基于可见光与红外热图像的行车环境复杂场景分割

陈武阳 [1, 2] 赵于前 [1, 3, 4] 阳春华 [1] 张 帆 [1, 3] 余伶俐 [1, 3] 陈白帆 [1, 4]

摘 要 复杂场景分割是自动驾驶领域智能感知的重要任务, 对稳定性和高效性都有较高的要求. 由于一般的场景分割方法主要针对可见光图像, 分割效果非常依赖于图像获取时的光线与气候条件, 且大多数方法只关注分割性能, 忽略了计算资源. 本文提出一种基于可见光与红外热图像的轻量级双模分割网络 (DMSNet), 通过提取并融合两种模态图像的特征得到最终分割结果. 考虑到不同模态特征空间存在较大差异, 直接融合将降低对特征的利用率, 本文提出了双路特征空间自适应 (DPFSA) 模块, 该模块能够自动学习特征间的差异从而转换特征至同一空间. 实验结果表明, 本文方法提高了对不同模态图像的利用率, 对光照变化有更强的鲁棒性, 且以少量参数取得了较好的分割性能.

关键词 场景分割, 可见光图像, 红外热图像, 双模分割网络, 双路特征空间自适应模块

引用格式 陈武阳, 赵于前, 阳春华, 张帆, 余伶俐, 陈白帆. 基于可见光与红外热图像的行车环境复杂场景分割. 自动化学报, 2022, **48**(2): 460−469

DOI 10.16383/j.aas.c210029

Complex Scene Segmentation Based on Visible and Thermal Images in Driving Environment

CHEN Wu-Yang[1, 2] ZHAO Yu-Qian[1, 3, 4] YANG Chun-Hua[1] ZHANG Fan[1, 3]
YU Ling-Li[1, 3] CHEN Bai-Fan[1, 4]

Abstract Complex scene segmentation is an important task of intelligent perception in the field of autonomous driving, which has high requirements for stability and efficiency. Since general scene segmentation methods mainly focus on visible images, the segmentation result is highly dependent on the light and weather conditions at the time of image acquisition, and most methods only focus on segmentation performance and ignore computing resources. This paper proposes a lightweight dual model segmentation network (DMSNet) based on visible and thermal images, which can extract and fuse the features of the two modal images to obtain a final segmentation result. For large differences in the feature spaces of different modalities, direct fusion will reduce the utilization of features. This paper proposes a dual-path feature space adaptation (DPFSA) module, which can automatically learn the differences among features and convert them to the same space. The experimental results show that this method can better utilize the inherent information between different modal images. Moreover, the proposed method is more robust to illumination changes and can achieve good segmentation performance with only a small number of parameters.

Key words Scene segmentation, visible images, thermal images, dual modal segmentation network, dual-path feature space adaptation module

Citation Chen Wu-Yang, Zhao Yu-Qian, Yang Chun-Hua, Zhang Fan, Yu Ling-Li, Chen Bai-Fan. Complex scene segmentation based on visible and thermal images in driving environment. *Acta Automatica Sinica*, 2022, **48**(2): 460−469

环境感知作为自动驾驶系统的重要环节, 对于车辆与外界环境的理解、交互起关键作用. 然而, 真实情景中的行车环境感知, 需要解决复杂场景下感知精度不高、实时性不强等关键技术问题. 行车环境感知主要包括目标检测与语义分割[1]. 语义分割在像素级别上理解所捕获的场景, 与目标检测相比, 能够产生更加丰富的感知信息, 并且分割结果可以

收稿日期 2021-01-09 录用日期 2021-04-16
Manuscript received January 9, 2021; accepted April 16, 2021
国家自然科学基金 (62076256), 中南大学研究生校企联合创新项目 (2021XQLH048) 资助
Supported by National Natural Science Foundation of China (62076256), Graduate School-enterprise Joint Innovation Project of Central South University (2021XQLH048)
本文责任编委 张向荣
Recommended by Associate Editor ZHANG Xiang-Rong
1. 中南大学自动化学院 长沙 410083 2. 中南大学计算机学院 长沙 410083 3. 湖南省高强度坚固件智能制造工程技术研究中心 常德 415701 4. 湖南湘江人工智能学院 长沙 410005
1. School of Automation, Central South University, Changsha 410083 2. School of Computer Science and Engineering, Central South University, Changsha 410083 3. Hunan Engineering & Technology Research Center of High Strength Fastener Intelligent Manufacturing, Changde 415701 4. Hunan Xiangjiang Artificial Intelligence Academy, Changsha 410005

进一步用来识别、检测场景中的视觉要素, 辅助行车环境感知系统进行判断. 目前, 相关的公共图像分割数据集与语义分割网络大多数都是基于可见光图像. 可见光图像能够记录物体丰富的颜色和纹理特征, 但在光照条件不足或光照异常时 (如: 暗黑中迎面的大灯照射), 可见光图像的质量会大幅降低, 导致网络无法正确分割对象, 进而影响行车环境感知系统在这些环境下的准确性. 红外热成像相机与可见光相机不同, 其通过探测物体热量获取红外辐射信息, 因此对光线与天气的变化更加鲁棒, 缺点在于红外热图像提供的信息量较少, 视觉效果模糊. 由此可见, 若仅依靠单一传感器, 难以精确分割不同环境下的场景. 本文主要研究行车环境下基于可见光与红外热图像的复杂场景分割, 尝试利用深度学习技术挖掘不同传感器之间的互补信息提升分割性能, 使车辆能够充分感知其周围环境.

场景分割作为行车环境感知的基本技术需求, 一直以来受到研究人员的关注. 目前, 绝大部分研究集中在可见光图像上, 分割方法从初期的基于阈值、区域、边缘等由人工设计特征的传统算法, 向基于深度学习的语义分割网络过渡; 研究内容则根据可见光图像分割的难点大致从增加分割精细度、增强网络对多尺度的泛化能力和学习物体空间相关性三个方向提升网络性能. 如文献 [2] 利用膨胀卷积模块用来保留特征图中的细节信息, 预测更加准确的结果; 文献 [3] 使用一个共享参数的卷积神经网络训练不同尺度的图像获得多尺度特征; 文献 [4] 利用循环神经网络适用于序列数据编码的特性, 捕捉物体的空间关系等. 虽然上述研究提高了分割准确率并解决了某些技术难题, 但大多数方法只注重提升精度而忽略了网络大小和分割速度, 导致所提出的方法难以在行车环境感知系统中落地. 此外, 基于可见光图像的分割方法无论如何改进, 其输入数据来源决定了这些方法无法避免因光线不足、分割对象与背景颜色纹理一致等导致的分割误差.

红外热成像相机由于其能够全天时、全天候有效工作的特性, 在车辆驾驶领域中的应用越来越广泛[5−6]. 例如, 对红外图像中的行人进行识别, 能提供危险区域、安全距离等重要信息, 从而辅助行车系统更好地进行路径规划, 提高其可靠性与鲁棒性. 一般来说, 面向红外图像的分割算法都是通过人工设计特征来描述前景与背景的差异, 如基于阈值、模糊集和最短路径等方法, 但它们通常对场景变化和噪声很敏感, 无法适应车辆所处的复杂环境.

近年来, 有学者开始关注基于多种传感器的感知方法[7], 尝试通过融合多模态数据充分挖掘信息, 提高行车感知系统的性能[8]. Ha 等[9] 首次尝试结合可见光与红外热图像进行场景分割, 提出了基于卷积神经网络的 MFNet 分割模型, 并创建了一个可见光与红外热图像的场景分割数据集. RTFNet[10] 在 MFNet 的基础上引入残差结构[11] 进一步加强了信息的融合, 提高了场景分割结果的准确性, 由于该网络结构过于庞大且参数数量显著增加, 与行车环境感知系统需要轻量级、实时性高的分割模型相违背, 有待进一步改进. 在此之前, 针对多传感器感知的研究集中在应用点云与可见光融合进行目标检测[12−13], 可见光与深度图像进行分割[14], 以及针对多光谱图像进行目标检测[15−16] 等.

本文提出一种基于可见光与红外热图像的复杂场景分割模型 DMSNet (Dual modal segmentation network), 该模型通过构建轻量级的双路特征空间自适应 (Dual-path feature space adaptation, DPFSA) 模块, 将红外热特征与可见光特征变换到同一空间下进行融合, 然后学习融合后的多模态特征, 并提取这些特征中的低层细节与高层语义信息, 从而实现对复杂场景的分割. 实验结果表明, 该模型可减少由于不同模态特征空间的差异带来的融合误差, 即使在光线发生变化时也表现出较强的鲁棒性, 分割结果相对其他方法也有明显改进.

1 方法

本文所构建的模型以复杂场景的可见光与红外热两种模态图像作为输入, 输出该场景中不同类别物体的分割结果, 我们因此将它命名为双模分割网络 (Dual modal segmentation network, DMSNet), 总体结构如图 1 所示.

1.1 场景分割模型

该网络主要包括编码器与解码器. 编码器使用两条路径分别提取可见光与红外热图像特征. 两条路径除了输入图像分别为彩色图像与灰度图像外, 其余部分结构一致, 均包含五组操作. 每组内包含一到三个 3×3 卷积层, 卷积层后紧接着批归一化 (Batch normalization) 层[17], 用来保持特征在网络内分布的相对稳定, 然后是激活层. 每组之间采用步长为 2 的最大池化层缩小特征图空间尺寸, 同时增加卷积核数目, 由编码器的浅层至深层逐步学习到图像内更加丰富的语义信息. 由于 DMSNet 是面向行车环境感知的轻量级网络, 特征通道数目在编

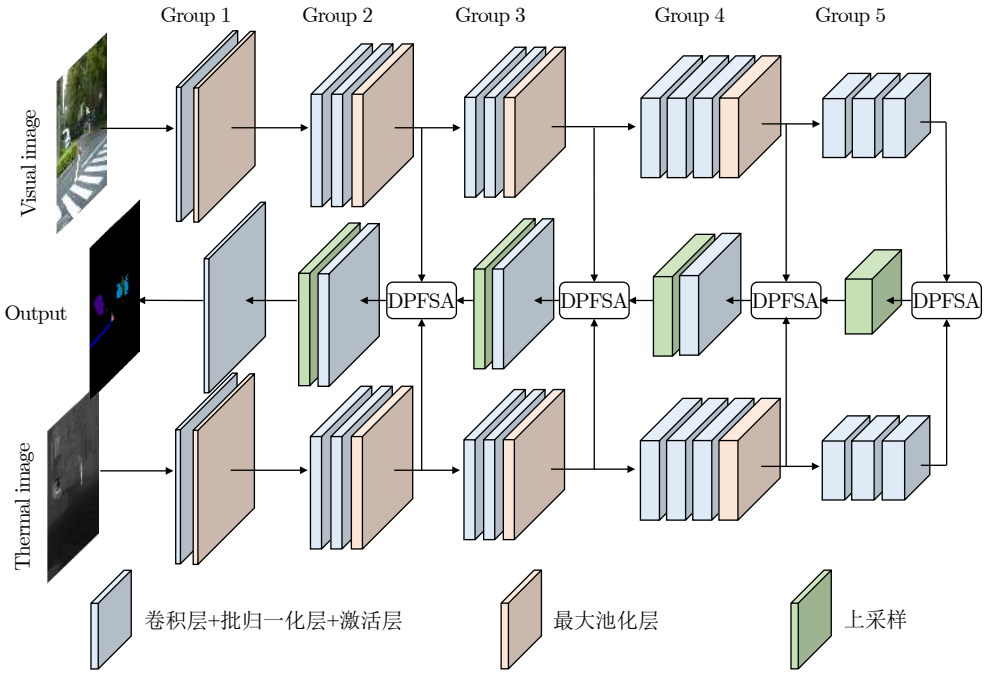


图 1 DMSNet 模型结构图

Fig. 1 The architecture of DMSNet

码器最深层也未超过 96, 因此采用 leaky-ReLU [18] 作为网络所有的激活函数, 这样做能够避免常用的 ReLU[19] 激活函数造成大量神经元失活的问题.

解码器负责融合两条编码路径学习到的特征, 依次通过五组操作逐步增加特征图空间尺寸, 并最终得到与输入图像尺寸一致的分割结果. 解码器每组内的操作与编码阶段类似, 包含卷积层、批归一化层与激活函数. 每组之间以缩放因子为 2 的最邻近插值法进行快速上采样, 以逐步恢复特征图空间尺寸. 进行上采样之前, 需要融合来自可见光编码器与红外热编码器同一尺寸的特征图. 为了缩小不同模态特征空间存在的差异, 本文提出双路特征空间自适应 (Dual-path feature space adaptation, DPFSA) 模块, 用来自动转换两种模态特征至同一空间, 并对它们进行融合. 该模块的详细设计将在第 1.2 节中阐述.

1.2 特征融合方法

文献 [13] 指出, 目前利用激光雷达数据与可见光图像融合进行道路检测的方法, 相对于仅基于可见光图像的算法, 正确率并没有明显提升. 这种现象主要是由于两种信息在数据空间与特征空间存在差异, 进而影响了二者的融合. 数据空间的差异是指激光雷达数据位于三维真实空间, 而可见光图像定义在二维平面上. 特征空间的差异来源于两种数据模态不同, 进而导致网络提取的特征也位于不同的空间, 这些都会对特征融合造成不利影响. 受该研究的启发, 本文将文献 [13] 中的特征空间转换 (Feature space transformation, FST) 模块进行改进并应用到 DMSNet 中.

FST 模块将激光雷达特征全部以逐点相加的方式融进可见光特征, 导致转换后的特征与未转换的特征发生混淆, 一定程度给激光雷达信息增加了噪声, 并有可能对可见光特征带来负面影响. 针对这种不足, 本文设计了 DPFSA 模块, 用来执行特征空间的转换. 该模块结构如图 2 所示, 相比 FST 模块, 最大的改进在于保留了不同模态数据的特征向量, 且增加了预适应步骤 (Pre-adaptation) 与逆转换层 (Reverse layer). 其中, 预适应步骤是为了增加模型的非线性能力; 逆转换层的设计则借鉴了文献 [17] 中的思想, 对转换完成的数据进一步执行卷积操作, 从而避免数据分布严重改变, 同时可增加模型的灵活性. 这些改进使得最终的场景分割模型在几乎不增加网络参数的情况下, 性能有了很大的提升.

该模块主要包含两个功能: 针对特征空间的转换, 以及将携带不同信息的特征进行融合. 对于特征空间转换, 首先使用一个 1×1 卷积层与 leaky-ReLU 激活层对红外热特征进行预适应, 然后将预适应后的红外热特征与可见光特征输入到转换网

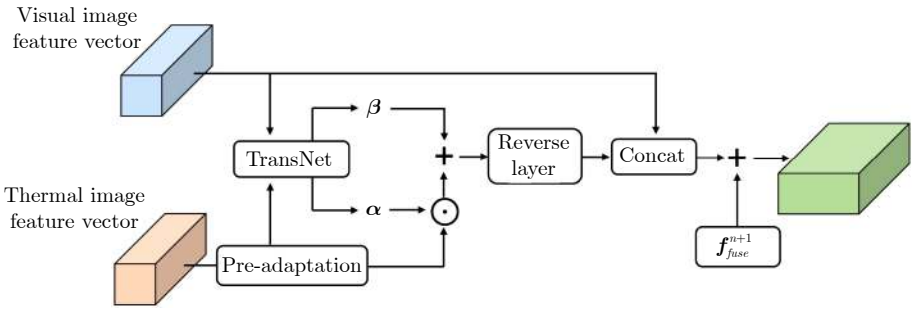


图 2　双路特征空间自适应模块 (DPFSA) 结构图

Fig. 2　The architecture of dual-path feature space adaptation module (DPFSA)

络 (TransNet) 学习转换参数, 最后经过逆转换层完成对红外热特征空间的转换:

$$\boldsymbol{f}_{adapt\_ther} = G_{rev}(\boldsymbol{\alpha}\boldsymbol{f}_{pre\_ther} + \boldsymbol{\beta}) \tag{1}$$

式中 $\boldsymbol{f}_{adapt\_ther}$ 为完成空间转换后的红外热特征, G_{rev} 代表逆转换层进行的操作, 逆转换层与预适应的结构相同, 仅包含单个 1×1 卷积层与激活层, 用来改变特征通道数, 同时增加模型的非线性; $\boldsymbol{f}_{pre\_ther}$ 是预适应后的红外热特征; $\boldsymbol{\alpha}$ 与 $\boldsymbol{\beta}$ 则代表 TransNet 输出的转换参数, 它们分别由 TransNet 内的两个转换子网络计算得到:

$$\boldsymbol{\alpha} = H_{\alpha}(\boldsymbol{f}_{pre\_ther}, \boldsymbol{f}_{vis}; \boldsymbol{W}_{\alpha}) \tag{2}$$

$$\boldsymbol{\beta} = H_{\beta}(\boldsymbol{f}_{pre\_ther}, \boldsymbol{f}_{vis}; \boldsymbol{W}_{\beta}) \tag{3}$$

其中 H_{α} 和 H_{β} 分别代表两个转换子网络计算 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 的全卷积运算, $\boldsymbol{W}_{\alpha}$ 和 $\boldsymbol{W}_{\beta}$ 则是对应的参数, $\boldsymbol{f}_{vis}$ 表示可见光特征.

完成对特征空间的转换后, 接着进行特征间的融合. 经过转换后的红外热特征首先与可见光特征进行拼接, 再与前一组已经融合的结果进行逐点相加达到融合效果, 得到双路特征. DPFSA 模块处理过程可表示为:

$$DPFSA(\boldsymbol{V}, \boldsymbol{T}; \boldsymbol{W}) = M_{fuse}^{n}\left(\boldsymbol{f}_{fuse}^{n+1}, \boldsymbol{f}_{vis}^{n}, \boldsymbol{f}_{adapt\_ther}^{n}\right) \tag{4}$$

其中, n 代表场景分割模型的第 n 组, $\boldsymbol{V}$、$\boldsymbol{T}$ 分别为可见光与红外热图像, $\boldsymbol{f}_{fuse}$ 代表 DPFSA 模块输出的双路特征经过解码器某一组卷积运算后的结果, $\boldsymbol{W}$ 泛指该模块所有参数, M_{fuse} 为逐点相加的融合过程. 需要注意, 在 $n=5$ 时, DPFSA 模块仅接收两个输入, 处理过程变为 $M_{fuse}^{5}(\boldsymbol{f}_{vis}^{5}, \boldsymbol{f}_{adapt\_ther}^{5})$, 结合图 2 与式 (4) 可见, DPFSA 模块不仅保留了两种模态信息形成双路特征, 而且该双路特征经过处理后能够继续作为下一个 DPFSA 模块的输入, 这种方式最大程度地减少了信息的杂糅与损失, 增加了对红外热图像的利用率.

1.3　损失函数

考虑到交叉熵损失在反向传播中更易优化, 而 Dice[20] 损失善于处理数据集中的类别不平衡问题, 本文构建新的损失函数 L_{mix} 如下:

$$L_{mix} = L_{CE} + L_{Dice} \tag{5}$$

$$L_{CE} = -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N}\pi\left(G\right)_i^k p_i^k \tag{6}$$

$$L_{Dice} = 1 - \frac{2}{K}\sum_{k=1}^{K}\frac{\sum\limits_{i=1}^{N} p_i^k \pi\left(G\right)_i^k}{\sum\limits_{i=1}^{N} p_i^k + \sum\limits_{i=1}^{N}\pi\left(G\right)_i^k} \tag{7}$$

其中 L_{CE} 表示交叉熵损失, L_{Dice} 表示 Dice 损失, K 为分割类别总数, G 代表图像对应的分割标签, N 为图像像素总个数, $\pi\left(G\right)_i^k$ 将图像 I 中第 k 类像素点 i 的分割标签映射为独热 (one-hot) 编码形式, p_i^k 映射任意数值到 $[0,1]$ 范围内, 其计算公式如下:

$$p_i^k = \frac{\exp\left(a_i^k\right)}{\sum\limits_{k=1}^{K}\exp\left(a_i^k\right)} \tag{8}$$

2　实验结果与分析

本文所有实验均通过基于 CUDA10.0 和 cuDNN7.6.0 的 PyTorch1.2.0 框架实现, 使用搭载了 Intel Xeon Bronze 3104 CPU (1.70 GHz) 和 NVIDIA GeForce RTX 2080 Ti (11 GB) 硬件的 Windows 10 电脑训练. 模型初始学习速率设置为 0.01, 每经过一轮迭代学习速率减少 1 %, 模型通过 SGD 随机梯度下降算法进行迭代优化, 并使用动量为 0.9、权重衰减系数为 0.0005 的策略避免模型过拟合. 本节首先介绍实验使用的数据集与评价指标, 然后通过消融实验验证 DMSNet 中 DPFSA 模块与混合

损失函数的有效性, 并分析它们对模型产生的影响及可能原因, 最后与其他分割模型进行对比.

2.1 数据集与评价指标

1) 数据集

本文主要使用文献 [9] 中公开的数据集 (后面统称为 "数据集 A"), 一共包含 1 569 幅行车环境下的城市场景图像, 其中 820 张拍摄于白天, 749 张拍摄于夜晚. 该数据集使用 InfRec R500 红外热成像相机拍摄, 该设备能够同时获取可见光与红外热图像. 数据集中一共有 8 个类别被标注, 分别是汽车 (Car)、行人 (Person)、自行车 (Bike)、路缘石 (Curve)、车辆停止标识 (Car stop)、护栏 (Guardrail)、路障 (Color cone) 和突出物 (Bump), 不属于上述类别的物体均以未标记 (Unlabeled) 处理. 由于场景中只有少量类别被标记, 未标记像素占据整体的 93 % 以上, 而已被标记的像素中, 不同类别像素占比相差达到 43 倍以上. 因此, 该数据集有较严重的类别不平衡问题. 在实际训练中, 本文采用了与文献 [9] 相同的数据划分策略, 50 % 的图像用于训练, 25 % 用于验证, 剩余的用作测试, 所有图像均被缩放至 480×640 固定尺寸.

由于面向行车环境的可见光与红外热多模态图像公开数据集稀缺, 本文使用 PST900 数据集[21] (后面统称为 "数据集 B") 作为实验补充. 该数据集面向机器人自主环境感知, 共包含 894 对 $720\times1\,280$ 大小的可见光与红外热图像, 具体有 5 个类别: 背景 (Background)、灭火器 (Fire-extinguisher)、背包 (Backpack)、手钻 (Hand-drill) 和幸存者 (Survivor). 数据划分策略与数据集 A 保持一致.

2) 评价指标

本文采用两个指标衡量分割结果的性能, 分别为正确率 (Acc) 和交并比 (IoU). 两个指标在所有类别上的平均结果分别以 mAcc、mIoU 指代, 计算公式如下:

$$mAcc = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{P_{ii}}{\sum\limits_{j=1}^{K}P_{ij}}\right) \tag{9}$$

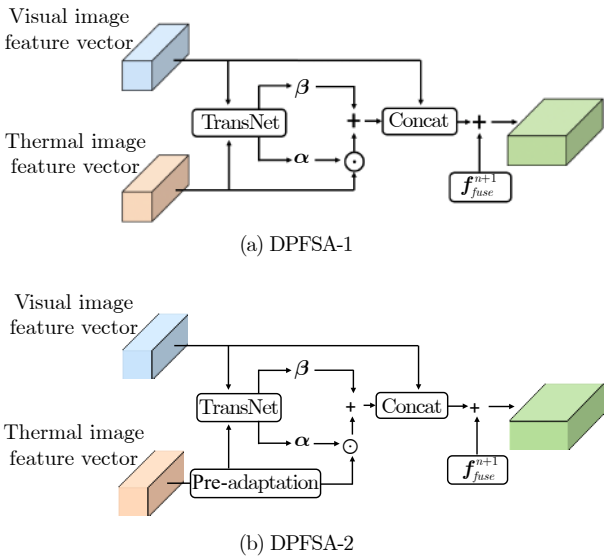
$$mIoU = \frac{1}{K-1}\sum_{i=2}^{K}\left(\frac{P_{ii}}{\sum\limits_{j=2}^{K}(P_{ij}+P_{ji})-P_{ii}}\right) \tag{10}$$

本文 K 在数据集 A、B 上分别取为 9 和 5, 即包含了未被标注的类别. P_{ij} 代表类别为 i 的像素被预测为类别 j 的数目. 在 mIoU 的计算中, 由于未被标注的像素占据绝大部分, 不同分割模型计算得到的 IoU 值非常接近, 因此该类别未被纳入考虑.

2.2 消融实验

1) DPFSA 模块分析

为了验证 DPFSA 模块的有效性, 现通过调整该模块内部结构得到另外两个模块, 并将它们和 MFNet、FuseNet[14] 进行对比实验. 两个调整后的模块如图 3 所示, 其中图 3 (a) 是在 DPFSA 基础上去掉逆转换层与预适应步骤, 为了表示方便, 将之命名为 DPFSA-1, 该模块的提出是为了证明对特征空间进行转换的思路是可行的; 图 3 (b) 是在 DPFSA 基础上去除逆转换层 (或者说在 DPFSA-1 的基础上增加了预适应步骤), 将之命名为 DPFSA-2, 该模块的提出是为了证明单纯增加网络参数或层数不一定能提升分割精度.



(a) DPFSA-1

(b) DPFSA-2

图 3 调整 DPFSA 内部结构得到的另外两个模块

Fig. 3 The other two modules obtained by adjusting the internal structure of DPFSA

为了保证比较的公平性, 排除损失函数对模型性能的影响, 将 MFNet 使用的交叉熵损失函数作为表 1 中 DMSNet 及其变种模型的损失函数, 并对白天与夜晚所有时间段内的图像进行测试. 由表 1 可知, 使用 DPFSA-1 模块的分割结果优于 MFNet 与 FuseNet, 表明不同模态特征空间的差异能够通过这种方式缩小, 对特征空间进行转换的思路可行; 使用 DPFSA-2 模块的实验结果虽然提升了 mAcc 指标, 但 mIoU 指标却有所下降, 表明单纯通过增

表 1　不同模块在数据集 A 上的 mAcc、mIoU 值与参数量比较

Table 1　Comparison of mAcc and mIoU values and parameter values of different modules on dataset A

Models	mAcc	mIoU	Parameters
MFNet	63.5	64.9	2.81 MB
FuseNet	61.9	63.8	46.4 MB
DMSNet (DPFSA-1)	65.6	68.1	5.45 MB
DMSNet (DPFSA-2)	68.9	65.1	5.54 MB
DMSNet (DPFSA)	**69.7**	**69.6**	5.63 MB

注: Parameters 代表整个分割模型的参数量, 而非模块的参数量

加网络参数或层数并不能保证模型正确率的提升, 更深的模型往往需要更多的训练数据, 且更难收敛. DPFSA 模块相比于 DPFSA-2 模块, 主要的不同在于将转换后的特征进一步输入到逆转换层, 由表 1 可以看出这种方式显著提升了模型性能, 且相比于未改进的 DPFSA-1, 参数量仅多出了 0.18 MB, 此外, 模型参数量也只有 FuseNet 的 12.1 %. 这也进一步证明, 模型性能的提升并非由于训练参数大量增多引起, 而是 DPFSA 模块起了关键作用.

2) 损失函数分析

本文基于交叉熵 (CE) 与 Dice 构建损失函数, 为了证明该损失函数的优越性, 在 DMSNet 上使用了四种不同的损失函数进行训练. 表 2 列出了不同损失函数在数据集 A 上各类别的 Acc 结果与 mAcc、mIoU 指标值. 其中, Focal 损失[22] 的提出即为了解决样本不均衡导致模型准确率降低的问题, 它通过调制系数 (Modulating factor) 减少易分类样本的权重, 从而使得模型在训练时更专注难分类的样本. 但从表 2 可以发现, Focal 损失在本文所使用的数据集上效果并不好, 很大程度是由于该数据集中不同类别的像素占比相差悬殊, 可达十几个数量级. 因此直接通过对难分类样本学习, 微小的噪声都将导致损失偏差严重, 影响模型收敛.

单独使用 Dice 损失函数效果也较差, 主要原因可能是 Dice 损失的梯度形式类似于 $2\pi(G)^2/(p+\pi(G))^2$, 在 p 与 $\pi(G)$ 均很小时, 该梯度会变得异常大, 导致整个训练过程不稳定. 虽然交叉熵损失不关注类别不平衡问题, 但其梯度更加简单、平稳, 并且能够学习到数据中主要类别的分布, 因此交叉熵损失相比 Dice 和 Focal 损失更适用于本文数据集.

基于以上分析, 为了让模型能学习到高频类别特征的同时也能兼顾低频类别, 本文使用了交叉熵与 Dice 相结合的混合损失函数. 由表 2 可知, 本文提出的混合损失函数在 mAcc 和 mIoU 指标上均表现最优, 可有效提升模型性能. 这在很大程度是由于交叉熵损失在网络训练前期起了主导作用, 而 Dice 损失作为辅助项, 进一步优化了在低频类别上的分割准确率.

2.3　对比实验

本节从准确性和鲁棒性角度将 DMSNet 分别与 SegNet、ENet、MFNet 和 FuseNet 的分割性能进行对比分析. 其中 SegNet 与 ENet 是针对可见光图像的分割网络, 之所以被选择为比较对象, 是因为这两种网络参数量适中, 并且 ENet 是专门针对嵌入式端的高速度分割网络. 其余大多数网络虽然在分割精度上表现更好, 却具有庞大的网络结构与参数量 (如 RTFNet, 模型参数量为 980.88 MB), 需要的硬件与计算条件也要求更高, 对于行车环境感知系统甚至可能无法承受. 为了确保对比实验的公平性, 分别使用两种图像训练并测试 SegNet 与 ENet. 第一种是可见光图像作为三通道输入, 用 3ch 表示; 第二种为结合了可见光与红外热信息的图像, 但由于 SegNet 与 ENet 本身不具备处理多模态数据的网络结构, 因此第二种直接由可见光图像与红外热灰度图像在色彩维度上拼接作为四通道输入, 用 4ch 表示.

表 3 展示了不同模型在数据集 A 上各个类别的 Acc 与 IoU 评价结果, 以及它们的平均值. 可以

表 2　不同损失函数在数据集 A 上的 Acc 结果与 mAcc、mIoU 值

Table 2　Acc results and mAcc and mIoU values of different loss functions on dataset A

Losses	Acc									mAcc	mIoU
	1	2	3	4	5	6	7	8	9		
CE	97.6	86.5	**84.9**	77.8	69.5	**53.3**	0.0	**79.8**	77.4	69.7	69.6
Focal	97.3	78.7	80.5	67.8	55.1	41.6	0.0	63.5	50.8	59.5	65.6
Dice	96.8	77.7	83.8	0.0	0.0	0.0	0.0	36.6	0.0	32.8	25.3
CE+Dice	97.6	**87.6**	83.5	**79.5**	**73.2**	47.5	0.0	74.7	**92.1**	**70.7**	**70.3**

注: 表中数字1 ~ 9为分割类别标号, 分别为 1: Unlabeled, 2: Car, 3: Pedestrian, 4: Bike, 5: Curve, 6: Car stop, 7: Guardrail, 8: Color cone, 9: Bump

表 3 不同模型在数据集 A 上的 Acc 与 IoU 结果对比
Table 3 Comparison of Acc and IoU results of different models on dataset A

Models	2		3		4		5		6		7		8		9		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
SegNet (3ch)	82.6	94.1	67.7	75.6	73.7	80.8	55.9	97.1	39.1	43.5	0.0	0.0	0.0	0.0	48.9	86.8	51.7	59.7
SegNet (4ch)	84.4	93.1	**85.5**	84.7	76.0	74.7	58.2	96.5	44.2	**43.6**	0.0	0.0	0.0	0.0	74.4	95.6	57.8	60.9
ENet (3ch)	85.3	92.3	53.8	68.4	67.7	71.7	52.2	95.7	16.9	24.2	0.0	0.0	0.0	0.0	0.0	0.0	41.5	43.8
ENet (4ch)	75.5	89.6	68.1	71.7	66.8	67.6	63.2	88.5	41.5	34.1	0.0	0.0	0.0	0.0	**93.2**	78.1	56.2	53.6
FuseNet	76.8	91.2	69.3	80.5	71.2	78.6	60.1	95.8	30.8	28.1	0.0	0.0	68.4	37.9	83.1	98.5	61.9	63.8
MFNet	78.9	92.9	82.7	84.8	68.1	75.7	64.4	97.2	31.6	29.7	0.0	0.0	71.8	40.6	77.1	98.4	63.5	64.9
DMSNet	**87.6**	**95.8**	83.5	**88.7**	**79.5**	**82.5**	**73.2**	**97.9**	**47.5**	35.7	0.0	0.0	**74.7**	**62.0**	92.1	**99.8**	**70.7**	**70.3**

注: 表中数字2 ~ 9为分割类别标号, 表示法同表 2

表 4 不同模型在数据集 B 上的 Acc 与 IoU 结果对比
Table 4 Comparison of Acc and IoU results of different models on dataset B

Models	2		3		4		5		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
SegNet (3ch)	0.0	0.0	71.2	79.3	0.0	0.0	21.6	47.1	38.4	31.6
SegNet (4ch)	0.0	0.0	62.9	70.1	0.0	0.0	30.5	46.8	38.5	29.2
ENet (3ch)	0.0	0.0	77.6	85.5	0.0	0.0	73.4	90.9	49.9	44.1
ENet (4ch)	0.0	0.0	72.9	74.9	0.0	0.0	74.8	89.6	49.1	41.1
FuseNet	**72.7**	43.1	**91.4**	**92.3**	74.4	78.9	**99.9**	99.8	87.4	78.5
MFNet	66.7	**47.0**	88.7	91.0	95.2	90.1	96.3	99.8	89.1	81.9
DMSNet	67.8	43.5	89.1	90.4	**96.3**	**97.5**	99.3	**99.9**	**90.2**	**82.8**

注: 表中数字2 ~ 5为分割类别标号, 分别为 2: Fire-Extinguisher, 3: Backpack, 4: Hand-Drill, 5: Survivor

看出, 除了个别类在其他模型上的分割结果具有优势外, 在绝大多数类别上 DMSNet 都更胜一筹, 且 mAcc 与 mIoU 指标相对于 MFNet 分别高出了 7.2 % 与 5.4 %, 相对于 FuseNet 则各高出了 8.8 % 与 6.5 %.

此外, 为了验证所提出模型在不同数据集上的适用性与鲁棒性, 表 4 展示了各模型在数据集 B 上的测试结果. 不难发现, 文本提出的方法同样具有较强的分割性能.

为了深入探究模型是否合理利用了两种模态信息, 本文进一步从时间角度比较不同模型对光照变化的鲁棒性. 表 5 列出了在白天与黑夜不同光线条件下不同模型在数据集 A 上的分割结果对比. 可以看出, 不经过任何处理直接将可见光与红外热图像拼接输入网络, 一定程度上影响了模型对于可见光数据的学习, 特别对于 SegNet 而言, 四通道输入相比三通道输入, 在白天的数据集上 mAcc 和 mIoU 有明显下降. 反观本文提出的 DMSNet, 在任意时间段的分割性能均有明显提高, 这也进一步说明 DMSNet 高效利用了两种模态数据的互补信息, 对光照的变化表现出较强鲁棒性.

表 5 不同模型在数据集 A 白天与黑夜环境下的 mAcc 与 mIoU 结果对比
Table 5 Comparison of mAcc and mIoU results of different models on dataset A in daytime and nighttime

Models	Daytime		Nighttime	
	mAcc	mIoU	mAcc	mIoU
SegNet (3ch)	47.8	55.5	52.6	61.3
SegNet (4ch)	45.4	49.3	58.2	62.9
ENet (3ch)	42.1	40.8	38.6	39.1
ENet (4ch)	44.1	45.9	57.1	54.3
FuseNet	50.6	61.2	63.4	64.7
MFNet	49.0	63.3	65.8	65.1
DMSNet	**57.7**	**69.1**	**71.8**	**71.3**

图 4 展示了 DMSNet、FuseNet 和 MFNet 在数据集 A 中 5 组测试图像上的分割结果, 其中第一行是可见光图像, 第二行是红外热图像, 第三行为分割标签, 前 3 幅拍摄于白天, 后 2 幅拍摄于夜晚. 第四、五、六行分别为 FuseNet、MFNet 和 DMS-

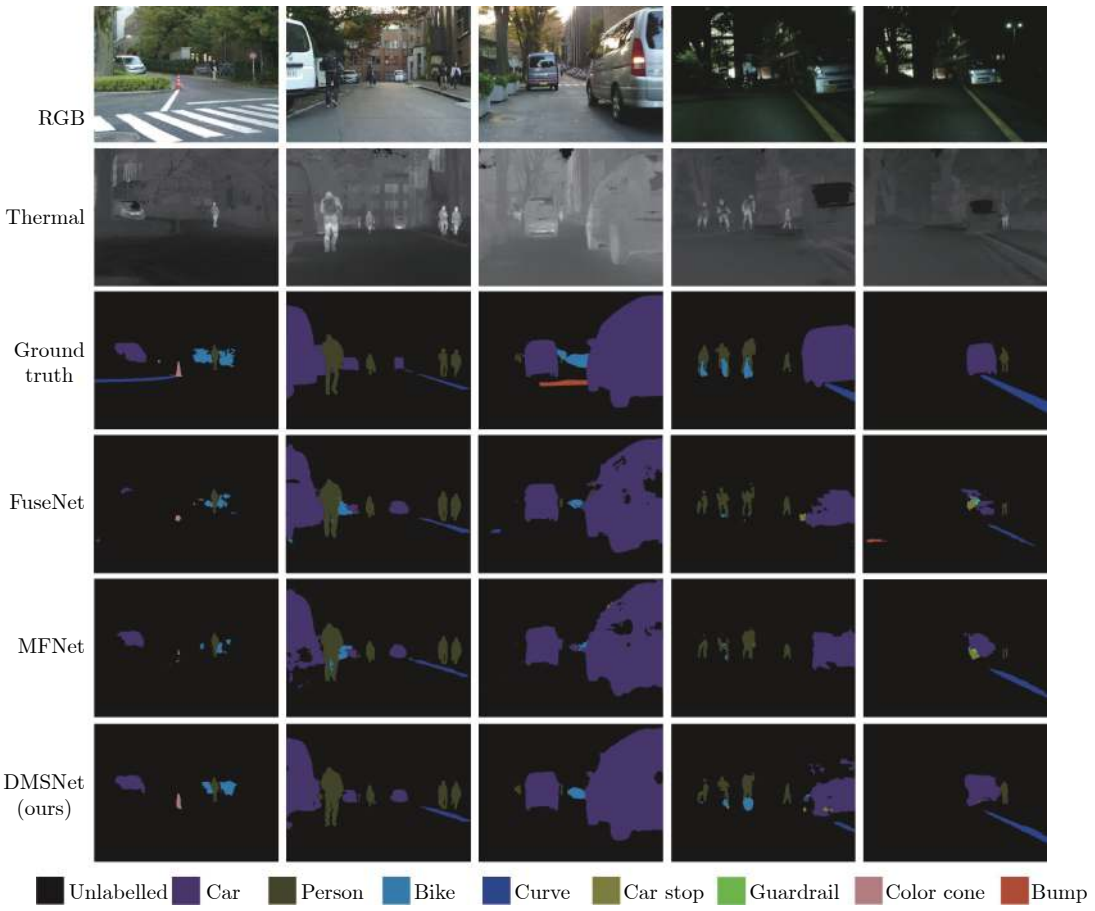


图 4　DMSNet、FuseNet 和 MFNet 在数据集 A 上的分割结果对比

Fig. 4　Comparison of segmentation results of DMSNet, FuseNet and MFNet on dataset A

Net 的分割结果. 可以看出, 相比于 MFNet 和 FuseNet, 本文提出的 DMSNet 对物体类别的判断更加准确, 如第一列中的路障与第四列中的自行车分割结果; 对边界细节的处理效果也更好, 如图中的行人; 另外分割结果的噪声也较少, 如第三列和第五列中的汽车分割结果.

3　结束语

针对现有场景分割模型大多基于可见光图像, 无法适应复杂环境变化, 且模型参数量庞大, 难以部署在行车环境感知系统中的问题, 本文构建了基于可见光与红外热图像的双模分割网络 DMSNet. 从可见光与红外热图像两种模态特征空间存在差异的角度入手, 提出了 DPFSA 模块. 该模块以十分轻量的操作对红外热图像特征进行转换, 缩小了两种模态特征空间的距离, 从而能够在几乎不增加模型参数的情况下, 有效改进模型性能. 另外, 使用本文提出的混合损失函数也可提升分割精度. 不足之处在于, 本文使用的数据集类别极其不平衡, 甚至存在错误标记、对类别划分标准不一致等情况, 导致场景中出现频率低的物体无法被准确分割, 因此, 下一步的工作需要从数据增强、模型优化等方面解决低频类别分割难的问题.

References

1　Feng D, Haase-Schütz C, Rosenbaum L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021, **22**(3): 1341−1360

2　Li C, Xu J X, Liu Q G, et al. Multi-view mammographic density classification by dilated and attention-Guided residual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020: 1−11

3　Chen L C, Yang Y, Wang J, Xu W, Yuille A L. Attention to scale: Scale-aware semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2016. 3640−3649

4　Lin D, Ji Y F, Lischinski D, Cohen-Or D, Huang H. Multi-scale context intertwining for semantic segmentation. In: Proceedings

of the European Conference on Computer Vision. Munich, Germany: Springer, 2018. 603−619

5 Kieu M, Bagdanov A D, Bertini M, Bimbo A D. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In: Proceedings of the 2020 European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 1−17

6 Liu T, Lam K M, Zhao R, Qiu G P. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, DOI: 10.1109/TCSVT.2021.3060162

7 Chen Hong, Guo Lu-Lu, Gong Xun, Gao Bing-Zhao, Zhang Lin. Automotive control in intelligent era. *Acta Automatica Sinica*, 2020, **46**(7): 1313−1332
(陈虹, 郭露露, 宫洵, 高炳钊, 张琳. 智能时代的汽车控制. 自动化学报, 2020, **46**(7): 1313−1332)

8 Zhang Xin-Yu, Zou Zhen-Hong, Li Zhi-Wei, Liu Hua-Ping, Li Jun. Deep multi-modal fusion in object detection for autonomous driving. *CAAI Transactions on Intelligent Systems*, 2020, **15**(4): 758−771
(张新钰, 邹镇洪, 李志伟, 刘华平, 李骏. 面向自动驾驶目标检测的深度多模态融合技术. 智能系统学报, 2020, **15**(4): 758−771)

9 Ha Q, Watanabe K, Karasawa T, Ushiku, Y, Harada, T. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, Canada: IEEE, 2017. 5108−5115

10 Sun Y X, Zuo W X, Liu M. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 2019, **4**(3): 2576−2583

11 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2016. 770−778

12 Zhao X M, Sun P P, Xu Z G, Min H G, Yu H K. Fusion of 3D LIDAR and camera data for object detection in autonomous vehicle applications. *IEEE Sensors Journal*, 2020, **20**(9): 4901−4913

13 Chen Z, Zhang J, Tao D C. Progressive lidar adaptation for road detection. *IEEE/CAA Journal of Automatica Sinica*, 2019, **6**(3): 693−702

14 Hazirbas C, Ma L, Domokos C, Cremers D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Proceedings of the Asian conference on Computer Vision. Taipei, China: Springer, 2016. 213−228

15 Zhang Y T, Yin Z S, Nie L Z, Huang S. Attention based multi-layer fusion of multispectral images for pedestrian detection. *IEEE Access*, 2020, **8**: 165071−165084

16 Guan D Y, Cao Y P, Yang J X, Cao Y L, Yang M Y. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 2019, **50**: 148−157

17 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning. Lille, France: ACM, 2015. 448−456

18 Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the International Conference on Machine Learning. Atlanta, USA: ACM, 2013. 1−6

19 Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, Florida, USA: JMLR, 2011. 315−323

20 Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the Fourth International Conference on 3D Vision. Stanford University, Stanford, CA, USA: IEEE, 2016. 565−571

21 Shivakumar S S, Rodrigues N, Zhou A, Miller L D, Kumar V, Taylor C J. Pst900: Rgb-thermal calibration, dataset and segmentation network. In: Proceedings of the International Conference on Robotics and Automation. Paris, France: IEEE, 2020: 9441−9447

22 Lin T Y, Goyal P, Girshick R, He K M, Dollar P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2980−2988

陈武阳 中南大学自动化学院和计算机学院硕士研究生. 主要研究方向为计算机视觉与智能感知.
E-mail: chenwuyanghn@163.com
(**CHEN Wu-Yang** Master student at the School of Automation, and School of Computer Science and Engineering, Central South University. Her research interest covers computer vision and intelligent perception.)

赵于前 中南大学自动化学院教授. 主要研究方向为计算机视觉, 智能感知, 机器学习, 精准医疗. 本文通信作者.
E-mail: zyq@csu.edu.cn
(**ZHAO Yu-Qian** Professor at the School of Automation, Central South University. His research interest covers computer vision, intelligent perception, machine learning, and precision medicine. Corresponding author of this paper.)

阳春华 中南大学自动化学院教授. 主要研究方向为复杂工业过程建模与优化控制, 智能自动化控制系统, 自动检测技术与仪器装置.
E-mail: ychh@csu.edu.cn
(**YANG Chun-Hua** Professor at the School of Automation, Central South University. Her research interest covers modeling and optimal control of complex industrial process, intelligent automation control system, automatic measurement technology and instrument.)

张 帆 中南大学自动化学院讲师. 主要研究方向为图像处理, 激光制造.
E-mail: zhangfan219@csu.edu.cn
(**ZHANG Fan** Lecturer at the School of Automation, Central South University. His research interest covers image processing and laser process.)

余伶俐 中南大学自动化学院教授. 主要研究方向为智能车辆路径规划与导航控制.
E-mail: llyu@csu.edu.cn
(**YU Ling-Li** Professor at the School of Automation, Central South University. Her research interest covers intelligent land vehicle path planning and navigation control.)

陈白帆 中南大学自动化学院副教授. 主要研究方向为智能驾驶, 环境感知, 计算机视觉.
E-mail: chenbaifan@csu.edu.cn
(**CHEN Bai-Fan** Associate professor at the School of Automation, Central South University. Her research interest covers intelligent vehicle, environment perception, and computer vision.)