Received: 22 February 2023 | Revised: 20 June 2023 | Accepted: 30 July 2023

DOI: 10.1002/mp.16703

RESEARCH ARTICLE

MEDICAL PHYSICS

SwinCross: Cross-modal Swin transformer for head-and-neck tumor segmentation in PET/CT images

Gary Y. Li[1] | **Junyu Chen**[2,3] | **Se-In Jang**[1] | **Kuang Gong**[1] | **Quanzheng Li**[1]

[1]Center for Advanced Medical Computing and Analysis, Massachusetts General Hospital/Harvard Medical School, Boston, Massachusetts, USA

[2]The Russell H Morgan Department of Radiology and Radiological Science, School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA

[3]Department of Electrical and Computer Engineering, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA

Correspondence
Gary Y. Li, 100 Cambridge St, Boston, MA 02114, USA.
Email: bettergary@gmail.com

Funding information
NIH, Grant/Award Numbers: C06 CA059267, R01AG078250

Abstract

Background: Radiotherapy (RT) combined with cetuximab is the standard treatment for patients with inoperable head and neck cancers. Segmentation of head and neck (H&N) tumors is a prerequisite for radiotherapy planning but a time-consuming process. In recent years, deep convolutional neural networks (DCNN) have become the de facto standard for automated image segmentation. However, due to the expensive computational cost associated with enlarging the field of view in DCNNs, their ability to model long-range dependency is still limited, and this can result in sub-optimal segmentation performance for objects with background context spanning over long distances. On the other hand, Transformer models have demonstrated excellent capabilities in capturing such long-range information in several semantic segmentation tasks performed on medical images.

Purpose: Despite the impressive representation capacity of vision transformer models, current vision transformer-based segmentation models still suffer from inconsistent and incorrect dense predictions when fed with multi-modal input data. We suspect that the power of their self-attention mechanism may be limited in extracting the complementary information that exists in multi-modal data. To this end, we propose a novel segmentation model, debuted, Cross-modal Swin Transformer (SwinCross), with cross-modal attention (CMA) module to incorporate cross-modal feature extraction at multiple resolutions.

Methods: We propose a novel architecture for cross-modal 3D semantic segmentation with two main components: (1) a cross-modal 3D Swin Transformer for integrating information from multiple modalities (PET and CT), and (2) a cross-modal shifted window attention block for learning complementary information from the modalities. To evaluate the efficacy of our approach, we conducted experiments and ablation studies on the HECKTOR 2021 challenge dataset. We compared our method against nnU-Net (the backbone of the top-5 methods in HECKTOR 2021) and other state-of-the-art transformer-based models, including UNETR and Swin UNETR. The experiments employed a five-fold cross-validation setup using PET and CT images.

Results: Empirical evidence demonstrates that our proposed method consistently outperforms the comparative techniques. This success can be attributed to the CMA module's capacity to enhance inter-modality feature representations between PET and CT during head-and-neck tumor segmentation. Notably, SwinCross consistently surpasses Swin UNETR across all five folds, showcasing its proficiency in learning multi-modal feature representations at varying resolutions through the cross-modal attention modules.

Conclusions: We introduced a cross-modal Swin Transformer for automating the delineation of head and neck tumors in PET and CT images. Our model incorporates a cross-modality attention module, enabling the

exchange of features between modalities at multiple resolutions. The experimental results establish the superiority of our method in capturing improved inter-modality correlations between PET and CT for head-and-neck tumor segmentation. Furthermore, the proposed methodology holds applicability to other semantic segmentation tasks involving different imaging modalities like SPECT/CT or PET/MRI. Code:https://github.com/yli192/SwinCross_CrossModalSwinTransformer_for_Medical_Image_Segmentation

KEYWORDS
network architecture, PEC/CT, Transformer, tumor segmentation

1 | INTRODUCTION

Head and neck (H&N) cancers are among the most common cancers worldwide,[1] accounting for about 4% of all cancers in the United States. FDG-PET and CT imaging are the gold standards for the initial staging and follow-up of H&N cancer. Quantitative image biomarkers from medical images such as radiomics have previously shown tremendous potential to optimize patient care, particularly for H&N tumors.[2] However, radiomics analyses rely on an expensive and error-prone manual process of annotating the Volume of Interest (VOI) in 3D. The automatic segmentation of H&N tumors from PET/CT images could therefore enable the validation of radiomics models on very large cohorts and with optimal reproducibility. Besides, automatic segmentation algorithms could enable a faster clinical workflow. By focusing on metabolic and anatomical features, respectively, PET and CT include complementary and synergistic information in the context of H&N primary tumor segmentation.

Recently, Transformer, a neural network based on self-attention mechanisms to compute feature representations and global dependencies, has flourished in natural language processing and computer vision.[3] In computer vision, Transformer-based architectures have achieved remarkable success and have demonstrated superior performance on a variety of tasks, including visual recognition,[4,5] objection detection,[6,7] semantic segmentation,[8,9] and more.[5,10–12] The success of vision transformers in the computer vision field has inspired their use in medical imaging, where they have shown promising potential in various applications, such as classification[13–15] segmentation,[16–18] and registration.[19,20] Chen et al. first proposed the TransUNet[16] for medical image segmentation, which used a 12-layer ViT for the bottleneck features and followed the 2D UNet design and adopted the Transformer blocks in the middle structure. Later that year, two improved versions of TransUNet, TransUNet+,[21] and Ds-TransUNet,[22] were proposed and achieved better results for CT segmentation tasks. For 3D segmentation where the computational cost for self-attention becomes very expensive, researchers have attempted to limit the use of transformer blocks, that is, only use self-attention at the bottleneck between the encoder and decoder network[23,24] or adopted a deformable mechanism which enables attention on a small set of key positions.[25] SegTran[26] proposed to leverage the learning tradeoff between larger context and localization accuracy by doing pairwise feature contextualization with squeeze and excitation blocks. More recently, more and more state-of-the-art performance has been refreshed by networks with pre-trained transformer backbone. Pre-training techniques have become a new area of research in transformers as the self-attention blocks commonly require pre-training data at a large scale to learn a more powerful backbone.[27,28] For example, self-supervised Swin UNETR[29] collects a large-scale of CT images (5000 subjects) for pretraining the Swin Transformer encoder, which derives significant improvement and state-of-the-art performance for BTCV[30] and Medical Segmentation Decathlon (MSD).[31] Self-supervised masked autoencoder (MAE)[32] investigates the MAE-based self-supervised pretraining paradigm designed for Transformers, which enforces the network to predict masked targets by collecting information from the context. Besides developing advanced architectures to better learn the data, researchers have also attempted to improve performance by providing additional data that is more specific to the task the network is given.

In representation learning, the advancement of multimodal learning has benefited numerous applications.[33,34] The utilization of fused features from multimodalities has largely improved performance in cross-media analysis tasks such as video classification,[35] event detection,[36,37] and sentiment analysis.[38,39] A characteristic that these works have demonstrated in common is that better features for one modality (e.g., audio) can be learned if multiple modalities (e.g., audio and video) are present at feature learning time. In,[40] Ngiam et al. proposed the cross-modality (audio + video) feature learning scheme for shared representation learning and demonstrated superior visual speech classification performance compared to the classifier trained with audio-only or video-only data. Wang et al. proposed a DNN-based model combining canonical correlated autoencoder and autoencoder-based terms to fuse multi-view for an unsupervised multi-view feature learning.[41] Following

this trend, deep learning-based multimodal methods have also gained attraction in the Medical Image Analysis community due to their remarkable performance in many Med Image Anal tasks including the classification,[42,43] diagnosis,[44,45] image-retrieval,[46] and segmentation.[47–50] Carneiro et al.[51] proposed to use of shared image features from unregistered views of the same region to improve classification performance. In,[44] Xu et al. proposed to jointly learn the nonlinear correlations between image and other non-image modalities for cervical dysplasia diagnosis by leveraging multimodal information, which significantly outperformed methods using any single source of information alone. In,[45] Suk et al. proposed to learn a joint feature representation from MRI and PET using a hierarchical DCNN for Alzheimer's disease diagnosis.

Although vision transformer models boast impressive representational capacities, contemporary vision transformer-based segmentation models continue to grapple with producing inconsistent and erroneous dense predictions when confronted with multi-modal input data. We speculate that the effectiveness of their self-attention mechanism is restricted in capturing the intrinsic complementary information within multi-modal data. To this end, we propose a dual-branch cross-attention Swin Transformer (SwinCross) to combine images from two different modalities at different scales to produce more complementary feature representations from the two modalities. To achieve this, we devise a cross-modal attention (CMA) module that draws inspiration from the cross-attention and shifted window self-attention mechanism found in Swin Transformer.[7] This adaptation not only streamlines computational demands but also enables effective cross-modal integration. To validate the effectiveness of the proposed method, we conducted experiments on a public dataset and compared the proposed method with state-of-the-art methods such as UNETR, Swin UNETR, and nnU-Net. The experimental results demonstrate that the proposed method surpasses the comparative segmentation methods (also with dual-modality input) by effectively capturing the inter-modality correlation between PET and CT, enhancing the performance of head-and-neck tumor segmentation.

2 | RELATED WORK

2.1 | The current state-of-the-art methods for H&N tumor segmentation

The top-five performing teams in the HECTOR 2021 challenge all used U-Net or its variants for the primary H&N tumor segmentation task.[52] In,[53] Xie et al. used a patch-based 3D nnU-Net with Squeeze and Excitation normalization and a novel training scheme, where the learning rate is adjusted dynamically using polyLR.[54] The approach achieved a five-fold average Dice score of 0.764 on the validation data dataset, which ranked them first on the leaderboard for the tumor segmentation task. They trained five models in a five-fold cross-validation manner with random data augmentation including rotation, scaling, mirroring, Gaussian noise, and Gamma correction. The final test results were generated by ensembling five test predictions via probability averaging. In,[55] An et al. proposed a coarse-to-fine framework using a cascade of three U-Nets. The first U-Net is used to coarsely segment the tumor and then select a bounding box. Then, the second U-Net performs a finer segmentation on the smaller region within the bounding box, which has been shown to often lead to more accurate segmentation.[56] Finally, the last U-Net takes as input the concatenation of PET, CT, and the previous segmentation to refine the predictions. The three U-Nets were trained with different objectives—the first one to optimize the recall and the rest two to optimize the Dice score. The final results were obtained via majority voting on three different predictions: an ensemble of five nnU-Nets, an ensemble of three U-Nets with squeeze-and-excitation (SE) normalization, and the predictions from the proposed model. In,[57] Lu et al. proposed a huge ensemble learning model, which consists of fourteen 3D U-Nets, including the eight models adopted in,[58] winner of the HECTOR 2020 challenge, five models trained with leave-on-center-out, and one model combining a prior and posteriori attention. The final ensembled prediction was generated by averaging all fourteen predictions and thresholding the resulting mask to 0.5. In,[59] Yousefirizi et al. used a 3D nnU-Net with SE normalization trained on a leave-one-center-out with a combination of a "unified" focal and Mumford-Shah losses, leveraging the advantage of distribution, region, and boundary-based loss functions. Lastly, Ren et al.[60] proposed a 3D nnU-Net with various PET normalization techniques, namely, PET-clip and PET-sin. The former clips the standardized uptake values (SUV) range in [0,5] and the latter transforms monotonic spatial SUV increase into onion rings via a sine transform of SUV, which ranked them fifth on the leaderboard. Despite the widespread success of CNNs in the H&N tumor segmentation task and medical imaging applications at large, there are still inherent constraints within the architecture that fundamentally impedes CNNs from attaining even greater levels of performance. Due to the prevalent use of small convolution kernels (3×3 or 5×5) in the majority of current CNNs, the convolution operations primarily focus on local spatial structures.[3] As a consequence, CNNs exhibit a bias towards capturing and emphasizing local information.[61,62] Although extensive efforts have been made to address such limitation by increasing the depth of the network,[63] introducing dilated convolutions,[54,64] deploying recurrent-,[65] or residual-[66] connections, the initial layers of CNNs still have very limited receptive fields (RFs), which restricts their ability to explicitly

capture long-range spatial dependencies. It is only at the deeper layers that such long-range dependencies can be implicitly modeled. However, it has been observed that as CNNs become deeper, the influence of distant voxels diminishes rapidly.[67] As a result, the effective receptive fields (RFs) of these CNNs are significantly smaller than their theoretical RFs, despite the theoretical RFs encompassing the entire input image.[3] This inherent limitation of receptive size sets an obstacle to learning global semantic information, which is critical for dense prediction tasks like segmentation.

2.2 | Transformers and multi-modal learning

Transformers have been widely applied in the fields of Natural Language Processing[68–70] and Computer Vision[71–76] primarily due to its excellent capability to model long-range dependency. Besides achieving impressive performance in a variety of language and vision tasks, the Transformer model also provides an effective mechanism for multi-modal reasoning by taking different modality inputs as tokens for self-attention.[77–86] For example, Prakash et al.[82] proposed to use a Transformer to integrate image and LiDAR representations using attention. Going beyond language and vision, we propose to utilize a cross-modal attention Swin Transformer to fuse 3D PET and CT images at multiple resolutions for the segmentation of H&N tumors. We build the SwinCross architecture based on the shifted window block from Swin Transformer, which only computes self-attention within local regions, unlike conventional ViTs, which are more computationally expensive. Although Swin Transformer is unable to explicitly compute correspondences beyond its field of view, similar to how ConvNets operate to some extent, the shifted window mechanism still yields much larger kernels than most ConvNets.[87]

3 | SWINCROSS

3.1 | Overall architecture of SwinCross

In this work, we propose an architecture for 3D multi-modal segmentation with two main components: (1) a cross-modal Swin Transformer for integrating information from multiple modalities (PET and CT), and (2) a cross-modal shifted window attention block for learning complementary information from the modalities. Our key idea is to exploit the cross-modal attention mechanism to incorporate the global context for PET and CT modalities given their complementary nature for the H&N tumor segmentation task. We illustrate the architecture of SwinCross in Figure 1. The input image to the SwinCross model is multi-channel 3D volume $F^{in} \in R^{H\times W\times D\times M}$, with a dimension of $H \times W \times D \times M$. The input image is first split channel-wise, forming a set of single-channel 3D images $F_{mod\_1}, \ldots, F_{mod\_k} \in R^{H\times W\times D}$. Then, we split each single-channel image into small non-overlapped patches with a patch size of $\frac{H}{H'} \times \frac{W}{W'} \times \frac{D}{D'}$, which corresponds to a patch resolution of $H' \times W' \times D'$. Each 3D patch is projected into an embedding space with dimension C to form a tokenized sequence $S_{mod\_k} \in R^{N\times C}$, where $N = H' \times W' \times D'$ is the number of tokens in the sequence and each token is represented by a feature vector of dimensionality C. The S^{mod_K} sequences are inputs to the encoder network.

3.2 | Cross-modal attention module

The cross-modal attention (CMA) module serves the purpose of learning complementary information from different modalities to accomplish a shared task. In the case of bimodal cross-attention, the CMA module utilizes features from one modality as the "key" and features from another modality as the "query." It then generates attention weights based on these features and employs them to filter/attend to the features from the "key" modality. The resulting attention weights represent a weighted sum that indicates the importance of features from the "key" modality that complements (queried by the other modality) the features from the "query" modality in achieving their common task.

For instance, in the context of PET/CT tumor segmentation, attention weights computed from a PET "query" and CT "key" aim to extract additional features from CT, such as fine boundaries, that may not be present in the PET features but are crucial for the final segmentation task. This is due to the inherent complementary nature of multimodal imaging procedures, which often relate to the distinctive image formation mechanisms of each modality. Conversely, attention weights derived from a CT "query" and PET "key" aim to extract additional features from PET, such as functional aspects, which compensate for the limited capability of CT in revealing functional features like tissue metabolism. Both streams of information are vital for the successful completion of the downstream task. A pictorial illustration of the cross-modal attention mechanism is in Figure 2.

Mathematically, for bimodal cross-attention, the CMA uses the scaled dot products between the query (Q) and key (K) of each modality to compute the attention weights and then aggregates the values for each modality,

$$A_{mod_1}\left(Q_{mod_2}, K_{mod_1}, V_{mod_1}\right) = \text{Softmax}\left(\frac{Q_{mod_2} K_{mod_1}^{T}}{\sqrt{D_k}}\right) V_{mod_1}, \quad (1)$$

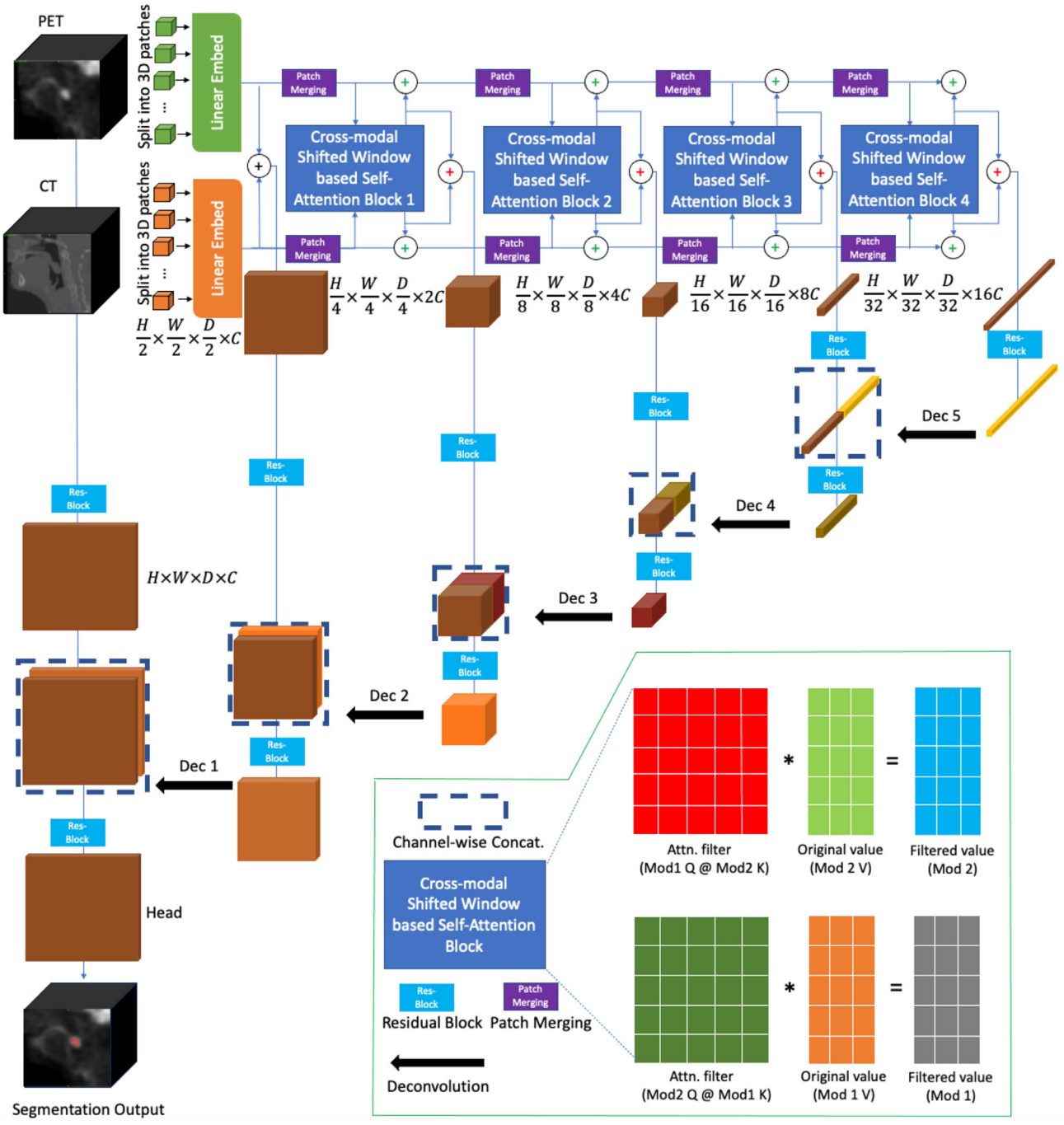

FIGURE 1 Architecture of SwinCross. 3D PET and CT volumes are used as inputs to our Cross-modal attention Swin Transformer (SwinCross) which adopts multiple cross-modal attention (CMA) modules for the fusion of intermediate feature maps between the two modalities. To effectively combine patch tokens from both modalities at different scales, we develop a fusion method based on the CMA blocks, which exchange information between two branches at multiple resolutions ($\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, *and* $\frac{1}{32}$ of the input resolution) throughout the two feature extracting branches resulting in five feature vectors ($\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, *and* $\frac{1}{32}$ of the input resolution) from both modalities, which are combined via element-wise summation. The five feature vectors constitute fused representations of the CT and PET image at five different resolutions. These feature vectors are then processed with a ConvNet decoder which predicts the final segmentation map. We channel-wise concatenate the decoded feature vectors from a previous resolution to the feature vector at the current resolution and use the resulting feature vectors as input to the deconvolution block to produce the feature vector at the next resolution.

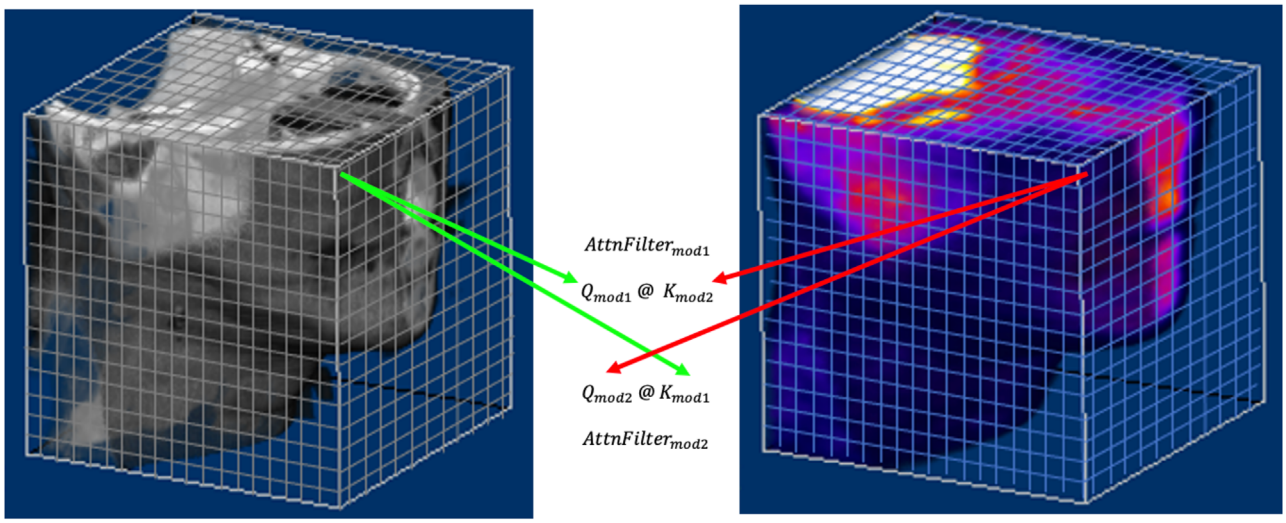

FIGURE 2 Illustration of the cross-modality attention mechanism in the proposed CMA module. In this illustration, the features corresponding to each modality within each window (visualized as individual cubes) serve a dual role as both queries and keys. To illustrate, considering the context of tumor segmentation using PET/CT data, the attention weights computed using a PET "query" and a CT "key" are intended to capture supplementary features from the CT data, such as intricate boundaries. These boundary features might not be prominent in the PET data but hold significance for achieving accurate segmentation results. Conversely, attention weights obtained from a CT "query" and a PET "key" are aimed at extracting supplementary features from the PET data. These additional features, such as functional characteristics, serve to compensate for the inherent limitations of CT in revealing functional attributes like tissue metabolism.

$$A_{mod_2}\left(Q_{mod_1}, K_{mod_2}, V_{mod_2}\right) = \text{Softmax}\left(\frac{Q_{mod_1} K_{mod_2}^T}{\sqrt{D_k}}\right) V_{mod_2}, \quad (2)$$

in which Q_{mod_1}, K_{mod_1}, V_{mod_1}, Q_{mod_2}, K_{mod_2}, V_{mod_2} denote queries, keys, and values from modality 1 and 2, respectively; D_k represents the size of the key and query.

3.3 | Network encoder

The encoder uses linear projections for computing a set of queries, keys, and values (Q, K, and V) for each input sequence S_{mod_k}.

$$Q_{mod\_k} = S_{mod\_k} M^q, \quad K_{mod\_k} = S_{mod\_k} M^k, \quad V_{mod\_k} = S_{mod\_k} M^v \quad (3)$$

where $M^q \in R^{D_f \times D_q}$, $M^k \in R^{D_f \times D_k}$, and $M^v \in R^{D_f \times D_v}$ are weight matrices.

As these are 3D tokens and the attention computation cost increases quadratically with respect to the number of tokens, we adopted the shifted window mechanism for the cross-attention calculation. Specifically, we utilized windows of size $M \times M \times M$ to evenly partition the patchified volume into $\frac{H'}{M} \times \frac{W'}{M} \times \frac{D'}{M}$ regions at a given layer l in the Transformer encoder. In the subsequent layers of l and $l+1$ of the encoder, the outputs are calculated as

$$\hat{A}^l_{mod\_k} = \text{W-MSA}\left(\text{LN}\left(A^{l-1}_{mod_k}\right)\right) + A^{l-1}_{mod_k} \quad (4)$$

$$A^l_{mod_k} = \text{MLP}\left(\text{LN}\left(\hat{A}^l_{mod\_k}\right)\right) + \hat{A}^l_{mod\_k} \quad (5)$$

$$\hat{A}^{l+1}_{mod\_k} = \text{SW-MSA}\left(\text{LN}\left(A^l_{mod_k}\right)\right) + A^l_{mod_k} \quad (6)$$

$$A^{l+1}_{mod_k} = \text{MLP}\left(\text{LN}\left(\hat{A}^{l+1}_{mod\_k}\right)\right) + \hat{A}^{l+1}_{mod\_k} \quad (7)$$

A 3D version of the cyclic-shifting[7] was implemented for efficient computation of the shifted window mechanism. SwinCross follows a standard four-stage structure[7] but has a cross-modality attention mechanism at each stage for the fusion of intermediate feature maps between both modalities. The fusion is applied at multiple resolutions ($\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times C$, $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 2C$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 4C$, $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 8C$) throughout the feature extractor of both modalities resulting in four filtered feature maps ($\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 2C$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 4C$, $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 8C$, $\frac{H}{32} \times \frac{W}{32} \times \frac{D}{32} \times 16C$) from each modality. The filtered feature maps from both modalities are summed element-wise and sent to the decoder, as indicated by the red plus signs on Figure 1. At each stage, these feature maps are fed back into each of the individual modality branches using an element-wise

TABLE 1 Ablation study of CMA module on HECKTOR 2021 dataset.

Model	Block composition	Embed dimension	Feature size	Number of blocks	Window size	Number of heads	Five-fold average dice score
Swin UNETR	W-MSA + SW-MSA	768	48	[2,2,2,2]	[7,7,7]	[3,6,12,24]	0.754 ± 0.032
	CMW-MSA + CMSW-MSA	768	48	[2,2,2,2]	[7,7,7]	[3,6,12,24]	**0.769 ± 0.026**

summation with the down-sampled (via patch merging) input feature maps, as indicated by the green plus signs on Figure 1.

We used an encoder with a patch size of $2 \times 2 \times 2$ and a feature dimension of $2 \times 2 \times 2 \times 2 = 16$, taking into account the multi-modal PET/CT images with two channels. The size of the embedding space C was set to 48 in our encoder. Furthermore, the encoder had four stages which comprise of [2, 4, 2, 2] cross-modal shifted window Transformer blocks at each stage. Hence, the total number of layers in the encoder was L = 10. Before stage 1, each single-channel image is split into small non-overlapped patches by a 3D convolution layer with stride size equal to 2 (patch size) and output channels equal to C, resulting in $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times C$ 3D tokens. To follow the hierarchical structure proposed in,[7] a patch merging layer was used on each modality branch to decrease the resolution of the feature representations by a factor of 2 at the beginning of each stage. In order to preserve fine details from the input image to the output segmentation, we sent the original input multi-channel 3D volume $F^{in} \in R^{H \times W \times D \times M}$ and its embedded version together with the feature map outputs from the 4 stages to the decoder, resulting in a total of 6 feature maps with dimensions of $H \times W \times D \times M$, $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times C$, $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 2C$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 4C$, $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 8C$, and $\frac{H}{32} \times \frac{W}{32} \times \frac{D}{32} \times 16C$.

3.4 | Network decoder

We adopted a ConvNet decoder as opposed to a Transformer decoder for the ease of cross-modal feature fusion and lower computational cost. SwinCross adopts a U-shaped network design in which the extracted feature representations of the encoder are used in the decoder via skip connections at each resolution. At each stage i ($i \in [0, 1, 2, 3, 4, 5]$) of the encoder, the output feature representations are reshaped into size $\frac{H}{2^i} \times \frac{W}{2^i} \times \frac{D}{2^i}$ and fed into a residual block comprising of two 3×3×3 convolutional layers that are normalized by instance normalization layers. Subsequently, the resolution of the feature maps is increased by a factor of 2 using a deconvolutional layer and the outputs are concatenated with the outputs of the previous stage. The concatenated features are then fed into another residual block as previously described. The final segmentation outputs are computed by using a 1×1×1 convolutional layer and a sigmoid activation function.

4 | RESULTS AND DISCUSSION

4.1 | Datasets

The study utilized two publicly available datasets for Head and Neck tumor segmentation: the HECTOR challenge dataset and the TCIA HNC dataset. In these datasets, the primary gross tumor volume (GTVt) for the patients was annotated by expert radiologists (in the case of HECTOR) and derived from histological data (in the case of TCIA).

The HECTOR challenge training dataset consisted of 224 cases, out of which 44 were reserved for validation. Each case in this dataset included two modalities: PET and CT, which were rigidly aligned and resampled to achieve a 1 × 1 × 1 mm isotropic resolution. The HECTOR data underwent preprocessing using the provided codes from the challenge website. The input image size for the HECTOR dataset was set to 144× 144×144.

The TCIA dataset comprised 122 cases, with 24 cases used for validation purposes. It contained the same imaging modalities as the HECTOR dataset (PET and CT), and the images were also resampled to the same isotropic resolution. The input image size for the TCIA dataset was defined as 128×128×128.

4.2 | Ablation studies on HECKTOR 2021 dataset

In Table 1, we ablate the CMA module block, which only concerns the attention mechanism of the Swin Transformer, and we keep everything else the same as in Swin UNETR (e.g., embed dimension, feature size, number of blocks in each stage, window size, and number of heads). We start from channel-wise concatenated input, which consists of two volume images from both modalities. This multi-modal input already gives Swin UNETR a strong five-fold average Dice Score of 0.754 ± 0.032. If we send in the images from two modalities in two separate branches (as shown in Figure 1) and use CMA

TABLE 2 Five-fold cross-validation benchmarks in terms of mean Dice score values from all methods on HECTOR 2021 dataset (PET+CT).

Dice score	nnU-Net	UNETR	Swin UNETR	SwinCross (proposed)
Fold0	0.714	0.702	0.715	**0.717**
Fold1	0.781	0.716	0.781	**0.788**
Fold2	**0.803**	0.727	0.752	0.800
Fold3	0.777	0.762	0.772	**0.779**
Fold4	**0.761**	0.708	0.748	**0.761**
Average	0.767	0.723	0.754	**0.769**

module block to fuse the learned features from each modality at each stage, the performance is improved to 0.769 ± 0.026. The output from each CMA module block has the same shape as the input and each filtered feature is added back to the corresponding modality's branch. At each stage, the sum of the filtered features from the CMA module block is sent to the decoder.

4.3 | Comparison to the state-of-the-art methods in medical image segmentation

We have compared the performance of SwinCross against the current SOTA methods in medical image segmentation such as Swin UNETR, UNETR, and nnU-Net, using a five-fold cross-validation split. Evaluation results (dual-modality) across all five folds are presented in Table 2. The proposed SwinCross model achieved the highest five-fold average Dice score of 0.769 among all the comparing methods. Note that SwinCross outperformed Swin UNETR across all five folds, which demonstrated its capability of learning multi-modal feature representations at multiple resolutions via the cross-modal

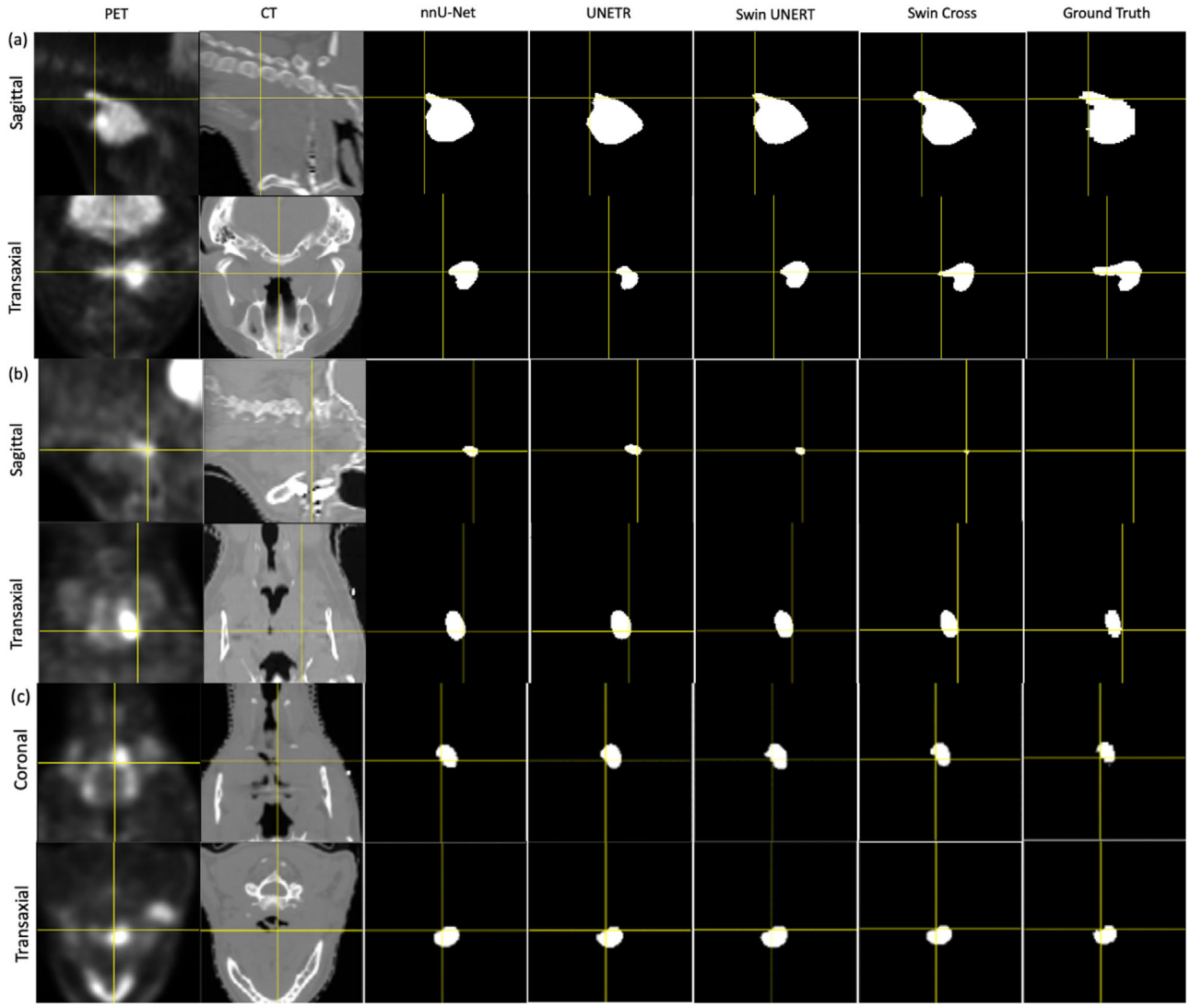

FIGURE 3 From left to right are input PET image, CT image, inferenced mask from nnU-Net, UNETR, Swin UNETR, SwinCross (proposed), and ground truth.

TABLE 3 Summary of mean Dice score values on two public datasets (PET+CT, single split).

Dice score	nnU-Net	UNETR	Swin UNETR	SwinCross (proposed)
TCIA	0.765 ± 0.082	0.708 ± 0.088	0.741 ± 0.083	**0.768 ± 0.079**
HECTOR2021	0.761 ± 0.154	0.708 ± 0.250	0.748 ± 0.191	**0.764 ± 0.151**

The bold value on each row marks/indicates the highest mean Dice score among all methods.

attention modules. These results are consistent with our previous findings,[88] in which we showed that nnU-Net outperformed Swin UNETR for H&N tumor segmentation on two public datasets. With the CMA block and dual-branch fusion mechanism, SwinCross demonstrates a slightly better segmentation performance than nnU-Net, as measured by the five-fold average Dice score. However, competitive performance is seen from nnU-Net, which again indicates that for small object segmentation, the improvement from modeling long-range dependency may be limited as a smaller effective field may be enough to capture all the foreground and background information of the small object such as a H&N tumor.[88]

In Figure 3, we present sample inference mask outputs from all the segmentation methods explored in this paper. Notably, for large tumors (depicted in Figure 3a), it is evident that our proposed method is the only network capable of capturing the precise tip of the tumor, as indicated by the yellow crosshair. This result highlights the advantage of leveraging Transformer-based models, which effectively model long-range dependencies in the data. Furthermore, for smaller tumors (illustrated in Figure 3c), SwinCross demonstrates its ability to capture the fine edges of the tumor by integrating complementary edge features from the CT image. This outperforms the other methods that rely on channel-wise concatenated inputs. The superiority of SwinCross can be attributed to its employment of the CMA module blocks, which facilitate feature fusion at multiple scales within the encoder. This enables SwinCross to effectively incorporate and leverage the benefits of complementary information, leading to enhanced segmentation results, particularly in capturing fine tumor edges. The observed improvement in performance can be attributed to the feature-level fusion design employed by SwinCross. This within-network feature fusion design has been shown to be generally better than alternative fusion mechanisms, such as those applied at the input (e.g., channel-wise concatenation) and output levels (e.g., voting), in the context of multimodal tumor segmentation.[34]

4.4 | Evaluation on TCIA dataset

We ran these experiments on one additional public head-and-neck primary tumor segmentation dataset. The summary of segmentation test results is summarized in Table 3. The results revealed that SwinCross outperformed the other three models on two public datasets. By incorporating the proposed CMA module, SwinCross demonstrated a significant performance advantage over Swin UNETR and ultimately surpassed nnU-Net in the task of head-and-neck primary tumor segmentation in PET and CT images. The results emphasize the effectiveness of SwinCross and highlight its superior performance in achieving accurate and precise tumor segmentation compared to the other models tested.

5 | CONCLUSION

A cross-modal Swin Transformer was introduced for the automatic delineation of head and neck tumors in PET and CT images. The proposed model has a cross-modal attention module that uses feature exchange between two modalities at multiple resolutions. A ConvNet-based decoder is connected to the encoder via skip connections at different resolutions. We have validated the effectiveness of our proposed model by comparing with the state-of-the-art methods using the HECTOR 2021 dataset. Through experimental validation, the proposed method demonstrates superior performance compared to the prevailing state-of-the-art segmentation techniques in the task of head-and-neck tumor segmentation, utilizing PET and CT images. The method proposed is generally applicable to other semantic segmentation tasks using dual imaging modalities such as SPECT/CT, or PET/ MRI.

ACKNOWLEDGMENTS

The authors extend their gratitude to the MGH and BWH Center for Clinical Data Science for providing the necessary computational resources.

This work was supported by NIH under grant number C06 CA059267 and R01AG078250. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CONFLICT OF INTEREST STATEMENT

No potential conflicts of interest relevant to this article exist.

REFERENCES

1. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin*. 2005;55(2):74-108.
2. Andrearczyk V, Oreiller V, Jreige M, et al. *Overview of the HECKTOR Challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT*. Springer International Publishing; 2021.
3. Li J, Chen J, Tang Y, Wang C, Landman BA, Zhou SK. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Med Image Anal*. 2023:102762.
4. Yuan Li, Hou Q, Jiang Z, Feng J, Yan S. VOLO: vision Outlooker for Visual Recognition. *IEEE Trans Pattern Anal Mach Intell*. 2022:1-13.
5. Yang C, Wang Y, Zhang J, et al. Lite Vision Transformer with enhanced self-attention. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
6. Wang W, Xie E, Li X, et al. *Pvtv2: Improved Baselines with Pyramid Vision Transformer*. Computational Visual Media; 2022.
7. Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
8. Cheng B, Schwing AG, Kirillov A. Per-Pixel classification is not all you need for semantic segmentation. *Adv Neural Inf Process*. 2021;34:17864-17875.
9. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with Transformers. *Adv Neural Inf Process*. 2021;34:12077-12090.
10. Luo ZL, Wang W, Xie E, et al. Panoptic SegFormer: delving deeper into panoptic segmentation with transformers. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
11. Wan Z, Zhang J, Chen D, Liao J. High-fidelity pluralistic image completion with Transformers. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
12. Wang H, Zhu Y, Adam H, Yuille A, Chen L-C. MaX-DeepLab: end-to-end panoptic segmentation with Mask Transformers. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
13. Hou B, Kaissis G, Summers RM, Kainz B. Ratchet: medical transformer for chest x-ray diagnosis and reporting. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference*, September 27–October 1, 2021, Proceedings, Part VII 24, Strasbourg, France. 2021. Springer.
14. Matsoukas C, Haslum JF, Söderberg M, Smith K, *Is it time to replace cnns with transformers for medical images?* arXiv preprint arXiv:2108.09038. 2021.
15. Park S, Kim G, Kim J, Kim B, Ye JC. Federated split task-agnostic vision transformer for COVID-19 CXR diagnosis. *Adv Neural Inf Process Sys*. 2021;34:24617-24630.
16. Chen J, Lu Y, Yu Q, et al. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv.org. 2021.
17. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth H, Xu D. Swin unetr: swin transformers for semantic segmentation of brain tumors in mri images. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event. Springer; 2022. *September 27, 2021, Revised Selected Papers, Part I*.
18. Hatamizadeh A, Tang Y, Nath V, et al. UNETR: transformers for 3D medical image segmentation. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022.
19. Chen J, Frey EC, He Y, Segars WP, Li Ye, Du Y. Transmorph: transformer for unsupervised medical image registration. *Med Image Anal*. 2022;82:102615.
20. Chen J, He Y, Frey EC, Li Y, Du Y, Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468, 2021.
21. Liu Y, Wang H, Chen Z, Huangliang K, Zhang H. *TransUNet+: Redesigning the skip connection to enhance features in medical image segmentation*. Knowledge-Based Systems; 2022:256.
22. Lin A, Chen B, Xu J, Zhang Z, Lu G. DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation. arXiv.org. 2021.
23. Chang Y, Menghan H, Guangtao Z, Xiao-Ping Z. TransClaw U-Net: Claw U-Net with Transformers for Medical Image Segmentation. ArXiv. 2021. abs/2107.05188.
24. Deng K, Meng Y, Gao D, et al. TransBridge: a lightweight transformer for left ventricle segmentation in echocardiography. *Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021, Held in Conjunction with MICCAI 2021*. 2021:63-72. *September 27, 2021, Proceedings*.
25. Xie Y, Zhang J, Shen C, Xia Y. CoTr: efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. *International conference on medical image computing and computer-assisted intervention*. 2021.
26. Li S, Sui X, Luo X, Xu X, Liu Y, Goh R. Medical image segmentation using Squeeze-and-Expansion Transformers. ArXiv. 2021. **abs/2105.09511**.
27. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. *arXiv 2020*. arXiv preprint arXiv:2010.11929, 2010.
28. Jiang J, Tyagi N, Tringale K, Crane C, Veeraraghavan H, Self-supervised 3D anatomy segmentation using self-distilled masked image transformer (SMIT). *Medical Image Computing and Computer Assisted Intervention–MICCAI2022: 25th International Conference*, Singapore, *September 18–22, 2022, Proceedings, Part IV*. 2022. Springer.
29. Tang Y, Yang D, Li W, et al. Self-supervised pre-training of swin transformers for 3d Med Image Anal. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
30. Landman B, Xu Zhoubing, Igelsias, J, Styner M, Langerak, T, Klein, A. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. Proc.*MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*. Vol 5. 2015:12.
31. Antonelli M, Reinke A, Bakas S, et al. The medical segmentation Decathlon. *Nat Commun*. 2022;13(1):4128.
32. Zhou L, Liu H, Bae J, He J, Samaras D, Prasanna P. Self pre-training with masked autoencoders for Med Image Anal. arXiv preprint arXiv:2203.05573, 2022.
33. Guo WZ, Wang JW, Wang SP. *Deep Multimodal Representation Learning: A Survey*. Ieee Access; 2019:63373-63394.
34. Guo Z, Li X, Huang H, Guo N, Li Q. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Trans Radiat Plasma Med Sci*. 2019;3(2):162-169.
35. Liu Y, Feng X, Zhou Z. Multimodal video classification with stacked contractive autoencoders. *Signal Process*. 2016;120:761-766.
36. Wu S, Bondugula S, Luisier F, Zhuang X, Natarajan P. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.
37. Habibian A, Mensink T, Snoek CGM. Video2vec embeddings recognize events when examples are scarce. *IEEE Trans Pattern Anal Mach Intell*. 2016;39(10):2089-2103.
38. Poria S, Cambria E, Howard N, Huang G-B, Hussain A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*. 2016;174:50-59.
39. Zadeh A, Chen M, Poria S, Cambria E, Morency L. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250. 2017.
40. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. *Multimodal Deep Learning*. ICML; 2011.

41. Wang W, Arora R, Livescu K, Bilmes J. On Deep Multi-View Representation Learning: Objectives and Optimization. ArXiv, 2016. abs/1602.01024.
42. Zhu Z, Luo P, Wang X, Tang X. Multi-view perceptron: a deep model for learning face identity and view representations. *28th Conference on Neural Information Processing Systems (NIPS)*. 2014. Montreal, Canada.
43. Carneiro G, Nascimento J, Bradley AP. Unregistered multiview mammogram analysis with pre-trained deep learning models. *18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2015. Munich, Germany.
44. Xu T, Zhang H, Huang X, Zhang S, Metaxas DN. *Multimodal Deep Learning for Cervical Dysplasia Diagnosis*. MICCAI; 2016:115-123.
45. Suk H-I, Lee S-W, Shen D. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage*. 2014;101:569-582.
46. Kang Y, Kim S, Choi S. Deep learning to hash with multiple representations. *12th IEEE International Conference on Data Mining (ICDM)*. 2012. Brussels, Belgium.
47. Zhu X, Wu Y, Hu H, et al. Medical lesion segmentation by combining multimodal images with modality weighted UNet. *Med Phys*. 2022;49(6):3692-3704.
48. Guo Z, Li X, Huang H, Guo N, Li Q, *Medical Image Segmentation Based On Multi-Modal Convolutional Neural Network: Study On Image Fusion Schemes*. *15th IEEE International Symposium on Biomedical Imaging (ISBI)*. 2018. Washington, DC.
49. Guo Z, Guo N, Gong K, Zhong S'A, Li Q. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. *Phys Med Biol*. 2019;64(20):205015.
50. Chen J, Li Y, Luna LP, et al. Learning fuzzy clustering for SPECT/CT segmentation via convolutional neural networks. *Med Phys*. 2021;48(7):3860-3877.
51. Carneiro G, Nascimento J, Bradley AP. Unregistered multiview mammogram analysis with pre-trained deep learning models. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2015.
52. Andrearczyk V, Oreiller V, Boughdad S, et al. *Overview of the HECKTOR Challenge at MICCAI 2021: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT Images*. Springer International Publishing; 2022.
53. Xie J, Peng Y. *The Head and Neck Tumor Segmentation Based on 3D U-Net*. Springer International Publishing; 2022.
54. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans Pattern Anal Mach Intell*. 2017;40(4):834-848.
55. An C, Chen H, Wang L. *A Coarse-to-Fine Framework for Head and Neck Tumor Segmentation in CT and PET Images*. Springer International Publishing; 2022.
56. Zhou Y, Xie L, Shen W, Wang Y, Fishman EK, Yuille AL. *A Fixed-Point Model for Pancreas Segmentation in Abdominal CT Scans*. Springer International Publishing; 2017.
57. Lu J, Lei W, Gu R, Wang G. *Priori and Posteriori Attention for Generalizing Head and Neck Tumors Segmentation*. Springer International Publishing; 2022.
58. Iantsen A, Visvikis D, Hatt M. *Squeeze-and-Excitation Normalization for Automated Delineation of Head and Neck Primary Tumors in Combined PET and CT Images*. Springer International Publishing; 2021.
59. Yousefirizi F, Janzen I, Dubljevic N, et al. *Segmentation and Risk Score Prediction of Head and Neck Cancers in PET/CT Volumes with 3D U-Net and Cox Proportional Hazard Neural Networks*. Springer International Publishing; 2022.
60. Ren J, Huynh B-N, Groendahl AR, Tomic O, Futsaether CM, Korreman SS. *PET Normalizations to Improve Deep Learning Auto-Segmentation of Head and Neck Tumors in 3D PET/CT*. Springer International Publishing; 2022.
61. Zhou H-Y, Lu C, Yang S, Yu Y. ConvNets vs. Transformers: whose visual representations are more transferable? *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
62. Naseer MM, Ranasinghe K, Khan S, Hayat M, Khan FS, Yang M. Intriguing properties of vision transformers. *Adv Neural Inf Process Sys*. 2021;34:23296-23308.
63. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
64. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.
65. Liang M, Hu X. Recurrent convolutional neural network for object recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
66. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
67. Luo W, Li Y, Zemel R. *Understanding the effective receptive field in deep convolutional neural networks*. Adv Neural Inf Process Sys; 2016:29.
68. Vaswani A, Shazeer N, Polosukhin I, et al. Attention Is All You Need. *31st Annual Conference on Neural Information Processing Systems (NIPS)*. 2017. Long Beach, CA.
69. Devlin J, Chang M, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
70. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Sys*. 2020;33:1877-1901.
71. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
72. Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer. *International conference on machine learning*. 2018. PMLR.
73. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
74. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. *European conference on computer vision*. 2020. Springer.
75. Chen M, Radford A, Sutskever I, et al. Generative pretraining from pixels. *Proceedings of the 37th International Conference on Machine Learning*. Hal D III, Aarti S, editors. 2020, PMLR: Proceedings of Machine Learning Research. pp. 1691-1703.
76. Chen C-FR, Fan Q, Panda R. Crossvit: cross-attention multi-scale vision transformer for image classification. *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
77. Tan H, Bansal M. Lxmert: learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019.
78. Li LH, Yatskar M, Yin D, Hsieh C-J, Chang K-W. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019.
79. Sun C, Myers A, Vondrick C, Murphy K, Schmid C. Videobert: a joint model for video and language representation learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
80. Chen Y-C, Li L, Yu L, et al. Uniter: universal image-text representation learning. *European conference on computer vision*. 2020. Springer.
81. Li X, Yin X, Li C, et al. Oscar: object-semantics aligned pre-training for vision-language tasks. *European Conference on Computer Vision*. 2020. Springer.
82. Prakash A, Chitta K, Geiger A. Multi-modal fusion transformer for end-to-end autonomous driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

83. Huang Z, Zeng Z, Huang Y, Liu B, Fu D, Fu J. Seeing out of the box: end-to-end pre-training for vision-language representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
84. Hu R, Singh A. Unit: multimodal multitask learning with a unified transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
85. Akbari H, Yuan L, Qian R, et al. Vatt: transformers for multimodal self-supervised learning from raw video, audio and text. *Adv Neural Inf Process Sys*. 2021;34:24206-24221.
86. Hendricks LA, Mellor J, Schneider R, Alayrac J-B, Nematzadeh A. Decoupling the role of data, attention, and losses in multimodal transformers. *Trans Assoc Comput Linguist*. 2021;9:570-585.
87. Ding X, Zhang X, Zhou Y, Han J, Ding G, Sun J. Scaling up your kernels to 31×31: revisiting large kernel design in cnns. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
88. Li Y, Chen J, Jang S, Gong K, Li Q. Investigation of Network Architecture for Multimodal Head-and-Neck Tumor Segmentation. Arxiv.org, 2022.

How to cite this article: Li GY, Chen J, Jang S-I, Gong K, Li Q. SwinCross: Cross-modal Swin transformer for head-and-neck tumor segmentation in PET/CT images. *Med Phys*. 2024;51:2096–2107.
https://doi.org/10.1002/mp.16703