SwallowReg: SAM-Assisted Deformable Registration with Adaptive Global-Local Features for Cine-MRI Swallowing Function Quantification

Zhiwen Tang[1], Minghao Sun[2], Chenghui Jiang[3], Xiaomeng Song[3], Lulu Liu[4], and Han Yu[1]

[1] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China
[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[3] Department of Oral and Maxillofacial Surgery, Affiliated Stomatological Hospital of Nanjing Medical University, Nanjing, China
[4] School of Mathematics and Statistics, Nanjing University of Science and Technology, Nanjing, China
han.yu@njupt.edu.cn

Abstract. Although medical image registration is an effective tool for quantifying organ motion, its application in assessing postoperative swallowing function in tongue cancer remains limited. Recognizing that existing methods often fail to fully exploit the semantic priors from segmentation tasks, we propose an end-to-end joint learning framework for segmentation and registration on dynamic cine-MRI sequences. By leveraging semantic priors from the Segment Anything Model to guide the registration, it becomes possible to precisely model the complex post-operative motion of the tongue. Our contributions can be generalized into three aspects:(1) We construct a deeply supervised SAM feature adaptor, where the U-Net encoder generates a multi-scale semantic feature pyramid to provide hierarchical anatomical priors for the registration task. (2) We design a novel convolutional unit combining DCNv3 and Swin-transformer to capture local-to-global features; a bidirectional coordinate attention fusion module is further introduced to adaptively fuse multi-scale features from both the SAM feature adaptor and registration encoder. (3) To directly enhance registration accuracy along anatomical boundaries, we propose an EdgeWeightedIoU loss that employs a Laplacian operator to extract and refine these boundaries from the segmentation output. Ablation studies show that our SAM scheme achieves a 7.55% Dice improve-ment on the sagittal tongue MRI dataset. In SOTA experiments on the coronal tongue MRI dataset, VoxelMorph, TransMorph, and our Baseline gain 3.93%, 35.22%, and 3.32% in Dice respectively after integrating our SAM scheme. Con-sistent improvements are also observed on the ACDC cardiac MRI da-taset . Our code is available at: https://github.com/FFigo/SwallowReg.

Keywords: Registration, Tongue cancer, Cine-MRI, SAM, EdgeWeightedIoU

1 Introduction

Medical image registration is essential to head and neck cancer patients in clinical practice, enabling precise anatomical alignment across scans [1] and supporting critical applications like postoperative evaluation, multimodal fusion, and disease tracking [2,3,4]. Rising tongue cancer incidence [5] often leads to swallowing disorders, with a 40%–60% postoperative rate [6], severely compromising patients' quality of life [7]. Therefore, this study analyses swallowing function of such patients with registration.

Quantification of tongue movement characteristics is critical for differentiating swallowing function, assessing treatment efficacy, and monitoring rehabilitation outcomes [8,9]. Videofluoroscopic swallowing study (VFSS) is hampered by radiation exposure and inadequate coronal visualization [10], precluding repeated monitoring. In contrast, Cine-MRI emerges as an optimal alternative, featuring non-invasiveness, radiation-free imaging, and high spatiotemporal resolution [11,12]. Current tongue movement quantification often relies on manual annotation of tongue structures and subsequent estimation of movement parameters by clinicians, which is time-consuming, inefficient, and associated with inter-observer variability [13]. Semi-automatic random-walker methods [14] require manual seeds, causing observer variability; fully automatic contour registration [15] fails under large deformations. Neither captures anatomical semantics, limiting generalization in postoperative tongues. Existing methods also exhibit three inherent limitations:

- The intricate head and neck anatomy is prone to non-target tissue interference [16]. Existing models sometimes lack tongue-specific perception, and especially for global consistency, they fail to maintain fine-grained alignment of the tongue region of interest, potentially degrading the accuracy of movement parameters [17].
- Mainstream registration methods, which rely on pixel-level intensity similarity, neglect anatomical semantics, leading to misregistration and suboptimal boundary alignment in regions where the tongue and adjacent tissues share structural similar intensity profiles [18,19,20].
- Tongue movement involves both local deformation and global coordination. However, pure CNNs struggle with long-range dependencies, while pure Trans-formers tend to lose local details [21]. Current hybrid models, with their static fusion strategies, fail to adaptively balance these features, hindering precise movement characterization [22].

To address the aforementioned challenges of manual annotation and registration limitations, this paper proposes an end-to-end joint segmentation-and-registration framework integrated with SAM semantic priors. By estimating the tongue deformation field, this framework offers a novel solution for tongue movement registration and quantification, facilitating swallowing function evaluation in head and neck cancer patients. Section 2 elaborates on the model architecture and core methodologies. Section 3 presents experimental findings and Section 4 concludes the paper.

2 Methodology

2.1 Network Architecture

The proposed end-to-end joint learning framework (Fig. 1) comprises three collaborative modules. (1) A deep supervision-based SAM [23] Feature Adaptor generates multi-scale anatomical features, providing tongue-specific guidance for registration. (2) The Registration Network uses Swin-Intern blocks (integrating DCNv3 [24] and Swin Transformer [25]) for global-local balance, and a Bidirectional Coordinate Attention Fusion (BCAF) module [26] to adaptively fuse semantic and geometric features. (3) A Joint Training Strategy employs a multi-task coupled loss, enabling segmentation-derived semantic information to guide the learning of the registration deformation field.

2.2 SAM Feature Adaptor

To enhance tongue region registration accuracy, we propose SAM Feature Adaptor that adapts high-dimensional anatomical semantic features from SAM to registration tasks, using segmentation-derived anatomical cues for precise guidance. This module integrates SAM's segmentation advantages with medical image registration requirements via lightweight structural adjustments and a selective freezing training strategy.

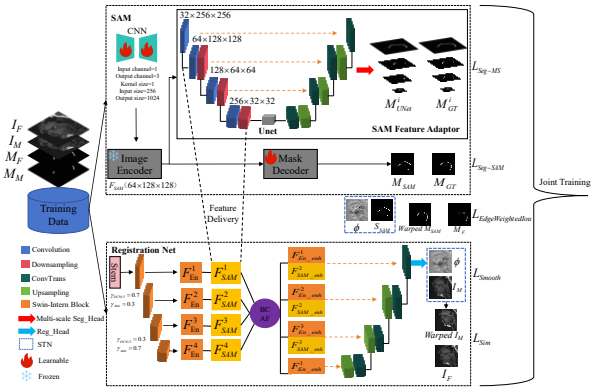

Fig. 1. Architecture of the proposed end-to-end joint segmentation and registration framework.

To mitigate the domain gap between medical images and the pre-training data of SAM, a learnable CNN is incorporated into the framework. It adjusts the channel and resolution of input images to ensure compatibility with the SAM image encoder. Specifically, the SAM image encoder parameters are fixed while the CNN and SAM mask decoder parameters are updated during training. This enables gradient sharing with the SAM Feature Adaptor, stabilizing its feature extraction capability for the subsequent registration task. This design helps the model adapt to medical anatomical feature learning. Processed images are fed into the frozen SAM image encoder, generating the feature $F_{\mathrm{SAM}} \in R^{64\times128\times128}$. We then employ a learnable U-Net for hierarchical feature

reconstruction on F_{SAM}. It converts F_{SAM} into a multi-scale feature sequence $\{F_{\text{SAM}}^{1-4}\}$, which aligns well with the hierarchical structure of the subsequent registration network. This design preserves fine-grained anatomical semantics of the tongue while achieving scale matching.

To ensure the learning quality of anatomical features at all levels, a deep supervision mechanism is embedded in the SAM Feature Adaptor. The U-Net outputs corresponding tongue segmentation masks M_{Unet}^{1-4} at each scale. The multi-scale segmentation loss is computed with the ground-truth segmentation mask M_{GT}^{i} at the corresponding scale:

$$L_{Seg\text{-}MS} = \sum_i^4 w_i \cdot Dice - BCE\left(M_{Unet}^{i}, M_{GT}^{i}\right), \tag{1}$$

where w_i denotes the weight for each scale loss, and $Dice - BCE(\cdot)$ denotes the combination of Dice loss and binary cross-entropy loss. This loss constrains each hierarchical feature to learn accurate anatomical structures. Meanwhile, the feature F_{SAM} is fed into the learnable mask decoder to produce the segmentation mask M_{SAM}, and $L_{Seg-SAM}$ is computed via Dice-BCE loss between M_{SAM} and M_{GT} to maintain the segmentation performance of the adapted model. The total loss of this module is obtained by weighted combination of the above two losses using hyperparameters λ_1 and λ_2 as follows

$$L_{Seg} = \lambda_1 \cdot L_{\text{Seg-}MS} + \lambda_2 \cdot L_{\text{Seg-}SAM}. \tag{2}$$

2.3 Registration Network

To jointly model fine-grained tongue deformations and global head-neck semantics, we propose the Swin-Intern Block (Fig. 2(a)) as the core innovation of our registration encoder. It employs a parallel architecture integrating Deformable Convolution DCNv3 and Swin Transformer window attention, adaptively fusing features via learnable gating weights to balance local deformation and global semantic modeling.

Given input feature $X \in R^{B\times 2C\times H\times W}$ (concatenated from fixed I_F and moving I_M images along channels), the module extracts features via two parallel branches: deformable convolution (F_{DCNv3}) and window attention (F_{Attn}). For adaptive fusion, a lightweight gating network with learnable weight matrices W_1 and W_2 generates channel-wise gating weights $[g_{DCNv3}; g_{Attn}]$ as follows

$$[g_{\text{DCNv3}}; g_{\text{Attn}}] = \mathcal{S}\left(W_2 \cdot \mathcal{R}(W_1 \cdot [F_{\text{DCNv3}}; F_{\text{Attn}}])\right), \tag{3}$$

where $\mathcal{S}$ and $\mathcal{R}$ denote the Softmax and the ReLU, respectively. These gating weights are further modulated by predefined, hierarchical specific weights γ_{DCNv3} and γ_{Attn}. Let $\hat{F}_k = \gamma_k \cdot g_k \cdot F_k$ ($k \in$ {DCNv3, Attn}), the final output can be calculated by

$$\text{Output} = \text{MLP}\left(\text{LayerNorm}\left(\hat{F}_{\text{DCNv3}} + \hat{F}_{\text{Attn}}\right)\right). \tag{4}$$

To prioritize capturing fine-grained deformations within the tongue region, we set γ_{DCNv3}=0.7 and γ_{Attn}=0.3 in the shallow layers of the registration encoder. In the

deeper encoder layers, these values are reversed to enhance the global anatomical semantic comprehension around the mouth. This hierarchical design enables the encoder to produce feature sequences encapsulating both local geometric deformation of the tongue and global semantic dependencies across the head and neck region.

To effectively fuse the feature sequences from SAM and the registration encoder, we design the BCAF module. Leveraging the advantages of coordinate attention in integrating context and positional awareness. Its unidirectional fusion process is shown in Fig. 2(b). The module performs collaborative computation across channel and spatial dimensions to accurately capture spatial-directional features of the tongue contour.

Given the primary input feature F, height and width descriptors Z_h and Z_w are first obtained via bidirectional global pooling. Such feature can be iteratively updated by

$$F' = F \oplus (\mathcal{A}_{\mathrm{h}} \otimes \mathcal{A}_{\mathrm{w}}), \text{ and } \mathcal{A}_{h/w} = \sigma\left(W_{h/w} \cdot \mathcal{R}(Conv_{1\times1}([Z_h, Z_w]))\right), \quad (5)$$

where W_w and W_h are the weight matrices for width and height attention; σ is the sigmoid function; $\otimes$ denotes element-wise multiplication, and $\oplus$ is for feature addition.

The feature sequences of the registration encoder and SAM are mutually modulated by the BCAF module and decoded by the registration network to produce the displacement field ϕ characterizing tongue motion. The registration loss L_{Reg} comprises two components: a similarity loss L_{Sim} between I_F and the warped moving image $I_M \circ \phi$, and a regularization loss L_{Smooth} that enforces the smoothness of ϕ by

$$L_{Reg} = L_{Sim}(I_M \circ \phi, I_F) + \lambda L_{Smooth}(\phi). \quad (6)$$

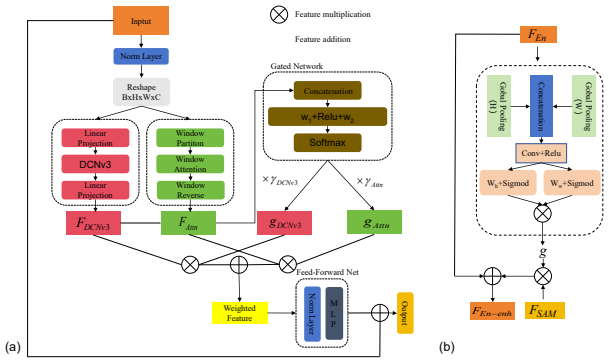

Fig. 2. Structural illustrations of core modules. (a) Architecture of the Swin-Intern Block; (b) Unidirectional fusion process of the Bidirectional Coordinate Attention module with F_{En} as the main feature and F_{SAM} as the auxiliary feature.

2.4 Joint Training

End-to-end joint segmentation and registration are realized via multi-loss supervision within a unified optimization framework, where bidirectional gradient propagation enables information interaction between branches. This allows the segmentation network

to learn more targeted anatomical representations, providing high-quality semantic priors that enhance registration accuracy and boundary consistency.

We propose an EdgeWeightedIoU loss to achieve coupled supervision for segmentation and registration. This loss is calculated between the fixed image ground truth M_F and the warped SAM segmentation mask $M_{SAM} \circ \phi$, using the Laplacian operator ∇^2 to extract edge responses and construct spatially adaptive weights. Such a design encourages the model to pay more attention to the alignment accuracy of anatomical boundaries. The loss function is defined as follows:

$$L_{EdgeWeightedIoU} = 1 - \frac{\sum_{i,j}(M_{\mathrm{SAM}}\circ\phi)^{(i,j)} \cdot M_F^{(i,j)} \cdot \left(1+\omega \cdot \left|\nabla^2 M_F^{(i,j)}\right|\right)}{\sum_{i,j}\left((M_{\mathrm{SAM}}\circ\phi)^{(i,j)} + M_F^{(i,j)} - (M_{\mathrm{SAM}}\circ\phi)^{(i,j)} \cdot M_F^{(i,j)}\right) \cdot \left(1+\omega \cdot \left|\nabla^2 M_F^{(i,j)}\right|\right)}, \quad (7)$$

where ω represents the edge weighting coefficient and is set to 2.0 in our experiments. This design enables the registration network to refine local boundary deformation while preserving global structure alignment. The total training loss is a weighted combination of segmentation loss, registration loss, and EdgeWeightedIoU loss, defined as:

$$L_{Total} = \mu_1 \cdot L_{Seg} + \mu_2 \cdot L_{Reg} + \mu_3 \cdot L_{EdgeWeightedIoU}. \quad (8)$$

We adaptively adjust μ_i according to different tasks to balance gradient contributions from various supervised objectives. In this way, the model maintains reliable anatomical semantic priors and improves registration accuracy and boundary consistency.

3 Experiments

3.1 Dataset & Criteria

We validate our method on two datasets: a private dynamic tongue MRI dataset and the public ACDC cardiac MRI dataset [27]. The tongue dataset comprises coronal (90 healthy + 90 patient frames) and sagittal (23 subjects: 14 healthy, 9 post-operative) dynamic head and neck scans. Each subject covers three swallowing cycles (8 frames/cycle), with all images annotated by two professional physicians.

Registration performance is evaluated using three metrics: Dice, GNCC, and SSIM. Dice measures anatomical overlap between warped moving and fixed masks, while GNCC and SSIM—both computed from warped moving and fixed images—quantify intensity similarity and structural consistency, respectively.

3.2 Ablation Experiments

Table 1. Ablation experiment results of different model variants on the sagittal and coronal tongue MRI dataset, values in parentheses corresponding to the results before registration.

F_{En}	F_{SAM}	BCAF	$L_{EdgeWeightedIoU}$	Sagittal Tongue MRI			Coronal Tongue MRI		
				DICE (24.47)	GNCC (83.89)	SSIM (72.03)	DICE (91.26)	GNCC (80.41)	SSIM (68.35)
√				35.17	**95.59**	**84.03**	92.51	92.39	74.49
	√			33.68	94.24	81.74	92.68	92.75	76.07
√			√	37.49	94.65	82.44	95.69	93.94	*77.50*
	√		√	37.92	93.58	*83.89*	95.71	94.06	77.41
√	√		√	*39.74*	*94.90*	82.65	*95.79*	*94.13*	77.34
√	√	√	√	**42.72**	94.73	82.19	**96.21**	**94.16**	**77.54**

Ablation results on tongue MRI datasets are shown in Table 1 and Fig. 3. Row 1 is our Baseline, which uses only the registration-encoder features F_{En} without F_{SAM}, BCAF, and EdgeWeightedIoU loss. In single-feature decoding of Rows 1 and 2,F_{En} outperforms F_{SAM} on sagittal data across all metrics, with DICE 35.17 versus 33.68 while F_{SAM}excels on coronal data at DICE 92.68 versus 92.51, indicating that two feature types each suit a different view and are thus complementary. Adding EdgeWeightedIoU loss (Rows 3 and 4) consistently improves DICE on both views, with sagittal rising to 37.49, 37.92 and coronal rising to 95.69, 95.71, confirming the loss's effectiveness in region alignment.

Row 5, which fuses two features by simple concatenation, further enhances performance (DICE: 39.74 sagittal, 95.79 coronal), verifying feature complementarity. Row 6, the complete model that replaces concatenation with the BCAF module, achieves optimal DICE of 42.72 on sagittal and top performance across all metrics on coronal data. The slight GNCC/SSIM drop on sagittal data reflects a trade-off where larger deformations improve region overlap at a minor cost to global similarity, rather than a BCAF deficiency, as all three metrics improve on coronal data (DICE +0.42, GNCC +0.03, SSIM +0.2). Compared with concatenation, the BCAF module better exploits the two feature types, yielding the highest DICE on both views.

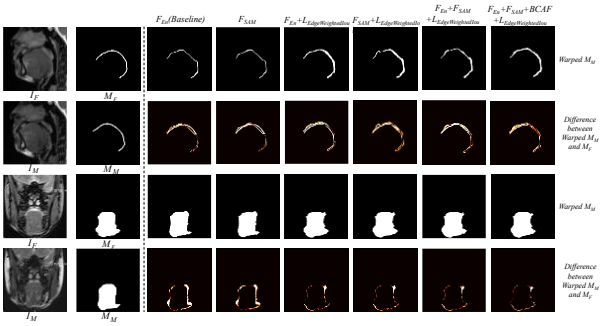

Fig. 3. Visualization comparison of warped moving labels, fixed labels, and corresponding difference maps on sagittal and coronal views of tongue MRI in ablation experiments.

3.3 Performance Comparisons with SOTA and Representative SAM Variants

Table 2. Registration performance of different models and their SAM-fused variants on coronal tongue MRI and ACDC cardiac MRI datasets (parentheses: pre-registration results)

Model（params）	Coronal Tongue MRI			*ACDC*		
	DICE (91.26)	GNCC (80.41)	SSIM (68.35)	DICE (89.11)	GNCC (94.72)	SSIM (93.57)
VoxelMorph （3.68M）	91.95	85.31	67.64	90.69	96.88	86.74
TransMorph （34.62M）	56.21	71.52	43.98	94.06	98.01	88.21
LDMMorph （35.11M）	92.11	92.59	73.61	94.52	98.79	90.76
our Baseline （7.76M）	92.69	*92.97*	*76.38*	*94.69*	**99.03**	**92.15**
VoxelMorph+SAM	*95.88*	92.68	74.56	94.54	98.11	87.42
TransMorph+SAM	91.43	85.13	59.61	94.24	97.69	86.42
our Baseline+SAM	**96.01**	**94.48**	**77.32**	**95.06**	*98.88*	*91.72*

As shown in Table 2 and Fig. 4, we compare our method with VoxelMorph, TransMorph [28], and LDM-Morph [29] on coronal tongue MRI and ACDC cardiac MRI

datasets and implement SAM-fused variants of these models and our Baseline. Here, each '+SAM' variant denotes a registration network equipped with our full SAM scheme: F_{SAM}, BCAF module, and the EdgeWeightedIoU loss. Parameter counts are reported to evaluate efficiency.

Without our SAM scheme, VoxelMorph (3.68M) shows moderate performance. TransMorph (34.62M) suffers from severe interference from complex head-neck tissues, with tongue DICE dropping to 56.21 (GNCC:71.52, SSIM:43.98), while cardiac performance remains relatively stable, indicating pure attention networks easily lose target features in complex images. LDM-Morph (35.11M) outperforms VoxelMorph and TransMorph, but still falls short of our Baseline. In contrast, our Baseline (7.76M) surpasses all competing methods on both datasets with far fewer parameters, validating the efficiency of our Swin-Intern encoder.

Integrating our SAM scheme consistently improves tongue MRI performance across all three registration networks. VoxelMorph+SAM shows steady improvements across metrics, confirming SAM's generalization ability. TransMorph+SAM boosts tongue DICE to 91.43, but its GNCC (85.13) and SSIM (59.61) gains lag far behind, reflecting that pure attention networks struggle to balance detail preservation and multi-metric optimization. Our Baseline+SAM achieves the best overall performance, ranking first on all coronal tongue MRI metrics and attaining the highest DICE on ACDC, demonstrating that our SAM scheme combines effectively with a well-designed registration encoder.

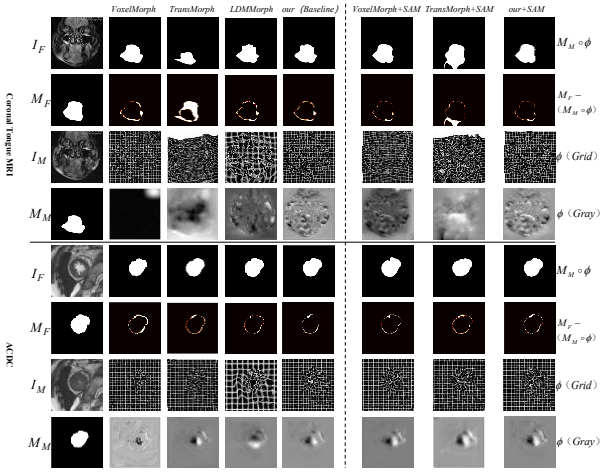

Fig. 4. Qualitative registration results and displacement field visualizations of all com-pared models on Coronal Tongue MRI and ACDC cardiac MRI datasets.

4 Conclusion

This study presents a novel SwallowReg architecture for dynamic tongue cine-MRI to assess postoperative swallowing function in tongue cancer patients. It extracts global-

local features by Swin-Intern blocks, and fuses them with SAM semantic priors through the BCAF module. It further adopts an EdgeWeightedIoU loss to refine anatomical boundary alignment. Experiments on a private tongue dataset and the public ACDC dataset demonstrate that it outperforms SOTA methods with efficient parameters, higher alignment accuracy and stronger motion field robustness, showing its practical value for clinical evaluation.

Acknowledgments. This work was funded in part by the Natural Science Foundation of China under Grant 12371440, in part by the Open Project Fund of the State Key Laboratory Cultivation Base of Research, Prevention and Treatment for Oral Disease under Grant JSKLOD-KF-2506, and in part by the Natural Science Foundation of Nanjing University of Posts and Telecommunications under Grant NY224142.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Reference

1. Oliveira, F.P., Tavares, J.M.R.: Medical image registration: a review. Computer Methods in Biomechanics and Biomedical Engineering. 17(2), 73–93 (2014).
2. Alam, F., Rahman, S.U., Ullah, S., Gulati, K.: Medical image registration in image guided surgery: Issues, challenges and research opportunities. Biocybernetics and Biomedical Engineering. 38(1), 71–89 (2018).
3. Azam, M.A., Khan, K.B., Ahmad, M., Mazzara, M.: Multimodal medical image registration and fusion for quality enhancement. Computers, Materials & Continua. 68(1), 821–840 (2021).
4. Gorbunova, V., Lo, P., Ashraf, H., Dirksen, A., Nielsen, M., de Bruijne, M.: Weight preserving image registration for monitoring disease progression in lung CT. In: International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008. LNCS, vol. 5242, pp. 863–870. Springer Berlin Heidelberg (2008).
5. Burus, T., Damgacioglu, H., Huang, B., Christian, W.J., Hull, P.C., Ellis, A.R., Lang Kuhs, K.A.: Trends in oral tongue cancer incidence in the US. JAMA Otolaryngology–Head & Neck Surgery. 150(5), 436–443 (2024).
6. Huang, Z.S., Chen, W.L., Huang, Z.Q., Yang, Z.H.: Dysphagia in tongue cancer patients before and after surgery. Journal of Oral and Maxillofacial Surgery. 74(10), 2067–2072 (2016).
7. Lin, P.C., Wang, W.C., Kao, Y.H., Matsuo, K., Lin, Y.C., Huang, H.L.: Long-Term Effects of an Oral Rehabilitation Programme on the Oral Function of Male Patients With or Without Tongue Cancer. Journal of Oral Rehabilitation. 52(10), 1810–1818 (2025).
8. Chien, C.Y., Chen, J.W., Chang, C.H., Huang, C.C.: Tracking dynamic tongue motion in ultrasound images for obstructive sleep apnea. Ultrasound in Medicine & Biology. 43(12), 2791–2805 (2017).
9. Yang, J., Mohamed, A.S., Bahig, H., Ding, Y., Wang, J., Ng, S.P., Joint Head and Neck Radiotherapy MRI Development Cooperative: Automatic registration of 2D MR cine images for swallowing motion estimation. PLoS One. 15(2), e0228652 (2020).

10. East, L., Nettles, K., Vansant, A., Daniels, S.K.: Evaluation of oropharyngeal dysphagia with the videofluoroscopic swallowing study. Journal of Radiology Nursing. 33(1), 9–13 (2014).
11. Hartl, D.M., Kolb, F., Bretagne, E., Bidault, F., Sigal, R.: Cine-MRI swallowing evaluation after tongue reconstruction. European Journal of Radiology. 73(1), 108–113 (2010).
12. Hartl, D.M., Albiter, M., Kolb, F., Luboinski, B., Sigal, R.: Morphologic parameters of normal swallowing events using single-shot fast spin echo dynamic MRI. Dysphagia. 18(4), 255–262 (2003).
13. Sun, M., Zhou, T., Jiang, C., Lv, X., Yu, H.: A new cine-MRI segmentation method of tongue dorsum for postoperative swallowing function analysis. In: MICCAI 2024. LNCS, vol. 15009, pp. 23–32. Springer, Cham (2024).
14. Lee, J., Woo, J., Xing, F., Murano, E.Z., Stone, M., Prince, J.L.: Semi-automatic segmentation for 3D motion analysis of the tongue with dynamic MRI. Computerized Medical Imaging and Graphics. 38(8), 714–724 (2014).
15. Bruijnen, T., Stemkens, B., Terhaard, C.H.J., Lagendijk, J.J.W., Raaijmakers, C.P.J., Tijssen, R.H.N.: Intrafraction motion quantification and planning target volume margin determination of head-and-neck tumors using cine magnetic resonance imaging. Radiotherapy and Oncology. 130(1), 82–88 (2019).
16. Bose, P., Brockton, N.T., Dort, J.C.: Head and neck cancer: from anatomy to biology. International Journal of Cancer. 133(9), 2013–2023 (2013).
17. Li, X., Zhang, Y., Shi, Y., Wu, S., Xiao, Y., Gu, X., Zhou, L.: Comprehensive evaluation of ten deformable image registration algorithms for contour propagation between CT and cone-beam CT images in adaptive head & neck radiotherapy. PLoS One. 12(4), e0175906 (2017).
18. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: Efficient non-parametric image registration. NeuroImage. 45(1), S61–S72 (2009).
19. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage. 54(3), 2033–2044 (2011).
20. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging. 38(8), 1788–1800 (2019).
21. Li, J., Chen, J., Tang, Y., Wang, C., Landman, B.A., Zhou, S.K.: Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. Medical Image Analysis. 85, 102762 (2023).
22. Meng, M., Fulham, M., Feng, D., Bi, L., Kim, J.: AutoFuse: Automatic fusion networks for deformable medical image registration. Pattern Recognition. 161, 111338 (2025).
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023).
24. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Qiao, Y.: InternImage: Exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14408–14419 (2023).
25. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021).
26. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722 (2021).

27. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Alvarez Sandoval, M.A., Sarasa, C.P., Tamez Urena, J.A., Frappart, F., Petit, C., Duchenne, G., Ducros Soler, J., Batchelor, P.G., Hawkes, D.J., Kohl, S.A.A., Isensee, F., Maier-Hein, K.H., Mansi, T.: Deep learning techniques for automatic MRI cardiac multi-structure segmentation and diagnosis: The ACDC challenge. IEEE Transactions on Medical Imaging. 38(2), 384–395 (2019).
28. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: TransMorph: Transformer for unsupervised medical image registration. Medical Image Analysis. 82, 102615 (2022).
29. Wu, J., Pan, T., Gong, K.: LDM-Morph: Latent diffusion model guided deformable image registration. Pattern Recognition. 174, 112925 (2026).