ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448218

PATCH-LEVEL KNOWLEDGE DISTILLATION AND REGULARIZATION FOR MISSING MODALITY MEDICAL IMAGE SEGMENTATION

Ruilin Wang[1] *Xiongfei Li*[2] *Mingjie Tian*[3] *Feiyang Yang*[2] *Xiaoli Zhang*[2,*]

[1] College of Software, Jilin University, Changchun 130012, China
[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[3] School of Artificial Intelligence, Jilin University, Changchun 130012, China

ABSTRACT

In the context of medical image segmentation, complementary information among multi-modality images can improve segmentation performance. However, acquiring the complete multi-modality data in clinical settings is difficult. To tackle this problem, we propose a novel multi-modality knowledge distillation segmentation framework, which allows the inference performance of single-modality closer to that of multi-modality. In order to facilitate the extraction of valuable information from the multi-modality teacher network, we first introduce a subtask named patch-selection to distill the patch-level knowledge and improve the generalization capacity of networks simultaneously. Moreover, we employ contrastive learning distillation by defining patch-level positive and negative pairs in embedding, which can encourage the student network to extract more potential information from single-modality input and better understand the similarities and differences with the teacher network in representations. The evaluation process on the BraTS 2018 dataset shows the state-of-the-art performance of our method.

***Index Terms*—** Knowledge distillation, medical image segmentation, missing modality, regularization, contrastive learning

1. INTRODUCTION

The synergy of complementary information from medical images of distinct modalities can significantly enhance the efficacy of medical diagnoses, and subsequently the medical image segmentation as well [1, 2, 3]. Unfortunately, obtaining a comprehensive set of modalities in real clinical scenarios poses considerable challenges because of the differences in scanning equipment, longer scanning time, and other disadvantages. For instance, certain modalities necessitate the administration of contrast agents, which can potentially harm the patient's health [4]. Consequently, there is a pressing need to research a robust segmentation method under the missing modality medical image circumstance.

Several methods have been proposed to deal with the problem of missing modality. This includes synthesizing the missing modality [1, 5, 6], which tends to perform inadequately under the extreme missing modality scenario. Another type of method focuses on mapping the available modalities into a shared latent space and recovering the missing information based on the constructed latent representation [7, 8, 9], these methods are usually ineffective when faced with the absence of more than one modality. The third type of method tries to transfer the knowledge from multi-modality teacher network to single-modality student network by using knowledge distillation [10, 11, 12]. These methods can get better performance when only one modality is available.

Most previous knowledge distillation methods on segmentation primarily focus on pixel-level distillation. In the pursuit of improved results, they often incorporate feature-level distillation as well [10, 11, 12]. However, these methods neglect the significance of distillation at patch-level, which is also an essential aspect in enhancing the performance of distillation because multiple levels distillation tends to offer broader knowledge. Furthermore, they only consider the similarities between single-modality and multi-modality at feature-level while ignoring differences, which has the potential to learn excessively biased or one-sided representations.

In this paper, we propose a novel patch-level knowledge distillation and regularization (PKDR) framework for missing modalities. Building upon the inspiration drawn from multi-task learning and subtask regularization in [14, 15, 16, 17], we first propose patch-selection subtask regularization such that the model could identify patches with specific class pixels. The subtask improves the generalization capacity of the network as additional output, and the patch-level prediction of the teacher model can be transferred to the student model as supplementary knowledge, with the particular focus on pixel location information. To facilitate feature-level distillation, we employ contrastive learning. We define positive and negative pairs at patch-level on the embedding according to the indication of patch-level lable, which can encourage the student model to learn the similarities and differences with the teacher model in representations, thereby extracting more potential

* Corresponding author: zhangxiaoli@jlu.edu.cn

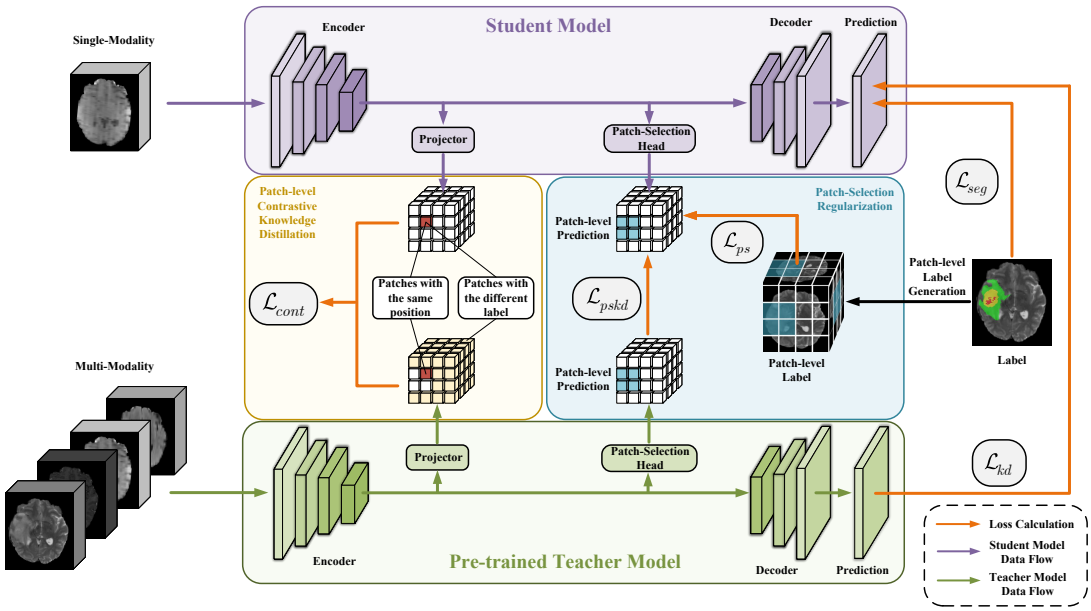

Fig. 1. Overview of our proposed framework. Teacher and Student network have the same architecture and both use VNet [13] except for different inputs. The teacher network is fixed after pre-training using all modalities.

information from single-modality input and finally improving the segmentation performance under the single-modality circumstance.

2. METHOD

The overall framework of our proposed method is shown in Fig. 1. The teacher model takes the multi-modality medical image as input, while the student model accepts one specific modality. The proposed framework aims to transfer the multi-modality knowledge from the teacher model to the single-modality student model, subsequently using only one modality through the student network for inference. The details of our method are as follows.

2.1. Patch-Selection Regularization

We define a subtask named patch-selection as the regularization of the model. Specifically, let $x \in \mathbb{R}^{H\times W\times D\times M}$ and $y \in \mathbb{R}^{H\times W\times D\times C}$ denote the training input with M modalities and corresponding label with C classes respectively. Dividing y into $\frac{H}{P}\times\frac{W}{P}\times\frac{D}{P}\times C$ patches without overlapping, where the patch size is $P\times P\times P$. We assign a label value to each patch. The patch-level label $y_p \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times\frac{D}{P}\times C}$ is generated from y. The definition of y_p is as follows:

$$y_p^{i,c} = \mathbb{1}\left\{N^{i,c} > 0\right\}, \tag{1}$$

where $y_p^{i,c}$ is the label value of i-th patch for class c, and $N^{i,c}$ denotes the number of pixels in i-th patch with the label value of 1 for class c. We add a patch-selection head to the model to output the patch-level prediction. The input of the patch-selection head is the output of the encoder. The binary cross entropy loss $\mathcal{L}_{bce}$ and the dice loss $\mathcal{L}_{dice}$ between the output of patch-selection head y'_p and patch-level label y_p for each class are calculated, which is formulated as follows:

$$\mathcal{L}_{bce} = \sum_{c=1}^{C}\left(-y_{p,c}log y'_{p,c} - (1-y_{p,c})log(1-y'_{p,c})\right), \tag{2}$$

$$\mathcal{L}_{dice} = \sum_{c=1}^{C}\left(1 - \frac{2\times|y'^{*}_{p,c}\cap y_{p,c}|}{|y'^{*}_{p,c}| + |y_{p,c}|}\right), \tag{3}$$

where $y'^{*}_{p,c}$ denotes binary $y'_{p,c}$. The network can detection the regions containing pixels of different classes by adding this subtask, leading to the enhancement in its generalization capacity. The patch-selection regularization loss is formulated as follows:

$$\mathcal{L}_{ps}(y'_p, y_p) = \mathcal{L}_{bce}(y'_p, y_p) + \mathcal{L}_{dice}(y'_p, y_p). \tag{4}$$

2.2. Generalized Pixel-Level Knowledge Distillation

We follow the generalized knowledge distillation in [18] to encourage the student to learn pixel-level knowledge from the teacher. We calculate the binary cross entropy loss shown in Eq. 2 between the predictions of student model and teacher model of each class, in order to adapt the situation in which some classes are mutually inclusive. The generalized pixel-level knowledge distillation loss is defined as:

$$\mathcal{L}_{kd}(p^s, p^t) = \mathcal{L}_{bce}(\sigma(p^s/T_1), \sigma(p^t/T_1)), \tag{5}$$

where p^s and p^t denote the output logits of student model and teacher model respectively. T_1 is the temperature hyper-parameter of pixel-level knowledge distillation and σ is the sigmoid operation.

2.3. Patch-Level Knowledge Distillation

In order to encourage student model to acquire more knowledge from teacher model at multiple levels, we introduce patch-level knowledge distillation with the help of patch-selection subtask. Specifically, the binary cross entropy loss shown in Eq. 2 between the patch-level predictions of student model and teacher model of each class is calculated, which is formulated as follows:

$$\mathcal{L}_{pskd}(p_p^s, p_p^t) = \mathcal{L}_{bce}(\sigma(p_p^s/T_2), \sigma(p_p^t/T_2)), \tag{6}$$

where p_p^s and p_p^t denote the output logits of patch-selection head of student model and teacher model respectively. T_2 is the temperature hyper-parameter of patch-level knowledge distillation.

2.4. Patch-Level Contrastive Knowledge Distillation

Multi-modality representation typically contains more information. For student model, it is significant to close the representation of teacher network, while simultaneously staying away from the representation with substantial disparities, which can better constrain the representation of the student network to converge in the correct direction. Specifically, we binarize the patch-level label to only distinguish the background and foreground, each vector of the embedding is corresponded to a patch of corresponding position. According to the indication of binarized patch-level label, for a vector at a specific position within the embedding of the student model, vectors from the embedding of the teacher model with different patch-level label are considered as negative samples. And the vector at the same position within the embedding of the teacher model is considered as positive sample. To prevent the regularization of the representation learning too heavily, we follow [19] to add a projector, which is a two-layer convolution. The patch-level contrastive knowledge distillation loss is formulated as follows:

$$\mathcal{L}_{cont} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{sim(z_i^s, z_+^t)/\tau}}{e^{sim(z_i^s, z_+^t)/\tau} + \sum_{z_-^t \in NS} e^{sim(z_i^s, z_-^t)/\tau}}, \tag{7}$$

where N denotes the number of patches, z_i^s denotes the i-th embedding vector from the student model, z_+^t is the unique positive sample of z_i^s from the embedding of the teacher model, z_-^t denotes the negative samples of z_i^s from the embedding of the teacher model, NS denotes the set of negative samples of z_i^s, sim denotes the cosine similarity function and τ is a temperature scaling parameter.

We use hybrid loss combining the binary cross entropy loss and the dice loss for medical image segmentation. The image segmentation loss is formulated as follows:

$$\mathcal{L}_{seg}(p, y) = \mathcal{L}_{bce}(p, y) + \mathcal{L}_{dice}(p, y), \tag{8}$$

where p is the segmentation prediction of network, y is the ground truth.

The objective function of our framework for training the student model is:

$$\mathcal{L} = \mathcal{L}_{seg} + \alpha\mathcal{L}_{ps} + \beta(\mathcal{L}_{kd} + \mathcal{L}_{pskd} + \mathcal{L}_{cont}), \tag{9}$$

where α and β are hyper-parameters to balance the loss of subtask regularization item and knowledge distillation item.

3. EXPERIMENTS

3.1. Setup

Dataset. We evaluate our method on the 2018 Brain Tumor Segmentation Challenge (BraTS 2018) dataset [2], which contains 210 HGG (high grade gliomata) and 75 LGG (low grade gliomata) MRI cases including T1, T2, T1ce and Flair modalities. There are three types of annotations including non-enhancing tumor, edema, and enhancing tumor. Following the challenge setting, we amalgamate the various tumor into whole tumor (WT) encompassing all tumor, tumor core (TC) composed of the other two tumor except for edema, and enhancing core (EC). For pre-processing, we normalize each volume and randomly crop each volume to 96 × 128 × 128 to save GPU memory. We split all cases into 70%, 10% and 20% for training, validating, and testing respectively.

Implementation Details. The teacher network is first pre-trained with hybrid loss combining the segmentation loss (Eq. 8) and the patch-selection loss (Eq. 4) for 1000 epochs. We set the batch size to 4 and use Adam optimizer with weight decay $1e^{-5}$. The learning rate is set to $1e^{-3}$ and reduced by multiplying with $(1 - epoch/max\_epoch)^{0.9}$ during the training. Subsequently, we fix the teacher model and train the student model for 1000 epochs using proposed framework. The temperature hyper-parameter T_1 in Eq. 5 and T_2 in Eq. 6 is set to 10 and 5 respectively. In the objective function (Eq. 9), α is set to 0.7 and β is set to 0.5 to balance the various components. The teacher and the student network have the same architecture, both based on the VNet [13].

3.2. Evaluation on Single-Modality Segmentation

To verify the effectiveness of proposed method, we compare it to some baseline models including **Unimodal** baseline, which is trained using single-modality, **U-HVED** [7], **KD-Net** [10], **PMKL** [11] and **ProtoKD** [12]. We use dice similarity coefficient (DSC) as the evaluation metric, which is the most commonly used for evaluating medical image segmentation. The network performs better if it has higher DSC. We choose

Table 1. Dice similarity coefficient (DSC) results on the BraTS 2018. The best result in each modality is **bold-faced**.

Method	T1				T2				T1ce				Flair			
	WT	CO	EC	Avg	WT	CO	EC	Avg	WT	CO	EC	Avg	WT	CO	EC	Avg
Teacher	87.22	79.02	80.06	82.10	-	-	-	-	-	-	-	-	-	-	-	-
Unimodal	72.96	65.59	37.77	58.77	82.65	66.76	45.32	64.91	71.41	73.30	76.36	73.69	81.91	63.57	40.74	62.07
U-HVED [7]	52.40	37.20	13.70	34.43	80.90	54.10	30.80	55.27	62.40	66.70	65.50	64.87	82.10	50.40	24.80	52.43
KD-Net [10]	**79.62**	59.83	33.69	57.72	**85.74**	66.79	33.63	62.05	**78.87**	80.83	70.52	76.74	**88.28**	64.37	33.39	62.01
PMKL [11]	71.31	64.26	41.37	58.98	81.00	67.92	47.09	65.34	70.50	76.92	75.54	74.32	84.11	62.21	41.35	62.56
ProtoKD [12]	74.46	67.34	**47.41**	63.07	81.83	68.29	47.35	65.82	74.67	81.48	76.01	77.39	84.64	65.56	**42.30**	64.17
PKDR (Ours)	74.30	**69.34**	46.47	**63.37**	84.01	**69.18**	**50.66**	**67.95**	76.50	**81.72**	**77.86**	**78.70**	86.92	**67.48**	42.28	**65.56**

Table 2. Comparison of different subtask regularization.

Method	DSC[%]
Baseline	79.95
Baseline + VAE [14]	80.06
Baseline + Boundary-Aware [16]	82.08
Baseline + Fusion [17]	80.83
Baseline + Patch-Selection (Ours)	**82.10**

Table 3. Ablation study on T2 modality.

Method	DSC[%]
w/o Patch-Selection Reg	67.31
w/o Patch-Level KD	67.23
w/o Patch-Level ContrKD	67.72
Full (Ours)	**67.95**

the model with the highest DSC on the validation set to test. The performance of different methods are shown in Table 1.

As presented in Table 1, our method exhibits a remarkable improvement over the unimodal baseline. For the four test modalities, our method leads to a mean DSC increase of 4.60%, 3.04%, 5.01%, and 3.49%, respectively, surpassing the performance of the compared methods. The results demonstrate that incorporating patch-level distillation can provide the student with more knowledge. Moreover, the process of learning the similarities and differences in representation proves to be advantageous, helping the student extract more potential information from single-modality data.

3.3. Evaluation on Subtask Regularization

We compare the proposed patch-selection subtask with other medical image segmentation regularization subtasks. Those subtasks including restoring the original input image with VAE [14], detecting the boundary of different class [16] and multi-modality image fusion [17]. We use the VNet [13] as the baseline and add different subtasks heads separately according to the original papers. The input of the model is all four modalities. As shown in Table 2, our patch-selection regularization achieves the best performance among all methods, which demonstrates the effectiveness of our proposed subtask in improving the generalization capacity of the network.

3.4. Ablation Study

We conduct an ablation study on T2 modality by removing the key components from the proposed PKDR framework. As shown in Table 3, the performance of the model dropped by 0.64%, 0.72% and 0.23% after removing patch-selection regularization, patch-level knowledge distillation and patch-level contrastive knowledge distillation respectively. The results prove that each component in the proposed framework plays a crucial role.

4. CONCLUSION

In this paper, we propose a patch-level knowledge distillation and regularization (PKDR) framework, which is an efficient method to segment the missing modality medical images. We introduce the patch-selection subtask regularization to improve the generalization capacity of the network and encourage the student to learn patch-level knowledge from the teacher. The patch-level contrastive knowledge distillation helps the student to learn the similarities and differences of representations from the multi-modality embedding of the teacher. The experiment results on the BraTS 2018 show the advanced performance of our method.

5. ACKNOWLEDGEMENTS

This work was supported by Natural Science Foundation of Jilin Province (NO. 20220101108JC), Young and Middle-aged Science and Technology Innovation and Entrepreneurship Outstanding Talents (Team) Project (Innovation Category) of Jilin Province (NO. 20230508052RC), the National Natural Science Foundation of China (NO. 20230508052RC), and the National Key Research and Development Project of China (NO. 20230508052RC).

6. REFERENCES

[1] Gijs Van Tulder and Marleen de Bruijne, "Why does synthesized data improve multi-sequence classification?," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2015, pp. 531–538.

[2] Bjoern H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[3] Spyridon Bakas et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[4] 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 *International Conference on Medical Imaging with Deep Learning*. PMLR, 2021, pp. 787–801.

[5] Anmol Sharma and Ghassan Hamarneh, "Missing mri pulse sequence synthesis using multi-modal generative adversarial network," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1170–1183, 2019.

[6] Liyue Shen et al., "Multi-domain image completion for random missing input data," *IEEE Transactions on Medical Imaging*, vol. 40, no. 4, pp. 1113–1122, Apl. 2021.

[7] 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*. Springer, 2019, pp. 74–82.

[8] 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*. Springer, 2019, pp. 447–456.

[9] 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*. Springer, 2022, pp. 107–117.

[10] 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*. Springer, 2020, pp. 772–781.

[11] 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*, vol. 41, no. 3, pp. 621–632, 2021.

[12] Shuai Wang, Zipei Yan, Daoan Zhang, Haining Wei, Zhongsen Li, and Rui Li, "Prototype knowledge distillation for medical segmentation with missing modality," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[13] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision*. IEEE, 2016, pp. 565–571.

[14] Andriy Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *MICCAI BrainLesion Workshop*. Springer, 2019, pp. 311–320.

[15] Ali Hatamizadeh, Demetri Terzopoulos, and Andriy Myronenko, "End-to-end boundary aware networks for medical image segmentation," in *MICCAI 10th International Workshop on Machine Learning in Medical Imaging*. Springer, 2019, pp. 187–194.

[16] Andriy Myronenko and Ali Hatamizadeh, "3d kidneys and kidney tumor semantic segmentation using boundary-aware networks," *arXiv preprint arXiv:1909.06684*, 2019.

[17] Yu Liu, Fuhao Mu, Yu Shi, and Xun Chen, "Sf-net: A multi-task model for brain tumor segmentation in multimodal mri via image fusion," *IEEE Signal Processing Letters*, vol. 29, pp. 1799–1803, 2022.

[18] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[19] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.