This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MMANet: Margin-aware Distillation and Modality-aware Regularization for Incomplete Multimodal Learning

Shicai Wei Chunbo Luo Yang Luo
School of Information and Communication Engineering
University of Electronic Science and Technology of China
shicaiwei@std.uestc.edu.cn {c.luo, luoyang}@uestc.edu.cn

Abstract

Multimodal learning has shown great potentials in numerous scenes and attracts increasing interest recently. However, it often encounters the problem of missing modality data and thus suffers severe performance degradation in practice. To this end, we propose a general framework called MMANet to assist incomplete multimodal learning. It consists of three components: the deployment network used for inference, the teacher network transferring comprehensive multimodal information to the deployment network, and the regularization network guiding the deployment network to balance weak modality combinations. Specifically, we propose a novel margin-aware distillation (MAD) to assist the information transfer by weighing the sample contribution with the classification uncertainty. This encourages the deployment network to focus on the samples near decision boundaries and acquire the refined inter-class margin. Besides, we design a modality-aware regularization (MAR) algorithm to mine the weak modality combinations and guide the regularization network to calculate prediction loss for them. This forces the deployment network to improve its representation ability for the weak modality combinations adaptively. Finally, extensive experiments on multimodal classification and segmentation tasks demonstrate that our MMANet outperforms the state-of-the-art significantly. Code is available at: https://github.com/shicaiwei123/MMANet

1. Introduction

Multimodal learning has achieved great success on many vision tasks such as classification [21, 33, 46], object detection [26, 45, 53], and segmentation [5, 23, 41]. However, most successful methods assume that the models are trained and tested with the same modality data. In fact, limited by device [32, 39], user privacy [13, 25], and working condition [3, 29], it is often very costly or even infeasible to collect complete modality data during the inference stage. There is thus substantial interest in assisting the incomplete or even single modality inference via the complete modality data during training.

Modality	Customized	Unified	Drop
RGB	10.01	11.75	-1.65
Depth	4.45	5.87	-1.42
IR	11.65	16.62	-4.97
RGB+Depth	3.41	4.61	-1.2
RGB+IR	6.32	6.68	-0.36
Depth+IR	3.54	4.95	-1.41
RGB+Depth+IR	1.23	2.21	-0.98

Table 1. The performance of customized models and the unified model for different modality combinations on the CASIA-SURF dataset using the average classification error rate. The 'customized' means to train a model for each combination independently while the 'unified' means to train only one model for all the combinations. The architectures of all the models are the same and the feature map of missing modality (such as the IR for RGB+Depth) is set as zero.

A typical solution is to reconstruct the sample or feature of the missing modalities from the available ones [10, 14, 15, 20, 29, 32]. Nevertheless, this needs to build a specific model for each modality from all possible modality combinations and thus has high complexity. Recent studies focus on learning a unified model, instead of a bunch of networks, for different modality combinations. Generally, many such approaches [6, 11, 12, 17, 51, 52] attempt to leverage feature fusion strategies to capture modality-invariant representation so that the model can adapt to all possible modality combinations. For example, RFNet [11] designs the region-aware fusion module to fuse the features of available image modalities.

Although the existing unified models are indeed able to increase the efficiency of training and deployment of the multimodal models, their performance is likely to be sub-

optimal. As shown in Table 1, the customized models consistently outperform the unified model for different modality combinations. This is because existing unified models usually focus on the modality-invariant features while ignoring the modality-specific information. Note that the complementary modality-specific information of multiple modalities can help refine the inter-class discrimination and improve inference performance [2, 18, 36]. This motivates us to propose the first research question of this paper: **Can a unified model consider the modality invariant and specific information simultaneously while maintaining robustness for incomplete modality input?**

To this end, we propose to guide the unified model to learn the comprehensive multimodal information from the teacher model trained with complete modality. This regularizes the target task loss to encourage the unified model to acquire complementary information among different modality combinations multimodal information while preserving the generalization to them. Specifically, we propose a novel margin-aware distillation (MAD) that trains the unified model by guiding it to mimic the inter-sample relation of the teacher model. MAD introduces the classification uncertainty of samples to re-weigh their contribution to the final loss. Since the samples near the class boundary are more likely to be misclassified and have higher classification uncertainty [8], this encourages the unified model to preserve the inter-class margin refined by the complementary cues and learn the modality-specific information.

Another limitation of existing unified approaches is that they struggle to obtain optimal performance for the unbalanced training problem. To be specific, conventional multimodal learning models tend to fit the discriminative modality combination and their performance will degrade significantly when facing weak modality combinations. To solve this issue, existing unified approaches introduce the auxiliary discriminator to enhance the discrimination ability of the unimodal combinations [6, 11, 51]. This utilizes a hypothesis that a single modality is weaker than multiple ones. However, as shown in Table 1, no matter for the customized model or the unified model, the single Depth modality outperforms the RGB, IR, and their combinations. This indicates the combination with multiple weak modalities may be harder to be optimized than a single strong modality. Moreover, as shown in Table 3, RGB becomes the strong modality while Depth and IR become the weak modalities. This indicates that the modality importance is not fixed but varies with scenarios. These findings motivate us to propose the second research question: **How to effectively optimize the weak modality combination in varying scenarios?**

To this end, we design a regularization network and MAR algorithm to assist the training of the unified network. Specifically, the regularization network generates additional predictions for all inputs. Then MAR mines and calculates prediction loss for the sample from the weak combinations. This forces the unified model to improve its representation ability for the weak combination. In detail, MAR mines the weak combination via the memorization effect [1, 16, 49] that DNNs tend to first memorize simple examples before overfitting hard examples. As shown in Fig. 5(a), the unified model tends to fit the samples containing Depth modality firstly at the early stage. Therefore, MAR first mines the strong modality via the memorization effect. Then it determines the combinations of rest modalities as the weak ones.

Finally, we develop a model and task agnostic framework called MMANet to assist incomplete multimodal learning by combining the proposed MAD and MAR strategies. MMANet can guide the unified model to acquire comprehensive multimodal information and balance the performance of the strong and weak modality combination simultaneously. Extensive comparison and ablation experiments on multimodal classification and segmentation tasks demonstrate the effectiveness of the MMANet.

2. Related work

2.1. Multimodal Learning for Missing Modalities

Most existing multimodal learning methods assume that all instances consist of full modalities. However, this assumption does not always hold in real-world applications due to the device [32, 39], user privacy [13, 25], and working condition [3, 29]. Recently, many incomplete multimodal learning methods have been proposed and can be roughly categorized into two types: customized methods and unified methods. Customized methods aim to train a specific model to recover the missing modality in each incomplete modality combination. According to the recovering target, the customized methods can be further divided into sample-based methods and representation-based methods. Sample-based methods focus on imputing the missing modality at the input space with generative adversarial networks [4,27,32,37]. Due to the complexity of sample reconstruction, it is usually unstable and may introduce noise to harm the primary task at hand [34]. Thus the representation-based methods are proposed to reconstruct the sample representation via the knowledge distillation [3, 14, 20, 29] or matrix completion [30, 35]. Although promising results are obtained, these methods have to train and deploy a specific model for each subset of missing modalities, which has high complexity in practical applications.

The unified methods aim to train one model to deal with different incomplete modality combinations by extracting the modality-invariant features. For example, HeMIS [17] learns an embedding of multimodal information by computing statistics (i.e., mean and variance) from any number of available modalities. Furthermore, Chen *et al*. introduce the feature disentanglement to cancel out the modality-

specific information. Besides, more recent methods, such as LCR [52] and RFNet [11] focus on extracting the modality-invariant representation via different attention mechanisms. Moreover, mmFormer [51] introduces the transformer block to model the global semantic information for the modality-invariant embedding. While these methods achieve promising results, they only consider the modality-invariant information while ignoring the modality-specific information. As a result, they usually perform much worse than the customized methods, especially when more than one modality is missing [48].

2.2. Knowledge Distillation

Knowledge distillation aims to transfer knowledge from a strong teacher to a weaker student network to facilitate supervised learning. Generally, the distillation method can be divided into three types: response-based distillation that matches the softened logits of teachers and students [19], the representation-based distillation that matches the feature maps [24,28,40], and the relation-based distillation that matches the sample relations. [38,47].

While originating from the resource-efficient deep learning, knowledge distillation has found wider applications in such areas as incomplete multimodal learning. Here, it is used to transfer the privileged modality information that can only be accessed during the training stage from the teacher to the student [3,29]. Since the input of the teacher and student network is different in incomplete multimodal learning, transferring knowledge by representation-based methods may lead to overfitting [15]. Recent methods focus on transferring the privileged modality information by the relation-based methods [7,22,48]. However, these prior arts usually consider different instances equally and ignore their specificity, which would lead to sub-optimal performance.

3. Method

3.1. MMANet

In this section, we introduce a general framework called MMANet to address the challenge of incomplete multimodal learning. As shown in Fig. 1, it consists of three parts: deployment network, teacher network, and regularization network. Specifically, the deployment network is the inference network. To make it robust to the modality incompleteness, MMANet introduces the Bernoulli indicator $\Delta = \{\delta_1 ... \delta_m\}$ after modality encoders and conducts modality dropout during the training stage by randomly setting some components of Δ as 0. For missing modalities, the corresponding encoded feature maps will be replaced by a zero matrix. Besides, MMANet introduces the teacher network that is pre-trained with complete multimodal data to transfer the comprehensive multimodal knowledge to the deployment network via the MAD. This helps the deployment network acquire the modality-invariant and specific features simultaneously. Finally, MMANet guides the deployment network to train together with the regularization network that produces additional predictions for the weak modality combination via the MAR. This alleviates the overfitting for strong modality combinations. The total loss to guide the training of the deployment network is defined as follows,

$$L_{total} = L_{TL} + \alpha L_{MAD} + \beta L_{MAR} \tag{1}$$

where α and β are the hyper-parameters. L_{LT} is task learning loss, which is determined by the primary task at hand. For example, L_{LT} may be the cross entropy loss when the primary task is classification. L_{MAD} and L_{MAR} are the loss of MAD and MAR respectively.

Besides, the other nations used in MMANet are defined as follows. Given a mini-batch multimodal input $x = \{x_1, ..., x_m\}$, $x_m \in R^b$ denotes the data of m_{th} modality. b is the batch size. E_m^t and E_m^d denote the encoders for the m_{th} modality in the teacher and deployment networks, respectively. F^t and F^d denote the fusion module used in the teacher and deployment networks, respectively. $\Delta^d \in R^{b \times m}$ is the vector of Δ. $z^t \in R^{b^t \times c^t \times h^t \times w^t}$ and $z^d \in R^{b^d \times c^d \times h^d \times w^d}$ denote the fused features of the teacher and deployment networks, respectively. Here, where b is the batch size, c is the number of output channels, and h and w are spatial dimensions. P^t, P^r, and P^d denote the task predictor of the teacher, regularization, and deployment networks, respectively. y^t,y^r, and y^d denote the $R^{b \times k}$ prediction matrix of the teacher, regularization, and deployment networks, respectively. Here, k is the class number.

3.2. MAD

This section introduces the proposed MAD strategy for transferring the comprehensive multimodal information from the teacher network to the deployment network. As shown in Fig. 1, MAD is conducted between the z^t and z^d. z^d of a simple is varying due to the random modality dropout. In contrast, the sample semantic is invariant. Thus, MAD proposes to transfer the teacher's knowledge via relation consistency instead of feature consistency. This helps avoid overfitting and harming the representation ability of deployment networks. Moreover, MAD proposes to measure the class boundaries and guide the unified model to pay more attention to the samples near them. This can encourage the development network to inherit the refined inter-class margin from the teacher network. Nevertheless, boundaries are usually difficult to detect due to their irregularity. To solve this issue, MAD introduces the classification uncertainty of each sample to re-weight its contribution for the total loss. Since the samples near the class boundaries

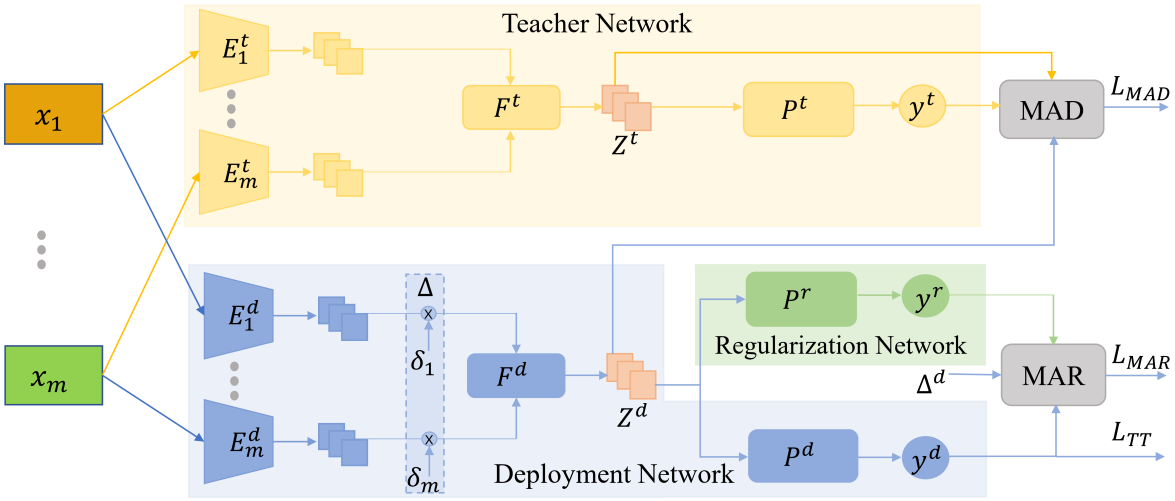

Figure 1. Overview of the proposed MMANet. It consists of three parts: the deployment network used for final inference, the teacher network transferring comprehensive multimodal knowledge to the deployment network, and the regularization network guiding the deployment network to balance weak modality combinations.

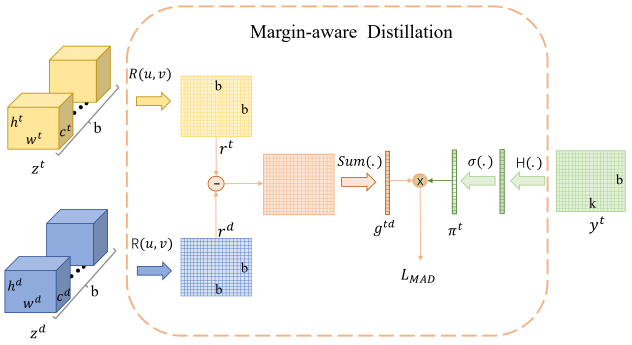

Figure 2. The illustration of the proposed MAD.

are more likely to be misclassified and have higher classification uncertainty, this can realize attention to them.

The overview of the MAD is shown in Fig. 2. It takes z^t, z^d, and y^t as the input and consists of three steps: (a) calculating the relation discrepancy vector $g^{td} \in R^b$, (b) calculating the classification uncertainty vector $\pi^t \in R^b$ and (c) calculating the total loss L_{MAD} for MAD.

(a) MAD calculates g^{td} from z^t and z^d. Specifically, MAD first reshape z^t and z^d into $z^{t'} \in R^{b^t \times c^t * h^t * w^t}$ and $z^{d'} \in R^{b^d \times c^d * h^d * w^d}$. Then MAD calculates the relation matrix $r^t \in R^{b \times b}$ and $r^d \in R^{b \times b}$ via the same relation function R(u,v), respectively. And the $r^t(i,j)$ that denotes the relation between i_{th} and j_{th} sample representations of the teacher network can be expressed as follows,

$$r^t(i,j) = R(z^{t'}(i,:), z^{t'}(j,:)) \tag{2}$$

Besides, the $r^d(i,j)$ that denotes the relation between i_{th} and j_{th} sample representations of the deployment network can be expressed as follows,

$$r^d(i,j) = R(z^{d'}(i,:), z^{d'}(j,:)) \tag{3}$$

Theoretically, $R(u,v)$ can be any metric for measuring the vector distance, such as the Euclidean distance and the cosine distance. Because the dimension of the feature vectors of the teacher and the deployment networks could be very high, to eliminate the curse of dimensionality, we choose cosine distance as the $R(u,v)$,

$$R(u,v) = \frac{u^T v}{\|u\|_2 \|v\|_2} \tag{4}$$

Furthermore, MAD calculates the discrepancy matrix between r^t and r^d and sum each row to get g^{td}.

$$g^{td} = \sum_{i=1}^{b} (r^t - r^d)_i \tag{5}$$

Here, $g^{td}(i)$ denotes the relation gap between the teacher and deployment networks from the i_{th} sample to other samples in the same mini-batch.

(b) MAD calculates π^t from y^t. In detail, it takes the information entropy of the logit output of each sample as its classification uncertainty. And the classification uncertainty for i_{th} sample, $\pi^t(i)$, can be expressed as follows,

$$\begin{cases} \pi^t(i) = H(y^t(i,:)) & (6) \\ H(x) = -\sigma(x) * log(\sigma(x)) & (7) \end{cases}$$

where $\sigma(.)$ is the softmax function for normalization. $H(x)$ is the information entropy of x. A sample that has a higher

classification uncertainty is usually closer to the decision boundaries, since it is more likely to be misclassified. Thus, $\pi^t(i), i \in [1, b]$ can also denote the margin from the i_{th} sample representation to the decision boundary.

(c) Finally, MAD takes $\pi^t(i)$ as the weight for the corresponding component $g^{td}(i)$ to calculate L_{MAD},

$$L_{MAD} = \sum_{i=1}^{b} \sigma(\pi^t)(i) * g^{td}(i) \tag{8}$$

This encourages the deployment network to focus on the samples near the decision boundaries and preserve the inter-class margin refined by the comprehensive multimodal information from the teacher network.

3.3. MAR

This section introduces the MAR algorithm that forces the deployment network to improve its discriminating ability for weak modality combinations adaptively. As shown in Fig. 1, MAR takes the y^r, y^d and Δ^d as as the input to calculate the L_{MAR}. Specifically, MAR first proposes a contrastive ranking strategy to mine the weak modality combinations. Compared to simply taking the combination with a single modality as the weak one, this further considers the combination with multiple modalities and can get more accurate mining results. Then, MAR calculates the prediction loss for the weak modality combinations, guiding the deployment network to pay more attention to them.

The overview of MAR is shown in Fig. 3. It consists of two steps: (a) when $E \leq N$, mining the weak modality combination set Ω, and (b) when $E > N$, calculating L_{MAR}. Here E is the current training epoch, and N is the number of warm-up epochs.

(a) MAR calculates Ω from y^d using contrastive ranking. MAR proposes to calculate the predicted output $Y^O \in R^{(m+1)\times n\times k}$ of Δ_i, $i \in [0, m]$, on the train set after each training epoch.

$$Y^O(i, :, :) = y^d(\Delta_i) \tag{9}$$

n is sample number. Δ_i means the i_{th} component of Δ is 0. Δ_0 means none component of Δ are 0, which must contain the strong modality. Since the deployment network tends to first memorize the samples with strong modality, $\Delta_w, w \in [1, m]$ that makes $Y^O(w, :, :)$ has the biggest distance with $Y^O(0, :, :)$ is the hard combination that does not contain the strong modality. And the element of Ω can be determined as Δ_w and the Δ consists of the modalities in it.

Specifically, to make Δ_w robust for the randomness of neural network learning, MAR introduces two innovations. Firstly, MAR calculates the prediction discrepancy from the prediction distribution $Y^d \in R^{(m+1)\times k}$ instead of Y^O,

$$\begin{cases} Y^d(i, j) = \sum(Y^D(i, :) == j) & (10) \\ Y^D = \arg\max(Y^O, dim = 2) & (11) \end{cases}$$

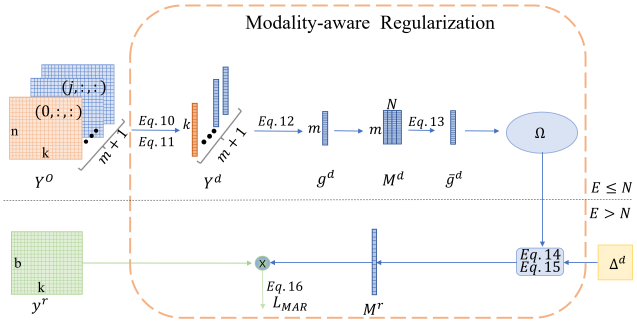

Figure 3. The illustration of the proposed MAR.

where $j \in [0, k-1]$. Compared with Y^O, Y^d needs only class-wise but not sample-wise consistency. Then the vector discrepancy vector $g^d \in R^m$ is defined as follows,

$$g^d(i) = KL(log(\sigma(Y^d(i))), \sigma(Y^d(0))) \tag{12}$$

where KL(,) means the KL divergence, $i \in [1, m]$.

Secondly, MAR introduces a memory bank $M^d \in R^{N\times m}$ to save the g^d among the warm-up epochs and performs average filtering to obtain $\overline{g}^d$,

$$\overline{g}^d = \sum_{i=1}^{N} \frac{1}{N} (M^d)_i \tag{13}$$

where $(M^d)_i$ is the g^d in the i_{th} epoch. And Δ_w can be determined as Δ_i where $i = argmax(\overline{g}^d)$.

(b) MAR calculates L_{MAR} from y^r, Δ^d and Ω. In detail, MAR first calculates the weak combination mask $M^r \in R^b$ from Δ^d and Ω,

$$\begin{cases} M^r(i) = FALSE \quad if \Delta^d(i) \notin \Omega & (14) \\ M^r(i) = TRUE \quad if \Delta^d(i) \in \Omega & (15) \end{cases}$$

where $i \in [0, b-1]$, $\Delta^d(i)$ is the Δ for the i_{th} sample in this mini-batch. Then, the L_{MAR} is defined as follows,

$$L_{MAR} = L_{TL}(y^r[M^r], l[M^r]) \tag{16}$$

where l is the groundtruth vector for y^r. [.] denotes the index operator.

4. Experiments

We conduct experiments on multimodal classification and segmentation tasks to evaluate the proposed MMANet. In the following, we first compare the MMANet architecture with the state-of-the-art on these two tasks. Then, we ablate the MAD and MAR strategies of MMANet.

4.1. Performance and Comparison on Multimodal Classification

Datasets: We take the face anti-spoofing task as the example of the multimodal classification and conduct exper-

Table 2. Performance on the multimodal classification task with CASIA-SURF. ↓ means that the lower the value, the better the performance.

Modalities			ACER(↓)						
RGB	Depth	IR	Customized	Unified					
			SF	SF-MD	HeMIS	LCR	RFNet	MMFormer	MMANet
●	○	○	10.01	11.75	14.36	13.44	12.43	11.15	**8.57**
○	●	○	4.45	5.87	4.70	4.40	4.17	4.67	**2.27**
○	○	●	11.65	16.62	16.21	15.26	14.69	13.99	**10.04**
●	●	○	3.41	4.61	3.23	3.32	2.23	1.93	**1.61**
●	○	●	6.32	6.68	6.27	5.16	4.27	4.77	**3.01**
○	●	●	3.54	4.95	3.68	3.53	3.22	3.10	**1.18**
●	●	●	1.23	2.21	1.97	1.88	1.18	1.94	**0.87**
Average			5.80	7.52	7.18	6.71	6.02	5.93	**3.94**

Table 3. Performance on the multimodal classification task with the CeFA dataset.

Modalities			ACER(↓)		
RGB	Depth	IR	Customized	Unified	
			SF	MMFormer	MMANet
●	○	○	27.44	28.51	**27.15**
○	●	○	33.75	33.58	**32.50**
○	○	●	36.17	39.56	**35.62**
●	●	○	35.62	29.47	**22.87**
●	○	●	31.62	27.66	**23.27**
○	●	●	36.62	32.17	**30.45**
●	●	●	24.15	30.72	**23.68**
Average			32.20	31.52	**27.94**

iments on the CASIA-SURF [50] and CeFA [31] datasets. Both of them consist of the RGB, Depth, and IR modalities. For CASIA-SURF, we follow the intra-testing protocol suggested by the authors and divide it into train, validation, and test sets with 29k, 1k, and 57k samples, respectively. For CeFA, we follow the cross-ethnicity and cross-attack protocol suggested by the authors and divide it into train, validation, and test sets with 35k, 18k, and 54k samples, respectively. Here we report the results on the test set using the metric of Average Classification Error Rate (ACER) [50].

Implementation: We use random flipping, rotation, and cropping for data augmentation. All models are optimized by an SGD for 100 epochs with a mini-batch 64. The learning rate is initialized to 0.001 with 5 epochs of linear warm-up and divided by 10 at 16, 33, and 50 epochs. Weight decay and momentum are set to 0.0005 and 0.9, respectively.

The hyper-parameters of comparison methods use the suggested ones in the original articles. The (α, β) for MMANet is set as $(30, 0.5)$ and $(30, 0.5)$ for CASIA-SURF and CeFA, respectively. The warm-up epoch N is set as 5.

Comparison: Here we compare MMANet with two different unified methods for incomplete multimodal learning. One is an early method that only focuses on extracting modality-invariant features, such as HeMIS [17] and LCR [52]. Another is the enhanced method that further considers improving the discrimination ability for single-modal combinations, such as RFNet [11], and mmFormer [51].

Besides, we introduce two baseline methods, SF and SF-MD. SF [50] is the benchmark method of the CASIA-SURF, which is a customized method that trains the model for each modality combination. SF-MD is the variant of SF by simply adding the Bernoulli indicator after its modality encoder. This enables SF-MD to become a unified model that trains a single model for all modality combinations.

Finally, for a fair comparison, we follow the basic implementation of SF for all the comparison methods. Specifically, we unify the modality encoders of HeMIS, LCR, RFNet, and mmFormer as the ResNet18 used in SF. Besides, we set the SF model trained with complete multimodal data as the teacher network and the SF-MD model as the development network.

Results: Table 2 and Table 3 show the comparison results with the state-of-the-art methods on the CASIS-SURF and CeFA datasets, respectively. Compared with the second-best unified method, i.e. mmFormer, MMANet decreases the average ACER by 1.99% and 3.58% on the CASIS-SURF and CeFA, respectively. Besides, we can see that MMANet achieves the best performance on both datasets for all the nine modality combinations for CASIA-SURF. This shows the superiority of our method on the incomplete multimodal classification task. More importantly, MMANet even outperforms the customized baseline method, i.e. SF, for all the modality combinations on the CASIA-SURF and CeFA, decreasing the average ACER by 1.86% and 4.26%. This demonstrates the effectiveness of the proposed MAD and MAR for the incomplete multimodal classification task.

Table 4. Performance on the multimodal segmentation task with NYUv2. ↑ means that the higher the value, the better the performance.

Modality		mIOU(↑)						
RGB	Depth	Customized	Unified					
		ESANet	ESANet-MD	HeMIS	LCR	RFNet	mmFormer	MMANet
●	○	44.22	41.34	33.23	41.91	42.89	43.22	**44.93**
○	●	40.55	39.76	31.23	39.88	40.76	41.12	**42.75**
●	●	49.18	47.23	37.77	47.46	48.13	48.45	**49.62**
Average		44.65	42.77	34.07	43.08	43.92	44.26	**45.58**

Table 5. Performance on the multimodal segmentation task with the Cityscapes dataset.

Modality		mIOU(↑)		
RGB	Depth	Customized	Unified	
		ESANet	mmFormer	MMANet
●	○	77.60	76.62	**77.61**
○	●	59.11	58.53	**60.12**
●	●	78.62	78.01	**78.89**
Average		71.77	71.05	**72.20**

4.2. Performance and Comparison on Multimodal Segmentation

Datasets: We take the semantic segmentation task as the example of multimodal segmentation and conduct experiments on the NYUv2 [43] and Cityscapes [9] datasets. Both of them consist of the RGB and Depth modalities. Specifically, NYUv2 contains 1,449 indoor RGB-D images, of which 795 are used for training and 654 for testing. We used the common 40-class label setting. Cityscapes is a large-scale outdoor RGB-D dataset for urban scene understanding. It contains 5,000 finely annotated samples with a resolution of 2048×1024, of which 2,975 for training, 500 for validation, and 1,525 for testing. Cityscapes also provides 20k coarsely annotated images, which *we did not use for training*. For both datasets, we report the results on the validation set using the metric of mean IOU (mIOU).

Implementation: We use random scaling, cropping, color jittering, and flipping for data augmentation. All models are optimized by Adam for 300 epochs with a mini-batch 8. The learning rate is initialized with 1e-2 and adapted by the PyTorch's one-cycle scheduler [44].

The hyper-parameters of the comparison methods use the suggested ones in their article.The hyper-parameter (α, β) for MMANet is set as $(4, 0.2)$ and $(10, 0.1)$ for the NYUv2 and Cityscapes datasets, respectively. The warm-up epoch N is set as 20.

Comparison: We also compare MMANet with the HeMIS [17], LCR [52], RFNet [11], and mmFormer [51]. Here, we set ESANet and ESANnet-MD as the baseline. ESANet [42] is an efficient and robust model for RGB-D segmentation, which trains the model for each modality combination. ESANet-MD is the variant of ESANet by simply adding the Bernoulli indicator after its modality encoder. ESANet-MD trains only a single model for all modality combinations. Finally, for a fair comparison, we unify the modality encoder of HeMIS, LCR, RFNet, and mmFormer as the ResNet50 with NBt1 used in ESANet. Besides, we set the ESANet model trained with complete multimodal data as the teacher network and the ESANet-MD model as the development network.

Results: Table 4 and Table 5 list the comparison results on the NYUv2 and Cityscapes datasets, respectively. From these results, we can see that MMANet achieves the best performance on both datasets for all the modality combinations. In particular, it outperforms the second-best method, mmFormer, by 1.32% and 1.05% in the NYUv2 and Cityscapes datasets, respectively. This demonstrates the effectiveness of the MMANet on the multimodal segmentation task. Moreover, MMANet improves the average performance of ESANet-MD by 2.81% in the NYUv2 dataset and even outperforms the customized baseline, ESANet, by 0.97% and 0.43% in NYUv2 and Cityscapes datasets. This shows the effectiveness of the proposed MAD and MAR on the incomplete multimodal segmentation task.

5. Ablation Study

This section will study the effectiveness of MAD and MAR and conduct extensive ablation experiments on four datasets. Limited by page, we only present the results of the CASIA-SURF dataset and other results can be seen in the supplementary material.

5.1. The effectiveness of MAD

To study the effect of MAD, we conduct experiments to compare the performance of the vanilla SF-MD and its variant with SP and MAD. Here, SP is the degradation method of MAD that transfers knowledge by directly matching the cosine distance of the sample representations between teacher and deployment networks. The results are shown in Table 6. We can see that the variant of SF-MD con-

Table 6. Ablation result of MAD on the CASIA-SURF dataset.

RGB	Depth	IR	SF-MD	+SP	+MAD
●	○	○	11.75	10.7	**10.36**
	●	○	5.87	3.3	**2.54**
	○	●	16.62	15.03	**11.67**
●	●	○	4.61	2.52	**1.23**
●	○	●	6.68	5.16	**4.09**
	●	●	4.95	3.18	**1.44**
●	●	●	2.21	1.13	**0.77**
Average			7.5	5.9	**4.57**

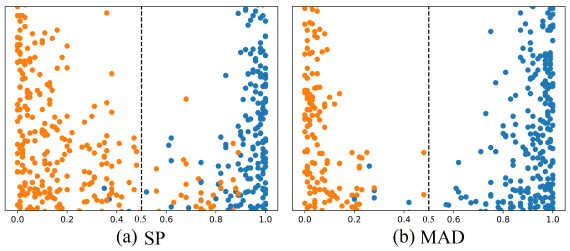

Figure 4. The prediction distribution of the SF-MD model assisted by SP and MAD on CASIA-SURF dataset. X-axis represents the normalized logit output and x=0.5 is the classification boundary. orange and blue dots denotes two different classes.

Table 7. Ablation result of MAR on the CASIA-SURF dataset.

RGB	Depth	IR	SF-MAD	+SR	+MAR
●	○	○	10.36	9.17	**8.57**
	●	○	2.54	**1.89**	2.27
	○	●	11.67	10.21	**10.04**
●	●	○	**1.23**	1.66	1.61
●	○	●	4.09	4.37	**3.01**
	●	●	1.44	2.12	**1.18**
●	●	●	**0.77**	0.92	0.87
Average			4.57	4.33	**3.94**

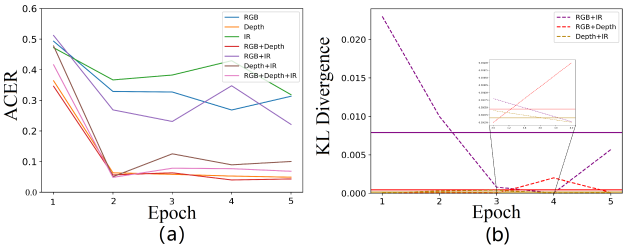

Figure 5. (a) The learning process of different modality combinations on the CASIA-SURF dataset during the warm-up stage. (b) The corresponding $g^d \in R^3$ for (dashed line) and its average result $\overline{g}^d$ (solid line) for the warm-up stage.

sistently outperforms the vanilla SF-MD in all the modality combination and improve its performance by 1.6% and 2.93% in average. This demonstrates the effectiveness of transferring comprehensive multimodal information from the teacher network to the deployment network. Furthermore, the proposed MAD outperforms SP by 1.33%, which demonstrates the validity of re-weighing sample loss via the classification uncertainty. This is because the classification uncertainty re-weighing can encourage the deployment to focus on the hard samples and thus acquire a more separable inter-class margin than the conventional SP (see Fig. 4).

5.2. The effectiveness of MAR

To study the effect of MAR, we conduct experiments to compare the performance of the SF-MAD, namely the SF-MD with the MAD, and its variant with SR and MAR. Here SR is the conventional modality regularization strategy considering only the single modality combination. As shown in Table 6, SR and MAR improve the performance of SF-MAD by 0.24% and 0.63% in average, respectively, showing the effeteness to regularize the single and weak modality combinations. Moreover, MAR outperforms SR by 0.39% in average, demonstrating the superiority of MAR.

Here the average gain of SR and MAR is less than SP and MAD since they aim to improve the performance of only the weak, not all combinations. Specifically, as shown in Table 6, the three worst-performing combinations are 'RGB', 'IR' and, 'RGB+IR'. However, SR only focuses on the combinations of single modality, RGB (1.19%), IR (1.46%), and Depth (0.65%), where 'Depth' is exactly a strong modality. In contrast, Fig. 5(b) shows that the prediction discrepancy between 'RGB+IR' and 'RGB+Depth+IR' is the largest. And the performance gain of MAR mainly comes from RGB (1.79%), IR (1.63%), as well as the combination of RGB and IR (1.02%). These results show that MAR can mine the weak modality combinations more accurately and force the deployment network to improve its discrimination ability for them.

6. Conclusion

This paper presents an MMANet framework to aid the deployment network for incomplete multimodal learning. Specifically, MMANet introduces a teacher network pretrained with complete multimodal data to transfer the comprehensive multimodal information to the deployment network via MAD. This helps it acquire modality-invariant and specific information while maintaining robustness for incomplete modality input. Besides, MMANet introduces a regularization network to mine and regularize weak modality combinations via MAR. This forces the deployment network to improve its discrimination ability for them effectively and adaptively. Finally, extensive experiments demonstrate the effectiveness of the proposed MMANet, MAD, and MAR for incomplete multimodal learning.

References

[1] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[3] Jonathan C, David A Stroud, Chen Ross, Jia Sun, Rahul Deng, and Sukthankar. D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2755–2764, 2020.

[4] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1158–1166, 2018.

[5] Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7088–7097, 2021.

[6] 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.

[7] 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.

[8] Zhuangzhuang Chen, Jin Zhang, Zhuonan Lai, Jie Chen, Zun Liu, and Jianqiang Li. Geometry-aware guided loss for deep crack recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4703–4712, 2022.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[10] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019.

[11] 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.

[12] 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.

[13] Kai Fan, Wei Jiang, Hui Li, and Yintang Yang. Lightweight rfid protocol for medical privacy protection in iot. *IEEE Transactions on Industrial Informatics*, 14(4):1656–1665, 2018.

[14] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision*, pages 103–118, 2018.

[15] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Learning with privileged information via adversarial discriminative modality distillation. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2581–2593, 2019.

[16] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

[17] 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.

[18] Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! *arXiv preprint arXiv:2010.06572*, 2020.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[20] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834, 2016.

[21] Danfeng Hong, Jingliang Hu, Jing Yao, Jocelyn Chanussot, and Xiao Xiang Zhu. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:68–80, 2021.

[22] 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.

[23] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019.

[24] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[25] Mimansa Jaiswal and Emily Mower Provost. Privacy enhanced multimodal neural representations for emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7985–7993, 2020.

[26] Wen-Da Jin, Jun Xu, Qi Han, Yi Zhang, and Ming-Ming Cheng. Cdnet: Complementary depth network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:3376–3390, 2021.

[27] Jiang Jue, Hu Jason, Tyagi Neelam, Rimner Andreas, Berry L Sean, Deasy O Joseph, and Veeraraghavan Harini. Integrating cross-modality hallucinated mri with ct to aid mediastinal lung tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 221–229. Springer, 2019.

[28] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

[29] Xiao Li, Lin Lei, Yuli Sun, and Gangyao Kuang. Dynamic-hierarchical attention distillation with synergetic instance selection for land cover classification using missing heterogeneity images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.

[30] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11174–11183, 2021.

[31] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1179–1187, 2021.

[32] Ajian Liu, Zichang Tan, Jun Wan, Yanyan Liang, Zhen Lei, Guodong Guo, and Stan Z Li. Face anti-spoofing via adversarial cross-modality translation. *IEEE Transactions on Information Forensics and Security*, 16:2759–2772, 2021.

[33] Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, et al. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[34] Haojie Liu, Shun Ma, Daoxun Xia, and Shaozi Li. Sfanet: A spectrum-aware feature augmentation network for visible-infrared person re-identification. *arXiv preprint arXiv:2102.12137*, 2021.

[35] Jiyuan Liu, Xinwang Liu, Yi Zhang, Pei Zhang, Wenxuan Tu, Siwei Wang, Sihang Zhou, Weixuan Liang, Siqi Wang, and Yuexiang Yang. Self-representation subspace clustering for incomplete multi-view data. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2726–2734, 2021.

[36] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389, 2020.

[37] Yongsheng Pan, Mingxia Liu, Chunfeng Lian, Yong Xia, and Dinggang Shen. Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages. *IEEE Transactions on Medical Imaging*, 39(9):2965–2975, 2020.

[38] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision*, pages 268–284, 2018.

[39] Allan Pinto, Helio Pedrini, William Robson Schwartz, and Anderson Rocha. Face spoofing detection through visual codebooks of spectral temporal cubes. *IEEE Transactions on Image Processing*, 24(12):4726–4740, 2015.

[40] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[41] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531. IEEE, 2021.

[42] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531. IEEE, 2021.

[43] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[44] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.

[45] Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1407–1417, 2021.

[46] Haiman Tian, Yudong Tao, Samira Pouyanfar, Shu-Ching Chen, and Mei-Ling Shyu. Multimodal deep representation learning for video classification. *World Wide Web*, 22(3):1325–1341, 2019.

[47] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

[48] Yixin Wang, Yang Zhang, Yang Liu, Zihao Lin, Jiang Tian, Cheng Zhong, Zhongchao Shi, Jianping Fan, and Zhiqiang He. Acn: Adversarial co-training network for brain tumor segmentation with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 410–420. Springer, 2021.

[49] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still)

requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[50] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 919–928, 2019.

[51] 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. *arXiv preprint arXiv:2206.02425*, 2022.

[52] 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.

[53] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Rgb-d salient object detection: A survey. *Computational Visual Media*, 7(1):37–69, 2021.