

Boosting Multimodal Learning via Disentangled Gradient Learning

Shicai Wei^{1,2}

Chunbo Luo^{1*}

Yang Luo¹

¹ University of Electronic Science and Technology of China

² National and Local Joint Engineering Research Center for Cloud Operating System

shicaiwei@std.uestc.edu.cn, {c.luo, luoyang}@uestc.edu.cn

Abstract

Multimodal learning often encounters the under-optimized problem and may have worse performance than unimodal learning. Existing methods attribute this problem to the imbalanced learning between modalities and rebalance them through gradient modulation. However, they fail to explain why the dominant modality in multimodal models also underperforms that in unimodal learning. In this work, we reveal the optimization conflict between the modality encoder and modality fusion module in multimodal models. Specifically, we prove that the cross-modal fusion in multimodal models decreases the gradient passed back to each modality encoder compared with unimodal models. Consequently, the performance of each modality in the multimodal model is inferior to that in the unimodal model. To this end, we propose a disentangled gradient learning (DGL) framework to decouple the optimization of the modality encoder and modality fusion module in the multimodal model. DGL truncates the gradient back-propagated from the multimodal loss to the modality encoder and replaces it with the gradient from unimodal loss. Besides, DGL removes the gradient back-propagated from the unimodal loss to the modality fusion module. This helps eliminate the gradient interference between the modality encoder and modality fusion module while ensuring their respective optimization processes. Finally, extensive experiments on multiple types of modalities, tasks, and frameworks with dense cross-modal interaction demonstrate the effectiveness and versatility of the proposed DGL. Code is available at <https://github.com/shicaiwei123/ICCV2025-GDL>

1. Introduction

With the growing availability of affordable sensors, multimodal learning, which harnesses data from various sources, has received significant interest in machine learning. It is evident in various domains, including classification tasks [11, 18, 30], object detection [15, 26, 41], and segmentation tasks [4, 13, 23]. Existing research on multimodal

learning mainly focuses on developing fusion techniques, such as tensor-based fusion [12, 37], and attention-based fusion [7, 40]. However, the simple combination of multiple modalities may not always yield satisfactory performance.

Recent studies [8, 21] observe that the multimodal models could be inferior to the unimodal models in some situations. These studies attribute the decline in performance to imbalanced learning between modalities, in which the dominant modality that has better performance will suppress the optimization of the weaker modality, leading to inadequate feature learning in the weaker modality. To address this issue, techniques such as teacher aid [8], gradient modulation [9, 17, 21, 36], and alternating optimization [14, 32, 33, 39] have been introduced to alleviate the imbalanced training and have shown promising results. However, these methods primarily focus on improving the weaker modality while overlooking the optimization of the dominant modality, which also exhibits sub-optimal performance compared to unimodal learning. As shown in Fig. 1, the audio branch that has better performance in the multimodal model (see Fig. 1 (a)) also underperforms its unimodal baseline (see Fig. 1 (b)).

To this end, this paper focuses on the question of what makes the dominant modality under-optimized in multimodal learning. Here, we reveal that each modality in the multimodal model, even the highest-performing one, learns less effectively than in unimodal models due to optimization conflicts with the fusion module. Specifically, compared with the unimodal model, we prove that the modality fusion module in multimodal models will suppress the gradient back-propagated to the modality encoder. Moreover, the suppression degree will increase with the optimization progress. As shown in Fig. 1(c), the gradient back-propagated to the audio encoder from the multimodal loss is smaller than that from the unimodal loss. And this gap becomes larger in the middle of the training. Consequently, the performance of each modality in the multimodal model is inferior to that in the unimodal model.

To tackle this issue, we introduce the disentangled gradient learning (DGL) strategy. Specifically, DGL first trun-

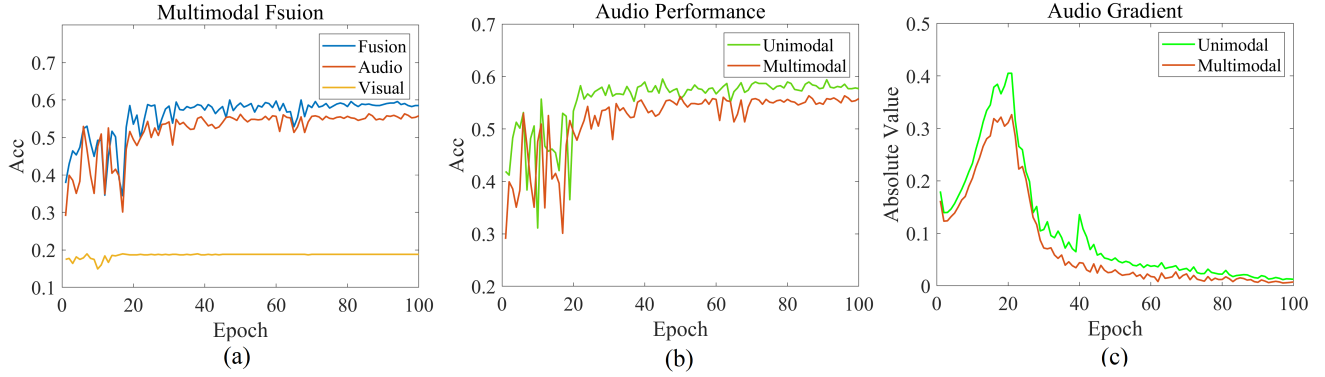


Figure 1. Visualization on the CREMA-D dataset. (a) illustrates the performance of each unimodal branch and their fusion in the multimodal model. (b) illustrates the performance of audio modality in the unimodal and multimodal models, respectively. (c) illustrates the gradient of audio modality back-propagated the unimodal and multimodal models, respectively.

cates the gradient from the modality fusion module propagating to the modality encoder. This helps avoid the problem of gradient suppression. Then DGL introduces independent unimodal loss for each modality encoder via the parameter-free modality dropout technique. This provides an independent gradient path for each modality encoder, eliminating inter-modality interference and enabling their optimization. Moreover, the gradient from the unimodal loss propagating to the modality fusion modules is also removed, avoiding the gradient interference between the unimodal loss and the multimodal loss. This enables the normal optimization of the fusion module in the multimodal model. Notably, the DGL relies solely on the representations of each modality, free from constraints of model structures and fusion methods, making it versatile for diverse scenarios. In summary, our contributions are as follows:

- We reveal that the cross-modal interaction in multimodal learning decreases the gradient back-propagating to each modality encoder, leading to their inferior learning. This helps explain the insufficient optimization phenomenon of the dominant modality in multimodal learning.
- We introduce DGL to truncate the gradient from the modality fusion module propagating to the modality encoder and replace it with the gradient from independent unimodal loss. This helps eliminate optimization conflicts between the fusion module and encoders, as well as interference among different encoders.
- Extensive experiments demonstrate that DGL can achieve considerable improvements over existing methods.

2. Related Work

2.1. Multimodal Learning

Multimodal learning utilizes complementary information contained in multimodal data to improve the performance of various tasks. One important direction in this area is the design of modality fusion methods, such as tensor-based

fusion [12, 37] and attention-based fusion [7, 40]. Moreover, considerable efforts have been dedicated to harnessing information from multiple modalities to enhance model performance in specific tasks compared to unimodal frameworks. These tasks include action recognition [6, 10, 19], semantic segmentation [4, 13, 23] and audio-visual speech recognition [1, 22, 27]. Yet, in joint training strategies, many multimodal methods often don't fully utilize all modalities, resulting in suboptimal unimodal representations. This leads to multimodal model performance falling short of expectations, even inferior to their unimodal counterparts.

2.2. Under-optimization in Multimodal Learning

Recent studies pointed out that most multimodal learning methods fail to enhance performance significantly even with more information [8, 9, 17, 21, 28, 36, 39]. Wang *et al.* [28] observed that different modalities exhibit varying convergence rates, leading to multimodal models that fail to surpass their unimodal counterparts. Peng *et al.* [21] further showed that the modality with superior performance tends to dominate the optimization process, leading to inadequate feature learning in weaker modalities. To this end, various methods have been developed to enhance the conventional multimodal learning framework and can be roughly categorized into two types: gradient modulation and alternating optimization.

Gradient modulation [9, 17, 21, 36] aims to enlarge the gradient of weaker modality in multimodal learning, balancing the optimization of different modality encoders. Specifically, OGM [21] proposes on-the-fly gradient modulation to manage the optimization of each modality adaptively. MMCosine [36] performs modality-wise L2 normalization to features and weights towards balanced and better multi-modal fine-grained learning. PMR [9] proposes the prototypical rebalancing strategy to hasten the learning of the slower modality and reduce the dominance of

the stronger one. AGM [17] introduces an adaptive gradient modulation method that can boost the performance of multimodal models with various fusion strategies. While gradient modulation methods show good results, improving weak modalities often degrades the performance of strong ones due to the inter-modality conflict.

To this end, the alternating optimization methods are proposed to improve the unimodal learning of each modality, including the domain one, in multimodal learning. MLA [39] transforms the conventional joint multimodal learning process into an alternating unimodal learning process to minimize inter-modality interference directly. ReconBoost [14] updates a fixed modality each time via a dynamical learning objective to overcome the competition with the historical models. MMPareto [32] leverages the Pareto integration technique to catch innocent unimodal assistance, avoiding its conflict with multimodal optimization. Besides, Wei *et al.* [33] proposes the Diagnosing & Re-learning method to overcome the intrinsic limitation of modality capacity via the network re-initialization technique. While these methods improve the performance of each modality in multimodal learning simultaneously, they still fail to explain why the dominant modality in multimodal models underperforms that in unimodal learning.

3. Method

In this section, we analyze the under-optimization problem in the conventional multimodal learning paradigm, and then we describe the details of our proposed disentangled gradient learning framework.

3.1. Under-optimization Analysis in Multimodal Learning.

Existing methods attribute the insufficient multimodal performance to imbalanced learning, where the dominant modality will suppress the learning of the weak one. We reveal that all modality in the multimodal model, including the dominant modality, is under-optimized compared to the unimodal model due to the optimization conflict between the modality encoder and fusion module in the multimodal model.

Without loss of generality, we consider two input modalities as m_1 and m_2 . As shown in Fig. 2 (a), conventional multimodal learning can be described as follows: given a training set $\mathcal{D} = \{x_i, y_i\}_{i \in [N]}$, where the inputs $x_i = (x_i^{m_1}, x_i^{m_2})$ and $y_i \in [1, 2, \dots, K]$, where K is the number of categories. We use two neural network encoders $\varphi_1(\theta_1, \cdot)$ and $\varphi_2(\theta_2, \cdot)$ to map each modality of the inputs to $z_i^{m_1} = \varphi_1(\theta_1, x_i^{m_1}) \in \mathbb{R}^{d_1}$ and $z_i^{m_2} = \varphi_2(\theta_2, x_i^{m_2}) \in \mathbb{R}^{d_2}$. Here θ_1, θ_2 are the parameters of φ_1 and φ_2 respectively. let $\varphi_\tau(\theta_\tau, \cdot, \cdot)$ denote the fusion module, $W \in \mathbb{R}^{K \times (d_1 + d_2)}$ and $b \in \mathbb{R}^K$ denote the parameters of the linear classifier to produce the logits output. The classification loss value of

input x_i in a multimodal model can be expressed as follows,

$$\begin{cases} L^{Multi} = L_{CE}(Wz_i^\tau + b, y_i) \\ z_i^\tau = \varphi_\tau(\theta_\tau, z_i^{m_1}, z_i^{m_2}) \end{cases} \quad (1)$$

where $L_{CE}(\cdot, \cdot)$ is the the cross-entropy loss function.

Then, according to the chain rule, the gradient $g_{\theta_1}^{Multi}$ passed back to the encoder $\varphi_1(\theta_1, \cdot)$ in the multimodal model can be expressed as follows,

$$g_{\theta_1}^{Multi} = \frac{\partial L^{Multi}}{\partial f(z_i^\tau)} \frac{\partial f(z_i^\tau)}{\partial z_i^\tau} \frac{\partial z_i^\tau}{\partial z_i^{m_1}} \quad (3)$$

where $f(z_i^\tau) = Wz_i^\tau + b$

Here, we set the fusion module as concatenation, which is the most widely used vanilla fusion method, thus $z_i^\tau = [z_i^{m_1}; z_i^{m_2}]$. According to the gradient formula for the cross-entropy loss, we can rewrite $g_{\theta_1}^{Multi}$ as follows,

$$g_{\theta_1}^{Multi} = \left(\frac{e^{(W_{y_i}^{m_1} z_i^{m_1} + b_1)}}{\sum_{k=1}^K e^{(W_k^{m_1} z_i^{m_1} + b_1)}} e^{(W_k^{m_2} - W_{y_i}^{m_2}) z_i^{m_2}} - 1 \right) W_{y_i}^{m_1} \quad (4)$$

where $W = [W^{m_1}, W^{m_2}]$, $W^{m_1} \in \mathbb{R}^{d_1}$, $W^{m_2} \in \mathbb{R}^{d_2}$, W_j denote the j_{th} row of W .

Since the y_{th} row of W is the class center of the y_{th} class [34, 35], the $z_i^{m_2}$ will gradually approach W_{y_i} and move away from W_k when optimizing the multimodal model. As a result, we can get the range of $e^{(W_k^{m_2} - W_{y_i}^{m_2}) z_i^{m_2}}$ as follow,

$$\begin{cases} e^{(W_k^{m_2} - W_{y_i}^{m_2}) z_i^{m_2}} < 1, k \neq y_i \\ e^{(W_k^{m_2} - W_{y_i}^{m_2}) z_i^{m_2}} = 1, k = y_i \end{cases} \quad (5)$$

Besides, let $W = W^{m_1}$ and $z^\tau = z^{m_1}$, we can get the classification result of the vanilla unimodal model. The gradient $g_{\theta_1}^{Uni}$ passed back to the encoder $\varphi_1(\theta_1, \cdot)$ in the vanilla unimodal model can be expressed as follows,

$$g_{\theta_1}^{Uni} = \left(\frac{e^{(W_{y_i}^{m_1} z_i^{m_1} + b_1)}}{\sum_{k=1}^K e^{(W_k^{m_1} z_i^{m_1} + b_1)}} - 1 \right) W_{y_i}^{m_1} \quad (7)$$

According to Equation 5 and 6, we can get the following inequality,

$$abs(g_{\theta_1}^{Uni}) > abs(g_{\theta_1}^{Multi}) > 0 \quad (8)$$

Because the parameters are updated along the negative gradient direction, the encoder $\varphi_1(\theta_1, \cdot)$ in the unimodal model will converge faster than that in the multimodal model. In other words, compared with unimodal models, multimodal models limit the optimization of modality encoders. Thus, each modality, including the dominant modality that has higher performance, in the multimodal

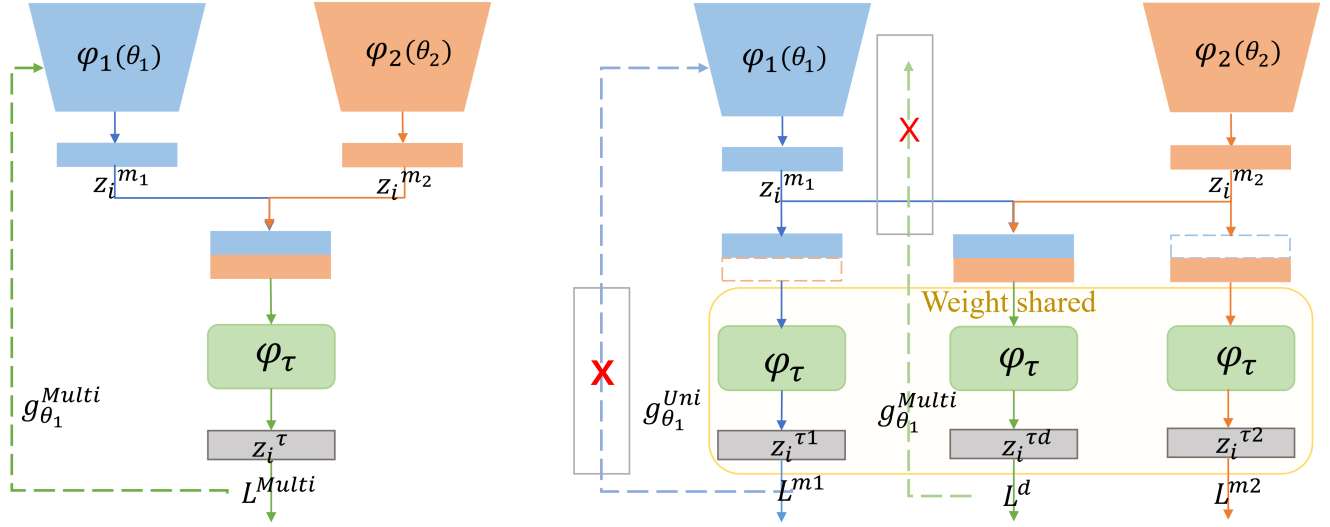


Figure 2. Architecture of vanilla multimodal model (a) and multimodal model with DGL (b). Compared with the vanilla model, DGL truncates the gradient back-propagated from the multimodal loss to the modality encoder and replaces it with the gradient from the unimodal loss, eliminating the gradient suppression on the modality encoder. Besides, DGL removes the gradient back-propagated from the unimodal loss to the modality fusion module, avoiding the gradient conflict between unimodal and multimodal losses. The red X means the operation of gradient truncation; the block rectangle denotes the truncation region.

model will suffer insufficient learning and has worse performance than that in the unimodal model (see Fig. 1 (b)). In addition, if m_2 is an easy-to-learn modality, $W_{y_i}^{m_2} z_i^{m_2}$ will be larger than $W_{y_i}^{m_1} z_i^{m_1}$, so $e^{(W_k^{m_2} - W_{y_i}^{m_2}) z_i^{m_2}}$ will be smaller than $e^{(W_k^{m_1} - W_{y_i}^{m_1}) z_i^{m_1}}$, which means that the gradient suppression caused by the easy-to-learn modality will be greater than that of hard-to-learn one. This explains why the easy-to-learn modality performs better than the hard-to-learn one when the modality gradients interfere with each other.

More importantly, even if the learning difficulty of m_1 and m_2 is the same, the optimization constraints on each modality encoder will also become increasingly severe when the multimodal model is optimized (see Fig. 1 (c)). This is because $W_{y_i}^{m_2} z_i^{m_2}$ and $W_{y_i}^{m_1} z_i^{m_1}$ becomes increasingly close to 1 and the absolute value of $g_{\theta_1}^{Multi}$ becomes increasingly smaller than that of $g_{\theta_1}^{Uni}$.

To ensure the optimization of modality encoders in the multimodal model, we should constraint $e^{(W_k^{m_2} - W_{y_i}^{m_2}) z_i^{m_2}} = 1$ so that $g_{\theta_1}^{Multi} = g_{\theta_1}^{Uni}$. Since each row of W is the class center vector of different types, the cross entropy loss will constrain them to stay away from each other to achieve good discrimination. Thus, the distance between $(W_k^{m_2}$ and $W_{y_i}^{m_2}$ will always be larger than 0. Consequently, the only way to ensure $e^{(W_k^{m_2} - W_{y_i}^{m_2}) z_i^{m_2}} = 1$ is to constrain $z_i^{m_2} = 0$. Nevertheless, this indicates that the fusion module entirely neglects the data from the modality m_2 , which deviates from the

Algorithm 1 Multimodal learning with DGL strategy

Input: Training dataset D , iteration number T , hyper-parameter α .

for $t = 0, \dots, T - 1$ **do**

- Feed-forward the batched data to the model.
- Calculate unimodal loss L^{m_1} and L^{m_2} via Equation 11 and 13.
- Calculate unimodal gradient.
- Eliminate the gradient passed back from the unimodal loss to the fusion module.
- Calculate multimodal loss L^d with detach representation via Equation 10.
- Calculate multimodal gradient.
- Update the model parameters.

purpose of multimodal fusion to integrate information from multiple sources. Therefore, there is an optimization conflict between the modality encoder and modality fusion module, which limits the learning of modality encoders.

3.2. Disentangled Gradient Learning

As discussed, compared with unimodal models, multimodal models constrain the optimization of modality encoders via $e^{(W_k^{m_2} - W_{y_i}^{m_2}) z_i^{m_2}}$. To this end, we introduce DGL to decouple the optimization process between the modality encoder and the modality fusion module in the multimodal model by detaching and reorganizing their gradient propagation paths, as shown in Fig. 2 (b). Compared with the vanilla

model, this can eliminate the optimization constraint on modality encoders completely while ensuring the optimization of the modality encoder and fusion module. Specifically, DGL consists of two parts: gradient detaching to truncate the gradient back-propagation of multimodal loss and gradient reorganizing to boost the optimization of each modality encoder.

Gradient Detaching. This stage aims to truncate the gradient $g_{\theta_1}^{Multi}$ transmitted from the modality fusion module to the modality encoder so that eliminates the suppression on the modality encoder optimization. To this end, we rewrite the Equation 2 as follows,

$$\begin{cases} z_i^{\tau d} = \varphi_{\tau}(\theta_{\tau}, z_i^{m_1}.detach(), z_i^{m_2}.detach()) \end{cases} \quad (9)$$

where ‘.detach()’ is the PyTorch function to truncate the gradient. Then we calculate the multimodal loss as follows,

$$L^d = L_{CE}(W(z_i^{\tau d}) + b, y_i) \quad (10)$$

Compared with vanilla loss L^{Multi} in Equation 1, $g_{\theta_1}^{Multi}$ calculated by the multimodal loss L^d will not be passed back to encoder $\varphi_1(\theta_1, \cdot)$, and thus will not affect its parameter optimization.

Gradient Reorganizing. Although gradient detaching eliminates the constraint on the modality encoder optimization, it also blocks the optimization of encoder $\varphi_1(\theta_1, \cdot)$ because no gradient can be passed back to the encoder. To address this problem, we consider introducing unimodal gradients for each encoder to optimize their parameters.

Unlike existing methods that introduce extra classifiers for modality representations, the proposed DGL calculates the unimodal loss via the modality dropout technique [29, 31]. This retains the fusion module and calculates the unimodal loss by setting other modality representations as 0 directly. This enables DGL to handle dense cross-modal interaction. More importantly, the module parameters are shared with unimodal and multimodal loss, introducing no extra structure complexity.

Specifically, take encoder $\varphi_1(\theta_1, \cdot)$ as an example, its unimodal loss is defined as follows,

$$\begin{cases} L^{m_1} = L_{CE}(W(z_i^{\tau 1}) + b, y_i) \end{cases} \quad (11)$$

$$\begin{cases} z_i^{\tau 1} = \varphi_{\tau}(\theta_{\tau}, z_i^{m_1}, 0) \end{cases} \quad (12)$$

where the parameters θ_{τ} , W and b are shared with those in L^d .

According to the chain rule, the gradient passed back to the encoder $\varphi_1(\theta_1, \cdot)$ is exactly the vanilla $g_{\theta_1}^{Uni}$ defined in Equation 7. Similarly, we can get the unimodal loss for encoder $\varphi_2(\theta_2, \cdot)$ as follows,

$$\begin{cases} L^{m_2} = L_{CE}(W(z_i^{\tau 2}) + b, y_i) \end{cases} \quad (13)$$

$$\begin{cases} z_i^{\tau 2} = \varphi_{\tau}(\theta_{\tau}, 0, z_i^{m_2}) \end{cases} \quad (14)$$

It is worth mentioning that the gradient of the unimodal loss also passes through the fusion module. To avoid conflicting with the multimodal gradient, we will set the unimodal gradient in the fusion module to 0 before calculating the gradient of the multimodal loss, as described in Algorithm 1.

So far, we have successfully eliminated the optimization conflict between the modality encoder and the modality fusion module when ensuring their respective optimization processes. More importantly, since the extra unimodal gradient paths for each modality encoder are independent, DGL also addresses the inter-modality competition that limits the performance of multimodal model [9, 21].

Total loss. As discussed above, the total loss L_{DGL} for multimodal learning with disentangled gradient learning is defined as follows,

$$L_{DGL} = [L^d, \alpha(L^{m_1} + L^{m_2})] \quad (15)$$

where L^d is the multimodal loss with gradient detaching. L^{m_1} and L^{m_2} are unimodal loss. Here we use the operation of concatenation instead of summation since the gradient of L^d and $L^{m_1} + L^{m_2}$ is decoupled, as described in Algorithm 1. Specifically, L^d only optimizes the modality fusion module and the fully connected layer in the multimodal model. $L^{m_1} + L^{m_2}$ only optimizes the modality encoders in the multimodal model. α is the hyper-parameter to calibrate the optimization of the modality encoder, which has been widely proven to improve the modality capacity and multimodal performance.

3.3. Relationship to Prior Work

Some concurrent works [28, 32] seem similar to our DGL, which also introduces unimodal loss to assist multimodal learning. However, there are still differences in motivations, implementations, and performance with ours. They focus on the optimization conflict between different modalities and leverage the unimodal loss to balance their optimization progress. Here, the unimodal loss and multimodal loss will be used to update all parameters simultaneously. In contrast, we address the optimization conflict between the modality encoder and the modality fusion module. And the unimodal model loss in DGL is only used to update the modality encoder, while multimodal loss is only used to update the fusion module and classifier, eliminating the interference between them. As a result, DGL achieves better performance in unimodal and multimodal evaluation than existing unimodal assistance methods (see Tables 1 and 2).

Besides, existing methods usually introduce extra classifiers for modality representations. In contrast, the proposed

Method	CREMA-D			KS			VGGSound		
	Audio	Visual	Multi	Audio	Visual	Multi	Audio	Visual	Multi
Audio-only	62.18	-	-	48.7	-	-	45.13	-	-
Visual-only	-	68.23	-	-	54.6	-	-	30.68	-
Concat	59.62	36.71	65.1	43.35	48.72	64.45	43.18	20.55	47.90
G-Blending	58.78	58.62	69.21	46.35	51.12	69.60	44.34	26.56	49.33
OGM_GE	57.76	40.09	68.82	44.23	45.81	66.89	41.85	27.41	48.71
PMR	55.11	38.34	67.44	43.61	46.67	65.70	42.33	25.12	48.17
AGM	56.37	43.54	69.61	46.12	47.65	68.88	41.87	27.34	49.10
MLA	60.46	64.23	73.12	<u>50.03</u>	<u>54.67</u>	<u>71.12</u>	<u>45.87</u>	<u>31.60</u>	<u>51.19</u>
MMPareto	59.43	61.09	70.12	48.40	52.42	69.83	42.44	30.07	49.12
D&R	<u>61.11</u>	<u>64.57</u>	<u>74.32</u>	49.78	54.88	69.10	45.18	31.23	50.84
DGL	63.12	69.11	77.48	52.89	60.11	74.78	47.13	33.45	52.53

Table 1. Comparison with existing modulation strategies on CREMA-D, Kinetics-Sounds, and VGGSound datasets. Bold and underline mean the best and second-best results, respectively. The proposed DGL achieves the best performance in both unimodal and multimodal performance comparisons. The metric is accuracy.

DGL overcomes those defects by retaining the fusion module and setting other modality representations as 0 to calculate the unimodal directly.

4. Experiments

4.1. Datasets

CREMA-D [3] serves as an audio-visual dataset crafted for speech emotion recognition. It contains 7442 video clips of 2-3 seconds from 91 actors for 6 emotions. The whole dataset is randomly divided into 6698 samples as the training set and 744 samples as the testing set.

Kinetics-Sounds (KS) [2] is a dataset formed by filtering the Kinetics dataset for 34 human action classes which have been chosen to be potentially manifested visually and aurally. This dataset contains 19k 10-second video clips (15k training, 1.9k validation, 1.9k test).

VGGSound [5] is a large-scale video dataset with 309 categories, capturing various audio events in daily life. For our experiment, we employed 168,618 videos for training and validation, and 13,954 videos for testing.

MOSI [38] is a popular Audio-Visual-Text dataset for the multimodal sentiment analysis task. It collects 2,199 utterance-video clips of 93 monologue videos from YouTube, each of which is labeled with a continuous sentiment score ranging from -3 (strongly negative) to 3 (strongly positive). In this paper, we use the two-class label setting that considers positive/negative results only. We use this dataset to prove the method’s generalization to the scene with multiple modalities.

NYUv2 [24] is a RGB-D dataset for indoor semantic segmentation task. It comprises 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. We use this dataset to prove the method’s effectiveness beyond the audio-visual dataset and classification task.

4.2. Comparison on the multimodal task

Implementation details. For a fair comparison, we adopt the same setting as the previous method [21] for the audio-visual task. For datasets CREMA-D, KS, and VGGSound, we adopt the ResNet18 as the encoder backbone, mapping the input data to 512-dimensional vectors. For CREMA-D, we extract 1 frame from each of the clips as the visual input. The whole audio data is transformed into a spectrogram of size 257×299 by librosa [20] using a window with a length of 512 and an overlap of 353. For the KS and VGGSound datasets, we extract 3 frames from each clip as visual inputs as visual input and process audio data into a spectrogram of size 257×1004. We use SGD with 0.9 momentum and 1e-4 weight decay as the optimizer. The learning rate is 2e-3 initially and multiplies 0.1 every 70 epochs. The training batch size and epoch are 100 and 64, respectively.

Comparison settings. To study the advantage of DGL, we make comparisons with three modulation approaches, OGM [21], AGM [17], PMR [9], and four alternating optimization methods, G-Blending [28], MLA [39], MM-Pareto [32] and D&R [33]. For a fair comparison, we unify the backbone as ResNet18 and the fusion method as concatenation in all experiments. Note that the original MLA uses epoch as 150 and batch size as 16. For a fair comparison, we unify them with other comparison methods as epoch 100 and batch size 64.

Results. As shown in Table 1, the proposed DGL framework achieves the best performance on all datasets with a significant improvement compared to existing methods. Compared to the second-best methods, MLA and D&R, it improves the multimodal performance by 3.15%, 3.66%, and 1.34% on CREMA-D, KS, and VGGSound datasets, respectively. These results demonstrate the superiority of the proposed DGL technique.

More importantly, the unimodal performance in the mul-

Method	CREMA-D			MOSI			
	Audio	Visual	Multi	Audio	Visual	Text	Multi
	MMTM-Fusion			MLP-Fusion			
Audio-only	59.18	-	-	44.25	-	-	-
Visual-only		68.34		-	54.28	-	-
Text-only	-	-	-		-	77.12	-
Multimodal Baseline	56.11	23.43	60.12	40.65	51.33	76.33	76.83
G-Blending	57.72	62.11	69.99	42.39	53.03	76.97	78.04
OGM_GE	55.61	40.61	65.31	42.09	52.68	75.87	77.97
PMR	54.71	39.21	64.89	41.89	52.43	75.67	77.54
AGM	55.29	44.18	66.65	42.24	52.97	76.04	77.95
MMPareto	56.79	63.21	69.88	42.92	53.08	76.87	78.15
D&R	<u>58.67</u>	<u>66.32</u>	<u>72.28</u>	<u>43.65</u>	<u>53.81</u>	<u>77.01</u>	<u>78.81</u>
DGL	60.30	69.50	75.00	44.78	55.11	77.39	79.78

Table 2. **Left:** Comparison with imbalanced multimodal learning methods with dense cross-modal interaction on the CREMA-D dataset. **Right:** Comparison with imbalanced multimodal learning methods on the MOSI dataset with three modalities.

timodal model trained with the DGL also outperforms that in the unimodal model. This verifies the effectiveness of DGL in promoting unimodal learning in multimodal learning. As a comparison, while existing gradient modulation methods, such as OGM_GE, can boost the weaker visual modality’s performance in CREMA-D above 36.71% compared to vanilla concatenation fusion, they also reduce the dominant audio modality’s performance below 59.62%. Then, the alternating optimization methods introduce unimodal assistance to improve the performance of audio and visual modalities in the multimodal model simultaneously. However, they overlook the optimization conflicts between the fusion module and encoders, failing to improve the unimodal performance in the multimodal model to be similar as that in the unimodal model. In contrast, DGL addresses this defect by truncating the gradient from the modality fusion module propagating to the modality encoder. As a result, the performance of the unimodal branch within the multimodal model trained with DGL surpasses that of the vanilla unimodal model.

4.3. Ablation Study

We first conduct experiments to discuss the influence of gradient truncation and hyper-parameters α , then we study the generalization of DGL to different fusion methods, modality types, modality numbers, and tasks. Limited by the page, more ablation experiments and discussion can be seen in the supplementary materials.

The effect of gradient truncation. We conduct experiments on the CREMA-D dataset to study the effectiveness of multimodal gradient truncation (MT) and unimodal gradient truncation (UT). As shown in Table 3, both MT and UT improve the performance significantly compared to the baseline model. The former avoids the gradient suppression from multimodal fusion to the unimodal encoder, and

Setting	Baseline	+MT	+UT	DGL(MT+UT)
ACC	65.10	73.31	74.22	77.48

Table 3. Ablation results of gradient truncation on the CREMA-D.

α	1	2	3	4	5
G-Blending	67.89	68.12	69.20	69.21	68.67
DGL	73.35	75.46	76.13	77.48	77.08

Table 4. Results of the DGL and G-Blending with different α on CREMA-D. The fusion strategy is concatenation.

the latter avoids the optimization conflict between multimodal and unimodal loss. Moreover, their combination can achieve better performance.

The effect of α . We investigate the influence of different α for DGL on the CREMA-D dataset. As shown in Table 4, the performance of DGL increases with parameter α , indicating that enhancing the optimization of unimodal learning can improve multimodal performance, which is consistent with the conclusion of existing methods [28, 32]. However, although existing methods, such as G-Blending, also introduce unimodal loss to boost unimodal learning in the multimodal model, their performance improvement is limited. This is because they ignore the optimization conflict between the modality encoder and the modality fusion module. In contrast, DGL eliminates this conflict and achieves significant improvement.

Generalization to dense cross-modal interaction. Existing experiments are conducted with vanilla concatenation fusion. To validate the applicability of DGL in the multimodal model with dense cross-modal interaction, we apply it to two intermediate fusion methods MMTM [16] and MLP-Mixer [25] for CREMA-D and MOSI datasets. Specifically, MMTM is a CNN-based architecture that fuses

Initialization	Method	
	ESANet	DGL
From Scratch	38.59	41.67
ImageNet Pre-train	48.48	50.10

Table 5. Model performance comparison of RGB-Depth semantic segmentation task on NYUv2 dataset. The metric is mIOU.

intermediate feature maps from different modalities via multimodal squeeze and excitation operations, while MLP-Mixer is an MLP-based architecture that performs fusion using affine transformations on intermediate feature maps. For both methods, we use only one frame for each video on each dataset to align their input on the benchmark.

Since MLA [39] only works with parameter-shared fusion, it is excluded from this experiment. For a fair comparison, we set the modality representation to 0 to obtain unimodal logits for G-Blending, OGM_GE, and MMPareto. As shown in Table 2, DGL consistently surpasses the multimodal baseline and other methods in both multimodal and unimodal performance. Additionally, the unimodal performance of the DGL multimodal model also exceeds that of the unimodal model. These results confirm the applicability of DGL in scenarios with dense cross-modal interaction.

Generalization to the multiple-modality case. Existing methods mainly only focus on the case of two modalities [9, 17, 21], limiting their applicability to broader, more complex scenarios. In contrast, DGL imposes no restrictions on the number of modalities, allowing for greater flexibility. Here, we conduct experiments on the MOSI dataset with three modalities: audio, vision, and text. For a comprehensive comparison, we retain the core uni-modal balancing strategy of G-Blending, OGM_GE, AGM, PMR, and MM-Pareto and extend them to more than two modality cases. As shown in the left section of Table 2, DGL achieves the best performance in multimodal and unimodal performance, demonstrating its flexibility in such scenarios.

Generalization to the dense prediction task. Existing methods require unimodal logit output or cluster center to calculate the modulation coefficients of different modalities. This limits their application in dense prediction scenarios. In contrast, DGL does not need these prerequisites, so it can be extended to non-classification tasks.

We take ESANet as the baseline, which is an efficient and robust model for RGB-D segmentation. We keep all hyper-parameters the same as the official implementation. Specifically, the model is optimized by Adam for 300 epochs with a mini-batch 8 and learning rate $1e-2$.

As shown in Table 5, we train the network on both scenarios: (1) training from scratch on NYUv2; (2) pre-training on ImageNet followed by fine-tuning on NYUv2. For both scenarios, the proposed DGL significantly enhances the performance of ESANet. This confirms the abil-

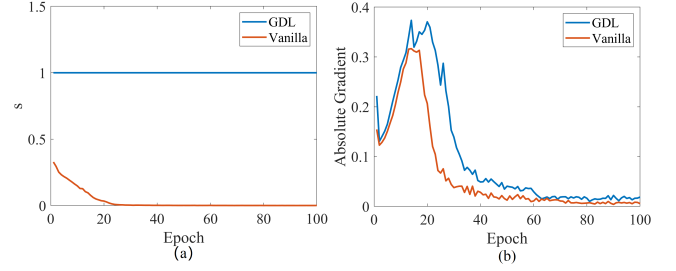


Figure 3. Visualization on the CREMA-D dataset. (a) illustrates $s = e^{(W_k^{m2} - W_{y_i}^{m2})z_i^{m2}}$ for the audio branch in the vanilla and DGL multimodal model. (b) illustrates the gradient of the audio branch in the vanilla and DGL multimodal model.

ity of DGL to handle tasks beyond the audio-visual data and classification.

Visualization. To further explain the mechanism of DGL, we visualize $s = e^{(W_k^{m2} - W_{y_i}^{m2})z_i^{m2}}$ and the training gradient for the audio branch in the vanilla and DGL multimodal model. As shown in Fig. 3(a), s of the vanilla model is always less than 1 and is close to 0 after the 20th epochs. As shown in Fig. 3(b), the gradient from vanilla multimodal back-propagated to the audio encoder also decreases rapidly after the 20th epoch. This is consistent with the derivation of Equation 3. On the contrary, DGL mitigates gradient suppression caused by $s < 1$ and maintains a large gradient to update the audio encoder after the 20th epoch. Therefore, as shown in Table 1 and 2, the audio branch in the DGL model outperforms that in the vanilla model significantly. This demonstrates the effectiveness of disentangled gradient learning.

5. Conclusion

In this study, we reveal the optimization conflict between the modality encoder and modality fusion module in the multimodal model. The cross-modal interaction will suppress the gradient back-propagated to each modality encoder and limit their optimization, including the dominant modality. To address this problem, we introduce DGL to decouple the optimization of the modality encoder and modality fusion module. First, DGL eliminates gradient suppression by truncating the gradient from the multimodal loss to the modality encoder and substituting it with the gradient from the unimodal loss. Additionally, DGL blocks the gradient from the unimodal loss to the modality fusion module, ensuring independent optimization of the modality encoder and fusion module. Extensive experiments show that DGL can achieve consistent performance gain on multimodal classification and segmentation tasks with different model structures and fusion methods.

References

- [1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018. 2
- [2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 6
- [3] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkov, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 6
- [4] 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. 1, 2
- [5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 6
- [6] 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. 2
- [7] Yuhang Ding, Xin Yu, and Yi Yang. Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3975–3984, 2021. 1, 2
- [8] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021. 1, 2
- [9] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multi-modal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023. 1, 2, 5, 6, 8
- [10] 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. 2
- [11] 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. 1
- [12] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. Deep multimodal multilinear fusion with high-order polynomial pooling. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2
- [13] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgb-d semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019. 1, 2
- [14] Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconciliation. *arXiv preprint arXiv:2405.09321*, 2024. 1, 3
- [15] 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. 1
- [16] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299, 2020. 7
- [17] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22214–22224, 2023. 1, 2, 3, 6, 8
- [18] 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. 1
- [19] Mengyuan Liu and Junsong Yuan. Recognizing human actions as the evolution of pose estimation maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1159–1168, 2018. 2
- [20] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015. 6
- [21] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022. 1, 2, 5, 6, 8
- [22] Stavros Petridis, Themis Stafylakis, Pinghuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6548–6552. IEEE, 2018. 2
- [23] 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. 1, 2
- [24] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 6
- [25] Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceed-*

- ings of the 30th ACM international conference on multimedia, pages 3722–3729, 2022. 7
- [26] Peng Sun, Wenhui 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. 1
- [27] Fei Tao and Carlos Busso. Aligning audiovisual features for audiovisual speech recognition. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 2
- [28] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 2, 5, 6, 7
- [29] Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20039–20049, 2023. 5
- [30] Shicai Wei, Yang Luo, Xiaoguang Ma, and Ren Peng. Msh-net: Modality-shared hallucination with joint adaptation distillation for remote sensing image classification using missing modalities. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 1
- [31] Shicai Wei, Yang Luo, Yuji Wang, and Chunbo Luo. Robust multimodal learning via representation decoupling. In *European Conference on Computer Vision*, pages 38–54. Springer, 2025. 5
- [32] Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730*, 2024. 1, 3, 5, 6, 7
- [33] Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *European Conference on Computer Vision*, pages 71–86. Springer, 2025. 1, 3, 6
- [34] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer vision—ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14*, pages 499–515. Springer, 2016. 3
- [35] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A comprehensive study on center loss for deep face recognition. *International Journal of Computer Vision*, 127:668–683, 2019. 3
- [36] Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. Mmcossine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1, 2
- [37] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 1, 2
- [38] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 6
- [39] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. *arXiv preprint arXiv:2311.10707*, 2023. 1, 2, 3, 6, 8
- [40] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. *arXiv preprint arXiv:2206.02425*, 2022. 1, 2
- [41] 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. 1