Modeling and Simulation 建模与仿真, 2025, 14(4), 205-216
Published Online April 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.144279

Hans汉斯

基于多尺度空谱交互网络的多光谱目标检测

陆召阳，张荣福，景　李，魏辉光

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年3月8日；录用日期：2025年4月1日；发布日期：2025年4月11日

摘　要

近年来，可见光图像与热红外图像结合的多光谱目标检测，因其互补特性得到了广泛应用。然后，现有的大多数多光谱目标检测模型主要关注图像的局部特性，忽视了图像全局特征的提取，同时在特征提取过程中往往会丢失关键信息，如纹理和边缘等细节，导致提取的图像特征信息不足。针对这些问题，本文提出了多光谱目标检测模型SSIDet。该模型通过构建多尺度编码网络，分别从热红外图像和可见光图像中提取不同尺度的局部－全局特征；接着设计了一种空间－光谱交互注意力网络，充分融合空间特征和光谱特征，同时通过减少特征之间的冗余来增强其互补性；最后引入多尺度重建网络，进一步实现空间特征与光谱特征的协同增强。通过在FLIR和LLVIP数据集上的大量实验验证，本文方法在性能上优于现有方法。

关键词

深度学习，多光谱目标检测，特征融合

Multi-Spectral Object Detection Based on Multi-Scale Spatial-Spectral Interaction Network

Zhaoyang Lu, Rongfu Zhang, Li Jing, Huiguang Wei

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Mar. 8th, 2025; accepted: Apr. 1st, 2025; published: Apr. 11th, 2025

Abstract

In recent years, multispectral object detection combining visible and thermal infrared images has been widely used due to its complementary characteristics. Then, most of the existing multispectral

文章引用: 陆召阳, 张荣福, 景李, 魏辉光. 基于多尺度空谱交互网络的多光谱目标检测[J]. 建模与仿真, 2025, 14(4): 205-216. DOI: 10.12677/mos.2025.144279

object detection models mainly focus on the local characteristics of the image, neglecting the extraction of global features of the image, and at the same time, key information, such as details of texture and edges, are often lost in the process of feature extraction, which leads to insufficient information of the extracted image features. Aiming at these problems, this paper proposes a multispectral object detection model SSIDet. The model extracts local-global features at different scales from thermal infrared images and visible images respectively by constructing a multiscale coding network; then a spatial-spectral interactive attention network is designed to fully integrate spatial and spectral features, and at the same time, its complementarity is enhanced by reducing redundancy between features; finally, a multiscale reconstruction network is introduced to further enhance feature extraction, and a multiscale reconstruction network is introduced to further enhance feature extraction. A multi-scale reconstruction network is introduced to further realize the synergistic enhancement of spatial and spectral features. Through extensive experimental validation on FLIR and LLVIP datasets, the method of this paper outperforms the existing methods in terms of performance.

Keywords

Deep Learning, Multi-Spectral Object Detection, Feature Fusion

1. 引言

目标检测是计算机视觉领域的基础研究之一，在监控[1] [2]，遥感[3]-[5]，自动驾驶[6] [7]等多个领域都得到了广泛的关注和应用。然而，传统可见光目标检测易受光照条件限制，例如在低光照或夜间条件下效果不佳，同时受雾、烟、灰尘等环境影响，检测性能会大幅下降。多光谱图像的引入可以有效解决上述问题。多光谱图像是通过捕捉物体在不同波长范围内的反射或辐射能量而获取的图像，通常包括红外和可见光的多个波段，比起常见图像，多光谱图像记录了更广泛的光谱信息。多光谱图像对于光照变化、阴影、大气散射等环境因素具有一定的鲁棒性，并且不同波段的光谱信息在一定程度上能够减轻环境因素对目标检测造成的干扰，提高检测结果的稳定性和可靠性。多光谱图像还具有光谱信息丰富、目标区分度高、上下文信息丰富等优点[8] [9]。例如，随着红外摄像头的普及，行人检测往往使用红外图像作为额外光谱信息，红外图像反应物体的热辐射，因此可以在光照条件差的环境下提供有效的信息，同时红外光能更好地穿透雨雾等遮挡物，从而侦测到遮挡条件下的目标。由于多光谱目标检测相较于传统目标检测鲁棒性更好、检测精度更高，其得到越来越多的关注[10]-[12]。

在多光谱目标检测中，有效的特征融合方法至关重要。大多数现有的多光谱目标检测方法分别从可见光图像和热红外图像中提取模态特征，然后直接对这些特征[13]-[15]进行相加或者拼接操作。由于没有明确的跨模态融合，这种融合策略在学习互补信息上受到了限制，导致性能较差。为了进一步探索融合策略，很多研究者开始考虑“中途融合”的策略，设计了各种模态特征之间的交互模块。具体而言，Zhou [16]等人提出了基于差异挖掘的 DMAF 模块，旨在通过挖掘可见光图像与热红外图像之间的差异信息，进一步提升特征的价值。Liu [17]等人则利用 GAN (生成对抗网络)技术，生成高空间分辨率的多光谱图像，并通过双子网络分别提取特征。Diao [18]等人提出了一种基于多尺度 GAN 的框架，通过逐步生成融合图像并逐级判别，从而实现更有效的特征融合。受多光谱图像融合方法的启发，Lee 等人[19]和 Fang 等人[20]分别提出了跨模态注意力 Transformer 和跨模态融合 Transformer，并将其应用于多光谱目标检测领

域。针对单一模态中的信息丢失和跨模态特征不对齐的问题，You 等人[21]设计了一个由多尺度聚合 Transformer 和跨模态合并融合机制共同构成的双流多尺度聚合网络，充分发挥 Transformer 和 CNN 的优势，从而实现对局部和全局上下文相关性的共同捕获。然而，现有方法在实际应用中仍存在以下局限性：多光谱图像包含不同尺度的物体特征，现有方法仅通过单一尺度卷积神经网络进行特征提取，导致模型在捕捉图像全局特征方面存在不足，并且在特征提取过程中会遗漏纹理、边缘等细节信息。此外，虽然通过叠加卷积层可以扩展模型的感受野，但这一策略也会导致模型在局部特征信息提取上出现偏差。综上所述，现有方法在特征融合策略上仍存在一定的局限性，亟需进一步优化以提升多光谱目标检测的性能。

为了切实地提取源图像中的局部细节特征与全局特征并借助这些特征生成高空间分辨率的融合图像，本文提出了一种基于多尺度空谱交互网络的热红外图像和可见光图像融合方法。通过构建多尺度 CNN-Transformer (CT)编码网络，从热红外图像和可见光图像中分别提取不同尺度的局部－全局特征，使提取的特征既具备通用性又具备全局性。设计了空间－光谱交互注意力网络，能够有效地融合空间特征和光谱特征。通过空间注意力和光谱注意力的交互作用，强化突出重点特征，降低特征之间的冗余，同时增强其互补性。此外，构建了多尺度重建网络，用于生成高空间分辨率多光谱图像。该模块通过对不同尺度特征的由粗粒度至细粒度的融合，逐步恢复融合图像中的空间信息和光谱信息，从而有效减少融合过程中的信息损失。

本文方法主要贡献概括如下：

1) 该方法通过构建多尺度 CT 编码网络，能够在热红外图像和可见光图像之间实现特征的多尺度提取。该网络能够粗到细逐步提取特征，从而全面描述源图像的空间特性和光谱特性。

2) 为实现对来自不同编码网络特征的有效融合，构建了一种空间－光谱交互注意力网络。此网络借助注意力机制，提取热红外图像与可见光图像的堆叠特征中的空谱细节信息。空间－光谱交互注意力网络有效降低了空谱特征之间的冗余度，强化了二者间的互补性。

3) 为集成多个尺度的空谱信息，构建起了一个多尺度重建网络。此网络对不同尺度的融合特征实施深度融合操作，逐步生成高空间分辨率的多光谱融合图像。

2. 网络模型

2.1. 模型概述

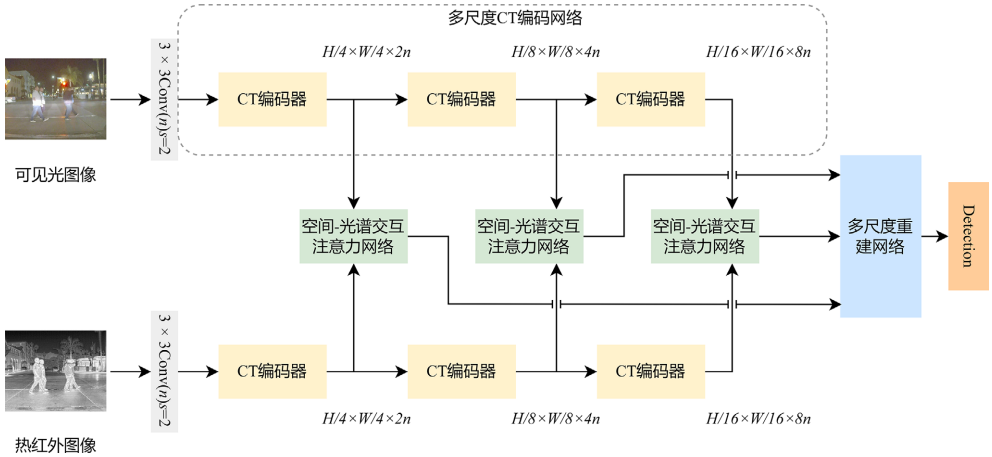

Figure 1. The overall structure of SSIDet
图 1. SSIDet 的整体结构

本文提出了一种多光谱目标检测网络 SSIDet，其模型架构如图 1 所示。该网络由多尺度 CT 编码网络、空间－谱交互注意力网络以及多尺度重建网络三个主要模块组成。SSIDet 采用了从粗到细的特征提取策略，通过两个多尺度 CT 编码网络分别从热红外图像和可见光图像中提取多尺度的空谱特征。其中，每个多尺度 CT 编码网络由一个卷积模块和 3 个 CT 编码器级联而成。随后，空间－谱交互注意力网络通过交互机制对不同编码网络的空间特征和光谱特征进行有效融合，从而去除信息冗余并实现信息互补。最终，多尺度重建网络对融合图像进行重建，并将重建后的图像输入目标检测头，完成目标检测任务。在本文方法中，$V \in R^{H\times W\times 3}$ 和 $I \in R^{H\times W\times 1}$ 分别表示可见光图像和热红外图像。n 表示卷积滤波器数量，设定为 32。W 和 H 分别代表图像的水平和垂直尺寸。

2.2. CNN-Transformer (CT)编码器

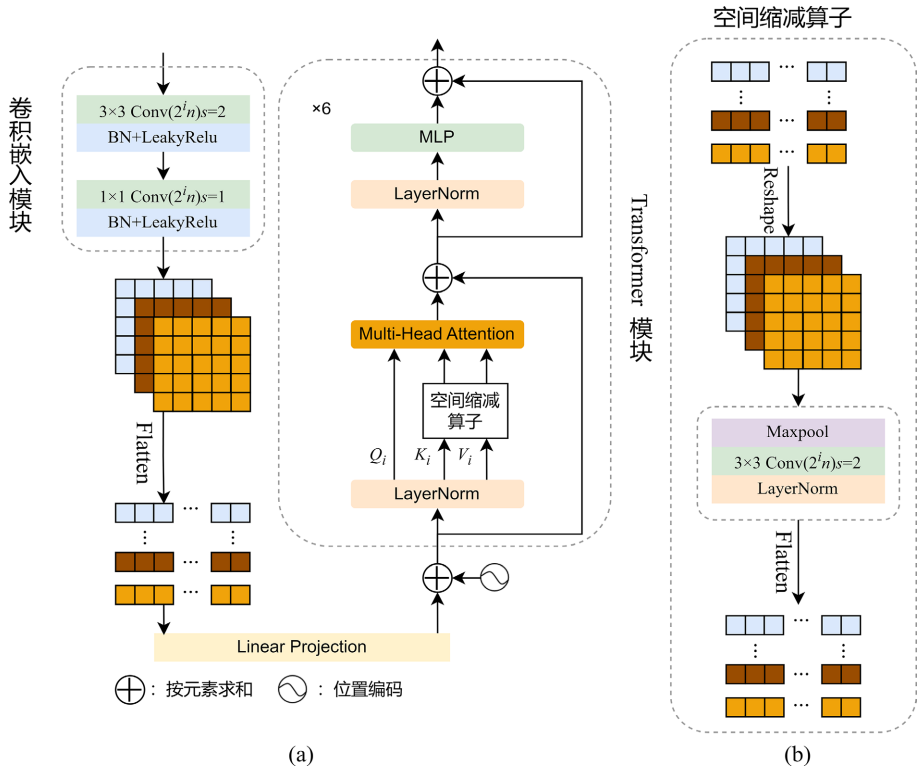

Figure 2. Illustration of CT Encoder module. (a) CT encoder; (b) Space reduction operator
图 2. CT 编码器模块示意图。(a) CT 编码器；(b) 空间缩减算子

为有效提取热红外图像和可见光图像中的全局特征和局部特征，本文设计了卷积嵌入模块和 Transformer 模块组成的 CT 编码器，其结构如图 2(a)所示。对于第 i 个 CT 编码器，将其对应的特征图经卷积嵌入模块处理后，能够有效地捕捉输入图像的局部特征并降低特征图的尺寸。具体而言，卷积嵌入模块包含两组卷积层，其滤波器尺寸为 3×3，步长分别为 2 和 1。接着，将特征图输入 Transformer 模块中，该模块旨在学习图像中的全局依赖关系。其中，Transformer 模块是由 6 个 Transformer 堆叠组成，输入是一维向量，为了将卷积嵌入模块输出的特征图转换为适合 Transformer 的输入形式，将特征图按照通道方向展平为一维向量，并添加位置编码。在经过层归一化得到 Q_i、K_i 和 V_i。为缓解 Transformer 常见的运算量过大问题，本文使用空间缩减算子对 K_i 和 V_i 进行下采样，如图 2(b)所示。空间缩减后的 $\mathcal{S}(K_i)$ 和 $\mathcal{S}(V_i)$ 计算公式如下：

$$\mathcal{S}\left(K_i\right)=\mathcal{R}^{-1}\left(\text{LayerNorm}\left(\text{Conv}\left(\text{Maxpool}\left(\mathcal{R}\left(K_i\right),\text{r}_i\right)\right)\right)\right)W_K^s \tag{1}$$

$$\mathcal{S}\left(V_i\right)=\mathcal{R}^{-1}\left(\text{LayerNorm}\left(\text{Conv}\left(\text{Maxpool}\left(\mathcal{R}\left(V_i\right),r_i\right)\right)\right)\right)W_V^s \tag{2}$$

其中，r_i 表示下采样率，编码网络的三个 CT 编码器里所包含的 r_i 分别设定为 4，2，1。公式(1)和(2)中的卷积层包含 $2^i n$ 个卷积核，卷积核大小为 3×3，步长为 1。$W_K^S\in R^{2^i n\times 2^i n}$ 和 $W_V^S\in R^{2^i n\times 2^i n}$ 代表对应的线性投影矩阵。$\mathcal{R}^{-1}\left(\cdot\right)$ 代表对多维输入进行展平操作，$\mathcal{R}^{-1}\left(\cdot\right)$ 和 $\mathcal{R}\left(\cdot\right)$ 互为逆操作。最终，将 Q_i、$\mathcal{S}\left(K_i\right)$ 和 $\mathcal{S}\left(V_i\right)$ 输入到多头注意力层和多层感知器层当中，有效地捕获特征间的全局依赖关系。多头注意力层具体计算公式如下：

$$\text{SRA}\left(Q_i,K_i,V_i\right)=\text{Concat}\left(\text{head}_{i,1},\cdots,\text{head}_{i,j},\cdots,\text{head}_{i,J}\right)W_i^O \tag{3}$$

$$\text{head}_{i,j}=\text{Attention}\left(Q_iW_{i,j}^Q,S\left(K_i\right)W_{i,j}^K,S\left(V_i\right)W_{i,j}^V\right) \tag{4}$$

$$\text{Attention}\left(q,k,\nu\right)=\text{softmax}\left(\frac{qk^T}{\sqrt{d}}\right)V \tag{5}$$

其中，$W_{i,j}^Q\in R^{2^i n\times d}$、$W_{i,j}^K\in R^{2^i n\times d}$ 和 $W_{i,j}^V\in R^{2^i n\times d}$ 表示第 i 个 CT 编码器中第 j 个注意力头的线性投影矩阵。$W_i^o\in R^{2^i n\times 2^i n}$ 表示多头注意力堆叠后的线性投影矩阵。J 表示注意力头的个数，在本文中设置为 8。d 设置为 $\frac{2^i n}{J}$。

2.3. 空间－光谱交互注意力网络

多尺度 CT 编码网络能够有效地提取热红外图像和可见光图像中的空谱信息。然而，大多数基于深度学习的融合方法在提取多尺度特征以后，只是对特征予以简单的拼接，忽视了特征之间的差异性，因此会导致特征之间出现部分冗余信息，进而对图像的融合结果产生影响。因此，本文设计了空间－光谱交互注意力网络，能够有效融合特征，减少信息冗余，其结构如图 3 所示。首先将特征图 F_P^i 和 F_L^i 分块并展开成序列得到相应的键和值。特征图 F_P^i 对应的 K_P^i 和 V_P^i 计算如下：

$$K_P^i=\text{LayerNorm}\left(\mathcal{F}\left(F_P^i\right)W_i^{K,P}\right) \tag{6}$$

$$V_P^i=\text{LayerNorm}\left(\mathcal{F}\left(F_L^i\right)W_i^{V,P}\right) \tag{7}$$

其中，$\left(\cdot\right)$ 表示为切分展开操作。$W_i^{K,P}$ 和 $W_i^{V,P}$ 表示为线性映射矩阵。同样，通过线性映射矩阵 $W_i^{K,L}$ 和 $W_i^{V,L}$，可得到特征图 F_L^i 对应的 K_L^i 和 V_L^i。Q_C^i 是将 F_P^i 和 F_L^i 堆叠后用同样计算方法得到：

$$Q_C^i=\text{LayerNorm}\left(\mathcal{F}\left(\text{Concat}\left(F_P^i,F_L^i\right)\right)W_i^Q\right) \tag{8}$$

其中，$\text{Concat}\left(\cdot\right)$ 表示堆叠操作。W_i^Q 是 $\text{Concat}\left(F_P^i,F_L^i\right)$ 的线性映射矩阵。然后，子网络输出特征间的空谱交互网络作用通过以下公式实现：

$$\text{Attention}\left(Q_C^i,K_P^i,V_P^i\right)=\text{Softmax}\left(\frac{Q_C^iK_P^{iT}}{\sqrt{d}}\right)V_P^i \tag{9}$$

$$\text{Attention}\left(Q_C^i,K_L^i,V_L^i\right)=\text{Softmax}\left(\frac{Q_C^iK_L^{iT}}{\sqrt{d}}\right)V_L^i \tag{10}$$

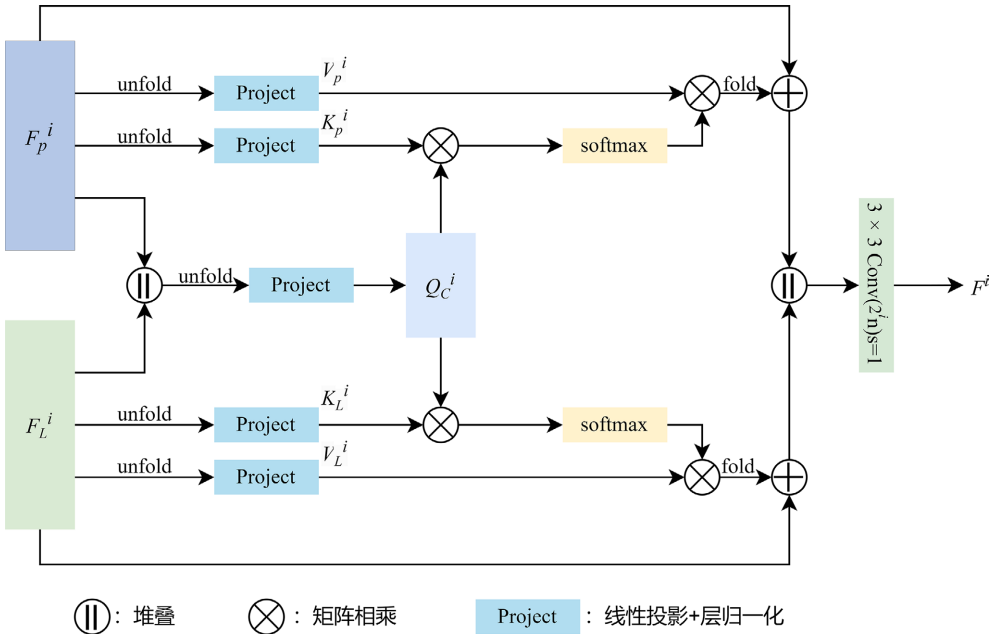

Figure 3. Illustration of spatial-spectral interaction module
图 3. 空间－光谱交互模块示意图

如图 3 所示，将 Q_C^i 与 K_P^i 相乘来捕获 Q_C^i 中的空间信息，并对输出结果进行编码。同样的，通过 Q_C^i 与 K_L^i 相乘来捕获 Q_C^i 的光谱信息。通过注意力估计降低特征间的信息冗余，同时进一步增强互补性。最终，将带有交互注意力的特征堆叠，并经过 3×3 的卷积层获取最终的融合结果，如公式(11)所示：

$$F^i = \mathrm{Conv}\left(\mathrm{Concat}\left(\mathrm{Attention}\left(Q_C^i, K_P^i, V_P^i\right), \mathrm{Attention}\left(Q_C^i, K_L^i, V_L^i\right)\right)\right) \tag{11}$$

2.4. 多尺度重建网络

多尺度重建网络采用多级特征融合策略，通过逐级上采样和特征重建，有效减少信息丢失。如图 4 所示，多尺度重建网络由多个上采样模块和跳跃连接组件构成。在特征融合过程中，每级特征图通过级联的上采样操作逐步提升空间分辨率。每个上采样模块包括卷积层和像素变换算子，卷积层用于调整特征图的通道数量，以增强空间信息与光谱信息的表征能力。随后，上采样得到的特征图与更高尺度的特征图进行融合，通过 3 × 3 卷积层和 LeakyReLU 激活函数，进一步优化特征表示，最终实现高空间分辨率多光谱图像的空间细节与光谱细节的重建。该网络显著提高了特征融合的精度，实现了空间特征与光谱特征的协同增强。

3. 实验结果与分析

3.1. 数据集与实验设置

FLIR (Forward-Looking Infrared)数据集[22]：FLIR 数据集是面向多光谱目标检测任务的权威基准数据集，由 Teledyne FLIR 公司发布，旨在推动红外热成像与可见光模态融合的算法研究。该数据集包含 14,452 张高分辨率热红外图像(分辨率 640 × 512)，同时提供部分对齐的可见光(RGB)图像，覆盖日间、夜间及复杂天气条件下的城市道路、高速公路等典型场景。数据标注涵盖行人(Pedestrian)、车辆(Car)、自行车(Cyclist)等 3 类目标，共标注约 10,228 个实例，以边界框(Bounding Box)及类别标签形式提供。

LLVIP (Low-Light Visible and Infrared paired)数据集[23]：LLVIP 数据集包含 30,976 张图像即

15,488 对，24 个夜晚场景，2 个白天场景。该数据集可用于图像转换(可见光转换到红外，红外转换到可见光)，图像融合，弱光行人检测以及红外行人检测等计算机视觉任务的研究。

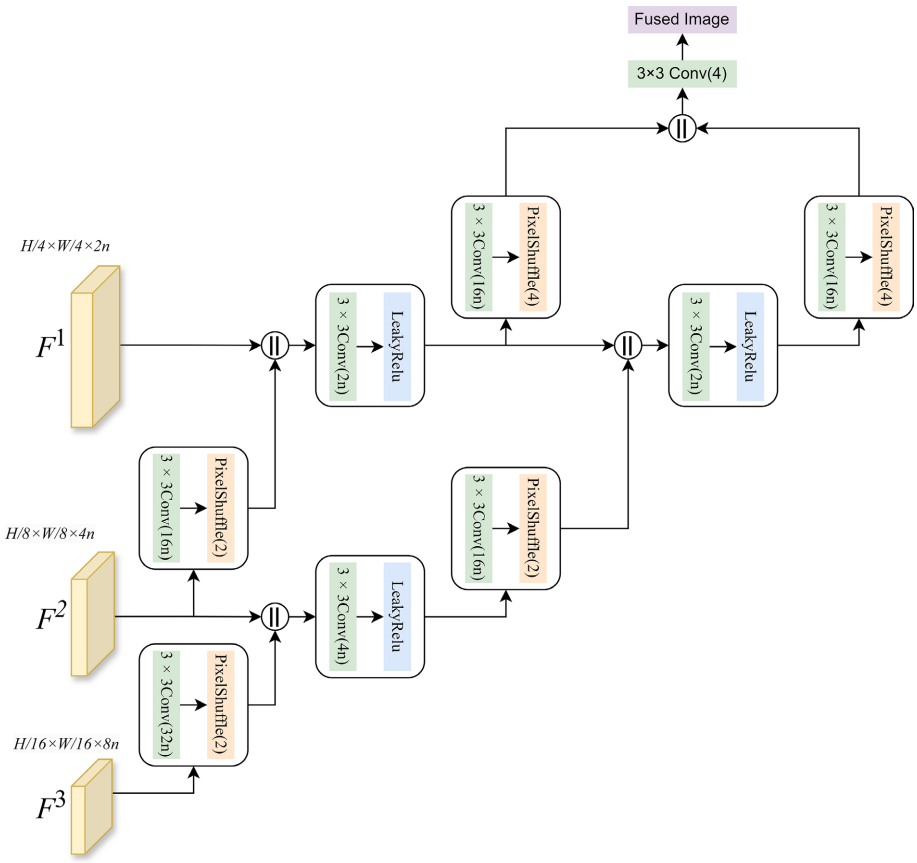

Figure 4. Illustration of multi-scale reconstruction module
图 4.多尺度重建模块示意图

实验设置：本文所有实验均在 Ubuntu 操作系统下运行，使用的硬件环境为 AMD EPYC 9754 CPU 和 24 GB 的 RTX 3090 GPU，深度学习框架为 PyTorch 1.10.0。实验设置的初始学习率为 0.01，优化器采用随机梯度下降(SGD)，动量参数设为 0.9，权重衰减系数为 0.0005，学习率使用余弦衰减策略。批量大小设置为 16，训练周期为 100 个 epoch。

评价指标：对于 FLIR 和 LLVIP 数据集，本文采用了常用的目标检测指标平均精度(AP)。根据分类的正确性和 IoU (Intersection over The Union)阈值进行正负样本的划分。平均平均精度(mAP)代表所有类别下平均精度的平均值。mAP_{50} 表示在 IoU = 0.50 时的平均准确率。mAP 指标表示 IoU 在 0.50 至 0.95 之间的平均准确率，跨度为 0.05。mAP 指标越高，表示性能越好。

3.2. 对比实验与分析

本节中，在 FLIR 和 LLVIP 数据集上对 SSIDet 与 4 种多光谱目标检测方法和 5 种单模态目标检测方法进行了对比实验。实验结果结果如表 1 和表 2 所示。如表 1 所示，本文提出的 SSIDet 模型在 FLIR 数据集上的性能普遍优于单模态方法。具体来说，与性能最好的 DDQ-DETR 相比，本文的方法提高了 7% 的 mAP_{50} 与 4.4%的 mAP，并且在 People、Cyclist 和 car 三个类别都取得了最佳的性能。结果表明，SSIDet

可以有效地融合两种模态的信息，提高检测结果。与多光谱检测方法相比，SSIDet 也取得了最优的性能。具体来说，与性能最好的 CSSA 相比，本文的方法提高了 1.7%的 mAP_{50} 与 0.2%的 mAP，并且在 People、Cyclist 和 car 三个类别也都取得了最佳的性能。这些指标的提升主要得益于 SSIDet 中 CT 编码器出色的跨模态互补特征提取能力、空间－光谱交互模块的优秀多模态融合能力，以及多尺度重建网络对全局特征和局部特征的全面感知和增强能力。

为进一步验证所提网络在其他场景下的泛化能力，本文在 LLVIP 数据集上进一步展开了实验。与 FLIR 数据集不同，LLVIP 数据集采用交通摄像头收集数据。因此，LLVIP 的实验证明了本文模型可以用于不同的现实场景，具有良好的鲁棒性。实验结果如表 2 所示，SSIDet 在 LLVIP 数据集上的性能也取得了最优的结果，分别取得了 95.5%的 mAP_{50} 和 59.8%的 mAP。这些结果证明了 SSIDet 的泛化能力及其在两个数据集上实现最佳性能的能力。

Table 1. Detection results of different methods on the FLIR dataset
表 1. 不同方法在 FLIR 数据集上的检测结果

Methods	Backbone	people	Cyclist	car	mAP50 ↑	mAP ↑	Modality
SSD [24]	VGG16	66.3	64.3	67.2	65.5	29.6	IR
RetinaNet [25]	ResNet50	67.5	65.2	66.4	66.1	31.5	IR
Cascade R-CNN [26]	ResNet50	69.5	72.3	71.6	71.0	34.7	IR
Faster R-CNN [27]	ResNet50	75.6	73.3	74.1	74.4	37.6	IR
DDQ-DETR [28]	ResNet50	72.6	74.5	73.2	73.9	37.1	IR
SSD [24]	VGG-16	51.1	52.6	52.9	52.2	21.8	RGB
RetinaNet [25]	ResNet50	50.8	52.5	51.4	51.2	21.9	RGB
Cascade R-CNN [26]	ResNet50	54.2	57.3	56.5	56.0	24.7	RGB
Faster R-CNN [27]	ResNet50	63.8	65.9	64.2	64.9	28.9	RGB
DDQ-DETR [28]	ResNet50	62.6	65.8	65.4	64.9	30.9	RGB
CAFF [29]	ResNet18	75.3	72.6	75.8	74.6	37.4	IR + RGB
probEn [30]	ResNet50	74.3	72.6	76.5	75.5	37.9	IR + RGB
LGADet [31]	ResNet50	75.3	72.6	74.2	74.5	39.6	IR + RGB
CSSA [32]	ResNet50	78.9	77.3	78.2	79.2	41.3	IR + RGB
SSIDet (ours)	CTNet	81.3	78.1	79.5	80.9	41.5	IR + RGB

Table 2. Detection results of different methods on the LLVIP dataset
表 2. 不同方法在 LLVIP 数据集上的检测结果

Methods	Backbone	mAP50 ↑	mAP ↑	Modality
SSD	VGG16	90.2	53.5	IR
RetinaNet	ResNet50	94.8	55.1	IR
Cascade R-CNN	ResNet50	95.0	56.8	IR
Faster R-CNN	ResNet50	94.6	54.5	IR
DDQ-DETR	ResNet50	93.9	58.6	IR

续表

SSD	VGG-16	82.6	39.8	RGB
RetinaNet	ResNet50	88.0	42.8	RGB
Cascade R-CNN	ResNet50	88.3	47.0	RGB
Faster R-CNN	ResNet50	87.0	45.1	RGB
DDQ-DETR	ResNet50	86.1	46.7	RGB
CAFF	ResNet18	94.0	55.8	IR + RGB
probEn	ResNet50	93.4	51.5	IR + RGB
LGADet	ResNet50	93.6	52.3	IR + RGB
CSSA	ResNet50	94.3	59.2	IR + RGB
SSIDet (ours)	CTNet	95.5	59.8	IR + RGB

3.3. 消融实验

模块消融实验：为验证本文提出的空间-光谱交互模块和多尺度重建网络的有效性，在这两个模块上进行消融实验。结果如表 3 所示，只添加空间－光谱交互模块时，FLIR 数据集上的 mAP 提高了 1.7%，LLVIP 数据集上的 mAP 提高了 1.6%；只添加多尺度重建网络时，FLIR 数据集上的 mAP 提高了 3.0%，LLVIP 数据集上的 mAP 提高了 2.1%；进一步将两个模块都添加时，mAP 达到了最佳水平。以上结果证明了模块的有效性。

Table 3. Ablation study on each module result on the FLIR and LLVIP dataset
表 3. 各模块在 FLIR 和 LLVIP 数据集上的消融实验结果

空间－光谱交互模块	多尺度重建网络	FLIR		LLVIP	
		mAP50 ↑	mAP ↑	mAP50 ↑	mAP ↑
×	×	74.9	36.9	94.3	55.2
√	×	76.5	38.6	94.6	56.8
×	√	78.2	39.9	95.0	57.3
√	√	80.9	41.5	95.5	59.8

CT 编码器数量消融实验：为研究不同的 CT 编码器数量对于检测结果的影响，在 FLIR 和 LLVIP 数据集进行消融实验。本节设置多尺度 CT 编码网络中的 CT 编码器数量分别为 2，3，4。不同数量 CT 编码器的实验结果如表 4 所示。从表 4 可知，3 个 CT 编码器时的检测性能最佳，而 CT 编码器为 2 和 4 时检测精度都有一定损失。具体来说，当 CT 编码器数量为 2 时，模型的特征表征能力会受到网络深度限制，导致图像特征融合时的误差较大。当 CT 编码器数量为 4 时，编码器提取的特征图大小为 32 × 32，提取的空间信息和光谱信息受到限制。因此，本文将 CT 编码器数量设置为 3。

3.4. 定性分析

FLIR 和 LLVIP 数据集：图 5 为 SSIDet 在 FLIR 和 LLVIP 数据集测试集上的目标检测可视化结果。如图所示，SSIDet 的检测结果几乎包括了 Ground Truth 标签的所有边界框，并且一致性很高。此外，本文的方法还检测出一些 Ground Truth 中遗漏的小目标，这表明了 SSIDet 在多尺度目标检测中的优越性。

Table 4. Ablation study on number of CT encoders result on the FLIR and LLVIP dataset
表 4. CT 编码器数量在 FLIR 和 LLVIP 数据集上的消融实验结果

CT 编码器数量	FLIR		LLVIP	
	mAP50 ↑	mAP ↑	mAP50 ↑	mAP ↑
2	75.2	37.7	94.5	56.3
3 (ours)	80.9	41.5	95.5	59.8
4	77.5	39.2	94.8	57.9

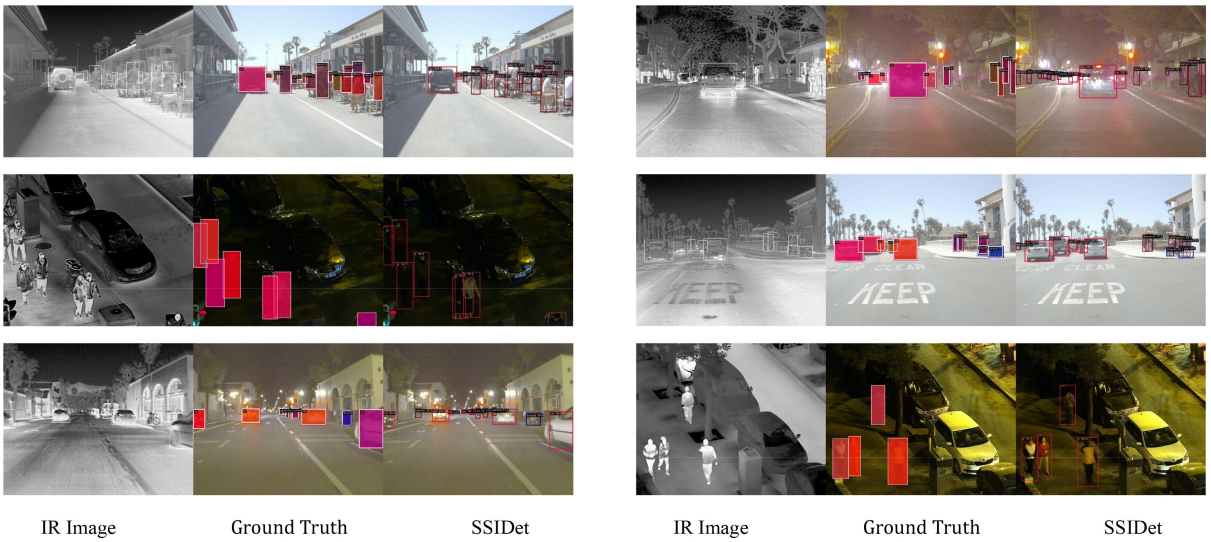

Figure 5. Visualization of detection results for FLIR and LLVIP datasets
图 5. FLIR 和 LLVIP 数据集检测结果可视化

特征融合结果可视化：图 6 为 FLIR 数据集的特征融合结果可视化。通过对比融合前后的特征，可以发现经空间－光谱交互注意力模块和多尺度重建网络处理后，原本图像中不显著的目标变得显著，并且本文方法的融合特征相对于其他单个特征更加突出。

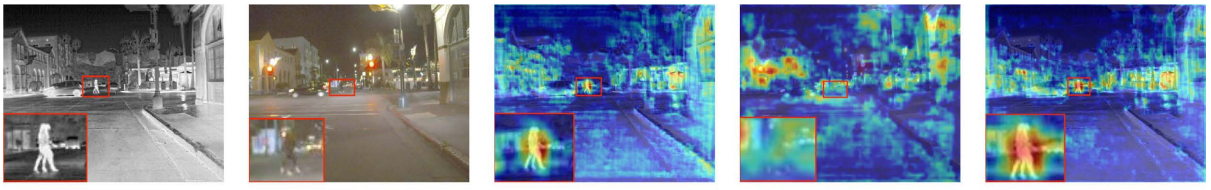

Figure 6. Visualization of feature fusion results
图 6. 特征融合结果可视化

4. 结论

本文提出了一种基于多尺度空谱特征交互的目标检测框架。首先采用双分支特征编码架构，通过级联的 CT 编码器捕获可见光图像和热红外图像的局部细节和全局上下文表征。然后，构建了空间－光谱交互注意力网络，该模块实现了双模态特征的跨域交互，有效减少了特征冗余并强化了互补特性。最终

通过多尺度重建网络进行空谱信息的渐进式融合，有效地重建了高分辨率多光谱图像，实现了空间特征与光谱特征的协同增强。在 FLIR 和 LLVIP 数据集上的实验表明，本文方法可以有效进行特征融合并具有最好的目标检测性能。然而，基于 Transformer 和 CNN 结合的主干网络依然需要较大的计算量。因此，可以进一步研究如何使模型在减少更多计算量的同时进一步提高多模态目标检测的精度。

基金项目

专项名称：基础科研条件与重大科学仪器设备研发；所属项目名称：全光纤非线性单光子显微光谱仪；所属课题编号：2022YFF0706003；所属课题名称：全光纤非线性单光子显微光谱仪研发。

参考文献

[1] Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M., Qi, H., *et al*. (2020) UA-DETRAC: A New Benchmark and Protocol for Multi-Object Detection and Tracking. *Computer Vision and Image Understanding*, **193**, Article 102907. https://doi.org/10.1016/j.cviu.2020.102907

[2] Nascimento, J.C. and Marques, J.S. (2006) Performance Evaluation of Object Detection Algorithms for Video Surveillance. *IEEE Transactions on Multimedia*, **8**, 761-774. https://doi.org/10.1109/tmm.2006.876287

[3] Li, B., Xie, X., Wei, X. and Tang, W. (2021) Ship Detection and Classification from Optical Remote Sensing Images: A Survey. *Chinese Journal of Aeronautics*, **34**, 145-163. https://doi.org/10.1016/j.cja.2020.09.022

[4] Xia, G., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., *et al*. (2018) DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 3974-3983. https://doi.org/10.1109/cvpr.2018.00418

[5] Yan, H., Li, B., Zhang, H. and Wei, X. (2022) An Antijamming and Lightweight Ship Detector Designed for Spaceborne Optical Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **15**, 4468-4481. https://doi.org/10.1109/jstars.2022.3179612

[6] Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J. and Lu, J. (2023) SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 21672-21683. https://doi.org/10.1109/iccv51070.2023.01986

[7] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., *et al*. (2020) BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 2633-2642. https://doi.org/10.1109/cvpr42600.2020.00271

[8] Wei, X. and Zhao, S. (2024) Boosting Adversarial Transferability with Learnable Patch-Wise Masks. *IEEE Transactions on Multimedia*, **26**, 3778-3787. https://doi.org/10.1109/tmm.2023.3315550

[9] Kim, J.U., Park, S. and Ro, Y.M. (2022) Uncertainty-Guided Cross-Modal Learning for Robust Multispectral Pedestrian Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, **32**, 1510-1523. https://doi.org/10.1109/tcsvt.2021.3076466

[10] Song, S., Miao, Z., Yu, H., Fang, J., Zheng, K., Ma, C., *et al*. (2022) Deep Domain Adaptation Based Multi-Spectral Salient Object Detection. *IEEE Transactions on Multimedia*, **24**, 128-140. https://doi.org/10.1109/tmm.2020.3046868

[11] Xie, Z., Shao, F., Chen, G., Chen, H., Jiang, Q., Meng, X., *et al*. (2023) Cross-Modality Double Bidirectional Interaction and Fusion Network for RGB-T Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, **33**, 4149-4163. https://doi.org/10.1109/tcsvt.2023.3241196

[12] Wang, K., Tu, Z., Li, C., Zhang, C. and Luo, B. (2024) Learning Adaptive Fusion Bank for Multi-Modal Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, **34**, 7344-7358. https://doi.org/10.1109/tcsvt.2024.3375505

[13] Liu, J., Zhang, S., Wang, S. and Metaxas, D. (2016) Multispectral Deep Neural Networks for Pedestrian Detection. *Procedings of the British Machine Vision Conference* 2016, New York, 19-22 September 2016, 1-13. https://doi.org/10.5244/c.30.73

[14] Li, C. Song, D. Tong, R. and Tang, M. (2018) Multispectral Pedestrian Detection via Simultaneous Detection and Segmentation. *British Machine Vision Conference* (*BMVC*) 2018, Newcastle, 3-6 September 2018, 225.

[15] Cao, Y., Guan, D., Wu, Y., Yang, J., Cao, Y. and Yang, M.Y. (2019) Box-Level Segmentation Supervised Deep Neural Networks for Accurate and Real-Time Multispectral Pedestrian Detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, **150**, 70-79. https://doi.org/10.1016/j.isprsjprs.2019.02.005

[16] Zhou, K., Chen, L. and Cao, X. (2020) Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems. *Computer Vision—ECCV* 2020, Glasgow, 23-28 August 2020, 787-803. https://doi.org/10.1007/978-3-030-58523-5_46

[17] Liu, Q., Zhou, H., Xu, Q., Liu, X. and Wang, Y. (2021) PSGAN: A Generative Adversarial Network for Remote Sensing Image Pan-Sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, **59**, 10227-10242. https://doi.org/10.1109/tgrs.2020.3042974

[18] Diao, W., Zhang, F., Sun, J., Xing, Y., Zhang, K. and Bruzzone, L. (2023) ZerGAN: Zero-Reference GAN for Fusion of Multispectral and Panchromatic Images. *IEEE Transactions on Neural Networks and Learning Systems*, **34**, 8195-8209. https://doi.org/10.1109/tnnls.2021.3137373

[19] Lee, W., Jovanov, L. and Philips, W. (2023) Cross-modality Attention and Multimodal Fusion Transformer for Pedestrian Detection. *Computer Vision—ECCV* 2022 *Workshops*, Tel Aviv, 23-27 October 2022, 608-623. https://doi.org/10.1007/978-3-031-25072-9_41

[20] Fang, Q., Han, D. and Wang, Z. (2022) Cross-Modality Fusion Transformer for Multispectral Object Detection. arXiv: 2111.00273. https://doi.org/10.48550/arXiv.2111.00273

[21] You, S., Xie, X., Feng, Y., Mei, C. and Ji, Y. (2023) Multi-Scale Aggregation Transformers for Multispectral Object Detection. *IEEE Signal Processing Letters*, **30**, 1172-1176. https://doi.org/10.1109/lsp.2023.3309578

[22] Zhang, H., Fromont, E., Lefevre, S. and Avignon, B. (2020) Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks. 2020 *IEEE International Conference on Image Processing* (*ICIP*), Abu Dhabi, 25-28 October 2020, 276-280. https://doi.org/10.1109/icip40778.2020.9191080

[23] Jia, X., Zhu, C., Li, M., Tang, W. and Zhou, W. (2021) LLVIP: A Visible-Infrared Paired Dataset for Low-Light Vision. 2021 *IEEE/CVF International Conference on Computer Vision Workshops* (*ICCVW*), Montreal, 11-17 October 2021, 3489-3497. https://doi.org/10.1109/iccvw54120.2021.00389

[24] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., *et al*. (2016) SSD: Single Shot Multibox Detector. *Computer Vision—ECCV* 2016, Amsterdam, 11-14 October 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[25] Lin, T., Goyal, P., Girshick, R., He, K. and Dollar, P. (2017) Focal Loss for Dense Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2999-3007. https://doi.org/10.1109/iccv.2017.324

[26] Cai, Z. and Vasconcelos, N. (2021) Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 1483-1498. https://doi.org/10.1109/tpami.2019.2956516

[27] Ren, S., He, K., Girshick, R. and Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/tpami.2016.2577031

[28] Zhang, S., Wang, X., Wang, J., Pang, J., Lyu, C., Zhang, W., *et al*. (2023) Dense Distinct Query for End-to-End Object Detection. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 7329-7338. https://doi.org/10.1109/cvpr52729.2023.00708

[29] Zhang, H., Fromont, E., Lefevre, S. and Avignon, B. (2021) Guided Attentive Feature Fusion for Multispectral Pedestrian Detection. 2021 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-8 January 2021, 72-80. https://doi.org/10.1109/wacv48630.2021.00012

[30] Chen, Y., Shi, J., Ye, Z., Mertz, C., Ramanan, D. and Kong, S. (2022) Multimodal Object Detection via Probabilistic Ensembling. *Computer Vision—ECCV* 2022, Tel Aviv, 23-27 October 2022, 139-158. https://doi.org/10.1007/978-3-031-20077-9_9

[31] Zuo, X., Wang, Z., Liu, Y., Shen, J. and Wang, H. (2022) LGADet: Light-Weight Anchor-Free Multispectral Pedestrian Detection with Mixed Local and Global Attention. *Neural Processing Letters*, **55**, 2935-2952. https://doi.org/10.1007/s11063-022-10991-7

[32] Cao, Y., Bin, J., Hamari, J., Blasch, E. and Liu, Z. (2023) Multimodal Object Detection by Channel Switching and Spatial Attention. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Vancouver, 17-24 June 2023, 403-411. https://doi.org/10.1109/cvprw59228.2023.00046