

Large Language Model-Augmented Auto-Delineation of Treatment Target Volume in Radiation Therapy

Praveenbalaji Rajendran, Yong Yang, Thomas R. Niedermayr, Michael Gensheimer, Beth Beadle, Quynh-Thu Le, Lei Xing, and Xianjin Dai

Abstract— Radiation therapy (RT) is one of the most effective treatments for cancer, and its success relies on the accurate delineation of targets. However, target delineation is a comprehensive medical decision that currently relies purely on manual processes by human experts. Manual delineation is time-consuming, laborious, and subject to interobserver variations. Although the advancements in artificial intelligence (AI) techniques have significantly enhanced the auto-contouring of normal tissues, accurate delineation of RT target volumes remains a challenge. In this study, we propose a visual language model-based RT target volume auto-delineation network termed Radformer. The Radformer utilizes a hierarchical vision transformer as the backbone and incorporates large language models to extract text-rich features from clinical data. We introduce a visual language attention module (VLAM) for integrating visual and linguistic features for language-aware visual encoding (LAVE). The Radformer has been evaluated on a dataset comprising 2985 patients with head-and-neck cancer who underwent RT. Metrics, including the Dice similarity coefficient (DSC), intersection over union (IOU), and 95th percentile Hausdorff distance (HD95), were used to evaluate the performance of the model quantitatively. Our results demonstrate that the Radformer has superior segmentation performance compared to other state-of-the-art models, validating its potential for adoption in RT practice.

Index Terms—medical image segmentation, large language model, vision language model, radiation therapy.

I. INTRODUCTION

RADIATION therapy (RT) is a widely used modality for the treatment of cancer [1]–[3]. In RT, ionizing radiation is used to kill the cancerous cells by inflicting damage to their DNA. The therapeutic efficacy of the RT is achieved by administering sufficient radiation doses to the tumor target while minimizing the exposure to normal tissues [4]. In RT, precision delineation of the treatment target plays an important role, and it directly impacts the treatment outcomes. Furthermore, advanced RT treatment plans, such as the volumetric-modulated arc therapy (VMAT), are more susceptible to contouring inaccuracies. However, manual contouring of target volume is a complex, laborious process, subject to intra- and inter-observer variations [5]. Moreover,

studies have demonstrated that a fair amount of the manually delineated target volumes are subjected to changes during the peer review process [6]–[8]. Over the last decade, deep learning (DL) has achieved significant progress in medical image segmentation tasks. Convolutional neural network (CNN) has demonstrated significant achievements in medical image segmentation [9]–[16]. Among CNNs, UNet stands out as one of the most extensively employed networks for segmentation tasks. For instance, a UNet-based hybrid densely connected network has been proposed for hepatic tumor segmentation from the combination of magnetic resonance (MR) and computed tomography (CT) images [17]. Similarly, a three-branch two-dimensional (2D) U-Net termed multiple branch UNet (MB-UNet) has been proposed for the concurrent segmentation of the prostate and lesions from the T2-weighted, diffusion weighted (DWI), and apparent diffusion coefficient (ADC) MR images [18]. To facilitate rapid target contouring, U-Net has also been extended to 3D volumetric data. For instance, 3D-UNet has been proposed for segmenting the prostate and the related organs for dose optimization in radiation therapy [19]. Moreover, a two-channel input 3D-UNet has also been proposed for the target volume segmentation in head and neck cancer [20]. In another implementation, CUNet, a modified 3D U-Net with residual block integrated with attention center block, has been proposed for prostate segmentation from CT images [21]. Similarly, DSD-UNet, a 3D-UNet-inspired architecture incorporating residual connection and dilated convolution, has been implemented for target volume segmentation in brachytherapy [22]. However, these methods suffer from limited delineation accuracy and are deemed unacceptable in routine clinical practice.

Over recent years, attention-based transformers have significantly advanced natural language processing (NLP) and computer vision domains [23]–[28]. Moreover, transformer-based backbones have achieved comparable or better performance than those of CNN-based backbones. Inspired by their success, attention-based transformers have also been adapted for various medical segmentation tasks. For instance, CoTr model interleaves transformers between the CNN decoder and encoder to enhance the tumor and organ segmentation performance [29]. In another development, TransUNet employs

stacked self-attention-based transformers as an encoder backbone for multi-organ segmentation from CT images [30]. Similarly, UNETR utilized a transformer-based encoder backbone for tumor and organ segmentation [31]. Recently, SWIN-UNETR, a more efficient U-shaped architecture incorporating hierarchical shifted windows-based transformers as the encoder, has been proposed for tumor and organ segmentation [32], [33].

The development of transformers also ushered the rise of vision language models (VLMs), where combined visual and text representation learning are utilized for various downstream tasks such as classification, object detection, and segmentation [34]–[36]. Notably, the segmentation of target objects described by the natural language expression has garnered much attention in computer vision, leading to the development of VLM architectures such as Vision-Language Transformer (VLT) [37], Referring image Segmentation using Transformer (ReSTR) [38], Language-Aware Vision Transformer (LAVT) [39], Clip-driven referring image segmentation (CRIS) [40], etc. However, the implementation of VLM for medical image segmentation is limited. In this study, we propose a VLM-based 3D medical image segmentation network called Radformer. The Radformer utilizes 3D medical images and text-rich clinical information to delineate the RT target volume. The Radformer employs a hierarchical shifted window (SWIN) transformer [33] as its backbone and integrates large language models (LLMs) comprising of Generative Pre-trained Transformer 4 (GPT-4) [41] and PubMed Bidirectional Encoder Representations from Transformers (PubMed-BERT) [42] to extract the text-rich clinical information. The language and visual information are integrated through language-aware visual encoding (LAVE) via the visual language attention module (VLAM) at each stage of the network. A CNN-based decoder is adopted to generate the 3D segmentation masks using the language-aware visual features. To evaluate the effectiveness of the proposed Radformer architecture, a public dataset comprising 2985 patients with head-and-neck cancer who underwent radiation therapy was used. Our Radformer achieves a Dice similarity coefficient (DSC) of 76%, an intersection over union (IOU) of 69%, and an 95th percentile Hausdorff distance (HD95) of 7.82 mm over a test set comprising 597 patients, respectively, significantly outperforms the state-of-the-art methods.

The summarization of this study is as follows:

1. Radformer, a VLM-based 3D segmentation approach utilizing the hierarchical vision transformer and LLMs has been developed for RT treatment target volume delineation.
2. The Radformer was evaluated on a large cohort of test data comprising 2985 patients to demonstrate its effectiveness and generalizability.
3. The Radformer outperformed the baseline 3D-UNETR by 15% with respect to mean DSC, and 17% with respect to mean IOU exhibiting high performance and excellent efficiency.

II. MATERIALS AND METHODS

A. Method

The overview of the Radformer is shown in Figure 1. The Radformer utilizes an encoder-decoder architecture, utilizing hierarchical vision transformers [33] in the encoder to generate cross-modal alignments between visual and textual information. A CNN-based decoder is utilized to generate 3D segmentation maps. In this section, we begin with the introduction of Language-Aware Visual Encoding (LAVE), followed by the Vision Language Attention Module (VLAM), the Language Gating Unit (LGU), and the CNN-based decoder.

1) Language-Aware Visual Encoding (LAVE)

The encoder of the Radformer is designed to extract and fuse the features from the 3D images and the text-rich clinical data. Typically, a patient’s clinical data comprises extensive unstructured information. To extract the features that are relevant to defining the treatment targets, we leverage the potential of the Generative Pre-trained Transformer 4 (GPT-4) [41]. We utilize tailored prompts to guide the GPT-4 to extract tumor-related information from the clinical data. A domain-specific Bidirectional Encoder Representations from Transformers (BERT) model pre-trained on the PubMed corpus called PubMed-BERT [42] was utilized to contextually embed the data extracted by GPT-4. The embedded high-dimensional word vector is represented as $L \in \mathbb{R}^{C_t \times T}$, where C_t represents the number of channels (768, corresponding to the size of the PubMed BERT hidden layer) and T represents the number of words.

Following the language feature extraction, we employ a SWIN-UNETR [33] based encoder to perform the combined visual feature fusion and visual language feature fusion. The encoder layer comprises four stages $i \in \{1, 2, 3, 4\}$. Each stage of the encoder has two transformer blocks, and it is designed to take two inputs, a 3D image of dimension $H \times W \times D$ ($V \in \mathbb{R}^{H \times W \times D}$) and vectorized tumor information of dimension $C_t \times T$ ($L \in \mathbb{R}^{C_t \times T}$). A patch partition layer is used to divide the 3D image volume into patches of dimension $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$ ($P \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}}$). The partitioned patches are then projected into an embedding space dimension C_i . In the encoder, the first three stages consist of two SWIN transformer blocks α_i , a multimodal feature fusion module β_i , and a gating unit ψ_i . In these stages, the generation of semantic-aware visual features involves a three-step process inspired by language-aware vision transformer (LAVT) model [39]. In each of the initial three stages, the SWIN transformer blocks α_i utilize the features generated by the preceding stage as input to generate visual features as output $I_i \in \mathbb{R}^{C_i \times H_i \times W_i \times D_i}$. The resultant visual features I_i are then integrated with the language features L via a multimodal feature fusion module termed VLAM to generate multimodal attention features $M_i \in \mathbb{R}^{C_i \times H_i \times W_i \times D_i}$. A learnable gating unit ψ_i called LGU weighs the attention features M_i ; the weighted features G_i are then combined elementwise to the

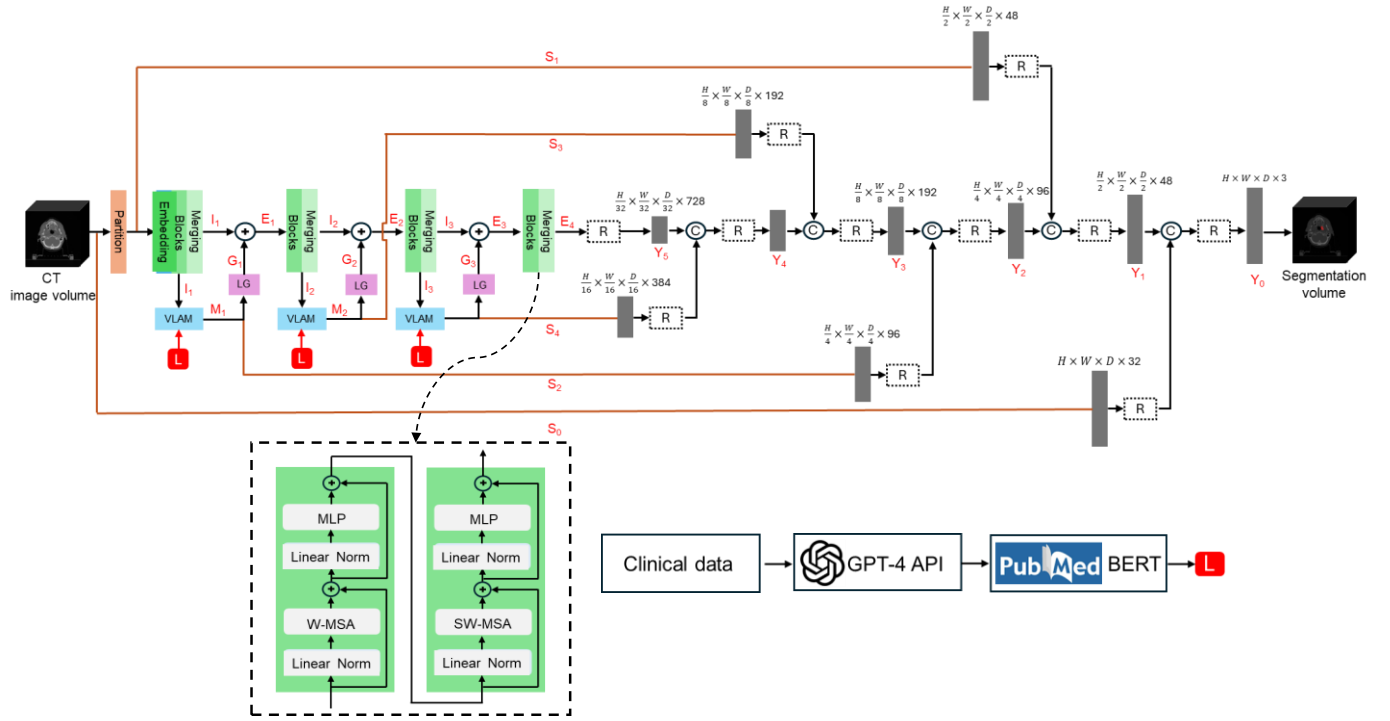


Fig. 1. Schematic of the proposed Radformer architecture. The Radformer utilizes 3D SWIN transformers as backbone to perform visual aware encoding. At each stage, the encoder generates visual features I_i $i \in \{1, 2, 3, 4\}$, which are used as the queries for the generation of position multimodal feature maps M_i $i \in \{1, 2, 3, 4\}$ through the visual language attention module (VLAM). These multimodal feature maps M_i are then adaptively fused with the visual features I_i . The fused features E_i are then used to generate the segmentation output through a CNN-based decoder employing skip connections.

visual feature I_i to produce language-enriched visual features E_i . The final stage of the encoder comprises only the SWIN transformer block without the VLAM and LGU.

2) Visual Language Attention Module (VLAM)

The schematic of the VLAM is shown in Figure 2. The VLAM utilizes the visual features as the query $I_i \in \mathbb{R}^{C_i \times H_i \times W_i \times D_i}$ and language features $L \in \mathbb{R}^{C_t \times T}$ as the key and value to generate position-specific sentence-level feature vectors $F_i \in \mathbb{R}^{C_i \times H_i \times W_i \times D_i}$ as follows.

$$I_{iq} = \text{flatten}(\rho_{iq}(I_i)) \quad (1)$$

$$L_{ik} = \rho_{ik}(L) \quad (2)$$

$$L_{iv} = \rho_{iv}(L) \quad (3)$$

$$F'_i = \text{softmax}\left(\frac{I_{iq}^T L_{ik}}{\sqrt{C_i}}\right) L_{iv}^T \quad (4)$$

$$F_{ij} = \rho_{iw}(\text{unflatten}(F_i'^T)) \quad (5)$$

Where, $\rho_{iq}, \rho_{iw}, \rho_{ik}, \rho_{iv}$, are projection functions. The projection functions ρ_{iq} and ρ_{iw} are implemented as $1 \times 1 \times 1$ convolutions followed by instance normalizations to generate projections resulting in channels C_i . The projection functions ρ_{ik} and ρ_{iv} are implemented as $1 \times 1 \times 1$ convolutions to

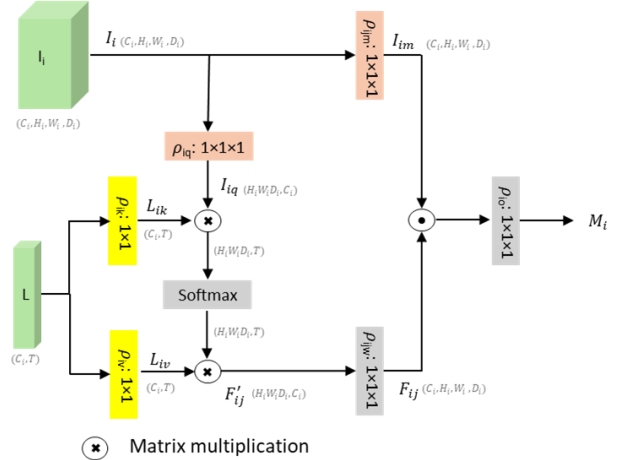


Fig. 2. Schematic of the Visual Language Attention Module (VLAM)

generate C_i number of output channels. The terms flatten and unflatten refer to the transformation of data from the original three spatial dimensions to a singular dimension and vice versa. The generated position-specific sentence-level feature vector F_i is then fused with the visual projections I_i to generate the multimodal feature maps M_i as follows:

$$I_{im} = \rho_{im}(I_i) \quad (6)$$

$$M_i = \rho_{io}(I_{im} \odot F_i) \quad (7)$$

Where, ρ_{im} and ρ_{io} are projection functions and \odot

represents element-wise multiplication. These projection functions are implemented as $1 \times 1 \times 1$ convolutions followed by rectified linear unit (ReLU) activation.

3) Language Gating Unit (LGU)

Figure 3 illustrates the schematic of the LGU. The LGU is designed as a language gate that adaptively regulates the amount of information flowing to the subsequent stage of the transformer. The LGU is implemented as follows:

$$H_i = \gamma_i(M_i) \quad (8)$$

$$G_i = H_i \odot F_i + I_i \quad (9)$$

Where γ_i refers to a two-layer perceptron; the first layer consists of $1 \times 1 \times 1$ convolution paired with RELU activation, and the second layer comprises $1 \times 1 \times 1$ convolution followed by hyperbolic tangent function. The \odot in Eq. (9) signifies the element-wise multiplication.

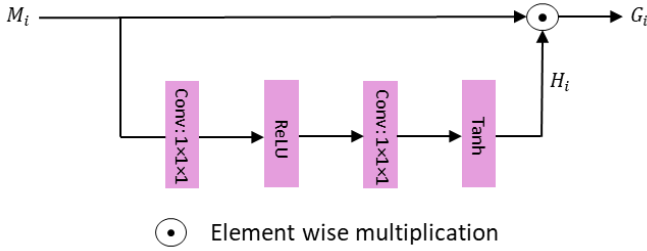


Fig. 3. Schematic of the Language Gating Unit (LGU)

4) Decoder

The multimodal feature maps M_i generated by the encoder are utilized to generate the 3D segmentation maps using a CNN-based decoder. The decoding process at each stage of the decoder is implemented as follows:

$$S_0 = \mathbf{R}(V), \quad V = V_1 + V_2 \quad (10)$$

$$S_1 = \mathbf{R}(P), \quad P = P_1 + P_2 \quad (11)$$

$$S_j = \mathbf{R}(M_j), \quad j \in \{2, 3, 4\} \quad (12)$$

$$Y_i = \mathbf{R}[\mathbf{v}(Y_{i+1}); S_i], \quad i \in \{0, 1, 2, 3, 4\} \quad (13)$$

$$Y_5 = S_5 \quad (14)$$

Where, \mathbf{R} represents a residual block implemented with two $1 \times 1 \times 1$ convolutions with instance normalization \mathbf{v} . The signifies the deconvolution based up sampling operation. The symbol $[\cdot]$ represents the concatenation among the channel dimensions. The final 3D segmentation outputs are generated from the feature maps Y_0 utilizing a $1 \times 1 \times 1$ convolution with a SoftMax activation.

B. Data

We evaluated the Radformer on a public head-and-neck cancer dataset (RADCURE) [43]. The dataset comprises CT volumes and gross tumor volume (GTV) contours of patients who underwent radiation therapy, along with clinical data such as demographic details, clinical histories, and treatment specifics. The average age of the patients in the dataset was 63 years and the dataset consists of individuals diagnosed with oropharyngeal, larynx, nasopharynx, and hypopharynx cancer. The clinical information was based on the 7th edition TNM staging system and was standardized to adhere to the American Association of Physicists in Medicine (AAPM) Task Group report no.263 (TG263) nomenclature. To optimize the Radformer, we preprocessed the CT images and the GTV contours. The patient cases with data inconsistencies and corrupt labels were removed during the preprocessing. The CT images were center cropped to a size of $336 \times 336 \times 64$ voxels to ensure all regions of interest were included. The contours were used to generate segmentation masks, including the background and targets. The dataset was randomly divided into two groups: one comprising 2,388 datasets for training and validating the model, and another group with 597 datasets designated for testing.

The prompt developed for our experiments to guide GPT-4 in summarizing clinical information, along with two example summaries used for the patient cases, is shown in Figure 4.

Prompt:	
"Please infer and summarize the following patient's cancer information: ///{file_contents}/// in approximately 50 words, emphasizing the patient's age, gender, smoking history, cancer type, specific location, and overall stage. Exclude technical details like TNM classification or specific categories. For example, the resultant response should look like this ///A 62.6-year-old female ex-smoker has advanced oropharyngeal cancer located on the posterior wall. The tumor is large and invasive, affecting nearby lymph nodes significantly, but without spreading to distant parts of the body, categorizing it as stage IVB///"	
Summarization:	
Example 1	The patient is an 89.4-year-old male ex-smoker who has been diagnosed with early-stage laryngeal cancer, specifically located in the glottis. There is no evidence of the cancer spreading to the lymph nodes or other parts of the body, categorizing it as stage 0
Example 2	The patient is a 76.6-year-old male who is a current smoker. He has been diagnosed with stage III nasopharyngeal cancer. Despite the cancer affecting nearby lymph nodes, it has not spread to distant parts of the body

Fig. 4. Prompt and examples of language Information used for patient cases.

C. Training

1) Loss Function

The loss function employed is a composite loss function comprising dice focal loss and Tversky loss from the MONAI library [44]. The composite loss function is formulated as:

$$L_{tot} = L_{DFC}(Y_G, Y_P) + L_{TV}(Y_G, Y_P) \quad (15)$$

Where, y_G denotes the true label and y_P denotes the predicted label. The dice focal loss, and Tversky loss are represented as L_{DFC} and L_{TV} .

2) Implementation

Besides our introduced multimodal model, we also implemented advanced models including SWIN-UNETR and 3D-UNETR for comparisons. Each model (Radformer, SWIN-UNETR, and 3D-UNETR) was implemented using the PyTorch deep learning library [45] on an Nvidia RTX 4090 GPU with a batch size of 1. Adam optimizer with a learning rate of 4×10^{-4} was utilized with a weight decay of 1×10^{-3} . All the models were trained for 100 epochs. We utilized HuggingFace’s Transformer library [46] to integrate the PubMed BERT into the Radformer. The PubMed BERT utilized is a base model comprising of 12 layers with a hidden size of 768. The SWIN transformer layers of the Radformer are initialized with pre-trained weights from the BraTS dataset [32]. The remaining weights are randomly initialized.

D. Evaluation Metrics

For evaluating the segmentation performance, we chose commonly used metrics including the dice similarity coefficient (DSC), intersection over union (IOU), and 95th percentile Hausdorff distance (HD95) [44]. These metrics were calculated on 3D volumes. Statistical analysis such as paired t-test was performed between our proposed method and other models across all metrics (DSC, IOU, and HD95).

III. RESULTS

The performance of the Radformer was assessed utilizing the public RADCURE dataset as described in the above session. We also compared the performance of the Radformer with other state-of-the-art DL-based segmentation approaches including SWIN-UNETR and 3D-UNETR. The SWIN-UNETR and 3D-UNETR networks are DL approaches widely utilized for 3D segmentation tasks. As summarized in Table 1, the Radformer significantly outperforms other networks over all the metrics ($p < 0.05$). In particular, Radformer achieved a mean DSC score of 0.76, a mean IOU of 0.69, and a mean HD95 of 7.82 mm. The Radformer outperformed the baseline 3D-UNETR by 15% in terms of mean DSC and 17% in terms of mean IOU. Furthermore, a 45% improvement in the boundary accuracy (HD95) was noted compared to that of the 3D-UNETR.

TABLE I
COMPARISON OF THE PERFORMANCE OF RADFORMER FOR GTV SEGMENTATION WITH OTHER METHODS INCLUDING RADFORMER WITHOUT LAVE, SWIN-UNETR, AND 3D-UNETR

Network	DSC	IOU	HD95 (mm)
Radformer (1)	0.76±0.09	0.69±0.08	7.82±6.87
SWIN-UNETR (2)	0.69±0.11	0.64±0.09	12.88±6.60
UNETR (3)	0.66±0.09	0.59±0.07	14.28±6.85
p-value (1 vs 2)	<0.05	<0.05	<0.05
p-value (1 vs 3)	<0.05	<0.05	<0.05

The results of the ablation study containing the Radformer implementation without LAVE are summarized in Table 2. Compared to the Radformer without LAVE (VLAM and LGU module), the Radformer with LAVE improves the segmentation performance by 8% in terms of mean DSC and 9% with respect to the mean IOU. Moreover, incorporating the language modules (VLAM and LGU) in the Radformer improved the segmentation accuracy (HD95) by 36%, with the improvements being statistically significant ($p < 0.05$).

TABLE II
COMPARISON OF THE PERFORMANCE OF RADFORMER FOR GTV SEGMENTATION WITH RADFORMER WITHOUT LAVE

Network	DSC	IOU	HD95 (mm)
Radformer (1)	0.76±0.09	0.69±0.08	7.82±6.87
Radformer without LAVE (2)	0.70±0.10	0.63±0.08	12.27±7.68
p-value (1 vs 2)	<0.05	<0.05	<0.05

The qualitative results of the Radformer's GTV segmentation on a patient with early-stage laryngeal cancer and a patient with stage III nasopharyngeal cancer are illustrated in Figures 5 and 6, respectively. These figures are arranged into columns depicting the CT images, manual contours, and the GTV predictions by the proposed Radformer, Radformer without LAVE, SWIN-UNETR, and 3D-UNETR. Figures 5 and 6 show that the Radformer precisely segmented the GTVs in both cases when the language and image information were integrated. Whereas the Radformer without LAVE, the SWIN-UNETR, and 3D-UNETR overestimated the GTVs in most slices.

IV. DISCUSSION

Delineation of the treatment target is a crucial step in RT treatment planning, and its accuracy significantly affects the quality of the RT plan. Currently, the delineation of target volume purely relies on human expert’s manual process, which is time-consuming, laborious, and subject to intra- and inter-observer variability depending on the individual's knowledge and experience. Although DL-based segmentation approaches have made considerable advancements in recent years, its success in RT target delineation compared to normal organ segmentation is limited. At present, the delineation of RT treatment targets is still clinically unacceptable in most situations. The inferior performance of DL-based approaches for RT target delineation compared to normal organ segmentation could be attributed to the complexity of target delineation in RT. In fact, RT target delineation is a comprehensive medical decision-making process. In routine clinical practice, radiation oncologists consider the entire medical history and diagnosis of the disease when performing target delineation. In this study, we tested our hypothesis that, in addition to images, radiation oncologists leverage clinical data for RT target delineation. Therefore, integrating both text-

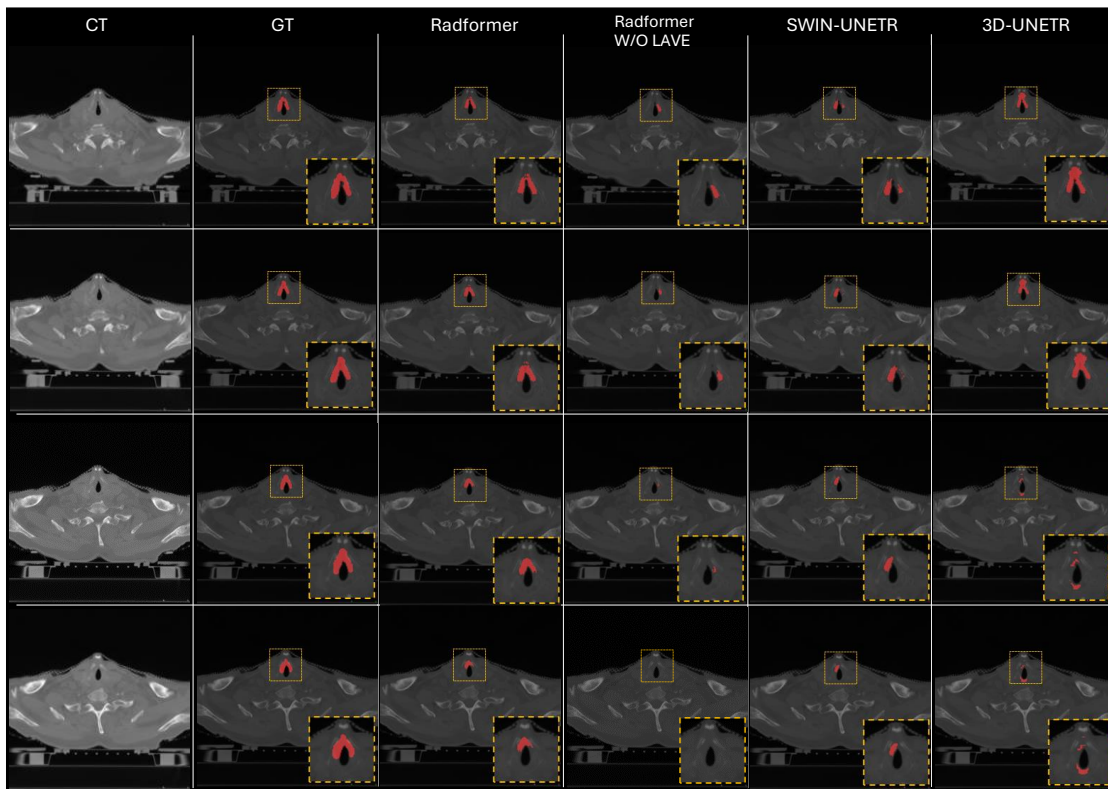


Fig. 5. Illustrative example of the GTV segmentation for a patient with early-stage laryngeal cancer

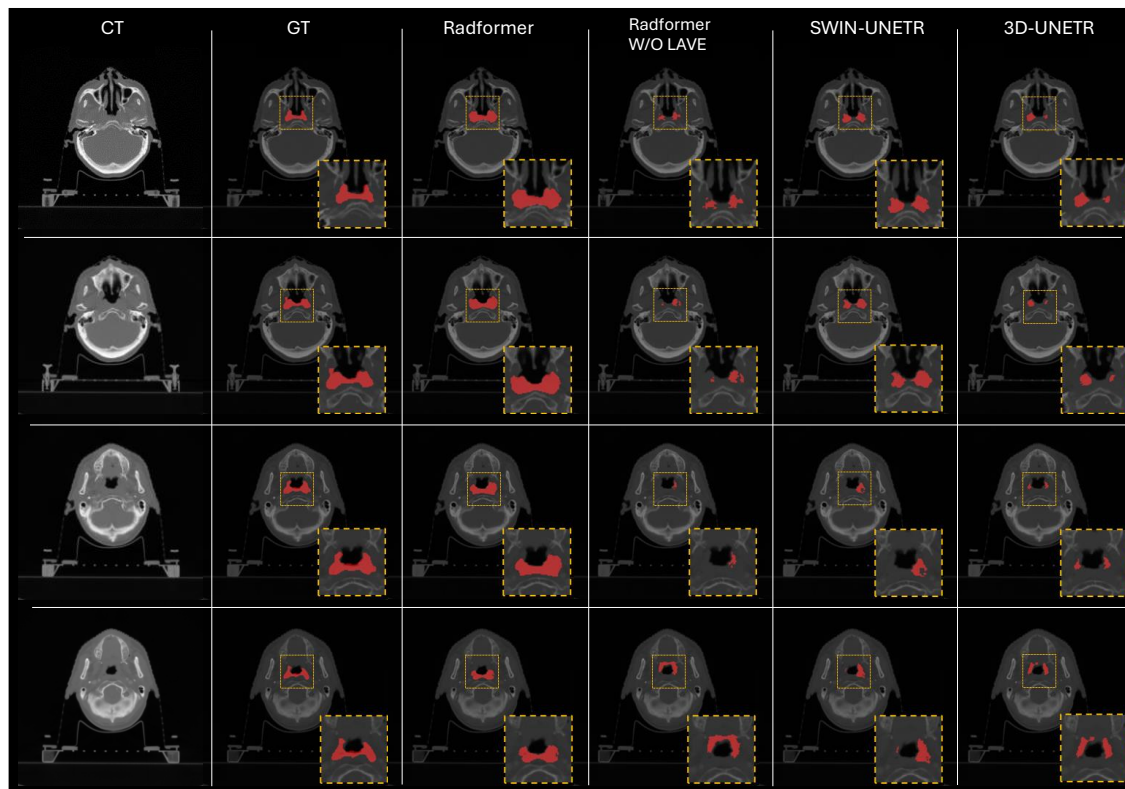


Fig. 6. Illustrative example of the GTV segmentation for a patient stage III nasopharyngeal cancer

rich clinical information and imaging data will improve DL-based auto-segmentation methods for RT target delineation.

We developed a VLM-based 3D medical image segmentation network called Radformer. The Radformer utilizes the SWIN transformer as the backbone to extract visual features from images and LLMs to extract text-rich features from clinical data. The visual and linguistic features are fused through language-aware visual encoding and then leveraged by a CNN-based decoder to output segments. The Radformer was trained on a dataset from 2388 patients and evaluated on a test dataset comprising of 597 patients. Our results demonstrate that the Radformer achieves superior performance in comparison with the other state-of-the-art architectures purely leveraging visual features (SWIN-UNETR, and 3D-UNETR). Moreover, the statistical analyses using paired t-tests over the metrics DSC, IOU, and HD95 further substantiate the significant improvement achieved by Radformer. Furthermore, the enhanced accuracy in segmenting the GTV between the Radformer and Radformer without LAVE (Table 2) underscores the importance of linguistic and visual feature integration for GTV segmentation.

Figure 7 depicts the violin plots of segmentation performance distribution on all patient test cases, evaluated over the metrics including DSC, IOU, and HD95. These plots reveal the spread and density of the segmentation performance scores across ~600 patient cases. Figures 7(a) and 7(b) highlight that the Radformer architecture generally surpasses the performance of other models in most patient cases for both DSC and IOU metrics. This suggests a consistent, robust, and generalized segmentation capability. Furthermore, over other architectures such as Radformer without LAVE, SWIN-UNETR, and 3D-UNETR, it can be noted that the performance scores are tightly clustered within the interquartile range and are characterized by lower median values, indicating a relative underperformance in certain patient scenarios. This variability in performance substantiates the advantages of integrating linguistic and visual features. Furthermore, the violin plots representing performance over the HD95 metric indicate that the majority of the Radformer's performance is skewed towards a lower median value, suggesting enhanced accuracy in contour segmentation compared to other methods. This skewing towards lower values indicates the tendency of the Radformer to yield more precise segmentations with less deviation from the ground truth, as lower HD95 values correspond to smaller distances between predicted and actual contours. Moreover, the improvement in segmentation performance when linguistic information can be noted by comparing the performance distribution between the Radformer and Radformer without LAVE over all the metrics.

At present, the clinical data associated with the medical images are unstructured and extensive. In this study, we leverage the potential of LLMs to extract valuable information from the clinical data through customized prompts. In particular, we used GPT-4 to extract tumor-related information from clinical data. As GPT-4 only generates output texts for public access, we utilized PubMed BERT (an open-source model) to contextually embed the lesion information provided by GPT-4. Ideally, the process of extracting tumor-related

information and contextual embedding can be masked by a single LLM. The proposed approach is generic and has the great

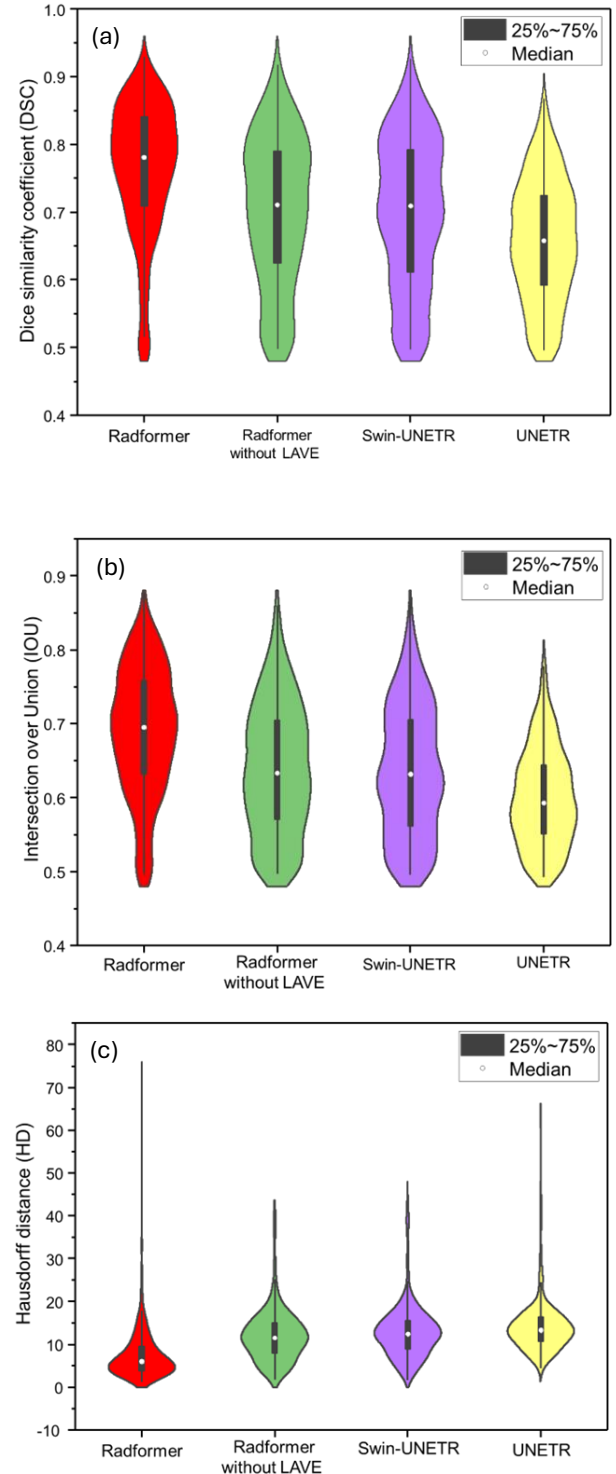


Fig. 7. Violin plots demonstrating the performance of the Radformer, Radformer without LAVE, SWIN-UNETR and 3D-UNETR on all patient test cases.

potential to be adaptable to the segmentation of various cancer types in radiation oncology.

Moving forward, integrating our proposed method into standard radiotherapy treatment planning could transform

clinical practices. The Radformer’s ability to delineate target volumes quickly and precisely could drastically decrease the manual effort and time typically required for this task. This increase in efficiency is especially important in the context of advanced RT treatment plans, where even small errors in contouring can lead to significant consequences.

V. CONCLUSION

In conclusion, in this study, we propose a large language model-augmented approach for RT treatment target auto-delineation. Breaking away from the conventional architectures relying only on visual features, our model leverages text-rich clinical information and visual features to enhance the accuracy of target delineation. The proposed method could be adopted into routine radiotherapy treatment planning, offering a means to rapidly contour the target volumes with high precision.

ACKNOWLEDGMENT

This research is supported in part by the Department of Defense Prostate Cancer Research Program Award W81XWH-19-1-0567, Stanford Radiation Oncology Prostate Cancer Research grant, Stanford Quality Incentive Metric Monies program, the Human-Centered Intelligence of Stanford University, and the National Institutes of Health (1R01CA223667, 1R01CA176553, and 1R01CA227713).

REFERENCES

- [1] L. Xing, B. Thorndyke, E. Schreibmann, Y. Yang, T.-F. Li, G.-Y. Kim, G. Luxton, and A. Koong, “Overview of image-guided radiation therapy,” *Medical Dosimetry*, vol. 31, no. 2, pp. 91–112, Jun. 2006.
- [2] K. Wang and J. E. Tepper, “Radiation therapy-associated toxicity: Etiology, management, and prevention,” *CA Cancer J Clin*, vol. 71, no. 5, pp. 437–454, Sep. 2021.
- [3] X. Kui, F. Liu, M. Yang, H. Wang, C. Liu, D. Huang, Q. Li, L. Chen, and B. Zou, “A review of dose prediction methods for tumor radiation therapy,” *Meta-Radiology*, vol. 2, no. 1, p. 100057, 2024.
- [4] X. Yu, F. Jin, H. Luo, Q. Lei, and Y. Wu, “Gross Tumor Volume Segmentation for Stage III NSCLC Radiotherapy Using 3D ResSE-UNet,” *Technol Cancer Res Treat*, vol. 21, p. 15330338221090848, 2022.
- [5] H. Bollen, S. Willems, M. Wegge, F. Maes, and S. Nuyts, “Benefits of automated gross tumor volume segmentation in head and neck cancer using multi-modality information,” *Radiotherapy and Oncology*, vol. 182, May 2023.
- [6] A. A. Albert, W. N. Duggar, R. P. Bhandari, T. Vengaloor Thomas, S. Packianathan, R. M. Allbright, M. R. Kanakamedala, D. Mehta, C. C. Yang, and S. Vijayakumar, “Analysis of a real time group consensus peer review process in radiation oncology: an evaluation of effectiveness and feasibility,” *Radiation Oncology*, vol. 13, no. 1, p. 239, 2018.
- [7] K. Brunskill, T. K. Nguyen, R. G. Boldt, A. V Louie, A. Warner, L. B. Marks, and D. A. Palma, “Does Peer Review of Radiation Plans Affect Clinical Care? A Systematic Review of the Literature,” *International Journal of Radiation Oncology*Biophysics*, vol. 97, no. 1, pp. 27–34, 2017.
- [8] E. Martín-García, F. Celada-Álvarez, M. J. Pérez-Calatayud, M. Rodríguez-Pla, O. Prato-Carreño, D. Farga-Albiol, O. Pons-Llanas, S. Roldán-Ortega, E. Collado-Ballesteros, F. J. Martínez-Arcelus, Y. Bernisz-Díaz, V. A. Macías, J. Chimeno, J. Gimeno-Olmos, F. Lliso, V. Carmona, J. C. Ruiz, J. Pérez-Calatayud, A. Tormo-Micó, and A. J. Conde-Moreno, “100% peer review in radiation oncology: is it feasible?,” *Clinical and Translational Oncology*, vol. 22, no. 12, pp. 2341–2349, 2020.
- [9] B. Ibragimov and L. Xing, “Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks,” *Med Phys*, vol. 44, no. 2, pp. 547–557, Feb. 2017.
- [10] X. Dai, Y. Lei, T. Wang, J. Zhou, S. Rudra, M. McDonald, W. J. Curran, T. Liu, and X. Yang, “Multi-organ auto-delineation in head-and-neck MRI for radiation therapy using regional convolutional neural network,” *Phys Med Biol*, vol. 67, no. 2, p. 025006, 2022.
- [11] X. Dai, Y. Lei, J. Wynne, J. Janopaul-Naylor, T. Wang, J. Roper, W. J. Curran, T. Liu, P. Patel, and X. Yang, “Synthetic CT-aided multiorgan segmentation for CBCT-guided adaptive pancreatic radiotherapy,” *Med Phys*, vol. 48, no. 11, pp. 7063–7073, Nov. 2021.
- [12] X. Li, L. C. Jia, F. Y. Lin, T. Liu, S. M. He, W. Zhang, M. Zhang, and Y. Wang, “Small Samples and Low-Cost Auto-Segmentation Method for Pelvic Organ-at-Risk Segmentation in Magnetic Resonance Images Using Deep-Learning,” *Int J Radiat Oncol Biol Phys*, vol. 117, no. 2, pp. e685–e686, Oct. 2023.
- [13] S. Wang, C. Li, R. Wang, Z. Liu, M. Wang, H. Tan, Y. Wu, X. Liu, H. Sun, R. Yang, X. Liu, J. Chen, H. Zhou, I. Ben Ayed, and H. Zheng, “Annotation-efficient deep learning for automatic medical image segmentation,” *Nat Commun*, vol. 12, no. 1, p. 5915, 2021.
- [14] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, “Segment anything in medical images,” *Nat Commun*, vol. 15, no. 1, p. 654, 2024.
- [15] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, “Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation,” *IEEE Trans Neural Netw Learn Syst*, vol. 32, no. 2, pp. 523–534, 2021.
- [16] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, “H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes,” *IEEE Trans Med Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.
- [17] H. R. Roth, H. Oda, X. Zhou, N. Shimizu, Y. Yang, Y. Hayashi, M. Oda, M. Fujiwara, K. Misawa, and K. Mori, “An application of cascaded 3D fully convolutional networks for medical image segmentation,” *Computerized Medical Imaging and Graphics*, vol. 66, pp. 90–99, 2018.
- [18] Y. Chen, L. Xing, L. Yu, H. P. Bagshaw, M. K. Buyyounouski, and B. Han, “Automatic intraprostatic lesion segmentation in multiparametric magnetic resonance images with proposed multiple branch UNet,” *Med Phys*, vol. 47, no. 12, pp. 6421–6429, Dec. 2020.
- [19] M. Kawula, D. Purice, M. Li, G. Vivar, S.-A. Ahmadi, K. Parodi, C. Belka, G. Landry, and C. Kurz, “Dosimetric impact of deep learning-based CT auto-segmentation on radiation therapy treatment planning for prostate cancer,” *Radiation Oncology*, vol. 17, no. 1, p. 21, 2022.
- [20] S. Kihara, Y. Koike, H. Takegawa, Y. Anetai, S. Nakamura, N. Tanigawa, and M. Koizumi, “Clinical target volume segmentation based on gross tumor volume using deep learning for head and neck cancer treatment,” *Medical Dosimetry*, vol. 48, no. 1, pp. 20–24, 2023.
- [21] J. Shen, Y. Tao, H. Guan, H. Zhen, L. He, T. Dong, S. Wang, Y. Chen, Q. Chen, Z. Liu, and F. Zhang, “Clinical Validation and Treatment Plan Evaluation Based on Autodelineation of the Clinical Target Volume for Prostate Cancer Radiotherapy,” *Technol Cancer Res Treat*, vol. 22, p. 15330338231164884, 2023.
- [22] D. Zhang, Z. Yang, S. Jiang, Z. Zhou, M. Meng, and W. Wang, “Automatic segmentation and applicator reconstruction for CT-based brachytherapy of cervical cancer using 3D convolutional neural networks,” *J Appl Clin Med Phys*, vol. 21, no. 10, pp. 158–169, 2020.
- [23] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in Vision: A Survey,” *ACM Comput. Surv.*, vol. 54, no. 10s, Sep. 2022.
- [24] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, “A Survey on Vision Transformer,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 01, pp. 87–110, 2023.
- [25] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, “Overview of the Transformer-based Models for NLP Tasks,” in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 2020, pp. 179–183.
- [26] T. Lin, Y. Wang, X. Liu, and X. Qiu, “A survey of transformers,” *AI Open*, vol. 3, pp. 111–132, 2022.
- [27] M. Ramprasath, K. Dhanasekaran, T. Karthick, R. Velumani, and P. Sudhakaran, “An Extensive Study on Pretrained Models for Natural Language Processing Based on Transformers,” in *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, 2022, pp. 382–389.
- [28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image is Worth 16x16 Words:

- Transformers for Image Recognition at Scale,” *ArXiv*, vol. abs/2010.11929, 2020.
- [29] J. and S. C. and X. Y. Xie Yutong and Zhang, “CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 2021, pp. 171–180.
- [30] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, L. Xing, L. Lu, A. Yuille, and Y. Zhou, “3D TransUNet: Advancing Medical Image Segmentation through Vision Transformers,” *ArXiv*, vol. abs/2310.07781, 2023.
- [31] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, “UNETR: Transformers for 3D Medical Image Segmentation,” in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1748–1758.
- [32] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and 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*, 2022, pp. 272–284.
- [33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10002.
- [34] J. Zhang, J. Huang, S. Jin, and S. Lu, “Vision-Language Models for Vision Tasks: A Survey,” *IEEE Trans Pattern Anal Mach Intell*, pp. 1–20, 2024.
- [35] J. Wu, X. Li, X. Li, H. Ding, Y. Tong, and D. Tao, “Toward Robust Referring Image Segmentation,” *IEEE Transactions on Image Processing*, vol. 33, pp. 1782–1794, 2024.
- [36] X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang, “Knowledge-enhanced visual-language pre-training on chest radiology images,” *Nat Commun*, vol. 14, no. 1, p. 4542, 2023.
- [37] H. Ding, C. Liu, S. Wang, and X. Jiang, “VLT: Vision-Language Transformer and Query Generation for Referring Segmentation,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 06, pp. 7900–7916, 2023.
- [38] N. Kim, D. Kim, S. Kwak, C. Lan, and W. Zeng, “ReSTR: Convolution-free Referring Image Segmentation Using Transformers,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18124–18133.
- [39] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. S. Torr, “LAVT: Language-Aware Vision Transformer for Referring Image Segmentation,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18134–18144.
- [40] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu, “CRIS: CLIP-Driven Referring Image Segmentation,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11676–11685.
- [41] OpenAI, “GPT-4 Technical Report.”
- [42] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, “Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing,” *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, Oct. 2021.
- [43] M. L. Welch, S. Kim, A. Hope, S. H. Huang, Z. Lu, J. Marsilla, M. Kazmierski, K. Rey-McIntyre, T. Patel, B. O’Sullivan, J. Waldron, J. Kwan, J. Su, L. Soltan Ghoraie, H. B. Chan, K. Yip, M. Giuliani, Princess Margaret Head And Neck Site Group, S. Bratman, B. Haibe-Kains, and T. Tadic, “Computed Tomography Images from Large Head and Neck Cohort (RADCURE),” *The Cancer Imaging Archive*, 2023.
- [44] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, V. Nath, Y. He, Z. Xu, A. Hatamizadeh, W. Zhu, Y. Liu, M. Zheng, Y. Tang, I. Yang, and A. Feng, *MONAI: An open-source framework for deep learning in healthcare*. 2022.
- [45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: an imperative style, high-performance deep learning library,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019.
- [46] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “HuggingFace’s Transformers: State-of-the-art Natural Language Processing,” *ArXiv*, vol. abs/1910.03771, 2019.