in segmentation networks could inadvertently sway the model toward volumetric information, at the peril of overshadowing crucial intra-slice features. Thus, there is a palpable necessity to devise a 2D-centric module that deftly captures the nuances within slices.

The L-MSCA module is improved based on [32], both modules for processing 2D images. However, what demands attention in the context of the NPC dataset is that NPC GTV generally occupies a minuscule fraction of the image and exhibits gray levels not significantly divergent from surrounding tissues, compounded by blurred boundaries. Thus, when constructing the L-MSCA module, it is necessary to give it a finer processing capability and slightly different corresponding parameters.

Fig. 2(A) delineates the structure of the L-MSCA module. For an input feature map $F \in \mathbb{R}^{C\times D\times H\times W}$, the module first applies a 5× 5× 5 depth-wise convolution (DWConv) to assimilate local information. This operation can be succinctly expressed as:

$$Z_1^{2D} = DW_{5\times5\times5}(F) \tag{1}$$

where $DW_{5\times5\times5}$ symbolizes the $5 \times 5 \times 5$ DWConv operation.

Subsequently, the feature maps Z_1^{2D} are processed through an ensemble of three parallel depth-wise strip convolution branches coupled with an additional identity branch. Within these three branches, pairs of depth-wise strip convolutions are deployed, with convolutional pairs specifically configured as 3×3 and 3×3, 1×5 and 5×1, as well as 1×7 and 7×1. The rationale behind utilizing pairs with larger kernels in strip convolutions is to optimize for flat or elongated targets, such as the heterogeneously shaped NPC GTV. This design also enables the inclusion of more information about the surrounding areas of tumors with variable shapes, thereby optimizing the handling of their ambiguous boundaries. Furthermore, employing $1 \times N$ and $N \times 1$ convolution sizes emulates the performance of $N \times N$ convolutions, thereby achieving a broader receptive field with minimal computational overhead. Lastly, to counterbalance any deficiencies that depth-wise strip convolutions may exhibit in block object scenarios, a complementary branch with a pair of 3×3 DWConvs is integrated. Taken together, this architecture, boasting four parallel branches, is exquisitely tailored to the context like NPC GTV, ensuring the rich excavation of multiscale and morphologically diverse information within the slices. The process is succinctly encapsulated in the following equation:

$$Z_2^{2D} = DW_{[3\times3]}\left(Z_1^{2D}\right) + DW_{[5\times1]}\left(Z_1^{2D}\right) + DW_{[7\times1]}\left(Z_1^{2D}\right) + Z_1^{2D} \tag{2}$$

where $DW_{[N\times M]}$ represents a pair of DWConvs, with strip convolution kernel sizes set at $N \times M$ and $M \times N$.

Finally, the feature maps from the four branches are amalgamated, serving as attention weights that undergo a Hadamard product with F to reweigh the input features. This series of operations can be articulated as:

$$Z_{out}^{2D} = Conv_{1\times1\times1}\left(Z_2^{2D}\right) \otimes F \tag{3}$$

where Z_{out}^{2D} denotes the output of the L-MSCA module, $Conv_{1\times1\times1}$ represents a $1 \times 1 \times 1$ convolution, and $\otimes$ symbolizes the Hadamard product.

2.3.3. V-MSCA module

As shown in Fig. 2(B), V-MSCA's network structure is akin to the L-MSCA module; however, it distinguishes itself by employing 3D depth-wise strip convolutions in its three branches. To further augment the expression of multi-scale features, it exhibits a unique set of convolution kernel sizes distinct from L-MSCA. The V-MSCA module's calculations are captured in the following:

$$Z_1^{3D} = DW_{5\times5\times5}(F) \tag{4}$$

$$Z_2^{3D} = DW_{3[5\times1]}\left(Z_1^{3D}\right) + DW_{3[7\times1]}\left(Z_1^{3D}\right) + DW_{3[11\times1]}\left(Z_1^{3D}\right) + Z_1^{3D} \tag{5}$$

$$Z_{out}^{3D} = Conv_{1\times1\times1}\left(Z_2^{3D}\right) \otimes F \tag{6}$$

where Z_1^{3D} and Z_2^{3D} represent intermediate outputs, Z_{out}^{3D} signifies the output of the V-MSCA module, and $DW_{K[N\times M]}$ denotes a pair of DWConvs with strip convolution kernel sizes set at $K \times N \times M$ and $K \times M \times N$.

2.3.4. Layer-Channel Attention

As depicted in Fig. 3, the slices that contain the NPC tumor constitute only a small fraction of the total slice count. Moreover, the small size of the tumor within these slices, combined with its indistinct boundaries when compared to normal tissue, presents a significant challenge. Indiscriminately applying sliding convolutional kernels across feature maps results in the extraction of information that is mixed with the complex background, negatively impacting the precision of GTV segmentation. Given these observations, a Layer-Channel Attention module is introduced, specifically tailored to the output of the L-MSCA module, as an effective countermeasure.

As shown in Fig. 4, inspired by the dual attention mechanism presented in Refs. [31,33], the conventional channel attention has been extended into a more sophisticated Layer-Channel Attention. Specifically, the input feature maps, denoted as $F \in \mathbb{R}^{C\times D\times H\times W}$, are sequentially passed through Channel Attention and Layer Attention, processing the features along the channel and layer dimensions. This endows the network with the ability to allocate varying weights to each channel and slice, thereby accentuating features that exhibit a strong correlation with the tumor and attenuating contributions from less relevant regions. Finally, a Layer-Channel Scale matrix of size $C \times D \times 1 \times 1$ is deployed, which further amplifies the importance of the strongly correlated channels and slices of the tumor to the overall segmentation. The specific formulation of the formula is as follows:

$$Z_{ca} = CA(F) + F \tag{7}$$

$$Z_{la} = LA(Z_{ca}) + Z_{ca} \tag{8}$$

$$Z_{out} = LCS(Z_{la}) \tag{9}$$

where CA represents Channel Attention, LA denotes Layer Attention, and LCS signifies the Layer-Channel Scale. Z_{ca} and Z_{la} are intermediate outputs after the application of Channel Attention and Layer Attention, respectively, whereas Z_{out} represents the final output subsequent to processing by the entire Layer-Channel Attention module.

2.4. Network decoder

LVPA-UNet embraces the architecture of the decoder design delineated in Ref. [24] which is organized in a four-stage cascade that works together to gradually improve resolution and define the segmentation outlines. Each stage comprises three Conv3D-GroupNorm-ReLU (CGR) operations, where Group Normalization is chosen over Instance Normalization. Between adjacent CGR operations, upsampling methods based on trilinear interpolation and skip connections are implemented. These skip connections

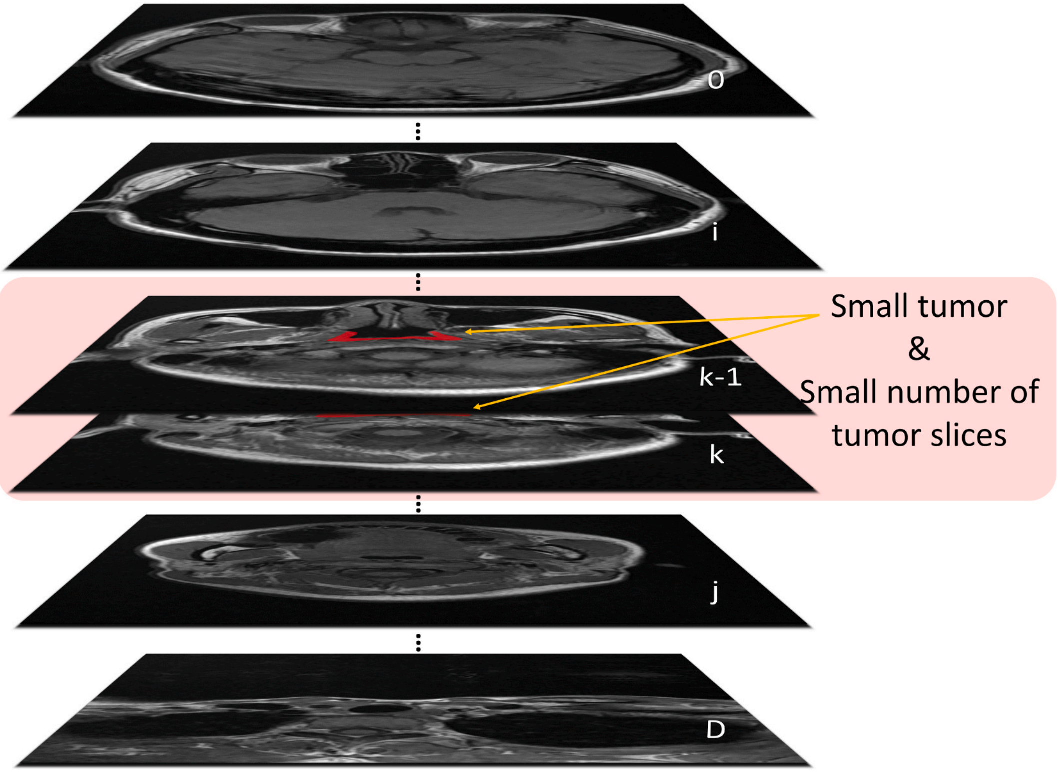

Fig. 3. Visualization of tumor characteristics in MRI imaging. D represents image depth, k represents the slice number containing the NPC tumor, and the red markings indicate areas with the tumor. Note that the tumor is present only in a small portion of the total number of slices, and it appears small in size within these slices. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

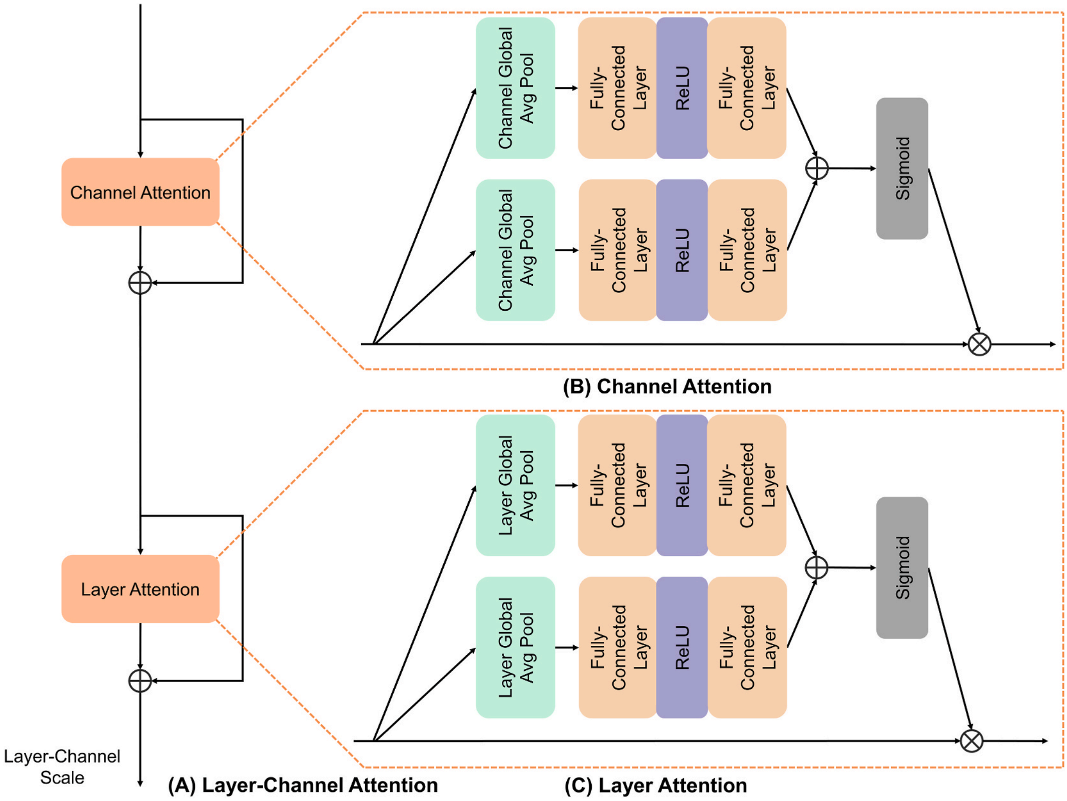

Fig. 4. The architecture of Layer-Channel Attention. Layer-Channel Attention is depicted in (A). Channel Attention (B) and Layer Attention (C) are similar in structure, but the scope of expression of their attention is different. This architecture automatically selects key features for multiple channels and layers.
Abbreviation: ReLU = rectified linear unit.

efficiently combine the high-level feature maps from the encoder with the corresponding stage of the decoder, effectively integrating multi-scale features. Notably, to safeguard the richness of depth information, The decoder treats D in the same way as the encoder, with stages I and II preserving depth congruent to the input features.

Furthermore, the segmentation results of the previous layer are often instructive. As shown in the green flow in Fig. 1, the decoder part starts from stage III, downscaling the features extracted by the CNN block, realizing cross-channel information interaction and integration, and generating a single-channel rough segmentation result; the rough segmentation result is scaled up to the size of the result of the next layer, and the output of the CNN block of the next layer is fused into a higher resolution and more spatially detailed segmentation result to achieve the effect that the segmentation result of the former layer guides the latter layer. Repeat the operation until the final prediction map $R \in \mathbb{R}^{1\times D\times H\times W}$ is generated.

3. Experiments and results

3.1. Experimental data

In this study, a retrospective review is carried out on 577 cases of stage II NPC from the Sun Yat-sen University Cancer Center (from May 2010 to July 2017), 286 cases from Fujian Province Cancer Hospital (from April 2008 to December 2016), and 212 cases from Jiangxi Province Cancer Hospital (from January 2010 to July 2017). As illustrated in Fig. 5, based on the inclusion and exclusion criteria, a total of 1010 NPC patients (median age, 45 years; 673 males, 337 females) with corresponding diagnostic MRIs are included in this study. Oncologists delineate the GTV on all MRI slices, serving as the ground truth for training and assessing the model's performance..

Experiments are conducted on T1-weighted MRIs of the head and neck region of 1010 NPC patients. The inter-voxel spacings in the MRIs are set to $0.5 \times 0.5 \times 6mm^3$ through linear interpolations, with average dimensions of $33.0 \times 464.8 \times 464.6$. The data are randomly divided into three subsets: a training set (606 cases, 60 %), a validation set (202 cases, 20 %), and a testing set (202 cases, 20 %). Statistical information on the dataset is shown in Table 1.

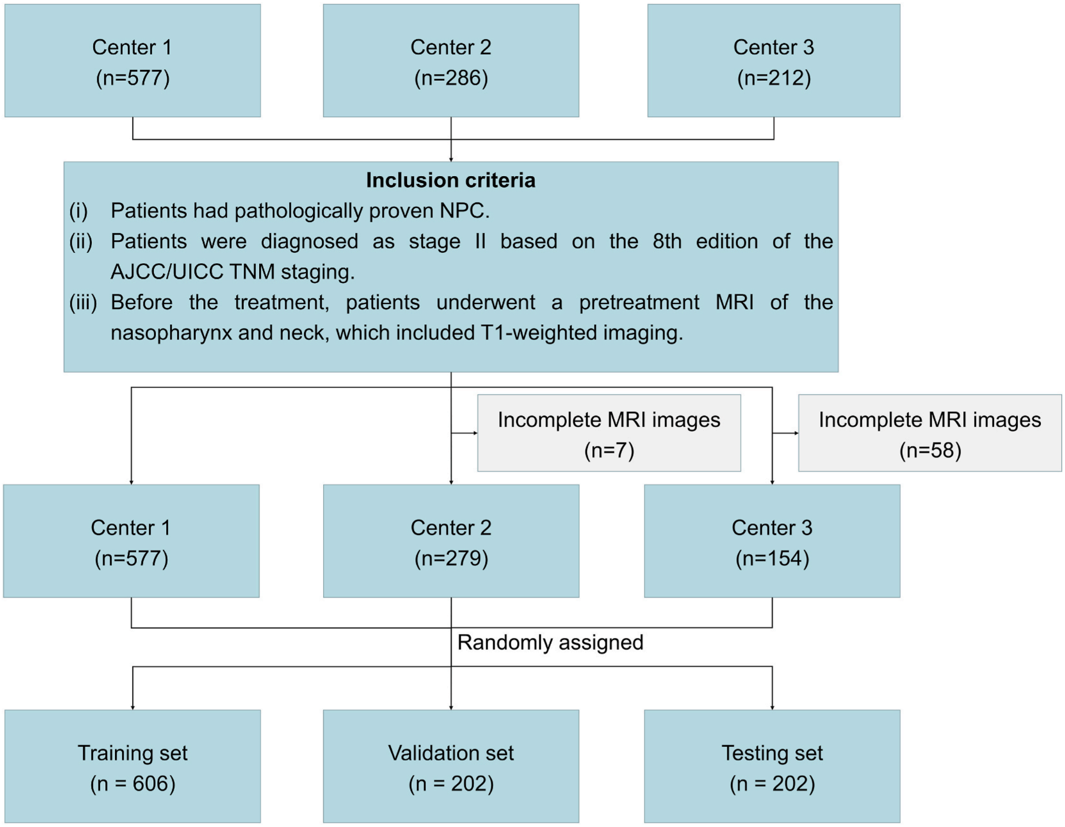
Center 1
(n=577)
Center 2
(n=286)
Center 3
(n=212)
Inclusion criteria
(i) Patients had pathologically proven NPC.
(ii) Patients were diagnosed as stage II based on the 8th edition of the AJCC/UICC TNM staging.
(iii) Before the treatment, patients underwent a pretreatment MRI of the nasopharynx and neck, which included T1-weighted imaging.
Incomplete MRI images
(n=7)
Incomplete MRI images
(n=58)
Center 1
(n=577)
Center 2
(n=279)
Center 3
(n=154)
Randomly assigned
Training set
(n = 606)
Validation set
(n = 202)
Testing set
(n = 202)

Fig. 5. Flowchart of patient enrollment.
AbbreviationAJCC/UICC = American Joint Committee on Cancer/Union for International Cancer Control.

Table 1
Statistical information of the dataset.

Characteristic	Training-Validation Set (n = 808)	Testing Set (n = 202)
Age (years), median (range)	45 (19, 82)	45 (15, 76)
Sex, No. (%)		
Male	533 (66.0)	140 (69.3)
Female	275 (34.0)	62 (30.7)
Histopathology, No. (%)		
WHO I	50 (6.2)	0 (0.0)
WHO II	108 (13.4)	13 (6.4)
WHO III	650 (80.4)	189 (93.6)
T, No. (%)		
T1	399 (49.4)	108 (53.5)
T2	409 (50.6)	94 (46.5)

Patients are categorized based on the 8th edition of the American Joint Committee on Cancer staging manual.
Abbreviations: WHO: World Health Organization.

3.2. Performance metrics

The segmentation performance is evaluated via four metrics, namely the Dice similarity coefficient (DSC) [35], 95 % Hausdorff distance (HD95) [36], precision, and recall.

The DSC is defined as follows:

$$DSC = \frac{2|Z_p \cap Z_g|}{|Z_p| + |Z_g|} \tag{10}$$

where Z_p denotes the predicted output mask, and Z_g denotes the ground truth mask. A higher DSC indicates a more accurate segmentation. In binary segmentation tasks, the DSC can be derived from precision and recall and is computationally equivalent to the F1-

score.

The HD95 assesses the spatial discrepancy between the predicted output and ground truth masks, with smaller values indicating better performance. Let C_p and C_g denote the boundaries of the predicted mask and the ground truth mask, respectively. The maximal HD is defined as:

$$max\{h(C_p, C_g), h(C_g, C_p)\} \tag{11}$$

where $h(C_p,C_g) = max_{a\in C_p} min_{b\in C_g}||a-b||$, $h(C_g,C_p) = max_{a\in C_g} min_{b\in C_p}||a-b||$. The value 95 % is employed to mitigate the influence of a very small subset of outliers.

In summary, among these four metrics, a higher DSC, precision, recall, and lower HD95 indicate superior segmentation performance.

3.3. *Implementation details*

In this study, for the sake of fair comparison, the settings for LVPA-UNet, as well as all other models used in comparative experiments, are consistent. Apart from differences in model architecture, all training and testing parameters are kept identical.

All experiments are conducted using PyTorch [37] and MONAI [38]. Training and testing are carried out on a Windows Server equipped with an Intel Xeon CPU E5-2650 v4 with 12 cores and an NVIDIA GeForce RTX 3090 with 24 GB memory. All data are subjected to min-max normalization to normalize pixel values to the range of 0–1 before use.

During the training process, in addition to normalization, random flipping and rotation are performed on the data to enhance the model's generalizability. Furthermore, random patches of 32 × 256 × 256 are cropped from 3D image volumes, an approach devised to exploit as many voxels as possible within the constraints of GPU memory. The dimensions of 32 × 256 × 256 essentially encompass all lesions, striving to optimize segmentation performance. All weights are initialized using the Kaiming method [39]. The loss function comprises a weighted sum of Dice loss and cross entropy loss [19]. An Adam optimizer and cosine scheduler are employed, and the learning rate is initially set to 0.0001. The training batch size is set to 1, with 300 training iterations. The model weights at the iteration with the highest Dice score on the validation set are selected as the final model weights.

For the testing process, the sliding window inference method from MONAI is used, with a window size of 32 × 256 × 256 and an overlap of 50 %. The final four metrics are then derived by averaging the evaluations across all test data.

3.4. *Comparison with typical models*

The proposed model is compared to eight typical segmentation models: 3D UNet [40], Lin et al. [23], Isensee et al. [24], the 2D implementation of Isensee et al. [24], TransBTS [20], VT-UNet [19], UNETR [18], and MPU-Net [41]. As shown in Table 2, LVPA-UNet outperforms all other models in every metric, reaching peak scores for DSC, HD95, precision, and recall at 0.7907, 1.8702 mm, 0.7929, and 0.8025, respectively. Compared to the top-performing model VT-UNet, LVPA-UNet demonstrates an increase in mean DSC (0.7907 vs 0.7758), precision (0.7929 vs 0.7690), and recall (0.8025 vs 0.7989), respectively. Compared to the previous advanced model by Lin et al. [23] from Sun Yat-sen University Cancer Center, LVPA-UNet showcases an increase in mean DSC (0.7907 vs 0.7754), precision (0.7695 vs 0.7633), and recall (0.8025 vs 0.7983), respectively. In terms of HD95, it achieves superior performance by recording a lower value compared to Isensee et al. [24] (1.8702 mm vs 2.4136 mm).

Fig. 6 depicts the visual comparison of segmentation outcomes between LVPA-UNet and other typical models. The first and second rows present two adjacent slices from the same MRI image. In the first row, the delineations of the thin tumor area by 3D UNet, Lin et al. [23], Isensee et al. [24], 2D implementation of Isensee et al. [24], and VT-UNet are visibly thinner than the ground truth, even fragmented. Only UNETR, MPU-Net, and LVPA-UNet manage to accurately depict this area, but UNETR shows a tendency toward overlapping boundaries. In the second row, 3D UNet's prediction noticeably deviates from the ground truth. While the predictions by Isensee et al. [24], Lin et al. [23], TransBTS, UNETR, MPU-Net, and VT-UNet are fairly similar and superior to 3D UNet, LVPA-UNet's prediction closely matches the ground truth, demonstrating optimal performance.

Upon scrutinizing the adjacent slices in the first and second rows, it's evident that the size and shape of the tumor vary significantly, even in consecutive slices. The eight typical models exhibit varying degrees of error in outlining the tumor across these slices. Notably,

Table 2
Comparison of evaluation metrics of various typical models.

Models	DSC ↑	HD95 (mm) ↓	Precision ↑	Recall ↑
3D UNet	0.7172 ± 0.0955	5.7052 ± 26.80	0.7354 ± 0.1185	0.7223 ± 0.1317
Isensee et al. [24]	0.7693 ± 0.0893	2.4136 ± 3.5872	0.7633 ± 0.1143	0.7924 ± 0.1163
2D Isensee et al. [24]	0.6520 ± 0.1463	7.9084 ± 10.2189	0.8081 ± 0.1134	0.5763 ± 0.1846
Lin et al. [23]	0.7754 ± 0.0954	2.6832 ± 4.7794	0.7695 ± 0.1177	0.7983 ± 0.1197
TransBTS	0.7744 ± 0.0929	8.6436 ± 87.3025	0.7583 ± 0.1107	0.8080 ± 0.1221
UNETR	0.7683 ± 0.0944	3.0822 ± 6.2588	0.7687 ± 0.1184	0.7850 ± 0.1199
VT-UNet	0.7758 ± 0.0906	2.5433 ± 5.1414	0.7690 ± 0.1073	0.7989 ± 0.1203
MPU-Net	0.7693 ± 0.0963	3.1479 ± 7.7695	0.7817 ± 0.1143	0.7747 ± 0.1279
LVPA-UNet	0.7907 ± 0.0937	1.8702 ± 2.8491	0.7929 ± 0.1088	0.8025 ± 0.1201

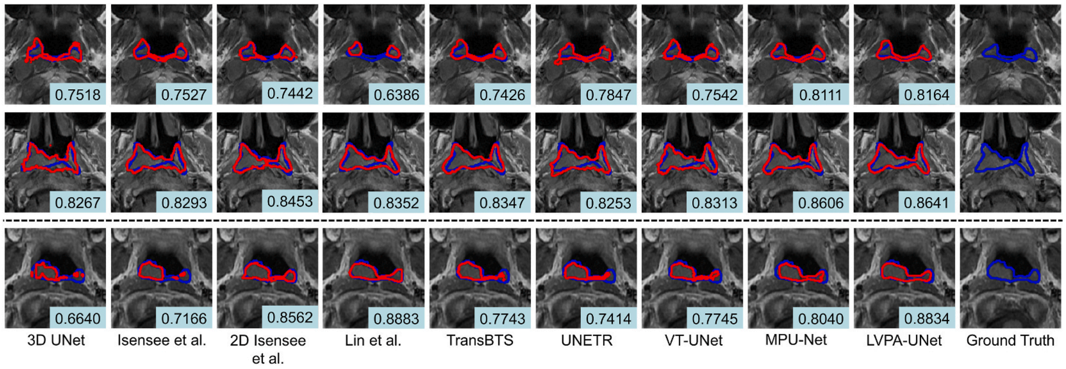

Fig. 6. Comparison of segmentation results between various representative segmentation models and LVPA-UNet. Ground truths for GTV in the current slice annotated by oncologists are marked with blue lines, while model predictions are outlined in red. The numbers at the bottom of the image represent the DSC score for the GTV in the current slice. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

the model proposed by Lin et al. [23] exhibits significant morphological distortion of the tumor between adjacent slices. The 2D implementation of Isensee et al. [24] exhibits inferior performance on the first slice and superior on the second compared to its 3D counterpart, indicating a loss of 3D features that leads to inconsistencies between adjacent slices. In stark contrast, LVPA-UNet's predictions align well with the ground truth in both slices. This illustrates the strength of LVPA-UNet's parallel and interactive 2D and 3D workflows, which integrate volumetric information into the fine edge details of 2D slices, adapting to the anisotropic image environment for enhanced segmentation performance.

As shown in the third row, the delineations of the thin tumor area by 3D UNet and the model proposed by Isensee et al. [24] are interrupted. TransBTS, UNETR, VT-UNet, and MPU-Net, despite connecting this thin tumor area, do not accurately represent its thickness, and they significantly miss the ambiguous region on the right. In contrast, LVPA-UNet proposed in this study adeptly handles both the thin tumor area and the ambiguous region on the right.

3.5. Ablation study on the effect of the LVPA module

To elucidate the effectiveness of the LVPA module in improving NPC GTV segmentation performance, different modules are integrated into the model in a stepwise manner. As Table 3 demonstrates, the work by Isensee et al. [24] denotes the baseline, 'V' indicates the exclusive application of V-MSCA in the encoder block, and 'LV' refers to the module featuring a parallel structure of L-MSCA and V-MSCA. 'LVPA' represents the LV module enhanced with the Layer-Channel Attention module. This progressive integration of components simultaneously improves segmentation performance. Notably, compared to the baseline model, LVPA-UNet achieves the best performance in terms of mean DSC (0.7907 vs 0.7693), mean HD95 (1.8702 mm vs 2.4136 mm), mean precision (0.7929 vs 0.7633) and mean recall (0.8025 vs 0.7924). These results manifest the superior performance of the LVPA module in NPC GTV segmentation.

Fig. 7 presents visual results that emerge from progressively integrating distinct modules into the model. As depicted in the first row, the baseline model delineates an area that deviates from the ground truth, interrupting the delineation in the thinned area of the tumor within the slice. The incorporation of V-MSCA brings the outlined area into closer alignment with the ground truth, and it bridges the delineation in the thinning target area within the slice, attributed to the strip design of MSCA's multi-branch depth-wise strip convolution that optimizes the segmentation of slender structures and their indistinct boundaries. Transitioning from V-MSCA to the parallel workflow of L-MSCA and V-MSCA enhances the detailed handling of the slices by leveraging the advantages of a 2D-3D parallel architecture. Ultimately, with the support of the Layer-Channel Attention module, LVPA adaptively promotes significant slices and channels while suppressing other non-important areas, resulting in segmentation outcomes that closely approximate the ground truth in the images.

In the second row, apart from the shrunken segmentation result of the baseline model, the proposed models closely align with the ground truth. While the V model marginally outperforms the LV model by 0.59 % in DSC, the fully integrated LVPA model achieves the best results. This underscores the necessity of integrating the complete set of modules in LVPA-UNet.

Table 3
Comparison of segmentation performance by stepwise integration of various modules into the model.

Models	DSC ↑	HD95 (mm) ↓	Precision ↑	Recall ↑
Isensee et al. [24]	0.7693 ± 0.0893	2.4136 ± 3.5872	0.7633 ± 0.1143	0.7924 ± 0.1163
V	0.7822 ± 0.0913	2.3478 ± 3.4176	0.7836 ± 0.1113	0.7952 ± 0.1168
LV	0.7865 ± 0.0925	2.3342 ± 6.0059	0.7827 ± 0.1106	0.8041 ± 0.1172
LVPA	0.7907 ± 0.0937	1.8702 ± 2.8491	0.7929 ± 0.1088	0.8025 ± 0.1201

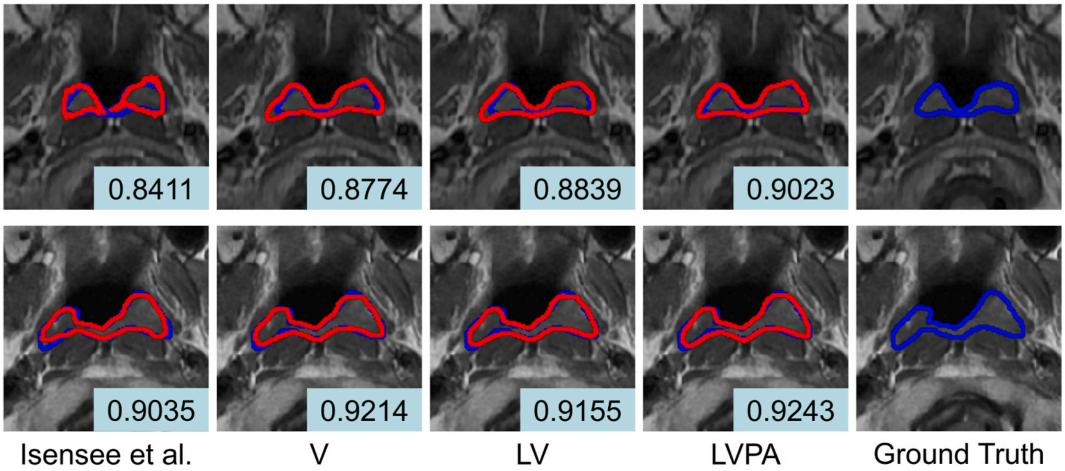

Fig. 7. Comparative visualization of segmentation results of different modules gradually integrated into the model. Ground truths for GTV in the current slice annotated by oncologists are marked with blue lines, while model predictions are outlined in red. The numbers at the bottom of the image represent the DSC score for the GTV in the current slice. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.6. Ablation study on different designs of L-MSCA and V-MSCA

This study undertakes an ablation study on the distinct kernel designs within the multiple branches of the L-MSCA and V-MSCA in the LVPA module. Here, $3[K \times K]$ denotes a $3 \times 1 \times K$ 3D depth-wise convolution and a 3D $3 \times K \times 1$ depth-wise convolution, while $[K \times K]$ signifies a $K \times 1$ 2D depth-wise convolution and a $1 \times K$ depth-wise convolution. Table 4 presents the evaluation results for these different designs. The convolution parameters in the second row achieve the optimal scores in DSC, HD95, and precision. By selecting convolutional kernel geometries as exemplified in the second row, it is possible to capture both the comprehensive attributes and specific details of tumors across various dimensions and configurations, achieving a more appropriate receptive field that enhances the perception of indistinct boundaries. Consequently, this strategic choice can improve the overall GTV segmentation accuracy.

4. Discussion

This study utilizes a dataset compiled from 1010 stage II NPC T1-weighted MRIs from three renowned medical centers, meticulously delineated by two experienced radiation therapy specialists. Utilizing deep learning technology, an LVPA-UNet segmentation network integrating three key strategic advantages is designed. Through training, an automated segmentation model targeting GTV is developed.

The results of LVPA-UNet show a DSC of 0.7907, precision of 0.7929, recall of 0.8025, and a Hausdorff Distance HD95 of 1.8702 mm, surpassing the best existing model, with higher DSC by 1.49 %, greater precision by 2.39 %, improved recall by 0.36 %, and a reduced HD95 by 0.6731 mm. The model performance is enhanced continuously with the integration of modules during the ablation study: Firstly, a 1.29 % increase in DSC is obtained from the base model (Isensee et al. [24]) to the V model, which verifies that LVPA-UNet's strategy of multi-branch depth-wise strip convolutions improves the adaptability of the base model to the complex issues of elongated tumors, size variations, and ambiguous boundaries. Subsequently, a further 0.43 % increase in DSC is achieved when V is improved to LV, which thoroughly demonstrates the synergistic advantages of 2D-3D parallel workflows and parallel multi-branch depth-wise strip convolutions. Finally, an additional 0.42 % DSC increase is achieved with the segmentation algorithm from LV to LVPA by incorporating the Layer-Channel Attention module which can focus on slices with high tumor correlation. Overall, LVPA-UNet improves in DSC, precision, and recall by 2.14 %, 2.96 %, and 1.01 % respectively to the base model, while reducing HD95 by 0.5434 mm. The model successfully overcomes challenges commonly faced by normal 2D, 3D, or hybrid 2D-3D models, enables efficient and accurate segmentation for GTV automatically and consistently, and eliminates subjective judgment disparities.

From a medical perspective, the accurate delineation by experts from renowned medical institutions ensures the professionalism of the ground truth data, making it a solid foundation for the reliability of the segmentation model. Notably, while Lin et al. [23] have

Table 4
The effect of different L-MSCA and V-MSCA designs of LVPA modules.

L-MSCA Design			V-MSCA Design			DSC ↑	HD95 (mm) ↓	Precision ↑	Recall ↑
[3 × 1]	[5 × 1]	[7 × 1]	3[3 × 1]	3[5 × 1]	3[7 × 1]	0.7890 ± 0.0923	2.0439 ± 3.1675	0.7814 ± 0.1095	0.8116 ± 0.1187
[3 × 1]	[5 × 1]	[7 × 1]	3[5 × 1]	3[7 × 1]	3[11 × 1]	0.7907 ± 0.0937	1.8702 ± 2.8491	0.7929 ± 0.1088	0.8025 ± 0.1201
[5 × 1]	[7 × 1]	[11 × 1]	3[3 × 1]	3[5 × 1]	3[7 × 1]	0.7901 ± 0.0933	1.9259 ± 2.9413	0.7842 ± 0.1079	0.8103 ± 0.1202
[7 × 1]	[11 × 1]	[21 × 1]	3[3 × 1]	3[5 × 1]	3[7 × 1]	0.7884 ± 0.0915	2.0829 ± 3.9428	0.7863 ± 0.1102	0.8049 ± 0.1172

demonstrated their model's NPC GTV segmentation capabilities to be comparable with those of senior radiation oncologists, LVPA-UNet presented in this study outperforms it across all essential metrics. This advancement offers substantial technical support for the accurate formulation of IMRT radiation plans. Especially for remote areas without seasoned IMRT specialists, the reliable LVPA-UNet segmentation model can bring expert-level medical services to local patients.

From a perspective of universality, LVPA-UNet is endowed with natural generalizing genes for different tissues and organs by the inherent transferability of deep learning models. The solutions to address the issues of spatial feature loss and anisotropy, variability in tumor characteristics and blurred boundaries, as well as background interference, can be readily adapted to other similar 3D MRI or CT segmentation tasks, such as BraTS [42,43] and SPPIN [44]. This offers precise segmentation technology support for a wider range of clinical scenarios.

This study has several limitations.

(a) The Diffusion Probabilistic Models (DPMs) [45–48] have recently shown outstanding performance in the field of medical image segmentation. By simulating the diffusion process between image voxels, DPMs refine image features and capture contextual information, allowing them to implicitly recognize complex patterns and subtle changes within images. These models introduce unique optimization paths for GTV segmentation that differ from LVPA-UNet, demonstrating their distinct advantages. Consequently, by integrating the strengths of DPMs to enhance LVPA-UNet, there is potential for significant advancements in GTV segmentation performance.

(b) The design of LVPA-UNet primarily focuses on optimizing the encoder stage, with limited depth optimization applied to skip connections and the decoder mechanism. There is a need to introduce attention mechanisms based on transformers or more complex fusion strategies in the future, to further enhance feature propagation [49,50]. Such advancements would not only allow for the full utilization of rich contextual information and long-distance dependencies during the decoding process but also more effectively restore spatial details in images.

(c) The dataset utilized in this study is relatively limited in size. Given the importance of large datasets for enhancing the cross-domain adaptability and generalization capability of deep learning models, acquiring more extensive datasets is necessary. Moreover, LVPA-UNet currently relies solely on T1-weighted MRIs for training and evaluation. Incorporating additional MRI modalities such as T2-weighted and CET1-weighted would significantly enrich the information available to the model, providing more detailed analyses. However, the legal sensitivities associated with medical imaging data related to specific diseases like nasopharyngeal carcinoma impose restrictions and barriers on data collection and sharing, limiting the expansion of dataset size and modality diversity. In this context, exploring methods such as generative adversarial networks (GANs), DPMs, and text-to-image generation models [51–54] to create synthetic images that closely resemble real data offers a viable solution to this challenge.

5. Conclusion

This study introduces LVPA-UNet, a network model for GTV segmentation. The model employs three strategies: parallel workflows, parallel multi-branch depth-wise strip convolutions, and Layer-Channel Attention mechanism, effectively addressing issues encountered in GTV segmentation, such as loss of spatial features and anisotropy, diversity in tumor size and shape, blurred boundaries, and background noise interference. By comparing the performance of the proposed model with eight typical models in NPC GTV segmentation tasks, the model's efficacy is successfully validated. From a medical application perspective, LVPA-UNet precisely delineates GTV, aiding in the creation of accurate radiation treatment plans, thereby improving patient treatment outcomes and quality of life. In terms of generalizability, LVPA-UNet applies to 3D segmentation tasks in similar images (e.g., MRI or CT). The synthetic images to expand multimodal information and increase data samples, investigate new skip connection and decoder network strategies will be the next work in the future, and the advantages of DPM will be also fully utilized to provide a more accurate and robust solution for the GTV segmentation task.

Ethics statement

This study receives ethics approval from the Ethics Committee of Sun Yat-Sen University Cancer Center, approval number B2021-229-01.

This study is a retrospective research project that analyzes medical records acquired from past diagnoses and treatments. The privacy and personal identity information of the subjects will be safeguarded. Consequently, this study is exempt from obtaining informed consent from patients.

Funding

This work is supported by Guangdong Basic and Applied Basic Research Foundation Grant (2023A1515011413).

Data availability statement

The source code of LVPA-UNet is available on https://github.com/XuHaoRran/LVPA-UNET. Data will be made available on request.

CRediT authorship contribution statement

Yu Zhang: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Conceptualization. **Hao-Ran Xu:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology. **Jun-Hao Wen:** Investigation. **Yu-Jun Hu:** Investigation. **Yin-Liang Diao:** Writing – review & editing. **Jun-Liang Chen:** Visualization. **Yun-Fei Xia:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] K.C.W. Wong, E.P. Hui, K.-W. Lo, W.K.J. Lam, D. Johnson, L. Li, Q. Tao, K.C.A. Chan, K.-F. To, A.D. King, B.B.Y. Ma, A.T.C. Chan, Nasopharyngeal carcinoma: an evolving paradigm, Nat. Rev. Clin. Oncol. 18 (2021) 679–695, https://doi.org/10.1038/s41571-021-00524-x.

[2] A.M. Chen, R. Chin, P. Beron, T. Yoshizaki, A.G. Mikaeilian, M. Cao, Inadequate target volume delineation and local–regional recurrence after intensity-modulated radiotherapy for human papillomavirus-positive oropharynx cancer, Radiother. Oncol. 123 (2017) 412–418, https://doi.org/10.1016/j.radonc.2017.04.015.

[3] K. Harrison, H. Pullen, C. Welsh, O. Oktay, J. Alvarez-Valle, R. Jena, Machine learning for auto-segmentation in radiotherapy planning, Clin. Oncol. 34 (2022) 74–88, https://doi.org/10.1016/j.clon.2021.12.003.

[4] Z. Chen, J. Yang, L. Chen, Z. Feng, L. Jia, Efficient railway track region segmentation algorithm based on lightweight neural network and cross-fusion decoder, Autom. ConStruct. 155 (2023) 105069, https://doi.org/10.1016/j.autcon.2023.105069.

[5] Z. Feng, J. Yang, Z. Chen, Z. Kang, LRseg: an efficient railway region extraction method based on lightweight encoder and self-correcting decoder, Expert Syst. Appl. 238 (2024) 122386, https://doi.org/10.1016/j.eswa.2023.122386.

[6] Y. Zhang, J. Niu, Z. Huang, C. Pan, Y. Xue, F. Tan, High-precision detection for sandalwood trees via improved YOLOv5s and StyleGAN, Agriculture 14 (2024), https://doi.org/10.3390/agriculture14030452.

[7] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, A. Zhou, Rethinking the learning paradigm for dynamic facial expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17958–17968.

[8] F. Liu, H.-Y. Wang, S.-Y. Shen, X. Jia, J.-Y. Hu, J.-H. Zhang, X.-Y. Wang, Y. Lei, A.-M. Zhou, J.-Y. Qi, Z.-B. Li, OPO-FCM: a computational affection based OCC-PAD-OCEAN federation cognitive modeling approach, IEEE Transactions on Computational Social Systems 10 (2023) 1813–1825, https://doi.org/10.1109/TCSS.2022.3199119.

[9] C.-X. Ren, G.-X. Xu, D.-Q. Dai, L. Lin, Y. Sun, Q.-S. Liu, Cross-site prognosis prediction for nasopharyngeal carcinoma from incomplete multi-modal data, Med. Image Anal. (2024) 103103.

[10] Y.-J. Hu, L. Zhang, Y.-P. Xiao, T.-Z. Lu, Q.-J. Guo, S.-J. Lin, L. Liu, Y.-B. Chen, Z.-L. Huang, Y. Liu, Y. Su, L.-Z. Liu, X.-C. Gong, J.-J. Pan, J.-G. Li, Y.-F. Xia, MRI-based deep learning model predicts distant metastasis and chemotherapy benefit in stage II nasopharyngeal carcinoma, iScience 26 (2023) 106932, https://doi.org/10.1016/j.isci.2023.106932.

[11] J. Cai, Y. Tang, L. Lu, A.P. Harrison, K. Yan, J. Xiao, L. Yang, R.M. Summers, Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: slice-propagated 3d mask generation from 2d recist, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11, Springer, 2018, pp. 396–404.

[12] R.P. Poudel, P. Lamata, G. Montana, Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation, in: reconstruction, segmentation, and analysis of medical images: first international workshops, RAMBO 2016 and HVSMR 2016, held in conjunction with MICCAI 2016, athens, Greece, october 17, 2016, in: Revised Selected Papers 1, Springer, 2017, pp. 83–94.

[13] W. Wang, C. Chen, J. Wang, S. Zha, Y. Zhang, J. Li, Med-DANet: dynamic architecture network for efficient medical volumetric segmentation, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI, Springer, 2022, pp. 506–522.

[14] J. Li, J. Chen, Y. Tang, C. Wang, B.A. Landman, S.K. Zhou, Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives, Med. Image Anal. (2023) 102762.

[15] P. Bilic, P. Christ, H.B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G.E.H. Mamani, G. Chartrand, others, the liver tumor segmentation benchmark (lits), Med. Image Anal. 84 (2023) 102680.

[16] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers Make Strong Encoders for Medical Image Segmentation, 2021 arXiv Preprint arXiv:2102.04306.

[17] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H.R. Roth, D. Xu, Swin unetr: swin transformers for semantic segmentation of brain tumors in mri images, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I, Springer, 2022, pp. 272–284.

[18] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, Unetr: transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 574–584.

[19] H. Peiris, M. Hayat, Z. Chen, G. Egan, M. Harandi, A robust volumetric transformer for accurate 3d tumor segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V, Springer, 2022, pp. 162–172.

[20] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: multimodal brain tumor segmentation using transformer, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 109–119.

[21] Y. Xie, J. Zhang, C. Shen, Y. Xia, Cotr: efficiently bridging cnn and transformer for 3d medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, Springer, 2021, pp. 171–180.

[22] X. Yu, Q. Yang, Y. Zhou, L.Y. Cai, R. Gao, H.H. Lee, T. Li, S. Bao, Z. Xu, T.A. Lasko, R.G. Abramson, Z. Zhang, Y. Huo, B.A. Landman, Y. Tang, UNesT: local spatial representation learning with hierarchical transformer for efficient medical segmentation, Med. Image Anal. (2023) 102939, https://doi.org/10.1016/j.media.2023.102939.

[23] L. Lin, Q. Dou, Y.-M. Jin, G.-Q. Zhou, Y.-Q. Tang, W.-L. Chen, B.-A. Su, F. Liu, C.-J. Tao, N. Jiang, others, Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma, Radiology 291 (2019) 677–686.

[24] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, K.H. Maier-Hein, Brain tumor segmentation and radiomics survival prediction: contribution to the brats 2017 challenge, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3, Springer, 2018, pp. 287–297.

[25] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, P.A. Heng, H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. http://arxiv.org/abs/1709.07330, 2018. (Accessed 5 December 2023).

[26] H. Tang, X. Liu, K. Han, X. Xie, X. Chen, H. Qian, Y. Liu, S. Sun, N. Bai, Spatial context-aware self-attention model for multi-organ segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 939–949.

[27] F. Xia, Y. Peng, J. Wang, X. Chen, A 2.5D multi-path fusion network framework with focusing on z-axis 3D joint for medical image segmentation, Biomed. Signal Process Control 91 (2024) 106049, https://doi.org/10.1016/j.bspc.2024.106049.

[28] J. Zhao, Z. Xing, Z. Chen, L. Wan, T. Han, H. Fu, L. Zhu, Uncertainty-Aware multi-dimensional mutual learning for brain and brain tumor segmentation, IEEE Journal of Biomedical and Health Informatics 27 (2023) 4362–4372, https://doi.org/10.1109/JBHI.2023.3274255.

[29] Y. Zhang, Q. Liao, L. Ding, J. Zhang, Bridging 2D and 3D segmentation networks for computation-efficient volumetric medical image segmentation: an empirical study of 2, 5 D solutions, Computerized Medical Imaging and Graphics 99 (2022) 102088.

[30] Y. Xie, B. Yang, Q. Guan, J. Zhang, Q. Wu, Y. Xia, Attention Mechanisms in Medical Image Segmentation: A Survey, 2023 arXiv Preprint arXiv:2305.17937.

[31] Y. Liu, Z. Zhang, J. Yue, W. Guo, SCANeXt: enhancing 3D medical image segmentation with dual attention network and depth-wise convolution, Heliyon 10 (2024) e26775, https://doi.org/10.1016/j.heliyon.2024.e26775.

[32] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, S.-M. Hu, Segnext: rethinking convolutional attention design for semantic segmentation, arXiv Preprint arXiv: 2209 (2022) 08575.

[33] P. Tang, C. Zu, M. Hong, R. Yan, X. Peng, J. Xiao, X. Wu, J. Zhou, L. Zhou, Y. Wang, DA-DSUnet: dual attention-based dense SU-net for automatic head-and-neck tumor segmentation in MRI images, Neurocomputing 435 (2021) 103–113.

[34] R.J. Wang, X. Li, C.X. Ling, Pelee: a real-time object detection system on mobile devices, Adv. Neural Inf. Process. Syst. 31 (2018).

[35] O. Ronneberger, P. Fischer, T. Brox, 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, 2015, pp. 234–241.

[36] D.P. Huttenlocher, G.A. Klanderman, W.J. Rucklidge, Comparing images using the Hausdorff distance, IEEE Trans. Pattern Anal. Mach. Intell. 15 (1993) 850–863.

[37] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic Differentiation in PyTorch, 2017.

[38] M.J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yangothers, MONAI: an open-source framework for deep learning in healthcare, arXiv Preprint arXiv:2211.02701 (2022). https://doi.org/10.48550/arXiv.2211.02701.

[39] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[40] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer, 2016, pp. 424–432.

[41] Z. Yu, S. Han, 3D medical image segmentation based on multi-scale MPU-net, arXiv Preprint arXiv:2307.05799 (2023). https://doi.org/10.48550/arXiv.2307.05799.

[42] A.F. Kazerooni, N. Khalili, X. Liu, D. Haldar, Z. Jiang, S.M. Anwar, J. Albrecht, M. Adewole, U. Anazodo, H. Anderson, S. Bagheri, U. Baid, T. Bergquist, A. J. Borja, E. Calabrese, V. Chung, G.-M. Conte, F. Dako, J. Eddy, I. Ezhov, A. Familiar, K. Farahani, S. Haldar, J.E. Iglesias, A. Janas, E. Johansen, B.V. Jones, F. Kofler, D. LaBella, H.A. Lai, K.V. Leemput, H.B. Li, N. Maleki, A.S. McAllister, Z. Meier, B. Menze, A.W. Moawad, K.K. Nandolia, J. Pavaine, M. Piraud, T. Poussaint, S.P. Prabhu, Z. Reitman, A. Rodriguez, J.D. Rudie, M. Sanchez-Montano, I.S. Shaikh, L.M. Shah, N. Sheth, R.T. Shinohara, W. Tu, K. Viswanathan, C. Wang, J.B. Ware, B. Wiestler, W. Wiggins, A. Zapaishchykova, M. Aboian, M. Bornhorst, P. de Blank, M. Deutsch, M. Fouladi, L. Hoffman, B. Kann, M. Lazow, L. Mikael, A. Nabavizadeh, R. Packer, A. Resnick, B. Rood, A. Vossough, S. Bakas, M.G. Linguraru, The brain tumor segmentation (BraTS) challenge 2023, Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs) (2024). https://doi.org/10.48550/arXiv.2305.17033.

[43] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F.C. Kitamura, S. Pati, others, The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification, arXiv Preprint arXiv:2107.02314 (2021). https://doi.org/10.48550/arXiv.2107.02314.

[44] M.A.D. Buser, A.F.W. van der Steeg, D.C. Simons, M.H.W.A. Wijnen, A.S. Littooij, A.H. ter Brugge, I.N. Vos, B.H.M. van der Velden, Surgical Planning in Pediatric Neuroblastoma, 2023, https://doi.org/10.5281/zenodo.7848306.

[45] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, Y. Xu, MedSegDiff: medical image segmentation with diffusion probabilistic model, in: I. Oguz, J. Noble, X. Li, M. Styner, C. Baumgartner, M. Rusu, T. Heinmann, D. Kontos, B. Landman, B. Dawant (Eds.), Medical Imaging with Deep Learning, PMLR, 2024, pp. 1623–1639, in: https://proceedings.mlr.press/v227/wu24a.html.

[46] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, Y. Xu, Medsegdiff-v2: Diffusion Based Medical Image Segmentation with Transformer, 2023 arXiv Preprint arXiv:2301.11798.

[47] Z. Xing, L. Wan, H. Fu, G. Yang, L. Zhu, Diff-unet: a diffusion embedded network for volumetric segmentation, arXiv Preprint arXiv:2303.10326 (2023). https://doi.org/10.48550/arXiv.2303.10326.

[48] A. Kazerouni, E.K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, D. Merhof, Diffusion models in medical imaging: a comprehensive survey, Med. Image Anal. 88 (2023) 102846, https://doi.org/10.1016/j.media.2023.102846.

[49] E.K. Aghdam, R. Azad, M. Zarvani, D. Merhof, Attention swin U-net: cross-contextual attention mechanism for skin lesion segmentation, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), 2023, pp. 1–5, https://doi.org/10.1109/ISBI53787.2023.10230337.

[50] M.M. Rahman, R. Marculescu, Medical image segmentation via cascaded attention decoding, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6222–6231.

[51] A. Makhlouf, M. Maayah, N. Abughanam, C. Catal, The use of generative adversarial networks in medical image augmentation, Neural Comput. Appl. 35 (2023) 24055–24068.

[52] Z. Chen, J. Yang, Z. Feng, H. Zhu, RailFOD23: a dataset for foreign object detection on railroad transmission lines, Sci. Data 11 (2024) 72.

[53] K. Kim, Y. Na, S.-J. Ye, J. Lee, S.S. Ahn, J.E. Park, H. Kim, Controllable text-to-image synthesis for multi-modality MR images, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7936–7945.

[54] S. Dayarathna, K.T. Islam, S. Uribe, G. Yang, M. Hayat, Z. Chen, Deep learning based synthesis of MRI, CT and PET: review and analysis, Med. Image Anal. 92 (2024) 103046, https://doi.org/10.1016/j.media.2023.103046.