UniMRSeg: Unified Modality-Relax Segmentation via Hierarchical Self-Supervised Compensation

Xiaoqi Zhao[1] **Youwei Pang**[2] **Chenyang Yu**[3]
Lihe Zhang[3] **Huchuan Lu**[3] **Shijian Lu**[2] **Georges El Fakhri**[1] **Xiaofeng Liu**[1]
[1]Yale University, USA
[2] Nanyang Technological University, Singapore
[3]Dalian University of Technology, China
xiaoqi.zhao@yale.edu

Abstract

Multi-modal image segmentation faces real-world deployment challenges from incomplete/corrupted modalities degrading performance. While existing methods address training-inference modality gaps via specialized per-combination models, they introduce high deployment costs by requiring exhaustive model subsets and model-modality matching. In this work, we propose a unified modality-relax segmentation network (UniMRSeg) through hierarchical self-supervised compensation (HSSC). Our approach hierarchically bridges representation gaps between complete and incomplete modalities across input, feature and output levels. First, we adopt modality reconstruction with the hybrid shuffled-masking augmentation, encouraging the model to learn the intrinsic modality characteristics and generate meaningful representations for missing modalities through cross-modal fusion. Next, modality-invariant contrastive learning implicitly compensates the feature space distance among incomplete-complete modality pairs. Furthermore, the proposed lightweight reverse attention adapter explicitly compensates for the weak perceptual semantics in the frozen encoder. Last, UniMRSeg is fine-tuned under the hybrid consistency constraint to ensure stable prediction under all modality combinations without large performance fluctuations. Without bells and whistles, UniMRSeg significantly outperforms the state-of-the-art methods under diverse missing modality scenarios on MRI-based brain tumor segmentation, RGB-D semantic segmentation, RGB-D/T salient object segmentation. The code will be released at `https://github.com/Xiaoqi-Zhao-DLUT/UniMRSeg`.

1 Introduction

Visual multi-modal image segmentation has become a cornerstone in critical applications such as autonomous driving [9], medical diagnostics [23], and robotics [67], where complementary visual cues (*e.g.*, RGB-D, MRI sequences) improve scene understanding. Advanced hybrid CNN-Transformer [14, 46, 39, 79, 68, 40], global-local attention [64, 66, 81, 82, 84, 78, 75, 41], multi-scale [38, 36, 37, 83], dynamic convolution [35, 77, 76] techniques have achieved remarkable success under idealized settings with complete modalities. However, real-world scenarios often suffer from incomplete modality inputs due to sensor failures, low-quality data and clinical constraints. For example, although it is ideal to use four complementary MRI modalities—fluid-attenuated inversion recovery (Flair), contrast-enhanced T1-weighted (T1ce), T1-weighted (T1), and T2-weighted (T2)—for brain tumor diagnosis, variations in scanning protocols and patient conditions may limit the ability to obtain all MRI scans.

Corresponding Authors: Youwei Pang and Lihe Zhang

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

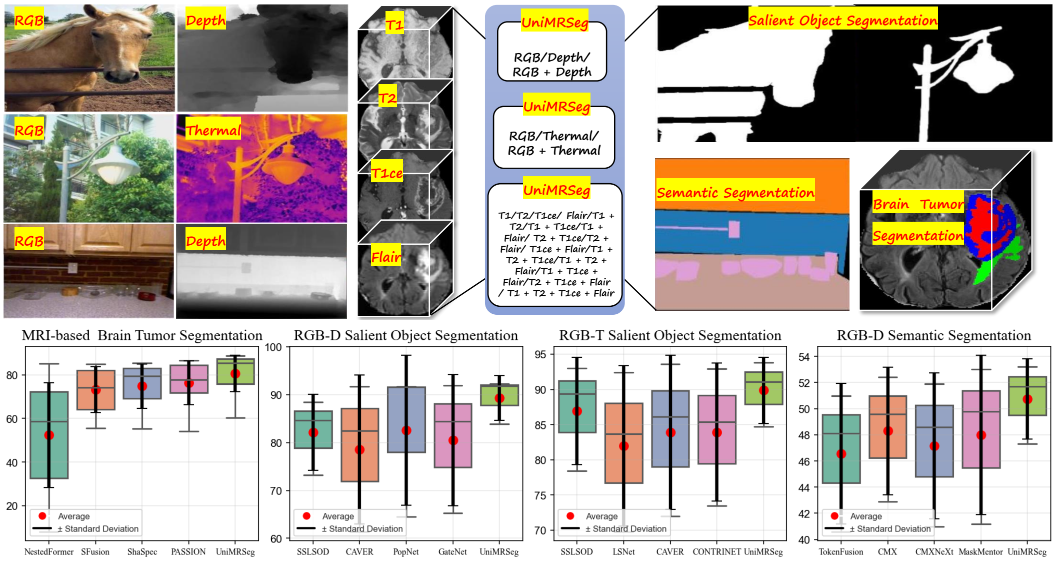


Figure 1: *Top*: UniMRSeg has a unified framework and parameters within each segmentation task to handle 21 modality combinations (RGB-Depth: 3, RGB-Thermal: 3, MRI: 15). *Bottom*: Box plots compare UniMRSeg with existing methods across four benchmarks, displaying average performance (red dots) and standard deviation (error bars).

Recent research on addressing missing modalities falls into two main challenges. ***Firstly***, most methods [71, 58, 72, 80] focus on designing adaptable cross-modal interaction to amalgamate multi-modal features while preserving generalizable architectures for single-modal scenarios. However, during inference, diverse modality combinations require separate model parameters [71, 58] or independent encoder parameters [69, 74], which not only increase resource consumption in practical deployment but also necessitate additional manual or automatic modality classification as a prerequisite. Although some works [25, 8] leverage the knowledge distillation from complete to incomplete modalities, they still demand multiple models for each modality subset, complicating clinical deployment. ***Secondly***, modality reconstruction-based methods [80, 69, 25, 85] aim to predict missing modality inputs to align features during training and inference. Since segmentation requires precise spatial features and boundary information, the pre-trained reconstruction model prioritizes global feature compression, resulting in insufficient feature representation. Therefore, it is difficult to directly inherit the ability to reduce the modality gap obtained by input-level [80, 25] or feature-level [69, 85] reconstruction for downstream tasks (*i.e.*, image segmentation). In particular, cascading low-quality reconstruction predictions [25] as input to the segmentation network will increase error propagation and degrade performance.

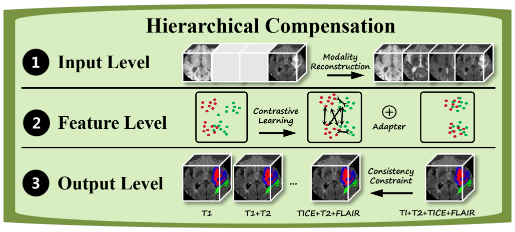


Figure 2: Illustration of the hierarchical compensation.

In this paper, we propose a unified modality-relax segmentation framework (UniMRSeg). As shown in Fig. 1, UniMRSeg shares 100% of its parameters across all possible modality input combinations in a given segmentation task. Since complete modality inputs typically yield the best prediction, our goal is to ensure that UniMRSeg, after training, can approach complete modality representation quality during inference with arbitrary modality inputs. To achieve this, we fully exploit the power of self-supervision [1, 16, 4, 57] in representation learning and propose a hierarchical compensation mechanism that operates at the input level, feature level, and output level, as shown in Fig. 2. ***First***, we adopt cross-modal reconstruction as a pretext task. Unlike previous methods [80, 2] that either discard complete modalities or employ partial masking strategies, our approach simultaneously applies global and local masking mechanisms. This dual masking design enables the model to capture both fine-grained local patterns and holistic semantic representations from intra-modality and cross-modality interactions.

Furthermore, we introduce a channel-wise modality shuffling operation that deliberately breaks the correspondence between input and reconstructed modalities. This operation implicitly formulates a modality classification task, compelling the model to disentangle modality-specific characteristics while eliminating dependence on modality category priors during inference. ***Next***, we leverage contrastive learning to finish the feature-level compensation. Specifically, we construct complete and randomly missing modalities from the same sample as positive pairs, and those from other samples as negative pairs. To enhance the inheritance of representation for downstream segmentation tasks [33, 7], we jointly optimize the spatial distance metric and object segmentation at this stage to guide feature clustering in a direction that is beneficial for the results of the segmentation. Considering the inevitable prediction errors in the aforementioned pixel-level reconstruction and feature contrast learning, they may still limit the boundaries of the compensation mechanism. Inspired by adapter-based methods [26, 5, 61], we design a lightweight reverse attention adapter to explicitly compensate for weak perceptual semantics in the frozen encoder, while embedding a 3D Swin Transformer [28, 52] to capture high-response mutual attention patterns across modalities. By adding feature-level consistency constraints, we ensure that the adapter is aware of the partial representation defects inherent in any missing modal combination. ***Last***, UniMRSeg is fine-tuned by enforcing segmentation consistency constraints, and the knowledge presented by the complete modality at the output will be distilled to all missing modality combinations as supervision information. Through the hierarchical three-level compensation strategy, UniMRSeg achieves optimal average performance, the highest full-modality performance, and minimal performance variance, as shown in Fig. 1.

Our main contributions can be summarized as follows:

- We propose a unified framework, *i.e.*, UniMRSeg, with one set of parameters adapting to varying modality-missing scenarios for general multi-modal segmentation.
- We integrate pixel-level modality reconstruction, feature-level contrastive learning, and prediction-level label distillation to construct a hierarchical annotation-free compensation mechanism, breaking through the previous isolated self-supervised research paradigm.
- Benefiting from the reverse attention adapter, UniMRSeg can explicitly obtain the compensation about difficult regions with weak perception of the complete modality, aligning missing and complete modality representations.
- Extensive comparisons conducted on brain tumor, salient object and semantic segmentation tasks in MRI (15 combinations), RGB-D (3 combinations), and RGB-T (3 combinations) modalities across 2D and 3D images within medical and natural scenes, show that our method consistently achieves the best performance in all individual modality combinations while attaining superior average accuracy with minimal standard deviation.

2 Related Works

2.1 Incomplete Multi-modal Image Segmentation

There are three popular research patterns to handle missing modalities situations. ***I) Generalizable Architectures.*** Some methods [58, 71, 73] aim to fully integrate multi-modal features while minimizing structural modifications in single-modal scenarios. CMX [71] introduces a cross-modal interaction attention mechanism that combines both the CNN and Transformer. Tokenfusion [58] focuses on efficient token-based fusion, pruning multiple single-modal Transformer and repurposing the pruned units for multi-modal fusion. ***II) Projection-based Methods.*** Hetero [11] and HeMIS [15] perform arithmetic operations (*e.g.*, averaging) in the projected space to obtain the final segmentation result. SFusion [30] proposes a self-attention-based fusion block, where extracted features from available modalities are projected as tokens and processed through a self-attention layer to capture cross-modal relationships. ***III) Reconstruction Strategies.*** M3AE [25] refines the segmentation network by reusing the reconstructed modality images as inputs for fine-tuning. MaskMentor [85] and SSLSOD [80] treat modality reconstruction as a pre-training representation task, where missing and complete modalities share weights for joint training. Zeng *et al.* [69] enforce consistency between missing modality feature reconstruction and complete modality features at the token level. Different from them, we aim to achieve a simple yet efficient architecture while completing the comprehensive compensation of multi-modal and multi-combinatorial representations with unified parameters by constructing multi-granularity self-supervised tasks that are not limited to pixel/token-level reconstruction.

2.2 Self-supervised Learning

Self-supervised learning (SSL) has emerged as a powerful paradigm for learning representations without relying on manual annotations. Three principal SSL mechanisms dominate current research: ***I) Mask-based SSL*** involves training models to predict missing or masked parts of an input. Early image inpainting works [42, 70] evolved into modern masked autoencoders [16, 60] using high-ratio masking and asymmetric encoder-decoder architectures to various visual tasks, including image segmentation, depth estimation, and low-level image restoration. ***II) Contrastive learning*** encourages models to bring similar data points closer while distancing dissimilar ones. SimCLR [4] establishes data augmentation principles with momentum encoders, while MoCo series [17, 6] address negative sample scarcity through dynamic queues. CLIP [44] leverages cross-modal contrastive learning to embed images and text into a unified semantic space, enabling zero-shot transfer capabilities and providing a universal representation foundation for multi-modal tasks like medical image retrieval and caption generation. ***III) Knowledge distillation*** [19] transfers supervision via teacher-student paradigms. In image segmentation, Chen *et al.* [49] propose to normalize the activation map of each channel to obtain a soft probability map. For cross-modal scenarios, Gupta *et al.* [13] transfer supervision from labeled RGB images to unlabeled depth and optical flow images. Existing studies usually develop these paradigms in isolation. We hope to take the incomplete multi-modal image segmentation task as an opportunity to effectively combine these three different levels of self-supervision techniques and demonstrate their synergy.

3 Approach

Preliminaries. In this section, we adopt the MRI-based brain tumor segmentation with the T1, T1ce, T2 and Flair modalities as the targeted task to describe the proposed multi-stage learning framework. Let $I \in \mathbb{R}^{1\times H\times W\times T}$ be the input modality-specific sequence with T slices for the model, where 1, H, and W are the channel, height, and width of the slice. The final model generates the segmentation tensor $P \in \mathbb{R}^{N\times H\times W\times T}$, where each channel corresponds to a specific class and indicates the probability of each spatial-temporal location being assigned to that class.

Multi-stage Learning Framework. Our framework aims to bridge the performance gap between complete and incomplete multi-modal inputs through three-stage progressive learning, specifically designed to empower flexible handling of diverse missing modality scenarios while maintaining segmentation robustness. The overall pipeline is shown in Fig. 3. The basic models in all stages follow a unified 3D U-Net-style [45] encoder-decoder structure and embed a 3D ASPP [3] with dilation rates of [1, 6, 12, 18] into the high-level feature. The following sections elaborate on the technical specifics of each learning stage.

3.1 Multi-granular Modality Reconstruction

This stage emphasizes enhancing the representation capabilities under diverse potential input-side missing modality scenarios. Existing work [16] has demonstrated that effective data perturbations facilitate learning more robust representations. To encourage the model to mine implicit contextual relationships across modalities, this stage integrates three strategies, including *modality dropout*, *modality shuffle* and *spatial masking*, to enable multi-granular information reconstruction.

Data Perturbation. For the complete multi-modal input, we first apply the **random modality dropout** strategy to randomly discard some modalities with a 50% probability and generate the output. And it also preserves at least one modality, thus retaining the fundamental information for overall reconstruction. Furthermore, we **randomly shuffle** the order of remaining modalities, which mitigates the model's reliance on fixed modality order and further decouples modality-agnostic representations from their inherent sequential dependencies in existing paradigms [85, 25]. Additionally, the following **spatial masking** strategy randomly masks a portion of the input data, thus simulating missing effects within the sequences from available modalities.

Data Reconstruction. We input these remaining perturbed samples into the 3D U-Net-based reconstruction network, and obtain the output through a ReLU function. The normalized slices from the original complete modalities are used as reconstruction objectives based on the combination loss of L1 and SSIM [59], thus overall achieving self-supervised pre-training in the first stage.

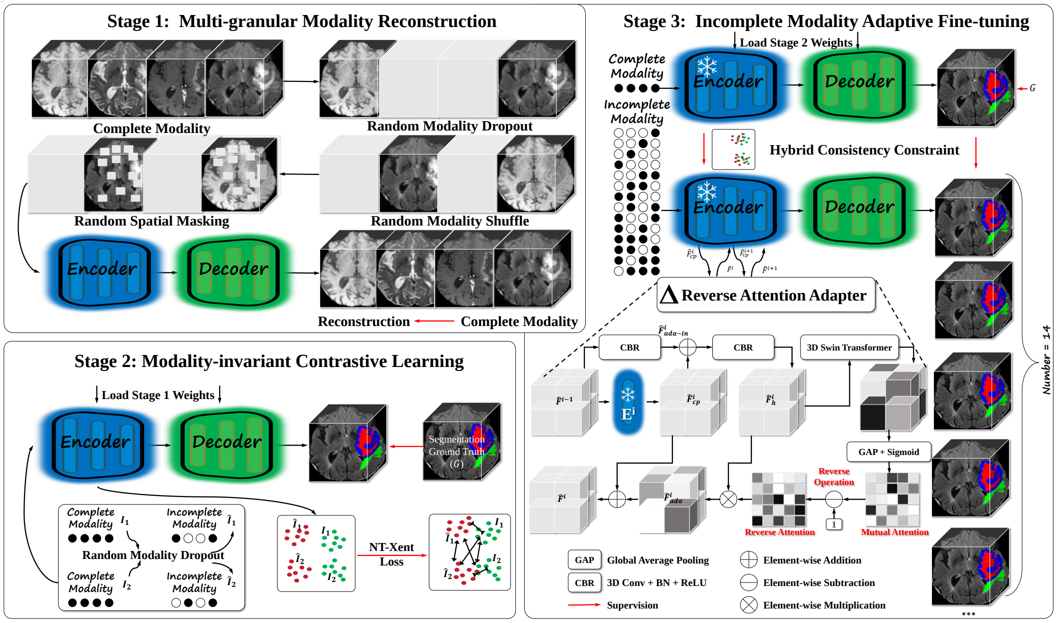


Figure 3: Multi-stage learning framework built on our encoder-decoder network. Stage 1: complete modality reconstruction based on multi-granular random perturbations. Stage 2: modality-invariant contrastive learning for enhancing incomplete-modality representation. Stage 3: incomplete modality adaptive fine-tuning via hybrid consistency constraints.

3.2 Modality-invariant Contrastive Learning

The core idea behind this stage is to implicitly compensate for the feature biases introduced by missing modalities through contrastive learning.

Representation Construction. Without loss of generality, we consider a batch containing two input samples with complete modalities, *i.e.*, I_1 and I_2. By applying random modality dropout to each of them while ensuring at least one and at most three modalities remain active, we can obtain two extended samples ($\hat{I}_1$ and $\hat{I}_2$) with missing modalities. And then, we initialize the segmentation model by load the weights of the first stage except for the last output layer. $\{I_1, \hat{I}_1, I_2, \hat{I}_2\}$ are then fed into the segmentation model to extract multi-level encoder features and generate the final predictions $\{P_1, \hat{P}_1, P_2, \hat{P}_2\}$, respectively. After the global average pooling, those features are converted into four vector sets $\{f_1^i\}_{i=1}^5$, $\{\hat{f}_1\}_{i=1}^5$, $\{f_2\}_{i=1}^5$, and $\{\hat{f}_2\}_{i=1}^5$, respectively. They are utilized for the following representation contrastive learning.

Contrastive Learning. The aforementioned representation vectors are utilized to construct positive-negative sample relationships. And we introduce the NT-Xent loss [4] in each feature level to minimize distances between positive pairs and maximize separation of negative pairs. Specifically, for the i^{th} level, positive pairs are from the same input sample and its augmented variant, *i.e.*, I_k^i and $\hat{I}_k$ ($k \in \{1, 2\}$). Negative pairs are from distinct input sources, *e.g.*, (I_1, I_2) and $(I_1, \hat{I}_2)$. Considering the vector set $\mathbf{f}^i = \{f_1^i, \hat{f}_1^i, f_2^i, \hat{f}_2^i\}$, we can calculate the NT-Xent loss as follows:

$$l^i(u, v) = -\log \frac{\exp(\mathrm{sim}(\mathbf{f}_u^i, \mathbf{f}_v^i)/\tau)}{\sum_{k=1}^{2B} \mathbb{I}_{[k \neq i]} \exp(\mathrm{sim}(\mathbf{f}_u^i, \mathbf{f}_k^i)/\tau)} \tag{1}$$

$$L_{\text{NT-Xent}} = \sum_{i=1}^{5} \sum_{k=1}^{B} \frac{l^i(2k-1, 2k) + l^i(2k, 2k-1)}{2B \times 5} \tag{2}$$

where $B = 2$ and sim are the batch size and the cosine similarity. $\mathbb{I}_{[k \neq i]}$ is an indicator function that results in 1 when $k \neq i$ otherwise 0. Such a contrastive learning encourages the model to learn modality-invariant representations, focusing on the underlying target semantic features rather than being biased by modality-specific characteristics.

Segmentation Constraint. The common Dice loss l_{Dice} is used to supervise the segmentation predictions as follows:

$$L_{\text{Dice}} = \frac{1}{B}\sum_{k=1}^{B}\sum_{P\in\{P_k,\hat{P}_k\}} l_{\text{Dice}}(P, G_k) \tag{3}$$

It is worth noting that despite applying the contrastive learning, the segmentation constraint is also necessary to guide the encoder's representation learning. This ensures the learned features are optimized and aligned for our ultimate objective, *i.e.*, the segmentation task.

3.3 Incomplete Modality Adaptive Fine-tuning

Due to potential errors in the reconstruction from the first stage and the positive-negative sample distance control in the second stage, we introduce the following adaptive fine-tuning guided by the complete modalities to dynamically compensate for these representation errors. As shown in Fig. 3, our framework implements parallel pipelines to handle complete modality samples and their all potential incomplete counterparts (14 valid variants for 4 MRI modalities). The model directly generates the ideal intermediate features $\{F^i\}_{i=1}^5$ and segmentation prediction P as the reference from complete modality samples via the forward propagation. And its prediction P is supervised by the GT mask using the Dice loss. The incomplete samples additionally require targeted optimization through lightweight adapters $\{\mathbf{A}^i\}_{i=1}^5$, which progressively aligns their features $\{\hat{F}^i\}_{i=1}^5$ and prediction $\hat{P}$ with those references. To maintain the representation capabilities pre-trained from the first two stages, we freeze the entire encoder during this stage and only fine-tune the decoder and adapters.

Reverse Attention Adapter. Taking the i^{th} encoder stage as an example, we attach the adapter component to the frozen encoder stage $\mathbf{E}^i$. In the feature propagation path for incomplete multi-modal samples, the feature $\hat{F}^{i-1}$ from the previous encoder layer is first processed by $\mathbf{E}^i$ to obtain the base features $\hat{F}^i_{cp}$. Meanwhile, $\hat{F}^{i-1}$ also undergoes 3D convolutions and output the initial adaptive feature $\hat{F}^i_{ada-in}$, which is then integrated with $\hat{F}^i_{cp}$ via element-wise addition and sequential 3D convolutions. The generated feature $\hat{F}^i_h$ is fed into a 3D Swin Transformer [29] block to establish global contextual correlations between the $\hat{F}^i_{cp}$ and $\hat{F}^i_{ada-in}$. And then the global average pooling is applied across channel and sequence dimensions, followed by the sigmoid to generate the mutual attention. We hope to capture the difficult semantic parts that cannot be perceived by the first two stages and then compensate $\hat{F}^i_{cp}$. Therefore, we apply the reverse operation to the mutual attention map and generate the reverse attention map. And it is multiplied to $\hat{F}^i_h$ and generate the adapted feature $\hat{F}^i_{ada}$, which highlights differential information between modality-complete and -incomplete representations. Finally, the sum $\hat{F}^i$ of $\hat{F}^i_{ada}$ and $\hat{F}^i_{cp}$ replaces original $\hat{F}^i_{cp}$ as the real input to the subsequent process. The roles played by the above features need to be activated with the help of feature-level consistency constraints.

Hybrid Consistency Constraints. The consistency constraint process depicted in Fig. 3 involves the levels of encoder features and final predictions. For the **feature-level consistency**, we compare the intermediate features $\{F^i\}_{i=1}^5$ from modality-complete samples and $\{\hat{F}^i\}_{i=1}^5$ compensated by the adapter from modality-incomplete samples:

$$L_{fc} = \frac{1}{B}\sum_{k=1}^{B}\sum_{m}^{M}\frac{1}{5}\sum_{i=1}^{5}\|F^i_k - \hat{F}^i_{k,m}\|_1 \tag{4}$$

where M denotes the number of potential valid modality-incomplete samples and it is 14 in our MRI experiments. Besides, the **prediction-level consistency** between the predicted segmentation maps from modality-complete and -incomplete samples can be formulated by:

$$L_{pc} = \frac{1}{B}\sum_{k=1}^{B}\sum_{m=1}^{M} l_{\text{Dice}}(P^i_k, \hat{P}^i_{k,m}) \tag{5}$$

By accumulating these consistency constraints, the parameters of the decoder and adapters are jointly optimized to minimize the difference between the representations from modality-complete and -incomplete samples.

Table 1: Quantitative comparison of brain tumor segmentation on BraTS2020 [32]. ↑ and ↓ indicate that the larger scores and the smaller ones are better, respectively. **Note**: Official implementations are used when available. All methods are evaluated under a unified missing modality setting when possible. Some methods use different training data and lack released code, making exact re-training infeasible. Extended results under various settings are in the Appendix.

Modality				Dice score(%)														
				Whole Tumor (Whole)					Tumor Core (Core)					Enhancing Tumor (Enhancing)				
Flair	T1	T1ce	T2	NestedFormer [63]	SFusion [30]	ShaSpec [55]	PASSION [48]	UniMRSeg	NestedFormer [63]	SFusion [30]	ShaSpec [55]	PASSION [48]	UniMRSeg	NestedFormer [63]	SFusion [30]	ShaSpec [55]	PASSION [48]	UniMRSeg
○	○	○	●	23.22	58.54	63.10	64.83	**75.14**	8.26	37.20	42.83	50.36	**56.31**	2.31	16.34	22.15	30.23	**35.19**
○	○	●	○	24.61	60.82	57.31	60.92	**68.87**	27.40	45.19	54.59	54.78	**79.20**	35.85	44.34	64.93	69.12	**79.20**
○	●	○	○	8.15	55.69	55.40	54.10	**60.41**	6.77	29.84	34.26	39.93	**45.70**	3.55	14.13	18.23	22.13	**28.70**
●	○	○	○	61.18	74.25	79.54	77.35	**85.33**	34.01	43.23	52.74	50.27	**61.81**	29.84	30.92	37.40	39.60	**45.61**
○	○	●	●	47.22	72.91	70.60	75.82	**81.64**	40.02	71.82	75.93	80.35	**82.01**	49.13	66.18	74.23	79.03	**79.58**
○	●	●	○	40.87	60.30	67.58	67.90	**72.98**	52.34	75.62	78.92	79.85	**84.60**	56.87	66.09	77.13	78.32	**80.23**
●	●	○	○	70.54	84.40	80.23	82.39	**86.22**	42.84	64.31	60.20	60.55	**67.90**	33.61	36.53	48.08	46.05	**51.14**
○	●	○	●	22.77	67.35	72.40	77.71	**77.98**	13.16	42.92	52.80	51.60	**59.71**	7.26	32.29	36.02	33.42	**41.49**
●	○	○	●	59.18	83.71	85.31	86.64	**87.78**	32.42	55.62	63.53	52.87	**70.00**	29.40	32.09	43.52	43.34	**53.12**
●	○	●	○	58.57	84.90	82.80	85.74	**87.96**	43.91	73.30	76.52	78.94	**85.71**	54.05	69.48	77.28	79.08	**81.59**
●	●	●	○	76.67	80.52	84.12	85.42	**86.91**	63.12	77.30	79.34	80.38	**85.92**	69.73	74.36	76.02	78.12	**81.28**
●	●	○	●	73.00	78.90	81.13	81.31	**85.50**	42.25	63.62	61.93	63.95	**67.43**	32.44	30.03	39.89	43.60	**49.57**
●	○	●	●	74.30	80.65	83.20	83.52	**87.51**	60.04	76.28	78.93	79.92	**82.78**	64.47	70.02	74.68	79.09	**80.50**
○	●	●	●	58.77	70.52	74.36	76.31	**76.57**	60.32	78.48	80.65	82.36	**84.92**	69.46	67.18	73.39	76.04	**77.15**
●	●	●	●	81.07	84.92	85.02	85.90	**88.74**	67.02	78.84	84.26	84.75	**86.01**	73.78	72.09	75.59	80.72	**82.18**
Average ↑				52.01	73.23	74.81	76.39	**80.64**	39.59	60.90	65.16	66.06	**73.33**	40.78	48.14	55.90	58.53	**63.10**
Std Dev ↓				23.09	10.47	10.08	10.07	**8.43**	19.53	17.07	15.45	15.34	**13.04**	24.20	21.80	21.62	21.84	**19.86**

Table 2: Quantitative comparison of segmentation performance across RGB, Depth and Thermal modalities.

Modality		RGB-D Salient Object Segmentation (STERE [34])					RGB-T Salient Object Segmentation (VT1000 [54])					RGB-D Semantic Segmentation (SUN-RGBD [50])				
RGB	Depth/Thermal	SSLSOD [80]	CAVER [39]	PopNet [62]	GateNet [82]	UniMRSeg	SSLSOD [80]	LSNet [86]	CAVER [39]	CONTRINET [51]	UniMRSeg	TokenFusion [58]	CMX [71]	CMXNeXt [72]	MaskMentor [85]	UniMRSeg
●	○	.846	.825	.916	.844	**.918**	.784	.698	.720	.735	**.847**	48.1	49.6	48.6	49.8	**51.7**
○	●	.732	.614	.645	.653	**.839**	.894	.837	.861	.854	**.911**	40.6	42.9	41.0	41.2	**47.3**
●	●	.885	.917	.917	.919	**.923**	.930	.924	.936	.929	**.938**	51.0	52.4	51.9	53.0	**53.2**
Average ↑		.821	.785	.826	.805	**.893**	.869	.820	.839	.839	**.899**	46.6	48.3	47.2	48.0	**50.7**
Std Dev ↓		.080	.155	.157	.137	**.047**	.076	.114	.110	.098	**.047**	5.4	4.9	5.6	6.1	**3.1**

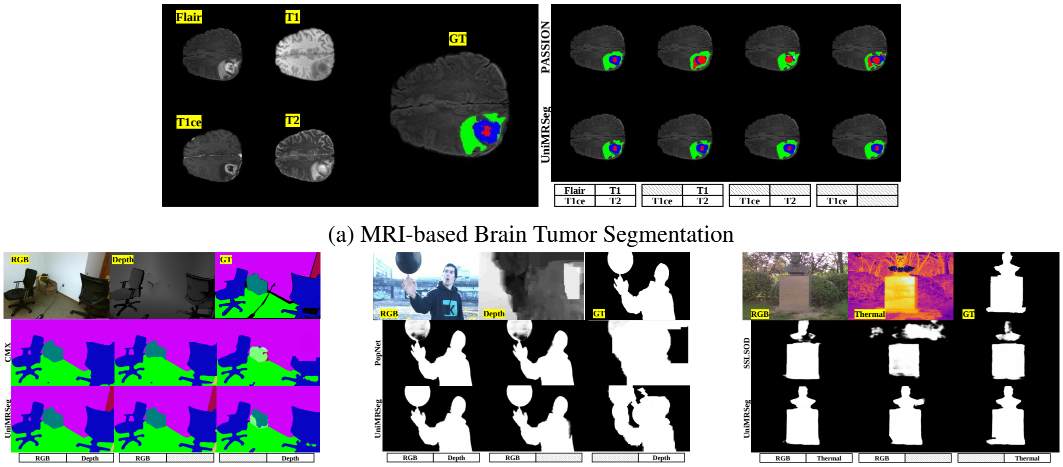
4 Experiments

4.1 Datasets and Metrics

In this work, we conduct experimental comparisons on four popular visual multi-modal image segmentation tasks to show the generalizability of the proposed method. ***I) Brain Tumor Segmentation***. We follow most brain tumor segmentation methods [22, 63, 10, 30] use the BraTS2020 dataset [32], which contains 369 images with four modality scans: T1ce, T1, T2, and Flair, along with three annotated regions: enhancing tumor, tumor core, and whole tumor, which are mutually inclusive. The dataset is split into training (315), validation (17), and test (37) sets. ***II) RGB-D Salient Object Segmentation***. We adopt the same training set as most methods [80, 82, 39], *i.e.*, 1,485 samples from the NJUD [20] and 700 samples from the NLPR [43]. The test dataset is STERE [34], which contains 1,000 RGB and depth image pairs with complex scenes. ***III) RGB-T Salient Object Segmentation***. We follow the setting of recent works [39, 86, 51], the training set only contains the 2,500 samples from VT5000 [53] and adopt the VT1000 [54] as the test set which contains 1,000 pairs of RGB-T images including more than 400 kinds of common objects collected in 10 types of scenes under different illumination conditions. ***IV) RGB-D Semantic Segmentation***. SUN-RGBD [50] is the popular indoor scene benchmark with 37 classes. It contains 10,335 pairs of RGB-D images, with 5,285 pairs allocated for training and 5,050 for testing. We introduce some widely used metrics in each field for fair evaluation, including Dice for brain tumor segmentation, S-measure [12] (S_m) for salient object segmentation and IoU for semantic segmentation.

4.2 Implementation Details

All experiments are conducted on one NVIDIA A800 GPU. We adopt basic image augmentation techniques to avoid overfitting, including random flipping, rotating and border clipping. We train the model for 300 epochs based on the AdamW optimizer [31] with a warmup schedule, an initial learning rate of 0.0001, and a weight decay of 0.00001. For RGB-D and RGB-T tasks, we concatenate the inputs into a 4-channel tensor to maintain the single-stream architecture. For fair comparison, we separately adopt ResNet-50 [18] and ConvNext-B [27] as the backbone for RGB-D/T salient object segmentation and RGB-D semantic segmentation, which are widely used in their respective fields.



(a) MRI-based Brain Tumor Segmentation

(b) RGB-D Semantic Segmentation

(c) RGB-D Salient Object Segmentation

(d) RGB-T Salient Object Segmentation

Figure 4: Qualitative comparisons of predictions from different methods across different modality combinations. Best viewed on screen.

4.3 Evaluation

Quantitative Results. We conduct thorough comparisons on all the four tasks, as shown in Tab. 1 and Tab. 2. All these models are either directly tested for performance or retrained based on their publicly released code. To simulate missing modalities in practice and keep consistent with training, we fill missing modalities in MRI with zero pixel values and directly copy other existing modalities to complete the dual input of any missing modality in RGB-D and RGB-T tasks. In addition, we follow some methods with depth estimation functions [80, 62, 85] to cascade the intrinsic depth predictor as the input of the models for RGB-D salient object segmentation and RGB-D semantic segmentation tasks. It can be seen that our proposed UniMRSeg achieves excellent performance on all tasks with the best average performance and lowest standard deviation. These high accuracy, strong robustness and generalization capabilities across multiple modalities show its potential as a unified, reliable, and efficient segmentation framework.

Qualitative Results. In Fig. 4a, PASSION [48] exhibits significant prediction fluctuations for enhancing tumor (**blue**) and tumor core (**red**) under missing modalities, while UniMRSeg maintains highly consistent predictions aligned with ground truth. T1 and Flair modalities notably degrade PASSION's boundary discrimination for whole tumor (**green**), revealing mutually exclusive fusion deficiencies. In Fig. 4b, UniMRSeg achieves precise door segmentation (**red**) by fusing color and geometric features across RGB/RGB-D modes, whereas CMX [71] fails completely with depth-only inputs. Leveraging spatial topology of chairs, UniMRSeg robustly segments blurred wall-adjacent boxes using pure depth data, demonstrating superior shape reasoning. Fig. 4c shows PopNet [62] produces identical spherical predictions in RGB/RGB-D modes via embedded depth prediction, but degenerates to depth map binarization without RGB. In contrast, UniMRSeg reconstructs RGB semantics from depth and fuses cross-modal features to accurately segment humans and spheres. Fig. 4d verifies thermal modality's advantage in resolving RGB's color-similarity-induced sticking issues. Remarkably, UniMRSeg achieves complete target segmentation using thermal data alone.

4.4 Ablation Study

In this section, we show the effectiveness of each component on the brain tumor segmentation task with the most complex modality combination. The baseline model is 3D-UNet without any pre-training.

Each Components. As shown in Tab. 3, the three data perturbations in stage 1 demonstrate the effectiveness of learning intra- and inter-modal features for multi-granular reconstruction. Their accumulation during pre-training leads to over 14.7% performance gain over the baseline in subsequent segmentation. Stage 2 highlights the importance of jointly training the segmentation decoder and the encoder guided by spatial distance clustering, which promote each other. Stage 3 further optimizes cross-modal alignment through reverse attention adapters and prediction-level consistency

Table 3: Ablation study of each component.

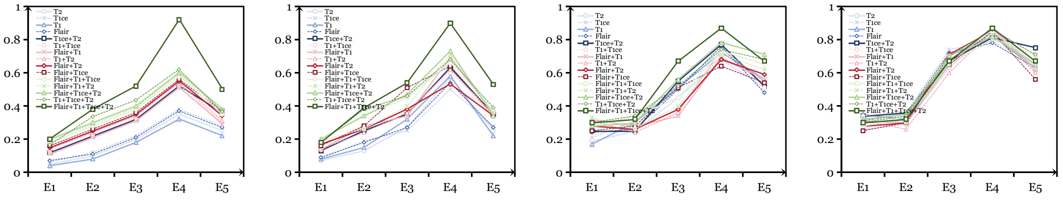
Models	Average Dice score(%)		
	Whole	Core	Enhancing
Baseline	63.31	51.60	38.40
Stage 1: Multi-granular Modality Reconstruction (Sec. 3.1)			
+ Random Modality Dropout	66.98	55.47	42.25
+ Random Modality Shuffle	67.78	56.85	44.17
+ Random Spatial Masking	69.35	59.89	47.12
Stage 2: Modality-invariant Contrastive Learning (Sec. 3.2)			
+ Contrastive Learning (Encoder)	72.45	64.02	51.45
+ Segmentation Constraint (Decoder)	74.53	65.25	53.97
Stage 3: Incomplete Modality Adaptive Fine-tuning (Sec. 3.3)			
+ Feature-Level Consistency (Adapter)	78.12	69.38	59.25
+ Prediction-Level Consistency (Segmentation)	80.64	73.33	63.10
Three-Stage Design vs. Unified Single Stage			
Three-Stage Design	80.64	73.33	63.10
Unified Single Stage	20.32	13.67	10.03

Table 4: Ablation study of the reverse attention adapter.

Models	Average Dice score(%)		
	Whole	Core	Enhancing
UniMRSeg	80.64	73.33	63.10
w/o Rervese Attention	78.28	69.26	60.45
w/o Rervese Attention+Mutual Attention	78.12	68.98	60.23
w/o 3D Swin Transformer+Rervese Attention+Mutual Attention	77.44	68.74	59.24
Fine-tune Encoder in Stage 3	77.05	68.15	58.20

Table 5: Evaluation of different compensation levels.

Input Level (Stage 1)	Feature Level (Stage 2 and Adapter)	Output Level (Segmentation Consistency)	Average Dice score(%)		
			Whole	Core	Enhancing
✓	-	-	69.35	59.89	47.12
-	✓	-	72.46	62.30	50.92
-	-	✓	69.47	57.25	48.31
✓	✓	-	78.12	69.38	59.25
✓	-	✓	74.45	66.40	54.38
-	✓	✓	75.65	67.48	54.95
✓	✓	✓	80.64	73.33	63.10
Baseline			63.31	51.60	38.40



(a) Baseline (b) Stage 1 (c) Stage 2 (d) Stage 3

Figure 5: Average response value of feature maps of each encoding layer under different modal combinations at different stages.

constraints. The latter alone surpasses the stage 2 decoder by 12.0%, showing the importance of enforcing prediction consistency across modalities. Gains in each stage are consistently and gradually improved, with the final model outperforming the baseline by over 44.6% on average.

Three-Stage Design vs. Unified Single Stage. UniMRSeg is built upon a self-supervised pretraining framework, with training initialized from scratch starting at Stage 1. For each task (e.g., BraTS 2020), self-supervised training is conducted solely on the task's own training set, rather than relying on externally labeled datasets such as ImageNet-pretrained weights for initialization. Our three-stage design includes: 1) Pretrain both the encoder and decoder through arbitrary modality reconstruction. 2) Pretrain the encoder with contrastive learning. The segmentation task here is only used to guide the contrastive objective, not as a final goal. 3) Perform the downstream segmentation task. In this way, our three-stage design explicitly aims to reduce the representation gap between complete and incomplete modalities in the encoder-decoder space during the final segmentation stage. If all these tasks are trained jointly in a single-stage model, it would fall into the scope of multi-task learning, rather than self-supervised pretraining.

In multi-task learning, the goal is to leverage multiple complementary clues for collaborative learning. These clues can come from the data level [21, 56, 77], or from structural supervision types [47, 24, 65]. Most approaches often incorporate deliberate designs for task-sharing and task-specific components to enable effective joint prediction across tasks. If the three stages are forcibly merged into a single training stage, two major issues will arise: 1) The model is trained with inputs that involve Random Modality Dropout, Random Modality Shuffle, and Random Spatial Masking. The encoder is supervised using the NT-Xent contrastive loss, an adapter is incorporated, and the model is simultaneously tasked with both segmentation and reconstruction. Such a fully entangled training setup lacks a clear task hierarchy, and there is no explicit coordination among the input, encoder, decoder, and output. 2) When all tasks are treated equally without ordering, joint optimization becomes highly difficult. In the unified single-stage training attempt, the total loss involves six parts. As shown in Tab. 3, we compare the three-stage model with the unified single-stage model. We observed two phenomena during the experiments: 1) The single stage model failed to converge, with loss plateauing early. 2) The optimization process was highly unstable, with different losses fluctuating in turn and failing to decrease consistently together. These phenomena and performace clearly indicate that unified single stage training is unable to achieve effective coordination among the various designs and supervision signals. The lack of a clear training order, combined with competing optimization objectives, leads to mutual interference.

Reverse Attention Adapter. Tab. 4 shows ablations on the reverse attention adapter. There are three key findings: 1) Removing reverse attention alone leads to a 4.2% average drop, confirming its role in semantic compensation for missing modalities. 2) Subsequent elimination of mutual attention yields negligible performance variation, indicating that only emphasizing high-response regions between adapter layers and frozen encoding layers cannot achieve meaningful feature compensation. 3) Replacing the 3D Swin Transformer with a 3D convolution of similar parameters significantly degrades performance, validating the Transformer's advantage in cross-modal correlation modeling. Additionally, we compare freezing vs. fine-tuning the encoder during stage 3. Fine-tuning causes over 6.4% performance drops, indicating that it destroys the adapter-encoder synergy and the compensation mechanism relies on stable encoder representations rather than task-specific tuning.

To further explain this observation, we provide two perspectives: the inheritance between Stage 2 and Stage 3, and the rationale behind the lightweight reverse attention adapter (RAA). 1) Unlike general SSL methods [4, 17, 16] that focus on learning generic representations without targeting specific downstream tasks, our contrastive learning in Stage 2 is task-aware. We co-train the segmentation head to guide the encoder toward modality-invariant features that are directly beneficial for segmentation, rather than general-purpose representations. This task-guided design ensures that the learned contrastive space aligns with the downstream segmentation objective. Therefore, fine-tuning the encoder in Stage 3 would undermine the task-guided contrastive representations that were carefully established. 2) The RAA itself is designed as a residual correction bridging the encoder representation gap between incomplete and complete modality inputs. Formally:

$$f_{\text{inc}} + \mathcal{A}(f_{\text{inc}}) \approx f_{\text{com}}, \tag{6}$$

where f_{inc} denotes encoder features from incomplete modalities, f_{com} represents encoder features from complete modalities, and $\mathcal{A}(\cdot)$ is the learnable adapter. During Stage 3, both f_{inc} and f_{com} are frozen, while only $\mathcal{A}$ is trained. This constrained setup ensures that the adapter focuses purely on compensating missing information, enabling stable and efficient optimization. Once the encoder is unfrozen, however, all three components become variables, making their roles unclear and the optimization unstable. In conclusion, freezing the encoder in Stage 3 is a deliberate and necessary design choice to preserve the task-guided contrastive representations learned earlier. The RAA thus works in synergy with a stable encoder to effectively compensate for missing modalities and maintain robust segmentation performance.

Hierarchical Compensation Mechanism. As shown in Tab. 5, single-level compensations each surpass the baseline. Cross-level combinations reveal nonlinear synergy. Notably, merely integrating input- and feature-level compensations outperforms existing methods in Tab. 1. The fully compensated model achieves 36.2% average gain over baseline across categories, demonstrating consistent collaborative synergy among all three levels without mutual exclusion. To quantify modality representation gaps, Fig. 5 visualizes the average feature activation values across different encoding layers under various modality combinations. Fig. 5a displays the baseline visual distributions, where single-modality, dual-modality, and triple-modality combinations exhibit pronounced representation gaps compared to the full-modality setup. Fig. 5b shows results after stage 1 pre-training, where the compensation capability for single-modality inputs is significantly enhanced. Fig. 5c illustrates the results under the joint constraints of contrastive learning and segmentation tasks, where the representations of all the modalities are further aligned. In Fig. 5d, the reverse attention adapter in stage 3 effectively compensates for cumulative errors from modality reconstruction in stage 1 and spatial discrepancy differentiation in stage 2, using only a small number of parameters.

5 Conclusion

In this work, we propose a novel modality-relax segmentation framework based on a hierarchical self-supervised compensation strategy. Through multi-granularity data perturbation, segmentation task-guided feature distance constraints, and the design of a reverse attention adapter, we simultaneously integrate three different self-supervised techniques into the proposed UniMRSeg model and complete the representation compensation of the missing modality from the input level, feature level, and output level. UniMRSeg is simple yet effective, achieving dominant performance in four different multi-modal image segmentation tasks. We hope that this research paradigm focusing on representation-level compensation can inspire more visual tasks that require modality-relax conditions in the future.

Acknowledgements: This work was supported by the National Natural Science Foundation of China under Grant 62431004, and by Dalian Science and Technology Innovation Foundation under Grant 2023JJ12GX015.

References

[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

[2] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Proceedings of European Conference on Computer Vision*, pages 348–367, 2022.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:834–848, 2017.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of International Conference on Machine Learning*, pages 1597–1607, 2020.

[5] T. Chen, L. Zhu, C. Deng, R. Cao, Y. Wang, S. Zhang, Z. Li, L. Sun, Y. Zang, and P. Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of IEEE International Conference on Computer Vision*, pages 3367–3375, 2023.

[6] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[7] Y. X. Chng, H. Zheng, Y. Han, X. Qiu, and G. Huang. Mask grounding for referring image segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 26573–26583, 2024.

[8] Y. Choi, M. A. Al-Masni, K.-J. Jung, R.-E. Yoo, S.-Y. Lee, and D.-H. Kim. A single stage knowledge distillation network for brain tumor segmentation on limited mr image modalities. *Computer Methods and Programs in Biomedicine*, 240:107644, 2023.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[10] Y. Ding, X. Yu, and Y. Yang. Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proceedings of IEEE International Conference on Computer Vision*, pages 3975–3984, 2021.

[11] R. Dorent, S. Joutard, M. Modat, S. Ourselin, and T. Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 74–82. Springer, 2019.

[12] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of IEEE International Conference on Computer Vision*, pages 4548–4557, 2017.

[13] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2827–2836, 2016.

[14] 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 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[15] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio. Hemis: Hetero-modal image segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477, 2016.

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[19] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[20] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu. Depth saliency based on anisotropic center-surround difference. In *Proceedings of International Conference on Image Processing*, pages 1115–1119, 2014.

[21] 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. In *Proceedings of IEEE International Conference on Computer Vision*, pages 4015–4026, 2023.

[22] A. J. Larrazabal, C. Martínez, J. Dolz, and E. Ferrante. Orthogonal ensemble networks for biomedical image segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 594–603, 2021.

[23] D. Li, B. Yang, W. Zhan, and X. He. Multi-category graph reasoning for multi-modal brain tumor segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 445–455, 2024.

[24] B. Lin, W. Jiang, P. Chen, Y. Zhang, S. Liu, and Y.-C. Chen. Mtmamba: Enhancing multi-task dense scene understanding by mamba-based decoders. In *Proceedings of European Conference on Computer Vision*, pages 314–330, 2024.

[25] H. Liu, D. Wei, D. Lu, J. Sun, L. Wang, and Y. Zheng. M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 1657–1665, 2023.

[26] W. Liu, X. Shen, C.-M. Pun, and X. Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 19434–19445, 2023.

[27] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[28] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[29] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3192–3201, 2022.

[30] Z. Liu, J. Wei, R. Li, and J. Zhou. Sfusion: Self-attention based n-to-one multimodal fusion block. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 159–169, 2023.

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

[32] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). 34(10):1993–2024, 2014.

[33] M. Mistretta, A. Baldrati, L. Agnolucci, M. Bertini, and A. D. Bagdanov. Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. In *Proceedings of International Conference on Learning Representations*, 2025.

[34] Y. Niu, Y. Geng, X. Li, and F. Liu. Leveraging stereopsis for saliency analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 454–461, 2012.

[35] Y. Pang, L. Zhang, X. Zhao, and H. Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *Proceedings of European Conference on Computer Vision*, pages 235–252, 2020.

[36] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2170, 2022.

[37] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[38] Y. Pang, X. Zhao, L. Zhang, and H. Lu. Multi-scale interactive network for salient object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 9413–9422, 2020.

[39] Y. Pang, X. Zhao, L. Zhang, and H. Lu. Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE Transactions on Image Processing*, 32:892–904, 2023.

[40] Y. Pang, X. Zhao, L. Zhang, and H. Lu. Comptr: Towards diverse bi-source dense prediction tasks via a simple yet general complementary transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[41] Y. Pang, X. Zhao, J. Zuo, L. Zhang, and H. Lu. Open-vocabulary camouflaged object segmentation. In *Proceedings of European Conference on Computer Vision*, 2024.

[42] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[43] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji. Rgbd salient object detection: A benchmark and algorithms. In *Proceedings of European Conference on Computer Vision*, pages 92–109, 2014.

[44] 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 *Proceedings of International Conference on Machine Learning*, pages 8748–8763, 2021.

[45] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[46] D. She, Y. Zhang, Z. Zhang, H. Li, Z. Yan, and X. Sun. Eoformer: Edge-oriented transformer for brain tumor segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 333–343, 2023.

[47] H. Shi, S. Ren, T. Zhang, and S. J. Pan. Deep multitask learning with progressive parameter sharing. In *Proceedings of IEEE International Conference on Computer Vision*, pages 19924–19935, 2023.

[48] J. Shi, C. Shang, Z. Sun, L. Yu, X. Yang, and Z. Yan. Passion: Towards effective incomplete multi-modal medical image segmentation with imbalanced missing rates. In *Proceedings of ACM International Conference on Multimedia*, pages 456–465, 2024.

[49] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of IEEE International Conference on Computer Vision*, pages 5311–5320, 2021.

[50] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015.

[51] H. Tang, Z. Li, D. Zhang, S. He, and J. Tang. Divide-and-conquer: Confluent triple-flow network for rgb-t salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[52] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.

[53] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 25:4163–4176, 2022.

[54] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang. Rgb-t image saliency detection via collaborative graph learning. *IEEE Transactions on Multimedia*, 22(1):160–173, 2019.

[55] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023.

[56] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.

[57] Y. Wang, P. Cao, Q. Hou, L. Lan, J. Yang, X. Liu, and O. R. Zaiane. Progressively correcting soft labels via teacher team for knowledge distillation in medical image segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 521–530, 2024.

[58] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang. Multimodal token fusion for vision transformers. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 12186–12195, 2022.

[59] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, volume 2, pages 1398–1402, 2003.

[60] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.

[61] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[62] Z. Wu, D. P. Paudel, D.-P. Fan, J. Wang, S. Wang, C. Demonceaux, R. Timofte, and L. Van Gool. Source-free depth for object pop-out. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1032–1042, 2023.

[63] Z. Xing, L. Yu, L. Wan, T. Han, and L. Zhu. Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 140–150, 2022.

[64] Y. Xu, Y. Bin, J. Wei, Y. Yang, G. Wang, and H. T. Shen. Multi-modal transformer with global-local alignment for composed query image retrieval. 25:8346–8357, 2023.

[65] Y. Yang, P.-T. Jiang, Q. Hou, H. Zhang, J. Chen, and B. Li. Multi-task dense prediction via mixture of low-rank experts. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 27927–27937, 2024.

[66] B. Yin, X. Zhang, Z. Li, L. Liu, M.-M. Cheng, and Q. Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. In *Proceedings of International Conference on Learning Representations*, 2024.

[67] C. Yin and Q. Zhang. A multi-modal framework for robots to learn manipulation tasks from human demonstrations. *Journal of Intelligent & Robotic Systems*, 107(4):56, 2023.

[68] Q. Yu, X. Zhao, Y. Pang, L. Zhang, and H. Lu. Multi-view aggregation network for dichotomous image segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3921–3930, 2024.

[69] Z. Zeng, Z. Peng, X. Yang, and W. Shen. Missing as masking: Arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–433, 2024.

[70] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, and C. C. Loy. Self-supervised scene de-occlusion. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3784–3792, 2020.

[71] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24(12):14679–14694, 2023.

[72] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023.

[73] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023.

[74] Y. Zhang, N. He, J. Yang, Y. Li, D. Wei, Y. Huang, Y. Zhang, Z. He, and Y. Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 107–117, 2022.

[75] X. Zhao, S. Chang, Y. Pang, J. Yang, L. Zhang, and H. Lu. Adaptive multi-source predictor for zero-shot video object segmentation. *International Journal of Computer Vision*, pages 1–19, 2024.

[76] X. Zhao, Y. Pang, Z. Chen, Q. Yu, L. Zhang, H. Liu, J. Zuo, and H. Lu. Towards automatic power battery detection: New challenge benchmark dataset and baseline. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 22020–22029, 2024.

[77] X. Zhao, Y. Pang, W. Ji, B. Sheng, J. Zuo, L. Zhang, and H. Lu. Spider: A unified framework for context-dependent concept understanding. In *Proceedings of International Conference on Machine Learning*, pages 60906–60926, 2024.

[78] X. Zhao, Y. Pang, J. Yang, L. Zhang, and H. Lu. Multi-source fusion and automatic predictor selection for zero-shot video object segmentation. In *Proceedings of ACM International Conference on Multimedia*, pages 2645–2653, 2021.

[79] X. Zhao, Y. Pang, L. Zhang, and H. Lu. Joint learning of salient object detection, depth estimation and contour extraction. *IEEE Transactions on Image Processing*, 31:7350–7362, 2022.

[80] X. Zhao, Y. Pang, L. Zhang, H. Lu, and X. Ruan. Self-supervised pretraining for rgb-d salient object detection. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 3463–3471, 2022.

[81] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang. Suppress and balance: A simple gated network for salient object detection. In *Proceedings of European Conference on Computer Vision*, pages 35–51, 2020.

[82] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang. Towards diverse binary segmentation via a simple yet general gated network. *International Journal of Computer Vision*, 132(10):4157–4234, 2024.

[83] X. Zhao, L. Zhang, and H. Lu. Automatic polyp segmentation via multi-scale subtraction network. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 120–130, 2021.

[84] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang. A single stream network for robust and real-time rgb-d salient object detection. In *Proceedings of European Conference on Computer Vision*, pages 646–662, 2020.

[85] Z. Zhao, J. Li, L. Wang, Y. Wang, and H. Lu. Maskmentor: Unlocking the potential of masked self-teaching for missing modality rgb-d semantic segmentation. In *Proceedings of ACM International Conference on Multimedia*, pages 1915–1923, 2024.

[86] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu. Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images. *IEEE Transactions on Image Processing*, 32:1329–1340, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's contributions and scope are detailed in the abstract and introduction. Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include the limitations of our work in Appendix E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This is not a theoretical paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in Appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have included the implementation details in Sec. 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be released to facilitate further research.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have included the implementation details and the hyperparameters in Sec. 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report average and standard deviation values in our main experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the computing cost in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and understood the code of ethics and have done our best to conform.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work does not present any potential negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This work does not incorporate any large language models as part of the core methods, any LLMs used were only for manuscript editing and formatting.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Table 6: Quantitative comparison of brain tumor segmentation on BraTS2020. ↑ and ↓ indicate that higher and lower scores are better, respectively. **Note:** The reported results of RFNet [10] and mmFormer [74] are taken from the PASSION paper [48], where they are evaluated using the same training and test splits of BraTS2020.

Modality				Dice score(%)											
				Whole Tumor (Whole)				Tumor Core (Core)				Enhancing Tumor (Enhancing)			
Flair	T1	T1ce	T2	RFNet [10]	mmFormer [74]	PASSION-RFNet [48]	UniMRSeg	RFNet [10]	mmFormer [74]	PASSION-RFNet [48]	UniMRSeg	RFNet [10]	mmFormer [74]	PASSION-RFNet [48]	UniMRSeg
○	○	○	●	82.25	83.33	81.35	**84.54**	66.77	63.74	53.69	**68.94**	40.08	38.20	32.13	**47.41**
○	○	●	○	69.27	69.87	75.02	**80.49**	77.24	76.40	77.25	**78.33**	65.81	67.25	69.15	**68.79**
○	●	○	○	68.53	69.60	72.20	**79.72**	59.53	56.00	57.20	**63.32**	31.99	28.66	30.42	**40.12**
●	○	○	○	82.28	83.34	83.95	**84.45**	64.30	61.74	58.60	**66.69**	36.67	33.76	26.40	**41.17**
○	○	●	●	83.94	85.47	82.60	**85.66**	81.90	81.36	78.16	**82.43**	69.18	69.83	69.52	**71.48**
○	●	●	○	74.07	74.74	76.45	**81.46**	81.45	80.50	79.83	**82.06**	68.56	70.70	70.34	**71.38**
●	●	○	○	85.51	86.60	**86.78**	86.27	70.28	67.07	65.57	**72.00**	41.19	38.51	36.62	**48.45**
○	●	○	●	84.90	85.81	82.30	**85.97**	70.39	66.61	61.70	**72.09**	43.61	41.02	37.26	**54.44**
●	○	○	●	86.42	**87.73**	86.85	86.76	70.70	68.66	61.54	**72.34**	43.62	42.17	31.47	**56.44**
●	○	●	○	85.18	**87.39**	85.57	86.54	80.16	79.67	78.38	**81.91**	68.38	68.11	68.60	**71.83**
●	●	●	○	86.61	87.99	87.46	**88.81**	81.67	80.71	79.54	**82.41**	69.47	70.86	72.24	**74.93**
●	●	○	●	87.63	88.60	87.91	**89.87**	72.83	69.89	66.50	**80.69**	45.21	43.22	37.48	**53.49**
●	○	●	●	87.36	88.84	87.45	**89.46**	81.97	81.25	78.88	**83.68**	68.76	69.91	70.18	**74.37**
○	●	●	●	85.40	86.40	83.04	**89.04**	83.28	82.23	80.00	**84.64**	71.04	70.82	69.90	**74.56**
●	●	●	●	88.30	89.27	88.20	**90.45**	82.80	81.90	80.79	**85.26**	69.64	70.62	71.35	**75.96**
Average ↑				82.51	83.67	83.15	**85.97**	75.02	73.18	70.51	**77.12**	55.55	54.91	52.87	**61.65**
Std Dev ↓				6.27	6.45	4.86	**3.26**	7.45	8.48	9.66	**6.95**	14.56	16.24	18.69	**12.81**

Table 7: Quantitative comparison of brain tumor segmentation on BraTS2018. ↑ and ↓ indicate that higher and lower scores are better, respectively. **Note:** The reported results of mmFormer [74], M3AE [25] and M3FeCon [69] are taken from the M3FeCon [69], where they are evaluated using the same training and test splits of BraTS2018.

Modality				Dice score(%)											
				Whole Tumor (Whole)				Tumor Core (Core)				Enhancing Tumor (Enhancing)			
Flair	T1	T1ce	T2	mmFormer [74]	M3AE [25]	M3FeCon [69]	UniMRSeg	mmFormer [74]	M3AE [25]	M3FeCon [69]	UniMRSeg	mmFormer [74]	M3AE [25]	M3FeCon [69]	UniMRSeg
○	○	○	●	81.43	84.22	85.13	**87.10**	64.61	69.14	72.48	**76.45**	41.92	46.93	49.32	**54.20**
○	○	●	○	72.62	75.16	75.26	**82.22**	75.93	**82.53**	82.31	82.25	71.37	73.04	**75.88**	75.46
○	●	○	○	67.92	73.83	75.24	**82.29**	56.96	65.77	65.72	**69.70**	31.38	36.54	44.35	**49.20**
●	○	○	○	86.37	88.04	89.05	**89.85**	61.61	66.02	69.42	**73.52**	37.98	34.96	46.59	**51.35**
○	○	●	●	83.25	85.58	86.61	**88.57**	79.07	83.85	84.75	**84.85**	73.13	74.39	76.55	**76.75**
○	●	●	○	74.75	76.50	79.16	**83.25**	79.01	83.11	82.88	**83.97**	72.77	74.58	76.78	**78.41**
●	●	○	○	87.46	88.49	90.17	**91.45**	66.47	70.53	72.82	**76.72**	41.75	48.16	48.47	**53.20**
○	●	○	●	82.46	86.34	86.03	**88.15**	69.89	71.46	72.01	**77.43**	43.85	44.73	50.17	**55.12**
●	○	○	●	87.99	89.31	90.46	**91.78**	70.18	70.59	72.81	**76.95**	46.47	40.57	51.12	**56.09**
●	○	●	○	87.61	88.85	90.35	**91.98**	78.34	84.03	84.14	**85.20**	73.93	74.08	76.23	**78.43**
●	●	●	○	87.77	88.07	89.62	**89.99**	80.22	83.72	84.81	**85.41**	74.31	73.42	76.74	**78.72**
●	●	○	●	88.08	89.24	90.24	**91.89**	71.97	72.41	75.98	**80.26**	46.51	44.15	52.63	**60.12**
●	○	●	●	88.47	89.46	90.07	**92.01**	79.94	84.25	84.64	**85.47**	74.53	74.68	77.64	**79.02**
○	●	●	●	83.08	85.06	86.79	**90.48**	80.89	84.13	84.39	**85.49**	73.49	73.37	77.33	**79.36**
●	●	●	●	89.93	89.56	90.69	**92.45**	86.23	84.24	85.03	**86.44**	76.36	74.85	77.81	**79.87**
Average ↑				83.28	85.18	86.32	**88.90**	73.42	77.05	78.28	**80.67**	58.65	59.23	63.84	**67.02**
Std Dev ↓				6.37	5.29	5.25	**3.50**	8.02	7.34	6.60	**5.05**	16.51	16.17	14.05	**12.25**

Appendix

A Performance Comparison on BraTS2020 and BraTS2018

To enable fair comparisons with a broader range of methods, we adopt the same data split settings as used in the PASSION paper [48] for BraTS2020, where the dataset is divided into 219 cases for training, 50 for validation, and 100 for testing. For BraTS2018, we follow the data split protocol from M3FeCon [69], using 200 cases for training and 85 for testing. As shown in Tab. 6 and Tab. 7, our method consistently achieves the best average performance and lowest standard deviation.

B Necessity of the Three-Stage Training Strategy

The proposed three-stage training pipeline, comprising Multi-granular Modality Reconstruction, Modality-Invariant Contrastive Learning, and Incomplete Modality Adaptive Fine-tuning, is designed to address the unique challenges of missing modality segmentation, which cannot be effectively tackled using standard end-to-end training.

Unlike single-modality RGB image tasks that benefit from large-scale pre-trained models, many other modalities such as depth, infrared, and medical imaging still lack general-purpose pre-trained encoders. As a result, most models initialized randomly suffer from suboptimal feature representations and unstable performance. Moreover, simple end-to-end optimization struggles to bridge the representation gap between complete and incomplete modality inputs, leading to poor generalization to unseen modality combinations.

To this end, each stage in our training strategy plays a complementary role in strengthening the model's robustness and adaptability:

- Stage 1 (Modality Reconstruction): Enables the model to learn cross-modality structural priors and recover semantically meaningful representations from incomplete inputs.
- Stage 2 (Contrastive Learning): Encourages the network to align features between complete and missing modality inputs, enhancing modality-invariant representations.
- Stage 3 (Unified Fine-tuning): Refines prediction consistency across different modality combinations and ensures deployment under a unified model.
- Each stage serves a distinct and complementary role (input-level structure recovery, feature-level modality alignment, and output-level consistency). Extensive ablation studies (Sec. 4.4) validate that each stage contributes distinct improvements in average accuracy. Without any of the stages, performance consistently degrade.

C Discussion on Training Complexity and Efficiency

Although UniMRSeg adopts a three-stage training strategy, it is deliberately designed for efficiency and simplicity. We analyze the training complexity from two perspectives: concise design and inference simplicity.

Concise Design: All three stages in UniMRSeg share a unified 3D U-Net-style encoder-decoder backbone, integrated with a lightweight 3D ASPP module. No additional complex or stacked modules are introduced. This ensures both architectural clarity and training efficiency. Stage 1 utilizes a masked autoencoding-based reconstruction head. This head operates on top of the shared encoder and does not require extra encoder-decoder branches, keeping the design compact. Stage 2 applies a contrastive loss to already-computed latent features, without introducing new network components. Stage 3 focuses on refining segmentation via a lightweight reverse attention adapter and decoder. The encoder is frozen, making this phase computationally efficient and requiring minimal fine-tuning.

Single-Stage Inference with Unified Weights:

Despite the multi-stage training, inference remains a single-stage process. The final model trained at stage 3 merges all learned representations into a unified network with shared parameters across all modality combinations. Unlike other methods that require ensemble inference [25, 8] or modality-specific branches [69, 74], UniMRSeg performs fast, unified inference without requiring modality recognition or dynamic model selection. As shown in Tab. 8, UniMRSeg achieves comparable computational efficiency to state-of-the-art methods.

Table 8: Efficiency comparison.

Metrics	RFNet [10]	mmFormer [74]	M3AE [25]	UniMRSeg
Parameters (MB)↓	34	106	167	87
FLOPs (G)↓	148	748	248	202

D Modality-Agnostic Robustness via Random Modality Shuffle

Previous multi-modal segmentation methods [69, 74, 10, 25] often rely on a fixed modality-to-channel correspondence, where each input modality (*e.g.*, T1, T2, T1ce, Flari) is assigned to a specific encoder or channel during both training and inference. This design imposes a strong prior assumption: the modality type of each input channel must be known in advance and must strictly align with the model's expected input structure during inference. However, such assumptions significantly limit the scalability and automation of AI-driven medical image analysis pipelines, especially in real-world applications where modality labels may be missing, inconsistent, or ambiguous.

To alleviate this constraint, we introduce a random modality shuffle strategy in stage 1, where the input modalities are randomly permuted at each training iteration. This forces the model to learn modality-invariant representations and reduces its reliance on fixed input orders.

As shown in Tab. 9, even after randomly shuffling the modality-channel order five times during inference, the segmentation performance remains stable with minimal variation. These results demonstrate that the random shuffle not only improves performance (see Tab. 3) but also enhances practical robustness and scalability. This aligns with the broader goal of building fully automated and generalizable medical AI systems.

Table 9: Ablation study on the impact of different modality-channel orders during inference. `Flair-T1-T1ce-T2` denotes the best-performing fixed input order, while `5-Ave` represents the average result of five inference runs using random modality shuffle.

Models	Average Dice score(%)		
	Whole	Core	Enhancing
`Flair-T1-T1ce-T2`	80.64	73.33	63.10
`5-Ave`	80.54	73.27	63.05

E Limitations and Future Work

While the three-stage training strategy significantly improves model generalizability and segmentation accuracy under missing modalities, it inevitably increases the training pipeline complexity. This additional training overhead may pose challenges for practitioners in time-constrained or resource-limited environments. We acknowledge this trade-off and consider streamlining or accelerating the training process as a critical future direction. Several promising avenues to achieve this include: ***I) Curriculum-based Optimization.*** Introducing a curriculum learning schedule that gradually transitions from reconstruction to segmentation may allow for progressive learning in a single stage. ***II) Modality-aware Parameter Sharing.*** Parameter-efficient fine-tuning (*e.g.*, adapters or LoRA) across stages may allow most model weights to be reused, reducing memory and time overhead. ***III) Cross-task Shared Training Objectives.*** Reformulating the three tasks under a shared objective (*e.g.*, information bottleneck or consistency maximization) may unify their gradients and reduce stage-specific training routines. These directions hold the potential to preserve the benefits of multi-stage training while simplifying the optimization process, making the proposed method more accessible to broader applications in multi-modal image analysis.