Online Self-distillation and Self-modeling for 3D Brain Tumor Segmentation

Yan Pang, *Member, IEEE*, Yunhao Li, Teng Huang*, Jiaming Liang, Zhen Wang, Changyu Dong, Dongyang Kuang, Ying Hu, Hao Chen, *Senior Member, IEEE*, Tim Lei, Qiong Wang*, *Member, IEEE*

***Abstract*— In the specialized domain of brain tumor segmentation, supervised segmentation approaches are hindered by the limited availability of high-quality labeled data, a condition arising from data privacy concerns, significant costs, and ethical issues. In response to this challenge, this paper presents a training framework that adeptly integrates a plug-and-play component, MOD, into current supervised learning models, boosting their efficacy in scenarios with limited data. The MOD consists of an Online Tokenizer and a Dense Predictor, which employs self-distillation and self-modeling on masked patches, promoting swift convergence and efficient representation learning. During the inference phase, the plug-and-play MOD component is excluded, preserving the computational efficiency of the original model without incurring extra processing costs. We substantiated the value of our approach through experiments on leading 3D brain tumor segmentation baselines. Remarkably, models augmented with the MOD consistently showcased superior results, achieving elevated Dice coefficients and HD95 scores on two datasets: BraTS 2021 and MSD 2019 Task-01 Brain Tumor. Code: https://github.com/aigzhusmart/MOD**

***Index Terms*— 3D Medical Image Segmentation, Image Generation, Representation Learning**

I. Introduction

Medical image segmentation on the brain serves as a cornerstone within the healthcare domain, specifically within the context of diagnostic and therapeutic procedures, thereby accentuating its role in facilitating accurate, patient-focused medical outcomes [1], [2]. In general, current predominant frameworks incorporate a backbone network linked to a segmentation head, both of which undergo supervised learning on comparatively smaller datasets of medical images [3]. The backbone network primarily undertakes the extraction of valuable representations from a vast pool of labeled images, subsequently enabling the segmentation head to produce masks based on these features [4], [5]. Generally, an increase in the volume of labeled data utilized for training tends to improve the overall segmentation performance. Nevertheless, complications associated with data privacy, substantial costs, and ethical considerations contribute to the limited high-quality labeled medical data [6], [7], presenting considerable obstacles to traditional supervised segmentation methodologies. This challenge highlights the pressing need for the advancement of more precise segmentation strategies.

Advancements in brain tumor segmentation have been largely fueled by progress in general computer vision, especially with hybrid transformer-based frameworks. These innovations have enabled more precise identification and delin-

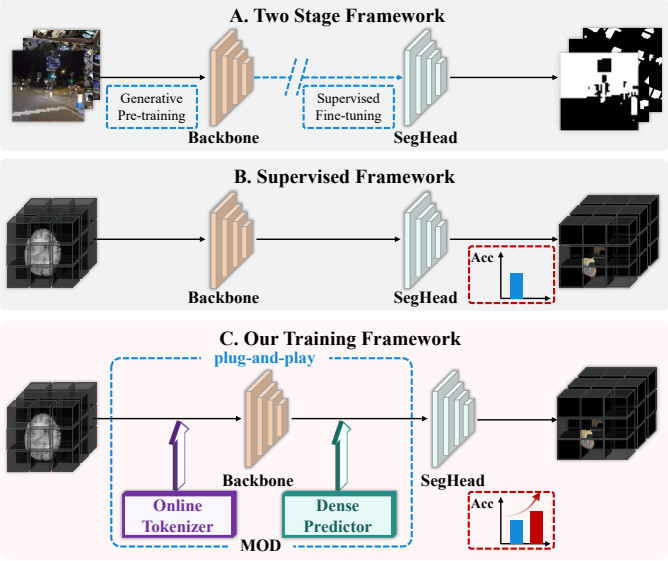

Fig. 1: Comparison of the architectural design between a conventional medical image segmentation framework and our proposed general framework, MOD, which is specifically designed to accelerate the training process.

This work was supported by the National Key R&D Program of China under grant No.2023YFB4705700; in part by NSFC Key Project U23A20391in part by the Regional Joint Fund of Guangdong (Guangdong–Hong Kong–Macao Research Team Project) under Grant 2021B1515130003.(*Corresponding authors: Teng Huang and Qiong Wang.)

Yan Pang, Ying Hu, and Qiong Wang are with the Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China (email: yanpang, ying.hu, wangqiong@siat.ac.cn)

Yunhao Li, Teng Huang, Jiaming Liang are with the School of Artificial Intelligence, Guangzhou University, China. (email: yunhaoli, huangteng1220, jiaming.liang, Changyu.dong@gzhu.edu.cn)

Zhen Wang is with the Zhejiang Lab, Hangzhou, China.(email: wangzhen@zhejianglab.com)

Dongyang Kuang is with the School of Mathematics (Zhuhai), Sun Yat-sen University, China. (email: kuangdy@mail.sysu.edu.cn)

Hao Chen is with the Department of Computer Science and Engineering and the Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, China. (email: jhc@cse.ust.hk)

Tim Lei is with the Department of Electrical Engineering, University of Colorado Denver. (email: tim.lei@ucdenver.edu)

eation of structures in current medical images [8]. To counter the limitation of high-quality labeled data, a two-stage process, depicted in Fig. 1A, involving generative pre-training followed by supervised fine-tuning, has been adopted. Generative pre-training leverages large volumes of unlabeled data to detect essential patterns and correlations [9], providing a foundational understanding of data distribution. This phase is particularly vital in situations with limited high-quality, labeled training data. Subsequently, the supervised fine-tuning stage refines this knowledge on a smaller, task-specific labeled dataset, aligning the model's capabilities with the specific requirements of the segmentation task. This two-stage approach ensures that the model not only grasps broad data characteristics but also hones in on specific features necessary for effective segmentation performance.

Transferring general computer vision advancements to medical image segmentation faces significant challenges, primarily due to the limited availability of diverse, quality brain tumor images. This issue stems largely from data privacy and ethical constraints. For instance, in general, computer vision, models like CLIP [10] benefit from extensive datasets of around 400 million internet-sourced images for generative pre-training. In contrast, assembling a similarly large and varied dataset of unlabeled medical brain tumor images is far less feasible. Additionally, the diversity shortage in medical imaging modalities, especially in 3D Magnetic Resonance Imaging (MRI), poses further challenges [11], [12]. The inherent lack of cross-plane contextual information in 3D MRI adds complexity to transfer learning processes. In addition, the evolution of task-specific objectives in computing has led to capability enhancements but also introduced cross-domain inconsistencies [13]–[15]. In the pretraining phase of general computer vision, the focus is on high-level semantic representations for accurate classification within deep hierarchical architectures. However, medical segmentation tasks require detailed pixel-level predictions, particularly in the context of 3D high-resolution medical images [16]–[18]. This shift in focus from high-level to pixel-level precision presents a challenge, as traditional brain tumor segmentation frameworks are more adept at learning high-level representations. An overemphasis on high-level learning can be counterproductive for tasks demanding pixel-level accuracy, highlighting the necessity for a balanced and adaptive approach [19], [20].

To effectively address the challenges associated with transferring advancements from general computer vision to medical image segmentation—particularly in the context of high-quality labeled data—it is crucial to conduct a comprehensive analysis of the architectural frameworks currently employed in this domain. These frameworks typically comprise two essential components in supervised learning: a backbone responsible for efficiently extracting high-level semantic representation and a segmentation head designed to achieve precise pixel-level classification. This dual-structure approach is illustrated in Fig. 1B and highlights the intricate interplay between feature abstraction and detailed segmentation. Given the intricate nature of medical imaging, a critical reassessment of the training framework is necessary, focusing on its ability to optimize representation extraction and achieve rapid convergence within the specific domain of 3D brain tumor segmentation. This necessitates a thorough understanding of how high-level features can be effectively translated into the pixel-level accuracy required for medical tasks. Moreover, considering the unique characteristics of 3D imaging modalities, such as the lack of cross-plane contextual information, it becomes evident that the architecture must be meticulously calibrated to overcome these specific challenges. Therefore, the development of a comprehensive and synergistic framework to current medical image segmentation models is essential, ensuring that they are adeptly equipped to navigate the unique complexities and idiosyncrasies inherent in medical data. Such a framework will enhance the models' capability to provide accurate and clinically relevant segmentation outcomes.

Given the challenges mentioned above, we introduce an effective training framework. This framework seamlessly embeds a plug-and-play component, MOD, into existing supervised models, enhancing their performance in data-limited situations. As illustrated in Fig. 1C, the MOD includes two integral modules: an Online Tokenizer and a lightweight Dense Predictor. The Online Tokenizer, powered by a self-distillation mechanism, facilitates the efficient learning of high-level semantics. This is particularly useful in scenarios with limited availability of labeled medical images, as it bolsters the backbone's ability to incorporate intricate feature representations during the training phase [21], [22]. Concurrently, the Dense Predictor employs a self-modeling strategy, scrutinizing regions at a lower semantic level. This strategy allows an evaluation of the backbone's ability to extract detailed information from the data [23]–[25]. Through the harmonious interplay of the Online Tokenizer and Dense Predictor at diverse semantic levels, the MOD cultivates an efficient training procedure, enhancing the effectiveness of current supervised learning backbones, particularly on small yet high-quality labeled training datasets available in the public domain. At the inference stage, the plug-and-play component, MOD, is detached, ensuring that the original model incurs no extra computational overhead. In order to showcase the efficacy of our proposed training framework, we conducted a series of experiments integrating MOD with several competitive 3D medical segmentation baselines featuring diverse structures. These experiments were performed on two publicly available brain tumor datasets both consisting of limited labeled images. The findings indicate that our training framework significantly boosts the performance of current supervised learning models, highlighting its utility as a beneficial addition to established approaches in brain tumor segmentation.

The main contributions are summarized as follows:

- The paper presents an innovative training framework incorporating the modular component MOD, designed to enhance the performance of existing supervised learning algorithms for segmenting 3D brain tumor images under conditions of limited training data. During the inference phase, the plug-and-play MOD component is excluded, preserving the computational efficiency of the original model without incurring extra processing costs.
- The MOD component integrates two key elements: the Online Tokenizer and the Dense Predictor. The Online

Tokenizer utilizes a self-distillation mechanism to streamline the learning of high-level semantic features. Concurrently, the Dense Predictor, through its self-modeling approach, focuses on capturing detailed aspects. This combination effectively bolsters swift convergence and efficient representation learning of current 3D medical segmentation models.

- In order to validate the effectiveness of the proposed framework, we conduct a series of experiments using various competitive 3D medical segmentation baselines with and without MOD. The results demonstrate that when integrated with MOD, these models achieve a substantial improvement in segmentation performance, as evidenced by higher Dice coefficients and HD95 on two public limited supervised labeled datasets: BraTS 2021 [26] and MSD 2019 Task-01 Brain Tumor [27].

II. Related Work

A. Two-Stage Segmentation Tasks

In computer vision, the two-stage process leveraging pre-trained models for downstream tasks has shown great success [22]. For example, CP^2, initially trained on the 1.28 million images of the ImageNet 1K dataset, achieved state-of-the-art dice accuracy when fine-tuned for tasks in PASCAL VOC 2012 [28]. The Swin Transformer, pre-trained on the 14 million images of ImageNet-22K, outperformed SETR by +3.2 mIoU when fine-tuned on the ADE20K dataset [29]. Zero-shot transfer learning, a method to save computational resources, uses pre-trained models on novel datasets without further fine-tuning [22], like the Clip simple pre-training task applying image representations for segmentation based on textual class descriptions. Efforts to increase labeled images led to the creation of Segment Anything (SA) [30], the largest segmentation dataset with over 1 billion masks across 11 million images. However, in medical imaging, assembling large high-quality datasets remains challenging due to data privacy, cost, and ethical issues. Even extensive datasets like MedSAM [31], with around 0.2 million masks across 11 modalities, are significantly smaller than natural image datasets like SA. Thus, while effective in general computer vision, applying this two-stage process in medical image analysis is limited by the scarcity of annotated datasets.

B. Supervised Brain Tumor MRIs Segmentation

Semantic segmentation is an essential component of medical image analysis, as it enables the accurate examination of anatomical structures [32]. since the introduction of the U-shaped encoder-decoder architecture with convolutional layers by U-Net [33], it has demonstrated remarkable performance in 3D brain tumor segmentation tasks [34], [35]. However, the convolutional layers often ignore global representations because of the limitations of receptive field size [3]. In order to alleviate this issue, TransBTS [36] uses a combination of the transformer, CNN, and encoder-decoder structure. Which encoder utilizes 3D CNN to extract the volumetric spatial feature maps. The ViTAutoEnc [37] integrates a transformer-based connection with convolutional layers as its Encoder. Meanwhile, the UNETR [3] utilizes a transformer-based encoder in conjunction with a convolution-based decoder. By expanding the number of parameters to develop larger-scale models, this kind of hybrid transformer-based algorithm exhibits significant progress in medical image analysis tasks. Nevertheless, the increased model size can lead to overfitting, high computational demands, and additional expenses, ultimately reducing the training efficiency for 3D medical images [20]. As an attempt to address this issue, Swin UNETR enhances efficiency to linear complexity by incorporating a Swin Transformer as the encoder, utilizing non-overlapping shifted windows for streamlined processing [38]. Edge U-Net [39] precisely localizes tumors by merging boundary-related MRI data with the main data from brain MRIs. Even though existing methods have brought about remarkable advancements in the field of medical image segmentation, they continue to grapple with a substantial challenge: improving the segmentation performance while dealing with limited supervised datasets, such as the BraTS and MSD Task-01 Brain Tumor. While these methods have demonstrated their capabilities, there remains notable room for further enhancement. Accordingly, the objective of our work is to construct a comprehensive framework that can be integrated with existing models, aiming to improve the overall accuracy of segmentation tasks. This goal is achieved by exploiting the meaningful information from the limited yet high-quality labeled training datasets available publicly, thereby addressing one of the key issues in the realm of medical image segmentation.

III. Methods

As illustrated in Fig. 1B, the existing medical image segmentation framework primarily hinges on two basic components: the backbone and the segmentation head [40], [41]. The backbone extracts pivotal features from brain images, upon which the segmentation head optimizes the loss function to delineate different medical image parts. Segmentation accuracy and effectiveness heavily depend on the quality of features extracted by the model's backbone and the segmentation head. Yet, the limited availability of brain tumor segmentation datasets, constrained by privacy and ethical considerations, presents a formidable challenge. Thus, it is vital to amplify the feature-extracting power of the model's backbone to overcome these limitations and improve segmentation outcomes [42].

A. Play-and-plug component: MOD

Addressing the prevalent challenges in brain tumor segmentation, we've introduced the play-and-plug component named MOD, which is illustrated in Fig. 2. This innovation is tailored to amplify the capability of current **M**edical image segmentation models, driving them towards heightened accuracy and efficiency. At its core, MOD is bifurcated into two pivotal components: the **O**nline tokenizer and the **D**ense predictor. The online tokenizer is meticulously designed to extract high-level semantic features, enriching the backbone's depth of understanding. Conversely, the dense predictor zeroes in on pixel-level details, ensuring granular accuracy. The synchronized

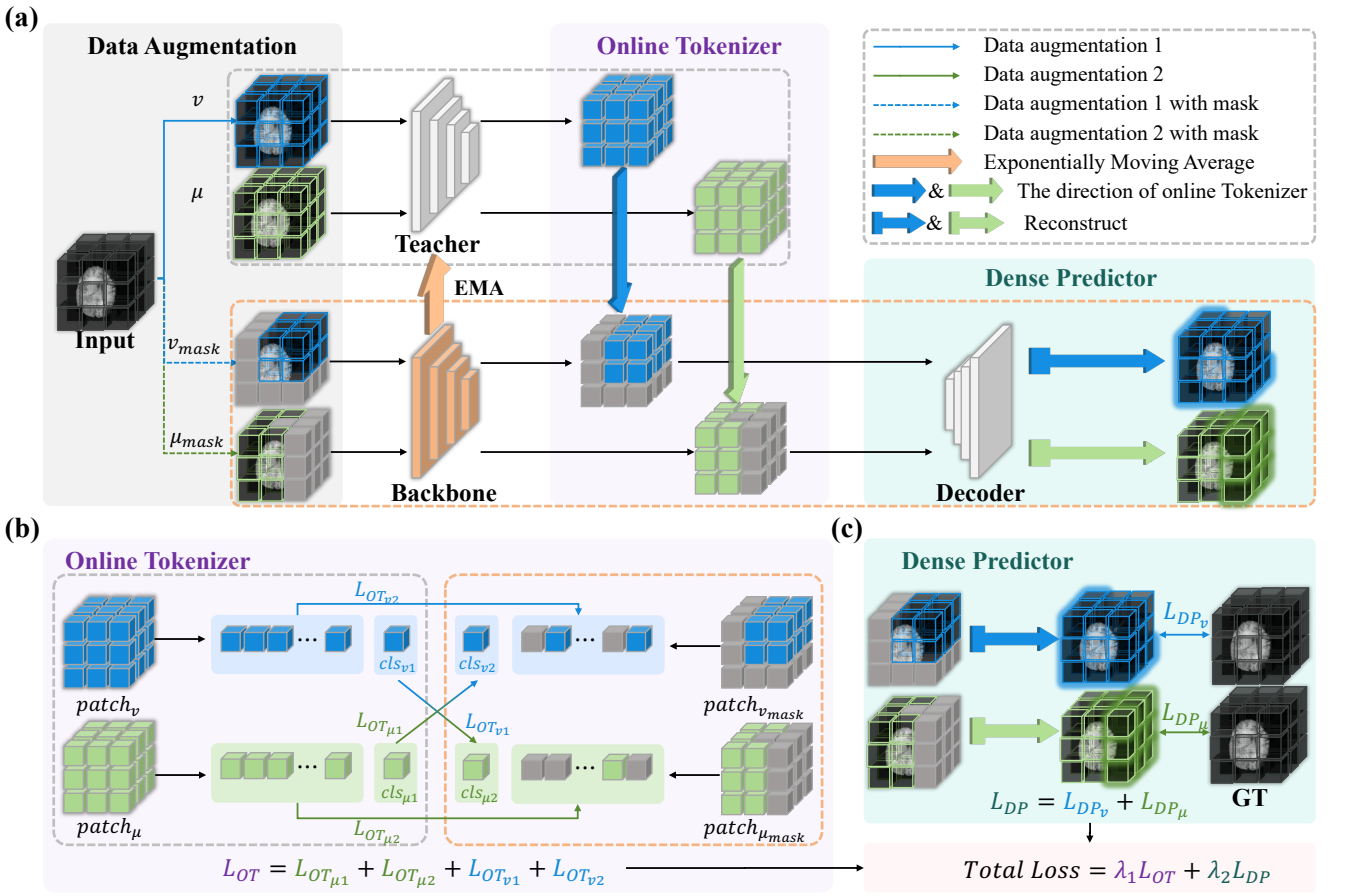

Fig. 2: The plug-and-play component, MOD, integrates an online tokenizer and a lightweight dense predictor, enhancing the performance of existing supervised learning algorithms across both semantic levels. The Online Tokenizer utilizes a self-distillation mechanism to streamline the learning of high-level semantic features. Concurrently, the Dense Predictor, through its self-modeling approach, focuses on capturing detailed aspects at a lower semantic level.

operation of the Online Tokenizer and Dense Predictor fortifies representation learning across various semantic strata of the backbone, culminating in an evident boost in the segmentation accuracy of the model.

1) Online Tokenizer: The online tokenizer is designed to enhance the backbone's capability in extracting high-level semantic features. Considering the self-distillation approach [22], the online tokenizer adopts a teacher-student model framework to advance the backbone's learning process. Specifically, a teacher backbone f' with parameters θ', structurally identical to the target backbone f with parameters θ, is initialized to guide the feature extraction evolution of the target backbone.

Advanced semantic feature extraction is facilitated through data augmentation techniques, including Gaussian blur and contrast enhancement, to produce augmented images μ and ν. These images are then subjected to blockwise masking, resulting in $\widehat{\mu}$ and $\widehat{\nu}$, which serve as masked inputs for the target backbone. The target backbone processes these inputs and outputs feature maps $f_{\theta}^{efm}(\widehat{\mu})$, whereas the teacher backbone processes the unmasked inputs to yield $f_{\theta'}^{efm}(\mu)$.

The training target is defined by a loss function $\mathcal{L}_{efm}$ in Equation 1, aiming to synchronize the learning of enhanced high-level semantics between the teacher and target backbones. This loss is calculated by averaging the cross-entropy loss between the teacher's outputs on unmasked images and the student's outputs on the corresponding masked variants. Consequently, the online tokenizer serves a dual function: it bolsters the target backbone's capability for feature extraction while simultaneously refining the model's performance through enhanced learning of high-level semantic information.

Moreover, to ensure semantic richness in feature capture as indicated by [43], the online tokenizer extracts depth feature maps from the student backbone f_{θ} and the teacher backbone $f_{\theta'}$, to produce classification tokens $f_{\theta}^{cls}(\widehat{\mu})$, $f_{\theta}^{cls}(\widehat{\nu})$, $f_{\theta'}^{cls}(\mu)$, and $f_{\theta'}^{cls}(\nu)$. These tokens are subjected to further distillation. Taking $f_{\theta}^{cls}(\widehat{\nu})$ and $f_{\theta'}^{cls}(\mu)$ as examples, the training target is formalized in Equation 2. Additionally, we compute the mean loss between $f_{\theta'}^{cls}(\mu)$ and $f_{\theta}^{efm}(\widehat{\nu})$ via another cross-entropy term, resulting in the loss $\mathcal{L}_{cls}$, which is designed to harmonize high-level semantics across images with the same ground truth. Thus, the total loss function of the online tokenizer, $\mathcal{L}_{ot}$, is finally calculated in Equation 3.

$$\mathcal{L}_{efm} = -\frac{1}{N}\sum_{i=1}^{N} m_i \cdot f_{\theta'}^{efm}(\mu_i)^T log f_{\theta}^{efm}(\widehat{\mu_i}) \quad (1)$$

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N} m_i \cdot f_{\theta'}^{cls}(\nu_i)^T log f_{\theta}^{cls}(\widehat{\mu}_i) \tag{2}$$

$$\mathcal{L}_{OT} = \alpha\mathcal{L}_{efm} + \beta\mathcal{L}_{cls} \tag{3}$$

where $m_i \epsilon \{0,1\}^N$ denotes the amount of mask for the ith sample and N denotes the total number of training data. In this case, we set $\alpha = 1.5, \beta = 0.5$.

2) Dense Predictor: The dense predictor is designed to refine the extraction of pixel-level semantic features within a model's backbone, a pivotal enhancement for the precise delineation required in 3D brain tumor segmentation. By leveraging the principles of masked image modeling, the dense predictor elevates the backbone's capability for pixel restoration, directly correlating with an uptick in segmentation fidelity [44]. This is achieved through a self-modeling mechanism that weaves the dense predictor into the existing segmentation framework, thereby sharpening the model's interpretative precision at the granular pixel level.

The dense predictor begins its process by integrating position embeddings with enhanced images μ and ν, which is a critical step towards achieving accurate localization of mask tokens. Following this, the images traverse through the backbone network, resulting in feature maps $f_{\theta}^{efm}(\widehat{\mu})$ and $f_{\theta}^{efm}(\widehat{\nu})$, which are then relayed to the decoder. The decoder's task is to accurately reconstruct the images into $recon(\widehat{\mu})$ and $recon(\widehat{\nu})$, where outputs are transformed into linear vectors that correspond to the positional encoding, enabling meticulous patch token localization within the reconstruction phase. From reconstruction, the dense predictor outputs the same size as the input image and optimizes the loss function with the original image.

In pursuit of achieving a reconstructed image that mirrors the ground truth, the dense predictor optimizes a masked L1-loss function [45] that compares the patch tokens of the reconstructed and ground truth images. The minimization of this loss, defined in the objective function $\mathcal{L}_{DP}$ in Equation 4, only calculates unmasked features, is pivotal for refining pixel-level feature extraction, consequently enhancing segmentation precision.

$$\mathcal{L}_{DP} = \frac{1}{N}\sum_{i=1}^{N} ||\mu_i - f_{\theta}^{efm}(\widehat{\mu_i}))||_1 * mask \tag{4}$$

Where N denotes the total number of patches in the training image, and $f_{\theta}^{efm}(\widehat{\mu_i}))$ represents the output feature maps from the student backbone, and μ_i corresponds to a single patch from the augmented image. The dense predictor loss is calculated only for the masked patches in the input image. The entire methodology encompassing MOD is concisely encapsulated in Algorithm 1.

Algorithm 1 The plug-and-play component MOD

Require: Backbone f_θ, dataset D
Ensure: Backbone f_θ

Initialization: Teacher Backbone $f_{\theta'}$
for $epoch$ in {1,...,E} **do**
  $x \leftarrow \{\mathrm{d}|\mathrm{d}\in D\}$ ▷ Load dataset
  $\mu \leftarrow A_\mu(x), \nu \leftarrow A_\nu(x)$ ▷ Data Augmentation
  $\mu\prime \leftarrow M(\mu), \nu\prime \leftarrow M(\nu)$ ▷ Mask
  Step 1. Compute total Loss for MOD:
    $\mathcal{L}_{OnlineTokenizer}(\theta) = \alpha\mathcal{L}_{efm}(\theta) + \beta\mathcal{L}_{cls}(\theta)$
    $\mathcal{L}_{DensePredictor}(\theta) = ||\mu - f_\theta^{efm}(\widehat{\mu}))||_1$
    $\mathcal{L}_{MOD}(\theta) = \lambda_1\mathcal{L}_{\mathrm{OT}}(\theta) + \lambda_2\mathcal{L}_{\mathrm{DP}}(\theta)$
  Step 2. Compute total Loss for Segmentation Head:
    $\mathcal{L}_{SEG}(\theta) = \lambda_1\mathcal{L}_{focal}(\theta) + \lambda_2\mathcal{L}_{dice}(\theta)$
  Step 3. Back-propagation and update gradient:
  Update parameter θ
    $\theta' \leftarrow \theta + lr \cdot \nabla(\mathcal{L}_{MOD}(\theta) + \mathcal{L}_{SEG}(\theta))$
  Update Teacher Network by EMA
end for
Return Prime Backbone $f_{\theta'}$

B. Loss Functions

The proposed innovative training framework incorporates a plug-and-play component named MOD into established supervised models. Consequently, the overall loss function within this system is bifurcated into two distinct components: the MOD-specific loss and the conventional loss associated with the original segmentation task.

MOD-specific Loss: As shown in Equation 5, the MOD-specific Loss is constituted by two specialized loss functions: the self-distillation loss, which stems from the Online Tokenizer's high-level semantic analysis, and the self-modeling loss, which arises from the Dense Predictor's detailed insight extraction process.

$$\mathcal{L}_{\mathrm{MOD}} = \lambda_1\mathcal{L}_{\mathrm{OT}} + \lambda_2\mathcal{L}_{\mathrm{DP}} \tag{5}$$

In this paper, we set $\lambda_1 = 1, \lambda_2 = 1$.

Segmentation Optimization: In order to achieve the end-to-end segmentation task, the loss function of original models still comprises the minimization of both Dice loss [46] and Focal loss [47], formalized in Equation 6 and Equation 7 respectively.

$$\mathcal{L}_{dice}(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{M}\frac{\sum_{i=1}^{N} G_{i,j}Y_{i,j}}{\sum_{i=1}^{N} G_{i,j} + \sum_{i=1}^{N} Y_{i,j}} \tag{6}$$

s

$$\begin{aligned}\mathcal{L}_{focal}(G,Y) = -\sum_{j=1}^{M}\sum_{i=1}^{N}(Y_{i,j}G_{i,j}^2 log(1-G_{i,j}) \\ +(1-Y_{i,j})(1-G_{i,j})^2 log(G_{i,j}))\end{aligned} \tag{7}$$

where I denotes the number of voxels and J denotes the number of classes; $Y_{i,j}$ and $G_{i,j}$ denote the probability of output and one-hot encoded ground truth for class$j$ at voxel i, respectively.

The segmentation loss function is calculated as Equation 8.

$$\mathcal{L}_{SEG} = \lambda_1\mathcal{L}_{focal} + \lambda_2\mathcal{L}_{dice} \tag{8}$$

where the optimal values of $\lambda_1 = \lambda_2 = 1$ were calculated using a grid-search hyper-parameter optimization.

Overall Loss of the proposed training framework is finally calculated in Equation 9.

$$\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{MOD} + \lambda_2 \mathcal{L}_{SEG} \tag{9}$$

where $\lambda_1 = \lambda_2 = 1$.

In addition, after every iteration involving loss in the backbone, we utilize Exponential Moving Average (EMA) techniques [43] to judiciously transfer a subset of the parameters from the backbone f_θ to the teacher backbone $f_{\theta'}$ within the Online Tokenizer, ensuring the teacher backbone provides credible guidance.

IV. Experiments and Discussion

Our experimental framework involves an extensive comparison of several leading 3D brain tumor baselines, both with the integration of the MOD component during training and without it. The goal is to evaluate the training efficiency and the segmentation results by excluding MOD during the inference stage. This comparison is designed to showcase the enhanced performance that our proposed general training framework contributes to the domain.

A. Datasets

We utilized two datasets with a limited number of high-quality labeled images: the BraTS 2021 [26] and the MSD 2019 Task-01 Brain Tumor [27] datasets. The BraTS 2021 dataset includes 1,251 data samples, whereas the MSD 2019 dataset offers 484 samples of multimodal MRI. The assessment of our MOD's performance on these smaller datasets is critical to determine its effectiveness in supervised learning scenarios with constrained data availability.

The **BraTS 2021 Brain Tumor** dataset encompasses $1,251$ multi-institutional magnetic resonance imaging (MRI) scans, each detailing a volumetric snapshot with dimensions of $240 \times 240 \times 155$. These scans are instrumental in identifying three pivotal tumor sub-regions: the enhancing tumor (ET), the peritumoral edema (ED), and the necrotic and non-enhancing tumor core (NCR/NET). Ground truth annotations within this dataset are meticulously segmented into three hierarchical categories: the Whole Tumor (WT), the Tumor Core (TC), and the Enhancing Tumor (ET). The BraTS 2021's inherent challenge lies in the inherent diversity of its data—stemming from various MRI scanners and institutions—compounded by the intricate complexity of the tumor sub-regions delineated within it.

The **MSD 2019 Task-01 Brain Tumor** dataset, a cornerstone of the Medical Segmentation Decathlon 2019, consists of 484 data samples with uniform voxel dimensions of $1.0 \times 1.0 \times 1.0\ mm^3$. Each sample comprises four modalities of 3D MRI scans, akin to those found in the BraTS 2021 dataset, and includes ground truth labels delineating glioma segmentation into the necrotic/active tumor and edema. Given the modest number of training samples, the ability of models to learn representations is expectedly diminished when compared to those trained on more extensive datasets. As such, evaluating model performance for brain tumor segmentation within the MSD 2019 dataset serves as a litmus test for the efficacy of representation learning, offering a direct comparison against a range of different methodologies.

B. Experimental Settings

Our proposed training framework, constructed on PyTorch [48] and MONAI [49], leverages the computational prowess of 4 NVIDIA V100 GPUs, with each processing a batch of 4 patches sized at $128 \times 128 \times 128$. Optimization is facilitated through AdamW [50], paired with a cosine scheduler for a warm-up and an initial learning rate set at 0.001, further fine-tuned with a weight decay of 0.05. We employed a cosine scheduler with a momentum value of 0.996 for the EMA update teacher model. And used a random masking method with a mask rate of 0.6 during the training phase. Our data augmentation strategy includes intensity normalization and random intensity adjustments, with a scale factor oscillating by ± 0.1 and shifts maintained at a 0 ± 0.1 range.

C. Evaluation Metrics

Our framework's efficacy in medical image segmentation is gauged using two principal metrics: Dice accuracy [46], which assesses the overlap between the predicted and ground truth segments (computed as per Equation 10), and HD95 [51], which measures the 95th percentile Hausdorff distance to evaluate shape similarities, detailed in Equation 11. While the former reflects segmentation precision, the latter offers insights into the contour accuracy, with lower HD95 values signifying a closer resemblance to the actual anatomy.

$$\mathrm{Dice}(G, Y) = \frac{2\sum_{i=1}^{I} G_i Y_i}{\sum_{i=1}^{I} G_i + \sum_{i=1}^{I} Y_i} \tag{10}$$

$$\begin{aligned}\mathrm{HD}(G', Y') = \max\{&\max \min \|g' - y'\|, \\ &\max \min \|y' - g'\|\}. \\ &(g' \in G', y' \in Y')\end{aligned} \tag{11}$$

where the notations G_i and Y_i represent the ground truth and predicted values for the i^{th} data sample, respectively. In the context of surface distances, G' and Y' correspond to the sets of surface points derived from the ground truth and the predicted segmentation.

D. Empirical Results

1) Segmentation Results on limited supervised datasets: In evaluating the effectiveness of the proposed training framework, we scrutinized nine distinct 3D brain tumor segmentation detection models utilized for medical image analysis on the limited supervised BraTS 2021 dataset, benchmarked by Dice and HD95 scores. For a more comprehensive understanding, we elected four representative models ViT, UNETR, UNETR++, and Swin-UNETR to demonstrate the performance alterations brought about by the incorporation of the MOD framework. As shown in Table I, ViT, which introduced long-range dependency in visual images, recorded the lowest performance among all evaluated models with a dice score of 61.80% and HD95 of 38.89 mm on the BraTS 2021 dataset. Conversely, UNETR, which uniquely

TABLE I: Evaluation on BraTS 2021 Brain Tumor with and without integration of MOD component. The Metrics of the Integration Framework are Highlighted in Gray. The increase and loss of metrics are recorded in **red** and **green** in this table. Moreover, the performance metrics of the baseline setting of MOD are highlighted in gray.

Baseline	Dice(%)↑				HD95(mm)↓			
	TC	WT	ET	Avg.	TC	WT	ET	Avg.
VIT	62.64	69.63	53.13	61.80	43.59	35.66	37.42	38.89
VIT(MOD)	66.30 +3.66	73.46 +3.83	58.65 +5.52	**66.14** +4.34	33.85 -9.47	33.31 -2.35	12.55 -24.87	**26.57** -12.32
UNETR	90.78	92.53	87.59	90.05	5.19	8.97	4.22	6.13
UNETR(MOD)	92.84+2.06	93.63 +1.10	89.56 +1.97	**92.01** +1.96	3.60 -1.59	4.20 -4.77	2.90 -1.32	**3.57** -2.56
UENTR++	91.49	92.87	88.82	91.06	3.81	6.12	3.45	4.42
UENTR++(MOD)	91.54 +0.05	92.87 +0.00	89.08 +0.26	**91.16** +0.10	3.94 +0.13	5.81 -0.31	3.01 -0.44	**4.25** -0.17
Swin-UNETR	91.69	93.22	89.08	91.36	5.81	4.64	4.65	5.03
Swin-UNETR(MOD)	93.15 +1.46	93.32 +0.10	89.85 +0.77	**92.11** +0.75	4.54 -1.27	3.83 -0.81	3.80 -0.85	**4.06** -0.97
Swin-UNETRV2	92.48	93.49	89.30	91.76	5.35	4.12	4.20	4.55
Swin-UNETRV2(MOD)	93.64 +1.16	93.52 +0.03	90.05 +0.75	**92.40** +0.64	4.11 -1.24	3.67 -0.45	3.53 -0.67	**3.77** -0.78
SegMamba	92.10	93.81	89.49	91.80	6.03	5.76	5.10	5.63
SegMamba(MOD)	93.28 +1.18	93.90 +0.09	90.70 +1.21	**92.62** +0.82	5.28 -0.75	4.90 -0.86	4.31 -0.79	**4.83** -0.80

TABLE II: Evaluation on MSD 2019 Task-01 Brain Tumor with and without integration of MOD component. The colors of the record in this table are labeled the same as in Table I.

Baseline	Dice(%)↑				HD95(mm)↓			
	TC	WT	ET	Avg.	TC	WT	ET	Avg.
VIT	55.08	62.56	33.65	50.43	41.42	44.89	37.31	41.20
VIT(MOD)	61.36 +6.28	70.38 +7.82	36.86 +3.21	**56.20** +5.77	40.74 -0.68	43.03 -1.86	36.88 -0.43	**40.22** -0.98
UNETR	79.87	87.18	55.93	74.33	17.34	19.74	13.74	16.94
UNETR(MOD)	83.44 +3.57	90.87 +3.69	60.82 +4.89	**78.37** +4.04	7.73 -9.61	15.37 -4.37	9.44 -4.30	**10.84** -6.10
UENTR++	83.65	87.19	62.38	77.74	8.00	6.99	8.31	7.76
UENTR++(MOD)	83.66 +0.01	90.05 +2.86	62.40 +0.02	**78.85** +1.11	7.01 -0.99	5.91 -1.08	8.43 -0.12	**7.11** -0.65
Swin-UNETR	81.76	92.54	62.16	78.82	6.34	11.90	7.95	8.73
Swin-UNETR(MOD)	84.22 +2.46	93.23 +0.69	63.02 +0.86	**80.15** +1.33	5.82 -0.52	6.36 -5.54	7.11 -0.84	**6.43** -2.30
Swin-UNETRV2	84.59	91.17	63.63	79.80	6.22	10.51	7.12	7.95
Swin-UNETRV2(MOD)	85.04 +0.45	92.92 +1.75	63.71 +0.08	**80.55** +0.75	5.38 -0.84	5.94 -4.57	6.59 -0.53	**5.97** -1.98
SegMamba	84.81	90.50	64.22	79.84	8.56	9.39	8.10	8.68
SegMamba(MOD)	84.99 +0.18	91.23 +0.73	64.86 +0.64	**80.36** +0.52	8.30 -0.26	8.48 -0.91	7.22 -0.88	**8.00** -0.68

integrated global and local representation through its U-shape architecture, delivered high segmentation accuracy but was marred by high computational complexity. In response to this, UNETR++ introduces efficient paired attention and channel-wise discriminative features, resulting in an improved dice score of 91.06% and HD95 of 4.42 mm. However, Swin-UNETR harnessed the Swin Transformer as an encoder, thus transforming global representation into a linear computational complexity reduction, and consequently achieving the highest dataset accuracy with a dice score of 91.36% and HD95 of 5.03mm.

Results from our experiments strongly indicate that the Online Tokenizer of the MOD framework significantly augmented the ability of target models to extract advanced semantic features. In Table I, ViT, after MOD integration, achieves an average Dice score improvement of 4.34%, jumping to 66.14%. UNETR's Dice score increased from 90.05% to 92.01% after MOD incorporation. Notably, the implementation of the MOD framework in UNETR++ resulted in a Dice score increase from 91.06% to 91.16%. While Swin-UNETR, already achieving a Dice score of 91.36%, still improved by 0.75% upon the inclusion of MOD. Furthermore, owing to the Dense Predictor component enhancing the pixel-level semantic feature extraction capability of the trained models, the application of MOD led to marked improvements in segmentation edge accuracy (HD95) across the three task models. ViT, UNETR, UNETR++, and Swin-UNETR witnessed HD95 reductions of 12.32mm, 0.67mm, 0.17mm, and 0.97mm, respectively, thereby emphasizing the potency of the MOD framework in enhancing the performance of hybrid transformer-based models in medical image segmentation tasks.

2) Segmentation Results on smaller supervised datasets: In settings with limited training datasets, the proficiency of a model in feature extraction is a critical determinant of its effectiveness. We executed a sequence of evaluations to gauge the power of pre-trained models from the MSD 2019 Task-01 Brain Tumor dataset, particularly when different frameworks are incorporated during training. Notably, this dataset is markedly less extensive than the BraTS 2021 dataset, containing merely 38.69% of the latter's volume of data.

Table II showcases the performance boost for the ViT model upon the incorporation of the MOD component during training. In the inference phase, there was a notable increase in

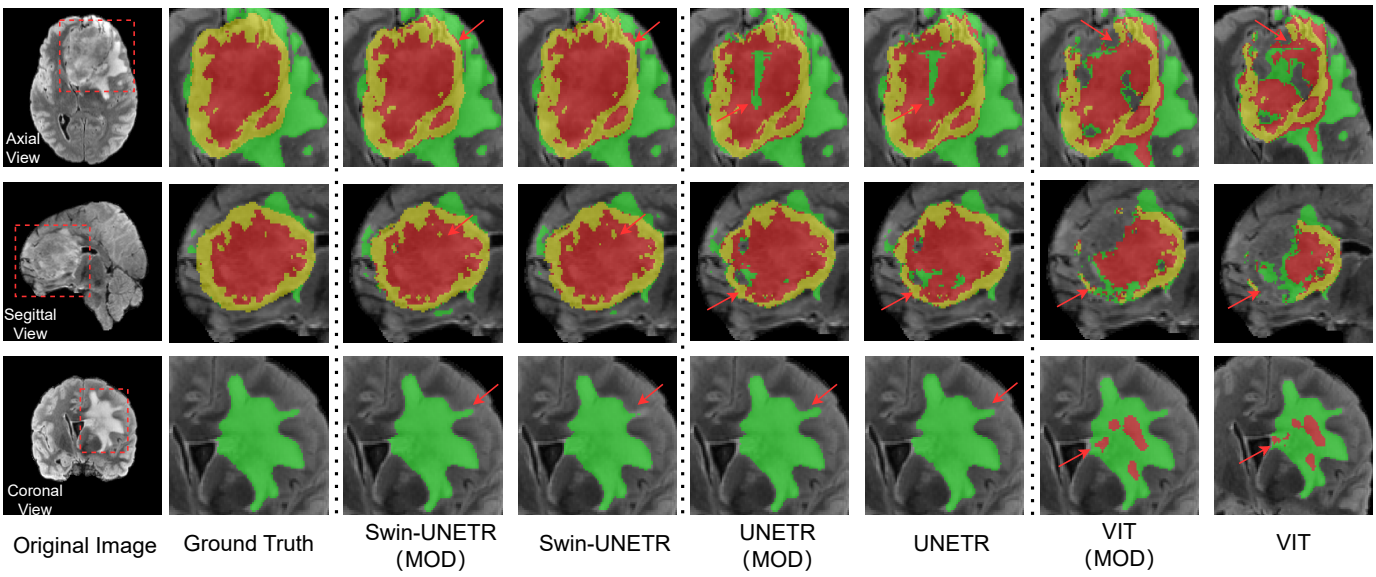

Fig. 3: Visual comparisons of model performance, highlighting the differences when MOD components are included versus excluded, offer qualitative insights into their impact on the MSD 2019 Task-01 dataset. The segmentation of necrotic tumor core (NT, highlighted in green), peritumoral edema (ED, highlighted in yellow), and enhancing tumor (ET, highlighted in red).

Dice scores, achieving 55.08% for Tumor Core (TC), 62.56% for Whole Tumor (WT), and 33.65% for Enhancing Tumor (ET). These scores signify improvements of 6.28%, 7.82%, and 3.21% compared to those obtained prior to the MOD incorporation. Furthermore, UNETR, UNETR++ and Swin-UNETR also exhibited improved Dice accuracy, with average Dice increases of 4.04%, 1.11% and 1.33% respectively. In terms of the HD95 metric, the integration of the MOD framework yielded reductions across the three experimental task models. Notably, ViT showed a slight average HD95 decrease from 41.20mm to 40.22mm. UNETR, on the other hand, showcased a remarkable reduction in HD95 for all three regions, with respective decreases of 9.61mm, 4.37mm, and 4.3mm, culminating in an overall average HD95 reduction of 6.10mm. As for UNETR++, it recorded an improvement in HD95 from 7.76 mm to 7.11 mm. Similarly, Swin-UNETR displayed a notable average HD95 decrease from 8.73mm to 6.43mm, a reduction of approximately 26%. These compelling results indicate that the MOD framework is capable of significantly enhancing the performance of baseline models, particularly in resource-constrained environments typified by limited data availability.

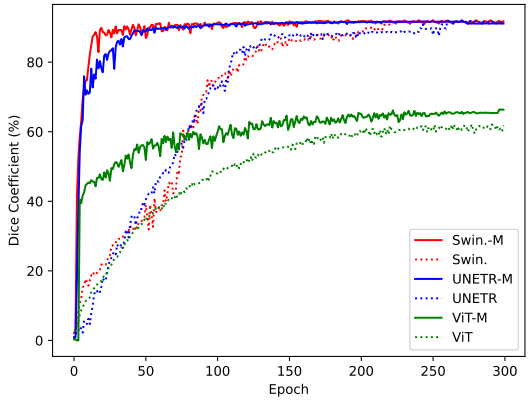

Fig. 4: This fig illustrates the convergence trends of existing segmentation models throughout training, both with and without the inclusion of our proposed training framework, using the BraTS 2021 Dataset. The solid line represents the progression of Dice scores over epochs within our framework, while the dashed line reflects the performance trends of the existing models.

3) Convergence Speed in the Training Process: UNETR, and Swin-UNETR demonstrated a notable increase in accuracy within a reduced number of training epochs. Fig. 4 compares the Dice coefficients of these models on the BraTS 2021 dataset both with and without the proposed MOD-enhanced training framework. The results showed that with MOD, models converged faster. For example, the Swin-UNETR model, which normally required 120 epochs to achieve a Dice score of 80%, reached this milestone in just 10 epochs after incorporating MOD. This marked improvement in training efficiency is attributed to the MOD framework's facilitation of faster and more effective extraction of high-level and pixel-level semantic features. This rapid convergence demonstrates the strength of the MOD-enhanced training framework in promoting a streamlined and potent training regime, which is crucial for complex tasks like high-resolution medical image segmentation.

4) Qualitative visualization: Fig. 3 presents the comparative segmentation outcomes for the primary tumor region across different planes—axial, sagittal, and coronal—utilizing baseline models with and without the incorporation of the MOD component, relative to the ground truth delineation. The various tumor regions, including the necrotic core, peritumoral edema, and the enhancing tumor, are differentiated by a color-

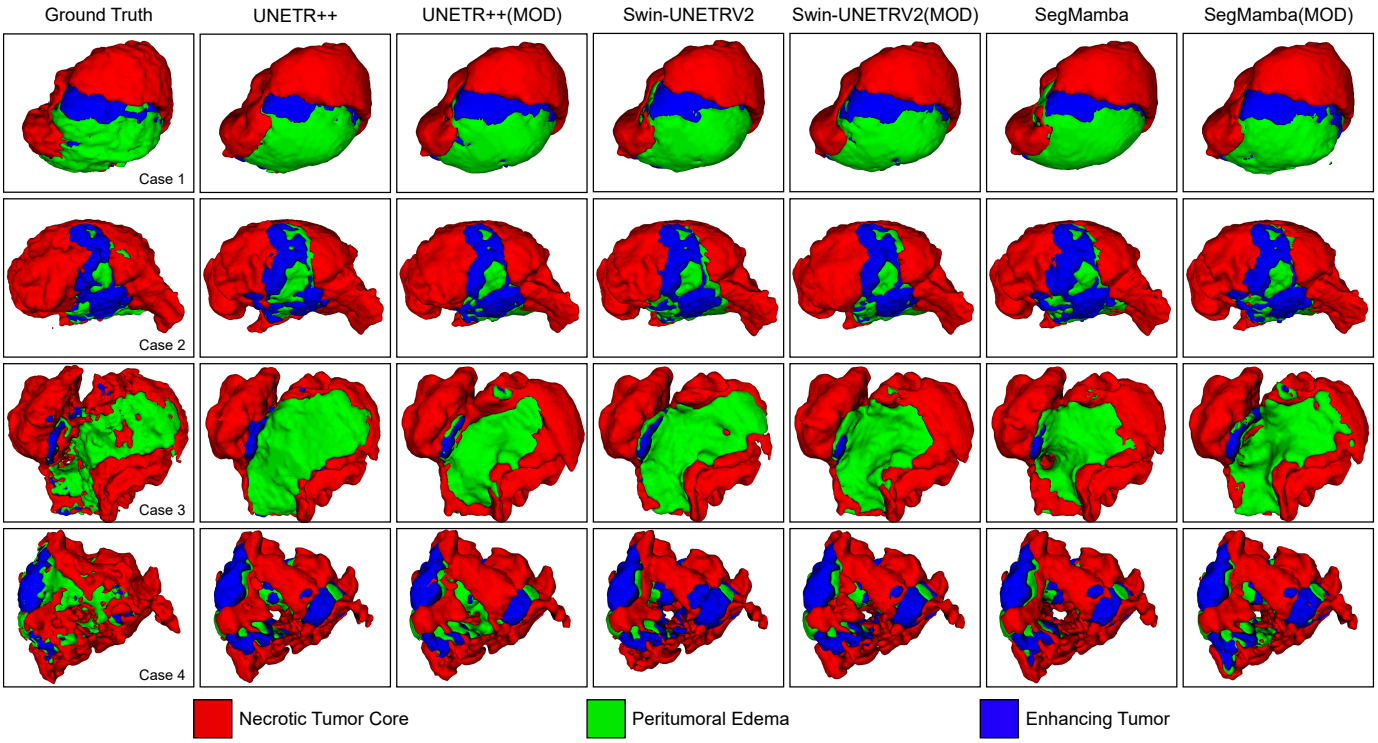

Fig. 5: Qualitative visualizations on 3D BraTS 2021 Brain Tumor with and without integration of MOD component. The inclusion of MOD specifically showcases its ability to delineate complex cardiac structures such as the necrotic tumor core (NT, highlighted in red), peritumoral edema (ED, highlighted in green), and enhancing tumor (ET, highlighted in blue).

coded scheme. The ground truth segmentation is depicted in the first column, while the subsequent columns demonstrate the segmentation achieved by the Swin-UNETR and UNETR baseline models, both prior to and following the application of the MOD component during the training phase. The integration of MOD with the baseline models results in markedly improved delineation of tumor boundaries and heightened accuracy in identifying intersecting regions, highlighting the enhanced performance of our proposed training framework in facilitating superior 3D segmentation for brain tumor tasks.

Fig. 5 provides a 3D visual comparison of brain tumor segmentation performance, comparing results with and without the inclusion of MOD components, using the BraTS 2021 dataset. The tumor regions, including the necrotic core, peritumoral edema, and enhancing tumor, are distinguished using a color-coded scheme. The first column shows the ground truth segmentation, while the subsequent columns display segmentation results from the UNETR++, Swin-UNETRV2, and SegMamba models, both before and after incorporating the MOD component during the training phase. The integration of MOD into the baseline models significantly improves the delineation of tumor boundaries and increases accuracy in identifying overlapping regions. This demonstrates that our proposed training framework effectively extracts both high-level and pixel-level features, leading to superior 3D segmentation for brain tumor tasks.

TABLE III: Evaluation outcomes of linear probing for different training frameworks on the BraTS 2021 dataset, utilizing UNETR as the baseline model. The colors of the record in this table are labeled the same as in Table I

Dice(%) ↑	Avg Acc	TC	WT	ET
Baseline	74.1	67.9	85.2	69.1
Baseline+Online Tokenizer	87.90+13.8	86.3+18.4	90.3+5.1	87.1+18.0
Baseline+Dense Predictor	89.2 +15.1	89.0 +21.1	90.8 +5.6	88.0 +18.9
Baseline+MOD	89.9 +15.8	90.0 +22.1	91.3 +6.1	88.2 +19.1

HD95(mm) ↓	Avg Score	TC	WT	ET
Baseline	34.0	33.7	34.3	34.1
Baseline+Online Tokenizer	13.2-20.8	13.6-20.1	16.2-18.1	9.8-24.3
Baseline+Dense Predictor	8.4 -25.66	7.4 -26.35	12.5 -21.8	5.2 -28.9
Baseline+MOD	6.0 -28.06	5.6 -28.15	8.8 -25.5	3.6 -30.5

E. Ablation Study

1) Evaluating the Effectiveness of MOD Architecture: Evaluating the effectiveness of the MOD architecture involved a meticulous analysis focusing on the architecture's two innovative modules: the Online Tokenizer and the Dense Predictor. The assessment employed the Linear Probe method to preserve the integrity of the UNETR backbone's parameters during the evaluation of downstream tasks. According to the data presented in Table III, the integration of the Dense Predictor enhanced the Dice score by a substantial 15.1% and concurrently achieved a reduction in the HD95 metric by 25.66 mm, showcasing its prowess in pixel-level image restoration. Moreover, the inclusion of the Online Tokenizer significantly improved the model's grasp of high-level semantics, achieving an impressive Dice accuracy of 89.9% and an HD95 of 6.0

mm.

Further examination through Fig. 6 indicated that adding the Online Tokenizer to the standalone backbone notably improved the Dice score from approximately 70% to 83% within a training span of 100 epochs. The full MOD framework registered the highest Dice scores, underscoring the incremental but critical enhancement brought by the Online Tokenizer, which refined the segmentation performance through advanced semantic feature extraction. In summation, the MOD architecture's components—Online Tokenizer and Dense Predictor—collectively strengthen the model's ability to interpret images more effectively, bolster learning efficiency, and lead to higher segmentation accuracy.

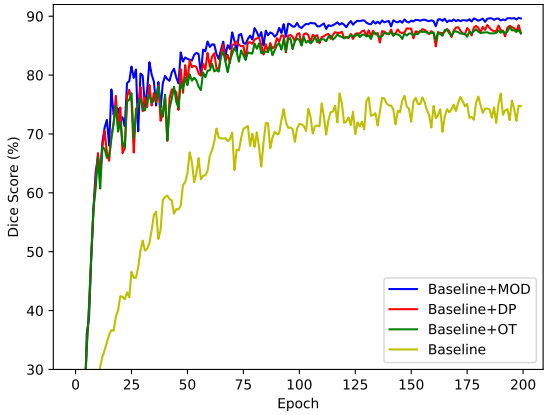

Fig. 6: Comparison of baseline model linear probing performance across various training frameworks using the BraTS 2021 dataset, with UNETR serving as the baseline model.

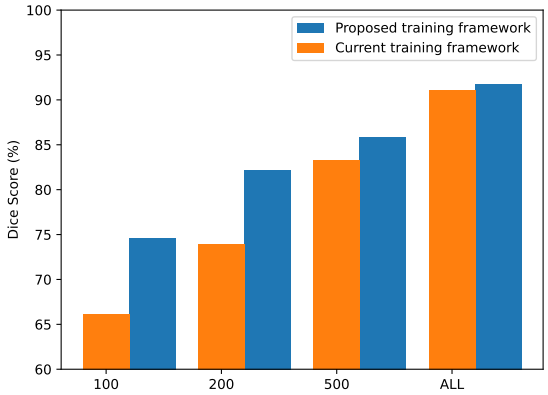

Fig. 7: Comparision of the standalone backbone (UNETR) and the backbone combined with MOD on fewer samples in the training process.

2) Evaluation of fewer samples in the training process: Fig. 7 compares the results of the standalone baseline (UNETR) and the baseline combined with MOD component across various dataset sizes for medical image segmentation. When utilizing a mere 100 images, the standalone baseline's Dice score reaches 66%, whereas the Dice of the baseline integrated with MOD achieves 75%, representing an enhancement of approximately 10% compared to the base model. When employing 500 images, the Dice of the standalone baseline attains 84%, while the baseline integrated with MOD achieves 87%. Upon using 200 images, the Dice of the baseline combined with MOD reaches 83%, signifying an improvement of roughly 10% compared to the standalone baseline with an equal number of training samples.

The modest 9% difference in Dice scores between the baseline integrated with MOD trained on 500 images and the one trained on 200 images demonstrates that our training framework excels in capturing both high-level and low-level semantic information during the learning process. This capability facilitates a more comprehensive understanding of medical images and enables the achievement of positive segmentation performance in scenarios with limited data.

V. Conclusion

In conclusion, this paper presents a training framework that adeptly integrates a plug-and-play component, MOD, into current supervised learning models, boosting their efficacy in scenarios with limited data. During the inference phase, the plug-and-play MOD component is excluded, preserving the computational efficiency of the original model without incurring extra processing costs. Our experiments demonstrated the superiority of models integrated with the MOD component over standalone models and the impact of pre-trained representation learning accuracy on subsequent segmentation tasks. The results highlight MOD's potential as a valuable plug-and-play component for high-quality 3D brain tumor segmentation, paving the way for advancements in medical image analysis.

References

[1] J. Wicaksana, Z. Yan, D. Zhang, X. Huang, H. Wu, X. Yang, and K.-T. Cheng, "Fedmix: Mixed supervised federated learning for medical image segmentation," *IEEE Transactions on Medical Imaging*, pp. 1955–1968, 2022.

[2] Z. Ning, S. Zhong, Q. Feng, W. Chen, and Y. Zhang, "Smu-net: Saliency-guided morphology-aware u-net for breast lesion segmentation in ultrasound image," *IEEE transactions on medical imaging*, vol. 41, no. 2, pp. 476–490, 2021.

[3] 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 *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[4] R. Zhang, L. Wang, Y. Qiao, P. Gao, and H. Li, "Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 769–21 780.

[5] Z. Tang, S. Ahmad, P.-T. Yap, and D. Shen, "Multi-atlas segmentation of mr tumor brain images using low-rank based image recovery," *IEEE transactions on medical imaging*, vol. 37, no. 10, pp. 2224–2235, 2018.

[6] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," *arXiv preprint arXiv:2210.10163*, 2022.

[7] C. Ma, G. Luo, and K. Wang, "Concatenated and connected random forests with multiscale patch driven active contour model for automated brain tumor segmentation of mr images," *IEEE transactions on medical imaging*, vol. 37, no. 8, pp. 1943–1954, 2018.

[8] J. Jiang, Y.-C. Hu, N. Tyagi, A. Rimner, N. Lee, J. O. Deasy, S. Berry, and H. Veeraraghavan, "Psigan: Joint probabilistic segmentation and image distribution matching for unpaired cross-modality adaptation-based mri segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 12, pp. 4071–4084, 2020.

[9] T. Lei, D. Zhang, X. Du, X. Wang, Y. Wan, and A. K. Nandi, "Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network," *IEEE Transactions on Medical Imaging*, 2022.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[11] Z.-X. Cui, S. Jia, J. Cheng, Q. Zhu, Y. Liu, K. Zhao, Z. Ke, W. Huang, H. Wang, Y. Zhu *et al.*, "Equilibrated zeroth-order unrolled deep network for parallel mr imaging," *IEEE Transactions on Medical Imaging*, 2023.

[12] M. Hu, J. Zhang, L. Matkovic, T. Liu, and X. Yang, "Reinforcement learning in medical image analysis: Concepts, applications, challenges, and future directions," *Journal of Applied Clinical Medical Physics*, vol. 24, no. 2, p. e13898, 2023.

[13] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[14] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 358–19 369.

[15] Z. Du, J. Li, H. Su, L. Zhu, and K. Lu, "Cross-domain gradient discrepancy minimization for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3937–3946.

[16] J. Du, X. Zhang, P. Liu, and T. Wang, "Coarse-refined consistency learning using pixel-level features for semi-supervised medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[17] S. Yang, L. Zhang, S. Liu, H. Lu, and H. Chen, "Real-time semantic segmentation via a densely aggregated bilateral network," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[18] Y. Pang, J. Liang, T. Huang, H. Chen, Y. Li, D. Li, L. Huang, and Q. Wang, "Slim unetr: Scale hybrid transformers to efficient 3d medical image segmentation under limited computational resources," *IEEE Transactions on Medical Imaging*, 2023.

[19] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," in *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*. IEEE, 2020, pp. 558–564.

[20] 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 *International MICCAI Brainlesion Workshop*. Springer, 2022, pp. 272–284.

[21] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.

[22] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "ibot: Image bert pre-training with online tokenizer," *International Conference on Learning Representations (ICLR)*, 2022.

[23] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[24] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," in *European Conference on Computer Vision*. Springer, 2022, pp. 348–367.

[25] A. Sheikhjafari, M. Noga, K. Punithakumar, and N. Ray, "Unsupervised deformable image registration with fully connected generative neural network," in *Medical imaging with deep learning*, 2022.

[26] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[27] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

[28] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, pp. 98–136, 2015.

[29] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, pp. 302–321, 2019.

[30] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[31] J. Ma and B. Wang, "Segment anything in medical images," *arXiv preprint arXiv:2304.12306*, 2023.

[32] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Drinet for medical image segmentation," *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2453–2462, 2018.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[34] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[35] 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 transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

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

[37] 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," *ICLR*, 2021.

[38] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[39] A. M. G. Allah, A. M. Sarhan, and N. M. Elshennawy, "Edge u-net: Brain tumor segmentation using mri based on deep u-net model with boundary information," *Expert Systems with Applications*, vol. 213, p. 118833, 2023.

[40] H. Messaoudi, A. Belaid, D. B. Salem, and P.-H. Conze, "Cross-dimensional transfer learning in medical image segmentation with deep learning," *Medical Image Analysis*, p. 102868, 2023.

[41] S. He, Y. Feng, P. E. Grant, and Y. Ou, "Segmentation ability map: Interpret deep features for medical image segmentation," *Medical image analysis*, vol. 84, p. 102726, 2023.

[42] S. Hansen, S. Gautam, S. A. Salahuddin, M. Kampffmeyer, and R. Jenssen, "Adnet++: A few-shot learning framework for multi-class medical image volume segmentation with uncertainty-guided feature refinement," *Medical Image Analysis*, p. 102870, 2023.

[43] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.

[44] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, "Self pre-training with masked autoencoders for medical image analysis," *arXiv preprint arXiv:2203.05573*, 2022.

[45] P. Bloomfield and W. L. Steiger, *Least absolute deviations: theory, applications, and algorithms*. Springer, 1983.

[46] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[49] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang *et al.*, "Monai: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701*, 2022.

[50] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[51] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

Yan Pang, received the Ph.D. degree from the University of Colorado Denver, USA in 2021. His research interests span machine learning, computer vision, and efficient deep learning. From April 2021 to May 2022, Dr. Pang was a machine learning scientist at Moffett AI, Los Altos, CA. From Aug. 2018 to May 2021, he was an instructor in the Department of Electrical Engineering at the University of Colorado Denver, and the Department of Electrical Engineering Technology at Metropolitan State University of Denver, USA. He is currently an Associate Professor at the Institute of Artificial Intelligence, Guangzhou University, China.

Yunhao Li, received his B.S. degree from Guilin University of Electronic Technology, Guilin, China, in 2022. He is currently pursuing his Master's degree at Guangzhou University, Guangzhou, China. His research areas of interest span computer vision and medical image analysis. As a Master's student, Yunhao Li focuses on advancing the field through innovative research and applications, leveraging his background in electronic technology and a keen interest in the medical domain.

Teng Huang, received his PhD degree in computer science in 2019 from Beihang University. Currently, he is working at the Institute of Artificial Intelligence at Guangzhou University. His research interests include Computer Vision, Medical Image Analysis, and Blockchain.

Jiaming Liang, received his B.S. degree from Foshan University in 2021. He is currently pursuing his Master's degree at Guangzhou University, Guangzhou, China. His research areas of interest span computer vision and medical image analysis. As a Master's student, He focuses on advancing the field through innovative research and applications, leveraging his background in electronic technology and a keen interest in the medical domain.

Zhen Wang, received the Ph.D. degree in computer science from the University of Wyoming, Laramie, WY, USA, in 2020. He is a researcher at the Zhejiang Lab, Hangzhou, China. He specializes in data analytics and machine intelligence, with research contributions featured in prominent academic conferences and esteemed journals, including ICCV, IJCAI, AAAI, ACM MM, CIKM, ICDM, ECML PKDD, and IEEE TNNLS.

Changyu Dong, received the Ph.D. degree from Imperial College London, UK. He is currently a professor with the Institute of Artificial Intelligence, at Guangzhou University. He has authored over 70 publications in international journals and conferences. His research interests include applied cryptography, trust management, data privacy, and security policies. His recent work focuses mostly on designing practical secure computation protocols. The application domains include secure cloud computing and privacy-preserving data mining.

Dongyang Kuang, received his B.S. degree in Applied Mathematics from the University of Science and Technology of China, Hefei, China, in 2011, and his Ph.D. degree in Applied Mathematics from the University of Wyoming, Laramie, WY, USA, in 2016, under the supervision of Professor Long Lee. He is currently an Assistant Professor at the School of Mathematics (Zhuhai), Sun Yat-sen University, Chin. His research interests include data analysis and mining, the development, application, and mathematical formalization of artificial intelligence models in cross-disciplinary sciences, and their interpretability.

Ying Hu, received the Ph.D. degree in mechanical engineering from the Harbin Institute of Technology, Harbin, China, in 2007. She is currently a Professor at the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, where she is the Executive Director of the Research Center of Cognitive and Human–Computer Interaction Technologies. She has published more than 140 scientific papers in refereed journals, book chapters, and conference proceedings. Her research interests include robot learning, medical assistant robots, and mobile robots. Prof. Hu has served as an Editor for the International Journal of Mechanisms and Robotic Systems and CAAI Transactions on Intelligence Technology.

Hao Chen, is an Assistant Professor at the Department of CSE&CBE, and Division of Life Science, The Hong Kong University of Science and Technology. He leads the Smart Lab focusing on trustworthy AI for healthcare. He received the Ph.D. degree from The Chinese University of Hong Kong (CUHK) in 2017. He has 200+ publications in MICCAI, IEEE-TMI, TNNLS, MIA, CVPR, ICCV, AAAI, Nature Communications, Lancet Digital Health, Nature Machine Intelligence, etc. He received several premium awards such as Asian Young Scientist Fellowship, MICCAI Young Scientist Impact Award, and several best paper awards. He serves as the Associate Editor of multiple journals including IEEE TMI, TNNLS, JBHI, CMIG, etc. He serves as the Program Committee of multiple international conferences including Area Chair of ICLR 2025, MICCAI 2021-2023, CVPR 2024-2025, ACM MM 2024, etc. He also led the team winning 15+ medical grand challenges.

Tim Lei, received the B.S. and M.S. degrees in electrical engineering from National Chiao Tung University, Taiwan, in 1995 and 1997, respectively, and the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, Michigan, USA, in 2003. He is currently an Associate Professor with the Department of Electrical Engineering at the University of Colorado Denver, Denver, CO, USA, holding secondary appointments in Bioengineering, Physiology Biophysics, Renal Medicine, and Ophthalmology. His research interests encompass biomedical photonics, optical engineering, and the development of imaging techniques for medical diagnostics, with a specific focus on optical imaging and spectroscopy applications in ophthalmology and cancer diagnostics.

Qiong Wang, received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2012. She is currently a Professor at the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. Her research interests include surgical robots, VR applications in medicine, visualization, medical imaging, human–computer interaction, and computer graphics.