MMSFORMER：多模态变换器用于材料和语义分割

Md Kaykobad Reza[1], **Ashley Prater-Bennette**[2], **M. Salman Asif**[1]
[1] 加利福尼亚大学河滨分校，CA 92508，美国
[2] 美国罗马空军研究实验室，NY 13441，美国
mreza025@ucr.edu, ashley.prater-bennette@us.af.mil, sasif@ucr.edu

摘要

利用不同模态之间的信息已知可以提高多模态分割任务的性能。 然而，由于每种模态的独特特征，有效地融合来自不同模态的信息仍然具有挑战性。 在本文中，我们提出了一种新颖的融合策略，可以有效地融合来自不同模态组合的信息。 我们还提出了一个名为多模态分割变换器（MMSFormer）的新模型，该模型将提出的融合策略纳入其中，以执行多模态材料和语义分割任务。 MMSFormer在三个不同数据集上优于当前最先进的模型。 由于我们从只有一个输入模态开始，随着逐渐引入更多模态，性能逐渐提高，展示了融合块在结合来自不同输入模态的有用信息方面的有效性。 消融研究表明，融合块中的不同模块对整体模型性能至关重要。 此外，我们的消融研究还突出了不同输入模态提高在识别不同类型材料方面性能的能力。 代码和预训练模型将在https://github.com/csiplab/MMSFormer上提供。

1 简介

图像分割 [7, 36] 方法将图像中的每个像素分配一个类标签。 分割地图可用于全面理解对象或场景的上下文。 图像分割可以进一步分为不同类型；例如语义分割 [15, 44]、实例分割 [14, 17]、全景分割 [12, 24] 和材料分割 [29, 42]。 这些分割任务各自设计用于解决特定挑战和应用。

多模态图像分割 [16, 52] 旨在通过利用多样化信息源来提高任务的准确性和完整性，可能导致对复杂场景的更稳健理解。 与单模态分割 [37] 相比，多模态方法 [53] 更复杂，因为需要有效地整合来自不同模态的异构数据。 关键挑战在于数据质量和属性的变化，每种模态的独特特征，以及需要创建能够准确和连贯地使用融合信息进行分割的模型。

大多数现有的多模态分割方法都是为特定的模态对设计的，例如RGB-深度 [5, 18, 20]，RGB-热像 [27, 28, 41]，和RGB-Lidar [26, 38, 55]。 由于它们是针对特定模态组合设计的，大多数情况下它们通常无法很好地处理与原始设计中使用的模态组合不同的情况。 最近，CMX [50] 提出了一种技术，可以融合来自RGB和另一种辅助模态的信息，但它无法同时融合超过两种模态。 一些最近的模型提出了技术，可以融合超过两种模态 [1, 29, 51]。 然而，它们要么使用非常复杂的融合策略 [1, 50]，要么需要像语义标签 [29] 这样的额外信息来执行底层任务。

在本文中，我们提出了一个新颖的融合块，可以融合来自不同模态的信息组合。
我们还提出了一个新的模型，用于多模态材料和语义分割任务，我们称之为MMSFormer。

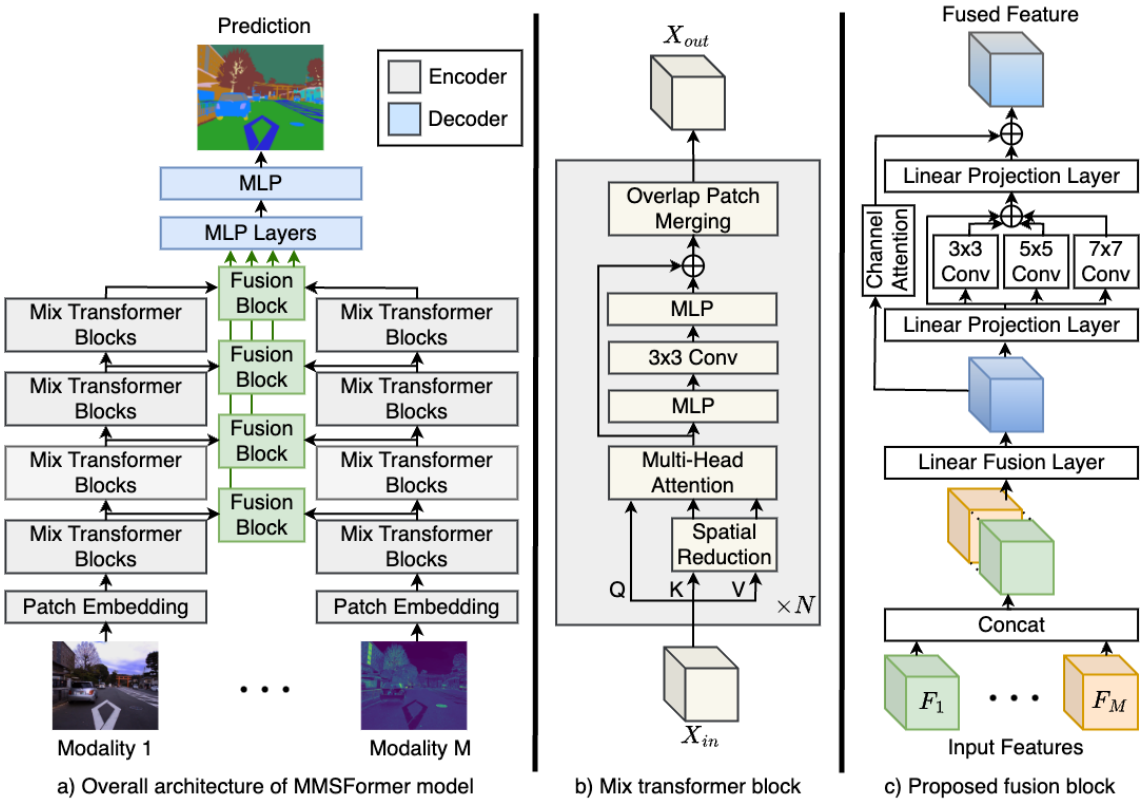


图**1:**(a) MMSFormer模型的整体架构。 每个图像都通过一个特定于模态的编码器，我们从中提取分层特征。 然后，我们使用提出的融合块融合提取的特征，并将融合的特征传递给解码器，以预测分割地图。(b) 混合变压器 [48] 块的示意图。 每个块在应用多头注意力之前都会进行空间缩减，以减少计算成本。 (c) 提出的多模态融合块。 我们首先沿着通道维度连接所有特征，并通过线性融合层将它们融合。 然后，特征张量被送入线性投影和并行卷积层，以捕获多尺度特征。 我们在残差连接中使用Squeeze and Excitation块 [19] 作为通道注意力，动态重新校准沿通道维度的特征。

我们的模型使用基于变压器的编码器 [48] 来捕获不同模态的层次特征，将提取的特征与我们的新型融合块融合，并利用MLP解码器执行多模态材料和语义分割。 特别是，我们提出的融合块使用并行卷积来捕获多尺度特征，通道注意力来重新校准沿通道维度的特征，线性层来结合跨多个模态的信息。 这样的设计提供了一个简单且计算效率高的融合块，可以处理任意数量的输入模态，并有效地组合不同模态组合的信息。我们在表9中比较我们的融合块与一些现有融合方法在参数数量和GFLOPs方面的差异。

为了评估我们提出的MMSFormer和融合块，我们专注于MCubeS上的多模态材料分割 [29] 数据集和FMB [33] 以及PST900 [40] 数据集上的多模态语义分割。 MCubeS数据集包括四种不同的模态：RGB、线偏振角（AoLP）、线偏振度（DoLP）和近红外（NIR）。 FMB数据集包括RGB和红外模态，而PST900数据集包括RGB和热红外模态。 我们在这些数据集的表1-5中展示了整体和每类性能比较。 一系列实验突出了提出的融合块有效地结合来自不同模态组合的特征，从而比当前最先进的方法表现出更优越的性能。 消融研究表明，不同的输入模态有助于识别不同类型的材料类别，如表8所示。 此外，随着添加新的输入模态，整体性能逐渐提高，突显了融合块将有用信息从新模态中整合的能力。 我们分别总结了FMB和MCubeS数据集的结果在表4和6中。

本文的主要贡献可以总结如下。

- 我们提出了一种名为MMSFormer的新型多模态分割模型。该模型包含一个新颖的融合块，可以融合来自任意（异构）模态的信息。

- 我们的模型在三个不同数据集上实现了新的最先进性能。 此外，与当前领先模型相比，我们的方法对所有模态组合都实现了更好的性能。
- 一系列消融研究表明，融合块上的每个模块对整体模型性能都有重要贡献，每个输入模态都有助于识别特定的材料类别。

本文的其余部分结构如下。 第2节简要回顾相关工作。 我们在第3节详细描述了我们的模型和融合块。 第4节展示了多模态材料和语义分割任务的实验结果和消融研究，包括定性和定量分析。

2 相关工作

图像分割在机器学习和计算能力的进步推动下取得了显著进展。 这一进展的重要改进是随着全卷积网络（FCNs）的出现 [34, 47]，它通过在卷积神经网络（CNNs）中利用分层特征实现了像素级预测。 这导致了基于CNN的各种模型用于不同的图像分割任务。 U-Net [39]是一种利用低分辨率和相应高分辨率特征图之间的跳跃连接的模型。 DeepLabV3+ [4]将扩张卷积（空洞卷积）引入编码器，允许扩展感受野而不显著增加计算复杂性。

PSPNet [54] 引入了全局上下文模块，使模型能够从广泛的空间尺度范围内收集信息，从而将局部和全局上下文整合到分割过程中。

最近，基于Transformer的模型已被证明在处理复杂的图像分割任务中非常有效。
一些值得注意的基于Transformer的模型包括金字塔视觉Transformer（PVT）[45]，SegFormer [48]和Mask2Former [6]。 PVT [45] 利用基于Transformer的设计来处理各种计算机视觉任务。 SegFormer [48] 利用高效的自注意力和轻量级MLP解码器进行简单高效的语义分割。Mask2Former
[6] 使用掩码注意力以及像素解码器和Transformer解码器来处理任何分割任务。 它们的成功证明了这些模型在各种分割任务中提供最先进解决方案的能力。

在多模态图像分割的背景下，来自不同来源的数据融合 [52] 已经成为提取更丰富信息并提高准确性的手段。 已经提出了各种模型和融合策略用于RGB-深度分割任务。 FuseNet [18] 模型将深度特征图集成到RGB特征图中，而SA-Gate [5] 在融合之前利用分离和聚合门来相互过滤和重新校准RGB和深度模态。 ACNet提出了注意力互补模块 [20]，用于提取加权的RGB和深度特征进行融合。 RGB-热像图分割领域也日益受到关注。 最近的模型包括RTFNet [41]，通过将热特征与RGB进行逐元素加法实现融合，RSFNet

[27] 提出了残差空间融合模块，用于融合RGB和热红外模态，而EAEFNet [28] 利用注意力交互和注意力补充机制来融合RGB和热红外特征。 许多方法还专注于融合RGB-Lidar数据，包括使用Transformer块的TransFuser [38]，而LIF-Seg [55] 则依赖于粗糙特征提取、偏移学习和细化以实现有效融合。

虽然先前提到的研究集中在特定的模态对上，但一些最近的研究展示了在融合任意模态方面取得了有希望的结果。 CMX [50] 引入了跨模态特征校正和融合模块，将RGB特征与补充模态进行融合。 对于多模态材料分割，MCubeSNet
[29] 模型被提出，可以无缝集成四种不同的模态以增强分割准确性。
在任意模态语义分割的背景下，CMNeXt [51] 引入了自查询中心和并行池混合器模块，为融合多样模态提供了一种多功能方法。 此外，HRFuser [1] 利用多窗口交叉注意力在不同分辨率上融合不同模态，从而丰富模型性能。

尽管这些模型中有一些可以融合不同的模态，但它们要么使用非常复杂的融合策略 [1, 50, 51]，要么需要额外的信息 [29] 来执行基础任务。 我们的目标是设计一个简单的融合模块，可以处理任意数量的输入模态，并能够有效地融合来自不同模态组合的信息。

3 提出的模型

我们提出的MMSFormer模型和融合块的整体架构如图1所示。 该模型有三个模块：（1）模态特定编码器；（2）多模态融合块；和（3）共享MLP解码器。 我们使用混合Transformer [48] 作为我们模型的编码器。 我们选择混合Transformer有多种原因。 首先，它可以提供分层特征而无需位置编码。 其次，在注意力之前使用空间缩减可以显著减少参数数量[45, 48]。 第三，它还可以很好地与简单轻量级的MLP解码器配合使用[48]。

3.1 整体模型架构

我们的整体模型架构如图1a所示。 假设我们有 M个不同的模态。 给定一组模态作为输入，每个模态特定的编码器通过将相应的图像映射到模态特定的分层特征来捕获每个输入模态的独特特征，如$F_m = \text{Encoder}_m(I_m)$,

(1)

其中 $I_m \in \mathbb{R}^{H\times W\times 3}$表示模态 m的输入图像，其中m∈ {1,2, . . . , M}，而Encoderm(•)表示该模态的编码器。 编码器在输入图像分辨率的{4 1,1 8, − − $\frac{1}{16}, \frac{1}{32}$}生成四个特征图。我们将它们表示为 $F_m = \{F_m^1, F_m^2, F_m^3, F_m^4\}$。为简单起见，我们将第i个编码器阶段的特征图形状表示为 (Hi×Wi×Ci)，其中$_i$ ∈ {1,2,3,4 }。

我们使用四个单独的融合块，每个对应一个编码器阶段，来融合来自编码器每个阶段的特征。 我们将提取的特征 F_{mi}对于所有模态传递到第 i个融合块，如下所示：Fi=FusionBlocki({Fm

i}m). (2)

每个融合块将从所有模态提取的特征融合在一起，生成一个组合特征表示$F = \{F^1, F^2, F^3, F^4\}$，其中 F^i表示第 i阶段的融合特征。 最后，我们将组合特征F传递给MLP解码器[48]来预测分割标签。

3.2 模态特定编码器

我们使用混合Transformer编码器[48]来从输入模态中捕获分层特征。 每个输入图像I_m经过补丁嵌入层，其中它被分成 4×4个补丁，遵循[48]，然后馈送到混合Transformer块中。 混合Transformer块的设计如图1b所示。 我们将任何混合Transformer块的输入表示为 $X_{in} \in \mathbb{R}^{H_i\times W_i\times C_i}$，它被重塑为 $N_i \times C_i$（其中 $N_i = H_iW_i$），并用作查询 Q，键 K和值 V。

为了减少计算开销，在[45]之后应用空间缩减，使用缩减比例 R。 K和 V 首先被重塑为 $\frac{N_i}{R} \times C_iR$ 矩阵，然后通过线性投影映射到$\frac{N_i}{R} \times C_i$矩阵。 标准的多头自注意力（MHSA）将Q, K, V映射到中间特征，如MHSA$(Q, K, V) = \text{Concat}(\text{head}_1, \ldots,$

$\text{head}_h)W^O$，其中headj=Attention(^{Q}Wj Q, ^{K}Wj K, ^{V}Wj V)。 (3)

其中 h表示注意力头的数量， $W_j^Q \in \mathbb{R}^{C_i\times d_K}$， $W_j^K \in \mathbb{R}^{C_i\times d_K}$， $W_j^V \in \mathbb{R}^{C_i\times}$\$d$V，而WO∈ Rh$d$V ×Ci是投影矩阵， dK, dV 分别表示K, V的维度。 我们可以将注意力函数表达为

$$\text{注意} \quad (Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_K}})V, \tag{4}$$

其中Q, K,和 V是输入的查询、键和值矩阵。 MHSA 后面跟着一个混合层（包括两个 MLP 和一个 3×3卷积层）。 卷积层为变压器编码器提供足够的位置编码，以实现最佳分割性能 [48]。 这一层可以写成$\hat{X}_{in} = \text{MH SA}(Q, K, V)$,

$$X_{out} = \text{MLP}(\text{GELU}(\text{Conv}_{3\times 3}(\text{MLP}(\hat{X}_{in})))) + \hat{X}_{in}, \tag{5}$$

最后，根据 [48] 进行重叠补丁合并，生成最终输出。

3.3 多模态融合块

在提取分层特征后，我们使用提出的融合块将它们融合。 如图 1c 所示的融合块负责融合从特定模态编码器中提取的特征。 对于每个四个编码器阶段，我们都有一个融合块。 对于第 i个融合块，假设输入特征图如下Fmi∈RHi×Wi×Ci∀m∈ {1,2, . . . , M}. 首先，我们沿着通道维度连接来自M个模态的特征图，得到组合特征图Fi∈RHi×Wi×M Ci. 然后，我们通过一个线性融合层将特征传递，并将通道维度减少到Ci. 我们将结果特征表示为 $\hat{F}$i∈RHi×Wi×Ci. 我们将操作表示为

$$\hat{F}^i = \text{Linear}(F_1^i || \cdots || F_M^i). \tag{6}$$

这里 ||表示沿着通道维度连接特征，线性层接收$M\,C_i$维输入并生成 C_i维输出。

在线性融合层之后，我们添加了一个模块来捕捉和混合多尺度特征。 该模块由两个线性投影层组成，在它们之间有并行卷积层。 首先，我们沿着通道维度对 $\hat{F}^i$进行线性变换，通过第一个线性投影层。 它可以改进和调整来自不同通道的特征。 然后，我们应用 3×3， 5×5和 7×7卷积，以有效捕获不同空间上下文中的多样特征。 通过使用不同尺寸的卷积，融合块可以关注局部模式，同时捕获更大的空间结构，从而增强其从输入数据中提取有意义特征的能力。 最后，我们通过第二个线性投影层沿着通道维度应用另一个线性变换，以巩固并提升并行卷积捕获的信息，促进特征一致性并增强融合特征的区分能力。 这些步骤可以执行为

$$\tilde{F}^i = \text{线性}(F^{\wedge i}), \tag{7}$$

$$F^i = \text{线性}\ (\tilde{F}^i + \sum_{k\in\{3,5,7\}} \text{Conv}_{k\times k}(\tilde{F}^i)). \tag{8}$$

我们发现，使用3个大小分别为3×3、5×5和7×7的并行卷积层可以提供最佳性能。
增加卷积核大小会降低性能，我们在表7中展示了这一点。 由于较大的卷积核会降低性能，我们的模型中没有添加超过3个并行卷积。

我们在残差连接中应用了Squeeze-and-Excitation块 [19] 作为通道注意力。 最终融合特征可以表示为

$$F^i = \text{ChannelAttention}(F^{\wedge}) + F^i. \tag{9}$$

通道注意力重新校准通道之间的相互依赖关系，并允许模型选择最相关的特征或通道，同时抑制不太重要的特征 [19]. 这导致更有效的特征表示，从而在基础任务上表现更好。

3.4 共享MLP解码器

从所有4个融合块生成的融合特征被发送到共享的MLP解码器。 我们使用[48]中提出的解码器设计。 图1a中显示的解码器可以表示为以下方程式:

$$\begin{aligned} F^i &= \text{Linear}(F^i),\ \ \forall i \in \{1,2,3,4\} \\ F^i &= \text{Upsample}(F^i),\ \ \forall i \in \{1,2,3,4\} \\ F &= \text{Linear}(F^1 || \cdots || F^4), \\ P &= \text{Linear}(F). \end{aligned} \tag{10}$$

第一个线性层获取不同形状的融合特征，并生成具有相同通道维度的特征。
然后将特征上采样到 ^{1} ^{4} $^{-th}$ 原始输入形状的通道维度上连接，并通过另一个线性层传递，生成最终融合特征 F。 最后，将 F通过最后一个线性层生成预测的分割地图 P。

4 实验和结果

我们评估了我们的模型和提出的融合块在多个数据集上以及不同模态组合下进行多模态语义和材料分割任务。 我们还定性和定量地比较了我们的方法与现有基线方法。 我们尽可能报告已经发表的结果。 ∗表示我们使用了论文中的代码和预训练模型生成结果。

4.1 数据集

多模态材料分割（MCubeS）数据集 [29] 包含来自42个街景的500组图像，具有四种模态：RGB、线性极化角（AoLP）、线性极化度（DoLP）和近红外（NIR）。 它为材料和语义分割提供了注释的地面真实标签，并分为训练集（302组图像集）、验证集（96组图像集）和测试集（102组图像集）。 该数据集有20个类别标签对应不同材料。

表1：在多模态材料分割（MCubeS）数据集 [29] 上的性能比较。 这里A、D和N分别代表线性极化角（AoLP）、线性极化度（DoLP）和近红外（NIR）。

方法	模态%	mIoU
DRConv [3]	RGB-A-D-N	34.63
DDF [59]	RGB-A-D-N	36.16
TransFuser [38]	RGB-A-D-N	37.66
DeepLabv3+ [4]	RGB-A-D-N	38.13
MMTM [23]	RGB-A-D-N	39.71
FuseNet [18]	RGB-A-D-N	40.58
MCubeSNet [29]	RGB-A-D-N	42.46
CBAM [46]	RGB-A-D-N	51.32
CMNeXt [51]	RGB-A-D-N	51.54
MMSFormer (我们的)	RGB-A-D-N	**53.11**

表2:FBM [33] 数据集上的性能比较。 我们展示了不同方法的性能，这些方法来自已经发表的作品。

方法	模态%	mIoU
CBAM [46]	RGB-红外	50.1
GMNet [64]	RGB-红外	49.2
LASNet [25]	RGB-红外	42.5
EGFNet [11]	RGB-红外	47.3
FEANet [8]	RGB-红外	46.8
DIDFuse [58]	RGB-红外	50.6
ReCoNet [21]	RGB-红外	50.9
U2Fusion [49]	RGB-红外	47.9
TarDAL [32]	RGB-红外	48.1
SegMiF [33]	RGB-红外	54.8
MMSFormer (我们的)	RGB-红外	**61.7**

FMB 数据集 [33] 是一个具有1500对校准的RGB-红外图像对的新的具有挑战性的数据集。 训练集和测试集分别包含1220和280个图像对。 该数据集涵盖了不同光照和天气条件下的各种场景（泰达尔效应、雨、雾和强光）。 它还为14个不同类别提供了每像素的地面真实注释。

PST900数据集 [40] 包含894对同步的RGB-热像对。 该数据集被分为训练集和测试集，每个像素都有五个不同类别的地面真实注释。

4.2 实现细节

为了确保与先前模型的公平比较，我们遵循了先前研究中采用的相同数据预处理和增强策略 [29, 51, 33]. 我们使用了在ImageNet上预训练的Mix-Transformer (MiT) [48] 编码器作为我们模型的骨干，以从不同的模态中提取特征。 我们使用了Mix-Transformer (MiT) [9] 数据集作为我们模型的骨干，以从不同的模态中提取特征。 每个模态都有一个单独的编码器。 我们使用了在SegFormer [48] 中引入的共享MLP解码器，并对其进行了随机初始化。我们使用两个NVIDIA RTX 2080Ti GPU对所有模型进行训练和评估，并使用PyTorch进行模型开发。

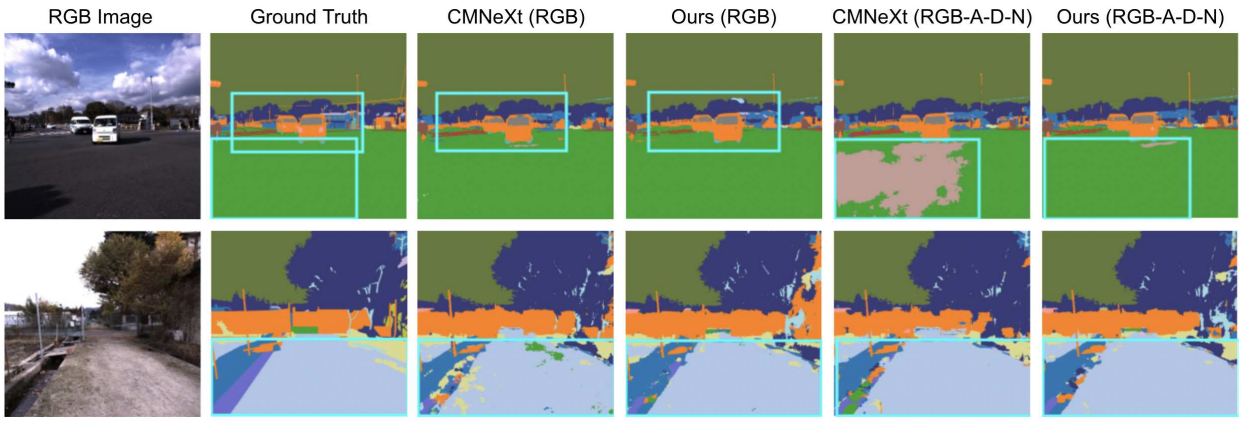
我们利用了一个多项式学习率调度器，幂为0.9，在训练过程中动态调整学习率。 前10个时代被指定为热身时代，学习率为原始速率的0.1倍。
为了损失计算，我们使用了交叉熵损失函数。 优化是使用AdamW [35]优化器进行的，epsilon值为 10^{-8}，权重衰减设置为0.01。对于CBAM [46]，我们使用相同的编码器、解码器和超参数，用CBAM[1]模块替换我们的融合块。 具体来说，在使用特定编码器从每个输入模态提取特征图后，我们将它们相加并将组合特征图传递给CBAM模块。

4.3 与现有方法的性能比较

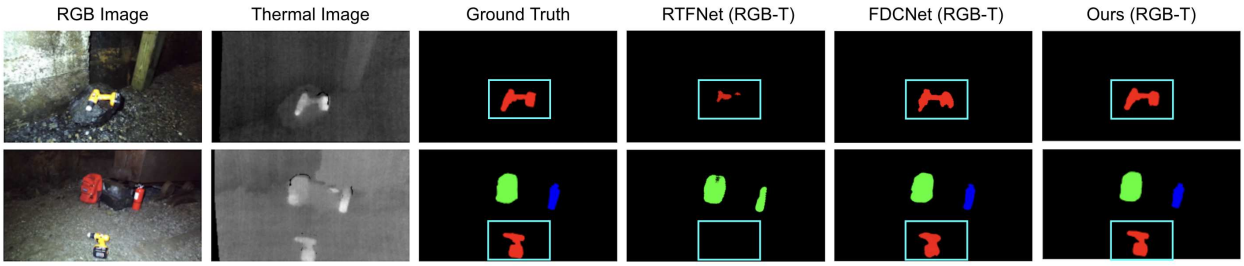
我们对我们的模型与三个数据集的已建立基线模型进行了严格的性能评估。 综合结果总结在表1–6中。 我们报告了来自我们实验的CBAM结果。
其他结果取自已发表的文献。

MCubeS数据集上的结果。表1显示了我们的模型与MCubeS数据集现有基线模型的整体性能比较。 我们的模型实现了53.11%的平均交并比(mIoU)，超过了当前最先进的模型。 它比CMNeXt [51]提高了1.57%，比CBAM提高了1.79%
[46] 并且相比MCubeSNet [29] 模型提高了10.65%。 为了进一步分析我们模型的性能，我们进行了逐类IoU分析，并在表3中呈现。 我们的模型在检测大多数材料类别方面表现比当前最先进的模型更好。 值得注意的是，我们的模型在检测塑料（+3.7%）、织物（+3.1%）、沥青（+2.3%）和鹅卵石（2.3%）类别方面表现出显著改善，同时在其他类别中保持竞争力或更好的性能。 这导致了整体更好的性能，并为该数据集设立了新的最先进水平。

[1] https://github.com/luuuyi/CBAM.PyTorch



(a) 在MCubeS数据集上的预测可视化



(b) 在PST900数据集上的预测可视化

图**2:**在MCubeS和PST900数据集上的预测可视化。 图2a展示了CMNeXt [51] 和我们模型在MCubeS数据集上的RGB和所有模态（RGB-A-D-N）预测。 为简洁起见，我们只展示了RGB图像和地面真实材料分割地图以及预测结果。 图2b展示了在PST900数据集上RGB-热输入模态的RTFNet [41]、FDCNet [57] 和我们模型的预测。 我们的模型在这两个数据集上都展示出更好的预测结果。

FMB数据集上的结果。 FMB数据集的性能比较见表2。 我们的模型相比当前最先进的模型提高了6.9%的mIoU。 该数据集的每类IoU分析见表4。 为了公平比较，我们仅比较文献中发布的8个类别（共14个类别）的性能。 T-Lamp和T-Sign分别代表交通灯和交通标志。 我们的模型在仅使用RGB进行预测时，相比最新的SegMiF，mIoU整体性能提高了6.7%。

[33] 模型。 除了卡车类别外，我们的模型在检测所有类别方面也表现出色，无论是仅使用RGB还是RGB-红外语义分割任务。 RGB-红外输入模式的性能比仅RGB性能好得多，这表明融合块能够有效地融合来自输入模式的信息。

在PST900数据集上的结果。我们还在PST900 [40] 数据集上测试了我们的模型，并在表5中总结了结果。 实验证明，我们的模型在这个数据集上的RGB-热像语义分割方面优于现有的基准模型。 它比最近的CACFNet [60] 模型在mIoU上高出0.89%。 我们的模型在检测数据集中可用的5个类别中的3个类别中也表现出更好的性能，并在其他两个类别中表现出竞争性能。

增量模态集成的性能比较

这项工作的一个关键方面涉及评估我们提出的融合块在结合来自不同模态的宝贵信息方面的有效性。 为了分析这种效果，我们在MCubeS数据集上训练了我们的模型，使用了各种模态的组合。 结果见表6。 我们的模型仅在RGB数据上训练，提供了50.44%的mIoU得分，比当前最先进的模型高出2.28%。 我们观察到随着我们逐渐整合额外的模态（AoLP，DoLP和NIR），性能逐步提高。集成导致性能逐步提升，mIoU从50.44%增加到51.30%，然后到52.03%，最终达到53.11%。这些发现作为一个有力的证据，表明我们的融合方法有效地利用和融合了来自不同模态组合的宝贵信息，从而显著提高了分割性能。

表3：在多模态材料分割（MCubeS）[29] 数据集上的每类IoU比较。 我们提出的MMSFormer模型在检测大多数类别时表现比当前最先进的模型更好。 ∗表示结果是使用作者提供的代码和预训练模型生成的。

方法	沥青	混凝土	金属	道路标线	织物	玻璃	灰泥	塑料	橡胶	沙子	砾石	陶瓷	鹅卵石	砖头	草地	木头	叶子	水	人类	天空	平均
MCubeSNet [29]	85.7	42.6	47.0	59.2	12.5	44.3	3.0	10.6	12.7	66.8	67.1	27.8	65.8	36.8	54.8	39.4	73.0	13.3	0.0	94.8	42.9
CBAM [46]	85.7	47.7	55.4	70.4	27.6	54.7	**0.9**	30.9	26.5	61.6	63.0	28.0	71.1	41.8	58.6	47.4	76.7	**56.3**	25.9	96.5	51.3
CMNeXt [51] ∗	84.3	44.9	53.9	**74.5**	32.3	54.0	0.8	28.3	**29.7**	**67.7**	66.5	27.7	68.5	42.9	58.7	**49.7**	75.4	55.7	18.9	96.5	51.5
MMSFormer（我们的）	**88.0**	**48.3**	**56.2**	72.2	**35.4**	**54.9**	0.5	**34.6**	29.4	67.2	**69.0**	**29.9**	**73.4**	**44.7**	**59.5**	47.8	**77.1**	50.5	**26.9**	**96.6**	**53.1**

表格 **4:**在 FMB [33] 数据集上，RGB 和 RGB-红外模式的每类IoU比较。 我们展示了已发布的 8 个类别（共 14 个）的比较。 T-Lamp 和 T-Sign 分别代表交通灯和交通标志。 我们的模型在除卡车类别外的所有类别上表现优异。

方法	模式	汽车	人类	卡车	交通灯	交通标志	建筑	植被	电线杆	% 平均IoU
SegMiF [33]	RGB	78.3	46.6	**43.4**	23.7	64.0	77.8	82.1	41.8	50.5
MMSFormer（我们的模型）	RGB	**80.3**	**56.7**	42.1	**31.6**	**77.8**	**77.9**	**85.4**	**48.1**	**57.2**
CBAM [46]	RGB-红外	71.9	49.3	20.9	25.8	67.1	75.8	80.9	19.7	50.1
GMNet [64]	RGB-红外	79.3	60.1	22.2	21.6	69.0	79.1	83.8	39.8	49.2
LASNet [25]	RGB-红外	72.6	48.6	14.8	2.9	59.0	75.4	81.6	36.7	42.5
EGFNet [11]	RGB-红外	77.4	63.0	17.1	25.2	66.6	77.2	83.5	41.5	47.3
FEANet [8]	RGB-红外	73.9	60.7	32.3	13.5	55.6	79.4	81.2	36.8	46.8
DIDFuse [58]	RGB-红外	77.7	64.4	28.8	29.2	64.4	78.4	82.4	41.8	50.6
ReCoNet [21]	RGB-红外	75.9	65.8	14.9	34.7	66.6	79.2	81.3	44.9	50.9
U2Fusion [49]	RGB-红外	76.6	61.9	14.4	28.3	68.9	78.8	82.2	42.2	47.9
TarDAL [32]	RGB-红外	74.2	56.0	18.8	29.6	66.5	79.1	81.7	41.9	48.1
SegMiF [33]	RGB-红外	78.3	65.4	**47.3**	43.1	74.8	82.0	85.0	49.8	54.8
MMSFormer（我们的）	RGB-红外	**82.6**	**69.8**	44.6	**45.2**	**79.7**	**83.0**	**87.3**	**51.4**	**61.7**

此外，我们的模型始终在所有模态组合上表现优异，超越了当前最先进的基准。 这种持续的优越性凸显了我们融合块的稳健性和多功能性，展示了其无论提供何种特定模态组合都能够适应和表现出色的能力。

4.5 预测的定性分析

除了定量分析外，我们还对预测的分割地图进行定性分析。 我们在图2a中展示了CMNeXt [51]模型和我们提出的MMSFormer模型预测的材料分割结果。
为简洁起见，我们在插图中仅展示RGB图像和地面真实材料分割地图。 我们展示了两种模型的RGB图像预测和所有模态（RGB-A-D-N）预测。 正如矩形边界框所示，我们提出的MMSFormer模型在仅使用RGB和所有模态（RGB-A-D-N）预测时，比CMNeXt [51]模型更准确地识别沥青、沙子和水。

我们还将在图2b中将我们在PST900 [40]数据集上的预测与RTFNet [41]和FDCNet [57]进行比较。 我们展示了输入的RGB图像、热像图、地面真实分割地图和模型的预测。 正如矩形边界框所示，与其他两种方法相比，我们的模型在检测具有更精确轮廓的对象时显示出更好的准确性。

4.6 融合块消融研究

我们进行了多项消融研究，旨在探讨融合块内各个组件对整体模型性能的贡献。 正如表7中详细描述的那样，研究结果揭示了这些组件的重要性。 在这些实验中，我们在训练和测试过程中使用了FMB数据集的RGB和红外模态。 首先，我们观察到残差连接中缺乏通道注意力会产生负面影响，导致性能降低3.36%。这表明沿通道维度进行特征校准在有效捕获和利用关键信息方面起着重要作用。 此外，在比较更大的

表5：PST900 [40] 数据集上的性能比较。 我们展示了所有类别的每类IoU百分比以及总体IoU百分比。

方法	模态	背景	灭火器	背包	手电钻	幸存者	% mIoU
ACNet [20]	RGB-热红外	99.25	59.95	83.19	51.46	65.19	71.81
CCNet [22]	RGB-热红外	99.05	51.84	66.42	32.27	57.50	61.42
高效FCN [30]	RGB-热红外	98.63	39.96	58.15	30.12	28.00	50.98
RTFNet [41]	RGB-热红外	99.02	51.93	74.17	7.07	70.11	60.46
PSTNet [40]	RGB-热红外	98.85	70.12	69.20	53.60	50.03	68.36
EGFNet [62]	RGB-热红外	99.26	71.29	83.05	64.67	74.30	78.51
MTANet [61]	RGB-热红外	99.33	64.95	87.50	62.05	79.14	78.60
MFFENet [63]	RGB-热红外	99.40	66.38	81.02	72.50	75.60	78.98
GMNet [64]	RGB-热红外	99.44	73.79	83.82	85.17	78.36	84.12
CGFNet [43]	RGB-热红外	99.30	71.71	82.00	59.72	77.42	78.03
GCNet [31]	RGB-热红外	99.35	77.68	79.37	82.92	73.58	82.58
GEBNet [10]	RGB-热红外	99.39	73.07	85.93	67.14	80.21	81.15
GCGLNet [13]	RGB-热红外	99.39	77.57	81.01	81.90	76.31	83.24
DHFNet [2]	RGB-热红外	99.44	78.15	87.38	71.18	74.81	82.19
MDRNet+ [56]	RGB-热红外	99.07	63.04	76.27	63.47	71.26	74.62
FDCNet [57]	RGB-热红外	99.15	71.52	72.17	70.36	72.36	77.11
CBAM [46]	RGB-热红外	99.43	73.81	82.75	80.00	69.60	81.12
EGFNet [11]	RGB-热红外	99.55	79.97	90.62	76.08	**80.88**	85.42
CACFNet [60]	RGB-热红外	99.57	**82.08**	89.49	80.90	80.76	86.56
MMSFormer（我们的）	RGB-热红外	**99.60**	81.45	**89.86**	**89.65**	76.68	**87.45**

表6：在多模态材料分割（MCubeS）[29]数据集上不同模态组合的性能比较（% mIoU）。 这里A、D和N分别代表线偏振角（AoLP）、线偏振度（DoLP）和近红外（NIR）。

模态	MCubeSNet [29]	CMNeXt [51]	MMSFormer (我们的)
RGB	33.70	48.16	**50.44**
RGB-A	39.10	48.42	**51.30**
RGB-A-D	42.00	49.48	**52.03**
RGB-A-D-N	42.86	51.54	**53.11**

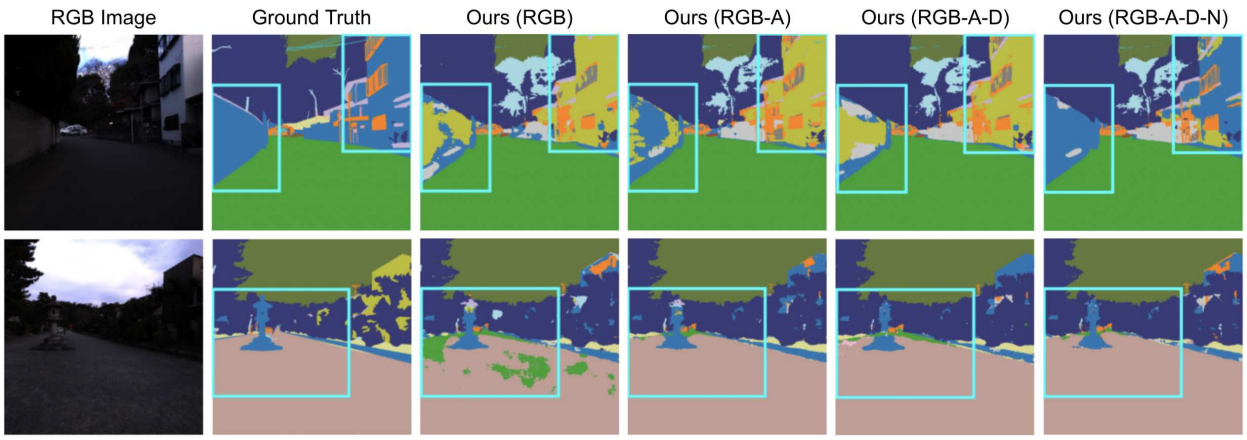
卷积核大小 $(3 \times 3, 7 \times 7,$ 和 $11 \times 11)$ 相对于最初使用的 $(3 \times 3, 5 \times 5,$ 和 $7 \times 7)$，我们注意到性能下降了5.36 %。这个结果强调了融合块中精心选择的卷积核大小的重要性。

此外，完全删除块中的并行卷积导致性能下降了4.51%，强调了它们在捕获多尺度特征和整体模型性能方面的重要贡献。 最后，如果我们只使用线性融合层来融合特征，并从融合块中删除并行卷积和通道注意力，性能显著下降了9.25%。这些研究表明，通过并行卷积捕获多尺度特征和使用通道注意力进行通道级特征校准对于学习更好的特征表示和因此对于整体模型性能至关重要。 这些全面的消融研究共同强调了融合块中每个组件的重要性，揭示了每个模块在实现我们模型的整体性能方面发挥着独特而重要的作用。

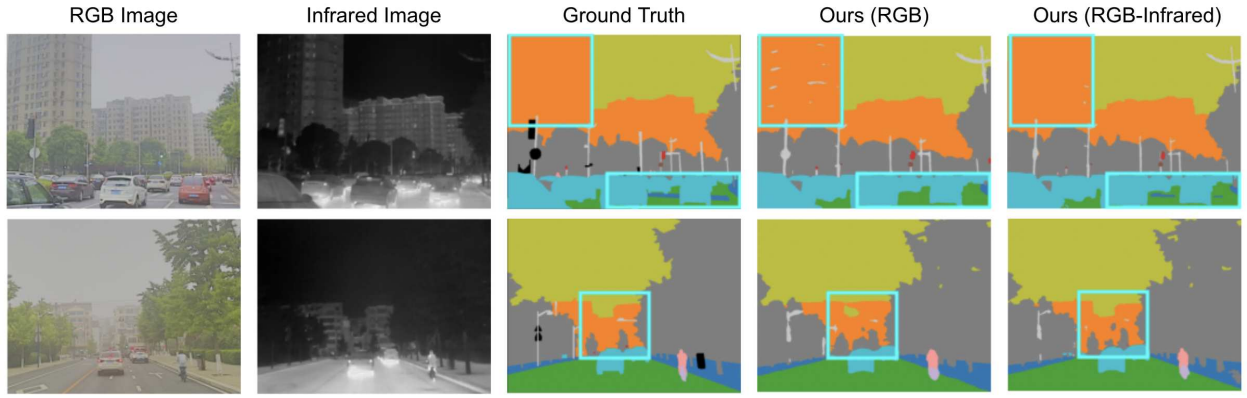
4.7 不同模态组合的消融研究

为了分析不同模态在识别不同材料中的贡献，我们进行了一系列消融研究，重点关注不同模态组合的每类IoU。 这些见解总结在表8中。 随着我们逐渐整合新的模态，特定类别的性能逐渐提高，其中包括草地、叶子、沥青、鹅卵石和塑料类别。 值得注意的是，NIR数据在分类沥青、混凝土、塑料、鹅卵石和人类等类别方面提供了帮助，在将NIR作为额外模态添加后，这些类别的性能显著提高。

相反，随着我们引入额外的模态，某些类别如水和砖逐渐表现下降。 这表明仅RGB数据就足以准确识别这些类别，而更多的包含



(a) 在MCubeS数据集上不同模态组合的预测可视化



(b) 在FMB数据集上不同模态组合的预测可视化

图 **3:**在MCubeS [29]和FMB [33]数据集上，不同模态组合的预测分割图的可视化。
两幅图都显示随着逐渐添加新的模态，预测准确性增加。 它们还展示了融合块有效地结合了不同模态组合的信息的能力。

表 **7:**在FMB [33]数据集上对融合块的消融研究。 在训练和测试过程中都使用了RGB和红外输入模态。 表格显示了融合块中不同模块对整体模型性能的贡献。

结构	参数数量 (M)	% mIoU (变化)
MMSFormer	61.26	61.68
- 没有通道注意力	61.21	58.32 (-3.36)
- 没有并行卷积	61.17	57.17 (-4.51)
- 使用3x3、7x7和11x11卷积	61.36	56.32 (-5.36)
- 仅线性融合	59.57	52.43 (-9.25)

模态的增加可能引入负面影响性能的噪音或冗余。 此外，AoLP似乎有助于增强对道路标线、玻璃和木材等材料的识别。 同样，DoLP提高了像灰泥、橡胶、沙子、砾石和陶瓷这样的类别的性能。 这些发现强调了不同成像模态之间的关系以及不同类型材料的独特特征，表明特定模态在检测特定类别时凭借其独特特征表现出色。

在图3a中，我们展示了一些示例，展示了如何通过添加不同的模态来提高分割性能。 我们展示了来自我们提出的MMSFormer模型的RGB、RGB-A、RGB-A-D和RGB-A-D-N输入的预测。 随着我们添加新的模态，预测变得更加准确，如边界框所示。 插图显示，通过添加额外的模态，对混凝土和砾石的识别变得更加准确。
图3b显示了来自FMB数据集的RGB和RGB-红外的预测。 如边界框所示，

表**8**：在多模态材料分割（MCu**be**S）[29]数据集上按类别的% IoU比较，用于不同的模态组合。 随着我们逐渐添加模态，整体性能逐渐提高。 这张表还显示了特定的模态组合有助于更好地识别特定类型的材料。

模态	沥青	混凝土	金属	道路标线	织物	玻璃	灰泥	塑料	橡胶	沙子	砾石	陶瓷	鹅卵石	砖头	草地	木头	叶子	水	人类	天空	平均
RGB	83.2	44.2	52.1	70.4	31.0	51.6	1.3	26.2	21.8	65.0	61.8	31.3	72.5	**45.0**	55.4	46.0	74.7	**56.0**	22.7	96.4	50.4
RGB-A	86.5	46.5	55.9	**73.0**	35.3	**56.0**	0.8	27.3	27.8	66.2	67.0	28.6	69.6	43.0	57.6	**49.6**	76.4	53.8	8.4	96.6	51.3
RGB-A-D	86.0	44.0	55.5	68.1	31.9	54.8	**2.3**	30.0	**29.7**	**69.4**	**73.7**	**32.2**	69.4	41.4	59.2	48.3	76.6	50.6	20.9	**96.7**	52.0
RGB-A-D-N	**88.0**	**48.3**	**56.2**	72.2	**35.4**	54.9	0.5	**34.6**	29.4	67.2	69.0	29.9	**73.4**	44.7	**59.5**	47.8	**77.1**	50.5	**26.9**	96.6	**53.1**

表**9**：融合块中参数数量和MCu**be**S数据集的GFLOPs比较，该数据集具有4个输入模态，每个模态的输入形状为$(3\times 512 \times 512)$。 我们的融合块显示出与现有方法相比显著较低的复杂性。

方法	融合块参数（M）	GFLOPs
CMNeXt [51]	16.63	6.47
MCubeSNet [29]	7.41	12.10
HRFuser [1]	**1.72**	17.50
CMX [50]	16.59	6.41
DDF (Resnet-101) [59]	28.10	4.10
MMSFormer（我们的模型）	3.23	**2.47**

添加新的模态有助于提高检测建筑物、道路和人行道的性能。 这也说明了融合块有效地融合了不同模态组合的信息的能力。

融合块的计算复杂度为4.8

除了更好的性能外，我们的融合块与大多数为这些数据集提出的融合块相比，在计算上也更有效率。 我们在表9中展示了MCubeS数据集上一些最近模型的融合块参数和GFLOPs的数量比较，该数据集具有4个输入模态，每个模态的输入形状为（$3 \times 512 \times 512$）。 从表中可以看出，我们提出的融合策略在参数数量和GFLOPs方面都明显低于现有方法。 HRFuser [1]的参数数量比我们的低，但需要超过7×GFLOPs。其他方法需要比我们的融合策略更多的参数$(2.3\times -8.7\times)$和GFLOPs$(1.6\times -7\times)$。 我们的比较仅包括这些结果在已发表文献中可用的模型。

5 结论

在本文中，我们介绍了一种新颖的融合模块，旨在将来自各种模态组合的有用信息结合在一起。 我们还提出了一个名为MMSFormer的新模型，该模型集成了所提出的融合块，以完成多模态材料和语义分割任务。 实验结果表明，该模型能够有效地融合来自不同模态组合的信息，从而在三个不同数据集上实现了新的最先进性能。 实验还表明，融合块能够从不同模态组合中提取有用信息，帮助模型始终胜过当前最先进的模型。 从仅有一个输入模态开始，随着我们添加新的模态，性能逐渐提高。 几项消融研究进一步突出了融合块的不同组件如何对整体模型性能做出贡献。 消融研究还揭示了不同模态有助于识别不同类型的材料类别。 然而，所提出模型的一个局限性是使用特定于模态的编码器，并且编码器的数量随着模态数量的增加而增长。

未来的工作将包括探索使用共享编码器对所有模态的可能性和有效性，研究并扩展模型与其他模态和多模态任务的能力。

致谢

这项工作部分得到AFOSR奖励FA9550-21-1-0330的支持。

参考文献

[1] Tim Broedermann, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Hrfuser: 用于2D目标检测的多分辨率传感器融合架构。 在2023年I*EEE*智能交通系统国际会议（ITSC）中。

[2] 蔡宇琪，周武杰，张丽婷，余璐，罗婷。Dhfnet: 用于RGB-热红外语义分割的双解码分层融合网络。《视觉计算》, 页码1-11, 2023年。

[3] 陈晋，王希军，郭子超，张祥宇，孙健。动态区域感知卷积。 在2021年I*EEE/CVF*计算机视觉与模式识别会议论文集中，页码8064-8073。

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 使用空洞可分离卷积的编码器-解码器用于语义图像分割。 在欧洲计算机视觉会议*(ECCV)*论文集中，2018年，第801-818页。

[5] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. 双向跨模态特征传播与分离聚合门用于RGB-D语义分割。 在欧洲计算机视觉会议*(ECCV)*中，2020年，第561-577页。

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 用于通用图像分割的掩蔽注意力掩蔽变换器。 在I*EEE/CVF*计算机视觉与模式识别会议论文集中，2022年，第1290-1299页。

[7] 程恒达，姜晓华，孙颖，王静丽。彩色图像分割：进展与展望。
模式识别，34(12)：2259–2281，2001年。

[8] 邓福勤，冯华，梁明健，王宏民，杨勇，高远，陈俊峰，胡俊杰，郭希悦，林天伦。Feanet：用于*R*GB-热红外实时语义分割的特征增强注意力网络。 在*2021*年*IEEE/RSJ*国际智能机器人与系统大会*(IROS)*，第4467–4473页。IEEE出版社，2021年。

[9] 邓嘉，董伟，Richard Socher，李立佳，李凯，李飞飞。Imagenet：一个大规模的分层图像数据库。 在*2009*年*IEEE*计算机视觉与模式识别大会，第248–255页，2009年。

[10] 邵华东，周武杰，钱晓红，于璐。Gebnet：用于RGB-T场景解析的图增强分支网络。*IEEE*信号处理通信，29：2273-2277，2022年。

[11] 邵华东，周武杰，徐才娥，严伟庆。Egfnet：用于RGB-热成像城市场景解析的边缘感知引导融合网络。*IEEE*智能交通系统交易，页1-13，2023年。

[12] Omar Elharrouss，Somaya Ali Al-Maadeed，Nandhini Subramanian，Najmath Ottakath，Noor Almaadeed，Yassine Himeur。全景分割：一项综述。 *ArXiv*，abs/2111.10250，2021年。

[13] 龚婷婷，周武杰，钱晓红，雷景生，于璐。全局上下文引导轻量级RGB-热成像城市场景理解网络。人工智能工程应用，117：105510，2023年。

[14] 顾文超，白爽，孔玲星。基于深度神经网络的2D实例分割综述。图像与视觉计算，120:104401，2022年。

[15] 郭艳明，刘宇，Theodoros Georgiou，Michael S Lew。使用深度神经网络进行语义分割综述。多媒体信息检索国际期刊，7:87–93，2018年。

[16] 郭哲，李翔，黄恒，郭宁，李全正。基于深度学习的多模态医学图像分割。*IEEE*辐射与等离子医学科学交易，3(2):162–169，2019年。

[17] Abdul Mueed Hafiz和Ghulam Mohiuddin Bhat。实例分割调查：现状。多媒体信息检索国际期刊，9(3):171–189，2020年。

[18] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture. 在亚洲计算机视觉会议, 2016年11月.

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. 在*IEEE*计算机视觉和模式识别会议论文集, 2018年, 7132–7141页.

[20] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. 在*IEEE*国际图像处理会议 *(ICIP)*, 2019年, 1440–1444页.

[21] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. 在Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, 和 Tal Hassner, 编辑,计算机视觉 – *ECCV 2022*, 2022年, 539–555页. Springer Nature Switzerland.

[22] 黄子龙，王兴刚，黄立超，黄畅，魏云超，刘文宇。Ccnet: 用于语义分割的交叉注意力。 在I*EEE/CVF*国际计算机视觉会议*(*ICC*V)*上，2019年，第603-612页。

[23] Hamid Reza Vaezi Joze，Amirreza Shaban，Michael L Iuzzolino和Kazuhito Koishida。Mmtm: 用于*C*NN融合的多模态传输模块。 在I*EEE/C*V*F*计算机视觉和模式识别会议论文集上，2020年，第13289-13299页。

[24] Alexander Kirillov，Kaiming He，Ross Girshick，Carsten Rother和Piotr Dollár。全景分割。 在I*EEE/C*V*F*计算机视觉和模式识别会议论文集上，2019年，第9404-9413页。

[25] 龚阳李，王一可，刘智，张新鹏，曾丹。带有位置、激活和锐化的RGB-*T*语义分割。*IEEE*交易视频技术电路与系统，33(3):1223–1235，2023年。

[26] 李杰，戴航，韩航，丁宇。Mseg3d：自动驾驶的多模态3D语义分割。 在*20*23年I*EEE/CVF*计算机视觉与模式识别会议（*CVPR*），页面21694–21704，洛杉矶，加利福尼亚，美国，2023年6月。IEEE计算机学会。

[27] 李平，陈俊杰，林彬彬，徐向华。RGB-热红外语义分割的残差空间融合网络。arXi*v:2306.10364*，*2023*年。

[28] 梁明健，胡俊杰，包晨宇，冯华，邓福勤，林天伦。用于*R*GB-热感知任务的显式注意增强融合。*I*EEE机器人自动化信函，8(7):4060–4067，2023年。

[29] 梁宇鹏，若瑟・若崎，野原翔平和西野康. 多模态材料分割。 在I*EEE/CVF*计算机视觉与模式识别会议论文集*(CVP*R*)*, 2022年6月，页码19800–19808。

[30] 刘建波，何俊俊，张佳伟，任杰米和李宏升. Efficientfcn: 全面引导的语义分割解码。 在计算机视觉 – E*CCV 20*20, *20*20年，页码1–17。 施普林格国际出版社，*20*20年。

[31] 刘金福，周武杰，崔月丽，余璐和罗婷. Gcnet: 网格状上下文感知网络用于RGB-热红外语义分割。神经计算。, 2022年9月，506(C):60–67。

[32] 刘金元，范鑫，黄占波，吴冠耀，刘日升，钟伟和罗中轩. 面向目标的双对抗学习和多场景多模态融合红外和可见光用于目标检测。 在I*EEE/CV*F计算机视觉与模式识别会议论文集, 2022年，页码5802–5811。

[33] 刘金元，刘竹，吴冠尧，马龙，刘日升，钟伟，罗忠轩，范鑫。多交互特征学习和全时多模态图像融合与分割基准。 在2023年计算机视觉国际会议。

[34] 乔纳森・朗，埃文・谢尔哈默，特雷弗・达雷尔。用于语义分割的全卷积网络。 在2015年IEEE计算机视觉与模式识别会议，第3431-3440页。

[35] 伊利亚・洛什奇洛夫和弗兰克・胡特。解耦权重衰减正则化。 在2019年学习表示国际会议。

[36] 谢尔文・米纳伊，尤里・博伊科夫，法蒂赫・波里克利，安东尼奥・普拉萨，纳瑟尔・凯塔纳瓦兹，德米特里・特尔佐普洛斯。使用深度学习的图像分割：一项调查。*IEEE*模式分析与机器智能交易，44(7)：3523-3542，2021年。

[37] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. 使用深度学习的图像分割: 一项调查。*IEEE*模式分析和机器智能交易, 44(7):3523–3542, 2021.

[38] Aditya Prakash, Kashyap Chitta, 和 Andreas Geiger. 多模态融合变压器用于端到端自动驾驶。 在*IEEE/CVF*计算机视觉和模式识别会议论文集, 页码7077–7087, 2021.

[39] Olaf Ronneberger, Philipp Fischer, 和 Thomas Brox. U-net: 用于生物医学图像分割的卷积网络。 在医学图像计算和计算辅助干预–*MICCAI 2015:* 第*18*届国际会议, 德国慕尼黑, *2015*年*10*月*5-9*日, 第三部分*18*, 页码234–241. Springer, 2015.

[40] Shreyas S. Shivakumar, Neil Rodrigues, Alex Zhou, Ian D. Miller, Vijay Kumar, and Camillo J. Taylor. Pst900: Rgb-thermal 校准, 数据集和分割网络。 在*2020*年*IEEE*国际机器人与自动化大会*(ICRA)*, 2020年第9441-9447页。

[41] 孙宇翔, 左伟勋, 刘明。RTFNet: 城市场景语义分割的RGB-热融合网络。*IEEE*机器人与自动化信函, 2019年7月, 4(3):2576-2583。

[42] Paul Upchurch 和 Ransen Niu。用于室内和室外场景解析的密集材料分割数据集。 在欧洲计算机视觉大会, 2022年第450-466页。Springer。

[43] 王杰, 宋可晨, 鲍彦琦, 黄黎明, 严云辉。CGFNet: 用于RGB-*T*显著目标检测的交叉引导融合网络。*IEEE*视频技术电路与系统交易, 2022年第32卷第5期:2949-2961。

[44] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. 理解卷积用于语义分割。 在*2018*年*IEEE*冬季计算机视觉应用会议*(WACV)*, 页码1451-1460。Ieee, 2018年。

[45] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao .金字塔视觉变换器: 一种多功能的密集预测骨干网络，无需卷积。 在*IEEE/CVF*国际计算机视觉会议, 页码568-578，2021年。

[46] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: 卷积块注意模块。在欧洲计算机视觉会议*(ECCV)*, 页码3-19，2018年。

[47] 吴辉凯，张俊格，黄凯琦，梁孔明，于一舟。Fastfcn：重新思考骨干网络中的扩张卷积用于语义分割。*arXiv*预印本 *arXiv:1903.11816*，2019年。

[48] 谢恩泽，王文海，于志鼎，Anima Anandkumar，Jose M Alvarez，和罗平。Segformer：具有transformers的简单高效的语义分割设计。 在神经信息处理系统（*N*eurIPS）中，2021年。

[49] 徐瀚，马佳怡，姜俊俊，郭晓杰，和凌海滨。U2fusion：统一的无监督图像融合网络。*IEEE*模式分析与机器智能交易，44(1)：502–518，2022年。

[50] 张佳明，刘华尧，杨凯伦，胡欣欣，刘瑞平，和Rainer Stiefelhagen。CMX：用于RGB-X语义分割的跨模态融合。 *IEEE*智能交通系统交易，2023年。

[51] 张佳明，刘瑞平，史浩，杨凯伦，西蒙・赖斯，彭坤宇，傅浩东，王凯伟，雷纳・施蒂费尔哈根。交付任意模态语义分割。 在2023年*IEEE/CVF*计算机视觉与模式识别会议论文集, 页码1136-1147。

[52] 张一飞，德西雷・西迪贝，奥利维尔・莫雷尔，法布里斯・梅里奥多。深度多模态融合用于语义图像分割：一项调查。图像与视觉计算, 105:104042, 2021.

[53] 张一飞，德西雷・西迪贝，奥利维尔・莫雷尔，法布里斯・梅里奥多。深度多模态融合用于语义图像分割：一项调查。图像与视觉计算, 105:104042, 2021.

[54] 赵恒爽，石建平，齐晓娟，王晓刚，贾宇亚。金字塔场景解析网络。 在2017年I*EEE*计算机视觉与模式识别会议论文集, 页码2881-2890。

[55] 林昭，周辉，朱星哥，宋晓，李洪升，陶文兵。Lif-seg：激光雷达和摄像头图像融合用于3D激光雷达语义分割。*IEEE*多媒体交易，页码1-11，2023年。

[56] 赵申露，刘一晨，焦强，张强，韩军功。减轻RGB-*T*语义分割的模态差异。*IEEE*神经网络与学习系统交易，页码1-15，2023年。

[57] 赵申露和张强。一种用于RGB-T语义分割的特征分割与征服网络。 *IEEE*视频技术电路与系统交易，33(6)：2892-2905，2023年。

[58] 赵子翔，徐爽，张春霞，刘俊敏，张江舍，李鹏飞。Didfuse：红外和可见光图像融合的深度图像分解。 在IJC*AI*会议上，页码970-976。ijcai.org，2020年。

[59] 周靖凯，Varun Jampani，皮志雄，刘琼和杨明寰。解耦动态滤波网络。在2021年*IEEE/CVF*计算机视觉与模式识别会议论文集，第6647-6656页。

[60] 周武杰，董少华，方美欣和余璐。Cacfnet：用于RGB-*T*城市场景解析的跨模态注意力级联融合网络。*IEEE*智能车辆交易，第1-10页，2023年。

[61] 周武杰，董少华，雷景生和余璐。Mtanet：用于RGB-*T*城市场景理解的具有分层多模态融合的多任务感知网络。*IEEE*智能车辆交易，第8卷第1期：48-58页，2023年。

[62] 周武杰，董少华，徐才娥和钱亚冠。边缘感知引导融合网络用于RGB–热场景解析。 在2022年*AAAI*人工智能会议论文集，第36卷，第3571-3579页。

[63] 周武杰，林新扬，雷景生，余璐，黄仁能。Mffenet：用于rgb-热成像城市道路场景解析的多尺度特征融合和增强网络。*IEEE*多媒体交易，24：2526-2538，2022年。

[64] 周武杰，刘金福，雷景生，余璐，黄仁能。Gmnet：用于rgb-热成像城市场景语义分割的分级特征多标签学习网络。*IEEE*图像处理交易，30：7790-7802，2021年。