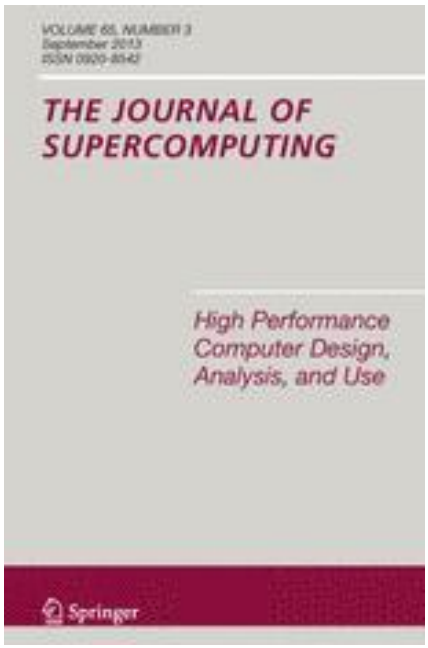

VOLUME 65, NUMBER 3
September 2013
ISSN 0920-8542
THE JOURNAL OF SUPERCOMPUTING
High Performance Computer Design, Analysis, and Use
Springer


Volume 82, Issue 3

February 2026

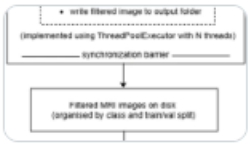
66 articles in this issue

## Acceleration of stochastic number generation using the segmentation method

Setareh Bazargan, Saadat Pour Mozafari & Houman Zarrabi

**OriginalPaper** | 04 February 2026 | Article: 124

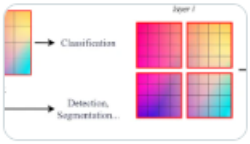
## Optimization of brain MRI preprocessing using parallel computing for efficient brain tumour classification




Lesia Mochurad, Khrystyna Melnychuk & Yulianna Mochurad

**OriginalPaper** | 03 February 2026 | Article: 123

## Swin-AFR-FPN: Attention and fusion refined object detection model with swin-transformer backbone for *Escherichia coli* (*E. coli*) bacterial cell cycle identification and antibiotic phenotyping

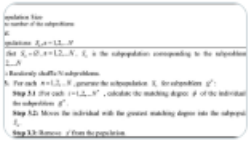



Yusuf Yargı Baydilli

**OriginalPaper** | Open access | 03 February 2026 | Article: 122

**Part of 1 collection:**
Section – Artificial Intelligence

## Decomposition-based multi-objective evolutionary algorithm with multi strategies for portfolio optimization problems

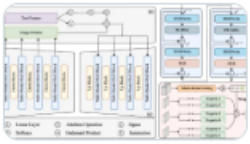


Chunlin Song, Bin Zhang ... Yingjie Song

**OriginalPaper** | 03 February 2026 | Article: 121

**Part of 1 collection:**
Section – Artificial Intelligence

## Sg-moe-net: semantic-guided mixture of experts with cross-modality attention for 3D medical segmentation



Yucai Qu & Canhui Xu

**OriginalPaper** | 03 February 2026 | Article: 120

**Part of 1 collection:**
Section – Artificial Intelligence

# SG-MoE-Net: Semantic-Guided Mixture of Experts with Cross-Modality Attention for 3D Medical Segmentation

Yucai Qu[1] and Canhui Xu[1*]

[1*]College of Information Science and Technology, Qingdao University of Science and Technology,Qingdao, 266061, Shandong, China.

*Corresponding author(s). E-mail(s): ccxu09@yeah.net;
Contributing authors: yucai.qu@qq.com;

**Abstract**

Multimodal 3D medical image segmentation remains challenging due to feature misalignment caused by intrinsic differences among imaging modalities. To address this issue, we propose the Semantic-Guided Mixture-of-Experts Network (SG-MoE-Net), a modality-aware framework built upon SwinUNETR. The proposed model incorporates a modality-decoupled Mixture-of-Experts architecture, in which heterogeneous experts are selectively activated to process CT, MRI, and semantic representations, thereby reducing cross-modality interference. Furthermore, a cross-level semantic guidance mechanism is introduced to integrate anatomical priors and promote semantic consistency across modalities during feature learning. Extensive experiments on three public multimodal benchmarks, namely CrossMoDA, MM-WHS, and HaN-Seg, demonstrate that SG-MoE-Net consistently outperforms existing state-of-the-art methods. In particular, the proposed approach achieves a Dice score of 84.3% on the CrossMoDA dataset. On the HaN-Seg benchmark, it attains a Dice score of 86.2%, outperforming existing state-of-the-art methods by 1.0% in Dice and 1.4% in Normalized Surface Dice.

**Keywords:** Deep Learning, 3D medical image segmentation, mixture of experts, semantic guidance

## 1 Introduction

Medical image segmentation forms the foundation of precision medicine, where segmentation accuracy directly determines the reliability of diagnosis, the design of personalized treatment strategies, and the advancement of biomedical research. In clinical practice, precise delineation of anatomical structures—covering their size, location, and boundaries—serves as a critical basis for decision-making and patient prognosis. However, the inherent complexity of human anatomy makes segmentation highly challenging: abdominal organs are spatially adjacent, and tumors often exhibit only subtle intensity differences from surrounding tissues. The heterogeneity across imaging modalities further compounds the difficulty: while CT excels at density resolution and MRI offers superior soft-tissue contrast, their fundamental disparities pose significant obstacles for multimodal fusion. Traditional approaches [1, 2] based on handcrafted features often lack robustness across diverse anatomical regions, while deep learning–based methods, despite notable progress, deep learning-based approaches still struggle to reconcile modality-specific characteristics with shared anatomical semantics, limiting their robustness in complex multimodal settings.

Early work on medical image segmentation predominantly relied on handcrafted features and heuristic thresholding strategies [3, 4]. While these methods achieved limited success in constrained scenarios, their generalization ability was inherently restricted. The emergence of deep learning marked a fundamental shift toward data-driven paradigms. Convolutional neural networks (CNNs), represented by U-Net [5] and its numerous extensions [6, 7], substantially improved segmentation performance by effectively capturing local structural patterns and refining boundary delineation. Nevertheless, the intrinsic locality of convolution operations limited their capacity to model long-range contextual dependencies.

To overcome this limitation, Transformer-based architectures have been introduced into medical image segmentation [8–10]. By leveraging global self-attention mechanisms, these models enable

# SG-MoE-Net: Semantic-Guided Mixture of Experts with Cross-Modality Attention for 3D Medical Segmentation

Yucai Qu[1] and Canhui Xu[1*]

[1*]College of Information Science and Technology, Qingdao University of Science and Technology,Qingdao, 266061, Shandong, China.

*Corresponding author(s). E-mail(s): ccxu09@yeah.net;
Contributing authors: yucai.qu@qq.com;

**Abstract**

Multimodal 3D medical image segmentation remains challenging due to feature misalignment caused by intrinsic differences among imaging modalities. To address this issue, we propose the Semantic-Guided Mixture-of-Experts Network (SG-MoE-Net), a modality-aware framework built upon SwinUNETR. The proposed model incorporates a modality-decoupled Mixture-of-Experts architecture, in which heterogeneous experts are selectively activated to process CT, MRI, and semantic representations, thereby reducing cross-modality interference. Furthermore, a cross-level semantic guidance mechanism is introduced to integrate anatomical priors and promote semantic consistency across modalities during feature learning. Extensive experiments on three public multimodal benchmarks, namely CrossMoDA, MM-WHS, and HaN-Seg, demonstrate that SG-MoE-Net consistently outperforms existing state-of-the-art methods. In particular, the proposed approach achieves a Dice score of 84.3% on the CrossMoDA dataset. On the HaN-Seg benchmark, it attains a Dice score of 86.2%, outperforming existing state-of-the-art methods by 1.0% in Dice and 1.4% in Normalized Surface Dice.

**Keywords:** Deep Learning, 3D medical image segmentation, mixture of experts, semantic guidance

## 1 Introduction

Medical image segmentation forms the foundation of precision medicine, where segmentation accuracy directly determines the reliability of diagnosis, the design of personalized treatment strategies, and the advancement of biomedical research. In clinical practice, precise delineation of anatomical structures—covering their size, location, and boundaries—serves as a critical basis for decision-making and patient prognosis. However, the inherent complexity of human anatomy makes segmentation highly challenging: abdominal organs are spatially adjacent, and tumors often exhibit only subtle intensity differences from surrounding tissues. The heterogeneity across imaging modalities further compounds the difficulty: while CT excels at density resolution and MRI offers superior soft-tissue contrast, their fundamental disparities pose significant obstacles for multimodal fusion. Traditional approaches [1, 2] based on handcrafted features often lack robustness across diverse anatomical regions, while deep learning–based methods, despite notable progress, deep learning-based approaches still struggle to reconcile modality-specific characteristics with shared anatomical semantics, limiting their robustness in complex multimodal settings.

Early work on medical image segmentation predominantly relied on handcrafted features and heuristic thresholding strategies [3, 4]. While these methods achieved limited success in constrained scenarios, their generalization ability was inherently restricted. The emergence of deep learning marked a fundamental shift toward data-driven paradigms. Convolutional neural networks (CNNs), represented by U-Net [5] and its numerous extensions [6, 7], substantially improved segmentation performance by effectively capturing local structural patterns and refining boundary delineation. Nevertheless, the intrinsic locality of convolution operations limited their capacity to model long-range contextual dependencies.

To overcome this limitation, Transformer-based architectures have been introduced into medical image segmentation [8–10]. By leveraging global self-attention mechanisms, these models enable

direct interactions among distant spatial locations, leading to notable performance improvements across a variety of benchmarks. Despite their success, the application of Transformers does not fully resolve the challenges posed by multimodal medical image segmentation.

A central difficulty lies in the effective integration of heterogeneous imaging modalities. Existing multimodal frameworks often adopt generic fusion strategies, in which multimodal inputs are either concatenated at early stages or processed through unified encoding pipelines, as exemplified by TransUNet [8] and SwinUNETR [9]. Although such designs can be effective under certain conditions, they implicitly assume a high degree of compatibility across modalities. In practice, however, modalities such as CT and MRI differ fundamentally in imaging principles, intensity distributions, and discriminative features. Forcing shared representations without sufficient decoupling may introduce modality interference, weakening modality-specific cues, including high-density structural information in CT or soft-tissue contrast in MRI.

Beyond feature-level interference, many current multimodal segmentation models lack explicit mechanisms to enforce high-level semantic consistency across modalities. Recent efforts, including dynamic fusion strategies [11] and text-driven approaches [12], attempt to incorporate prior knowledge into the learning process. However, semantic interactions are often simplified or restricted to late-stage decoding. Consequently, semantic alignment across modalities remains largely implicit during feature extraction, which can be particularly problematic in low-contrast or ambiguous regions where reliable semantic guidance is crucial for distinguishing anatomically related structures.

In parallel with these developments, Mixture-of-Experts (MoE) architectures have gained increasing attention as a means of scaling model capacity through conditional computation. Originally explored in natural language processing [13, 14] and later extended to computer vision tasks [15], MoE frameworks dynamically route inputs to specialized expert networks. Preliminary studies in medical imaging [16, 17] suggest that MoE-based designs may offer advantages in both efficiency and representation diversity. However, most existing implementations rely on homogeneous experts and routing strategies driven primarily by data distribution, without explicitly accounting for modality-specific characteristics or semantic alignment. As a result, directly adopting standard MoE formulations may be insufficient to address modality interference or to ensure anatomically consistent representations in multimodal segmentation.

Motivated by these observations, we propose Semantic-Guided Mixture-of-Experts Network (SG-MoE-Net) for multimodal medical image segmentation. The proposed framework aims to decouple modality-specific representations while simultaneously promoting semantic consistency across modalities. Rather than employing a fully shared encoder, SG-MoE-Net introduces a modality-aware Mixture-of-Experts mechanism, in which specialized experts are selectively activated to capture distinct modality characteristics alongside shared anatomical semantics. In addition, a cross-level semantic guidance strategy is incorporated to inject anatomical priors into the feature learning process, facilitating semantic alignment across modalities and improving robustness in challenging regions. Through joint optimization, spatial representations are not only effectively fused but also explicitly constrained to remain semantically consistent.

The main contributions of this work are summarized as follows:

- We propose a Mixture-of-Experts architecture specifically designed for multimodal medical image segmentation, enabling decoupled modality-specific processing to alleviate feature interference.
- A multi-level semantic guidance mechanism is introduced to promote semantic alignment across modalities beyond conventional decoder-level fusion.
- A unified training objective is developed by integrating cross-modal contrastive learning with semantic consistency constraints, explicitly regularizing feature alignment.
- Experiments conducted on three public multimodal datasets demonstrate that SG-MoE-Net consistently outperforms state-of-the-art methods in terms of segmentation accuracy and generalization.

# 2 Related Work

## 2.1 Deep learning based medical image segmentation

Early medical image segmentation methods were primarily based on classical image processing techniques, including handcrafted feature extraction and rule-based modeling [18, 19]. The introduction of convolutional neural networks (CNNs) marked a major shift in the field. U-Net [5] established a widely adopted encoder–decoder paradigm, which was subsequently extended to volumetric segmentation by V-Net [20] and further systematized by nnU-Net [21]. These CNN-based architectures significantly improved boundary delineation and training stability. However, their reliance on localized convolution operations inherently limits the receptive field, making it difficult to capture long-range contextual relationships that are critical for complex anatomical structures [22, 23].

To address this limitation, Vision Transformers (ViT) [24–27] were introduced to model global contextual dependencies through self-attention mechanisms. Hybrid architectures such as TransUNet [8] combined convolutional encoders with Transformer modules, enabling the integration of local and global representations for 2D segmentation tasks. This design philosophy was later extended to 3D volumetric segmentation by models including TransBTS [28] and SwinUNETR [9]. In particular, SwinUNETR [9] employs hierarchical shifted-window attention to balance global context modeling and computational efficiency, achieving strong performance on large-scale benchmarks such as BraTS [29].

Despite these advances, important challenges remain in multimodal medical image segmentation. Most existing Transformer-based frameworks [8, 9, 28] adopt straightforward fusion strategies, commonly based on modality concatenation or shared encoders, without explicitly accounting for the distributional heterogeneity across imaging modalities. Such unified processing can lead to modality interference, where modality-specific characteristics are partially suppressed during feature learning. Furthermore, while current architectures are effective at extracting spatial representations, they generally lack explicit mechanisms for modeling high-level semantic consistency across modalities. As a result, performance degradation is often observed in low-contrast or ambiguous regions, where spatial cues alone are insufficient to reliably distinguish anatomically related structures.

### 2.2 Mixture-of-Experts

Mixture-of-Experts (MoE) models [30, 31] have emerged as an effective paradigm for scaling model capacity without incurring proportional computational overhead. This efficiency stems from conditional computation with sparse activation, where only a subset of experts is selectively engaged for each input. Such a design has enabled substantial advances in natural language processing (NLP), underpinning the development of ultra-large models. Representative efforts include GShard [32], Switch Transformer [13], GLaM [33], Mixtral [34], and the DeepSeek family [14, 35–37], which scale to the trillion-parameter regime while retaining feasible training costs.

Building on these successes, MoE architectures have been extended beyond NLP to computer vision [15] and multimodal vision–language applications [38–40], achieving strong performance across diverse benchmarks. The modularity of MoE naturally accommodates heterogeneous data and multi-task learning, allowing individual experts to specialize in domain-specific subtasks. This flexibility has fueled applications in sentiment analysis [41], domain adaptation [42], multilingual translation [43], chart understanding [44], and multi-task vision problems [45, 46].

Despite these advantages, MoE systems face practical challenges, most notably expert load imbalance [32, 47], which has motivated the design of refined routing algorithms [13, 48] and theoretical analyses of their efficiency and generalization [49–51]. While the majority of research has concentrated on leveraging MoE's sparsity for large-scale representation learning in NLP, vision, and vision–language domains, interest is now shifting toward its application in medical image analysis [16, 17]. In this setting, the combination of MoE with semantic guidance holds particular promise. However, how best to integrate semantic priors with MoE's dynamic routing for domain-specific segmentation tasks—ranging from few-shot organ delineation and weakly supervised tumor detection to modality-aware multi-organ analysis—remains an open and compelling question.

## 3 Methodology

### 3.1 The Overall Architecture

The proposed SG-MoE-Net follows an encoder–decoder paradigm specifically designed for multimodal medical image segmentation. Multimodal inputs are first partitioned into patches and projected through modality-specific linear layers to obtain encoded representations. These features are then fed into the Multimodal Mixture-of-Experts (MMoE) Block, which leverages the Mixture-of-Experts (MoE) mechanism with dedicated CT, MRI, and semantic experts. A dynamic routing strategy selectively activates experts, enabling cross-modal feature disentanglement and targeted representation learning. Deeper within the encoder, a Swin Transformer is employed, whose shifted-window attention captures multi-scale contextual dependencies. The resulting hierarchical feature maps are passed to the decoder via skip connections. At each connection, a Semantic-Guided Module (SGM) fuses high-level semantic embeddings from deeper layers with low-level spatial features, while simultaneously aligning representations across modalities to reinforce semantic consistency. The aggregated features are progressively upsampled and refined in the decoder, and a final $1 \times 1$ convolution produces the multi-class segmentation probability map.

### 3.2 Multi-Modal MoE Block

Conventional multimodal segmentation models often treat modalities as independent or simply concatenate their features, overlooking both their complementary properties and inherent conflicts. Such designs frequently introduce semantic ambiguity and feature entanglement. To overcome these limitations, SG-MoE-Net introduces the Multi-Modal

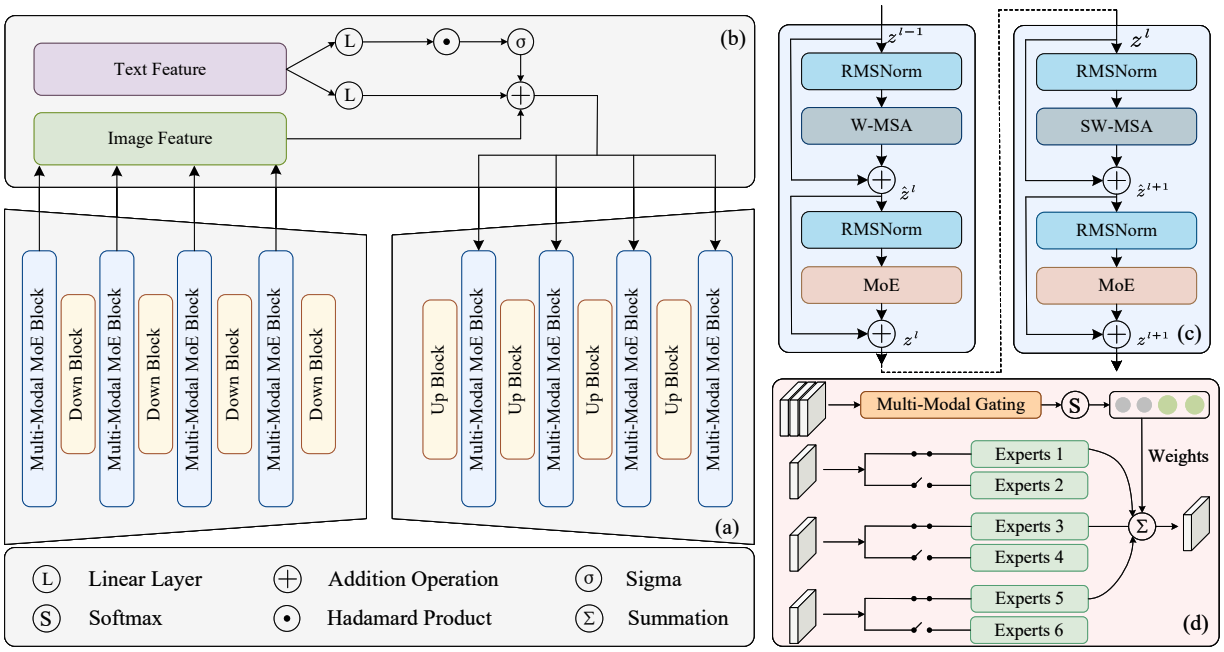



**Fig. 1** Illustration of the SG-MoE-Net model: (a) Overall UNet architecture, (b) Semantic-Guided Module, (c) Multi-Modal MoE Block, and (d) Modality-Specific Expert Networks with Dynamic Gating Network.

Mixture-of-Experts (MMoE) Block, which employs dynamic routing for adaptive feature decoupling and semantic enhancement.

As shown in Figure 1, the MMoE block consists of modality-specific expert networks and a dynamic gating module.

### 3.2.1 Modality-Specific Expert Networks

Each expert is implemented as a two-layer MLP with GELU activation:

$$Expert_i(x) = \text{GELU}(W_{i1}x + b_{i1})W_{i2} + b_{i2} \quad (1)$$

where $x$ is the input feature, $W_{i1}, W_{i2}$ are learnable weights, and $b_{i1}, b_{i2}$ are bias terms. The CT expert emphasizes high-density cues such as bone and calcification, while the MRI expert captures soft-tissue contrasts from T1/T2-weighted signals. The semantic expert takes fused multimodal features as input and produces cross-modality semantic embeddings. Unlike conventional MoE with homogeneous experts, this heterogeneous expert design simultaneously enables modality decoupling and explicit semantic modeling.

### 3.2.2 Dynamic Gating Network

The gating module computes the contribution of each expert through a single-layer MLP followed by a softmax activation:

$$g(x) = \text{softmax}(W_g x + b_g) \quad (2)$$

where $W_g$ and $b_g$ are learnable parameters. The MMoE output is the weighted combination of expert responses:

$$MMoE(x) = \sum_{i=1}^{K} g_i(x) \cdot Expert_i(x) \quad (3)$$

This mechanism adaptively selects modality-appropriate experts, effectively reallocating computation compared to shared MLPs.

### 3.2.3 Semantic-Guided Expert Collaboration

A key challenge in multimodal fusion is semantic inconsistency, where the same structure exhibits different representations across modalities. To address this, the semantic expert enforces cross-modal alignment via contrastive constraints. For CT features $x_{CT}$ and MRI features $x_{MRI}$, we impose:

$$\begin{aligned} &\text{sim}(E_{sem}(x_{CT}), E_{sem}(x_{MRI})) > \\ &\text{sim}(E_{sem}(x_{CT}), E_{sem}(x_{other})) \end{aligned} \quad (4)$$

where $\text{sim}(\cdot)$ denotes cosine similarity and $x_{other}$ refers to unrelated anatomical features. For example, in liver segmentation, this encourages semantic alignment of liver features across CT and MRI, reducing confusion caused by direct fusion.

The semantic expert further modulates modality-specific outputs through an attention-based mechanism:

$$Expert'_i(x) = Expert_i(x) \otimes \sigma(W_a E_{sem}(x) + b_a) \quad (5)$$

where $\otimes$ denotes element-wise multiplication, $\sigma$ is the sigmoid function, and $W_a, b_a$ are learnable

parameters. This design achieves effective disentanglement of modality-specific features while embedding global semantic cues, yielding robust feature representations for decoding.

## 3.3 Semantic-Guided Module

In standard encoder–decoder architectures, skip connections primarily deliver low-level spatial details but often lack high-level semantic context, leading to semantic drift—particularly severe in multimodal settings. To mitigate this, we introduce the Semantic-Guided Module (SGM) within each skip connection. The SGM bridges semantic and spatial features, aligns representations across modalities, and introduces domain priors to ensure anatomically plausible predictions.

Semantic embeddings $S$ are projected into two complementary subspaces:

$$f_{sem1} = \text{Linear}_1(S), \quad f_{sem2} = \text{Linear}_2(S) \tag{6}$$

where $\text{Linear}_1$ and $\text{Linear}_2$ are learnable mappings. For skip features $f_{skip}$, semantic modulation is performed as:

$$f_{guided} = f_{skip} \odot \sigma(f_{sem1}) + f_{sem2} \tag{7}$$

where $\odot$ is the Hadamard product. This enriches low-level features with task-relevant semantic guidance while suppressing irrelevant responses.

To further address cross-modal inconsistency, we enforce semantic alignment via a contrastive loss:

$$\begin{aligned}\mathcal{L}_{modality} = -\log \frac{\exp\left(\text{sim}(e^{CT}_{sem}, e^{MRI}_{sem})/\tau\right)}{\exp\left(\text{sim}(e^{CT}_{sem}, e^{MRI}_{sem})/\tau\right)} \\ + \sum_{i \neq j} \exp\left(\text{sim}(e^{i}_{sem}, e^{j}_{sem})/\tau\right)\end{aligned} \tag{8}$$

where $e^{CT}_{sem}$ and $e^{MRI}_{sem}$ are CT and MRI semantic embeddings, $\tau$ is a temperature coefficient, and $\text{sim}(\cdot)$ computes cosine similarity.

The SGM and MMoE interact bidirectionally: modality-specific experts provide priors to the SGM, while semantic modulation from the SGM refines expert responses, forming a complementary interplay between anatomical knowledge and modality-specific representation.

## 3.4 Loss Function

SG-MoE-Net is optimized with a joint objective balancing segmentation accuracy, cross-modal alignment, and semantic consistency:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \alpha \mathcal{L}_{modality} + \beta \mathcal{L}_{semantic} \tag{9}$$

### 3.4.1 Primary Segmentation Loss

We combine Dice, cross-entropy, and boundary terms:

$$\mathcal{L}_{seg} = \lambda_1 \mathcal{L}_{Dice} + \lambda_2 \mathcal{L}_{CE} + \lambda_3 \mathcal{L}_{Boundary} \tag{10}$$

Dice loss addresses class imbalance:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i + \varepsilon}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2 + \varepsilon} \tag{11}$$

where $p_i$ and $g_i$ denote predictions and ground truth. Cross-entropy loss enforces discriminative class prediction:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} g_i^c \log P_i^c \tag{12}$$

and boundary loss sharpens region boundaries:

$$\mathcal{L}_{Boundary} = -\sum_{i=1}^{N} g_i \log P_i \cdot d_i \tag{13}$$

where $d_i$ is the distance to the ground-truth boundary.

### 3.4.2 Cross-Modal Alignment Loss

This term enforces alignment of modality-specific embeddings using the contrastive form in Eq.8.

### 3.4.3 Semantic Consistency Loss

To stabilize semantic learning, we introduce a classification-based constraint:

$$\mathcal{L}_{semantic} = -\frac{1}{M} \sum_{m=1}^{M} \sum_{c=1}^{C} y_m^c \log(f_{semantic}(x_m)) \tag{14}$$

where $x_m$ is the input, $y_m^c$ the semantic class, and $f_{semantic}$ the classifier output.

# 4 Experiences

## 4.1 Dataset

To validate our method's effectiveness, we conducted volumetric segmentation experiments on three publicly available multimodal dataset: CrossMoDA2021 [58], MM-WHS [59], and HaN-Seg [60].

### 4.1.1 CrossMoDA2021 Dataset

The CrossMoDA2021 dataset provides brain MRI scans across two modalities: T1-CE and T2-HR. It includes expert annotations for vestibular schwannoma and cochlea masks. The training set consists of 105 T1-CE brain images with their corresponding segmentation masks, along with 105 unlabeled T2-HR brain images. The validation set contains 32

**Table 1** Quantitative Comparisons of Segmentation Performance in CrossMoDA, MM-WHS and HaN-Seg dataset. Our method is compared against current state-the-art models. The best result of each column is highlighted in bold.

| Method | CrossMoDA | | | | MM-WHS | | | | HaN-Seg | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DSC | NSD | HD95 | MASD | DSC | NSD | HD95 | MASD | DSC | NSD | HD95 | MASD |
| 3D UNet [52] | 0.642 | 0.597 | 1.050 | 0.605 | 0.709 | 0.700 | 1.903 | 1.416 | 0.574 | 0.571 | 18.035 | 20.512 |
| Swin-UNet [10] | 0.671 | 0.620 | 0.827 | 0.620 | 0.720 | 0.694 | 1.871 | 0.942 | 0.665 | 0.620 | 21.531 | 24.542 |
| TransUNet [8] | 0.680 | 0.626 | 0.603 | 0.503 | 0.796 | 0.702 | 1.890 | 1.241 | 0.696 | 0.648 | 15.456 | 12.346 |
| CoTr [53] | 0.723 | 0.775 | 0.529 | 0.424 | 0.876 | 0.794 | 1.752 | 0.801 | 0.741 | 0.746 | 10.735 | 9.238 |
| MISSFormer [54] | 0.735 | 0.760 | 0.513 | 0.416 | 0.872 | 0.772 | 1.705 | 0.780 | 0.784 | 0.761 | 9.376 | 9.804 |
| nnUNet [21] | 0.791 | 0.794 | 0.472 | 0.225 | 0.887 | 0.790 | 1.684 | 0.761 | 0.819 | 0.790 | 8.280 | 8.790 |
| TransBTS [28] | 0.769 | 0.808 | 0.420 | 0.247 | 0.870 | 0.791 | 1.749 | 0.762 | 0.801 | 0.808 | 8.372 | 8.803 |
| SwinUNETR [9] | 0.832 | 0.824 | 0.339 | 0.174 | **0.905** | 0.850 | 1.317 | 0.754 | 0.852 | 0.836 | **7.241** | 7.483 |
| SAM-Med3D [55] | 0.817 | 0.816 | 0.358 | 0.201 | 0.867 | 0.842 | 1.452 | **0.748** | 0.823 | 0.831 | 7.801 | 7.035 |
| SegMamba [56] | 0.835 | **0.829** | 0.332 | 0.173 | 0.895 | 0.848 | 1.315 | 0.759 | 0.855 | 0.821 | 7.471 | 6.950 |
| D-LKA Net [57] | 0.837 | 0.821 | 0.339 | 0.151 | 0.875 | 0.839 | 1.310 | 0.765 | 0.857 | 0.842 | 7.410 | 7.028 |
| Ours | **0.843** | 0.823 | **0.326** | **0.142** | 0.893 | **0.854** | **1.209** | 0.752 | **0.862** | **0.850** | 7.585 | **6.815** |

T2-HR brain MRI scans, and the test set includes 107 T2-HR brain MRI scans.

#### 4.1.2 MM-WHS Dataset

The MM-WHS dataset comprises 120 multimodal cardiac images, specifically 60 cardiac CT/CTA and 60 cardiac MRI scans. All images, encompassing the entire heart and its critical substructures, were acquired in real clinical settings and used for diagnostic purposes. The dataset is divided into a training set, consisting of 20 CT and 20 MRI samples, and a test set, which includes 40 CT and 40 MRI samples. For the training set, detailed manual annotations are provided for seven major cardiac substructures: the left and right ventricular blood pools, left and right atrial blood pools, left ventricular myocardium, ascending aorta, and pulmonary artery.

#### 4.1.3 HaN-Seg Dataset

HaN-Seg is a publicly accessible dataset for head and neck organ at risk(OAR) image segmentation. It features CT and MRI images from 42 patients. This dataset provides binary segmentation masks for 30 individual organs at risk.

### 4.2 Implementation Details

Our SG-MoE-Net was implemented using PyTorch 1.10.1 and MONAI 1.3.0. All training was conducted on a single NVIDIA RTX 4090 GPU. We used the AdamW optimizer with a momentum of 0.99. The initial learning rate was set to 0.0001 and was reduced to 80% of its current value every 100 epochs. Training proceeded for 500 epochs with a batch size of 2. During training, we cropped random $128 \times 128 \times 128$ patches from the 3D image volumes. We applied random axis mirroring with a 0.5 probability along all three axes. Furthermore, we implemented data augmentation transformations on the input image channels, including random per-channel intensity shifts within a range of $(-0.1, 0.1)$ and random intensity scaling within a range of $(0.9, 1.1)$. For inference, we employed a sliding window approach with an overlap of 0.7 between adjacent voxels.

### 4.3 Evaluation Metrics

We evaluate the performance of all models using for complementary metrics: Dice Similarity Coefficient (DSC) [61], Normalized Surface Dice(NSD) [62], 95% Hausdorff Distance (HD95) [63], and Mean Average Surface Distance(MASD) [64].

The DSC quantifies the degree of overlap between the volumetric segmentation predictions and the ground truth voxels, defined as follows:

$$DSC\,(Y, P) = 2 * \frac{|Y \cap P|}{|Y| \cup |P|} = 2 * \frac{Y \cdot P}{Y^2 + P^2} \quad (15)$$

here, $Y$ and $P$ denote the ground truth and the output probabilities for all voxels, respectively.

NSD quantifies the overlap between the predicted and ground-truth boundaries within the segmentation voxels, providing a normalized evaluation of their alignment. This metric accounts for both the boundaries and border regions, offering a comprehensive assessment of how well the predicted

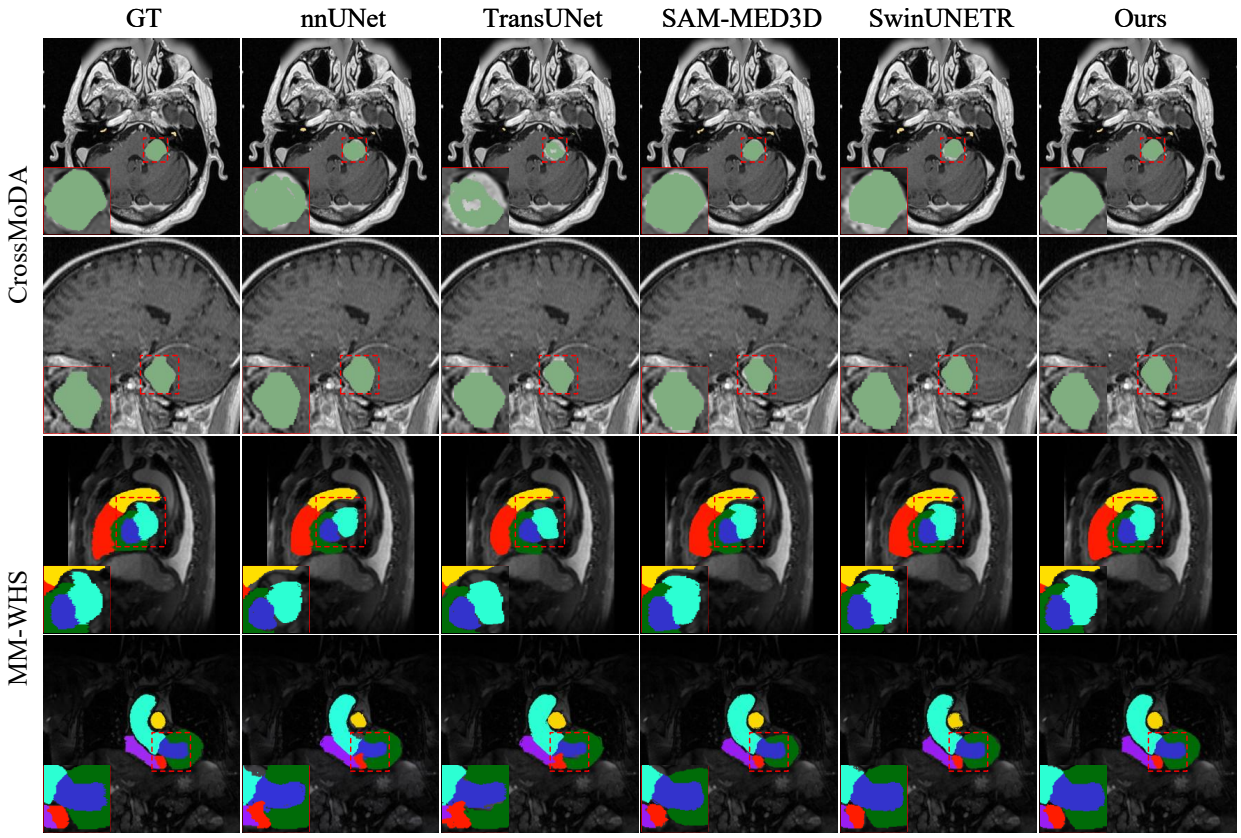



**Fig. 2** Qualitative comparison on the CrossMoDA and MM-WHS dataset. We compare our AgentUNETR with existing methods.

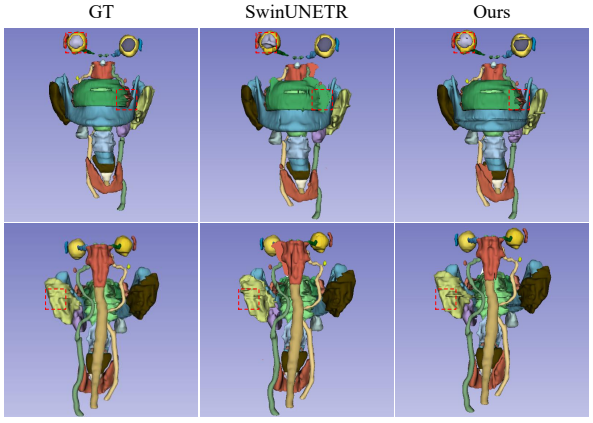



**Fig. 3** On the HaN-Seg test set, a comparison of visualizations with SwinUNETR demonstrates the performance of the segmentation results.

surface aligns with the reference. It is formally defined as:

$$NSD(Y,P) = \frac{|S_Y \cap B_P| + |S_P \cap B_Y|}{|S_Y| + |S_P|} \quad (16)$$

where $S_Y$ and $S_P$ represent the boundary voxels of the ground truth and prediction, respectively, while $B_Y$ and $B_P$ denote the corresponding border regions around these boundaries.

HD95 is commonly used as boundary-based metric to measure the $95^{th}$ percentile of the distances between boundaries of the volumetric segmentation predictions and the voxels of the ground truths. It is defined as follows:

$$HD_{95}(Y,P) = max\left\{d_{95}(Y,P), d_{95}(P,Y)\right\} \quad (17)$$

where, $d_{95}(Y,P)$ is the maximum $95^{th}$ percentile distance between the ground truth and predicted voxels, and $d_{95}(P,Y)$ is the maximum $95^{th}$ percentile distance between the predicted and ground truth voxels.

MASD evaluates the average spatial deviation between the predicted and ground-truth surfaces by computing the mean of distances from each surface to the other. This bidirectional metric captures spatial dissimilarity and quantifies how far, on average, the predicted voxels deviate from their ground-truth counterparts. It is defined as:

$$MASD(Y,P) = \frac{1}{2}\left(\frac{\sum_{y\in Y} d(y,P)}{|Y|} + \frac{\sum_{p\in P} d(Y,p)}{|P|}\right) \quad (18)$$

Here, $d(y,P)$ denotes the shortest distance from a ground-truth voxel $y \in Y$ to the predicted surface $P$, and $d(Y,p)$ represents the shortest distance from a predicted voxel $p \in P$ to the ground-truth surface $Y$.

### 4.4 Comparison with state-of-the-art methods

We evaluated the performance of our proposed SG-MoE-Net against state-of-the-art methods on the CrossMoDA, MM-WHS, and HaN-Seg dataset. All reported results reflect single-model performance without any pre-training, model ensembles, or additional data.

#### 4.4.1 Experiments of the CrossMoDA dataset

Table 1 summarizes the quantitative evaluation on the CrossMoDA dataset. Segmentation performance is assessed using the Dice Similarity Coefficient (DSC), Normalized Surface Distance (NSD), the 95th percentile Hausdorff Distance (HD95), and Mean Average Surface Distance (MASD). SG-MoE-Net achieves an average DSC of 84.3%, outperforming SwinUNETR [9], SAM-Med3D [55], as well as recent methods such as SegMamba [56] and D-LKA Net [57]. In addition to improved region overlap, SG-MoE-Net exhibits lower boundary-related errors, achieving an HD95 of 0.326 and a MASD of 0.142. These results indicate that decoupling modality-specific representations contributes to more accurate and stable surface delineation, suggesting improved robustness against cross-modality feature interference compared with sequential or purely convolutional architectures.

The top rows of Figure 2 present a qualitative comparison between SG-MoE-Net and existing methods on different cases from the CrossMoDA dataset. To facilitate a clearer inspection of subtle differences in boundary delineation, magnified views of the regions of interest (ROIs) are provided in the bottom-left corner of each sample and highlighted by red dashed boxes. The first row illustrates a relatively simple case, where existing approaches show limited accuracy in delineating tumor boundaries. In contrast, as observed in the magnified ROIs, SG-MoE-Net achieves a more precise and sharper segmentation. The second row presents a more challenging example: both TransUNet [8] and SAM-MED3D [55] exhibit under-segmentation, failing to fully capture the vestibular schwannoma(VS), while nnUNet [21] and SwinUNETR [9] tend to over-segment the VS region. By comparison, SG-MoE-Net provides more accurate delineations, achieving a better balance between completeness and precision.

#### 4.4.2 Experiments of the MM-WHS dataset

The comparative evaluation on the MM-WHS dataset is presented in Table 1. SG-MoE-Net achieves competitive performance across all metrics, with particularly strong results on boundary-sensitive measures, attaining an NSD of 85.4% and an HD95 of 1.209. Although SegMamba [56] and SwinUNETR [9] report comparable or slightly higher DSC values, SG-MoE-Net consistently maintains lower boundary errors. This behavior is especially relevant for cardiac structure segmentation, where accurate boundary delineation is critical. The stable performance on boundary-related metrics suggests that the proposed semantic guidance contributes to more coherent anatomical representations in regions with limited contrast.

The bottom two rows of Figure 2 illustrate the performance of SG-MoE-Net in comparison with state-of-the-art methods on representative cases from the MM-WHS dataset. The magnified ROIs provide a clearer view of regions with increased structural complexity, facilitating a more detailed visual assessment. In the third row, nnUNet [21] and TransUNet [8] fail to accurately segment the left atrial cavity (light blue), exhibiting clear under-segmentation. Similarly, SAM-MED3D [55] and SwinUNETR [9] underperform in delineating the pulmonary artery (yellow), also showing signs of under-segmentation. The fourth row highlights further challenges: nnUNet [21] and TransUNet [8] capture excessive non-target regions in the right ventricular cavity (red), resulting in over-segmentation. While SAM-MED3D [55] performs relatively well overall, it struggles to precisely define the boundaries of the right ventricular cavity (purple). SwinUNETR [9] similarly fails to accurately model the boundaries of the pulmonary artery. In contrast, SG-MoE-Net consistently achieves higher accuracy and robustness across cardiac structures, demonstrating superior capability in segmenting anatomically complex regions.

#### 4.4.3 Experiments of the HaN-Seg

Table 1 summarizes the results on the HaN-Seg dataset. SG-MoE-Net achieves the highest performance on most evaluation metrics, including DSC of 86.2%, NSD of 85.0%, and MASD of 6.815. These results represent a clear improvement over recent methods such as D-LKA Net [57] and SegMamba [56]. While SwinUNETR [9] and SegMamba [56] obtain slightly lower HD95 values, qualitative inspection shows that SG-MoE-Net produces more anatomically complete segmentations for complex head-and-neck structures, indicating a favorable balance between boundary precision and overall structural integrity.

Figure 3 presents a 3D visualization of the segmentation results on the HaN-Seg dataset, providing an intuitive view of segmentation quality that is particularly relevant for clinical assessment. To enable a more detailed comparison of complex anatomical structures, several representative regions are highlighted using red dashed boxes, where notable differences among methods can be observed. As illustrated in Figure 3, SwinUNETR [9] exhibits certain limitations in multi-organ segmentation tasks, often producing blurry boundaries

**Table 2** Ablation experiments on CrossMoDA dataset.

| Model | DSC | MASD |
|---|---|---|
| baseline | 0.832 | 0.174 |
| baseline+SGM | 0.836 | 0.155 |
| baseline+MMoE | 0.840 | 0.138 |
| baseline+SGM+MMoE | 0.843 | 0.142 |

and incomplete organ segmentations. Specifically, in the regions indicated by the red boxes, it shows a noticeable bulge at the edge of the right eye and misses information for the oral mucosa. In contrast, our SG-MoE-Net achieves clearer boundaries and more complete organ segmentation.

### 4.5 Ablation Experiment

To evaluate the effectiveness of each core component in SG-MoE-Net, we conducted a series of ablation studies on the CrossMoDA dataset. By comparing different model variants, we quantitatively assessed the individual contributions of each module to the overall segmentation performance.

#### 4.5.1 With SGM Module

As shown in Table 2, equipping the baseline model with the SGM leads to a 0.4% improvement in DSC (from 83.2% to 83.6%). This performance gain is likely attributed to the SGM's dual-branch semantic transformation design, which injects high-level anatomical priors into low-level feature representations, thereby enhancing the semantic separability between target structures and background regions.

#### 4.5.2 With MMoE Module

When the baseline is augmented with MMoE, DSC increases by 0.8% (from 83.2% to 84.0%), and MASD is reduced by 0.036 (from 0.174 to 0.138). This improvement stems from the MMoE's gated expert mechanism, which assigns modality-specific experts to CT and MRI inputs. This design enables the model to dynamically capture modality-specific representations and mitigates feature confusion arising from inter-modal discrepancies.

#### 4.5.3 With SGM and MMoE Module

Table 2 further demonstrates that our full SG-MoE-Net achieves a DSC of 84.3% and a MASD of 0.142, outperforming all other model variants. The fine-grained modality-specific feature extracted by MMoE serve as a foundation for SGM's semantic guidance, while the anatomical priors introduced by SGM help mitigate potential modality biases from MMoE, leading to an overall improvement in segmentation performance.

## 5 Discussion

The experimental results support the effectiveness of SG-MoE-Net. Ablation studies indicate that the proposed Mixture-of-Experts design reduces the Mean Average Surface Distance (MASD) on the CrossMoDA dataset from 0.174 to 0.142, suggesting that decoupling modality-specific representations can effectively alleviate cross-modality interference. In addition, qualitative results demonstrate improved delineation of anatomically ambiguous boundaries, such as the oral mucosa, where intensity contrast is limited. These observations indicate that the proposed semantic guidance mechanism contributes to more stable and semantically coherent representations in low-contrast regions.

Despite these improvements, several limitations remain. Incorporating a Mixture-of-Experts architecture introduces additional computational cost and increases parameter complexity compared with single-stream networks. Moreover, the effectiveness of semantic guidance depends on the availability and quality of anatomical annotations, which may constrain performance in scenarios with limited labeled data. Future work will explore more lightweight expert designs to improve computational efficiency and investigate weakly supervised or semi-supervised strategies to reduce reliance on dense annotations, thereby enhancing the practical applicability of the proposed framework in clinical settings.

## 6 Conclusion

This study tackles the issue of semantic inconsistency caused by modality discrepancies in multimodal medical image segmentation. We present SG-MoE-Net, a novel framework that couples a multimodal mixture-of-experts mechanism with explicit semantic guidance. The architecture integrates three key components: a Semantic Guidance Module (SGM), a Multimodal Mixture-of-Experts (MMoE) block, and a jointly optimized loss function, enabling accurate feature fusion across modalities while enforcing semantic alignment. Comprehensive evaluations on three public benchmarks verify the effectiveness and robustness of our approach. Looking ahead, we plan to investigate lightweight architectural designs to facilitate clinical deployment and pursue fine-grained anatomical semantic modeling to further enhance the clinical utility of the framework.

## 7 Funding

The authors declare that they have no affiliations with or involvement in any organization or entity with any financial interest or nonfinancial interest in the subject matter or materials discussed in this manuscript.

# 8 Data availability statements

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

# References


[1] Heimann, T., Meinzer, H.-P., Wolf, I.: A statistical deformable model for the segmentation of liver ct volumes. In: Proc. MICCAI Workshop 3-D Segmentation in the Clinic: A Grand Challenge, p. 161 (2007)

[2] Zikic, D., Glocker, B., Konukoglu, E., Criminisi, A., Demiralp, C., Shotton, J., Thomas, O.M., Das, T., Jena, R., Price, S.J.: Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel mr. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 369–376 (2012). Springer

[3] Moltz, J.H., Bornemann, L., Dicken, V., Peitgen, H.: Segmentation of liver metastases in ct scans by adaptive thresholding and morphological processing. In: MICCAI Workshop, vol. 41, p. 195 (2008)

[4] Sezgin, M., Sankur, B.l.: Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic imaging **13**(1), 146–168 (2004)

[5] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241 (2015). Springer

[6] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)

[7] Cai, S., Tian, Y., Lui, H., Zeng, H., Wu, Y., Chen, G.: Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. Quantitative imaging in medicine and surgery **10**(6), 1275 (2020)

[8] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[9] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop, pp. 272–284 (2021). Springer

[10] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218 (2022). Springer

[11] Zhang, J., Xie, Y., Xia, Y., Shen, C.: Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1195–1204 (2021)

[12] Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7086–7096 (2022)

[13] Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)

[14] Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., et al.: Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. arXiv preprint arXiv:2401.06066 (2024)

[15] Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. Advances in Neural Information Processing Systems **34**, 8583–8595 (2021)

[16] Liu, Y., Pei, J., He, Z., Yang, G., Jiang, Z., Lao, Q.: Medical language mixture of experts for improving medical image segmentation. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2210–2216 (2024). IEEE

[17] Jiang, S., Zheng, T., Zhang, Y., Jin, Y., Yuan, L., Liu, Z.: Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. arXiv preprint arXiv:2404.10237 (2024)

[18] Tsai, A., Yezzi, A., Wells, W., Tempany, C., Tucker, D., Fan, A., Grimson, W.E., Willsky, A.: A shape-based approach to the segmentation of medical imagery using level sets. IEEE transactions on medical imaging **22**(2), 137–154 (2003)

[19] Held, K., Kops, E.R., Krause, B.J., Wells, W.M., Kikinis, R., Muller-Gartner, H.-W.: Markov random field segmentation of brain mr images. IEEE transactions on medical imaging **16**(6), 878–886 (1997)

[20] Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016). Ieee

[21] Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

[22] Feng, S., Zhao, H., Shi, F., Cheng, X., Wang, M., Ma, Y., Xiang, D., Zhu, W., Chen, X.: Cpfnet: Context pyramid fusion network for medical image segmentation. IEEE transactions on medical imaging **39**(10), 3008–3018 (2020)

[23] Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging **38**(10), 2281–2292 (2019)

[24] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[25] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

[26] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020). Springer

[27] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021). PMLR

[28] Wenxuan, W., Chen, C., Meng, D., Hong, Y., Sen, Z., Jiangyun, L.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 109–119 (2021)

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

[30] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)

[31] Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the em algorithm. Neural computation **6**(2), 181–214 (1994)

[32] Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020)

[33] Du, N., Huang, Y., Dai, A.M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A.W., Firat, O., *et al.*: Glam: Efficient scaling of language models with mixture-of-experts. In: International Conference on Machine Learning, pp. 5547–5569 (2022). PMLR

[34] Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)

[35] Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D., et al.: Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434 (2024)

[36] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)

[37] Li, P., Shi, C.: Selectgan: Mamba based explicit selectivity gan for generating structured content. Signal, Image and Video Processing **19**(7), 562 (2025)

[38] Lin, B., Tang, Z., Ye, Y., Cui, J., Zhu, B., Jin, P., Huang,

J., Zhang, J., Pang, Y., Ning, M., et al.: Moe-llava: Mixture of experts for large vision-language models. arXiv preprint arXiv:2401.15947 (2024)

[39] Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024)

[40] Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)

[41] Guo, J., Shah, D.J., Barzilay, R.: Multi-source domain adaptation with mixture of experts. arXiv preprint arXiv:1809.02256 (2018)

[42] Gururangan, S., Lewis, M., Holtzman, A., Smith, N.A., Zettlemoyer, L.: Demix layers: Disentangling domains for modular language modeling. arXiv preprint arXiv:2108.05036 (2021)

[43] Chirkova, N., Nikoulina, V., Meunier, J.-L., Bérard, A.: Investigating the potential of sparse mixtures-of-experts for multi-domain neural machine translation. arXiv preprint arXiv:2407.01126 (2024)

[44] Xu, Z., Qu, B., Qi, Y., Du, S., Xu, C., Yuan, C., Guo, J.: Chartmoe: Mixture of diversely aligned expert connector for chart understanding. In: The Thirteenth International Conference on Learning Representations (2025)

[45] Chen, Z., Shen, Y., Ding, M., Chen, Z., Zhao, H., Learned-Miller, E.G., Gan, C.: Mod-squad: Designing mixtures of experts as modular multi-task learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11828–11837 (2023)

[46] Jain, Y., Behl, H., Kira, Z., Vineet, V.: Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. Advances in Neural Information Processing Systems **36**, 69625–69637 (2023)

[47] Nie, X., Cao, S., Miao, X., Ma, L., Xue, J., Miao, Y., Yang, Z., Yang, Z., Cui, B.: Dense-to-sparse gate for mixture-of-experts (2021)

[48] Zhu, J., Zhu, X., Wang, W., Wang, X., Li, H., Wang, X., Dai, J.: Uni-perceiver-moe: Learning sparse generalist models with conditional moes. Advances in Neural Information Processing Systems **35**, 2664–2678 (2022)

[49] Baykal, C., Dikkala, N., Panigrahy, R., Rashtchian, C., Wang, X.: A theoretical view on sparsely activated networks. Advances in Neural Information Processing Systems **35**, 30071–30084 (2022)

[50] Dikkala, N., Ghosh, N., Meka, R., Panigrahy, R., Vyas, N., Wang, X.: On the benefits of learning to route in mixture-of-experts models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 9376–9396 (2023)

[51] Mu, S., Lin, S.: A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. arXiv preprint arXiv:2503.07137 (2025)

[52] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, pp. 424–432 (2016). Springer

[53] Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, pp. 171–180 (2021). Springer

[54] Huang, X., Deng, Z., Li, D., Yuan, X.: Missformer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162 (2021)

[55] Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., *et al.*: Sam-med3d: towards general-purpose segmentation models for volumetric medical images. In: European Conference on Computer Vision, pp. 51–67 (2025). Springer

[56] Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation (2024) arXiv:2401.13560 [cs.CV]

[57] Azad, R., Niggemeier, L., Huttemann, M., Kazerouni, A., Aghdam, E.K., Velichko, Y., Bagci, U., Merhof, D.: Beyond self-attention: Deformable large kernel attention for medical image segmentation. arXiv preprint arXiv:2309.00121 (2023)

[58] Shapey, J., Kujawa, A., Dorent, R., Wang, G., Dimitriadis, A., Grishchuk, D., Paddick, I., Kitchen, N., Bradford, R., Saeed, S.R., *et al.*: Segmentation of vestibular schwannoma from mri, an open annotated dataset and baseline algorithm. Scientific Data **8**(1), 286 (2021)

[59] Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(12), 2933–2946 (2019) https://doi.org/10.1109/TPAMI.2018.2869576

[60] Podobnik, G., Strojan, P., Peterlin, P., Ibragimov, B., Vrtovec, T.: Han-seg: The head and neck organ-at-risk ct and mr segmentation dataset. Medical physics **50**(3), 1917–1927 (2023)

[61] Crum, W.R., Camara, O., Hill, D.L.: Generalized overlap measures for evaluation and validation in medical image analysis. IEEE transactions on medical imaging **25**(11), 1451–1461 (2006)

[62] Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., Fauw, J.D., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., Kelly, C., Karthikesalingam, A., Chu, C., Carnell, D., Boon, C., D'Souza, D., Moinuddin, S.A., Garie, B., McQuinlan, Y., Ireland, S., Hampton, K., Fuller, K., Montgomery, H., Rees, G., Suleyman, M., Back, T., Hughes, C., Ledsam, J.R., Ronneberger, O.: Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy (2021) arXiv:1809.04430 [cs.CV]

[63] Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence **15**(9), 850–863 (2002)

[64] Gerig, G., Jomier, M., Chakos, M.: Valmet: A new validation tool for assessing and improving 3d object segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 516–523 (2001). Springer