基于增强特征对比度的多光谱融合目标检测

摘要

行人检测在计算机视觉中起着至关重要的作用，因为它有助于确保交通安全。由于缺乏有用的信息，仅依赖 RGB 图像的现有方法在弱光条件下的性能会下降。为了解决这个问题，最近的多光谱检测方法结合了热图像以提供互补信息，并获得了更好的性能。然而，很少有方法关注由嘈杂的融合特征图引起假阳性的负面影响。与它们不同的是，我们综合分析了假阳性对检测性能的影响，发现增强特征对比度可以显著降低这些假阳性。在本文中，我们提出了一种基于增强特征对比度的多光谱融合目标检测，名为 EFCDet。这篇论文采用特征融合-特征优化模式。在特征融合阶段，我们揭示了 RGB 图像和热图像的平行通道相似性和跨通道相似性，并学习了自适应感受野以从这两个特征中收集有用的信息。在特征优化阶段，我们使用分割 (Seg) 分支来区分行人特征和背景特征，而且我们提出了一个相关-最大损失函数来增强行人特征和背景特征之间的对比度。因此，多光谱目标检测的性能得到了显著提高。在多个数据集上进行的广泛实验表明，这篇论文所提方案是有效的，并获得了最先进的检测性能。

关键词：多光谱目标检测；跨模态；特征融合；特征增强

1 引言

行人检测是计算机视觉中的一个关键问题，具有自动驾驶和视频监控等各种应用。现代研究主要依赖于 RGB 图像，并且在有利的照明条件下表现良好。由于此类场景中 RGB 图像的信噪比较低，因此它们在弱光条件下的性能会下降。相比之下，即使在光线不足的条件下，热图像也可以清晰地捕捉人体的形状。热图像无法记录颜色和纹理细节，从而限制了它们区分令人困惑的结构的能力。为了提高不同照明条件下的检测性能，多光谱行人检测成为一种很有前途的解决方案 [1]，[2]。它使用 RGB 和热图像的互补信息来检测行人。这种互补性如图 1 的左列所示。当同时使用 RGB 和热图像时，我们可以更轻松地定位行人（由绿色箭头标记）。

最近的工作表明，多光谱特征的组合提高了单模态行人检测的准确性 [3]–[9]。此外，精度的提高在很大程度上取决于融合阶段和所采用的策略。以前的研究表明，中途融合方法的性能优于早期和晚期融合方法 [5]、[8]、[10]、[11]。术语“早期”、“中途”和“晚期”融合分别是指双分支网络的低、中、高阶段多光谱信息的融合。基于这一发现，最近的工作设计了复杂的中途融合策略来应对各种挑战，如错位 [8]、[12]、[13]、模态不平衡 [9]、[14] 和集成学习 [15] 问题。然而，这些作品中的一个问题是产生嘈杂的融合特征。造成这种现象的原因是，以前的工作主要集中在结合两种模态的互补特征，而忽略了区分目标和非目标特征。我们观察到，这些噪声特征可能会导致大量假阳性（FP）并降低性能，如图 1 （a）所示。

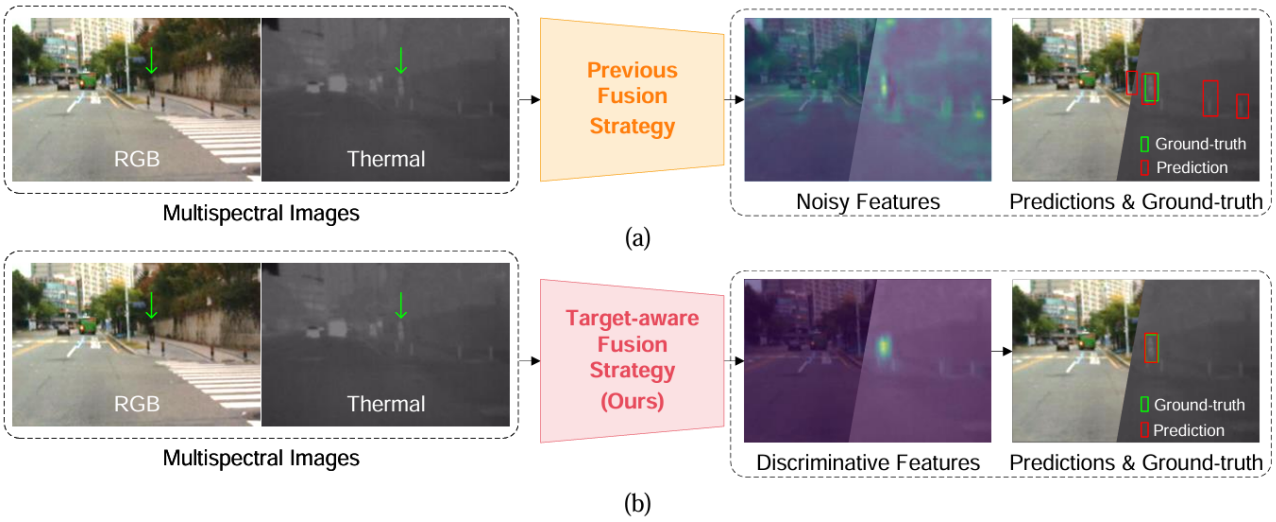

图 1. 使用不同融合策略的一对多光谱图像生成的特征和检测结果的可视化。(a) 以前的融合策略会产生噪声特征，并在背景区域诱发许多假阳性。(b) 我们的增强特征对比度的策略能产生判别特征并有效减少假阳性。左列中的绿色箭头标记行人的位置。右列中的绿色框和红色框分别表示真实边界框和预测框。

为了减轻噪声特征的不利影响，在这项工作中，我们提出了一种基于增强特征对比度的多光谱融合目标检测。在该目标检测框架内，我们引入了一种新的多光谱特征融合模块，该模块利用了成对多光谱特征中固有的平行通道相似性和跨通道相似性。此外，我们提出了一个特征优化模块，其中包括一个特征分割单元和特征对比度增强单元。此外，我们提出了一种新的相关-最大损失函数。这些增强功能不仅使模型能够组合互补信息，还可以区分目标和非目标特征，并改进目标区域中的表示，同时抑制非目标区域中的表示。受益于这些模块，我们的增强特征对比度的策略使模型能够有效地关注目标区域，从而减少背景区域的假阳性 (FP)，如图 1 (b) 所示。

我们的基于增强特征对比度的策略非常灵活。它可以应用于单级和两级探测器，并且可以轻松扩展到涉及多个类别的多光谱目标检测场景，例如行人、汽车和其他目标。EFCDet 在两个多光谱行人检测基准 [1]、[2] 以及两个多光谱目标检测基准 [16]、[17] 上实现了最先进的性能。更重要的是，在弱光条件下，它比以前的方法具有明显的优势。此外，我们的融合策略计算效率很高，使 TFDet 能够在交通场景中更快地生成预测。

2 相关工作

2.1 目标检测与行人检测

根据目标探测器是否需要目标建议，目标探测器可以分为两组：单级探测器 [18]、[19] 和两级探测器 [20]、[21]。一级探测器直接预测对象框，通过使用锚框 [23] 或潜在对象中心网格 [24]。YOLO [18] 是一种流行的单阶段检测器，它以全尺寸图像作为输入，同时输出对象框和特定于类的置信度分数。尽管一级检测器往往很快，但它们的精度在某种程度上是有限的。相比之下，两级探测器将物体检测定义为“从粗到细”的过程。他们根据区域建议预测对象框 [20]。在第一阶段，检测器生成多个对象建议，每个建议指示图像区域中的潜在对象。在第二

阶段，检测器进一步优化位置并预测这些提议的类分数。Faster R-CNN [20] 是一种众所周知的两级探测器。它由两部分组成：RPN [20] 和 R-CNN [25]。RPN 组件经过端到端训练，可生成高质量的区域建议，然后由 R-CNN 用于对象检测。尽管两级检测器具有很高的精度，但它们往往速度较慢。

行人检测是物体检测的一个重要应用。随着目标检测技术的进步，行人探测器也可以分为一级和两级探测器 [26]、[27]。但是，这些行人探测器仅依赖 RGB 图像，这会导致它们在弱光条件下的性能下降。这个问题的一个有前途的解决方案是使用多光谱图像。

2.2 多光谱行人检测

多光谱行人检测使用 RGB 和热图像来定位行人 [1]。该技术的一个重要研究方向是如何融合 RGB 和热图像，以便探测器能够在各种照明条件下实现稳定检测。为了实现自适应特征融合，最近的工作开发了照明感知特征融合方法 [28]–[32]、基于注意力的特征融合方法 [33]–[38] 和非局部特征融合方法 [6]。照明感知特征融合方法通常会引入一个分类分支，以根据照明条件确定 RGB 特征的重要性。考虑到图像的分类结果不能反映单个区域的重要性，基于注意力的方法使用空间注意力、通道注意力或 transformer 的交叉注意力来辅助特征融合 [11]、[13]、[33]–[35]、[38]。空间注意力和通道注意力可以为多光谱特征生成元素和通道加权因子，而交叉注意力可以模拟全局上下文相关性。交叉注意力机制能够解决多光谱特征之间的错位问题，但它的计算成本很高。这个问题可以通过非局部特征聚合来缓解 [6]。

与以前的方法不同，我们使用了一种改进的可变形卷积，以显式模拟 RGB 和热特征之间的偏移量。我们还揭示了多光谱特征中固有的平行通道相似性和跨通道相似性，并提出了一种自适应特征融合的两步法。随着弱监督学习在目标检测中的成功 [39]，最近的多光谱行人检测方法采用箱级掩码来提高检测性能 [3]、[4]、[34]、[40]。盒级掩码是一个二进制掩码，行人区域用 1 填充，其他区域用 0 填充。MSDS-RCNN [4] 引入了一个分割分支，并利用二进制掩码来监督其训练。GAFF [40] 使用盒级掩码来引导模态间和模态内的注意力。由于这些方法不关注来自行人的全局信息，LG-FAPF [3] 和 M2FNet [34] 利用箱级掩码和 transformer 的交叉注意力来增强全局建模能力。我们注意到，以前的方法忽略了嘈杂特征图对检测性能的影响，而嘈杂的融合特征可能会导致背景区域出现假阳性。在这项工作中，我们关注假阳性对检测性能的负面影响，并使用箱级掩码来增强行人和背景区域之间的特征对比度。

3 本文方法

3.1 本文方法概述

我们的 EFCDet 的整体架构如图 2 所示。我们将一对 RGB 和热图像送到 CNN 主干网络中，以生成相应的多光谱特征，表示为 F_{rgb} 和 $F_{thermal}$。随后，利用多光谱特征中固有的平行和跨通道相似性在颈部部分融合多光谱特征。在这种融合策略中，特征融合模块 (FFM) 首先自适应地组合多光谱特征，生成表示为 F_x 的初始融合特征。基于这个初始融合特征，然后特征优化模块 (FRM) 通过采用分割 (seg) 分支来区分目标和背景特征。之后，特征优化模块 (FRM) 通过相关-最大损失函数增强特征对比度。我们的特征优化模块最终产生判别特征 F_y，随后将其输入检测头。这些检测头可以是一级头（如 YOLO 头 [23]）或两级头（如 Faster

R-CNN 头 [20]，包括 RPN 和 R-CNN），此过程会预测框和相应的分数。

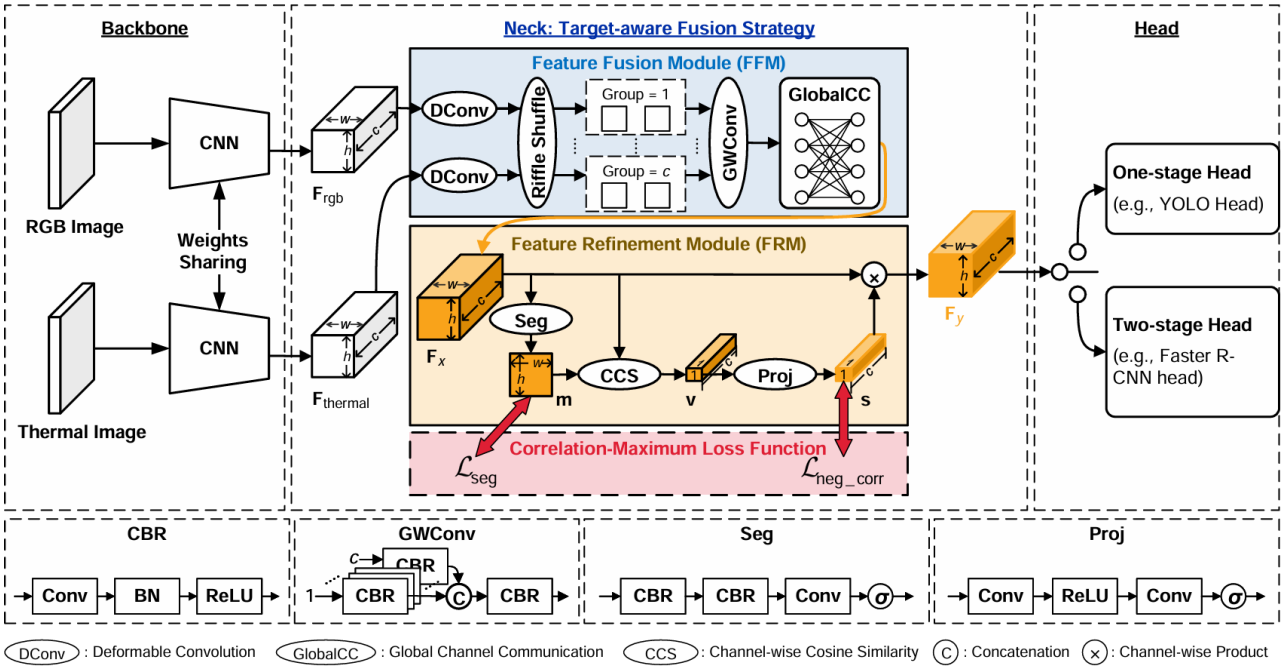

图 2. 我们的 TFDet 架构图示。它由三个部分组成：骨干网络 (Backbone)、颈部网络 (Neck) 和头部网络 (Head)。主干从成对的多光谱图像中提取特征。颈部使用通道相似性和增强特征对比度的策略融合这些多光谱特征。head 根据融合特征生成框和相应的分数。

3.2 特征融合模块 (FFM)

我们提出的特征融合模块 (FFM) 从多光谱特征的相似性中获得见解。具体来说，给定多光谱特征 $F_{rgb}\in R^{h\times w\times c}$ 和 $F_{thermal}\in R^{h\times w\times c}$，每个特征都包含分辨率为 h × w 的 c 个通道图，我们穷举计算所有通道图中的矩阵内积，得到大小为 c × c 的多光谱特征关系矩阵。这个过程如图 3 (a)所示，其中我们表示特征 F_{rgb} 和 $F_{thermal}$ 为 $\{f_v^i|f_v^i\in \mathrm{R}^{h\times w}, i\in[1,c]\}$ $\{\mathrm{f}_t^i|f_t^i\in \mathrm{R}^{h\times w}, i\in[1,c]\}$

从图 3 （b）中，我们可以观察到两个启发性的现象：（1）沿对角线的元素具有较大的值，表明同一通道位置的多光谱特征之间存在很强的相似性，我们称之为“平行通道相似性”，以及（2）在非对角线区域也有较大的元素，表明不同通道的特征也可能表现出很强的相似性，我们称之为“跨通道相似性”。

为了全面验证多光谱特征中的相似性是否普遍，我们定义了两个指标：平均平均值比（ANR）和平均中位数比（AIR）。这些量度旨在总结对角线元素和非对角线元素之间的差异。ANR 计算整个数据集中对角线位置元素与非对角线位置元素的平均值之间的平均比率。同时，AIR 采用类似的方法，但使用非对角线元素的中值作为分母。数学过程可以表述为

$$\begin{aligned}\mathbf{NR}_n &= \frac{1}{c}\sum_{i=1}^{c}\frac{\mathbf{f}_v^i\cdot\mathbf{f}_t^i}{\mathrm{Mean}\left(\left\{\mathbf{f}_v^i\cdot\mathbf{f}_t^j|j\in[1,c]\backslash\{i\}\right\}\right)}\\ \mathbf{IR}_n &= \frac{1}{c}\sum_{i=1}^{c}\frac{\mathbf{f}_v^i\cdot\mathbf{f}_t^i}{\mathrm{Median}\left(\left\{\mathbf{f}_v^i\cdot\mathbf{f}_t^j|j\in[1,c]\backslash\{i\}\right\}\right)}\end{aligned} \tag{1}$$

和

$$\mathrm{ANR}=\frac{1}{N}\sum_{n=1}^{N}\mathrm{NR}_n, \mathrm{AIR}=\frac{1}{N}\sum_{n=1}^{N}\mathrm{IR}_n \tag{2}$$

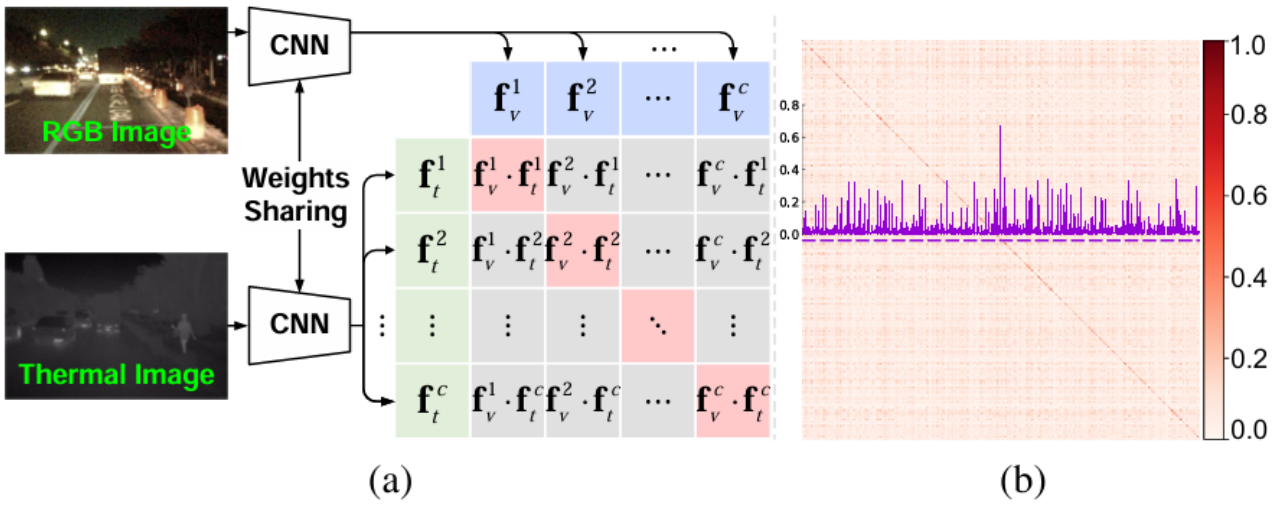

图 3. 成对多光谱图像的特征关系矩阵。(a) 计算关系矩阵的数学过程。(b) 生成的关系矩阵。在 (b) 中，紫色虚线标记了矩阵的中间行，该行的一维分布以紫色显示。

其中 N 表示数据集中成对的 RGB-T 图像的数量，n 表示图像对的索引。对于每个图像对，$f_v^i \cdot f_t^i$ 计算通道位置 i 处多光谱特征的矩阵内积。运算符 Mean (·) 和 Median (·) 分别返回给定集的平均值和中位数。我们选择平均值和中位数作为分母有两个原因：(1) 平均值反映了非对角线位置元素的总体预期，以及 (2) 中位数避免了异常值的影响。

基于上述定义，我们在两个代表性数据集上计算 ANR 和 AIR：FLIR [16] 和 LLVIP [2]。表 I 中提供的结果揭示了两种现象：(1) ANR 和 AIR 都超过 1，表明平行通道相似性强，以及 (2) ANR 始终倾向于小于 AIR，这意味着跨通道相似性强。

表 1. Comparison of Datasets

Dataset	FLIR [1]		LLVIP [2]	
# Channels	512	1024	512	1024
ANR	3.52	1.60	1.48	1.46
AIR	8.65	2.95	1.76	1.97

基于多光谱特征中的平行通道相似性和跨通道相似性，我们提出了一种融合这些特征的两步法，如图 4 所示。在第一步中，我们利用平行通道相似性的性质来融合两个相应的通道特征。考虑到行人在距离摄像头的不同距离上的不同尺度，我们采用可变形卷积 (DConv) [45] 来自适应地学习每个输出响应的感受野。然而，由于 DConv 最初是为单模态输入而设计的，因此我们提出了 Riffle shuffle 和分组卷积 (GWConv)，Riffle Shuffle 将两个多光谱特征图分组到同一通道位置，从而生成 c 组特征，GWConv 使用分组卷积融合通道的局部特征。在第二步中，我们利用跨通道相似性的特性来重新校准通道级特征。为此，我们引入了一个全局通道通信 (GlobalCC) 模块，它显式构建了通道之间的相互依赖关系，从而生成一组调制权重。然后将这些权重应用于步骤 I 中获得的特征，产生初始融合特征 F_x。

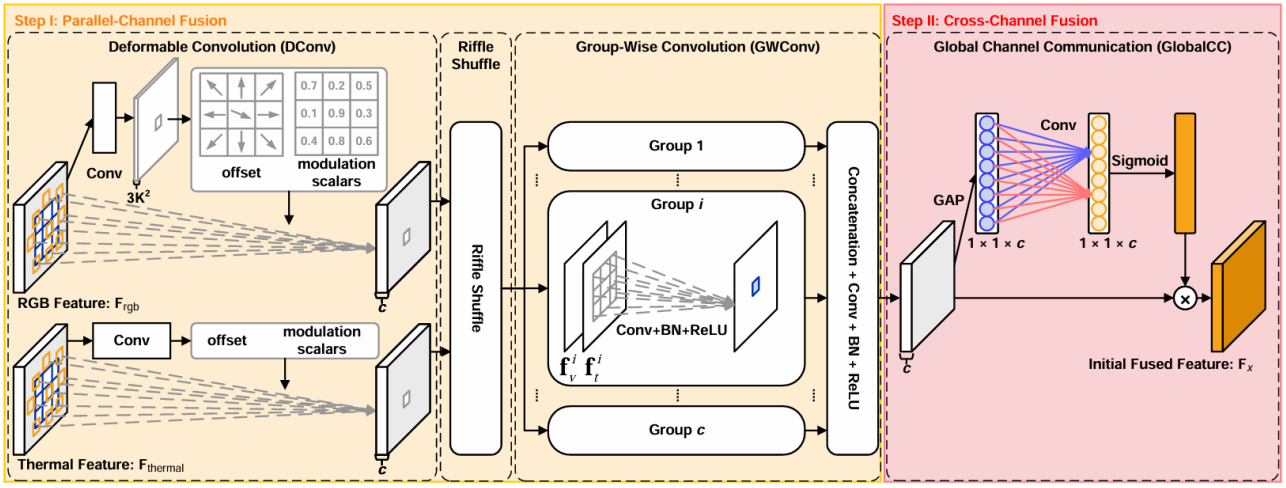

图 4. 特征融合模块（FFM）的图示。该模块分两步实现：平行通道融合和跨通道融合。它由四个部分组成。DConv 自适应地学习感受野。Riffle Shuffle 将两个多光谱特征图分组到同一通道位置，从而生成 c 组特征。GWConv 使用分组卷积融合通道的局部特征。GlobalCC 使用通道注意力机制学习跨通道特征。

3.3 特征优化模块 (FRM)

我们使用 FRM 来增强目标特征和背景特征之间的对比度。FRM 采用最初融合的特征 $F_x \in R^{h\times w\times c}$ 输入，并输出对比度增强的特征 $F_y \in R^{h\times w\times c}$。具体来说，我们首先根据边界框标签生成真实箱级掩码，方法是用 1 填充目标区域，用 0 填充剩余区域。接下来，我们使用 F_x 通过分割（Seg）分支生成预测的箱级掩码。此过程定义为

$$\mathbf{m} = \sigma\left(\mathcal{H}\left(\mathbf{F}_x\right)\right) \tag{3}$$

其中函数 H (·) 表示分割分支，函数 (·) 表示 sigmoid 激活。

然后，我们计算预测的箱级掩码 m 与特征 F_x 通道方向之间的相关性 v，定义为

$$\begin{aligned}\mathbf{v} &= [v_1, v_2, \ldots, v_c] \\ &= \left[\mathbf{m}\cdot\mathbf{f}_x^1, \mathbf{m}\cdot\mathbf{f}_x^2, \ldots, \mathbf{m}\cdot\mathbf{f}_x^c\right]\end{aligned} \tag{4}$$

其中运算符 "." 表示矩阵内积。$f_x^i \in R^{h\times w}$ (1 i c) 是 F 的第 i 个通道特征。由于 m 和 f_x^i 都已归一化，因此内积可以表明它们的相似性。

此相关性由投影（Proj）层处理，以生成处理后的相关性，定义为

$$\mathbf{s} = \sigma(\mathcal{P}(\mathbf{v})) \tag{5}$$

其中函数 P (·) 表示投影层，它由两个卷积运算组成。然后，我们使用它来增强特征 F_x 生成优化后的特征

$$\mathbf{F}_y = \mathbf{s} \otimes \mathbf{F}_x \tag{6}$$

其中，操作 执行通道乘积。在整个优化过程中，我们监督预测的箱级掩码 m $\in R^{h\times w}$ 并使用我们的相关-最大损失函数来处理的相关性 s $\in R^{1\times 1\times c}$ 。我们注意到，特征优化模块 (FRM) 可以很容易地扩展到多类对象检测场景。在这样的场景下，我们用 1 填充所有目标区域，用 0 填充背景区域来构建真实的箱级掩码，然后使用上述过程来增强特征对比度。

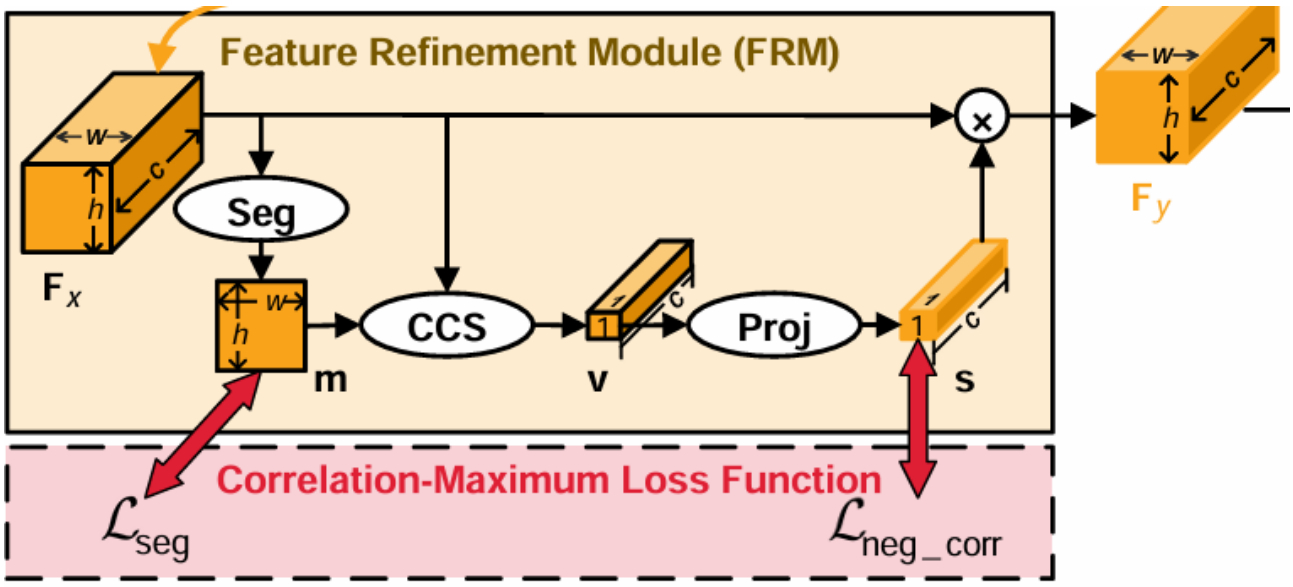

图 5. 特征优化模块

3.4 相关-最大损失函数

我们的特征对比度增强方法需要确保：(1) 预测的箱级掩码 m 尽可能准确，以及 (2) 预测的箱级掩码和初始融合特征之间的处理相关性 s 尽可能高。

为此，我们提出了一个名为相关-最大损失函数 L_{corr_max} 的损失函数来正则化模型。它定义为

$$\mathcal{L}_{\text{corr\_max}}(\tilde{\mathbf{m}}, \mathbf{m}, \mathbf{s}) = \mathcal{L}_{\text{seg}}(\tilde{\mathbf{m}}, \mathbf{m}) + \alpha \mathcal{L}_{\text{neg\_corr}}(\mathbf{s}), \tag{7}$$

其中 $\tilde{m}$ 表示真实箱级掩码，$\alpha = 0.1$。该函数由两个部分组成：分割损失函数 L_{seg} 和负相关损失函数 L_{neg_corr}。

分割损失函数 L_{seg} 目的在于最大化预测的箱级掩码的准确性。它定义为

$$\mathcal{L}_{\text{seg}}(\tilde{\mathbf{m}}, \mathbf{m}) = \mathcal{L}_{\text{bce}}(\tilde{\mathbf{m}}, \mathbf{m}) + \mathcal{L}_{\text{dice}}(\tilde{\mathbf{m}}, \mathbf{m}). \tag{8}$$

二进制交叉熵损失函数 L_{bce} 定义为

$$\mathcal{L}_{\text{bce}}(\tilde{\mathbf{m}}, \mathbf{m}) = \frac{1}{h \times w} \sum_{p=1}^{h \times w} l(\tilde{m}_p, m_p) \tag{9}$$

其中

$$l(\tilde{m}_p, m_p) = -\left[\tilde{m}_p \log m_p + (1 - \tilde{m}_p) \log (1 - m_p)\right] \tag{10}$$

损失函数 L_{dice} 定义为

$$\begin{aligned} \mathcal{L}_{\text{dice}}(\tilde{\mathbf{m}}, \mathbf{m}) &= 1 - \frac{2|\tilde{\mathbf{m}} \cap \mathbf{m}| + \epsilon}{|\tilde{\mathbf{m}}| + |\mathbf{m}| + \epsilon} \\ &= 1 - \frac{\left(2 \times \sum_{p=1}^{h \times w} \tilde{m}_p m_p\right) + \epsilon}{\left(\sum_{p-1}^{h \times w} \tilde{m}_p\right) + \left(\sum_{p-1}^{h \times w} m_p\right) + \epsilon} \end{aligned} \tag{11}$$

(14) 其中 m 和 $\widetilde{m}$ 分别表示位置 p 处的 m 和 $\widetilde{m}$ 的元素。 按照 [47] 中的方法设置为 1.0。我们使用最邻近方法对真实的箱级掩码进行下采样，以匹配预测掩码的形状。$\widetilde{m}$ m 计算它们的交集，而 $|\cdot|$ 计算掩码的大小。

负相关损失函数 L_{neg_corr} 旨在使处理后的相关性 s 中的每个元素都接近 1.0。它定义为

$$\mathcal{L}_{\text{neg\_corr}}(\mathbf{s}) = -\frac{1}{c}\sum_{i=1}^{c}\log s_i \tag{12}$$

4 复现细节

4.1 与已有开源代码对比

在本研究中，我们对现有的开源代码进行了有限的改进，主要集中在特征融合模块的优化上。我们以 Xue 等人开发的 TFDet 代码库为参考，该代码库实现了目标感知融合策略，用于 RGB-T 行人检测。以下是我们在特征融合模块方面的对比分析：

TFDet 代码库：TFDet 的特征融合模块主要通过自适应地调整感受野和通道间的相互依赖关系来实现特征的融合。这种方法在一定程度上能够有效地结合 RGB 和热图像的互补信息，但在处理复杂场景时，可能会产生一些噪声特征，导致假阳性增加.

EFCDet 代码库：在 TFDet 的基础上，我们对特征融合模块进行了优化。具体来说，我们引入了一种改进的可变形卷积（DConv）来更精确地模拟 RGB 和热特征之间的偏移量。此外，我们还增加了一个全局通道通信（GlobalCC）模块，该模块通过显式构建通道之间的相互依赖关系，进一步增强了特征的判别能力。这些优化措施有效地减少了噪声特征的产生，降低了假阳性的发生率，从而提高了检测的准确性. 通过上述优化，EFCDet 在特征融合模块的性能上得到了显著提升。尽管改进相对较小，但这些细节上的调整对于提高整体检测性能具有重要意义。我们的工作验证了在特征融合阶段进行精细优化的必要性和有效性，为后续的研究提供了有价值的参考.

4.2 数据集

LLVIP [2] 数据集是一个更具挑战性的多光谱行人检测数据集。它是在弱光条件下拍摄的，因此很难在 RGB 模式下检测到行人。该数据集在训练集中包括 12025 对对齐的 RGB-T 图像，在验证集中包括 3463 对对齐的 RGB-T 图像，每张图像的分辨率为 1024 × 1280。

FLIR [16] 数据集是多光谱物体检测的基准。最近的工作 [63] 更新了这个数据集，因为原始版本包含未对齐的图像对。我们使用遵循先前工作 [31]、[64] 的更新、对齐版本进行公平的比较。对齐版本由三个类别组成：'person'、'car' 和'bicycle'。它有 4 个训练集中有 129 对 RGB 和热图像，验证集中有 1013 对 RGB 和热图像，每张图像的分辨率为 512 × 640。

M3FD [17] 数据集是一个多光谱目标检测基准。它由六类组成：“人”、“汽车”、“公共汽车”、“摩托车”、“灯” 和 “卡车”。这个数据集没有提供官方的数据分割，因此我们使用了最近工作中介绍的数据分割 [65]。训练集和验证集分别由 2,905 和 1,295 个图像对组成。由于该数据集中的图像分辨率不同，我们调整长边的大小并填充较短的边，以获得分辨率为 640 × 640 的图像。

4.3 实验细节

在我们的实验中，我们采用 Faster R-CNN [20] 和 YOLOv5 [23] 分别作为两级和一级检测器。我们使用 MMDetection [66] 工具箱实现 Faster R-CNN，并使用其官方存储库 [23] 实现 YOLOv5。除非另有说明，否则所有实验均在两个 GTX 3090 GPU 上运行。对于 Faster R-CNN，我们训练了 12 个 epoch，总批次大小为 6，初始学习率为 0.01。学习率在第 8 个和第 11 个阶段衰减了 0.1 倍。除非另有说明，否则我们使用 SGD 作为优化器，并使用随机水平翻转作为数据增强技术。对于 YOLOv5，我们使用官方配置文件中指定的超参数。此外，为了确保与现有方法的公平比较，我们采用了与先前工作相同的检测器、骨干网络和评估指标。

对于 LLVIP 数据集，我们使用带有 ResNet50 的 Faster R-CNN[68] 和 YOLOv5-Large [23] 作为基线检测器，遵循 [2]、[6]、[69] 中的做法。我们注意到 YOLOv5 训练了 100 个 epoch。使用平均精度指标评估模型性能，其中 AP50 表示 IoU = 0.50 时的平均精度。AP 的计算方法是对 10 个 IoU 阈值进行平均，范围从 0.50 到 0.95，间隔为 0.05。AP 越高，检测性能越好。

对于 FLIR 数据集，我们使用 Faster R-CNN [20] 作为基线检测器，使用 AP 作为评估指标。我们使用 Swin-Transformer-Tiny [70] 作为主干网络，使用 AdamW 作为优化器。我们将初始学习率设置为 10^{-4}，并使用随机水平翻转和随机裁剪作为数据增强技术，遵循以前最先进工作的实现 [31]。

对于 M3FD 数据集，我们使用 YOLOv5-Small [23] 作为基线检测器，使用 AP 作为评估指标。我们训练 YOLOv5-Small 36 个 epoch。我们使用相同的数据拆分重新实施所有比较方法，以进行公平的比较。对于基于图像融合的方法，我们首先根据其代码生成融合图像，然后使用与我们的方法相同的设置来训练 YOLOv5-Small。

4.4 创新点

本文介绍了一种创新的方法—基于增强特征对比度的多光谱融合目标检测（EFCDet），该方法能够有效地利用 RGB 图像和热图像的互补信息，以提高在各种光照条件下的检测性能。通过提出的特征融合-特征优化模式，不仅增强了特征的判别能力，还显著降低了假阳性的发生率。

在实验部分，本文在多个具有代表性的多光谱检测数据集上进行了广泛的实验验证，包括 LLVIP、FLIR 和 M3FD 数据集。实验结果表明，EFCDet 在这些数据集上均取得了最先进的性能，特别是在弱光条件下的表现尤为突出。此外，本文还对 EFCDet 的不同组件进行了消融研究，进一步证明了所提出方法的有效性和各个模块对性能提升的贡献。

5 实验结果分析

5.1 在 LLVIP 数据集上的比较

在表 2 中，我们在 LLVIP 数据集上评估了 EFCDet，并将其与以前最先进的方法（包括单模态和多光谱检测方法）进行了比较。我们使用 Faster R-CNN 和 YOLOv5 作为基线检测器来构建 EFCDet。结果表明，使用 YOLOv5 的 EFCDet （EFCDet-YOLOv5）实现了最佳

性能。具体来说，EFCDet-YOLOv5 的性能比最好的单模态方法高 4.1%AP, 比最好的多光谱方法 LRAF-Net [10] 高 4.8% AP。有趣的是，使用热图像的 YOLOv5 在以前的方法中表现出良好的性能。这是因为 LLVIP 数据集中的大多数图像都是在低光场景下捕获的，其中热图像足以检测该数据集中的大多数行人。尽管如此，我们的 TFDet-YOLOv5 大大超过了以前的方法，受益假阳性 (FP) 的显著减少。

表 2. LLVIP 数据集上 AP （%）的比较。最佳结果以粗体突出显示并标记为红色，而次佳结果则以绿色突出显示并标记。

Method	Publication Year	AP(↑)	P50 (↑)
RGB-Based Detection Approaches			
IEGOD [71]	TNNLS 2023	-	87.6
DeformableDETR [49]	ICLR 2021	45.5	88.7
FasterRCNN [20]	NeurIPS 2015	49.2	90.1
DINO [50]	ICLR 2023	52.3	90.5
YOLOv5 [23]	version 7.0	52.7	90.8
Thermal-Based Detection Approaches			
DINO [50]	ICLR 2023	51.3	90.2
HalluciDet [72]	WACV 2024	57.8	90.1
FasterRCNN [20]	NeurIPS 2015	58.3	94.3
DeformableDETR [49]	ICLR 2021	61.9	96.1
TIRDet [33]	ACM MM 2023	64.2	96.3
YOLOv5 [23]	version 7.0	67.0	96.5
Multispectral Detection Approaches			
PoolFuser [73]	AAAI 2023	38.4	80.3
DetFusion [69]	ACM MM 2022	-	80.7
ProbEn [15]	ECCV 2022	51.5	93.4
LENFusion [75]	TIM 2024	53.0	81.6
DM-Fusion [76]	TNNLS 2023	53.1	88.1
DDFM [81]	ICCV 2023	58.0	91.5
DCMNet [6]	ACM MM 2022	58.4	-
CSAA [82]	CVPR 2023	59.2	94.3
YOLO-Adaptor [62]	TIV 2024	-	96.5
Fusion-Mamba [64]	arXiv 2024	62.8	96.8
CALNet [84]	ACM MM 2023	63.4	-
LRAF-Net [10]	TNNLS 2023	66.3	97.9
EFCDet-FasterRCNN (Ours)	-	59.4	96.0
EFCDet-YOLOv5 (Ours)	-	71.1	97.9

5.2 扩展到多类对象检测方案

我们在两个多光谱目标检测数据集 FLIR [16] 和 M3FD [17] 上进行了实验，以评估 EFCDet 对多类目标检测的效果。我们还将 EFCDet 与单模态和多光谱检测方法进行了比较。表 3 和表 4 分别显示了 FLIR 和 M3FD 数据集上的实验结果。

表 3. FLIR 数据集上 AP （%）的比较。最佳结果以粗体突出显示并标记为红色，而次佳结果则以绿色突出显示并标记。

Method	**Publication Year**	**AP** (↑)	**AP50** (↑)
RGB-Based Detection Approaches			
FasterRCNN [20]	NeurIPS 2015	30.2	67.6
DINO [50]	ICLR 2023	30.5	65.3
DeformableDETR [49]	ICLR 2021	31.2	68.4
YOLOv5 [23]	version 7.0	32.3	67.9
Thermal-Based Detection Approaches			
EGMT [85]	ICRA 2023	23.9	40.8
CE-RetinaNet [86]	TGRS 2023	30.0	62.3
DeformableDETR [49]	ICLR 2021	39.8	76.6
DINO [50]	ICLR 2023	40.3	78.7
YOLOv5 [23]	version 7.0	42.2	78.4
TIRDet [33]	ACM MM 2023	44.3	81.4
Multispectral Detection Approaches			
MoE-Fusion [78]	ICCV 2023	-	55.8
ProbEn [15]	ECCV 2022	37.9	75.5
MSAT [36]	IEEE SPL 2023	39.0	76.2
CSAA [82]	CVPR 2023	41.3	79.2
ICAFusion [87]	Pattern Recognition 2024	41.4	79.2
MFPT [13]	TITS 2023	-	80.0
YOLO-Adaptor [62]	TIV 2024	-	80.1
CMX [11]	TITS 2023	42.3	82.2
LRAF-Net [10]	TNNLS 2023	42.8	80.5
IGT [31]	Knowledge-Based Systems 2023	43.6	85.0
Fusion-Mamba [64]	arXiv 2024	44.4	84.3
EFCDet (Ours)	-	46.6	86.6

从表 3 中，我们可以看到以下观察结果：（1）与最好的单模态检测方法 TIRDet 相比，我们的 EFCDet 显示出显著的改进（+2.3% AP）[33]。（2）我们的 EFCDet 明显超过了之前最先进的多光谱检测方法 Fusion-Mamba [64]，高出 2.2% AP。表 4 显示了 M3FD 数据集的结果。我们的 EFCDet 还实现了最先进的性能，比第二好的方法 TarDAL [17] 高出 1.9%AP。这

些改进验证了我们的增强特征对比度的策略有利于多类对象检测任务。

表 4. M3FD 数据集上 AP （%）的比较。所有结果均使用 YOLOv5 [23] 作为检测器获得。最佳结果以粗体突出显示并标记为红色，而次佳结果则以绿色突出显示并标记。

Method	**Publication Year**	**AP** (↑)	**AP50** (↑)
Single-Modality Detection Baselines			
RGB [23]	version 7.0	36.1	60.2
Thermal [23]	version 7.0	34.9	57.2
Multispectral Detection Approaches			
DIVFusion [74]	Information Fusion 2023	37.1	60.8
PSFusion [88]	Information Fusion 2023	38.0	61.1
AUIF [89]	TCSVT 2022	38.3	62.0
CDDF [90]	CVPR 2023	38.6	61.9
U2Fusion [91]	TPAMI 2020	38.7	61.9
TarDAL [17]	CVPR 2022	39.1	61.9
EFCDet (Ours)	-	41.0	64.8

5.3 消融实验

我们在 FLIR 数据集上进行消融研究，以评估我们方法中每个组成部分的效果。具体来说，我们在融合过程中逐步包括每个成分（FFM、FRM、L_{seg} 和 L_{neg_corr}），并在表 5 中报告相应的检测结果。同时，我们在图 6 中显示了每个消融研究设置下假阳性 (FPs) 的置信度分数直方图。在此图中，我们将置信度分数大于 0.3 的条形标记为红色，并显示它们的数量以便于比较。选择 0.3 的原因是，分数大于此值的 FP 对检测性能的负面影响要大得多。

从表 5 和图 6 中，我们可以看到以下观察结果：（1）当同时包括 FFM 和 FRM 时，与仅使用 FFM 相比，MR 降低了 1.75%，并且分数大于 0.3 的 FP 数量减少了 564 个，（2）当包括 FFM、FRM 和 L_{seg} 时，MR 比仅使用 FFM 减少了 3.16%，分数大于 0.3 的 FP 数量减少了 615 个，（3）当包括所有组件时，与仅使用 FFM 相比，MR 减少了 4.1%，FP 数量减少了 664。这些结果证实了 FP 的降低与检测性能的提高之间的相关性。

表 5. 基于 Faster R-CNN 的方法在 FLIR 数据集上以 512 × 640 的分辨率进行的推理时间。

Method	**FFM**	**FRM**	$\mathcal{L}_{\mathbf{seg}}$	$\mathcal{L}_{\mathbf{neg\_corr}}$	**MR** (↓)
Ablation	✓				8.57
	✓	✓			6.82
	✓	✓	✓		5.41
EFCDet (Ours)	✓	✓	✓	✓	**4.47**

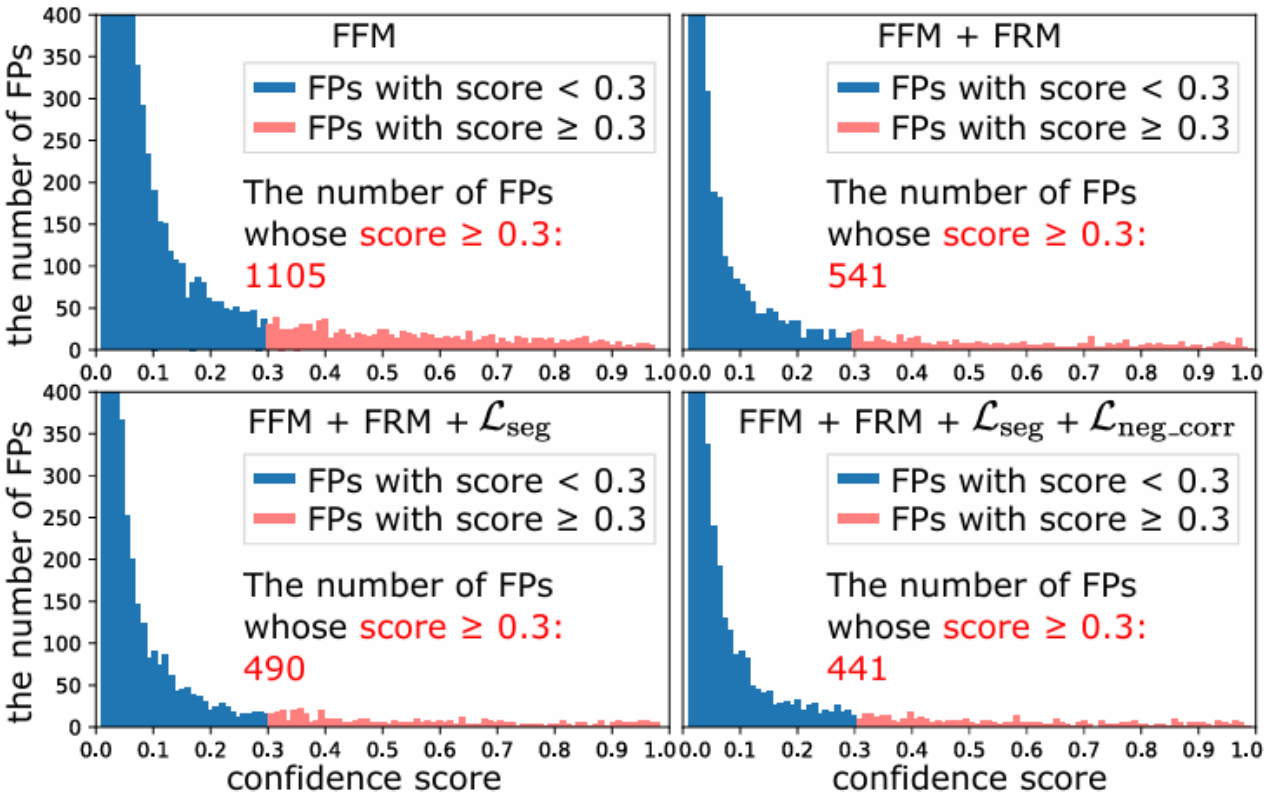

图 6. 在各种消融研究设置下 FLIR 数据集上假阳性（FP）的置信度得分直方图。

6 总结与展望

在这项工作中，我们解决了多光谱行人检测任务中特征融合的挑战。我们发现，由噪声特征引起的假阳性（FP）显著降低了检测性能，这在以前的工作中被忽视了。为了解决噪声特征的问题，我们为检测任务提出了一种基于增强特征对比度的多光谱融合目标检测，名为 EFCDet。在此策略中，我们引入了一个特征融合模块（FFM）来收集互补特征，一个特征优化模块（FRM）来区分行人和背景特征，以及一个相关-最大损失函数来增强特征对比度。可视化结果表明，我们的融合策略产生了判别性特征并显著降低了 FP。因此，EFCDet 在三个基准测试中实现了最先进的性能。此外，我们的策略灵活且计算效率高，可以应用于单级和两级检测器，同时保持较短的推理时间。基于目标感知融合策略，我们将在未来探索多光谱目标跟踪。

参考文献

[1] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral Pedestrian Detection: Benchmark Dataset and Baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "LLVIP: A Visible Infrared Paired Dataset for Low-Light Vision," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[3] Y. Cao, X. Luo, J. Yang, Y. Cao, and M. Y. Yang, "Locality Guided Cross-Modal Feature Aggregation and Pixel-Level Fusion for Multispectral Pedestrian Detection," *Information Fusion*, 2022.

[4] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral Pedestrian Detection via Simultaneous Detection and Segmentation," in *British Machine Vision Conference (BMVC)*, 2018.

[5] J. Liu, S. Zhang, S. Wang, and D. Metaxas, "Multispectral Deep Neural Networks for Pedestrian Detection," in *British Machine Vision Conference (BMVC)*, 2016.

[6] J. Xie, R. M. Anwer, H. Cholakkal, J. Nie, J. Cao, J. Laaksonen, and F. S. Khan, "Learning a Dynamic Cross-Modal Network for Multispectral Pedestrian Detection," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2022.

[7] X. Yang, Y. Qian, H. Zhu, C. Wang, and M. Yang, "BAANet: Learning Bi-Directional Adaptive Attention Gates for Multispectral Pedestrian Detection," in *International Conference on Robotics and Automation (ICRA)*, 2022.

[8] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[9] K. Zhou, L. Chen, and X. Cao, "Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[10] H. Fu, S. Wang, P. Duan, C. Xiao, R. Dian, S. Li, and Z. Li, "LRAF-Net: Long-Range Attention Fusion Network for Visible-Infrared Object Detection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023.

[11] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[12] J. U. Kim, S. Park, and Y. M. Ro, "Uncertainty-Guided Cross-Modal Learning for Robust Multispectral Pedestrian Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[13] Y. Zhu, X. Sun, M. Wang, and H. Huang, "Multi-Modal Feature Pyramid Transformer for RGB-Infrared Object Detection," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[14] K. Dasgupta, A. Das, S. Das, U. Bhattacharya, and S. Yogamani, "Spatio-Contextual Deep Network-Based Multimodal Pedestrian Detection for Autonomous Driving," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[15] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multi-modal Object Detection via Probabilistic Ensembling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[16] "FREE FLIR Thermal Dataset for Algorithm Training," https://www.flir.com/oem/adas/adas-dataset-form/.

[17] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-Aware Dual Adversarial Learning and A Multi-Scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5802–5811.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[21] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse R-CNN: End-to-End Object Detection with Learnable Proposals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[22] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," *Proceedings of the IEEE*, 2023.

[23] G. Jocher, "YOLOv5 by Ultralytics," 2020. [Online]. Available: https://github.com/ultralytics/yolov5.

[24] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[25] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[26] W. Liu, I. Hasan, and S. Liao, "Center and Scale Prediction: Anchor Free Approach for Pedestrian and Face Detection," *Pattern Recognition*, 2023.

[27] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable Pedestrian Detection: The Elephant in the Room," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[28] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-Aware Faster R-CNN for Robust Multispectral Pedestrian Detection," *Pattern Recognition*, 2019.

[29] Y. Liu, C. Hu, B. Zhao, Y. Huang, and X. Zhang, "Region-Based Illumination-Temperature Awareness and Cross-Modality Enhancement for Multispectral Pedestrian Detection," *IEEE Transactions on Intelligent Vehicles*, pp. 1–12, 2024.

[30] Q. Xie, T.-Y. Cheng, Z. Dai, V. Tran, N. Trigoni, and A. Markham, "Illumination-Aware Hallucination-Based Domain Adaptation for Thermal Pedestrian Detection," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[31] K. Chen, J. Liu, and H. Zhang, "IGT: Illumination-Guided RGB-T Object Detection with Transformers," *Knowledge-Based Systems*, vol. 268, p. 110423, 2023.

[32] Y. Zhang, H. Yu, Y. He, X. Wang, and W. Yang, "Illumination-Guided RGBT Object Detection with Inter- and Intra-Modality Fusion," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[33] Z. Wang, F. Colonnier, J. Zheng, J. Acharya, W. Jiang, and K. Huang, "TIRDet: Mono-Modality Thermal InfraRed Object Detection Based on Prior Thermal-To-Visible Translation," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2023.

[34] X. Li, S. Chen, C. Tian, H. Zhou, and Z. Zhang, "M2FNet: Mask-guided Multi-level Fusion for RGB-T Pedestrian Detection," *IEEE Transactions on Multimedia*, pp. 1–13, 2024.

[35] W.-Y. Lee, L. Jovanov, and W. Philips, "Cross-Modality Attention and Multimodal Fusion Transformer for Pedestrian Detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[36] S. You, X. Xie, Y. Feng, C. Mei, and Y. Ji, "Multi-Scale Aggregation Transformers for Multispectral Object Detection," *IEEE Signal Processing Letters*, 2023.

[37] L. Zhang, Z. Liu, X. Zhu, Z. Song, X. Yang, Z. Lei, and H. Qiao, "Weakly Aligned Feature Fusion for Multimodal Object Detection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[38] M. Yuan and X. Wei, "C²Former: Calibrated and Complementary Transformer for RGB-Infrared Object Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[39] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly Supervised Object Localization and Detection: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5866-5885, 2021.

[40] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Guided Attentive Feature Fusion for Multispectral Pedestrian Detection," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[41] L. Van der Maaten and G. Hinton, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research*, 2008.

[42] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection,"in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[43] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[44] R. Zhang, L. Li, Q. Zhang, J. Zhang, L. Xu, B. Zhang, and B. Wang, "Differential Feature Awareness Network within Antagonistic Learning for Infrared-Visible Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[45] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More Deformable, Better Results," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[46] X. Zhang, Z. Sheng, and H.-L. Shen, "FocusNet: Classifying Better by Focusing on Confusing Classes," *Pattern Recognition*, 2022.

[47] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *International Conference on 3D Vision (3DV)*, 2016.

[48] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[49] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection,"in *Proceedings of International Conference on Learning Representations (CVPR)*, 2021.

[50] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2023.

[51] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of Multispectral Data Through Illumination-Aware Deep Neural Networks for Pedestrian Detection," *Information Fusion*, 2019.

[52] L. Zhang, Z. Liu, S. Zhang, X. Yang, H. Qiao, K. Huang, and A. Hussain, "Cross-Modality Interactive Attention Network for Multispectral Pedestrian Detection," *Information Fusion*, 2019.

[53] J. U. Kim, S. Park, and Y. M. Ro, "Towards Versatile Pedestrian Detector with Multisensory-Matching and Multispectral Recalling Memory," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[54] T. Kim, S. Shin, Y. Yu, H. G. Kim, and Y. M. Ro, "Causal Mode Multiplexer: A Novel Framework for Unbiased Multispectral Pedestrian Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[55] Q. Li, C. Zhang, Q. Hu, H. Fu, and P. Zhu, "Confidence-Aware Fusion using Dempster-Shafer Theory for Multispectral Pedestrian Detection," *IEEE Transactions on Multimedia*, 2022.

[56] J. Kim, H. Kim, T. Kim, N. Kim, and Y. Choi, "MLPD: Multi-Label Pedestrian Detector in Multispectral Domain," *IEEE Robotics and Automation Letters*, 2021.

[57] Q. Li, C. Zhang, Q. Hu, P. Zhu, H. Fu, and L. Chen, "Stabilizing Multispectral Pedestrian Detection with Evidential Hybrid Fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[58] R. Li, J. Xiang, F. Sun, Y. Yuan, L. Yuan, and S. Gou, "Multiscale Cross-Modal Homogeneity Enhancement and Confidence-Aware Fusion for Multispectral Pedestrian Detection," *IEEE Transactions on Multimedia*, vol. 26, pp. 852–863, 2024.

[59] C. Tian, Z. Zhou, Y. Huang, G. Li, and Z. He, "Cross-Modality Proposal Guided Feature Mining for Unregistered RGB-Thermal Pedestrian Detection," *IEEE Transactions on Multimedia*, vol. 26, 2024.

[60] X. Zou, T. Peng, and Y. Zhou, "UAV-Based Human Detection With Visible-Thermal Fused YOLOv5 Network," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 3, pp. 3814–3823, 2024.

[61] N. Chen, J. Xie, J. Nie, J. Cao, Z. Shao, and Y. Pang, "Attentive Alignment Network for Multispectral Pedestrian Detection," in *Proceedings of ACM International Conference on Multimedia (ACM MM)*, 2023.

[62] H. Fu, H. Liu, J. Yuan, X. He, J. Lin, and Z. Li, "YOLO-Adaptor: A Fast Adaptive One-Stage Detector for Non-Aligned Visible-Infrared Object Detection," *IEEE Transactions on Intelligent Vehicles*, pp. 1–14, 2024.

[63] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks," in *Proceedings of International Conference on Image Processing (ICIP)*, 2020.

[64] W. Dong, H. Zhu, S. Lin, X. Luo, Y. Shen, X. Liu, J. Zhang, G. Guo, and B. Zhang, "Fusion-Mamba for Cross-Modality Object Detection," *arXiv preprint arXiv:2404.09146*, 2024.

[65] X. Zhang, S.-Y. Cao, F. Wang, R. Zhang, Z. Wu, X. Zhang, X. Bai, and H.-L. Shen, "Rethinking Early-Fusion Strategies for Improved Multispectral Object Detection," *arXiv preprint arXiv:2405.16038*, 2024.

[66] MMDetection Contributors, "OpenMMLab Detection Toolbox and Benchmark," 2018. [Online]. Available: https://github.com/open-mmlab/mmdetection.

[67] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[69] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "DetFusion: A Detection-Driven Infrared and Visible Image Fusion Network,"in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2022.

[70] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[71] H. Liu, F. Jin, H. Zeng, H. Pu, and B. Fan, "Image Enhancement Guided Object Detection in Visually Degraded Scenes," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023.

[72] H. R. Medeiros, F. A. Guerrero Peña, M. Aminbeidokhti, T. Dubail, E. Granger, and M. Pedersoli, "HalluciDet: Hallucinating RGB Modality for Person Detection Through Privileged Information," in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[73] Y. Cao, Y. Fan, J. Bin, and Z. Liu, "Lightweight Transformer for Multi-Modal Object Detection (Student Abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 37, no. 13, 2023, pp. 16172–16173.

[74] L. Tang, X. Xiang, H. Zhang, M. Gong, and J. Ma, "DIVFusion: Darkness-Free Infrared and Visible Image Fusion," *Information Fusion*, vol. 91, pp. 477–493, 2023.

[75] J. Chen, L. Yang, W. Liu, X. Tian, and J. Ma, "LENFusion: A Joint Low-Light Enhancement and Fusion Network for Nighttime Infrared and Visible Image Fusion," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[76] G. Xu, C. He, H. Wang, H. Zhu, and W. Ding, "DM-Fusion: Deep Model-Driven Network for Heterogeneous Image Fusion," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.

[77] L. Tang, Z. Chen, J. Huang, and J. Ma, "CAMF: An Interpretable Infrared and Visible Image Fusion Network Based on Class Activation Mapping," *IEEE Transactions on Multimedia*, 2023.

[78] B. Cao, Y. Sun, P. Zhu, and Q. Hu, "Multi-Modal Gated Mixture of Local-to-Global Experts for Dynamic Image Fusion," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023, pp. 23555–23564.

[79] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "MetaFusion: Infrared and Visible Image Fusion via Meta-Feature Embedding from Object Detection," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13955–13965.

[80] F. Chu, J. Cao, Z. Song, Z. Shao, Y. Pang, and X. Li, "Toward Generalizable Multispectral Pedestrian Detection," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[81] Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, and L. Van Gool, "DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023, pp. 8082–8093.

[82] Y. Cao, J. Bin, J. Hamari, E. Blasch, and Z. Liu, "Multimodal Object Detection by Channel Switching and Spatial Attention," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 403–411.

[83] X. Yi, L. Tang, H. Zhang, H. Xu, and J. Ma, "Diff-IF: Multi-Modality Image Fusion via Diffusion Model with Fusion Knowledge Prior," *Information Fusion*, p. 102450, 2024.

[84] X. He, C. Tang, X. Zou, and W. Zhang, "Multispectral Object Detection via Cross-Modal Conflict-Aware Learning," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2023, pp. 1465-1474.

[85] D.-G. Lee, M.-H. Jeon, Y. Cho, and A. Kim, "Edge-Guided Multi Domain RGB-to-TIR Image Translation for Training Vision Tasks with Challenging Labels," in *Proceedings of the International Conference on Robotics and Automation (ICRA). IEEE, 2023*, pp. 8291–8298.

[86] Y. Zhang and Z. Cai, "CE-RetinaNet: A Channel Enhancement Method for Infrared Wildlife Detection in UAV Images," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[87] J. Shen, Y. Chen, Y. Liu, X. Zuo, H. Fan, and W. Yang, "ICAFusion: Iterative Cross-Attention Guided Feature Fusion for Multispectral Object Detection," *Pattern Recognition*, vol. 145, p. 109913, 2024.

[88] L. Tang, H. Zhang, H. Xu, and J. Ma, "Rethinking the Necessity of Image Fusion in High-Level Vision Tasks: A Practical Infrared and Visible Image Fusion Network Based on Progressive Semantic Injection and Scene Fidelity," *Information Fusion*, vol. 99, p. 101870, 2023.

[89] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang, and J. Liu, "Efficient and Model-Based Infrared and Visible Image Fusion via Algorithm Unrolling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1186–1196, 2022.

[90] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "CDDFuse: Correlation-Driven Dual-Branch Feature Decomposition for Multi-Modality Image Fusion," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5906–5916.

[91] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A Unified Unsupervised Image Fusion Network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2020.