Uncertainty-aware Multi-modal Learning via Cross-modal Random Network Prediction

Hu Wang[1], Jianpeng Zhang[2], Yuanhong Chen[1], Congbo Ma[1], Jodie Avery[1], Louise Hull[1], and Gustavo Carneiro[1]

[1] The University of Adelaide, Australia
[2] Northwestern Polytechnical University, China

Abstract. Multi-modal learning focuses on training models by equally combining multiple input data modalities during the prediction process. However, this equal combination can be detrimental to the prediction accuracy because different modalities are usually accompanied by varying levels of uncertainty. Using such uncertainty to combine modalities has been studied by a couple of approaches, but with limited success because these approaches are either designed to deal with specific classification or segmentation problems and cannot be easily translated into other tasks, or suffer from numerical instabilities. In this paper, we propose a new Uncertainty-aware Multi-modal Learner that estimates uncertainty by measuring feature density via Cross-modal Random Network Prediction (CRNP). CRNP is designed to require little adaptation to translate between different prediction tasks, while having a stable training process. From a technical point of view, CRNP is the first approach to explore random network prediction to estimate uncertainty and to combine multi-modal data. Experiments on two 3D multi-modal medical image segmentation tasks and three 2D multi-modal computer vision classification tasks show the effectiveness, adaptability and robustness of CRNP. Also, we provide an extensive discussion on different fusion functions and visualization to validate the proposed model[3].

Keywords: Multi-modal Learning, Uncertainty-aware, Image Segmentation, Image Classification

1 Introduction

Multi-modal data analysis, where the input data comes from a wide range of sources, is a relatively common task. For instance, automatic driving vehicles may take actions based on the fusion of the information provided by multiple sensors. In the medical domain, automated diagnosis often relies on data from multiple complementary modalities. Recently, we have seen the development of successful multi-modal techniques, such as vision-and-sound classification [5], sound source localization [4], vision-and-language navigation [35] or organ segmentation from multiple medical imaging modalities [9,38,40]. However, current

[3] This project received grant funding from the Australian Government through the Medical Research Future Fund - Public Health Research Development Infrastructure PHRDI 000014 Grant and the Australian Research Council through grants DP180103232 and FT190100525.

multi-modal models typically rely on complex structures that neglect the uncertainty present in each modality. Although they can obtain promising results under specific scenarios, they are fragile when facing situations where modalities contain high uncertainties due to noise in the data or the presence of abnormal information. Such issue can reduce their prediction accuracy and limit their applicability in safety-critical applications [14].

Uncertainty is a crucial issue in many machine learning tasks because of the inherent randomness of machine learning processes. For instance, the randomness of data collection, data labeling, model initialization and training are sources of uncertainty that can result in large disagreements between models trained under similar conditions. According to [1,13,18], total uncertainty comprise: 1) aleatoric uncertainty (also known as data uncertainty), representing inherent noise in the data due to issues in data acquisition or labeling; and 2) epistemic uncertainty (i.e., model or knowledge uncertainty), which is related to the model estimation of the input data that may be inaccurate due to insufficient training steps/data, poor convergence, etc. Total uncertainty is defined as:

$$\underbrace{\mathbb{D}_{p(y|x,\theta)}[y]}_{\text{Total Uncertainty}} = \underbrace{\mathbb{E}_{p(\theta|D)}\left[\mathbb{D}_{p(y|x,\theta)}[y]\right]}_{\text{Aleatoric Uncertainty}} + \underbrace{\mathbb{D}_{p(\theta|D)}\left[\mathbb{E}_{p(y|x,\theta)}[y]\right]}_{\text{Epistemic Uncertainty}}, \tag{1}$$

where D indicates the given dataset, x and y are the inputs and outputs of the model, and $\mathbb{D}[\cdot]$ represents the measurement of disagreement (e.g., entropy). The estimation of aleatoric uncertainty is considered as the expectation of the predicted disagreement for each model on data points posterior parameterized by θ; while the epistemic uncertainty is shown by the disagreement of different models parameterized by θ sampled from the posterior. In this paper, we focus on estimating total uncertainty.

In multi-modal methods, existing methods typically assume that each modality contributes equally to the prediction outcome [9,27,33]. This strong assumption may not hold if one of the modalities leads to a highly uncertain prediction, which can damage the model performance. In general, deep learning models that can estimate uncertainty [2, 19, 20] were not designed to deal with multi-modal data. These models are usually based on Bayesian learning that have slow inference time and poor training convergence, or on abstention mechanisms [32] that may suffer from the low representational power of characterising all types of uncertainties with a single abnormal class. Recently, there have been a couple of methods designed to model multi-modal uncertainty [14, 26], but they are limited to work with very specific classification and segmentation problems, or they show numerical instabilities.

In this paper, we propose a novel approach to estimate the total uncertainty present in multi-modal data by measuring feature density via Cross-modal Random Network Prediction (CRNP). CRNP measures uncertainty for multi-modal Learning using random network predictions (RNP) [3], where the model is designed to be easily adaptable to disparate tasks (e.g., classification and segmentation) and training is based on a stable optimization that mitigates numerical instabilities. To summarize, the main contributions of this paper are:

- We propose a new uncertainty-aware multi-modal learning model through a feature distribution learner based on RNP, named as Cross-modal Random Network Prediction (CRNP). CRNP is designed to be easily adapted to disparate tasks (e.g. classification and segmentation) and to be robust to numerical instabilities during optimization.
- This paper introduces a novel uncertainty estimation based on fitting the output of an RNP, which from a technical viewpoint, represents a departure from more common uncertainty estimation methods based on Bayesian learning or abstention mechanisms.

The adaptability of CRNP is shown by its application on two 3D multi-modal medical image segmentation tasks and three multi-modal 2D computer vision classification tasks, where the proposed model achieves state-of-the-art results on all problems. We perform a thorough analysis of multiple CRNP fusion strategies and present visualization to validate the effectiveness of the proposed model.

2 Related Work

2.1 Multi-modal Learning

Multi-modal learning has attracted increasing attention from computer vision (CV) and medical image analysis (MIA). In MIA, Jia et al. [16] introduced a shared-and-specific feature representation learning for semi-supervised multi-view learning. Dou et al. [9] proposed a chilopod-shaped multi-modal learning architecture with separate feature normalization for each modality and a knowledge distillation loss function. In CV, Shen et al. [4] defined a trusted middle-ground for video-and-sound source localization. In video-and-sound classification, Chen et al. [5] proposed to distill multi-modal image and sound knowledge into a video backbone network through compositional contrastive learning. Also in video-and-source classification, Patrick et al. [29, 30] brought the idea of self-supervision learning into multi-modal by training the networks on external data, which boosted classification accuracy greatly. By exchanging channels, Wang et al. [39] showed that the multi-modal features are able to fuse in a better manner. Analyzing existing multi-modal learning methods, even though successful on several tasks, they do not consider that when reaching a decision, some modalities may be more reliable than others, which can damage the accuracy of the model.

2.2 Uncertainty-based Learning Models

Uncertainty also has been widely studied in deep learning. Corbiere et al. [7] proposed to predict a single uncertainty value by an external confidence network via training on the ground-truth class. Sensoy et al. [32] introduced the Dirichlet distribution for an overall classification uncertainty measurement based on evidence. Kohl et al. [19] proposed a probabilistic UNet segmentation architecture to optimize a variant of the evidence lower bound (ELBO) objective. Based on the probabilistic UNet model, Kohl et al. [20] and Baumgartner et al. [2] further updated the model in a hierarchical manner from either the backbone network or prior/posterior networks. Jungo et al. [17] used two medical datasets to compare several uncertainty measurement models, namely: softmax entropy [12], Monte Carlo dropout [12], aleatoric uncertainty [18], ensemble methods [21] and auxiliary network [8,31]. In MIA, multiple uncertainty measurements have been proposed as well [22,24,36,37]. However, none of the methods above are designed for multi-modal tasks and some of them contain long and complex pipelines that are not easily adaptable to new tasks. Bayesian or ensemble-based methods demand long training and inference times and have slow convergence. Evidential methods have drawbacks too, where the main issue is the representational power of the abstention class. In contrast, our proposed model, by introducing random network fitting for cross-modal uncertainty measurement, is not only technically novel, but it is also simple and easily adaptable to many tasks without requiring any restrictive assumption about uncertainty representation.

2.3 Combining Uncertainty and Multi-modal Analysis

Some methods have studied the combination of uncertainty modeling and multi-modal learning. For example, a trusted multi-view classification model has been developed by modeling multi-view uncertainties through Dirichlet distribution and merging multi-modal features via Dempster's Rule [14]. However, it is rigidly designed for classification problems, and cannot be easily translated to other tasks, such as segmentation. Monteiro et al. [26] took pixel-wise coherence into account by optimizing low-rank covariance metrics to apply on lung nodules and brain tumor segmentation. Nevertheless, the method by Monteiro et al. [26] requires a time-consuming step to generate binary brain masks to remove blank areas, and the method is also numerically unstable when training in areas of infinite covariance such as the air outside the segmentation target[4]. From an implementation perspective, this method [26] is also memory intensive when indexing the identity matrix to create one-hot encodings. Differently, in our model, the uncertainty is measured by modeling the overall distribution directly from features without constructing any second-order relation matrix, leading to a numerically more stable optimization and a smaller memory consumption.

3 Cross-modal Random Network Prediction

Below, we first introduce the Random Network Prediction (RNP), with a theoretical justification for its use to measure uncertainties. Then we present the CRNP model training and inference with the cross-modal uncertainty measuring mechanism to take the RNP uncertainty prediction from one modality to enhance or suppress the outputs for other modalities when producing a classification or segmentation prediction.

[4] As stated by SSN implementation [26] at https://github.com/biomedia-mira/stochastic_segmentation_networks.

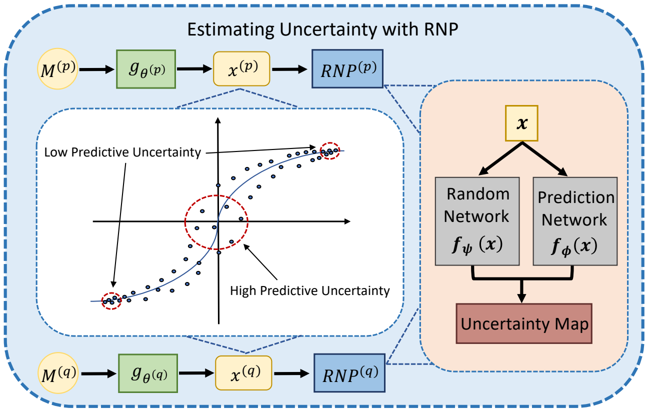

Fig. 1. The input data $M^{(p)}$ and $M^{(q)}$ are first processed by backbone models $g_{\theta^{(p)}}$ and $g_{\theta^{(q)}}$ that produce the features $x^{(p)}$ and $x^{(q)}$. Then the RNP modules have a fixed-weight random network $f_\psi(x)$ and a learnable prediction network $f_\phi(x)$ that tries to fit the output of the random network. The prediction network will fit better (i.e., with low predictive uncertainty) at more densely populated regions of the feature space, as shown in the graph. Hence, the difference between the outputs by $f_\psi(x)$ and $f_\phi(x)$ can be used to estimate uncertainty when processing a test input data.

3.1 Random Network Prediction

The uncertainty of a particular modality is estimated with the RNP depicted in Fig. 1. Specifically, for each RNP, we train a prediction network to fit the outputs of a weight-fixed and randomly-initialized network for feature density modeling. The intuition is that the prediction network will fit better the random network outputs of samples (i.e., with low uncertainty), populating denser regions of the feature space; but the fitting will be worse (i.e., with high uncertainty) for samples belonging to sparser regions. This phenomenon is depicted in the graph inside Fig. 1.

Formally, we consider input images from two modalities $M^{(p)}, M^{(q)} \in \mathcal{M}$, where p and q represent the modalities. After the input image $M^{(p)}$ pass through the encoder $g_{\theta^{(p)}} : \mathcal{M} \rightarrow \mathcal{X}$ (similarly for $g_{\theta^{(q)}}(.)$), the features of the two modalities $x^{(p)}, x^{(q)} \in \mathcal{X} \subset \mathbb{R}^N$ are analyzed by each RNP module. The RNP module feeds $x^{(p)}$ and $x^{(q)}$ to a randomly initialized neural network $f_\psi : \mathcal{X} \rightarrow \mathcal{Z}$, where $\mathcal{Z} \subset \mathbb{R}^M$, with fixed weights $\psi \in \Psi$. Meanwhile, $x^{(p)}$ and $x^{(q)}$ are fed to a learnable prediction network $f_\phi : \mathcal{X} \rightarrow \mathcal{Z}$ with parameters $\phi \in \Phi$. The prediction network has the same output space but a different structure from the random network, where the capacity of f_ϕ is smaller than f_ψ to prevent potential trivial solutions. The cost function used to train the RNP module is based on the mean square error (MSE) between the outputs of the prediction and random networks:

$$\phi^* = \arg\min_{\phi} \sum_{i=1}^{n} \ell_{MSE}(f_\phi(x_i), f_\psi(x_i)) + \mathcal{R}(\phi), \quad (2)$$

where n denotes the number of training samples, $\ell_{MSE}(f_\phi(x_i), f_\psi(x_i)) = \|f_\phi(x_i) - f_\psi(x_i)\|_2^2$, and $\mathcal{R}(\phi) = \|\phi\|_2^2$. The cost function in (2) provides a simple yet pow-

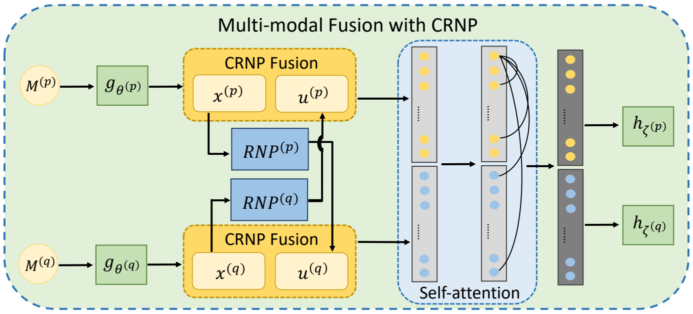

Fig. 2. The overall framework of multi-modal fusion with our CRNP.

erful supervisory signal to enable the prediction network to learn the uncertainty measuring function.

3.2 Theoretical Support for Uncertainty Measurement

The RNP has a strong relation with uncertainty measurement. Let us consider a regression process from a set of perturbed data $\tilde{\mathcal{D}} = \{(x_i, \tilde{y}_i)\}_{i=1}^n$. Considering a Bayesian setting, the objective is to minimize the distance between the ground truth $\tilde{y}_i$ and a sum made up of a generated prior $f_\psi(x_i)$ randomly sampled from a Gaussian and an additive posterior term $f_\phi(x_i)$ with a regularization $\mathcal{R}(\phi)$. Formally, the optimization is as follows:

$$\phi^{(*)} = \arg\min_{\phi} \sum_{i=1}^{n} \|\tilde{y}_i - [f_\psi(x_i) + f_\phi(x_i)]\|_2^2 + \mathcal{R}(\phi), \tag{3}$$

where, according to Lemma 3 in [28], the sum $[f_\psi(x_i)+f_\phi(x_i)]$ is an approximator of the genuine posterior. If we fix the target $\tilde{y}_i$ with zeros, then the objective to be optimized would be equivalent to minimize the distance between the posterior $f_\phi(x_i)$ and the randomly sampled prior $f_\psi(x_i)$. Thus, each output element within the randomized function or the predict function can be viewed as a member of a set of weight-shared ensemble functions [3]. The predicted error, therefore, can be viewed as an estimate of the variance of the ensemble uncertainty.

3.3 Training and Inference of CRNP

This section introduces our proposed CRNP, which fuses the multiple modalities with their inferred uncertainties to produce the final predictions (e.g., classification or segmentation), as shown in Fig. 2. During the multi-modal fusion phase, the features of the two modalities $x^{(p)}$ and $x^{(q)}$ are cross-attended by the uncertainty maps produced by the RNP module from both modalities. The uncertainty map for modality p is represented as:

$$u^{(p)} = \|f_{\phi^{(q)}}(x^{(q)}) - f_{\psi^{(q)}}(x^{(q)})\|_2^2, \tag{4}$$

and similarly for $u^{(q)}$ for modality q. The feature cross-attended by the uncertainty maps is represented by:

$$\tilde{x}^{(p)} = \mathrm{fusion}(x^{(p)}, \hat{u}^{(p)} \odot x^{(p)}), \tag{5}$$

where fusion(., .) represents the operator that fuses the original and cross-attended features, $\hat{u}^{(p)}$ is the channel-wise normalized CRNP uncertainty map, and $\odot$ is the element-wise product operator. $\tilde{x}^{(q)}$ is similarly defined as in (5). Different fusion operations are thoroughly discussed in Sec. 4.5.

We utilize self-attention to further fuse features $\tilde{x}^{(p)}$ and $\tilde{x}^{(q)}$, taking both uni-modal and cross-modal relations between feature elements into consideration. As shown in Fig. 2 , we first concatenate $\tilde{x}^{(p)}$ and $\tilde{x}^{(q)}$ to form the query, key and value inputs for the self-attention module with $Q = K = V = \mathrm{concatenate}(\tilde{x}^{(p)}, \tilde{x}^{(q)})$. Then the output of the self-attention is denoted by:

$$l = \mathrm{softmax}\left(\frac{(QW_q)(KW_k^T)}{\sqrt{d_k}}\right) VW_v, \tag{6}$$

where $l \in \mathcal{L}$, W_q,W_k and W_v are linear projection weights for queries, keys and values, respectively. d_k refers to the dimensions of queries, keys and values. The decoder after the multi-modal fusion is denoted by $h_{\zeta^{(p)}} : \mathcal{L} \to \Delta_{C-1}$ (similarly for $h_{\zeta^{(q)}}$), where $\mathcal{L}$ is the space of the output from the cross-modal RNP module and input to the decoder, and Δ_{C-1} is the classification simplex (output from softmax). Note that although the annotations of multi-modal data are similar, they can have significant differences, particularly in segmentation tasks. Hence, without losing generality, we may need to have multiple separate decoders, one for each modality. But multi-decoders are not needed in tasks where the multi-modal annotation is exactly the same. For segmentation problems, the output of $h_{\zeta^{(p)}}$ is the space Δ_{C-1} per pixel. The training of CRNP alternates the training of the RNP modules using (2) and the training of the whole model. During RNP training, only the weights of the prediction network inside the RNP are updated by minimising (2), and all other CRNP weights are kept fixed. During the training of the whole model, all CRNP weights are updated, except for the weights of the prediction network of the RNP. The whole model training minimizes the multi-class cross-entropy loss for a classification problem or the Dice and element-wise cross-entropy losses for a segmentation model.

During inference, CRNP receives multi-modal inputs, where each modality branch estimates an uncertainty output that will weight the other modality, and the results of both modalities will be fused to produce the final prediction. CRNP works by assigning large weights to the other modality when the current modality is uncertain. When both modalities have large uncertainties, the final prediction will rely on a balanced analysis of both modalities. For the analysis of more than two modalities, the uncertainty map for a particular modality, say p, in (4) is computed by summing the MSE results produced by all other modalities, with $u^{(p)} = \sum_{q \neq p} \|f_{\phi^{(q)}}(x^{(q)}) - f_{\psi^{(q)}}(x^{(q)})\|_2^2$. The decoders $g_{\theta^{(p)}}(.)$ and $g_{\theta^{(q)}}(.)$ from two modalities can be separated or share-weighted, depending on the corresponding output requirements.

4 Experiments

4.1 Datasets

Medical Image Segmentation Datasets. We conduct experiments on two publicly available multi-modal 3D segmentation datasets: Multi-Modality Whole Heart Segmentation dataset (MMWHS) and Multimodal Brain Tumor Segmentation Challenge 2020 dataset (BraTS2020). The MMWHS dataset contains 20 CTs and 20 MRs for training/validation and other 40 CTs and 40 MRs for testing [41]. Seven classes (background excluded) are considered for each pixel. The two modalities have individual ground-truth (GT) for each CT or MR. The BraTS2020 dataset has 369 cases for training/validation and other 125 cases for evaluation, where each case (with four modalities, namely: Flair, T1, T1CE and T2) share one segmentation GT. The evaluation is performed online[5]. Four classes (background included) are considered for each pixel.

Computer Vision Classification Datasets. We also validate our method on three computer vision classification datasets, namely: Handwritten[6], CUB [34] and Scene15 [10]. Each sample of the Handwritten dataset contains 2000 samples from six views and it is a ten-class classification problem, CUB contains 11,788 bird images from 200 different categories. Following Han et al. [14], we also adopt the first ten classes and two modalities (image and text features) extracted by GoogleNet and doc2vec. Three modalities are included in Scene15, which contains 4,485 images from 15 indoor and outdoor classes.

4.2 Implementation Details

Medical Image Segmentation Tasks. To keep a fair comparison, the implementation of all models evaluated on MMWHS and BraTS2020 is based on the 3D UNet (with 3D convolution and normalization) as our backbone network. On MMWHS, we adopt the official test set proposed by Zhuang et al. [41] (40 CTs and 40 MRs) for testing; on BraTS2020, we evaluate all models on the online validation set. For overall performance evaluation, the models were trained for 100,000 iterations on MMWHS and 180,000 iterations on BraTS2020 without model selection. Following Dou et al. [9], our hyper-parameter tuning and ablation are conducted on MMWHS with 16 CTs and 16 MRs for training, 4 CTs and 4 MRs for validation. The batch size is set to 2. Stochastic gradient descent optimizer with a momentum of 0.99 is chosen for the model training. The initial learning rate is set to 10^{-2} on both datasets with cosine annealing [23] learning rate tuning strategy. For the reproduction of Probability UNet [19], we use prior/posterior mean instead of random sampling a latent variable z for prediction. The evaluation of the methods is based on the Dice score and Jaccard index for MMWHS; and the Dice score and Hausdorff95 index for BraTS2020. For cross-modal RNP modules training, the randomized network is made up of

[5] https://ipp.cbica.upenn.edu/categories/brats2020
[6] https://archive.ics.uci.edu/ml/datasets/Multiple+Features

3 depth-wise convolutional hidden layers; the prediction network has 2 depth-wise convolutional hidden layers. Between every two layers, both the randomized network and the prediction network adopt Leaky-ReLU as their activation function, where the negative slope is set to 2.5×10^{-1}. We set 256 as RNP output dimension for both tasks. For performance evaluation, the CRNP is placed at the bottleneck of our 3D UNet backbone. For the ensemble version of CRNP on both datasets, following Wang et al. [40], we average the logits of 3 CRNP models to reduce the prediction variance.

Computer Vision Classification Tasks. For the model evaluation on computer vision datasets, we follow [14] to split the data into 80% for training and 20% for testing. To keep a fair comparison, we uniformly trained all models for 500 epochs without model selection and then evaluated them on the test set. The learning rate is set to 3×10^{-4}; Adam optimizer with 1×10^{-5} weight decay and coefficients (0.9, 0.999) are adopted. Following Han et al. [14], we apply accuracy and multi-class AUROC as evaluation metrics. We used similar setups for cross-modal RNP modules as on the medical data, with the following differences: the RNP output dimension is set to 32 for computer vision classification tasks and CRNP is placed at the layer before the fully connected layer. The training of CRNP model is conducted in an end-to-end manner without any pre-training or post-processing. Also, the hyper-parameters do not require much effort to tune.

4.3 Medical Image Segmentation Model Performance

Performance on MMWHS Dataset. We compare our approach with: Individual (CT or MR single modality segmentation with separate 3D UNet), 3D UNet (multi-modal fusion by concatenation), the multi-modal learning model Ummkd [9], and the uncertainty model Probability UNet[7] [19], which proposes a prior net to approximate the posterior distribution, combining the knowledge of inputs and ground truth, in a latent space. The evaluation is based on the Dice scores of the segmentation of the left ventricle blood cavity (LV), the myocardium of the left ventricle (Myo), the right ventricle blood cavity (RV), the left atrium blood cavity (LA), the right atrium blood cavity (RA), the ascending aorta (AA), the pulmonary artery (PA) and Whole Heart (WH). All results on MMWHS data are obtained by using the official evaluation toolkit[8].

As shown in Tab. 1, our proposed CRNP and its ensemble version have 7 out of the 8 best Dice results on both CT and MR. On CT (Tab. 1), CRNP raises LV Dice score from 0.9297 to 0.9369 and PA Dice score from 0.8425 to 0.8628, when compared to the second-best models. On whole heart segmentation Dice score, CRNP outperforms the second-best model by 1.9%. The ensemble version of CRNP further improves segmentation accuracy. A similar result is observed on MR. On LV, CRNP raises the Dice score from 0.8850 to 0.8962 and AA Dice score from 0.8551 to 0.8736 when compared to the second-best models.

[7] We also tried SSN [26], but it requires the creation of one-hot encodings that are memory intensive for seven classes on MMWHS dataset.

[8] http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs/

Table 1. The performance of different models on CT/MR segmentation of MMWHS dataset. The best results for each column within either CT or MR section are in bold. $*$ indicates the result with the ensemble model.

	Models	LV	Myo	RV	LA	RA	AA	PA	WH
CT	Individual	0.9297	0.8943	0.8597	0.9254	0.8701	0.9335	0.7833	0.8989
	3D UNet	0.9138	0.8781	0.8822	0.9274	0.8680	0.9088	0.8239	0.8957
	Ummkd	0.9145	**0.9066**	0.8410	0.9157	0.8853	0.8928	0.7579	0.8734
	Prob-UNet	0.9071	0.8775	0.8978	0.9262	0.8657	0.9318	0.8425	0.8997
	CRNP (Ours)	0.9369	0.9036	0.9076	**0.9375**	0.8885	**0.9538**	0.8628	0.9187
	CRNP* (Ours)	**0.9373**	0.9060	**0.9085**	0.9366	**0.8910**	0.9503	**0.8629**	**0.9193**
MR	Individual	0.8777	0.7923	0.6146	0.5686	0.7528	0.5854	0.3993	0.6729
	3D UNet	0.8850	0.7723	0.8559	0.8548	0.8676	0.8551	0.7964	0.8535
	Ummkd	0.8721	**0.7966**	0.8086	0.8577	0.8278	0.7998	0.7224	0.8211
	Prob-UNet	0.8742	0.7389	0.8332	0.8495	0.8531	0.8537	0.7895	0.8386
	CRNP (Ours)	0.8962	0.7787	0.8605	0.8637	**0.8748**	**0.8736**	0.7969	0.8615
	CRNP* (Ours)	**0.8963**	0.7811	**0.8742**	**0.8850**	0.8688	0.8692	**0.8329**	**0.8758**

Table 2. The performance comparison of CRNP and different challenge models on both CT and MR segmentation of MMWHS dataset. The best results for each column are in bold. ↑ sign indicates the higher value the better.

	CT		MR	
Models	Dice ↑	Jaccard ↑	Dice ↑	Jaccard ↑
GUT	0.9080	0.8320	0.8630	0.7620
KTH	0.8940	0.8100	0.8550	0.7530
CUHK1	0.8900	0.8050	0.7830	0.6530
CUHK2	0.8860	0.7980	0.8100	0.6870
UCF	0.8790	0.7920	0.8180	0.7010
SIAT	0.8490	0.7420	0.6740	0.5320
UT	0.8380	0.7420	0.8170	0.6950
UB1	0.8870	0.7980	0.8690	0.7730
UB2	-	-	0.8740	0.7780
UOE	0.8060	0.6970	0.8320	0.7200
Ours	**0.9193**	**0.8486**	**0.8758**	**0.7814**

On whole heart segmentation, CRNP increases MR Dice from 0.8535 to 0.8615. Model ensemble further improves the performance.

Interestingly, the Individual model obtains accurate results on CT (0.8989 for WH score). However, performance (0.6729 for WH score) drops drastically on MR evaluation, with particularly poor accuracy on RV, AA and PA. But when considering both modalities (3D UNet model), the model performance increases substantially. This shows the bounds of considering a single modality, especially for MR segmentation. The proposed CRNP outperforms the 3D Unet by a large margin. Ummkd [9] performs consistently well on Myo on both CT and MR. We hypothesize that the domain-specific normalization and knowledge distillation loss contribute more to Myo segmentation than to other organs. Probability UNet tries to model posterior latent space rather than a deterministic prediction, which may explain its performance. In general, we note that the CT segmentation results are better than MR, which resonates with the conclusion from [41].

From the number of parameters perspective, the randomized network is made up of 3 convolutional hidden layers and the prediction network has 2 convolutional hidden layers. So the change in number of parameters is minimal. More specifically, the number of parameters of competing methods are: 1) UNet:

Table 3. The performance of different models on BraTS2020 Online validation set. The best results for each column are in bold. $*$ indicates models with ensemble. ↑ sign indicates the higher value the better; while ↓ means the lower value the better.

	Dice ↑			Hausdorff95 ↓		
Models	ET	WT	TC	ET	WT	TC
3D UNet [6]	0.6876	0.8411	0.7906	50.9830	13.3660	13.6070
Basic VNet [25]	0.6179	0.8463	0.7526	47.7020	20.4070	12.1750
Deeper VNet [25]	0.6897	0.8611	0.7790	43.5180	14.4990	16.1530
Residual 3D UNet	0.7163	0.8246	0.7647	37.4220	12.3370	13.1050
ProbUNet [19]	0.7392	0.8782	0.7955	36.2458	6.9518	7.7183
SSN [26]	0.6795	0.8420	0.7866	43.6574	14.6945	19.5171
Modal-Pairing* [40]	0.7850	0.9070	0.8370	35.0100	4.7100	5.7000
TransBTS [38]	0.7873	0.9009	0.8173	**17.9470**	4.9640	9.7690
CRNP (Ours)	0.7887	0.9086	0.8372	26.5972	**4.0490**	6.0040
CRNP* (Ours)	**0.7902**	**0.9109**	**0.8550**	26.4682	4.1096	**5.3337**

41.05M, 2) Ummkd (with UNet backbone for fair comparison): 41.05M, and 3) ProbUNet: 57.44M. Our CRNP has 42.18M parameters, where the RNP module has 0.29M, and the attention module has 0.84M parameters.

We also compare the proposed CRNP model with the state-of-the-art models reported by the official challenge report [41]. The results are shown in Tab. 2. On whole heart segmentation, CRNP has a particularly accurate Dice score and Jaccard index for CT and MR. Compared to the second-best models, our CRNP model increases the Dice score from 0.9080 to 0.9193 and from 0.8740 to 0.8758 on CT and MR, respectively. Similar results are shown forJaccard index.

Performance on BraTS2020 dataset. Developing automated segmentation models to delineate intrinsically heterogeneous brain tumors is the main goal of BraTS2020 Challenge. Following [38], we compare the proposed CRNP model with many other strong methods, including 3D UNet [6], Basic VNet [25], Deeper VNet [25], Residual 3D UNet, Modal-Pairing [40], TransBTS [38], as well as uncertainty-aware models ProbUNet [19] and SSN [26] that models aleatoric uncertainty by considering spatially coherence. We evaluate the Dice and Hausdorff95 indexes of all models on four organs: enhancing tumor (ET); tumor core (TC) that consists of ET, necrotic and nonenhancing tumor core; and whole tumor (WT) that contains TC and the peritumoral edema.

In Tab. 3, our models have 5 out of the 6 best results. The CRNP improves the ET Dice score, compared with the second-best model, from 0.7873 to 0.7887; and from 0.9070 to 0.9086 on WT. Similar results are shown on Hausdorff95 indexes. Note that the Modal-Pairing model adopts an ensemble strategy. When applying the ensemble strategy to CRNP, the results improved even further. The WT Dice of CRNP* can reach 0.9109; the TC Dice can reach 0.8550, which is one more percent increment; and improves the TC Hausdorff95 to 5.3337. The performance improvements show the effectiveness of the proposed CRNP model.

4.4 Computer Vision Classification Model Performance

In this section, we show results that demonstrate the effectiveness of CRNP on multiple CV classification tasks. The evaluation metrics include accuracy

Table 4. The performance of different models on computer vision classification datasets. The best results for each row are in bold.

Data	Metric	MCDO [11]	DE [21]	UA [15]	EDL [32]	TMC [14]	CRNP
Handwritten	Acc	0.9737	0.9830	0.9745	0.9767	0.9851	**0.9925**
	AUROC	0.9970	0.9979	0.9967	0.9983	**0.9997**	0.9996
CUB	Acc	0.8978	0.9019	0.8975	0.8950	0.9100	**0.9167**
	AUROC	0.9929	0.9877	0.9869	0.9871	0.9906	**0.9961**
Scene15	Acc	0.5296	0.3912	0.4120	0.4641	0.6774	**0.7057**
	AUROC	0.9290	0.7464	0.8526	0.9141	0.9594	**0.9734**

and multi-class AUROC on Handwritten, CUB and Scene15 datasets. Following Han et al. [14], the comparison models include multiple uncertainty-aware models: Monte Carlo dropout (MCDO) [11] that adopts dropout at inference as a Bayesian approximator; deep ensemble (DE) [21], which uses an ensemble strategy to reduce uncertainty; uncertainty-aware attention (UA) [15] that creates uncertainty attention maps from a learned Gaussian distribution; evidential deep learning (EDL) [32] that predicts an extra Dirichlet distribution for all logits based on evidence; and trusted multi-view classification (TMC) [14], which is a multi-view version of EDL.

As shown in Tab. 4, CRNP model can outperform its counterparts on 5 out of 6 measures across datasets. CRNP performs particularly well on Scene15, increasing the accuracy from 0.6774 to 0.7057 (a 2.83% improvement) and AUROC from 0.9594 to 0.9734 (a 1.4% improvement). CRNP also has promising results on Handwritten and CUB data. On AUROC of Handwritten, CRNP gets slightly worse but comparable results than TMC (0.9996 vs. 0.9997).

4.5 Ablation Study

Effectiveness of Each Component In the ablation study, we examine each component of the proposed CRNP. The "Base" model is the plain multi-modal 3D UNet with dual branches; "CA" means cross-attention by assigning the query from one modality, while keep the key and value the other modality; "SA" means applying self-attention as we propose. We conducted the ablation on the validation set split of the MMWHS dataset and we measured the average Dice scores of each organ on CT and MR. As shown in Tab. 5, compared with the Base 3D UNet model, the CRNP model is able to improve (around 1% increment of Dice scores) the performance across multiple organs, where the improvements are especially obvious on Myo, LA, RA, AA and WH. From the table, we can perceive that, with the help of either cross-attention or self-attention, the model performance can be further boosted. But applying the self-attention as described in Sec. 3.3, causes the model to produce the best results (6 best results out of 8) across multiple organs. This is mainly because the self-attention on the multi-modal feature fusion not only models the cross-modal relations, but also considers uni-modal attentions.

Discussion of Different CRNP Fusion functions In terms of different CRNP fusion functions that can be applied in fusion(., .) (Sec. 3.3), we compare and discuss three types, as shown in Tab. 6: (a) "Replace" represents a

Table 5. Ablation study on MMWHS dataset. Best results per row are in bold.

Models	LV	Myo	RV	LA	RA	AA	PA	WH
Base	0.9334	0.8596	0.8876	0.8932	0.8794	0.8239	0.8168	0.8706
CRNP	0.9324	0.8685	0.8644	0.9007	0.8957	**0.9216**	0.8225	0.8865
CRNP+CA	0.9323	0.8683	0.8802	0.9147	**0.9116**	0.9098	0.8194	0.8909
CRNP+SA	**0.9356**	**0.8891**	**0.8814**	**0.9232**	0.8987	0.9148	**0.8277**	**0.8958**

Table 6. Analysis of different fusion functions of CRNP on MMWHS dataset. Best results per row are in bold.

Models	LV	Myo	RV	LA	RA	AA	PA	WH
Replace	**0.9342**	**0.8688**	0.8688	0.897	0.8812	0.9074	0.8128	0.8815
Concat	0.9327	0.8676	**0.8798**	**0.9031**	0.8781	0.9098	0.8042	0.8822
Residual	0.9324	0.8685	0.8644	0.9007	**0.8957**	**0.9216**	**0.8225**	**0.8865**

naive replacement of the original modality features by the uncertainty map attended features; (b) "Concat" applies the concatenation operation on the original modality features and the uncertainty map attended features; and (c) "Residual", which is the default fusion strategy of the proposed CRNP, denotes an addition operation performed between two feature tensors. This experiment is conducted on the MMWHS dataset and averages both CT and MR Dice results. From the results, we note that all three types of fusion functions have pros and cons. However, the "Residual" model performs better (4 best results out of 8) than other functions. This advantage is more noticeable on RA, AA and PA, on which more than 1% improvement is gained on Dice score.

4.6 Visualization

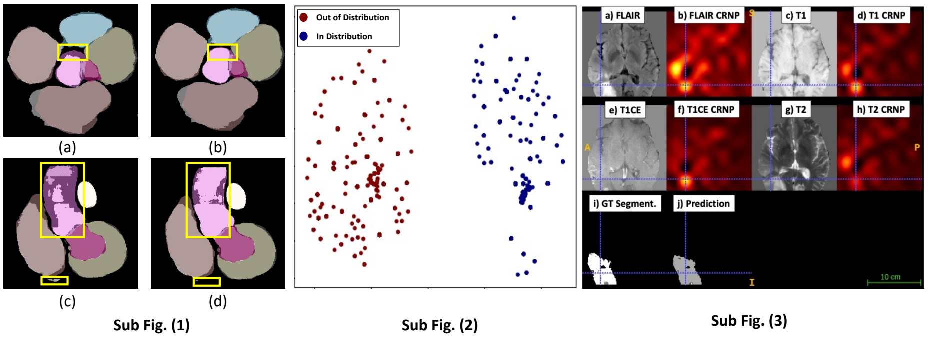

Fig. 3. Visualization experiments of CRNP. Sub Fig.(1) shows a comparison between the segmentation of the proposed CRNP ((b) and (d)) and its Base model ((a) and (c)). Sub Fig.(2) shows the T-SNE graph of the in and out of distribution data points produced by the cross-modal RNP module. In the Sub Fig.(3), we show the CRNP uncertainty heat-maps.

We also conduct a visualization experiment in Fig. 3 that shows the MMWHS segmentation visualization (Sub Fig.1), T-SNE visualization of in and out of distribution data points produced by the uncertainty maps from the RNP module on the CT images from MMWHS (Sub Fig.2), and the CRNP uncertainty heat-maps for BraTS2020 images (Sub Fig.3). As the two cases from validation set shown in Sub Fig.(1), (a) (c) are segmented by the Base model and (b) (d) are from CRNP. The color masks denote the segmentation results (e.g., pink) overlaid on the ground truth (e.g., purple). The obvious segmentation differences are highlighted by yellow boxes. When comparing segmentation from two models, we can notice that our CRNP has better segmentation results, especially on the organ edges. This is mainly because organ edges contain more uncertain regions. The proposed CRNP can perceive uncertain segmented regions within one modality and assign more weights to the other one. By leveraging this information, CRNP is able to alleviate segmentation uncertainties in organ edges. Moreover, we visualize the in and out of distribution uncertainty maps processed by T-SNE in Sub Fig.(2). Following Han et al. [14], we consider the original features as the in distribution data and noisy features modified by additive Gaussian noise as the out of distribution data. Then, these samples are fed into the cross-modal RNP modules to get the uncertainty map predictions. The T-SNE is able to clearly split these uncertainty predictions into two clusters. This shows further evidence of the effectiveness of our CRNP model to estimate uncertainties. In Sub Fig.(3), we show the CRNP uncertainty heat-maps for a BraTS image, where the maps are estimated in the feature space and mapped back to the original image space. In this figure, (a)(c)(e)(g) are the flair, t1, t1ce and t2 modalities; (b)(d)(f)(h) are the CRNP uncertainty maps for the modalities above (brighter pixel = higher uncertainty); and (i)(j) are the ground truth (GT) segmentation and CRNP prediction. Note that the high uncertainty regions are concentrated around the areas with brain tumors, which is reasonable since tumors are sparsely represented in the feature space, resulting in a large difference between RNP's random and prediction networks. Also note that the flair image has a stronger tumor signal than the other modalities, producing a larger uncertainty for the other modalities. In particular, this larger uncertainty will notify the other modalities to pay more attention to these areas.

5 Conclusions

In this paper, we proposed the Uncertainty-aware Multi-modal Learning model, named Cross-modal Random Network Prediction (CRNP). CRNP measures the total uncertainty in the feature space for each modality to better guide multi-modal fusion. Moreover, technically speaking, the proposed CRNP is the first approach to explore random network prediction to estimate uncertainty and fuse multi-modal data. CRNP has a stable training process compared with a recent multi-modal approach that uses potentially unstable covariance measures to estimate uncertainty [26], and CRNP can also be easily translated between different prediction tasks. Through experiments on two medical image segmentation datasets and three computer vision classification datasets, the effectiveness of the proposed CRNP model is verified. Also, ablation and visualization studies further validate CNRP as an effective multi-modal analysis method.

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion **76**, 243–297 (2021)
2. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 119–127. Springer (2019)
3. Burda, Y., Edwards, H., Storkey, A., Klimov, O.: Exploration by random network distillation. arXiv preprint arXiv:1810.12894 (2018)
4. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16867–16876 (2021)
5. Chen, Y., Xian, Y., Koepke, A., Shan, Y., Akata, Z.: Distilling audio-visual knowledge by compositional contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7016–7025 (2021)
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
7. Corbière, C., Thome, N., Bar-Hen, A., Cord, M., Pérez, P.: Addressing failure prediction by learning model confidence. Advances in Neural Information Processing Systems **32** (2019)
8. DeVries, T., Taylor, G.W.: Leveraging uncertainty estimates for predicting segmentation quality. arXiv preprint arXiv:1807.00502 (2018)
9. Dou, Q., Liu, Q., Heng, P.A., Glocker, B.: Unpaired multi-modal segmentation via knowledge distillation. In: IEEE Transactions on Medical Imaging (2020)
10. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 2, pp. 524–531. IEEE (2005)
11. Gal, Y., Ghahramani, Z.: Bayesian convolutional neural networks with bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158 (2015)
12. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
13. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al.: A survey of uncertainty in deep neural networks. arXiv preprint arXiv:2107.03342 (2021)
14. Han, Z., Zhang, C., Fu, H., Zhou, J.T.: Trusted multi-view classification. arXiv preprint arXiv:2102.02051 (2021)
15. Heo, J., Lee, H.B., Kim, S., Lee, J., Kim, K.J., Yang, E., Hwang, S.J.: Uncertainty-aware attention for reliable interpretation and prediction. Advances in neural information processing systems **31** (2018)
16. Jia, X., Jing, X.Y., Zhu, X., Chen, S., Du, B., Cai, Z., He, Z., Yue, D.: Semi-supervised multi-view deep discriminant representation learning. IEEE transactions on pattern analysis and machine intelligence **43**(7), 2496–2509 (2020)
17. Jungo, A., Reyes, M.: Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 48–56. Springer (2019)
18. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)

19. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems **31** (2018)
20. Kohl, S.A., Romera-Paredes, B., Maier-Hein, K.H., Rezende, D.J., Eslami, S., Kohli, P., Zisserman, A., Ronneberger, O.: A hierarchical probabilistic u-net for modeling multi-scale ambiguities. arXiv preprint arXiv:1905.13077 (2019)
21. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
22. Li, Y., Luo, L., Lin, H., Chen, H., Heng, P.A.: Dual-consistency semi-supervised learning with uncertainty quantification for covid-19 lesion segmentation from ct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 199–209. Springer (2021)
23. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
24. Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 318–329. Springer (2021)
25. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
26. Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. Advances in Neural Information Processing Systems **33**, 12756–12767 (2020)
27. Nie, D., Wang, L., Gao, Y., Shen, D.: Fully convolutional networks for multi-modality isointense infant brain image segmentation. In: 2016 IEEE 13Th international symposium on biomedical imaging (ISBI). pp. 1342–1345. IEEE (2016)
28. Osband, I., Aslanides, J., Cassirer, A.: Randomized prior functions for deep reinforcement learning. Advances in Neural Information Processing Systems **31** (2018)
29. Patrick, M., Asano, Y.M., Kuznetsova, P., Fong, R., Henriques, J.F., Zweig, G., Vedaldi, A.: Multi-modal self-supervision from generalized data transformations. arXiv preprint arXiv:2003.04298 (2020)
30. Patrick, M., Huang, P.Y., Misra, I., Metze, F., Vedaldi, A., Asano, Y.M., Henriques, J.F.: Space-time crop & attend: Improving cross-modal video representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10560–10572 (2021)
31. Robinson, R., Oktay, O., Bai, W., Valindria, V.V., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., et al.: Real-time prediction of segmentation quality. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 578–585. Springer (2018)
32. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in Neural Information Processing Systems **31** (2018)
33. Valindria, V.V., Pawlowski, N., Rajchl, M., Lavdas, I., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B.: Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 547–556. IEEE (2018)
34. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
35. Wang, H., Wu, Q., Shen, C.: Soft expert reward learning for vision-and-language navigation. In: European Conference on Computer Vision. pp. 126–141. Springer (2020)

36. Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y.: Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 450–460. Springer (2021)
37. Wang, L., Ju, L., Zhang, D., Wang, X., He, W., Huang, Y., Yang, Z., Yao, X., Zhao, X., Ye, X., et al.: Medical matting: a new perspective on medical segmentation with uncertainty. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 573–583. Springer (2021)
38. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119. Springer (2021)
39. Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., Huang, J.: Deep multimodal fusion by channel exchanging. Advances in Neural Information Processing Systems **33**, 4835–4845 (2020)
40. Wang, Y., Zhang, Y., Hou, F., Liu, Y., Tian, J., Zhong, C., Zhang, Y., He, Z.: Modality-pairing learning for brain tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 230–240. Springer (2020)
41. Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Medical image analysis **58**, 101537 (2019)