TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 11/21 pp1082–1091
DOI: 10.26599/TST.2023.9010070
Volume 29, Number 4, August 2024

Self-Aligning Multi-Modal Transformer for Oropharyngeal Swab Point Localization

Tianyu Liu and Fuchun Sun*

Abstract: The oropharyngeal swabbing is a pre-diagnostic procedure used to test various respiratory diseases, including COVID and Influenza A (H1N1). To improve the testing efficiency of testing, a real-time, accurate, and robust sampling point localization algorithm is needed for robots. However, current solutions rely heavily on visual input, which is not reliable enough for large-scale deployment. The transformer has significantly improved the performance of image-related tasks and challenged the dominance of traditional convolutional neural networks (CNNs) in the image field. Inspired by its success, we propose a novel self-aligning multi-modal transformer (SAMMT) to dynamically attend to different parts of unaligned feature maps, preventing information loss caused by perspective disparity and simplifying overall implementation. Unlike preexisting multi-modal transformers, our attention mechanism works in image space instead of embedding space, rendering the need for the sensor registration process obsolete. To facilitate the multi-modal task, we collected and annotate an oropharynx localization/segmentation dataset by trained medical personnel. This dataset is open-sourced and can be used for future multi-modal research. Our experiments show that our model improves the performance of the localization task by 4.2% compared to the pure visual model, and reduces the pixel-wise error rate of the segmentation task by 16.7% compared to the CNN baseline.

Key words: multi-modal perception; robotic perception; transformer; segmentation; localization

1 Introduction

The oropharyngeal swab (OP-swab) procedure is a pre-diagnostic measure for testing respiratory infectious diseases such as rhinovirus, adenovirus, influenza, respiratory syncytial virus (RSV), and others. However, it poses a risk of medical worker exposure and cross-infection between subjects. As a result, the COVID-19 pandemic and the surge of Influenza A (H1N1) have led to significant research interest in the field of oropharyngeal swab robots[1, 2]. While most existing robot systems use a semi-automatic or teleoperation approach[3], the semi-invasive sampling procedure requires a highly accurate, real-time, and robust localization method. However, the current oropharynx localization algorithm in these systems typically uses existing CNN architecture or pre-trained facial landmark detection algorithms designed only for RGB images. Depth information is either used for look-up only or directly dropped[4]. However, these practices are usually inaccurate when facing motion blur or lens glare.

Multi-modal sensory input has the potential to enhance reliability and accuracy but is hindered by the fact that most multi-modal cameras are primarily designed for long-distance tasks[5]. Aligning different sensors for close-up tasks results in significant disparities between sensors[6], and the registration

• Tianyu Liu and Fuchun Sun are with Department of Computer Science and Technology, Tsinghua University, Beijing 100083, China. E-mail: lty22@mails.tsinghua.edu.cn; fcsun@tsinghua.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2023-04-05 ; revised: 2023-07-04; accepted: 2023-07-11

process can cause significant information loss due to clipping caused by the wall of the oral cavity, as shown in Fig. 1.

To avoid performance loss due to alignment, we explored the use of transformers. Recent patch-based visual transformers have achieved comparable or even better performance than state-of-the-art convolutional neural networks (CNNs) in multiple computer vision tasks[7]. The transformer architecture has already been widely adopted in various multi-modal tasks when processing sequence data. This inspired us to leverage the dynamic weighting potential of the transformer architecture in camera space. To address this, we proposed a novel self-aligning multi-modal transformer (SAMMT) that can dynamically attend to different modalities without the need for explicit alignment.

We evaluated the performance of our network by creating a dataset with segmentation and key-point localization labels and making it open-source. Our experiments on this dataset show that our model is able to effectively combine different modalities, producing reliable and accurate localization results. When compared to a vision-only model, our proposed model achieved a 4.22% improvement in performance (reducing the localization error from 3.89 mm to 3.72 mm). Additionally, for the segmentation task, our model achieved a 16.7% decrease in error rate (from 4.44% to 3.70%) compared to CNN solutions that only rely on vision.

This paper presents the following main innovations and contributions:

- **SAMMT.** The proposed novel multi-modal transformer can automatically attend to different fractions of every sensory input. It can work on unaligned 2D feature maps without the need for explicit sensor registration.
- **First open-access OP-swab point localization dataset.** The dataset includes both the sampling region and the optimal sampling point, which have been annotated by trained medical personnel. This multi-modal dataset can also be utilized for future research on multi-modal algorithms.

The paper is organized as follows: Section 2 covers recent works in related fields, including visual transformers, multi-modal transformers, and existing oropharynx detection and localization methods. In Section 3, we propose our self-aligning multi-modal transformer and provide a formal description. The dataset and the data collection process are described in Section 4. In Section 5, we report a series of experiments conducted to evaluate the localization/segmentation performance of our proposed model. Finally, the conclusion and possible future work are presented in Section 6.

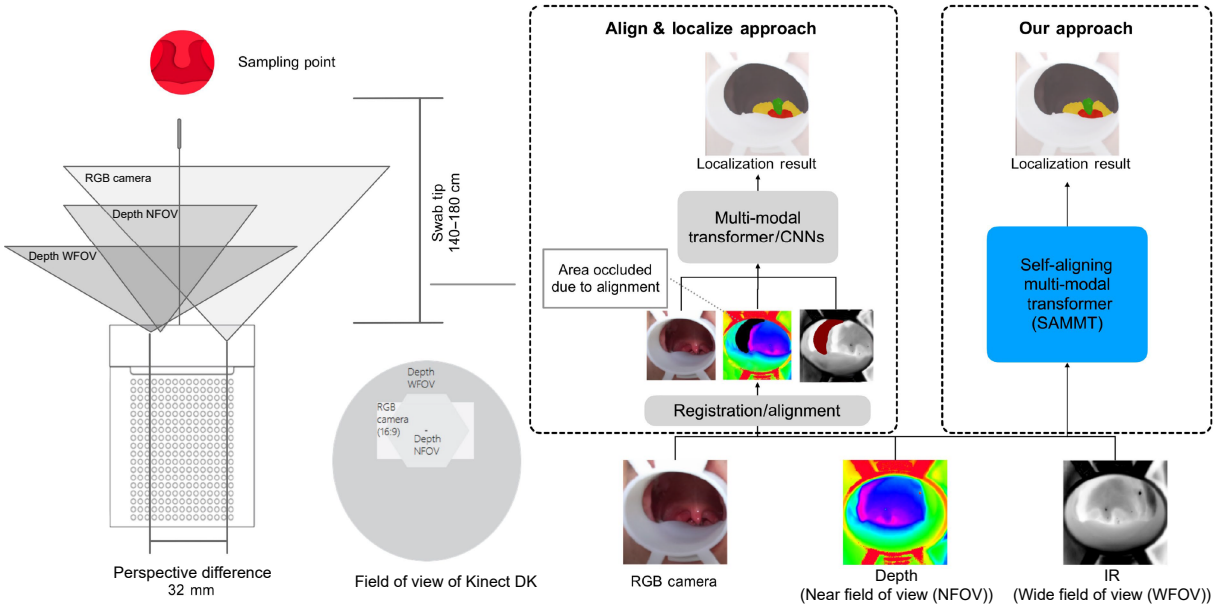

Fig. 1 Traditional alignment and localization paradigms suffering from information loss caused by perspective clipping (indicated by black areas in the depth image and red areas in the IR image). Our proposed self-aligning multi-modal transformer can dynamically attend to different parts of each modality and generate localization results without requiring alignment.

2 Related Work

2.1 Visual transformer

Transformers were initially proposed for natural language processing (NLP) tasks such as machine translation in 2017[8]. They have gradually become the state-of-the-art architecture for most NLP tasks. Both bidirectional encoder representations from transformers (BERT)[9] and generative pre-trained transformer (GPTs)[10] are different variants of transformers trained on large corpora with different pre-training tasks.

Transformer application in computer vision has recently gained extraordinary progress, achieving state-of-the-art performance in various computer vision tasks and outperforming convolution-based neural networks[11]. The naive application of self-attention transformer works in the pixel space with a meticulously designed sliding window[12]. The architecture achieved acceptable performance with small images. However, the limited sliding window compared with the gradually growing image size causes the quadratic cost in the number of pixels to quickly become a problem.

Extracting fixed-size patches and applying self-attention on patch space has recently become a common practice[13], achieving competitive performance compared to state-of-the-art CNNs. Combined with effective pre-training tasks, this approach yields better performance than CNNs in various computer vision tasks such as image classification[13–15], object detection[16, 17], video processing[18], image segmentation[19], and image generation[20, 21].

Recent research has explored the performance of models on extensive datasets using unsupervised or self-supervised methods[22, 23]. While the pyramid vision transformer (PVT)[24] was the first to use a pyramid structure for dense prediction tasks like semantic segmentation, subsequent methods like Swin[21], CvT[25], Twins[23], and SegFormer[19] have introduced tailored changes to the pyramid structure to further improve segmentation performance. Our work extends the capabilities of the visual transformer to other modalities, such as depth and infrared (IR) information, to improve the model's robustness. Inspired by all patch-based visual transformers, we focus on improving localization and segmentation accuracy with a transformer-based multi-modal transformer. Previous research has already demonstrated the potential of transformer architecture in sequence modeling.

2.2 Transformers in multi-modal tasks

The transformer has been extensively used for various multi-modal tasks, with a primary focus on sequence modeling and alignment. These tasks include visual question answering (VQA)[26, 27], image captioning[28], vision-and-language navigation[29], and joint video and language modeling[30]. Most of the aforementioned work concentrates on aligning a sequence to another sequence or a 2D image to an existing sequence. However, few of them concentrate on different 2D image-like sensory inputs.

Most approaches to multi-modal localization rely on constructing a multi-head attention layer on top of different extracted embedding features/tokens, constituting an attention mechanism in feature space. This is similar to the native application of the transformer in the field of NLP, where the vanilla multi-layer transformer can be applied without introducing too much architectural change. However, such an implementation relies on a feature extraction network to extract spatial information, which is usually not meticulous and accurate. A related model is cross-model attention[31], which constructs pair-wise cross-modal attention blocks on top of each extracted embedding pair and applies full self-attention on top.

In contrast, our model is fundamentally different from these multi-modal transformers. Our model goes further than constructing a transformer on top of embedding space to facilitate a more accurate segmentation head in the original camera space. Additionally, we discard the pair-wise design to simplify the overall architecture and reduce model complexity.

Another challenge for multi-modal applications is registration and alignment. Traditionally, multi-modal segmentation tasks in the field of CNN are done by aggregating multiple feature maps, including low-level high-res ones, which require different sensory inputs to be registered in the same camera space. The alignment accuracy can significantly affect the final segmentation performance. Our model utilizes the dynamic weighting potential of the transformer to allow the model to attend to the relevant part of different models dynamically, eliminating the need for sensor

registration and hence reducing information loss in the process.

2.3 OP-swab localization/segmentation methods

Traditionally, OP-swab robots have used a teleoperation approach, which requires extensive operator training, and is therefore not suitable for large-scale deployment[3]. Other robots with sampling point localization potential[32] use only RGB images, and depth data are either dropped or clipped for locomotion purposes[4], which could have been used to improve performance and even robot safety.

As shown in Fig. 1, the cone-like structure of the human oral cavity and the limited camera distance make the clipping caused by the parallax of different sensors particularly unacceptable. Additionally, mist due to breathing and lens glare on the RGB camera can further impact localization accuracy.

3 Method

In this section, we propose a multi-modal transformer that can dynamically attend to a 2D feature map of different modalities without the need for specific registration or alignment. The architecture of the proposed transformer is shown in Fig. 2. Inspired by the success of the vision transformer (ViT)[13], we split each sensory input into patches to attain a more fine-grained spatial structure framework. On top of the extracted patch embedding, we can construct our transformer in sensor space instead of feature space to attain a more precise spatial feature representation and dynamically attend to different parts of different sensory input, eliminating the need for complicated sensor registration processes.

Details of the self-aligned multi-modal transformer's backbone are introduced in Section 3.1. The self-attention building block used in our backbone is further elaborated in Section 3.2. We construct both segmentation and regression heads to train our network and evaluate the sampling point accuracy, which is shown in Section 3.3.

3.1 Self-aligned multi-modal vision transformer

Our model closely follows the original transformer design[8] and the original visual transformer[13], enabling us to use existing efficient and scalable NLP transformer architectures directly off the shelf. An overview of the model is shown in Fig. 3. Each sensory input $x_1, x_2, \ldots, x_M \in \mathbb{R}^{H_i W_i C_i}$ is reshaped into a sequence of fixed-sized, flattened 2D patches $x_{i,j} \in \mathbb{R}^{N\times(P^2\dot{C})}$, where $i \in [1,2,\ldots,M]$ and $j \in [1,2,\ldots,N]$, with C as the number of modalities and $N = HW/P^2$ as the number of patches. N represents the number of patches into which an image is divided. H and W are the height and width of the image, respectively, representing the image's dimensions. P is the side length of a square patch. The resolution of each modality is uniformly denoted by (H, W) for ease of explanation. Therefore, the input sequence length is $M \times N$. Since the transformer can naturally accept varying length input sequences, our trained model can work on different sizes of input images from different modalities, provided that the mix of modalities remains unchanged.

Due to the relatively small size of each patch, a trainable linear projection layer is used to extract patch embeddings in our practice:

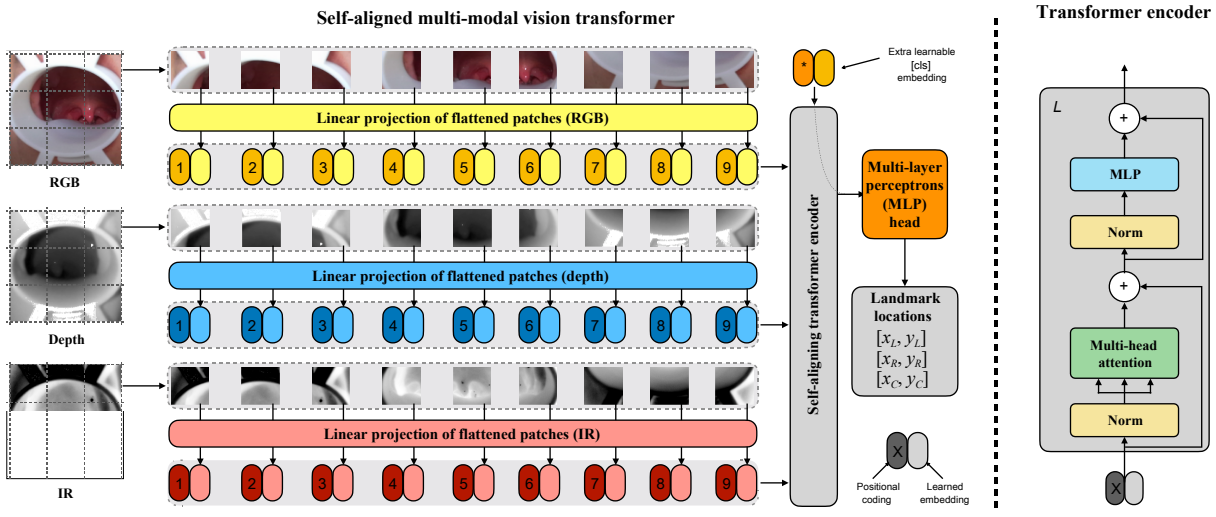

Fig. 2 Our proposed self-aligning multi-modal transformer (SAMMT) for OP-swab sampling region localization.

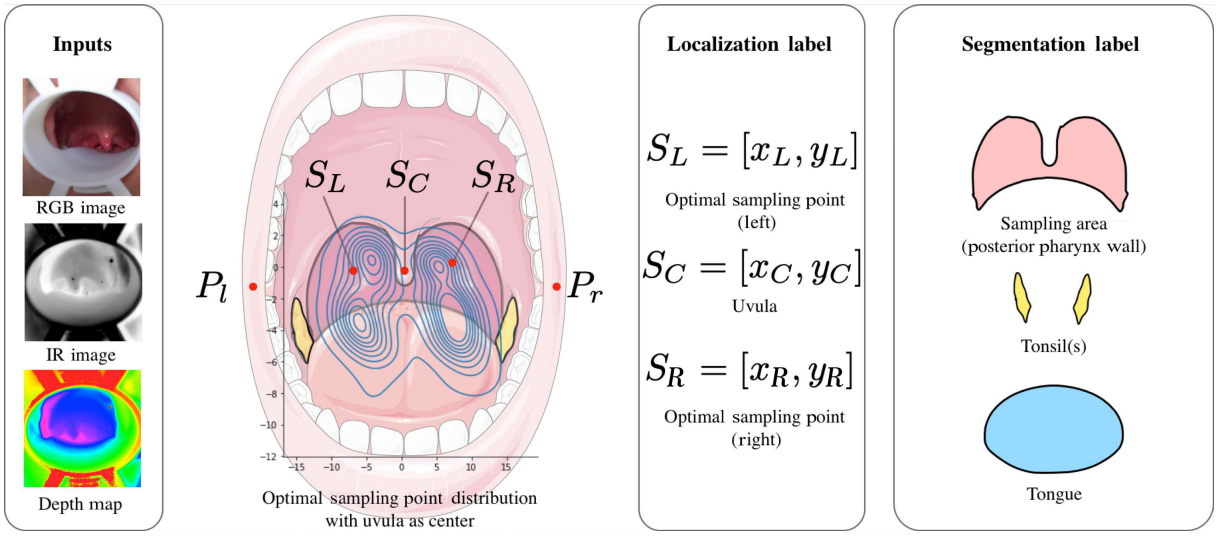

Fig. 3 Our collected multi-modal oropharynx localization dataset and annotation example.

$$z_0 = [\boldsymbol{x}_{\text{cls}}; \boldsymbol{x}_p^1 \boldsymbol{U}; \boldsymbol{x}_p^2 \boldsymbol{U}; \ldots; \boldsymbol{x}_p^{M \times N} \boldsymbol{U}] + \boldsymbol{U}_{\text{pos}} \tag{1}$$

where $\boldsymbol{U} \in \mathbb{R}^{(P^2 \dot{C}) D}$ and $\boldsymbol{U}_{\text{pos}} \in \mathbb{R}^{(NM+1) \times D}$ represent trainable weights for linear projection and positional encoding, respectively. We use standard learnable 1-D embeddings since limited modalities do not require more advanced 2D embeddings. Additionally, we add a standalone learnable embedding ($z_0^0 = \boldsymbol{x}_{\text{cls}}$) into the pool of patch embeddings, similar to BERT's [class] token[9]. The corresponding output (z_L^0) serves as the input of the sampling point localization head.

3.2 Transformer encoder

The embedding and positional encoding pairs are then fed into a transformer encoder[8], which consists of different attention blocks composed of multi-head self-attention (MSA) and MLP layers. Furthermore, layer normalization is applied before each block, and residual links after every block[33, 34].

The specific transformer encoder is also shown in Fig. 3. The encoder consists of MSA and MLP layers. Residual connections are connected between each block to allow a more direct gradient flow.

$$z_l' = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1} \tag{2}$$

$$z_l = \text{MLP}(\text{LN}(z_l')) + z_l' \tag{3}$$

$$\boldsymbol{y}^i = \text{LN}(z_L^i),\ i \in [0, M \times N + 1) \tag{4}$$

where z_l' and z_l both represent an intermediate or updated hidden state in the transformer model, and $\boldsymbol{y}^i$ is generated hidden states after layer normalization.

The multi-head self attention layers use standard $\boldsymbol{qkv}$ self-attention block[8]. For easier understanding, we first introduce single-headed self-attention, where each pair of embedding and positional encoding $z \in \mathbb{R}^{N \times D}$ ($N = 3 \times W \times H + 1$ for the first layer) generates a tuple of query ($\boldsymbol{q}$), key ($\boldsymbol{k}$), and value ($\boldsymbol{v}$) are computed by

$$[\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}] = \boldsymbol{z} \boldsymbol{U}_{qkv} \tag{5}$$

where $\boldsymbol{U}_{qkv} \in \mathbb{R}^{D \times 3 D_h}$ is a trainable weight matrix used to generate the tuple, D_h is the dimension of each hidden $\boldsymbol{qkv}$ value. And the attention weight matrix $\boldsymbol{A}$ are obtained by calculating the pairwise similarity between key-value $\boldsymbol{k}_j$ and query value $\boldsymbol{q}_j$ each two elements of the patch pool

$$\boldsymbol{A} = \text{softmax}(\boldsymbol{q}\boldsymbol{k}^{\text{T}} / \sqrt{D_h}) \in \mathbb{R}^{N \times N} \tag{6}$$

The final self-attention result is then easily obtained by

$$\text{SA}(z) = \boldsymbol{A}\boldsymbol{v} \tag{7}$$

We employed a multi-head self-attention mechanism in our model, consisting of h self-attention operations in each layer, which run in parallel. This approach allows for a more diverse attention flow to pass through. The results of each single-headed self-attention are then concatenated and projected, resulting in a constant output dimension, regardless of the value of h.

$$\text{MSA}(z) = [\text{SA}_1(z), \text{SA}_2(z), \ldots, \text{SA}_h(z)] \boldsymbol{U}_{\text{msa}} \tag{8}$$

where $\boldsymbol{U}_{\text{msa}} \in \mathbb{R}^{h \times D_h \times D}$.

3.3 Regression head and localization head

Our transformer backbone is augmented by two

prediction heads that simultaneously output optimal sampling points and segmentation results, improving accuracy and safety.

A three-layer MLP with ReLU activation is responsible for localizing the sampling points. The head takes z_L^0 as input and outputs keypoint coordinates. To handle small localization errors, we use wing loss during training, as shown in Eq. (9)[35].

$$\mathcal{L} = \begin{cases} w \times \ln(1 + |x|/\epsilon), & \text{if } |x| < w; \\ |x| - C, & \text{otherwise} \end{cases} \tag{9}$$

where w is a parameter used to adjust the behavior of the loss function. x typically represents the actual prediction error, which is the difference between the model's predicted value and the true value. In the context of Wing Loss, x signifies the measure of error. ϵ is a parameter used to normalize the error x. Its role is to scale the error value to an appropriate range, aiding in the computation of the loss function.

For the segmentation head, an all-MLP head is used to reduce computation costs. Since our data are unaligned, we only need the segmentation mask results in our annotated modality. Hence only $\boldsymbol{y}^{1-N}$ are fed into the segmentation head. And the network is trained with cross-entropy loss.

4 Dataset

To the best of our knowledge, our dataset is the first oropharynx localization/segmentation dataset for OP-swab robot systems. This open-source dataset can be used for future oropharynx sampling point localization/segmentation tasks. Additionally, our multi-modal sensory input supports future multi-modal perception research.

During the swab sampling process, the perspective of each sensor changes dramatically. The RGB camera captures images that are noticeably affected by motion blur and lens glare. This poses a new challenge for existing deep-learning models, which must leverage information gathered by other sensors.

Our data were collected using a Kinect Azure mounted on the end actuator of an OP-swab robot. Trained personnel operated the robot to collect OP-swab samples while recording sensory inputs. The dataset includes RGB data collected in 4K format, and IR and depth data in WFOV format. After cleaning and reprocessing the data, trained medical personnel annotated it. Both the valid sampling areasegmentation (posterior wall of the pharynx) and optimal sampling points are labeled. The sampling area can be used for semantic segmentation, while the optimal sampling points can be used for localization.

The dataset consists of 844 instances of sensory input from various sampling stages, aimed at ensuring the stability and accuracy of our trained model's guidance ability. Our robot collected 647 of these data from an external perspective of the oral cavity across 13 participants. Furthermore, 197 instances provide fine-grained support from an intraoral perspective. Additionally, there is a set of 1000 data where the posterior wall of the pharynx is not sufficiently exposed, which can be used as a negative set.

The dataset can be accessed through https://github.com/Acce1erat0rS/MMSWAB.

5 Experiment

5.1 Experimental setup

The image is fed into a self-aligning multi-modal transformer model trained by PyTorch 1.8 using one NVIDIA RTX 3090 GPU. Model inputs are reshaped to 256 × 256 to ensure fast inference speed; each patch size is 32 × 32, making a total of $8 \times 8 \times 3 = 192$ patches. The transformer has 16 attention heads with dimensions of 128. The final MLP width is set to 128 with a dropout of 0.1.

Since the innate difference between depth/IR images and RGB images, concatenating these images into a big image and using a ViT pre-trained on Image-Net does not give us much performance gain. During training, the image is augmented with mean subtraction, random resizing, shearing, shifting, and zooming with a ratio between 0.5 and 2.0, and, as a final step, random left-right flipping.

5.2 Sampling point localization result

According to the anatomical characteristics of the human body, the average width of the pharynx is around 10 mm, and the height is no more than 20 mm. Both annotated location and localized position are converted to camera space through application programming interface (API) provided by the Kinect camera. After evaluating on 100 samples, sampling point localization results on different modalities are shown in Table 1. The overall sampling point deviation based on visual input is 3.89 mm. Even though IR and depth themselves contain limited information (achieving an accuracy of around 10 mm),by adding IR

Table 1 Result about average displacement error (ADE) of OP-swab point generated from different modalities and trained with different losses.

Modality	ADE (cm)		
	L2 loss	L1 loss	Wing loss
V (RGB)	0.7264	0.5615	0.3890
I (IR)	1.1323	1.0296	0.9802
D (depth)	1.0763	0.9710	0.9049
V+I	0.7022	0.5359	0.3833
V+D	0.6897	0.5235	0.3770
V+I+D	0.7134	0.5224	**0.3726**

and depth information, the mean deviation still improved to 3.726 mm (+0.164, 4.2%), which indicates that the self-aligning multi-modal transformer can leverage unregistered multi-modal sensory inputs.

Also, other losses are used for training; the results show that wing loss is significantly better for fine-scale localization tasks than traditional L1 and L2 losses.

In order to gain a more direct image of the performance of our model, The kernel density estimation (KDE) of localization error is shown in Fig. 4, which is the localization error of S_L and S_R. As we can see in these estimations, compared with the size of the pharynx wall, the final localization accuracy is enough for OP-swab sampling. This particular setup results in an inference speed on a single NVIDIA 3090 of around 19 ms, enough for real-time sampling robot guidance.

5.3 Segmentation result

The segmentation results are shown in Fig. 5. The results shown here consist of two parts. The first seven images were captured as the robot gradually moved closer to the subject, and the last seven are segmentation results of intraoral images.

The first row shows the RGB input, and the second row shows the ground truth annotated by trained medical personnel. The third and fourth rows show the segmentation results attained by a UNet++ structure with ResNet34 as the backbone. The last two rows show the segmentation results generated by our multi-modal transformer.

Based on our observations, depth information plays a vital role in semantic segmentation. Due to poor lighting conditions and lack of depth information, the UNet++ baseline achieved a pixel-wise accuracy of 95.65%. Our proposed transformer achieved a pixel-wise accuracy of 96.30%, which is a 16.7% reduction

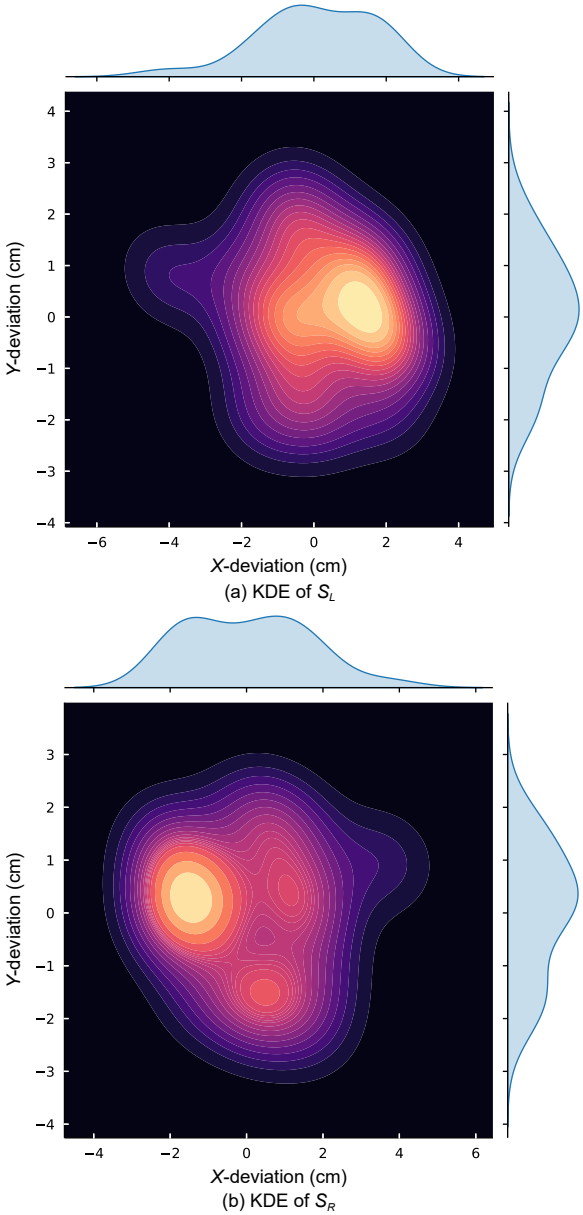

Fig. 4 KDE result of localization error.

in error rate. The mean intersection over union (mIOU) of our model is shown in Table 2. Our model achieved a mean IoU of 63.29% with a manually adjusted threshold of 0.27. It can also be noticed that the accuracy on intraoral images is significantly higher than on the external perspective. We assume that the relatively small segmentation area of the external perspective and unbalanced dataset pose a challenge for existing models, which could be researched further.

6 Conclusion

Aiming to address typical problems of multi-modal perception, which are particularly prominent in the

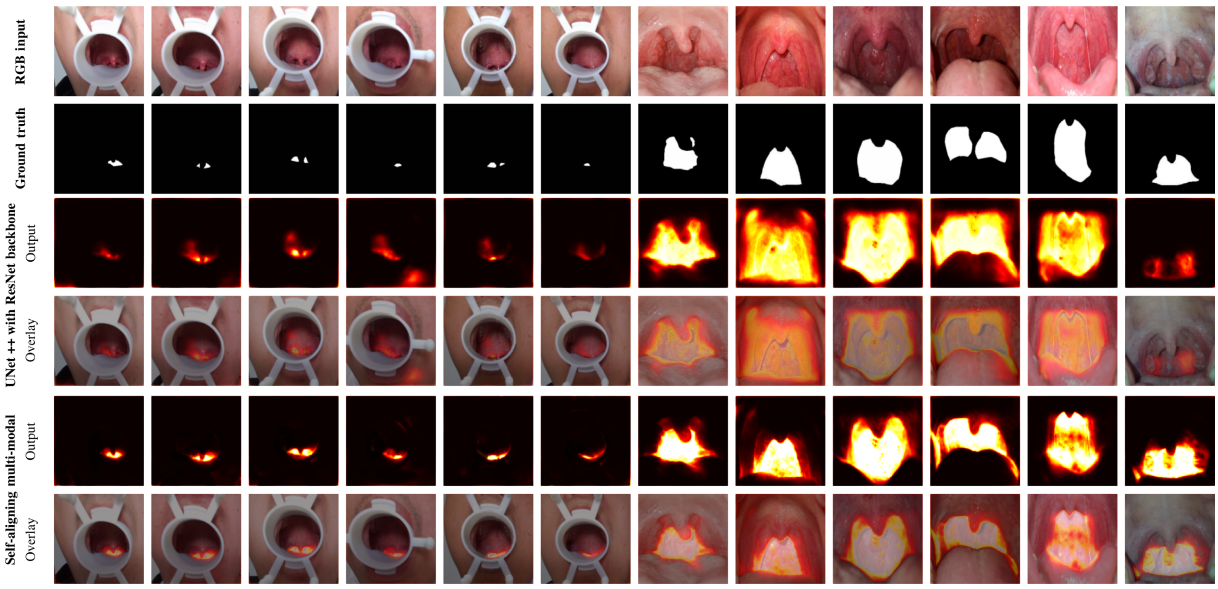

Fig. 5 Segmentation results of sampling regions in different stages of sampling and with different models.

Table 2 Segmentation results in different circumstances compared with traditional methods.

Model	Modality	mIOU (%)		
		External	Intraoral	Overall
UNet++	Visual	32.92	43.48	36.84
UNet++	Multi-modal	34.82	42.33	37.04
SAMMT	Multi-modal	**55.78**	**70.14**	**63.29**

field of oropharynx localization, this paper introduces a pure transformer-based multi-modal backbone for both segmentation and localization tasks. We observed that the transformer can dynamically capture different patches from different modalities. Therefore, we deployed a transformer on top of extracted patch embeddings of different modalities. We have also constructed both localization heads for key-point localization and a segmentation head for semantic segmentation. The constructed model was trained on our collected oropharynx localization dataset and performed better than pure visual and CNN-based methods.

In future research, we can consider substituting our backbone with a pyramid-like structure such as SwinTransformer and Segformer to acquire better segmentation results and study the imbalance problem of small target segmentation. Our dataset also provides a foundation for various future multi-modal research.

We believe that our proposed SAMMT is the first step towards an end-to-end multi-modal system without the need for explicit registration and alignment, and could bring about more exciting advancements.

Acknowledgment

This work was supported in part by the Sino-German Collaborative Research Project Crossmodal Learning (No. NSFC 62061136001/DFG TRR169).

References

[1] Y. L. Chen, F. J. Song, and Y. J. Gong, Remote human-robot collaborative impedance control strategy of pharyngeal swab sampling robot, in *Proc. 2020 5th Int. Conf. Automation, Control and Robotics Engineering* (*CACRE*), Dalian, China, 2020, pp. 341–345.

[2] G. Z. Yang, B. J. Nelson, R. R. Murphy, H. Choset, H. Christensen, S. H. Collins, P. Dario, K. Goldberg, K. Ikuta, N. Jacobstein, et al., Combating COVID-19—The role of robotics in managing public health and infectious diseases, *Sci. Robot.*, vol. 5, no. 40, p. eabb5589, 2020.

[3] S. Q. Li, W. L. Guo, H. Liu, T. Wang, Y. Y. Zhou, T. Yu, C. Y. Wang, Y. M. Yang, N. S. Zhong, N. F. Zhang, et al., Clinical application of an intelligent oropharyngeal swab robot: Implication for the COVID-19 pandemic, *Eur. Respir. J.*, vol. 56, no. 2, p. 2001912, 2020.

[4] Z. Xie, B. Chen, J. Liu, F. Yuan, Z. Shao, H. Yang, A. G. Domel, J. Zhang, and L. Wen, A tapered soft robotic oropharyngeal swab for throat testing: A new way to collect sputa samples, *IEEE Robot. Autom. Mag.*, vol. 28, no. 1, pp. 90–100, 2021.

[5] M. Draelos, N. Deshpande, and E. Grant, The Kinect up close: Adaptations for short-range imaging, in *Proc. 2012 IEEE Int. Conf. Multisensor Fusion and Integration for Intelligent Systems* (*MFI*), Hamburg, Germany, 2012, pp. 251–256.

[6] C. D. Herrera, J. Kannala, and J. Heikkilä, Joint depth and color camera calibration with distortion correction, *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 34, no. 10, pp. 2058–2064, 2012.

[7] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y.

Tang, A. Xiao, C. Xu, Y. Xu, et al., A survey on visual transformer, arXiv preprint arXiv: 2012.12556, 2020.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805, 2018.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv: 2005.14165, 2020.

[11] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, Visual transformers: Token-based image representation and processing for computer vision, arXiv preprint arXiv: 2006.03677, 2020.

[12] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, Image transformer, in *Proc. ICML 2018*: *Int. Conf. Machine Learning*, Sanya, China, 2018, pp. 4055–4064.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16×16 words: Transformers for image recognition at scale, arXiv preprint arXiv: 2010.11929, 2020.

[14] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, Attention augmented convolutional networks, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Seoul, Republic of Korea, 2020, pp. 3285–3294.

[15] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, Training dataefficient image transformers & distillation through attention, in *Proc. ICML 2021*: *Int. Conf. Machine Learning*, Virtual Event, 2021, pp. 10347–10357.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, End-to-end object detection with transformers, in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J. M. Frahm, eds. Cham, Switzerland: Springer, 2020, pp. 213–229.

[17] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, Deformable DETR: Deformable transformers for end-to-end object detection, arXiv preprint arXiv: 2010.04159, 2020.

[18] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, Video transformer network, arXiv preprint arXiv: 2102.00719, 2021.

[19] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, arXiv preprint arXiv: 2105.15203, 2021.

[20] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, Zero-shot text-to-image generation, arXiv preprint arXiv:, 2102.12092, 2021.

[21] 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, arXiv preprint arXiv: 2103.14030, 2021.

[22] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, Generative pretraining from pixels, in *Proc. 37th Int. Conf. Machine Learning*, Virtual Event, 2020, pp. 1691–1703.

[23] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, Twins: Revisiting the design of spatial attention in vision transformers, arXiv preprint arXiv: 2104.13840, 2021.

[24] W. Wang, E. Xie, X. Li, D. P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, arXiv preprint arXiv: 2102.12122, 2021.

[25] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, CvT: Introducing convolutions to vision transformers, arXiv preprint arXiv: 2103.15808, 2021.

[26] H. Tan and M. Bansal, LXMERT: Learning cross-modality encoder representations from transformers, arXiv preprint arXiv: 1908.07490, 2019.

[27] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, VL-BERT: Pre-training of generic visual-linguistic representations, arXiv preprint arXiv: 1908.08530, 2019.

[28] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, Unified vision-language pre-training for image captioning and VQA, *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 13041–13049, 2020.

[29] W. Hao, C. Li, X. Li, L. Carin, and J. Gao, Towards learning a generic agent for vision-and-language navigation via pre-training, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 13134–13143.

[30] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, VideoBERT: A joint model for video and language representation learning, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision* (*ICCV*), Seoul, Republic of Korea, 2020, pp. 7463–7472.

[31] S. Yao and X. Wan, Multimodal transformer for multimodal machine translation, in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, Virtual Event, 2020, pp. 4346–4350.

[32] Design motor drive systems with ease, https://www.therobotreport.com/danish-startupdevelops-throat-swabbing-robot-for-covid-19-testing/, 2020.

[33] A. Baevski and M. Auli, Adaptive input representations for neural language modeling, arXiv preprint arXiv: 1809.10853, 2018.

[34] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. F. Wong, and L. S. Chao, Learning deep transformer models for machine translation, arXiv preprint arXiv: 1906.01787, 2019.

[35] Z. H. Feng, J. Kittler, M. Awais, P. Huber, and X. J. Wu, Wing loss for robust facial landmark localisation with convolutional neural networks, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 2235–2245.

Fuchun Sun received the BEng and MEng degrees in automation from Naval Aeronautical Engineering Institute in China in 1986 and 1989, respectively, and the PhD degree in computer science and technology from Tsinghua University, China in 1997. He is a full professor at Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interest includes robotic perception and cognition. He was a recipient of the National Science Fund for Distinguished Young Scholars. He serves as an associate editor of a series of international journals including the *IEEE Transactions on Fuzzy Systems*, *Mechatronics*, and *Robotics and Autonomous Systems*.

Tianyu Liu received the BEng degree in computer science from Beijing Forestry University, China in 2019. He is currently pursuing the PhD degree in computer science at Tsinghua University, China. His main field of work includes neural network architecture, multi-modal perception and fusion, and computer vision algorithms in the autonomous driving scenario.