Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(2), 638-650
Published Online March 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.152061

Hans汉斯

RTRFNet：面向RGB-T分割的鲁棒融合网络

——应对传感器模态缺失的鲁棒性研究

谈焜宇，洪智勇*，熊利平*

五邑大学电子与信息工程学院，广东 江门

收稿日期：2026年2月3日；录用日期：2026年3月13日；发布日期：2026年3月25日

摘　要

RGB-热红外(RGB-T)语义分割对于在弱光或黑暗等复杂环境中运行的机器人系统至关重要。然而，传统的多模态融合方法往往导致不同模态特征的高度耦合，使得模型在真实场景下遭遇传感器信号缺失时极度脆弱，性能发生严重退化。为此，本文提出了RTRFNet，一个基于教师－学生学习机制的多模态鲁棒网络，从根本上打破了推理阶段对双源输入的强依赖。在训练阶段，网络通过一个精简的通道注意力特征融合模块(CA-FFM)高效聚合跨模态互补线索，为中枢的轻量化感知头(教师分支)构建完备的联合语义表示；随后，引入多模态知识蒸馏(MKD)策略，利用教师分支输出的高质量软分布，隐式地监督并引导完全独立的RGB与热红外双流网络(学生分支)，促使其充分获得并内化跨模态的丰富上下文知识。这种联合训练机制赋予了系统在推理阶段极高的灵活性：系统既能在全模态下移除教师网络并执行极具参数效率的决策层均值融合，也能在单传感器失效时仅激活存活链路进行高精度的独立推理。在主流基准数据集上的大量实验证明，RTRFNet不仅维持了全模态下的前沿精度，更在单模态缺失的极端条件下展现出了卓越的鲁棒性与轻量化部署优势。

关键词

RGB-T语义分割，多模态融合，鲁棒性，模态缺失

RTRFNet: A Robust Fusion Network for RGB-T Segmentation

—Robustness Research against Missing Sensor Modalities

Kunyu Tan, Zhiyong Hong*, Liping Xiong*

School of Electronics and Information Engineering, Wuyi University, Jiangmen Guangdong

*通讯作者。

文章引用: 谈焜宇, 洪智勇, 熊利平. RTRFNet: 面向 RGB-T 分割的鲁棒融合网络[J]. 人工智能与机器人研究, 2026, 15(2): 638-650. DOI: 10.12677/airr.2026.152061

Received: February 3, 2026; accepted: March 13, 2026; published: March 25, 2026

Abstract

RGB-Thermal (RGB-T) semantic segmentation is crucial for robotic systems operating in complex environments such as low-light or dark conditions. However, traditional multimodal fusion methods often lead to highly coupled modal features, making models extremely vulnerable to severe performance degradation when encountering missing sensor signals in real-world scenarios. To address this, this paper proposes RTRFNet, a robust multimodal network based on a teacher-student learning mechanism, which fundamentally breaks the strong reliance on dual-source inputs during the inference phase. During training, the network efficiently aggregates cross-modal complementary cues through a lightweight Channel Attention Feature Fusion Module (CA-FFM) to build a comprehensive joint semantic representation for a central lightweight perception head (teacher branch). Subsequently, a Multimodal Knowledge Distillation (MKD) strategy is introduced. It utilizes the high-quality soft distributions output by the teacher branch to implicitly supervise and guide the completely independent RGB and thermal dual-stream networks (student branches), prompting them to acquire and internalize rich cross-modal contextual knowledge. This joint training mechanism endows the system with extremely high flexibility during inference: by removing the teacher network, the system can perform highly parameter-efficient decision-level mean fusion under full modalities, or solely activate the surviving link for high-precision independent inference when a single sensor fails. Extensive experiments on mainstream benchmark datasets demonstrate that RTRFNet not only maintains state-of-the-art accuracy under full modalities but also exhibits exceptional robustness and lightweight deployment advantages under extreme missing modality conditions.

Keywords

RGB-T Semantic Segmentation, Multimodal Fusion, Robustness, Missing Modality

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

在自动驾驶等真实场景的部署中，多模态感知能力对于实现可靠的环境理解至关重要。具体而言，可见光(RGB)与热红外(Thermal)传感器的联合应用并非信息的简单叠加，而是模态间的优势互补：RGB图像提供了高分辨率的纹理和几何细节，而热红外成像则在低照度和极端天气条件下表现出极强的环境鲁棒性。两种模态信息的有效交互与融合，为实现全天候、复杂场景下的稳健感知提供了关键支持。

尽管现有的 RGB-T 语义分割研究在提升全模态精度方面取得了显著进展，但绝大多数工作主要聚焦于融合架构的设计，而忽视了模型在动态环境下的鲁棒性。这些传统的主流方法大致可分为两类：一类是基于直接操作的简单融合如图 1(a)，比如 MFNet [1]，通常通过特征的逐元素相加或通道拼接来整合信息；另一类则是追求极致互补性的深度特征交互融合如图 1(b)，利用复杂的注意力机制或 Transformer 模块来促进模态间的信息流动[2]-[4]。然而，无论是简单的叠加还是复杂的交互，这些策略在本质上都导致了模态特征之间的紧密结合。他们都有一个看似十分合理的基本假设：所有传感器输入始终是可用的且完美对齐的。预设所有传感器数据流在时空上是连续且完美对齐在现实中并不完全合理，因为硬件故障、

信号遮挡或环境干扰随时可能导致传感器数据流中断，此时该传感器模态将完全丢失。本文观察到一个现象：当一种模态丢失时，主流的 RGB-T (RGB-Thermal)模型性能会出现巨大幅度的下降。这充分暴露了现有系统在面对传感器失效时，严重缺乏工程落地所必需的鲁棒性机制。

为了最大化联合增益，这些网络倾向于将不同模态的特征深度交织，从而牺牲了各模态独立提取完整语义的能力，这种高度纠缠的特征表示使得模型对两种模态的同时存在产生了极强的依赖性，因此，当面对真实场景中常见的单模态丢失(如热像仪关闭或摄像头故障)时，这种强依赖性会致使高度融合的特征空间瞬间失效，进而导致模型性能出现大幅度的断崖式下滑，甚至远低于仅使用单一模态训练的基础网络。

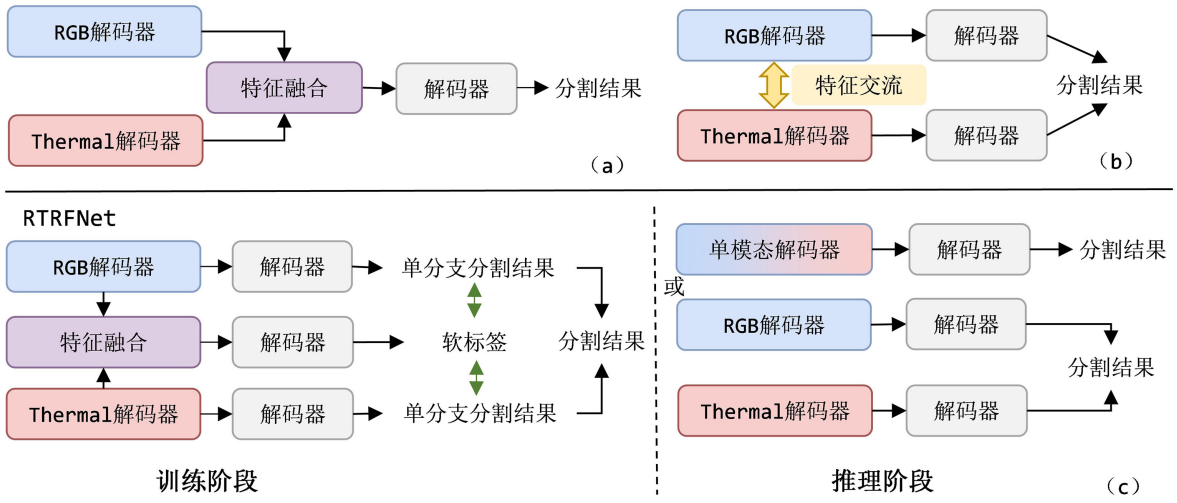

Figure 1. Comparison of RGB-T segmentation training paradigm architectures
图 1. RGB-T 分割训练范式架构对比

为解决这些局限性，本文提出了 RTRFNet 模型，如图 1(c)所示，引入了非对称的教师–学生学习机制：将具备全面多模态感受野的轻量化联合感知头作为教师分支，利用其输出的高质量联合语义分布，隐式地监督和增强 RGB 与热红外两个独立单模态学生分支的特征表征能力。具体而言，我们设计了一个轻量级的通道注意力特征融合模块(CA-FFM)，以高效提取多模态互补信息并构建联合特征空间，随后输入至教师头产生软标签；在此基础上，提出了多模态知识蒸馏(MKD)策略，通过最小化分布散度，促使单模态分支的预测分布向多模态中枢的最优解对齐。在推理阶段，该网络结构支持参数的灵活分离，无需运行教师分支即可完成推理：在全模态条件下直接对双流预测结果进行均值融合，而当某一模态失效时，系统可独立运行对应的单模态编码器–解码器路径。这一设计在保证全模态融合精度的同时，显著提升了模型在单一模态失效条件下的感知鲁棒性与推理效率。

本文的贡献主要有三点：

- 提出了一种面向 RGB-T 分割模态缺失场景的非对称教师–学生架构范式：训练阶段利用多模态教师分支建模跨模态互补信息，推理阶段则将系统解耦为可独立运行的单模态学生分支，从而兼顾全模态精度与单模态失效下的鲁棒性。
- 在该范式下，构建了一个轻量级实现方案，将 CA-FFM 与 MKD 相结合，用于生成联合语义表征并向单模态分支传递跨模态知识，以较小额外开销实现灵活推理。
- 在 MFNet 和 FMB 数据集上的实验表明，该架构在保持全模态竞争力的同时，能够显著缓解模态缺失带来的性能退化，验证了所提出范式在鲁棒感知中的有效性。

2. 相关工作

2.1. RGB-T 语义分割中的特征融合

利用热红外图像作为 RGB 图像的补充信息，已经在语义分割任务中取得了显著的精度提升。早期的研究主要基于卷积神经网络(CNN)，致力于验证多模态数据在复杂环境下的有效性。诸如 MFNet [1]、RTFNet [5]和 FuseSeg [6]等开创性工作，采用了较为直接的融合策略。这些方法通常在编码器或解码器阶段，通过对 RGB 和热红外特征图进行逐元素相加或通道拼接来整合信息。尽管这种简单的线性叠加方式在一定程度上证明了热红外信息在光照不足场景下的价值，但它们往往忽略了不同模态在不同场景下的信息差异性与噪声分布，导致在某一模态失效时，融合模型的鲁棒性较差。

为了解决直接融合带来的局限性，随后的工作聚焦于增强更细粒度的特征融合，旨在促进更全面的跨模态交互。研究重点逐渐转向设计更复杂的融合架构，以动态地筛选和整合有效信息。例如，ABMDRNet [7]等方法引入了通道注意力或空间注意力机制，通过计算模态间的相关性权重，自适应地增强互补特征并抑制模态特有的噪声。另一些工作如 EGFNet [8]，则引入边缘检测等辅助任务来指导融合过程，利用热红外图像中清晰的轮廓信息来优化分割边界。此外，GMNet [9]和 MFFENet [10]等方法致力于在不同分辨率层级上进行特征融合，确保了深层语义信息与浅层纹理细节的充分结合。

随着视觉 Transformer (ViT)的兴起，基于 Transformer 的融合架构成为了当前的研究热点。与传统 CNN 局限于局部感受野不同，HAPNet [11]、CMX [12]、CMNext [13]以及 StitchFusion [14]等方法充分利用了 Transformer 强大的长距离依赖建模能力。这些模型通常采用双流 Transformer 编码器结构，在提取各自模态特征的同时，通过交叉注意力模块进行显式的特征对齐与交互。这种机制不仅能够捕捉全局上下文信息，还能在特征空间中更有效地校正模态间的空间错位，从而极大地提升了分割性能。尽管上述跨模态交互策略不断刷新着分割精度的上限，但它们通常预设所有传感器数据都是完整且高质量的。这种对完整输入的强依赖使得融合模型在面对现实世界中常见的传感器故障时变得异常脆弱，一旦一种模态缺失，高度耦合的融合特征便会失效，导致性能急剧下降。

2.2. 面向模态缺失的鲁棒策略

为了在输入不完整的条件下维持分割精度，现有的研究主要致力于打破全模态输入与模型推理之间的刚性绑定，其主流解决方案可归纳为以下两种范式。首先知识蒸馏范式，这一范式的核心思想是将全模态数据视为一种“特权信息”，通过教师–学生架构将其蕴含的丰富语义“蒸馏”至单模态网络中。代表性工作如 CRM [15]和 LCM [16]等，通常首先训练一个高精度的多模态融合模型作为“教师”，随后利用其输出的特征图或概率分布来指导仅接收单一模态输入的“学生”网络进行学习。然而，这类方法通常需要复杂的多阶段训练流程，且“学生”网络往往是独立于融合主干之外的附加模块或独立网络，这种分离式的架构割裂了模态间的协同优化过程，使得单一模态分支难以直接从融合动力学中获益。另一种范式则侧重于根据输入的完整性动态调整网络参数或结构。例如，Adapted [17]提出了一种“冻结–适配”策略，即保持预训练好的融合主干网络参数不变，针对特定的模态缺失场景(如仅 RGB 或仅热红外)训练轻量级的适配器(Adapter)模块。该方法试图通过微调少量参数来校正因模态丢失引起的特征分布漂移。尽管这种方法具有一定的灵活性，但其局限性在于过分依赖预训练的融合表示：由于主干网络被冻结，其底层特征提取器无法针对单模态输入的特性进行重新适应或重构。

与先前的工作不同，本文的工作提出了一种可解耦机制，旨在同时强化融合主干网络，并实现在推理阶段按模态进行解耦，同时三条支线的同步训练也提升了模型训练的效率。

3. 方法

3.1. 框架概述

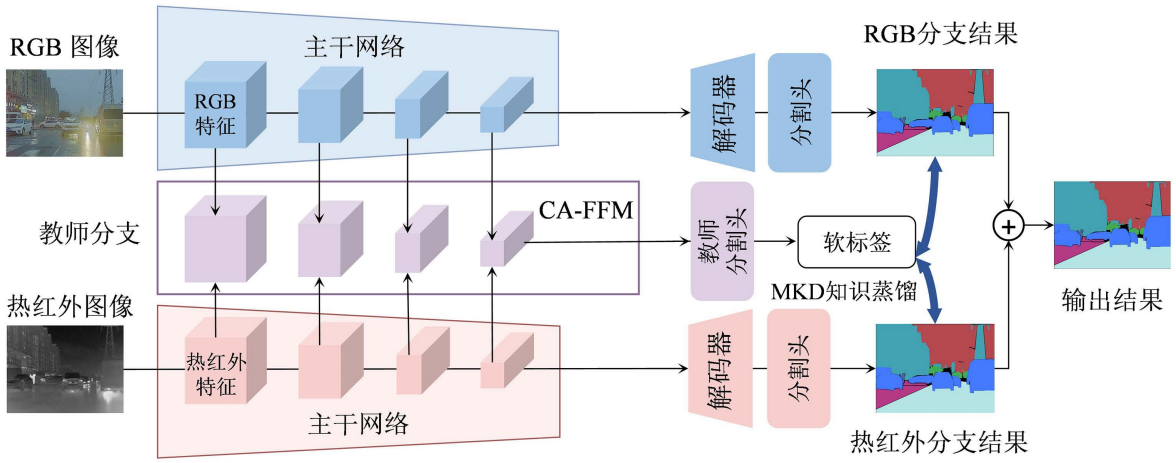

Figure 2. The RTRFNet schematic model is shown in the diagram, where CA-FFM performs modality fusion, and CMDR/RDR decouples the fused features to regularize and guide each single modality branch
图 2. RTRFNet 示意图模型，其中 CA-FFM 执行模态融合，CMDR/RDR 则解耦融合后的特征，以正则化并指导每一个单模态分支

为了从根本上解决由于传感器部分失效导致的多模态分割性能退化问题，本文提出了一种基于非对称教师–学生范式的 RGB-T 多模态鲁棒网络结构(RTRFNet)。如图 2 所示，该架构基于 SegFormer 构建了双流主网络与辅助教师头的系统。在训练阶段，RGB 与热红外(Thermal)图像首先被馈入各自独立的编码器中以提取多尺度空间特征。随后，一个轻量级的通道注意力特征融合模块(CA-FFM)在特征通道维度上执行高效的跨模态交互，并将融合后的完备特征传递给处于中心位置的轻量化感知头(即“教师”分支)。同时，两侧的单模态解码器(即“学生”分支)可以同时在自己的分支上独立处理各自的单模态特征。本文引入了多模态知识蒸馏(MKD)策略，强制约束单模态感知空间向多模态中枢的最优语义分布对齐，促使单模态分支能够获得多模态的丰富上下文知识，从而大幅提升其单一模态下的语义分割表现。这种设计使得网络在推理阶段拥有极高灵活性：在全模态正常运作时对两学生分支的预测概率进行均值融合，而在面临传感器缺失的极端场景下独立运作单分支时，仍能保持极高的感知精度与鲁棒性。

3.2. 通道注意力特征融合模块(CA-FFM)

为了在保证特征交互效率的同时，为轻量化教师提供来自多模态的互补信息的联合语义表征，如图 3 所示，本文设计了一个精简且高效的通道注意力特征融合模块(CA-FFM)。该模块旨在通过跨模态的通道维度重标定，自适应地放大互补线索并抑制冗余噪声。

令在第 i 个编码器阶段提取的 RGB 特征与热红外特征分别表示为 $F_{rgb}^{(i)} \in \mathbb{R}^{C\times H\times W}$ 和 $F_{t}^{(i)} \in \mathbb{R}^{C\times H\times W}$。首先，为了保留两个模态的完整原始信息特征，我们在通道维度上对其进行显式拼接，构建多模态联合特征张量 $F_{cat}^{(i)}$：

$$F_{cat}^{(i)} = \left[F_{rgb}^{(i)}, F_{t}^{(i)} \right] \in \mathbb{R}^{2C\times H\times W}$$

单一的池化操作容易导致部分关键高频信号的丢失。因此，我们同时采用全局平均池化(GAP)和全局最大池化(GMP)来分别聚合联合特征中的背景上下文与高频显著性线索：

$$z_{avg} = \mathrm{GAP}\left(F_{cat}^{(i)}\right) \in \mathbb{R}^{2C\times 1\times 1}$$

$$z_{max} = \mathrm{GMP}\left(F_{cat}^{(i)}\right) \in \mathbb{R}^{2C\times 1\times 1}$$

随后，为了捕捉这 $2\times C$ 个通道之间的非线性跨模态依赖关系，我们将 z_{avg} 和 z_{max} 分别输入到一个共享权重的多层感知机。该 MLP 包含一个降维比例为 r 的隐藏层，使网络学习模态间最核心的协同表示。具体计算过程如下：

$$E_{avg} = W_1\left(\delta\left(W_0\left(z_{avg}\right)\right)\right)$$

$$E_{max} = W_1\left(\delta\left(W_0\left(z_{max}\right)\right)\right)$$

其中，$W_0 \in \mathbb{R}^{\frac{2C}{r}\times 2C}$ 和 $W_1 \in \mathbb{R}^{2C\times\frac{2C}{r}}$ 分别代表降维和升维全连接层的可学习权重矩阵，δ 代表 ReLU 激活函数。将聚合后的特征描述符进行逐元素相加，并通过 Sigmoid 激活函数 σ 生成最终的跨模态通道注意力权重向量 w：

$$w = \sigma\left(E_{avg} + E_{max}\right) \in \mathbb{R}^{2C\times 1\times 1}$$

利用生成的高维权重矩阵 w，我们对拼接后的联合特征 $F_{cat}^{(i)}$ 进行通道级的特征重新标定，以自适应地激发有用的多模态特征通道：

$$F_{recal}^{(i)} = w \otimes F_{cat}^{(i)}$$

此处，$\otimes$ 表示沿通道维度的逐元素乘法。

最后，我们引入了一个1×1的逐点卷积(Point-wise Convolution)作为跨模态信息瓶颈层，将特征维度平滑降维至 C 维，输出最终的高质量融合特征 $F_{fused}^{(i)}$：

$$F_{fused}^{(i)} = \mathrm{Conv}_{1\times 1}\left(F_{recal}^{(i)}\right) \in \mathbb{R}^{C\times H\times W}$$

通过上述的计算，CA-FFM 在仅引入极少量参数(W_0,W_1及逐点卷积)的前提下，实现了对可见光与热红外特征在通道级别的高效提纯与深度融合，为后续教师分支生成精确的软标签提供了优越的融合特征。

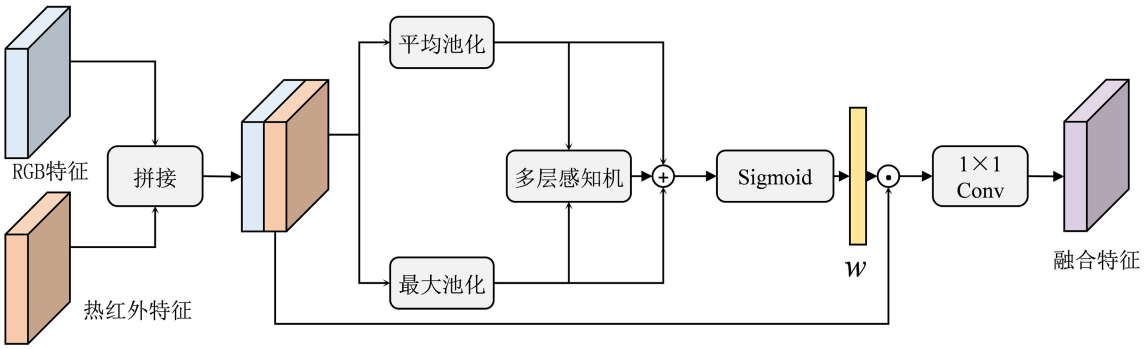

Figure 3. FFM schematic model
图 3. FFM 示意图模型

3.3. 多模态知识蒸馏(MKD)

本文引入了多模态知识蒸馏(MKD)策略。该策略旨在利用具备全局感受野的多模态融合分支作为

“教师”，将包含跨模态互补信息的高质量语义分布作为“软标签”(Soft Targets)，非对称地迁移并注入到 RGB 和热红外两个独立的“学生”分支中。

具体而言，多模态融合分支(教师分支)输出的最终分类概率值(Logits)为 $Z_{fused} \in \mathbb{R}^{K\times H\times W}$，其中 K 表示语义类别的总数。同理，RGB 学生分支与热红外学生分支输出的逻辑值分别记为 Z_{rgb} 和 Z_t。为了灵活调整网络输出的概率分布，我们引入了温度超参数 τ。经过带有温度缩放的 Softmax 函数处理后，各分支的概率分布可被统一形式化为：

$$P_{fused}^{(k)} = \frac{\exp\left(Z_{fused}^{(k)}/\tau\right)}{\sum_{j=1}^{K}\exp\left(Z_{fused}^{(j)}/\tau\right)}$$

$$P_{rgb}^{(k)} = \frac{\exp\left(Z_{rgb}^{(k)}/\tau\right)}{\sum_{j=1}^{K}\exp\left(Z_{rgb}^{(j)}/\tau\right)} \quad P_{t}^{(k)} = \frac{\exp\left(Z_{t}^{(k)}/\tau\right)}{\sum_{j=1}^{K}\exp\left(Z_{t}^{(j)}/\tau\right)}$$

其中，上标 (k) 表示第 k 个语义类别。由于教师头已经通过 CA-FFM 模块预先融合了来自可见光的高频几何结构与热红外的补充特征，其生成的软概率分布 P_{fused} 蕴含了更为完备的场景理解边界。

我们采用 Kullback-Leibler (KL)散度来强制单模态学生网络的预测分布逼近教师网络的联合最优解。为了确保在反向传播过程中，教师分支的表征不被学生分支的低质量特征负面优化，且优化梯度仅呈现从教师到学生的单向流动，我们在 P_{fused} 上应用了停止梯度算子(Stop-Gradient, $\mathrm{Sg}(\cdot)$)。RGB 分支与热红外分支的独立蒸馏损失分别定义为：

$$L_{KD\_rgb} = \tau^2 \cdot \mathcal{D}_{KL}\left(\mathrm{Sg}\left(P_{fused}\right) \| P_{rgb}\right)$$

$$L_{KD\_t} = \tau^2 \cdot \mathcal{D}_{KL}\left(\mathrm{Sg}\left(P_{fused}\right) \| P_{t}\right)$$

其中，$\|$ 符号表示两个概率分布之间的散度量度，用于精确量化学生网络分布(P_{rgb} 或 P_t)相对于教师网络分布(P_{fused})的信息差异。由于在计算带有温度参数 τ 的梯度时幅度会按 $1/\tau^2$ 的比例缩小，我们在上述公式中乘以 τ^2 以保持蒸馏梯度与硬标签监督梯度在量级上的一致性。

全局的多模态知识蒸馏损失 L_{KD} 为两个学生分支损失的联合总和：

$$L_{KD} = L_{KD\_rgb} + L_{KD\_t}$$

通过最小化 L_{KD}，两个独立运作的单模态网络被强制要求模拟轻量化联合感知头的决策过程。这种跨模态的隐式监督机制促使学生分支在特征提取阶段“内化”了缺失模态的上下文信息。最重要的是，这一知识迁移过程完全发生在训练阶段，在推理阶段无需引入教师网络的任何参数，从根本上兼顾了模型在单模态失效条件下的鲁棒性与计算效率。

3.4. 总目标函数与自适应推理机制

对于多类别分割任务，本文采用标准的交叉熵损失函数(Cross-Entropy Loss, $\mathcal{L}_{CE}$)来衡量各分支预测概率与真实标签之间的差异：

$$\mathcal{L}_{task} = \mathcal{L}_{CE}\left(P_{fused}, Y\right) + \mathcal{L}_{CE}\left(P_{rgb}, Y\right) + \mathcal{L}_{CE}\left(P_{t}, Y\right)$$

将上述基础分割任务损失与 3.3 节中定义的知识蒸馏损失结合，RTRFNet 的总体联合优化目标函数 $\mathcal{L}_{total}$ 可形式化表达为：

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \alpha \mathcal{L}_{KD}$$

其中，α 为平衡多任务学习与特征蒸馏强度的超参数。通过联合优化这一全局目标函数，网络在保证各独立感知通路基础分类准确性的同时，实现了跨模态高维语义向单模态空间的平滑迁移。

RTRFNet 在实际工程部署与推理阶段展现出了极高的运行灵活性与架构轻量化优势，其两个分支可以从框架中解耦的设计使得其在推理阶段可以应对任意的模态缺失的情况。

在模型完成训练并切入测试阶段时，用于提取联合表征的 CA-FFM 模块以及负责输出软标签的轻量化教师感知头将被移除。当可见光与热红外数据流均完好且同步时，两侧的 RGB 与热红外学生分支并行独立运作。在网络的最终输出端，系统采用无参数的决策层均值融合策略，直接对两条独立链路输出的软概率分布进行均值运算，以获取最终的分割结果：

$$P_{final} = \frac{1}{2}\left(P_{rgb} + P_t\right)$$

这种简单的决策层均值融合方式在参数轻便的同时也表现出优越的性能。

当遭遇突发硬件故障、极端恶劣天气或物理遮挡导致某一模态信号完全丢失时(如 RGB 摄像头失效)，系统将立即挂起并切断损坏的链路，仅激活存活传感器的单边推理路径(即直接输出 $P_{final} = P_t$ 或 $P_{final} = P_{rgb}$)。由于在训练阶段，单边通路已通过 MKD 获得了多模态协同表征的全局先验知识，因此即便在纯单源数据驱动的极端假设下，其依然能够独立输出稳健且高精度的语义分割预测。

4. 实验

本文进行了详细的实验，以评估本文提出的方法在不同任务和数据集上的性能。本文还将其与现有的、对模态缺失具有鲁棒性的方法进行了比较。

4.1. 数据集

本文使用三个数据集来评估 RTRFNet 在 RGB-热红外分割任务上的性能。

MFNet [1]数据集：该数据集记录了城市街道场景中的 9 个语义类别，包含 1569 对分辨率为 640 × 480 的 RGB 和热红外图像。本文遵循 MFNet 的划分方式，使用 784 张图像进行训练，392 张图像进行验证。

FMB [18]数据集：包含 1500 对高分辨率(800 × 600)图像，像素级标注率超过 98%。该库涵盖了自动驾驶场景中常见的 14 类目标，其中 1220 对用于模型训练，剩余 280 对用于测试。

4.2. 实现细节

本文提出的 RTRFNet 框架基于 SegFormer 构建了双流架构，其中教师和学生分支均采用在 ImageNet 上预训练的 Mix Transformer (MiT-B2) [19]编码器作为主干网络，以提取多尺度的特征表示。所有的实验均基于 PyTorch 深度学习框架实现，并在单张 NVIDIA A100 GPU 上完成训练与推理过程，以确保高效的计算性能。本文使用 AdamW 优化器对模型进行优化，权重衰减设置为 0.0001，动量设置为 0.9。在超参数配置上，编码器和解码器的初始学习率分别设定为 1×10^{-4} 和 6×10^{-4}，并采用标准的交叉熵损失函数进行监督训练。

为了与现有的先进方法进行公平比较，本文采用了完全一致的数据增强策略，包括比例范围在 0.5 到 2.0 之间的随机缩放、水平翻转以及随机尺寸裁剪。针对不同的数据集特性，本文调整了分辨率和训练周期：对于 MFNet 数据集，输入分辨率调整为 480 × 640，训练时长为 300 个周期，并选取在验证集上表现最佳的检查点进行评估；对于 FMB 数据集，输入分辨率设定为 512 × 512。

在评价指标方面，本文主要采用平均交并比(mIoU)来量化分割性能。mIoU 作为语义分割中最关键的指标，它计算了预测区域与真实标签区域的交集与并集之比在所有类别上的平均值，从而能够全面且严苛地评估模型对物体边界的定位精度与分类一致性。此外，为了评估模型在真实场景下的鲁棒性，本文设计了三种推理评估设置：(1) 全模态设置，即 RGB 和热红外数据均可用；(2) RGB 缺失设置，模拟摄像头故障；以及(3) 热红外缺失设置，模拟热成像仪失效的情况。

4.3. 实现结果

Table 1. Performance comparison with existing robust methods on the MFNet dataset
表 1. 与现有鲁棒方法在 MFNet 数据集上的性能对比

方法	主干网络	参数量(M)	仅 RGB	仅热红外	RGB-T	平均
			mIoU	mIoU	mIoU	mIoU
FuseNet [20]	VGG-16	284	10.31	36.85	45.6	30.92
MFNet [1]	DCNN	8.4	24.78	16.64	39.7	27.04
RTFNet [5]	ResNet-152	254.51	37.3	24.57	53.2	38.36
FEANet [2]	ResNet-152	337.1	8.69	48.72	55.3	37.57
EAEFNet [3]	ResNet-50	200.4	35.23	41.72	58.95	45.3
CMNexXt [13]	Mit-B4	116.56	53.55	35.46	59.77	49.59
CRM [15]	Swin-T	74.92	50.98	50.22	59.1	53.43
StitchFusion [14]	Swin-T	65.27	48.78	41.42	58.04	49.41
HKDNet [21]	Swin-S	-	52.5	–	56.5	–
Adapted [17]	MiT-B4	-	55.22	50.89	–	–
Ours (MiT-B2)	MiT-B2	57	55.42	**53.23**	57.97	**55.54**

Table 2. Performance comparison with existing robust methods on the FMB dataset
表 2. 与现有鲁棒方法在 FMB 数据集上的性能对比

方法	主干网络	参数量(M)	仅 RGB	仅热红外	RGB-T	平均
			mIoU	mIoU	mIoU	mIoU
SegMiF [22]	MiT-B2	-	50.50	-	54.8	52.65
StitchFusion [14]	Swin-T	65.27	56.75	34.82	63.32	51.63
HKDNet [21]	Swin-T	-	54.0	-	62.4	58.2
Ours (MiT-B2)	MiT-B2	57	61.03	**57.53**	65.34	**61.30**

表 1 详细列出了各模型在不同模态缺失条件下的性能对比。在全模态(RGB-T)的基准测试中，基于 MiT-B2 的 RTFDNet 取得了 57.97%的 mIoU，这一成绩超越了大多数现有方法，且与当前的 SOTA 模型持平。更关键的突破在于单模态推理场景：当遭遇热红外传感器失效(仅 RGB)时，本模型的 mIoU 仅出现微幅下滑(−2.35%)，维持在 55.42%的高位，显著优于 CMNeXt (53.55%)、CRM (50.98%)及 StitchFusion (48.78%)。同样，在仅依赖热红外输入的极端条件下，RTFDNet 依然保持了 53.23%的 mIoU，性能降幅

控制在 5.74%以内，超越了大多数现有方法。

本文框架的鲁棒性在 FMB 数据集上得到了进一步验证，如表 2 所示。在 FMB 数据集上，本文的 RTRFNet 在仅 RGB 和仅热红外设置下分别达到了 61.03%和 57.53%的 mIoU，显著优于其他方法。

4.4. 定性分析

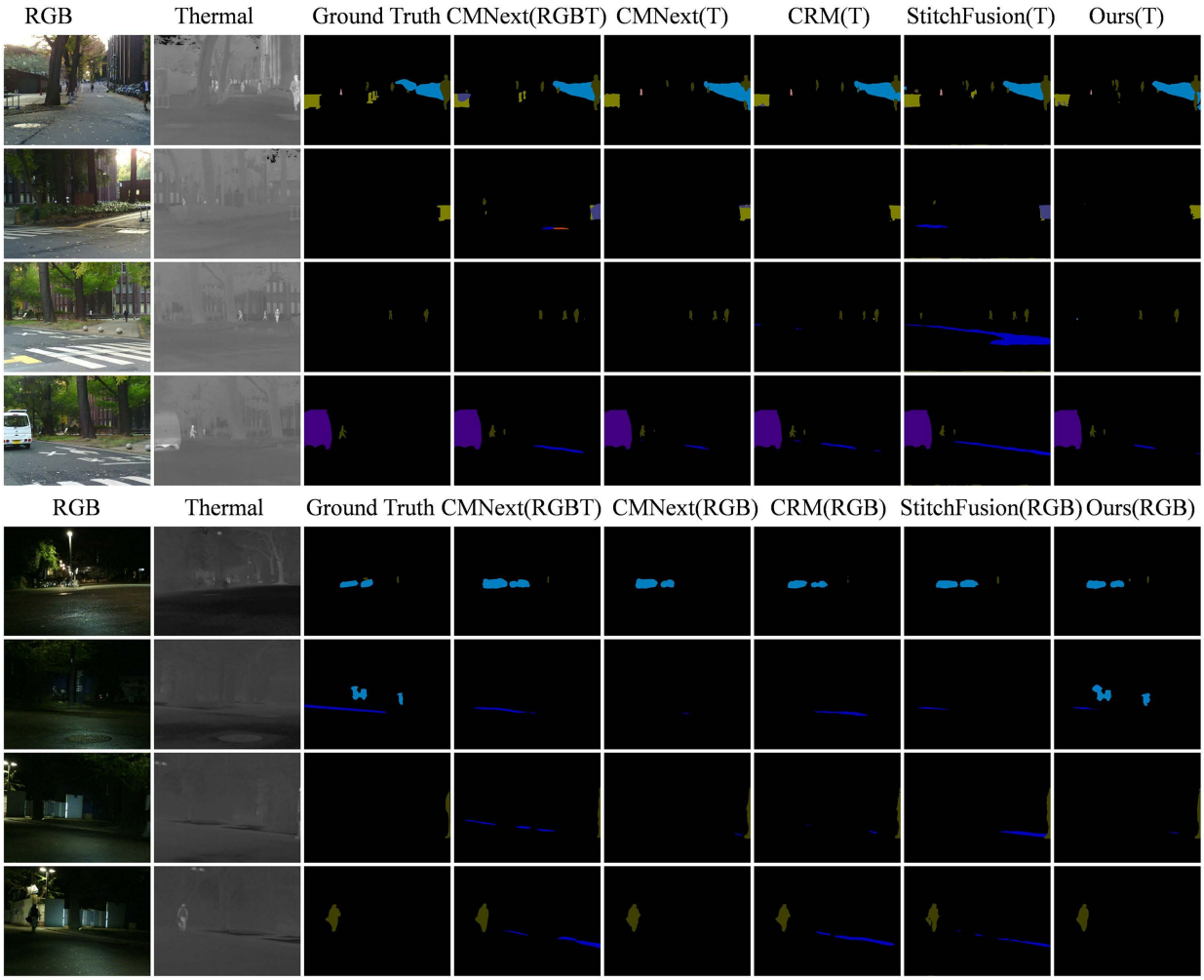

Figure 4. Qualitative results on the MFNet dataset
图 4. 在 MFNet 数据集上的定性结果

为了直观展示模型在极端环境下的鲁棒性，图 4 对比了 MFNet 数据集上不同方法在模态缺失时的分割结果。在光照充足但缺乏热红外辅助的场景中(如第三行)，大多数对比模型难以从单一模态中提取足够的细节，导致对“自行车”等细粒度目标的分割出现断裂或漏检。相比之下，RTFDNet 能够精准地重构出物体的完整轮廓。更具挑战性的是夜间场景(如第五至第八行)，此时 RGB 图像几乎全黑，信息量极低。在热红外模态缺失的极端假设下，大多数现有方法彻底失效，无法识别图像边缘的行人。然而，得益于跨模态正则化机制的约束，RTFDNet 依然能够从极其微弱的 RGB 信号中恢复出目标的语义结构，准确分割出行人区域。图 5 在 FMB 数据集上的可视化结果同样表明，即便在单传感器模式下，本模型仍能保持清晰的边界界定与极低的误检率。

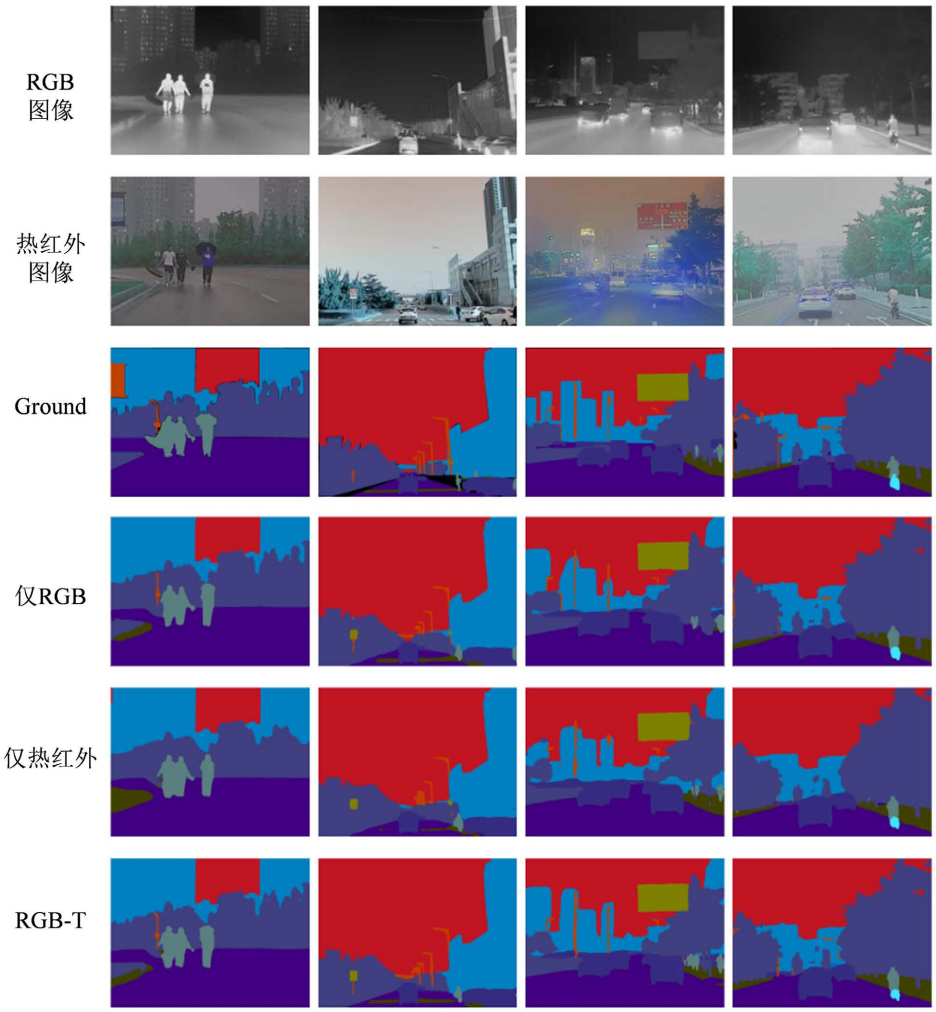

Figure 5. Qualitative results on the FMB dataset
图 5. 在 FMB 数据集上的定性结果

4.5. 消融实验分析

Table 3. Ablation experimental results on the MFNet dataset
表 3. 在 MFNet 数据集上的消融实验结果

基线	CA-FFM	MKD	RGB mIoU	热红外 mIoU
√			51.2	47.1
√	√		52.7	49.3
√	√	√	55.42	53.23

为了系统地评估 RTRFNet 中各核心组件的贡献，本文在 MFNet 数据集上执行了详尽的消融实验，旨在验证通道注意力特征融合模块(CA-FFM)与多模态知识蒸馏(MKD)策略的有效性，结果汇总于表 3。

实验从一个仅包含基础双流编解码器(即完全独立、无任何跨模态交互)的基线模型开始。如表所示，该基线模型在 RGB 和热红外单模态推理上仅分别获得了 51.2%和 47.1%的 mIoU。这一基础性能表明，在缺乏合理的模态交互机制与跨模态知识传递时，孤立的网络分支难以充分理解复杂场景的深层语义，

尤其是对热红外等低频信号的利用率明显不足。

随后，本文在基线模型中引入了 CA-FFM 模块构建多模态联合感知头。当单独加入 CA-FFM 时，RGB 和热红外分支的性能分别提升至 52.7%和 49.3%。这一提升证明了 CA-FFM 能够高效地在通道维度上聚合多模态互补线索；而在联合端到端训练过程中，这种高质量的融合损失梯度通过共享主干网络回传，隐式地增强了各个单模态底层特征提取器的判别力。

最终，当同时集成 CA-FFM 与 MKD 策略时，RTRFNet 展现出了最优的跨模态协同效应。在教师分支软标签的显式约束下，完整模型在 RGB 和热红外的单分支推理上分别取得了 55.42%和 53.23%的优异 mIoU，显著超越了基线模型(分别大幅提升 4.22%和 6.13%)。这一结果不仅验证了两个模块并非简单的功能叠加，更深刻揭示了其内在机制的完美闭环：CA-FFM 负责在特征空间提炼出高质量的联合语义分布，而 MKD 则精准地将这份高维知识跨模态“蒸馏”并内化到学生网络中。这种“高质融合–隐式下放”的互补策略，共同确保了模型在推理阶段彻底脱离联合分支后，即使面临单一模态输入的极端条件，依然能保持极高的感知精度与鲁棒性。

5. 结论

本文针对多模态语义分割在单一传感器失效或模态缺失条件下性能急剧退化的挑战，提出了一种基于知识蒸馏的鲁棒 RGB-T 分割框架 RTRFNet。与依赖复杂特征修补或高度耦合的传统融合方法不同，RTRFNet 创新性地构建了非对称的教师–学生学习范式。其中，本文设计的轻量级通道注意力特征融合模块(CA-FFM)高效地提取并重构了跨模态的互补语义；而多模态知识蒸馏(MKD)策略则将这一高质量的联合分布作为软标签，隐式地指导并内化增强了 RGB 与热红外双流学生网络的独立感知能力。该机制彻底打破了推理阶段对多源数据绝对完整的强依赖：系统在全模态输入时可通过无参数的均值后融合实现高精度预测；而在面对模态缺失的极端场景下，系统能够完全剥离教师头，仅保留单边存活链路即可完成稳健且极具参数效率的独立推理。本文仍存在一定局限性。首先，当前实验主要在 MFNet 与 FMB 两个 RGB-T 数据集上开展，数据场景与传感器配置的多样性仍然有限；其次，本文主要考察了较为理想化的“单一模态完全缺失”情形，而对噪声污染、局部退化、时序不同步等更复杂的模态退化模式尚未进行系统分析。未来工作将进一步在更多数据集和更贴近真实部署的退化设定下验证并扩展该架构范式。

参考文献

[1] Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y. and Harada, T. (2017) MFNet: Towards Real-Time Semantic Segmentation for Autonomous Vehicles with Multi-Spectral Scenes. 2017 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Vancouver, 24-28 September 2017, 5108-5115. https://doi.org/10.1109/iros.2017.8206396

[2] Deng, F., Feng, H., Liang, M., Wang, H., Yang, Y., Gao, Y., *et al.* (2021) FEANet: Feature-Enhanced Attention Network for RGB-Thermal Real-Time Semantic Segmentation. 2021 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Prague, 27 September-1 October 2021, 4467-4473. https://doi.org/10.1109/iros51168.2021.9636084

[3] Liang, M., Hu, J., Bao, C., Feng, H., Deng, F. and Lam, T.L. (2023) Explicit Attention-Enhanced Fusion for RGB-Thermal Perception Tasks. *IEEE Robotics and Automation Letters*, **8**, 4060-4067. https://doi.org/10.1109/lra.2023.3272269

[4] Tang, L., Yuan, J., Zhang, H., Jiang, X. and Ma, J. (2022) PIAFusion: A Progressive Infrared and Visible Image Fusion Network Based on Illumination Aware. *Information Fusion*, **83**, 79-92. https://doi.org/10.1016/j.inffus.2022.03.007

[5] Sun, Y., Zuo, W. and Liu, M. (2019) RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes. *IEEE Robotics and Automation Letters*, **4**, 2576-2583. https://doi.org/10.1109/lra.2019.2904733

[6] Sun, Y., Zuo, W., Yun, P., Wang, H. and Liu, M. (2021) FuseSeg: Semantic Segmentation of Urban Scenes Based on RGB and Thermal Data Fusion. *IEEE Transactions on Automation Science and Engineering*, **18**, 1000-1011. https://doi.org/10.1109/tase.2020.2993143

[7] Zhang, Q., Zhao, S., Luo, Y., Zhang, D., Huang, N. and Han, J. (2021) ABMDRNet: Adaptive-Weighted Bi-Directional Modality Difference Reduction Network for RGB-T Semantic Segmentation. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 2633-2642. https://doi.org/10.1109/cvpr46437.2021.00266

[8] Dong, S., Zhou, W., Xu, C. and Yan, W. (2024) EGFNet: Edge-Aware Guidance Fusion Network for RGB-Thermal Urban Scene Parsing. *IEEE Transactions on Intelligent Transportation Systems*, **25**, 657-669. https://doi.org/10.1109/tits.2023.3306368

[9] Zhou, W., Liu, J., Lei, J., Yu, L. and Hwang, J. (2021) GMNet: Graded-Feature Multilabel-Learning Network for RGB-Thermal Urban Scene Semantic Segmentation. *IEEE Transactions on Image Processing*, **30**, 7790-7802. https://doi.org/10.1109/tip.2021.3109518

[10] Zhou, W., Lin, X., Lei, J., Yu, L. and Hwang, J. (2022) MFFENet: Multiscale Feature Fusion and Enhancement Network for Rgb-Thermal Urban Road Scene Parsing. *IEEE Transactions on Multimedia*, **24**, 2526-2538. https://doi.org/10.1109/tmm.2021.3086618

[11] Li, J., Yun, P., Chen, Q. and Fan, R. (2024) HAPNet: Toward Superior RGB-Thermal Scene Parsing via Hybrid, Asymmetric, and Progressive Heterogeneous Feature Fusion. arXiv: 2404.03527.

[12] Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R. and Stiefelhagen, R. (2023) CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers. *IEEE Transactions on Intelligent Transportation Systems*, **24**, 14679-14694. https://doi.org/10.1109/tits.2023.3300537

[13] Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., *et al.* (2023) Delivering Arbitrary-Modal Semantic Segmentation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 1136-1147. https://doi.org/10.1109/cvpr52729.2023.00116

[14] Li, B., Zhang, D., Zhao, Z., Gao, J. and Li, X. (2025) StitchFusion: Weaving Any Visual Modalities to Enhance Multimodal Semantic Segmentation. *Proceedings of the* 33*rd ACM International Conference on Multimedia*, Dublin, 27-31 October 2025, 1308-1317. https://doi.org/10.1145/3746027.3755110

[15] Shin, U., Lee, K., Kweon, I.S. and Oh, J. (2024) Complementary Random Masking for RGB-Thermal Semantic Segmentation. 2024 *IEEE International Conference on Robotics and Automation* (*ICRA*), Yokohama, 13-17 May 2024, 11110-11117. https://doi.org/10.1109/icra57147.2024.10611200

[16] Zheng, X., Xue, H., Chen, J., Yan, Y., Jiang, L., Lyu, Y., Yang, K., Zhang, L. and Hu, X. (2024) Learning Robust Anymodal Segmentor with Unimodal and Cross-Modal Distillation. arXiv: 2411.17141.

[17] Reza, M.K., Prater-Bennette, A. and Asif, M.S. (2025) Robust Multimodal Learning with Missing Modalities via Parameter-Efficient Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **47**, 742-754. https://doi.org/10.1109/tpami.2024.3476487

[18] Liu, J., Liu, Z., Wu, G., Ma, L., Liu, R., Zhong, W., *et al.* (2023) Multi-Interactive Feature Learning and a Full-Time Multi-Modality Benchmark for Image Fusion and Segmentation. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 8081-8090. https://doi.org/10.1109/iccv51070.2023.00745

[19] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P. (2021) SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Advances in Neural Information Processing Systems*, **34**, 12077-12090.

[20] Hazirbas, C., Ma, L., Domokos, C. and Cremers, D. (2017) FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. In: Lai, S.H., Lepetit, V., Nishino, K. and Sato, Y., Eds., *Computer Vision—ACCV* 2016, Springer, 213-228. https://doi.org/10.1007/978-3-319-54181-5_14

[21] Sun, Y., Dong, W., Wang, S., Wu, P., Feng, M., Li, X., *et al.* (2025) Distilling Hierarchical Knowledge from Multimodal Fusion for Unimodal Image Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, **35**, 11797-11809. https://doi.org/10.1109/tcsvt.2025.3579580

[22] Lin, B., Lin, Z., Guo, Y., Zhang, Y., Zou, J. and Fan, S. (2023) Variational Probabilistic Fusion Network for RGB-T Semantic Segmentation. arXiv: 2307.08536.