arXiv:2304.08028v1 [cs.CV] 17 Apr 2023

MMANet：面向不完整多模态学习的边界感知蒸馏与模态感知正则化

Shicai Wei Chunbo Luo Yang Luo 电子科技大学信息与通信工程学院shicaiwei@std.uestc.edu.cn {c.luo, luoyang}@uestc.edu.cn

摘要

近年来，多模态学习在众多场景中展现出巨大的潜力，并日益受到关注。然而，它常常面临模态数据缺失的问题，因此在实际应用中性能会严重下降。为此，我们提出了一种名为 *MMANet* 的通用框架，以辅助不完整多模态学习。该框架由三个组件构成：用于推理的部署网络、向部署网络传递全面多模态信息的教师网络，以及指导部署网络平衡弱模态组合的正则化网络。具体而言，我们提出了一种新颖的边界感知蒸馏（*MAD*），通过依据分类不确定性对样本贡献进行加权，辅助信息传递。这促使部署网络关注决策边界附近的样本，并学习更精细的类间间隔。此外，我们设计了一种模态感知正则化（*MAR*）算法，用于挖掘弱模态组合，并指导正则化网络为其计算预测损失。这促使部署网络自适应地提升其对弱模态组合的表征能力。最后，在多模态分类和分割任务上的大量实验表明，我们的 *MMANet* 显著优于当前最先进的方法。代码见：*https://github.com/shicaiwei123/MMANet*

1. 引言

多模态学习已在分类等众多视觉任务中取得巨大成功 [21, 33, 46], 目标检测 [26, 45, 53], 和分割 [5, 23, 41]。然而，大多数成功的方法都假设模型在训练和测试时使用相同的模态数据。事实上，受设备 [32, 39], 用户隐私 [13, 25], 和工作条件 [3, 29], 的限制，在推理阶段收集完整模态数据往往成本高昂，甚至不可行。因此，人们对利用训练期间的完整模态数据来辅助不完整模态乃至单模态推理产生了浓厚兴趣。

模态	定制	统一	Drop
RGB	10.01	11.75	-1.65
深度	4.45	5.87	-1.42
IR	11.65	16.62	-4.97
RGB+深度	3.41	4.61	-1.2
RGB+红外	6.32	6.68	-0.36
深度+红外	3.54	4.95	-1.41
RGB+深度+红外	1.23	2.21	-0.98

表1. 在 CASIA-SURF 数据集上使用平均分类错误率评估不同模态组合下定制模型和统一模型的性能。‘定制’是指针对每种组合分别训练一个模型，而‘统一’是指只为所有组合训练一个模型。所有模型的架构均相同，缺失模态的特征图（例如 RGB+深度组合中的红外模态）设为零。

一种典型的解决方案是根据可用模态重建缺失模态的样本或特征 [10, 14,15, 20, 29, 32]。然而，这需要针对所有可能的模态组合分别构建特定模型，因此复杂度较高。近期研究着重于为不同模态组合学习统一模型，而非一组网络。通常，这类方法 [6, 11, 12, 17, 51, 52] 尝试利用特征融合策略来捕捉模态不变表示，使模型能够适应所有可能的模态组合。例如，RFNet [11]设计了区域感知融合模块，用于融合可用图像模态的特征。

尽管现有统一模型确实能够提高多模态模型的训练和部署效率，但其性能可能不尽如人意。

如表1所示，对于不同的模态组合，定制模型的表现始终优于统一模型。这是因为现有的统一模型通常侧重于模态不变特征，而忽略了模态特定信息。需要注意的是，多种模态中互补的模态特定信息有助于优化类间判别并提升推理性能 [2, 18, 36]。这促使我们提出本文的第一个研究问题：**统一模型能否同时考虑模态不变信息和模态特定信息，同时保持对不完整模态输入的鲁棒性?**

为此，我们提出引导统一模型从使用完整模态训练的教师模型中学习全面的多模态信息。通过这种方式对目标任务损失进行正则化，促使统一模型获取不同模态组合之间的互补信息，同时保持对这些组合的泛化能力。具体而言，我们提出一种新颖的边界感知蒸馏（MAD），通过引导统一模型模仿教师模型的样本间关系来训练它。MAD 引入样本的分类不确定性，对其对最终损失的贡献进行重新加权。由于靠近类别边界的样本更容易被误分类，且具有更高的分类不确定性 [8], ，这会促使统一模型保留由互补线索细化的类间间隔，并学习模态特定信息。

现有统一方法的另一项局限是难以在不平衡训练问题上达到最佳性能。具体而言，传统多模态学习模型往往会拟合具有判别性的模态组合；面对弱模态组合时，其性能会显著下降。为解决这一问题，现有统一方法引入辅助判别器，以增强单模态组合的判别能力 [6, 11, 51]。这种做法基于这样一种假设：单一模态弱于多个模态。然而，如表 1 所示，无论是定制模型还是统一模型，单一深度模态的表现都优于 RGB、红外及其组合。这表明，由多个弱模态组成的组合可能比单一强模态更难优化。此外，如表 3 所示，RGB 成为强模态，而深度和红外则成为弱模态。这表明模态重要性并非固定不变，而是随场景变化。这些发现促使我们提出第二个研究问题：**如何在不同场景下有效优化弱模态组合?**

为此，我们设计了一个正则化网络和 MAR算法，以辅助统一网络的训练。具体而言，正则化网络会为所有输入生成额外预测。随后，MAR 挖掘并计算来自弱组合的样本的预测损失。这促使统一模型提升其对弱组合的表征能力。具体来说，MAR 利用记忆效应挖掘弱组合 [1, 16, 49]：深度神经网络往往会先记住简单样本，再对困难样本过拟合。如图5(a)所示，在训练早期，统一模型倾向于先拟合包含深度模态的样本。因此，MAR 首先利用记忆效应挖掘强模态，再将其余模态的组合确定为弱组合。

最后，我们提出了一种与模型和任务无关的框架 MMANet，通过结合所提出的 MAD 和 MAR 策略，助力不完整多模态学习。MMANet 能够引导统一模型获取全面的多模态信息，同时平衡强势与弱势模态组合的性能。针对多模态分类和分割任务开展的大量对比实验和消融实验，验证了 MMANet 的有效性。

2. 相关工作

2.1. 缺失模态下的多模态学习

大多数现有的多模态学习方法都假设所有实例均包含完整模态。然而，由于设备 [32, 39], 用户隐私 [13, 25], 和工作条件 [3, 29]等因素，这一假设在现实应用中并非总是成立。近年来，研究者提出了许多不完整多模态学习方法，大致可分为两类：定制方法和统一方法。定制方法旨在针对每一种不完整模态组合训练一个专门的模型，以恢复缺失模态。根据恢复目标，定制方法还可进一步分为基于样本的方法和基于表示的方法。基于样本的方法侧重于利用生成对抗网络在输入空间中填补缺失模态 [4,27,32,37]。由于样本重建过程复杂，这类方法通常不够稳定，且可能引入噪声，从而损害当前的主要任务 [34]。因此，研究者提出了基于表示的方法，通过知识蒸馏 [3, 14, 20, 29] 或矩阵补全 [30, 35]来重建样本表示。尽管这些方法取得了不错的效果，但对于每种缺失模态子集，都必须训练并部署一个专门的模型，这在实际应用中复杂度较高。

统一方法旨在通过提取模态不变特征，训练一个模型来处理不同的不完整模态组合。例如，HeMIS [17]通过对任意数量的可用模态计算统计量（即均值和方差），学习多模态信息的嵌入。此外，Chen 等人引入特征解耦，以消除模态特定

信息。此外，较新的方法（如 LCR） [52] 和 RFNet [11] 则通过不同的注意力机制，专注于提取模态不变表示。此外，mmFormer [51]引入 Transformer 模块，对模态不变嵌入中的全局语义信息进行建模。尽管这些方法取得了令人鼓舞的结果，但它们只考虑模态不变信息，而忽略了模态特定信息。因此，它们的表现通常远逊于定制方法，尤其是在缺失不止一种模态时 [48]。

2.2. 知识蒸馏

知识蒸馏旨在将知识从能力较强的教师网络迁移到较弱的学生网络，以促进监督学习。通常，蒸馏方法可分为三类：基于响应的蒸馏，通过匹配教师网络和学生网络的软化逻辑值来实现 [19],；基于表征的蒸馏，通过匹配特征图来实现 [24,28,40], ；以及基于关系的蒸馏，通过匹配样本关系来实现。 [38,47]。

知识蒸馏最初源于资源高效的深度学习，如今已在不完整多模态学习等领域得到更广泛的应用。在此，知识蒸馏用于将仅能在训练阶段获取的特权模态信息从教师网络迁移到学生网络 [3,29]。由于在不完整多模态学习中，教师网络和学生网络的输入不同，通过基于表征的方法迁移知识可能会导致过拟合 [15]。近期方法着重采用基于关系的方法迁移特权模态信息 [7,22,48]。然而，这些已有方法通常将不同实例一视同仁，忽略了它们的特异性，从而导致性能欠佳。

3. 方法

3.1. MMANet

本节介绍一种名为 MMANet 的通用框架，用于应对不完整多模态学习的挑战。如图 1所示，该框架由三部分组成：部署网络、教师网络和正则化网络。具体而言，部署网络即推理网络。为使其能够适应模态不完整的情况，MMANet 在模态编码器之后引入伯努利指示器 $\Delta = \{\delta_1 ... \delta_m\}$，并在训练阶段通过随机将 Δ 中的某些分量设为 0 来进行模态丢弃。对于缺失模态，相应的编码特征图将由零矩阵替代。此外，MMANet 引入了一个使用完整多模态数据预训练的教师网络，通过 MAD 将全面的多模态知识迁移至部署网络。这有助于部署网络同时获得模态不变特征和模态特定特征。最后，MMANet 通过 MAR 引导部署网络与正则化网络共同训练，后者针对较弱的模态组合生成额外预测。这缓解了较强模态组合的过拟合问题。用于引导部署网络训练的总损失定义如下：

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

其中，α 和 β 是超参数。L_{LT} 是任务学习损失，由当前的主要任务决定。例如，当主要任务是分类时，L_{LT} 可以是交叉熵损失。L_{MAD} 和 L_{MAR} 分别是 MAD 和 MAR 的损失。

此外，MMANet 中使用的其他符号定义如下。给定一个小批量多模态输入$x = \{x_1, ..., x_m\}$，$x_m \in R^b$ 表示 m_{th} 模态的数据。b 是批次大小。E_m^t 和 E_m^d 分别表示教师网络和部署网络中用于 m_{th} 模态的编码器。F^t 和 F^d 分别表示教师网络和部署网络使用的融合模块。$\Delta^d \in R^{b \times m}$ 是 Δ的向量。$z^t \in R^{b^t \times c^t \times h^t \times w^t}$和 $z^d \in R^{b^d \times c^d \times h^d \times w^d}$ 分别表示教师网络和部署网络的融合特征。其中，b 是批次大小，c 是输出通道数，h和 w是空间维度。P^t、P^r和 P^d分别表示教师网络、正则化网络和部署网络的任务预测器。y^t、y^r和 y^d 分别表示教师网络、正则化网络和部署网络的 $R^{b \times k}$ 预测矩阵。其中，k 是类别数。

3.2. MAD

本节介绍所提出的 MAD 策略，该策略旨在将教师网络中的全面多模态信息迁移到部署网络。如图 1 所示，MAD 在 z^t 和z^d之间进行。z^d 由于随机模态丢弃，简单模型的输出会发生变化。相比之下，样本语义保持不变。因此，MAD 提出通过关系一致性而非特征一致性来迁移教师的知识。这有助于避免过拟合，并防止损害部署网络的表征能力。此外，MAD 提出测量类别边界，并引导统一模型更加关注边界附近的样本。这可以促使开发网络继承教师网络经过优化的类间间隔。然而，由于类别边界通常不规则，因此很难检测。为解决这一问题，MAD 引入每个样本的分类不确定性，对其对总损失的贡献进行重新加权。由于类别边界附近的样本

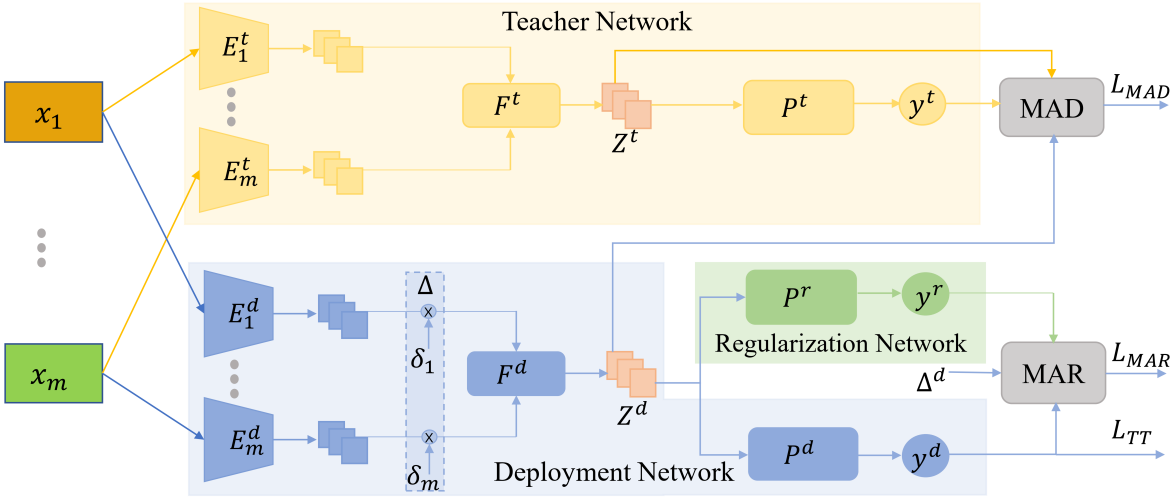


图 1. 所提出的 MMANet 概览。它由三个部分组成：用于最终推理的部署网络、将全面的多模态知识传递给部署网络的教师网络，以及引导部署网络平衡弱模态组合的正则化网络。

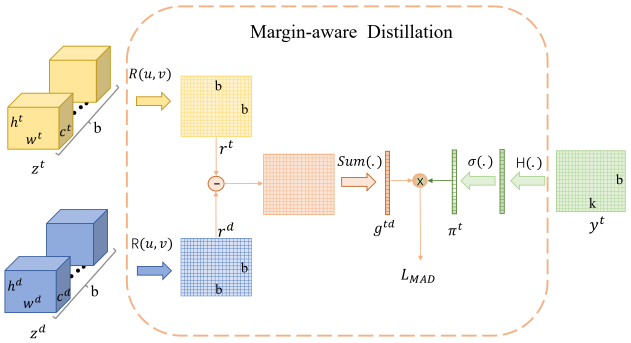


图 2. 所提出的 MAD 示意图。

更有可能被误分类，且具有更高的分类不确定性，因此可以借此实现对它们的关注。

MAD 的概览如图2所示。它以z^t、 z^d和 y^t 作为输入，并由三个步骤组成：(a) 计算关系差异向量 $g^{td} \in R^b$；(b) 计算分类不确定性向量 $\pi^t \in R^b$；以及 (c) 计算 MAD 的总损失 L_{MAD}。

(a) MAD 根据 z^t 和 z^d 计算 g^{td}。具体而言，MAD 首先将 z^t 和 z^d 重塑为 $z^{t'} \in R^{b^t \times c^t * h^t * w^t}$ 和 $z^{d'} \in R^{b^d \times c^d * h^d * w^d}$。然后，MAD 通过相同的关系函数 R(u,v) 分别计算关系矩阵 $r^t \in R^{b \times b}$ 和 $r^d \in R^{b \times b}$。表示教师网络的 i_{th} 和 j_{th} 个样本表示之间关系的 $r^t(i, j)$ 可表示如下：

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

此外，表示部署网络的 i_{th}和 j_{th} 样本表示之间关系的 $r^d(i, j)$ 可表示如下，

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

理论上， $R(u, v)$可以是用于度量向量距离的任意度量，例如欧氏距离和余弦距离。由于教师网络和部署网络的特征向量维度可能非常高，为了消除维度灾难，我们选择余弦距离作为 $R(u, v)$，

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

此外，MAD计算 r^t 与 r^d 之间的差异矩阵，并将每一行求和以得到 g^{td}。

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

这里， $g^{td}(i)$表示教师网络与部署网络从 i_{th} 样本到同一小批量中其他样本的关系差异。

(b) MAD根据 y^t计算 π^t。具体而言，它将每个样本的逻辑值输出的信息熵作为其分类不确定性。对于 i_{th} 样本，其分类不确定性 $\pi^t(i)$可表示如下，

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

其中， $\sigma(。)$是用于归一化的 softmax 函数。 $H(x)$是 x 的信息熵。具有较高

分类不确定性的样本通常更接近决策边界，因为它更有可能被误分类。因此，$\pi^t(i)$ $i \in [1,b]$ 也可以表示从 i_{th}样本表征到决策边界的间隔。

(c) 最后，MAD 将 $\pi^t(i)$ 作为相应组件 $g^{td}(i)$ 的权重来计算 L_{MAD}，

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

这促使部署网络关注决策边界附近的样本，并保留由教师网络的全面多模态信息优化得到的类间间隔。

3.3. MAR

本节介绍 MAR 算法，该算法促使部署网络自适应地提升对弱模态组合的判别能力。如图 1 所示，MAR 将 y^r、 y^d 和 Δ^d 作为输入来计算 L_{MAR}。具体而言，MAR 首先提出一种对比排序策略，用于挖掘弱模态组合。与简单地将仅包含单一模态的组合视为弱组合相比，该策略还考虑了包含多种模态的组合，因此能够获得更准确的挖掘结果。随后，MAR 计算弱模态组合的预测损失，引导部署网络更加关注这些组合。

MAR 的概览如图 3所示。它包含两个步骤：(a) 当 $E \leq N$ 时，挖掘弱模态组合集合 Ω；(b) 当 $E > N$ 时，计算L_{MAR}。其中， E 是当前训练轮次，N 是预热轮次的数量。

(a) MAR 使用对比排序从 y^d 中计算 Ω。MAR 提议在每个训练轮次结束后，计算训练集上 Δ_i，$i \in [0,m]$ 的预测输出 $Y^O \in R^{(m+1)\times n\times k}$ 。

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

n 是样本编号。 Δ_i 表示 Δ 的 i_{th} 分量为 0。Δ_0 表示 Δ 的所有分量均不为 0，这必然包含强模态。由于部署网络倾向于优先记住具有强模态的样本， Δ_w， $w \in [1,m]$ 使得 $Y^O(w,:,:)$ 与$Y^O(0,:,:)$ 之间的距离最大，这是一个不包含强模态的困难组合。Ω 的元素可确定为 Δ_w ，而其中的 Δ 由其包含的模态组成。

具体而言，为了使 Δ_w 能够适应神经网络学习的随机性，MAR 引入了两项创新。首先，MAR 根据预测分布 $Y^d \in R^{(m+1)\times k}$ 而非 Y^O计算预测差异，

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

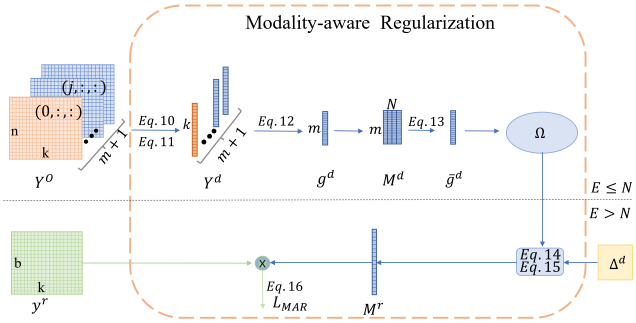


图 3. 所提出的 MAR 示意图。

其中 $j \in [0,k-1]$。与 Y^O相比， Y^d 只需要类别层面的一致性，而不需要样本层面的一致性。随后，向量差异向量 $g^d \in R^m$ 定义如下，

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

其中 KL(,) 表示 KL散度， $i \in [1,m]$。

其次，MAR 引入记忆库 $M^d \in R^{N\times m}$，用于保存预热轮次期间的 g^d ，并执行平均滤波以获得 g^d，

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

其中 $(M^d)_i$ 是第 i_{th} 训练轮次的 g^d 。而 Δ_w 可定义为 Δ_i ，其中 $i = argmax(g^d)$。

(b) MAR 根据 y^r、 Δ^d 和 Ω 计算 L_{MAR}。具体而言，MAR 首先根据 Δ^d 和 Ω 计算弱组合掩码 $M^r \in R^b$，

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

其中， $i \in [0,b-1]$、 $\Delta^d(i)$是该小批量中第 i_{th} 个样本的 Δ 。然后， L_{MAR} 定义如下，

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

其中， l 是 y^r的真实标签向量。 $[.]$ 表示索引算子。

4. 实验

我们在多模态分类和分割任务上进行实验，以评估所提出的 MMANet。接下来，我们首先将 MMANet 架构与这两项任务中的最先进方法进行比较。然后，我们对 MMANet 的 MAD 和 MAR 策略进行消融实验。

4.1. 多模态分类的性能与比较

数据集：我们以人脸反欺骗任务为例进行多模态分类实验，

表2. 在 m 上进行多模态分类任务的性能。CASIA-SURF。↓表示数值越低，性能越 ce.

模态			平均分类错误率(↓)						
			定制	统一					
RGB	深度	红外	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**
平均值			5.80	7.52	7.18	6.71	6.02	5.93	**3.94**

表3. 在 CeFA 数据集上的多模态分类任务性能。

模态			平均分类错误率(↓)		
			定制	统一	
RGB	深度	红外	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**
平均值			32.20	31.52	**27.94**

实验在 CASIA-SURF [50]和 CeFA [31] 数据集上进行。这两个数据集都包含 RGB、深度和红外模态。对于 CASIA-SURF，我们遵循作者提出的域内测试协议，将数据集划分为训练集、验证集和测试集，分别包含 29k、1k 和 57k 个样本。对于 CeFA，我们遵循作者提出的跨种族和跨攻击协议，将数据集划分为训练集、验证集和测试集，分别包含 35k、18k 和 54k 个样本。这里我们报告测试集上的结果，采用平均分类错误率（ACER）[50]。

实现细节：我们采用随机翻转、旋转和裁剪进行数据增强。所有模型均使用随机梯度下降优化，训练 100 轮次，每个小批量包含 64 个样本。学习率初始值设为 0.001，先进行 5 轮次的线性预热，并在第 16、33 和 50 轮次时降至原来的十分之一。权重衰减和动量分别设为 0.0005 和 0.9。

对比方法的超参数采用原论文推荐的设置。MMANet 的 (α，β) 在 CASIA-SURF 和 CeFA 上分别设为 (30，0。5) 和 (30，0。5)。预热轮次 N 设为 5。

比较：在此，我们将 MMANet 与两种不完整多模态学习的统一方法进行比较。一种是早期方法，仅专注于提取模态不变特征，例如 HeMIS [17] 和 LCR [52]。另一种是增强方法，进一步考虑提升单模态组合的判别能力，例如 RFNet [11],和 mmFormer [51]。

此外，我们引入了两种基线方法：SF 和 SF-MD。SF [50]是 CASIA-SURF 的基准方法，是一种定制方法，会针对每种模态组合分别训练模型。SF-MD 是 SF 的一种变体，仅在其模态编码器之后添加伯努利指示变量。这样，SF-MD 就成为一个针对所有模态组合训练单个模型的统一模型。

最后，为了公平比较，我们对所有对比方法均采用 SF 的基本实现。具体而言，我们将 HeMIS、LCR、RFNet 和 mmFormer 的模态编码器统一为 SF 中使用的 ResNet18。此外，我们将使用完整多模态数据训练的 SF 模型设为教师网络，并将 SF-MD 模型设为开发网络。

结果：表2 和表3 分别展示了在 CASIS-SURF 和 CeFA 数据集上与最先进方法的比较结果。与排名第二的统一方法 mmFormer 相比，MMANet 在 CASIS-SURF 和 CeFA 上的平均分类错误率分别降低了 1.99% 和 3.58%。此外，可以看到，在 CASIA-SURF 的全部九种模态组合上，MMANet 在两个数据集上均取得了最佳性能。这表明我们的方法在不完整多模态分类任务中的优越性。更重要的是，在 CASIA-SURF 和 CeFA 的所有模态组合上，MMANet 的表现甚至优于定制基线方法 SF，平均分类错误率分别降低了 1.86% 和 4.26%。这证明了所提出的 MAD 和 MAR 对不完整多模态分类任务的有效性。

表4. 性能 在 NYUv2 多模态分割任务上的表现。↑表示数值越高，性能越好 .

模态		平均交并比(↑)						
RGB	深度	定制	统一					
		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**
平均值		44.65	42.77	34.07	43.08	43.92	44.26	**45.58**

表 5. 在 Cityscapes 数据集上的多模态分割任务性能。

模态		平均交并比(↑)		
RGB	深度	定制	统一	
		ESANet	mmFormer	MMANet
●	○	77.60	76.62	**77.61**
○	●	59.11	58.53	**60.12**
●	●	78.62	78.01	**78.89**
平均值		71.77	71.05	**72.20**

4.2. 多模态分割的性能与比较

数据集：我们以语义分割任务为多模态分割的示例，并在 NYUv2 [43] 和 Cityscapes [9]数据集上开展实验。两者均由 RGB 和深度模态组成。具体而言，NYUv2 包含 1,449 张室内 RGB-D 图像，其中 795 张用于训练，654 张用于测试。我们采用常用的 40类标签设置。Cityscapes 是一个用于城市场景理解的大规模户外 RGB-D 数据集，包含 5,000 个精细标注的样本，分辨率为 2048×1024，其中 2,975 个用于训练，500 个用于验证，1,525 个用于测试。Cityscapes 还提供 20k 张粗略标注的图像，我们未将其用于训练。对于这两个数据集，我们使用平均交并比（mIOU）指标报告验证集上的结果。

实现：我们使用随机缩放、裁剪、颜色抖动和翻转进行数据增强。所有模型均使用 Adam 优化器训练 300 个训练轮次，每个小批量包含 8 个样本。学习率初始化为 1e-2，并通过 PyTorch 的单周期调度器 [44]进行调整。

对比方法的超参数采用其论文中建议的设置。MMANet 的超参数 (α，β) 在 NYUv2 和 Cityscapes 数据集上分别设为 (4，0。2) 和 (10，0。1)。预热轮数 N 设为 10。

比较：我们还将 MMANet 与 HeMIS [17], LCR [52], RFNet [11], 和 mmFormer [51]进行了比较。在此，我们将 ESANet 和 ESANnet-MD 设为基线。ESANet [42] 是一种高效且稳健的 RGB-D 分割模型，会针对每种模态组合分别训练模型。ESANet-MD 是 ESANet 的一个变体，只需在其模态编码器后添加伯努利指示变量即可。ESANet-MD 只需为所有模态组合训练一个模型。最后，为了进行公平比较，我们将 HeMIS、LCR、RFNet 和 mmFormer 的模态编码器统一为 ESANet 中采用的带有 NBt1 的 ResNet50。此外，我们将使用完整多模态数据训练的 ESANet 模型设为教师网络，并将 ESANet-MD 模型设为开发网络。

结果：表4 和表5分别列出了 NYUv2 和 Cityscapes 数据集上的对比结果。从这些结果可以看出，对于所有模态组合，MMANet 在两个数据集上均取得了最佳性能。具体而言，与排名第二的方法 mmFormer 相比，MMANet 在 NYUv2 和 Cityscapes 数据集上的表现分别高出 1.32% 和 1.05%。这证明了 MMANet 在多模态分割任务中的有效性。此外，MMANet 在 NYUv2 数据集上的平均性能比 ESANet-MD 提高了 2.81%，甚至在 NYUv2 和 Cityscapes 数据集上分别比定制基线 ESANet 高出 0.97% 和 0.43%。这表明所提出的 MAD 和 MAR 在不完整多模态分割任务中的有效性。

5. 消融实验

本节将研究 MAD 和 MAR 的有效性，并在四个数据集上进行广泛的消融实验。受篇幅限制，我们仅展示 CASIA-SURF 数据集的结果，其他结果请参阅补充材料。

5.1. MAD 的有效性

为研究 MAD 的效果，我们进行实验，比较原始 SF-MD与采用 SP 和 MAD 的变体的性能。这里，SP 是 MAD 的一种退化方法，通过直接匹配教师网络与部署网络之间样本表示的余弦距离来传递知识。结果如表6所示。我们可以看到，SF-MD 的变体始终

表6. CASIA-SURF数据集上MAD的消融实验结果。

RGB	深度	红外	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**
平均值			7.5	5.9	**4.57**

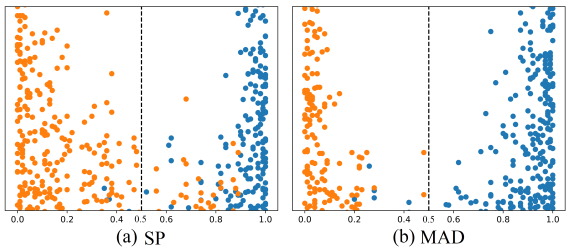


图4. 在SP和MAD辅助下，SF-MD模型在CASIA-SURF数据集上的预测分布。X轴表示归一化的逻辑值输出，x=0.5为分类边界。橙色和蓝色的点表示两个不同的类别。

表7. MAR 在 CASIA-SURF数据集上的消融实验结果。

RGB	深度	红外	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
平均值			4.57	4.33	**3.94**

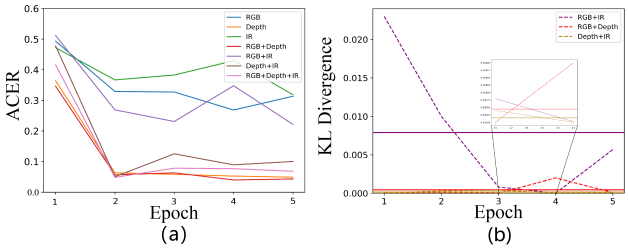


图5. (a) 预热阶段，CASIA-SURF数据集上不同模态组合的学习过程。(b) 预热阶段相应的 $g^d \in R^3$（虚线）及其平均结果 g^d _(实线）。

优于原始 SF-MD，在所有模态组合下均如此，平均性能分别提升了 1.6% 和 2.93%。这证明了将全面的多模态信息从教师网络传递到部署网络的有效性。此外，所提出的 MAD 比 SP 高出 1.33%，这证明了通过分类不确定性重新加权样本损失的有效性。这是因为基于分类不确定性重新加权可以促使部署网络关注困难样本，从而获得比传统 SP 更具可分性的类间间隔（见图4）。

5.2. MAR的有效性

为研究 MAR 的效果，我们进行了实验，对比 SF-MAD（即结合 MAD 的 SF-MD）及其采用 SR 和 MAR 的变体的性能。这里的 SR 是传统的模态正则化策略，仅考虑单模态组合。如表 6所示，SR 和 MAR 分别使 SF-MAD 的平均性能提升了 0.24% 和 0.63%，表明它们对单一模态组合和弱模态组合具有正则化作用。此外，MAR 的平均表现比 SR 高出 0.39%，证明了 MAR 的优越性。

SR 和 MAR 的平均增益低于 SP 和 MAD，因为它们旨在提升的仅是弱模态组合，而非所有模态组合。具体而言，如表6所示，表现最差的三种模态组合是‘RGB’、‘IR’和‘RGB+IR’。然而，SR 仅关注单一模态组合：RGB（1.19%）、IR（1.46%）和深度（0.65%），其中‘深度’恰好是一种强模态。相比之下，图5(b)显示，‘RGB+IR’与‘RGB+深度+IR’之间的预测差异最大。MAR 的性能增益则主要来自 RGB（1.79%）、IR（1.63%），以及 RGB 和 IR 的组合（1.02%）。这些结果表明，MAR 能够更准确地挖掘弱模态组合，并促使部署网络提升对这些组合的判别能力。

6. 结论

本文提出了MMANet框架，以辅助不完整多模态学习中的部署网络。具体而言，MMANet引入了一个使用完整多模态数据进行预训练的教师网络，通过MAD将全面的多模态信息传递给部署网络。这有助于部署网络在保持对不完整模态输入的鲁棒性的同时，获取模态不变信息和特定信息。此外，MMANet通过MAR引入正则化网络，以挖掘并正则化较弱的模态组合。这促使部署网络有效且自适应地提升对这些模态组合的判别能力。最后，大量实验表明，所提出的MMANet、MAD和MAR在不完整多模态学习中具有良好的有效性。

参考文献

[1] Devansh Arpit、Stanisław Jastrzebski、Nicolas Ballas、David Krueger、Emmanuel Bengio、Maxinder S Kanwal、Tegan Maharaj、Asja Fischer、Aaron Courville、Yoshua Bengio等。深入探究深度网络中的记忆。发表于国际机器学习会议，第233–242页。PMLR，2017年。[2] Tadas Baltrušaitis、Chaitanya Ahuja和Louis-Philippe Morency。多模态机器学习：综述与分类。*IEEE*模式分析与机器智能汇刊，41(2)：423–443，2018年。[3] Jonathan C、David A Stroud、Chen Ross、Jia Sun、Rahul Deng和Sukthankar。D3d：用于视频动作识别的蒸馏三维网络。发表于*IEEE/CVF*冬季计算机视觉应用会议论文集，第2755–2764页，2020年。[4] Lei Cai、Zhengyang Wang、Hongyang Gao、Dinggang Shen和Shuiwang Ji。用于多模态缺失数据补全的深度对抗学习。发表于*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第1158–1166页，2018年。[5] Jinming Cao、Hanchao Leng、Dani Lischinski、Daniel Cohen-Or、Changhe Tu和Yangyan Li。Shapeconv：用于室内RGB-D语义分割的形状感知卷积层。发表于*IEEE/CVF* 国际计算机视觉会议论文集，第7088–7097页，2021年。[6] Cheng Chen、Qi Dou、Yueming Jin、Hao Chen、Jing Qin和Pheng-Ann Heng。通过特征解耦与门控融合实现鲁棒的多模态脑肿瘤分割。发表于医学图像计算与计算机辅助干预国际会议，第447–456页。施普林格，2019年。[7] Cheng Chen、Qi Dou、Yueming Jin、Quande Liu和Pheng Ann Heng。利用特权多模态知识进行单模态分割。*IEEE* 医学影像汇刊，41(3)：621–632，2021年。[8] Zhuangzhuang Chen、Jin Zhang、Zhuonan Lai、Jie Chen、Zun Liu和Jianqiang Li。用于深度裂缝识别的几何感知引导损失。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第4703–4712页，2022年。[9] Marius Cordts、Mohamed Omran、Sebastian Ramos、Timo Rehfeld、Markus Enzweiler、Rodrigo Benenson、Uwe Franke、Stefan Roth和Bernt Schiele。用于城市场景语义理解的Cityscapes数据集。发表于*IEEE*计算机视觉与模式识别会议论文集，第3213–3223页，2016年。[10] Nieves Crasto、Philippe Weinzaepfel、Karteek Alahari和Cordelia Schmid。Mars：用于动作识别的运动增强RGB流。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第7882–7891页，2019年。[11] Yuhang Ding、Xin Yu和Yi Yang。RFNet：用于不完整多模态脑肿瘤分割的区域感知融合网络。发表于*IEEE/CVF* 国际计算机视觉会议论文集，第3975–3984页，2021年。

[12] Reuben Dorent、Samuel Joutard、Marc Modat、S´ebastien Ourselin 和 Tom Vercauteren。用于联合模态补全与分割的异模态变分编码器-解码器。载于医学图像计算与计算机辅助干预国际会议，第 74–82 页。施普林格，2019年。[13] Kai Fan、Wei Jiang、Hui Li 和 Yintang Yang。用于医疗隐私保护的轻量级物联网RFID协议。*IEEE*工业信息学汇刊，14 1665，2018 4 1656( ): – 。[14] Nuno C Garcia、Pietro Morerio 和 Vittorio Murino。用于动作识别的多流网络模态蒸馏。载于欧洲计算机视觉会议论文集，第 103–118 页，2018年。[15] Nuno C Garcia、Pietro Morerio 和 Vittorio Murino。通过对抗性判别式模态蒸馏利用特权信息进行学习。*IEEE*模式分析与机器智能汇刊，42(10):2581–2593，2019年。[16] Bo Han、Quanming Yao、Xingrui Yu、Gang Niu、Miao Xu、Weihua Hu、Ivor Tsang 和 Masashi Sugiyama。Co-teaching：利用极度噪声标签进行深度神经网络的鲁棒训练。神经信息处理系统进展，31，2018年。[17] Mohammad Havaei、Nicolas Guizard、Nicolas Chapados 和 Yoshua Bengio。HeMIS：异模态图像分割。载于医学图像计算与计算机辅助干预国际会议，第 469–477 页。施普林格，2016年。[18] Jack Hessel 和 Lillian Lee。我的多模态模型是否学会了跨模态交互？这比你想象的更难判断！*arXiv*预印本 *arXiv:2010.06572*，2020年。[19] Geoffrey Hinton、Oriol Vinyals 和 Jeff Dean。在神经网络中蒸馏知识。*arXiv*预印本 *arXiv:1503.02531*，2015年。[20] Judy Hoffman、Saurabh Gupta 和 Trevor Darrell。通过模态幻觉利用辅助信息进行学习。载于*IEEE*计算机视觉与模式识别会议论文集，第 826–834 页，2016年。[21] Danfeng Hong、Jingliang Hu、Jing Yao、Jocelyn Chanussot 和 Xiao Xiang Zhu。用于土地覆盖分类的多模态遥感基准数据集及共享与特有特征学习模型。*ISPRS*摄影测量与遥感杂志，178:68–80，2021年。[22] Minhao Hu、Matthis Maillard、Ya Zhang、Tommaso Ci- ceri、Giammarco La Barbera、Isabelle Bloch 和 Pietro Gori。将知识从多模态分割网络蒸馏到单模态分割网络。载于医学图像计算与计算机辅助干预国际会议，第 772–781 页。施普林格，2020年。[23] Xinxin Hu、Kailun Yang、Lei Fei 和 Kaiwei Wang。ACNet：一种利用互补特征进行RGB-D语义分割的基于注意力的网络。载于*2019*年*IEEE*国际图像处理会议（*ICIP*），第 1440–1444 页。IEEE，2019年。[24] Zehao Huang 和 Naiyan Wang。喜欢你所喜欢的：通过神经元选择性迁移进行知识蒸馏。*arXiv*预印本 *arXiv:1707.01219*，2017年。

[25]Mimansa Jaiswal 和 Emily Mower Provost。用于情绪识别的隐私增强型多模态神经表征。载于*AAAI*人工智能会议论文集，第 34 卷，第 7985–7993 页，2020年。[26] Wen-Da Jin、Jun Xu、Qi Han、Yi Zhang 和 Ming-Ming Cheng。Cdnet：用于RGB-D显著性目标检测的互补深度网络。*IEEE*图像处理汇刊，30:3376–3390，2021年。[27] Jiang Jue、Hu Jason、Tyagi Neelam、Rimner Andreas、Berry L Sean、Deasy O Joseph 和 Veeraraghavan Harini。融合跨模态幻觉MRI与CT，以辅助纵隔肺肿瘤分割。载于医学图像计算与计算机辅助干预国际会议，第 221–229 页。施普林格，2019年。[28] Nikos Komodakis 和 Sergey Zagoruyko。更加关注注意力：通过注意力迁移提升卷积神经网络的性能。载于国际学习表征会议，2017年。[29] Xiao Li、Lin Lei、Yuli Sun 和 Gangyao Kuang。利用缺失异质性图像进行土地覆盖分类的动态层次注意力蒸馏与协同实例选择。*IEEE*地球科学与遥感汇刊，60:1–16，2021年。[30] Yijie Lin、Yuanbiao Gou、Zitao Liu、Boyun Li、Jiancheng Lv 和 Xi Peng。Completer：通过对比预测实现不完整多视图聚类。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 11174–11183 页，2021年。[31] Ajian Liu、Zichang Tan、Jun Wan、Sergio Escalera、Guodong Guo 和 Stan Z Li。CASIA-SURF CeFA：多模态跨种族人脸反欺骗基准。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，第 1179–1187 页，2021年。[32] Ajian Liu、Zichang Tan、Jun Wan、Yanyan Liang、Zhen Lei、Guodong Guo 和 Stan Z Li。通过对抗性跨模态转换实现人脸反欺骗。*IEEE*信息取证与安全汇刊，16:2759–2772，2021年。[33] Ajian Liu、Jun Wan、Sergio Escalera、Hugo Jair Escalante、Zichang Tan、Qi Yuan、Kai Wang、Chi Lin、Guodong Guo、Isabelle Guyon 等。CVPR2019多模态人脸反欺骗攻击检测挑战赛。载于*IEEE/CVF*计算机视觉与模式识别会议研讨会论文集，第 0–0 页，2019年。[34]Haojie Liu、Shun Ma、Daoxun Xia 和 Shaozi Li。SFANet：用于可见光-红外行人重识别的频谱感知特征增强网络。*arXiv*预印本 *arXiv:2102.12137*，2021年。[35] Jiyuan Liu、Xinwang Liu、Yi Zhang、Pei Zhang、Wenxuan Tu、Siwei Wang、Sihang Zhou、Weixuan Liang、Siqi Wang 和 Yuexiang Yang。面向不完整多视图数据的自表征子空间聚类。载于第*29*届*ACM*多媒体国际会议论文集，第 2726–2734 页，2021年。[36] Yan Lu、Yue Wu、Bin Liu、Tianzhu Zhang、Baopu Li、Qi Chu 和 Nenghai Yu。通过共享-特有特征迁移实现跨模态行人重识别。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 13379–13389 页，2020年。

[37] Yongsheng Pan、Mingxia Liu、Chunfeng Lian、Yong Xia 和 Dinggang Shen。利用空间约束 Fisher 表征对不完整多模态神经影像进行脑疾病识别。*IEEE* 医学影像汇刊，39(9):2965–2975，2020年。[38] Nikolaos Passalis 和 Anastasios Tefas。利用概率知识迁移学习深度表征。载于欧洲计算机视觉会议论文集，第268–284页，2018年。[39] Allan Pinto、Helio Pedrini、William Robson Schwartz 和 Anderson Rocha。通过时频立方体的视觉码本进行人脸欺骗检测。*IEEE* 图像处理汇刊，24(12):4726–4740，2015年。[40] Adriana Romero、Nicolas Ballas、Samira Ebrahimi Kahou、Antoine Chassang、Carlo Gatta 和 Yoshua Bengio。FitNets：为轻量深度网络提供提示。*arXiv* 预印本 *arXiv:1412 6550*，2014。。[41] Daniel Seichter、Mona K¨ohler、Benjamin Lewandowski、Tim Wengefeld 和 Horst-Michael Gross。面向室内场景分析的高效 RGB-D 语义分割。载于*2021 IEEE* 国际机器人与自动化会议（*ICRA*），第13525–13531页。IEEE，2021年。[42] Daniel Seichter、Mona K¨ohler、Benjamin Lewandowski、Tim Wengefeld 和 Horst-Michael Gross。面向室内场景分析的高效 RGB-D 语义分割。载于*2021 IEEE* 国际机器人与自动化会议（*ICRA*），第13525–13531页。IEEE，2021年。[43] Nathan Silberman、Derek Hoiem、Pushmeet Kohli 和 Rob Fergus。从 RGB-D 图像进行室内分割和支撑关系推断。载于欧洲计算机视觉会议，第746–760页。施普林格，2012年。[44] Leslie N Smith 和 Nicholay Topin。超收敛：利用大学习率实现神经网络的超快速训练。载于多域作战应用中的人工智能与机器学习，第11006卷，第369–386页。国际光学工程学会，2019年。[45] Peng Sun、Wenhu Zhang、Huanyu Wang、Songyuan Li 和 Xi Li。结合深度敏感注意力与自动多模态融合的深度 RGB-D 显著性检测。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第1407–1417页，2021年。[46] Haiman Tian、Yudong Tao、Samira Pouyanfar、Shu-Ching Chen 和 Mei-Ling Shyu。用于视频分类的多模态深度表征学习。万维网，22(3):1325–1341，2019年。[47] Frederick Tung 和 Greg Mori。保持相似性的知识蒸馏。载于*IEEE/CVF*国际计算机视觉会议论文集，第1365–1374页，2019年。[48] Yixin Wang、Yang Zhang、Yang Liu、Zihao Lin、Jiang Tian、Cheng Zhong、Zhongchao Shi、Jianping Fan 和 Zhiqiang He。ACN：用于缺失模态脑肿瘤分割的对抗协同训练网络。载于医学图像计算与计算机辅助干预国际会议，第410–420页。施普林格，2021年。[49] Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht 和 Oriol Vinyals。理解深度学习（仍然）

需要重新思考泛化能力。*ACM*通信，64(3):107–115，2021年。[50] Shifeng Zhang、Xiaobo Wang、Ajian Liu、Chenxu Zhao、Jun Wan、Sergio Escalera、Hailin Shi、Zezheng Wang 和 Stan Z Li。用于大规模多模态人脸反欺骗的数据集与基准。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第919–928页，2019年。[51] Yao Zhang、Nanjun He、Jiawei Yang、Yuexiang Li、Dong Wei、Yawen Huang、Yang Zhang、Zhiqiang He 和 Yefeng Zheng。mmFormer：用于脑肿瘤分割不完整多模态学习的多模态医学Transformer。*arXiv* 预印本 *arXiv:2206.02425*，2022年。[52] Tongxue Zhou、St´ephane Canu、Pierre Vera 和 Su Ruan。通过潜在多源相关性表征，实现缺失模态下的脑肿瘤分割。载于医学图像计算与计算机辅助干预国际会议，第533–541页。施普林格，2020年。[53] Tao Zhou、Deng-Ping Fan、Ming-Ming Cheng、Jianbing Shen 和 Ling Shao。RGB-D显著性目标检测：综述。计算视觉媒体，7(1):37–69，2021年。