Scratch Each Other's Back: Incomplete Multi-modal Brain Tumor Segmentation Via Category Aware Group Self-Support Learning

Yansheng Qiu[1,2†] Delin Chen[1,2†] Hongdou Yao[1,2] Yongchao Xu[1,2] Zheng Wang[1,2‡]
[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Wuhan University [2]Hubei Key Laboratory of Multimedia and Network Communication Engineering
{qiuyansheng, wangzwhu}@whu.edu.cn

Abstract

*Although Magnetic Resonance Imaging (MRI) is very helpful for brain tumor segmentation and discovery, it often lacks some modalities in clinical practice. As a result, degradation of prediction performance is inevitable. According to current implementations, different modalities are considered to be independent and non-interfering with each other during the training process of modal feature extraction, however they are complementary. In this paper, considering the sensitivity of different modalities to diverse tumor regions, we propose a Category Aware **G**roup **S**elf-**S**upport Learning framework, called **GSS**, to make up for the information deficit among the modalities in the individual modal feature extraction phase. Precisely, within each prediction category, predictions of all modalities form a group, where the prediction with the most extraordinary sensitivity is selected as the group leader. Collaborative efforts between group leaders and members identify the communal learning target with high consistency and certainty. As our minor contribution, we introduce a random mask to reduce the possible biases. GSS adopts the standard training strategy without specific architectural choices and thus can be easily plugged into existing incomplete multi-modal brain tumor segmentation. Remarkably, extensive experiments on BraTS2020, BraTS2018, and BraTS2015 datasets demonstrate that GSS can improve the performance of existing SOTA algorithms by 1.27-3.20% in Dice on average. The code is released at https://github.com/qysgithubopen/GSS.*

1. Introduction

Magnetic resonance image (MRI) segmentation of brain tumors is becoming increasingly important in clinical evaluation and diagnosis. MRI is designed for different tissues of brain structures and brain tumors with multiple imaging modalities, such as Fluid Attenuation Inversion Recovery (FLAIR), contrast enhanced T1-weighted (T1c), T1-weighted (T1) and T2-weighted (T2). Combining multi-modal images for brain tumor segmentation can significantly improve segmentation accuracy. Most existing methods stitch multi-modal images on channels and input them into the network [18, 46, 27, 30, 10]. However, in clinical practice, the problem of lost modalities is pervasive due to data corruption, various scanning protocols, and unsuitable conditions of the patient [34, 22, 31, 47]. Therefore, there is a great need for a robust multi-modal approach for flexible and practical clinical applications to address the problem of missing one or more modalities.

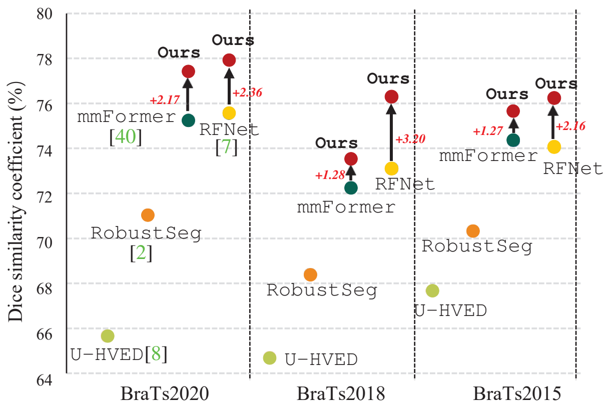

Figure 1. Average accuracy on BraTS2020, BraTS2018, and BraTS2015 datasets. Our GSS enables consistent performance improvements over state-of-the-arts, *i.e.* mmFormer [40], RFNet [7], without bringing any change to base networks during inference.

The current main direction for incomplete medical image segmentation tasks involves multiple stages network to deal with all incomplete modalities cases [11, 8, 7, 2, 40]. This approach considers improving the network's ability to extract features of interest for individual modalities, which plays a pivotal role in the subsequent fusion phase. However, these efforts only focus on learning invariant features and lack inter-modal interactions. It is worth noting

†Equal Contribution
‡Corresponding Author

that Ding *et al.* [7] find out that different modalities contain distinct appearances and thus have different sensitivities to diverse tumor regions. In particular, Flair is more sensitive to the background (BG), and T1c is more sensitive to necrotic, non-enhancing tumor cores (NCR/NET) and GD-enhancing tumors (ET), while Flair and T2 are more sensitive to peritumoral edema (ED), and unfortunately, the above approach does not consider it as a priori information. Consequently, multi-modal feature interaction could be a knowledge transfer process between sensitive and non-sensitive modalities. Fortunately, knowledge distillation is an effective method for addressing this issue. However, existing knowledge distillation-based methods [13, 32, 33, 1, 3, 17], often employ another, more complex model in order to convey complete modalities feature information, resulting in a tremendous computational effort during training. Meanwhile, there is a risk of conveying inaccurate information through direct mutual knowledge distillation between modalities. In addition, if the result after fusion is used for distillation, it can not match the modalities' input states, and model collapse is inevitable.

To alleviate these problems, we propose Category Aware **G**roup **S**elf-**S**upport Learning framework for incomplete multi-modal brain tumor segmentation, which is called **GSS**. As shown in Fig. 2, during the model training phase, we establish self-support groups among students (modalities) instead of the teacher of previous knowledge distillation methods, while no new models and parameters are introduced. Specifically, we establish the groups for each label and, based on how sensitive each modal is to each label, choose one or more top of them to serve as the group leaders. Group leaders can decide on this category with just one affirmative vote. When the group leaders are irresolute, the other group members can make another decision by their voting results to assist in making the right decision. If both group leaders and members are in doubt about this category, then take the category for which estimated minimum probability. Ultimately, based on the votes of the self-support group and after normalization, the soft labels of each category for knowledge distillation are determined. Considering that the decision of the self-help group is positively correlated with the quality of the initial prediction of the model at the early stage of training, which is still flawed at some locations, we introduce a random masking strategy to reduce the possible biases. Overall, our contributions are threefold: 1) We propose Category Aware **G**roup **S**elf-**S**upport Learning framework for incomplete multi-modal brain tumor segmentation. The dominating characteristics of several modalities are utilized to direct the distillation of mutual knowledge between modalities without expanding the complexity of the initial network. 2) In the optimal soft label selection, we set up a novel self-support group, abandoning the direct mutual constraint of modalities through the pseudo-labels generated between each modal, but refining to categories, maximizing the use of information from each modal. 3) Taking advantage of the proposed random mask strategy, GSS could improve the performance of the state-of-the-art segmentation framework on the widely used BraTs2020, BraTs2018, and BraTs2015 benchmarks (Fig. 1).

2. Related Work

Incomplete Multi-modal Brain Tumor Segmentation. Incomplete multi-modal learning, also known as hetero-modal learning [11], aims at designing robust methods with any subset of available modalities at inference [40], which is very common in practical applications, such as scarce annotation [42, 34, 6] and missing modal problems [38, 31, 22, 45]. The incomplete multi-modal brain tumor segmentation task involves segmenting brain tumors from hetero-modal MRI images with various missing components. Therefore, compared with the standard brain tumor segmentation, segmenting brain tumors from incomplete multi-modal data is more practical but challenging. Zhang *et al.* [40] bridged Transformer and CNN to build the long-range dependencies within and across different modalities of MRI images for a modal-invariant representation. Ding *et al.* [7] proposed a Region-aware Fusion Module by aggregating multi-modal features of different regions adaptively to model the relations of modalities and tumor regions. However, it is essential to note that such methods used separate encoders for each modal. There was ground truth that could guide the network to extract similar distributions from incomplete modal features at training time. Nonetheless, these efforts focused on learning shared/invariant features and lacked inter-modal interactions. To address this problem, Zhao *et al.* [44] recently proposed an adaptive feature interaction strategy based on the graph structure, which interacts with multi-modal features to accommodate multi-modal segmentation with different missing modalities. However, this approach introduced a more complex graph structure on the one hand, which increases the computational cost and can not be migrated to other algorithms. On the other hand, the features cannot learn information beyond the ground truth.

In contrast, Our GSS achieves inter-modal information complementarity by selecting an inter-modal consensus optimal learning pseudo-target. In addition, GSS only acts on the training phase of the model and does not require changes to the inference process.

Knowledge Distillation. The concept of knowledge distillation (KD) was first proposed by Hinton *et al.* in [13]. KD defines a learning manner where a bigger teacher network is employed to guide the training of a smaller student network for many tasks [13, 19, 21, 35]. The "dark knowledge" is transferred to students via soft labels from teach-

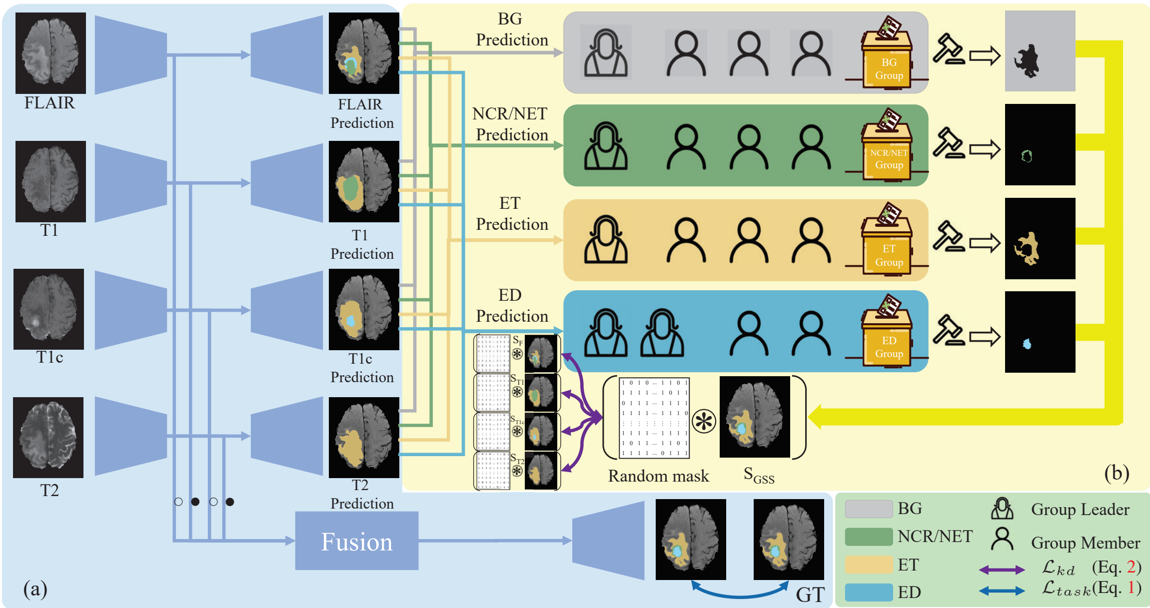

Figure 2. The overall architecture and design details for deploying GSS (b) in existing incomplete brain tumor segmentation methods (a).

ers. For raising the attention on negative logits, the hyper-parameter temperature was introduced. Previous works of logit distillation mainly focus on proposing effective regularization and optimization methods rather than novel methods [26, 26, 4, 29, 16, 15, 23, 14]. Zhang *et al.* [41] proposes a mutual learning manner to train students and teachers simultaneously. Zhang *et al.* [43] provides a novel viewpoint to interpret logit distillation by decoupling the classical KD loss into two parts. Yang *et al.* [37] presents a novel cross-image relational KD to transfer global pixel correlations from the teacher to the student for semantic segmentation. Phan *et al.* [28] proposes a new class similarity weighted knowledge distillation method to eliminate the forgetting of visually similar classes in continual semantic segmentation. However, these methods can not do away with the requirement for a huge computationally intensive teacher network.

There is one major difference between our GSS and previous distillation algorithms: rather than using a larger network as a teacher. It instead employs a self-support student group as a teacher while not introducing new models and parameters.

3. Approach

3.1. Preliminaries

Knowledge Distillation for Segmentation. Unlike traditional image classification, segmentation needs to classify each pixel from N category species to an individual category label. Assume that the input to the network is $\mathbf{F} \in \mathbb{R}^{C\times D\times H\times W}$, where C, D, H and W denote the number of channel, depth, height, and width. The segmentation network transforms the $\mathbf{F}$ into a categorical logit map $\mathbf{S} \in \mathbb{R}^{N\times D\times H\times W}$. The segmentation task loss is to train each pixel with its ground-truth label using cross-entropy:

$$L_{task} = \frac{1}{D\times H\times W}\sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{w=1}^{W} CE(\sigma(\mathbf{S}_{d,h,w}), \mathbf{G}_{d,h,w}). \tag{1}$$

Here, CE denotes the cross-entropy loss, σ denotes the softmax function and $\mathbf{G}_{d,h,w}$ denotes the ground-truth label of the (d, h, w)-th pixel.

Motivated by Hinton's KD [13], a direct method is to align the class probability distribution of each pixel from the student to the teacher. The formulation is expressed as:

$$L_{kd} = \frac{1}{D\times H\times W}\sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{w=1}^{W} KL(\sigma(\frac{\mathbf{S}^s_{d,h,w}}{\tau})||\sigma(\frac{\mathbf{S}^t_{d,h,w}}{\tau})). \tag{2}$$

Here, $\sigma(\mathbf{S}^s_{d,h,w}/\tau)$ and $\sigma(\mathbf{S}^t_{d,h,w}/\tau)$ represent the soft class probabilities of the (d, h, w)-th pixel produced from the student and teacher, respectively. KL denotes the Kullback-Leibler divergence, and τ is a temperature. The overall loss is formulated as:

$$L_{all} = \alpha * L_{task} + \beta * \tau^2 * L_{kd}. \tag{3}$$

where α and β are the balanced parameters, which are set to 0.7 and 0.3 in this paper.

Incomplete Multi-modal Brain Tumor Segmentation Baseline. As shown in Fig. 2 (a), RFNet [7] and mm-Forer [40] as the SOTA paradigm for incomplete modalities consists of four separate encoder-decoders and a fusion module. The complete MRI is used for training the encoder-decoder of each modal separately, and the features extracted by the encoder are fed to the fusion module. A feature map

Algorithm 1 GSS With Single Group Leader Algorithm

Input: L_0: group leader with size of $D \times H \times W$;M_0, M_1, M_2: members with size of $D \times H \times W$;T_L: the threshold of group leader;T_M: the threshold of group members; $T_L \leqslant T_M$

Output: Multi-modal students optimal soft target S_{vote};

set S_{vote} size with $D \times H \times W$, value with 0
for $j = 1; j < D \times H \times W; j++$ **do**
  if $L_{0_j} > T_L$ **then**
    $S_{vote_j} = L_{0_j}$
  else if $M_{0_j} > T_M$ and $M_{1_j} > T_M$ and $M_{2_j} > T_M$ **then**
    $S_{vote_j} = Average(M_{0_j}, M_{1_j}, M_{2_j})$
  else
    $S_{vote_j} = Minimum(L_{0_j}, M_{0_j}, M_{1_j}, M_{2_j})$
  end if
end for

Algorithm 2 GSS With Double Group Leaders Algorithm

Input: L_0, L_1: group leader with size of $D \times H \times W$; M_0, M_1: members with size of $D \times H \times W$; T_L: the threshold of group leader; T_M: the threshold of group members; $T_L \leqslant T_M$; ρ: penalties coefficients

Output: Multi-modal students optimal soft target S_{vote};

set S_{vote} size with $D \times H \times W$, value with 0
for $j = 1; j < D \times H \times W; j++$ **do**
  if $L_{0_j} > T_L$ and $L_{1_j} > T_L$ **then**
    $S_{vote_j} = Maximum(L_{0_j}, L_{1_j})$
  else if ($L_{0_j} \leqslant T_L$ and $L_{1_j} > T_L$) or ($L_{0_j} > T_L$ and $L_{1_j} \leqslant T_M$) **then**
    $S_{vote_j} = Maximum(L_{0_j}, L_{1_j}) - \rho * |L_{0_j} - L_{1_j}|$
  else if $M_0 > T_M$ and $M_1 > T_M$ **then**
    $S_{vote_j} = Average(M_{0_j}, M_{1_j})$
  else
    $S_{vote_j} = Minimum(L_{0_j}, L_{1_j}, M_{0_j}, M_{1_j})$
  end if
end for

of the same size value of 0 is used instead of one or more feature map inputs to simulate the modal missing condition. In the test phase, the missing modal is replaced by an all-0 input, and the decoder is discarded. In this paper, we focus on improving the feature extraction capability of the encoder and, therefore, only act in the training phase of the model.

Multi-modal Knowledge Distillation. Existing incomplete multi-modal brain tumor segmentation methods tend to deal with individual modalities separately, lacking information interaction between modalities. Different modalities have differential sensitivity to different tumor regions. A model's performance can be improved if incomplete modalities acquire information about complete modalities. Zhang *et al.* [39] proposes a self-distillation training technique to improve model performance, arguing that there is still room for progress in knowledge transfer methods within a model. Inspired by this motivation, GSS extracts the optimal knowledge through modal interactions during the training phase of the network, which is passed to each student as a teacher (Fig. 2 (b)). The self-support student group mechanism for multi-modal interactions is developed for each category based on the sensitivity of the different modalities to the tumor region, which is then supplemented by information from the group members to prevent incorrect decisions by the group leaders.

3.2. Multi-modal Self-Support Student Group

In order to accommodate multiple modalities sensitive to a single class at once, we have created two types of group leaders composition.

Single Group Leader. For categories, *i.e.* BG, NCR/NE, and ET, where there is only one sensitive modal, we give the segmentation mask of that modal as the only group leader $L_0 \in \mathbb{R}^{1 \times D \times H \times W}$. The remaining modalities segmentation masks are set as group members $M_0, M_1, M_2 \in \mathbb{R}^{1 \times D \times H \times W}$. For prediction, the sensitive modal is set as group leader. Compared to insensitive members, the prediction of the leader is more accurate at the same threshold. Thus, a sensitive group leader can accurately predict the current class at a relatively low prediction probability T_L, while the requirements T_M for members should be higher. The group leader decides the category when it predicts above T_L,

$$S_{vote_c} = L_0,\ where\ L_0 > T_L. \tag{4}$$

When the group leader's prediction does not reach T_L, the group leader's decision can only be overturned if all members vote above the T_M

$$\begin{aligned} S_{vote_c} = &Averge(M_1, M_2, M_3),\ where\ L_0 <= T_L \\ &and\ M_0 > T_M\ and\ M_1 > T_M\ and\ M_2 > T_M. \end{aligned} \tag{5}$$

The minimum of all modalities should be taken to amplify the self-support group's judgment when both the group leader and members believe the current position is inferior to the category,

$$\begin{aligned} S_{vote_c} = &Minimum(L_0, M_1, M_2, M_3), \\ &where\ L_0 <= T_L\ and\ not\ (M_0 > T_M \\ &and\ M_1 > T_M\ and\ M_2 > T_M). \end{aligned} \tag{6}$$

A detailed description of the GSS with a single group leader algorithm is given in Algorithm 1.

Double Group Leaders. For categories, *i.e.* ED, where there are two sensitive modalities, we give the two modalities as the group leaders L_0, $L_1 \in \mathbb{R}^{1 \times D \times H \times W}$. The remaining modalities are set as group members $M_0, M_1 \in$

$\mathbb{R}^{1\times D\times H\times W}$. It is necessary to select the most appropriate prediction within the group leaders since there are two seats. When the probability of the two group leaders are above T_L, then we consider that the current position must belong to that category, and therefore the value with the highest prediction should be retained,

$$S_{vote_c} = Maximum(L_0, L_1),\ where\ L_{0_j} > T_L\ and\ L_{1_j} > T_L. \tag{7}$$

If only one of L_0 and L_1 reached the T_L, the predicted value for that position was set to the maximum of L_0 and L_1 minus a penalty value,

$$\begin{aligned} S_{vote_c} &= Maximum(L_0, L_1) - \rho * |L_0 - L_1|, \\ &where\ (L_0 \leqslant T_L\ and\ L_1 > T_L)\ or \\ &(L_0 > T_L\ and\ L_1 \leqslant T_M). \end{aligned} \tag{8}$$

The penalty value is used to penalize group leaders for inconsistent results. However, it should not be so large as to potentially reduce the motivation of top students who meet the threshold. This paper sets the penalty value to 30% of the difference between L_0 and L_1. Other cases are processed similarly to Algorithm 1. A detailed description of the GSS With Double Group Leaders Algorithm is given in Algorithm 2.

Even though the selected soft labels are co-optimal across modalities for each category, their distribution across categories is chaotic, and it is easy to have unsmoothed predictions. Therefore, it needs to be normalized before they are used for knowledge distillation [20]. In order not to change the relative difference in predicted probabilities for each category, we use the vector normalization:

$$S_c^t = \frac{S_{vote_c}}{\sum_i^N S_{vote_i}} \tag{9}$$

where S_{vote_c} denotes the soft labels selected for category c by GSS, N denotes the total number of categories, and S_c^t denotes the normalized soft labels for category c.

3.3. Further enhancements

Why GSS need Random Mask? The soft labels elected by the self-support student group are the optimal solution in the current modal interaction. However, in some more challenging areas, the students can not make the same judgments as the teachers. Moreover, such regions are difficult to judge quantitatively. In recent years, several mask-based pre-training methods [12, 36, 25, 9] have consistently demonstrated that randomizing features to mask does not degrade network performance but also improves the robustness of the model. Inspired by them, we use the random mask to discard inaccurate positions of soft labels. Unlike MAE-based methods, the random mask sets all predictions to 0 at a random position (also in the same position as other modalities prediction) to discard inaccurate predictions. Specifically, we randomly generate a 0-1 matrix of size $D \times H \times W$, where the ratio of value 0 to value 1 is α : 1-α. Then N identical matrices are stitched on the channel to obtain $Mask \in R^{N\times D\times H\times W}$. The pixel multiplication of the GSS selected results S_{GSS} and other modalities prediction(S_F, S_{T1}, S_{T1c}, S_{T2}) with $Mask$ respectively provides the input used for distillation:

$$\begin{aligned} &S^t = S_{GSS} * Mask, \\ &S^{sF} = S_F * Mask,\ S^{sT1} = S_{T1} * Mask, \\ &S^{sT1c} = S_{T1c} * Mask,\ S^{sT2} = S_{T2} * Mask. \end{aligned} \tag{10}$$

4. Experiments

4.1. Datasets and Evaluation Metric

In this section we use three different datasets from Multimodal Brain Tumor Segmentation Challenge (BraTS) [24]: BraTs2020, BraTs2018 and BraTs2015. The subjects in the three datasets all contain four distinct MRI modalities, *i.e.*, Flair, T1c, T1 and T2. We split each dataset into train set, validate set and test set, respectively, in the same scheme as RFNet [7]. Dice coefficient [5] is used to measure the segmentation performance of the proposed method, defined as:

$$\text{Dice}_{\bar{k}}(\hat{y}, y) = \frac{2 \cdot \|\hat{y}_{\bar{k}} \bigcap y_{\bar{k}}\|_1}{\|\hat{y}_{\bar{k}}\|_1 + \|y_{\bar{k}}\|_1} \tag{11}$$

Where $\bar{k}$ denotes different tumor classes, including BG, NCR/NE, ED, and ET. Whole tumor, tumor core, and the enhancing tumor are composed of their combinations [7]. $Dice_{\bar{k}}$ denotes the Dice score of the tumor class $\bar{k}$. Larger Dice scores represent that predictions are more similar to the ground truth and thus indicate better segmentation accuracy.

4.2. Implementation Details

Group **S**elf-**S**upport Learning framework (*) is implemented in the official code of RFNet [7] and mmFormer [40], and then compare with their ontologies, U-HVED [8] and RobustSeg [2]. Due to mmFormer's dataset partitioning being different from our reference RFNet, we utilize the official code to retrain on the new dataset partitioning. For a quick and fair comparison, mmFormer is trained with an initial official hyperparameter. In contrast to the official, both mmFormer and our GSS deployment, in which batch size is set as 2, are implemented with PyTorch 1.10 on two Nvidia GeForce RTX 3090Ti GPUs. The GSS reloads the model every 300 epochs of training and starts training from the 0th epoch, reloading it a total of 4 times. In this paper, we use the 300th epoch result of mmFormer as the baseline. When GSS deployment on RFNet, We use the PyTorch environment provided by RFNet officials and train it on two NVIDIA Tesla V100 16GB GPUs.

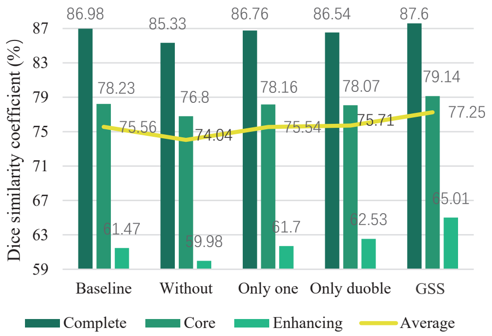

Figure 3. The influence on the number of group leader. Without represents direct mutual distillation.

Operation	Dice(%)		
	Complete	Core	Enhancing
Baseline	86.98	78.23	61.47
fusion dis.	54.63	53.61	52.83
GSS	**87.60**	**79.14**	**65.01**

Table 1. Compare with distillation from fusion result (fusion dis.).

Norm. Opera.	Dice(%)		
	Complete	Core	Enhancing
Baseline	86.98	78.23	61.47
w/o norm.	87.76	79.08	64.48
norm.	**87.60**	**79.14**	**65.01**

Table 2. The influence on the accuracy of the normalization, where 'norm.' denotes normalization operation (Norm. Opera.).

ρ	Dice(%)		
	Complete	Core	Enhancing
0.1	87.63	79.17	65.07
0.2	87.60	79.14	65.01
0.3	**87.63**	**79.25**	**65.32**
0.4	87.60	79.18	65.25
0.5	87.61	79.11	65.30

Table 3. The influence of the penalties coefficients ρ.

T_L	T_M	Dice(%)		
		Complete	Core	Enhancing
0.65	0.65	87.64	79.19	65.21
0.65	0.75	**87.66**	**79.24**	**65.39**
0.65	0.85	87.64	79.15	65.34
0.75	0.75	87.62	79.12	65.12
0.75	0.85	87.60	79.14	65.01
0.85	0.85	87.63	79.11	64.93

Table 4. The influence of the threshold T_L and T_M.

4.3. Ablation Study

For our initial attempts at GSS, we used RFNet [7] as the baseline. When variables are not discussed, the epoch is set to 300, T_L is set to 0.75, T_M is set to 0.85, Temperature τ for L_{kd} is set at 10, and penalties coefficients ρ is set at 0.2. As shown in Fig. 3, the number of group leaders is tailored by the sensitivity of the modal to the label and is better than using single or multiple group leaders directly. However, direct mutual distillation between modalities will reduce the accuracy of the baseline. Moreover, model performance will be drastically reduced if the fusion predictions are used to distill the model predictions for each modal (as is shown in Table. 1). As shown in Table 2, normalization is necessary to smooth out the prediction before distillation for the GSS-selected teachers. Additionally, we perform ablation experiments on the individual hyperparameters introduced by GSS, specifically, when penalties coefficients ρ is set to 0.3, threshold T_L and T_M to 0.65 and 0.75, and temperature τ to 8 the network obtained the optimal solution, which are presented in Table 3, 4, and 5.

τ	Dice(%)		
	Complete	Core	Enhancing
1	87.54	79.11	65.07
2	87.57	79.09	65.12
4	87.59	79.15	65.14
6	87.56	79.09	65.13
8	**87.61**	**79.24**	**65.10**
10	87.60	79.14	65.01

Table 5. The influence on the accuracy of the temperature τ.

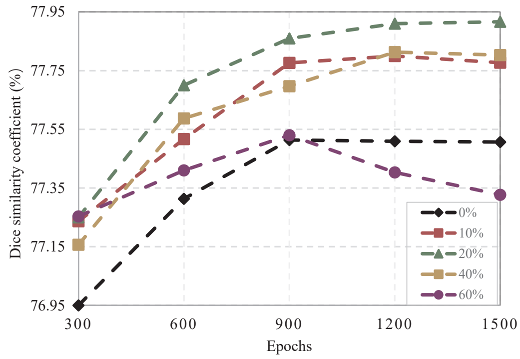

Figure 4. The influence on the random mask rate α.

As shown in Fig. 4, the random mask further improves the performance of the model as the training continues, and its ability gradually decreases when a certain number of iterations is reached, which is since GSS can select the correct predictions at most positions in this time, but most of the predictions are discarded by the random mask, failing to improve the distillation effect. In particular, when the mask rate is 60%, the performance decreases considerably instead, which we believe results from discarding high confidence predictions due to the extensive mask range. The experimental results show that a mask rate with 20% is optimal for GSS. As a final observation, GSS outperforms the baseline in all parameter scenarios, indicating its reliability.

4.4. Comparison with the State-of-the-art Methods

GSS (*) is implemented in the official code of RFNet [7] and mmFormer [40], and then compare with their ontologies, U-HVED [8] and RobustSeg [2]. As shown in Ta-

Modalities				Dice(%)																	
				Complete						Core						Enhancing					
F	T1	T1c	T2	[8]	[2]	[40]	[40]*	[7]	[7]*	[8]	[2]	[40]	[40]*	[7]	[7]*	[8]	[2]	[40]	[40]*	[7]	[7]*
○	○	○	●	81.19	85.49	87.37	**87.89**	86.89	**88.55**	53.40	58.66	62.21	**63.28**	63.81	**65.77**	29.05	37.66	54.71	**56.14**	40.07	**41.48**
○	○	●	○	67.48	71.86	74.33	**76.48**	74.95	**77.90**	68.24	72.87	73.39	**74.70**	72.64	**73.06**	71.54	70.22	69.30	**71.14**	**81.40**	81.00
○	●	○	○	53.58	68.40	74.07	**75.79**	74.20	**77.02**	41.14	50.00	61.73	**64.46**	61.27	**66.30**	19.16	22.67	54.38	**57.63**	29.44	**42.67**
●	○	○	○	83.82	83.02	86.87	**87.64**	86.91	**88.04**	51.37	46.67	56.39	**58.30**	58.71	**63.60**	22.18	28.30	48.95	**51.38**	35.23	**38.86**
○	○	●	●	84.77	87.53	87.92	**89.09**	88.39	**89.39**	73.18	78.46	77.59	**78.93**	77.50	**79.78**	83.54	76.82	73.63	**75.50**	86.97	**86.98**
○	●	●	○	69.65	74.59	77.19	**78.54**	78.13	**80.08**	68.85	76.40	74.65	**75.02**	74.06	**74.29**	76.96	73.95	70.79	**71.45**	**82.48**	82.05
●	●	○	○	85.82	87.66	89.15	**90.12**	88.51	**89.72**	58.39	58.39	65.63	**66.30**	66.88	**70.47**	26.65	35.28	58.31	**59.79**	40.95	**45.42**
○	●	○	●	82.17	87.87	88.33	**89.20**	88.25	**89.36**	57.58	64.88	64.79	**68.00**	67.24	**69.53**	33.94	41.05	58.14	**61.69**	40.58	**46.69**
●	○	○	●	87.74	89.08	90.63	**90.76**	89.62	**90.82**	59.13	63.51	65.98	**66.28**	68.74	**70.64**	30.31	39.72	58.50	**59.41**	44.64	**49.03**
●	○	●	○	87.48	88.01	89.21	**90.44**	88.45	**90.14**	74.27	78.09	78.56	**78.95**	79.30	**81.74**	84.30	76.62	74.83	**75.49**	86.15	**86.84**
●	●	●	○	87.91	87.73	89.98	**90.88**	88.75	**90.28**	75.82	80.68	80.00	**81.02**	80.46	**81.59**	84.33	78.81	76.25	**77.41**	87.30	**87.62**
●	●	○	●	87.59	89.07	90.99	**91.35**	89.93	**90.83**	62.43	65.99	66.94	**67.91**	69.75	**72.07**	33.21	43.04	60.33	**61.59**	44.21	**52.05**
●	○	●	●	89.85	89.06	90.78	**91.58**	90.07	**91.16**	75.10	79.47	79.05	**79.57**	79.29	**81.88**	86.03	78.07	75.05	**75.92**	87.34	**87.64**
○	●	●	●	84.72	88.26	87.92	**89.11**	88.41	**89.45**	74.85	80.84	78.84	**80.52**	79.18	**79.99**	84.03	78.56	74.98	**76.85**	87.47	**87.50**
●	●	●	●	89.79	89.07	90.99	**91.76**	90.49	**91.11**	76.48	81.19	80.28	**80.88**	80.16	**81.72**	86.12	79.13	76.39	**77.15**	87.68	**87.97**
Average				81.57	84.45	86.38	**87.38**	86.13	**87.59**	64.68	69.19	71.07	**72.27**	71.93	**74.16**	56.76	57.33	65.64	**67.24**	64.13	**66.92**

Table 6. Results of state-of-the-art unified models (mmFormer [40], RFNet [7], U-HVED [8], RobustSeg [2]) and the GSS (*) deployment on their basis, on BraTS 2015 dataset. Dice similarity coefficient (DSC) [%] is employed for evaluation with every combination settings of modalities.Complete, Core and Enhancing denote the Dice scores of the whole tumor, the tumor core and the enhancing tumor, respectively.

Modalities				Dice(%)																	
				Complete						Core						Enhancing					
F	T1	T1c	T2	[8]	[2]	[40]	[40]*	[7]	[7]*	[8]	[2]	[40]	[40]*	[7]	[7]*	[8]	[2]	[40]	[40]*	[7]	[7]*
○	○	○	●	80.90	82.24	83.90	**85.88**	84.30	**86.40**	54.10	57.49	66.20	**66.98**	67.62	**69.43**	30.80	28.97	38.81	**38.86**	40.17	**45.76**
○	○	●	○	62.40	73.31	74.77	**77.18**	74.93	**78.47**	66.70	76.83	79.92	**81.15**	80.99	**82.32**	65.50	67.07	72.28	**75.75**	69.43	**77.10**
○	●	○	○	52.40	70.11	74.24	**77.27**	74.68	**78.79**	37.20	47.90	62.26	**64.36**	64.42	**67.47**	13.70	17.29	31.34	**35.53**	34.43	**42.37**
●	○	○	○	82.10	85.69	86.00	**87.01**	86.46	**87.65**	50.40	53.57	60.82	**64.60**	64.89	**68.60**	24.80	25.69	33.47	**36.21**	33.92	**42.88**
○	○	●	●	82.70	85.19	85.48	**86.88**	86.39	**87.94**	73.70	80.20	82.46	**83.26**	83.27	**84.35**	70.20	69.71	73.64	**75.47**	73.01	**79.39**
○	●	●	○	66.80	77.18	78.35	**80.77**	78.59	**81.90**	69.70	78.72	81.82	**82.87**	82.22	**83.71**	67.00	69.06	74.81	**76.65**	70.73	**77.90**
●	●	○	○	84.30	88.24	88.26	**88.86**	88.78	**89.56**	55.30	60.68	68.67	**70.81**	71.59	**73.81**	24.20	32.13	35.96	**40.41**	39.68	**47.34**
○	●	○	●	82.20	84.78	85.35	**86.64**	86.15	**87.48**	57.20	62.19	68.51	**69.85**	70.89	**73.24**	30.70	32.01	40.83	**42.75**	41.42	**48.98**
●	○	○	●	87.50	88.28	88.72	**89.54**	89.12	**89.93**	59.70	61.16	67.90	**69.82**	70.82	**73.38**	34.60	34.60	40.20	**41.82**	43.77	**48.59**
●	○	●	○	85.50	88.51	88.61	**89.35**	89.17	**89.90**	72.90	88.54	81.66	**82.90**	82.94	**83.71**	70.30	70.30	74.09	**76.61**	72.84	**77.84**
●	●	●	○	86.20	88.73	88.54	**89.33**	89.71	**90.25**	74.20	81.06	82.63	**83.26**	83.77	**84.73**	71.10	70.78	74.45	**75.91**	73.17	**78.42**
●	●	○	●	88.00	88.81	89.20	**89.69**	89.68	**90.23**	61.50	64.38	70.24	**71.20**	73.09	**75.37**	34.10	36.41	39.67	**41.93**	44.79	**50.17**
●	○	●	●	88.60	89.27	89.39	**90.12**	90.06	**90.73**	75.60	80.72	82.41	**82.96**	83.54	**84.42**	71.20	70.88	74.08	**76.39**	73.13	**78.69**
○	●	●	●	83.30	86.01	85.78	**87.20**	86.78	**88.04**	86.01	80.33	82.70	**83.29**	83.97	**84.56**	71.10	71.10	74.81	**75.07**	72.56	**78.51**
●	●	●	●	88.80	89.45	89.39	**89.94**	90.26	**90.74**	76.40	80.86	83.03	**83.04**	84.02	**84.61**	71.70	71.70	75.52	**76.41**	73.21	**78.33**
Average				80.10	84.39	85.07	**86.38**	85.67	**87.20**	64.00	69.78	74.75	**76.02**	76.53	**78.25**	50.00	51.02	56.95	**59.03**	57.12	**63.49**

Table 7. Results of state-of-the-art unified models (mmFormer [40], RFNet [7], U-HVED [8], RobustSeg [2]) and the GSS (*) deployment on their basis, on BraTS 2018 dataset. Dice similarity coefficient (DSC) [%] is employed for evaluation with every combination settings of modalities.Complete, Core and Enhancing denote the Dice scores of the whole tumor, the tumor core and the enhancing tumor, respectively.

Modalities				Dice(%)																	
				Complete						Core						Enhancing					
F	T1	T1c	T2	[8]	[2]	[40]	[40]*	[7]	[7]*	[8]	[2]	[40]	[40]*	[7]	[7]*	[8]	[2]	[40]	[40]*	[7]	[7]*
○	○	○	●	80.75	82.20	85.51	**86.55**	86.05	**87.62**	57.43	61.88	63.36	**70.88**	71.02	**72.26**	28.70	36.46	49.09	**49.20**	46.29	**51.28**
○	○	●	○	68.54	71.39	78.04	**79.10**	76.77	**80.14**	73.01	76.68	81.51	**83.68**	81.51	**83.38**	66.59	67.91	78.30	**79.61**	74.85	**78.62**
○	●	○	○	54.93	71.41	76.24	**78.97**	77.16	**79.79**	36.73	54.30	63.23	**66.88**	66.02	**66.39**	12.33	28.99	37.62	**41.40**	37.30	**39.74**
●	○	○	○	82.69	82.87	86.54	**87.95**	87.32	**88.07**	51.15	60.72	64.60	**71.10**	69.19	**72.45**	20.87	34.68	36.68	**42.61**	38.15	**42.29**
○	○	●	●	83.37	85.97	87.52	**88.11**	87.74	**88.70**	77.85	82.44	82.69	**85.56**	83.45	**84.55**	68.74	71.42	77.20	**80.79**	75.93	**80.49**
○	●	●	○	71.58	76.84	80.70	**82.27**	81.12	**83.09**	76.49	80.28	82.81	**85.18**	**83.40**	83.18	67.82	70.11	81.71	**82.44**	78.01	**80.82**
●	●	○	○	85.01	88.10	88.76	**89.65**	89.73	**90.10**	55.10	68.18	71.76	**75.09**	73.07	**73.72**	22.53	39.67	42.98	**46.60**	40.98	**49.50**
○	●	○	●	81.58	85.53	86.94	**87.93**	87.73	**88.72**	85.53	66.46	67.76	**73.59**	73.13	**73.43**	28.73	39.92	49.12	**50.84**	45.65	**53.05**
●	○	○	●	87.40	88.09	89.49	**89.97**	89.87	**90.38**	61.87	68.20	70.34	**74.30**	74.14	**75.66**	30.48	42.19	49.06	**51.57**	49.32	**54.36**
●	○	●	○	86.13	87.33	89.31	**89.38**	89.89	**90.64**	76.86	81.85	83.79	**85.49**	84.65	**85.96**	69.53	70.78	79.44	**81.48**	76.67	**80.99**
●	●	●	○	87.10	88.87	89.79	**90.03**	90.69	**91.08**	79.51	82.76	84.44	**86.07**	85.07	**85.75**	71.32	71.77	80.65	**81.58**	76.81	**82.27**
●	●	○	●	88.07	89.24	89.83	**90.40**	90.60	**91.05**	63.46	70.46	72.42	**75.99**	75.19	**75.69**	30.60	43.90	50.08	**50.73**	49.92	**53.87**
●	○	●	●	88.33	88.68	90.49	**90.51**	90.68	**91.33**	78.68	81.89	83.94	**85.47**	84.97	**86.04**	69.84	71.17	78.73	**80.85**	77.12	**81.14**
○	●	●	●	84.27	86.63	87.64	**88.38**	88.25	**89.01**	79.99	82.85	83.66	**85.64**	83.47	**84.34**	69.74	71.87	77.34	**81.76**	76.99	**81.24**
●	●	●	●	88.81	89.47	90.54	**90.72**	91.11	**91.60**	80.40	82.87	84.61	**85.71**	85.21	**85.75**	70.50	71.52	79.92	**81.59**	78.00	**83.00**
Average				81.24	84.17	86.49	**87.33**	86.98	**88.09**	67.19	73.45	76.06	**79.38**	78.23	**79.24**	48.55	55.49	63.19	**65.54**	61.47	**66.42**

Table 8. Results of state-of-the-art unified models (mmFormer [40], RFNet [7], U-HVED [8], RobustSeg [2]) and the GSS (*) deployment on their basis, on BraTS 2020 dataset. Dice similarity coefficient (DSC) [%] is employed for evaluation with every combination settings of modalities.Complete, Core and Enhancing denote the Dice scores of the whole tumor, the tumor core and the enhancing tumor, respectively.

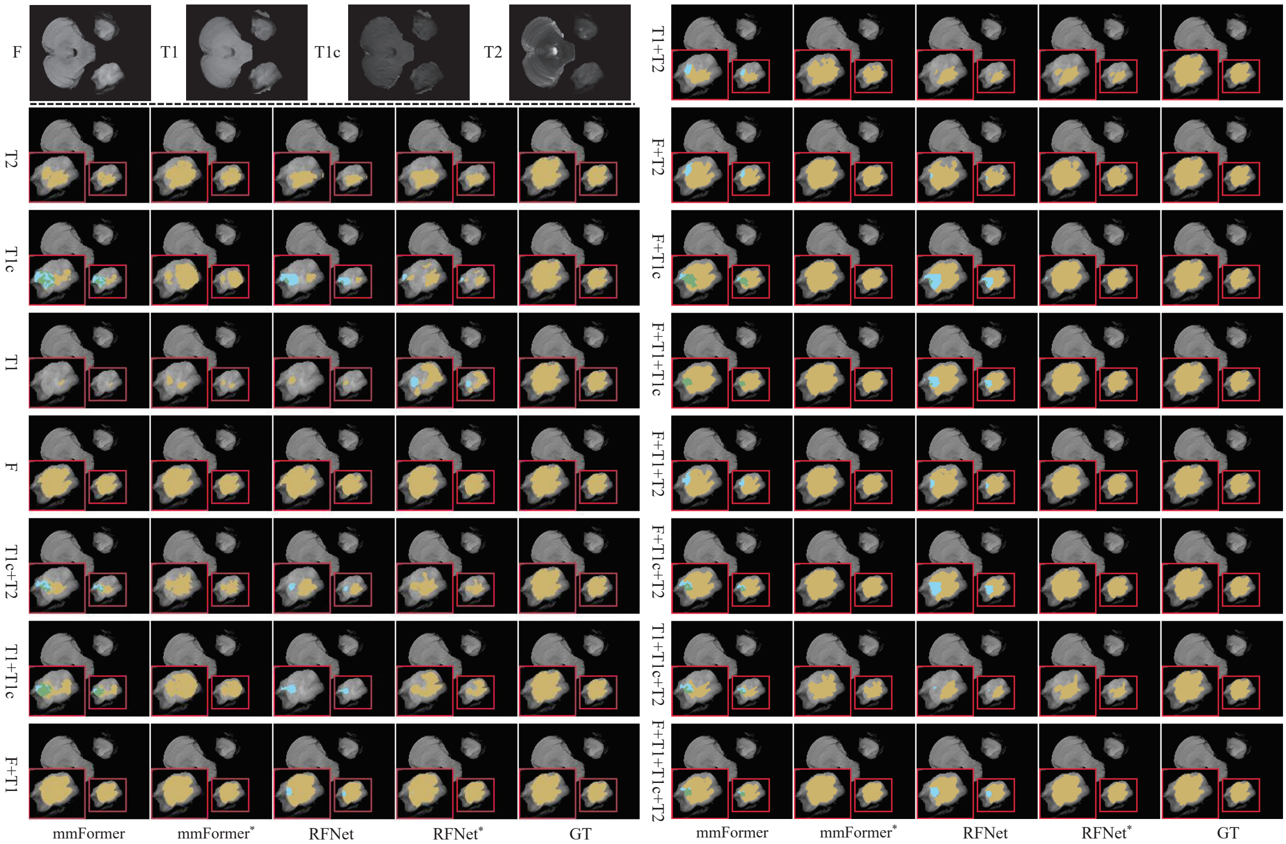

Figure 5. Visualization comparison of different baselines. NCR/NET, ED and ET are illustrated in green, yellow and blue, respectively.

Methods	sensitivity(%)↑			Hausdorff95(mm)↓		
	Com	Cor	Enh	Com	Cor	Enh
mmFormer	**99.61**	99.82	**99.82**	2.78	4.96	3.73
mmFormer*	**99.61**	**99.83**	99.81	**2.41**	**4.13**	**3.28**
RFNet	99.60	**99.84**	99.82	3.47	5.02	4.37
RFNet*	**99.67**	**99.84**	**99.84**	**2.30**	**4.12**	**3.24**

Table 9. Comparisons under two testing criteria on BRATS2020. 'Com', 'Cor' and 'Enh' denotes the whole tumor, the tumor core and the enhancing tumor.

ble 4.4, 4.4 and 4.4, GSS significantly improves the average scores for each category of RFNet [7] and mmFormer [40] in *Dice* on each dataset (BraTs2020, BraTs2018 and BraTs2015) and achieves a new SOTA for the incomplete modalities. In particular, GSS significantly improves the predictive accuracy of baseline for enhancing tumor (*i.e.* maximum 8.52% improvement for RFNet and maximum 5.93% improvement for mmFormer in BraTs2020). Table 9 reports that our GSS(*) also improves the SOTA in terms of sensitivity and Hausdorff distance (95%) on BRATS2020. Specifically, as shown in Fig. 1, GSS can improve the performance of existing SOTA algorithms by 1.27-3.20% in Dice on average. As shown in Fig. 5, compared to RFNet [7] and mmFormer [40], GSS can significantly improve segmentation accuracy and accurately distinguish between ED and other categories, which indicates that GSS can assist models in capturing better spatial context.

5. Conclusion

In this work, we propose a novel category aware group self-support learning framework (GSS) for incomplete multi-modal brain tumor segmentation. As part of our process, we divide the groups according to categories, decide the number of group leaders according to the sensitivity of the modal, and design special algorithms for the different kinds of group divisions to finally identify a pseudo-target for cross-modal knowledge distillation. In order to prevent GSS from failing at certain locations, we introduced the random mask method, which randomly discards these locations during training. Extensive experiments demonstrate GSS could improve the performance of state-of-the-art segmentation framework on the widely used BraTs2020, BraTs2018, and BraTs2015 benchmarks.

Acknowledgements. This work was supported by National Key R&D Project (2021YFC3320301), National Natural Science Foundation of China (62222112, 62176186 and 62171325), Hubei Key R&D Project (2022BAA033) and CAAI-Huawei MindSpore Open Fund.

References

[1] Reza Azad, Nika Khosravi, and Dorit Merhof. Smu-net: Style matching u-net for brain tumor segmentation with missing modalities. *arXiv preprint arXiv:2204.02961*, 2022. 2

[2] Cheng Chen, Qi Dou, Yueming Jin, Hao Chen, Jing Qin, and Pheng-Ann Heng. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 447–456. Springer, 2019. 1, 5, 6, 7

[3] Cheng Chen, Qi Dou, Yueming Jin, Quande Liu, and Pheng Ann Heng. Learning with privileged multimodal knowledge for unimodal segmentation. *IEEE Transactions on Medical Imaging*, 41(3):621–632, 2021. 2

[4] Xu Cheng, Zhefan Rao, Yilan Chen, and Quanshi Zhang. Explaining knowledge distillation by quantifying the knowledge. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12925–12935, 2020. 3

[5] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 5

[6] Yuhang Ding, Xin Yu, and Yi Yang. Modeling the probabilistic distribution of unlabeled data for one-shot medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1246–1254, 2021. 2

[7] Yuhang Ding, Xin Yu, and Yi Yang. Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3975–3984, 2021. 1, 2, 3, 5, 6, 7, 8

[8] Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 74–82. Springer, 2019. 1, 5, 6, 7

[9] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 5

[10] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022. 1

[11] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477. Springer, 2016. 1, 2

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 5

[13] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2, 3

[14] Mengshun Hu, Kui Jiang, Liang Liao, Zhixiang Nie, Jing Xiao, and Zheng Wang. Progressive spatial-temporal collaborative network for video frame interpolation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2145–2153, 2022. 3

[15] Mengshun Hu, Kui Jiang, Liang Liao, Jing Xiao, Junjun Jiang, and Zheng Wang. Spatial-temporal space hand-in-hand: Spatial-temporal video super-resolution via cycle-projected mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3574–3583, 2022. 3

[16] Mengshun Hu, Kui Jiang, Zheng Wang, Xiang Bai, and Ruimin Hu. Cycmunet+: Cycle-projected mutual learning for spatial-temporal video super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[17] Minhao Hu, Matthis Maillard, Ya Zhang, Tommaso Ciceri, Giammarco La Barbera, Isabelle Bloch, and Pietro Gori. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 772–781. Springer, 2020. 2

[18] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 1

[19] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 6356–6364, 2017. 2

[20] Xin-Chun Li, Wen-shu Fan, Shaoming Song, Yinchuan Li, De-Chuan Zhan, et al. Asymmetric temperature scaling makes larger networks teach well again. In *Advances in Neural Information Processing Systems*, 2022. 5

[21] Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, and Zhigeng Pan. Online knowledge distillation for efficient pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11740–11750, 2021. 2

[22] Yanbei Liu, Lianxi Fan, Changqing Zhang, Tao Zhou, Zhitao Xiao, Lei Geng, and Dinggang Shen. Incomplete multimodal representation learning for alzheimer's disease diagnosis. *Medical Image Analysis*, 69:101953, 2021. 1, 2

[23] Xianzheng Ma, Zhixiang Wang, Yacheng Zhan, Yinqiang Zheng, Zheng Wang, Dengxin Dai, and Chia-Wen Lin. Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18922–18931, 2022. 3

[24] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 5

[25] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Voxel-mae: Masked autoencoders for pre-training large-scale point clouds. *arXiv preprint arXiv:2206.09900*, 2022. 5

[26] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020. 3

[27] Himashi Peiris, Munawar Hayat, Zhaolin Chen, Gary Egan, and Mehrtash Harandi. A robust volumetric transformer for accurate 3d tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 162–172. Springer, 2022. 1

[28] Minh Hieu Phan, Son Lam Phung, Long Tran-Thanh, Abdesselam Bouzerdoum, et al. Class similarity weighted knowledge distillation for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16866–16875, 2022. 3

[29] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, pages 5142–5151. PMLR, 2019. 3

[30] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022. 1

[31] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1405–1414, 2017. 1, 2

[32] Saverio Vadacchino, Raghav Mehta, Nazanin Mohammadi Sepahvand, Brennan Nichyporuk, James J Clark, and Tal Arbel. Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images. In *Medical Imaging with Deep Learning*, pages 787–801. PMLR, 2021. 2

[33] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1828–1838, 2020. 2

[34] Shuxin Wang, Shilei Cao, Dong Wei, Renzhen Wang, Kai Ma, Liansheng Wang, Deyu Meng, and Yefeng Zheng. Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9162–9171, 2020. 1, 2

[35] Xiao Wang, Zheng Wang, Wu Liu, Xin Xu, Jing Chen, and Chia-Wen Lin. Consistency-constancy bi-knowledge learning for pedestrian detection in night surveillance. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4463–4471, 2021. 2

[36] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 5

[37] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022. 3

[38] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[39] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019. 4

[40] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 107–117. Springer, 2022. 1, 2, 3, 5, 6, 7, 8

[41] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018. 3

[42] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8553, 2019. 2

[43] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022. 3

[44] Zechen Zhao, Heran Yang, and Jian Sun. Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 183–192. Springer, 2022. 2

[45] Tan Zhi-Xuan, Harold Soh, and Desmond Ong. Factorized inference in deep markov models for incomplete multimodal time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10334–10341, 2020. 2

[46] Chenhong Zhou, Changxing Ding, Zhentai Lu, Xinchao Wang, and Dacheng Tao. One-pass multi-task convolutional neural networks for efficient brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 637–645. Springer, 2018. 1

[47] Tongxue Zhou, Stéphane Canu, Pierre Vera, and Su Ruan. Brain tumor segmentation with missing modalities via latent multi-source correlation representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 533–541. Springer, 2020. 1