引用格式：LI Jiale，HUI Bingwei，WANG Kaiyue，et al. Zero-Shot Infrared Object Detection via Frequency-Domain Feature Decoupling of Visible Images[J]. Acta Photonica Sinica，2026，55(8)：0810001
李佳乐，回丙伟，王凯月，等．基于可见光图像频域特征解耦的零样本红外目标检测方法[J]．光子学报，2026，55(8)：0810001

基于可见光图像频域特征解耦的零样本红外目标检测方法

李佳乐，回丙伟，王凯月，张军
（中国人民解放军国防科技大学 电子科学学院，长沙 410028）

摘　要：针对典型应用场景下红外图像训练样本匮乏导致的零样本检测难题，提出一种基于可见光图像频域特征解耦的零样本红外目标检测方法。首先，设计频域处理模块，通过削弱高频纹理以降低源域特有风格干扰，并对低频分量实施扰动以提升源域样本多样性，从而为跨模态域不变特征学习提供先验约束。其次，构建权重共享的双分支特征提取框架，并在浅层嵌入特征解耦模块，以显式分离域不变特征与域私有特征，增强模型对跨波段共性表征的建模能力。最后，在颈部网络中引入注意力引导的特征拼接策略，优化多尺度信息融合过程，提升小目标感知与定位性能。在DroneVehicle和FLIR公开数据集上的实验结果表明，本方法与最新的单域泛化目标检测算法相比，mAP@50分别提升了2.9%和5%，验证了所提方法在跨波段零目标域样本条件下的有效性与鲁棒性，为单域泛化目标检测提供了新的解决思路。

关键词：跨域泛化；零样本红外目标检测；频域增强；特征解耦；小目标检测
中图分类号：TP391.4　　**文献标识码**：A
doi：10.3788/gzxb20265508.0810001

0　引言

红外成像具备全天时探测和隐蔽目标识别能力，在精确打击、战场侦察、灾害预警、自动驾驶及工业安防等领域具有重要应用价值。然而，在非合作条件下，红外数据难以获取，面向红外场景的检测模型常面临目标域样本缺失甚至零样本训练的问题。相比之下，天基光学遥感卫星能够获得大量可见光数据，且可见光与红外图像在语义层面具有高度一致性，因此，利用可见光数据训练可泛化至红外场景的目标检测模型，为缓解红外样本缺失问题提供了一种可行思路。针对这类跨域目标检测问题，传统模型微调和域适应方法通常依赖目标域样本，难以适用于零样本红外检测场景，而单域泛化仅利用源域数据训练模型，即可提升模型对未知目标域的泛化能力，更符合零样本红外目标检测的应用需求。

单域泛化技术已在图像分类等经典视觉任务中得到广泛应用[1][2]，在目标检测领域也展现出极大的潜力。例如，史彩娟等[3]针对X-ray乳腺肿瘤检测问题，提出了单域泛化模型，通过域特征增强与实例泛化策略，有效缓解了模型在未知域上的性能衰减问题。现有的单域泛化目标检测方法主要分为纯视觉特征泛化和跨模态语义对齐两条路径。

在纯视觉特征泛化研究中，根据处理的思想，可将现有方法分成四类，即域不变特征表示学习、多专家混合方法、元学习方法和深度图匹配方法[4]。域不变特征表示学习旨在提取跨域稳定的判别特征，以提升模型在未见域上的泛化能力。JIN Xin等[5]采用风格归一化与恢复模块，分离身份特征与身份无关特征，可以弥补实例归一化层带来的判别信息的损失；ZHENG Kecheng等[6]则引入校准实例归一化，通过过滤身份无

第一作者：李佳乐(2002－)，女，硕士研究生，研究方向为目标特性数据工程。Email：lijiale24a@nudt.edu.cn
通讯作者：回丙伟(1985－)，男，副教授，研究方向为目标特性数据工程。Email：huibingwei@nudt.edu.cn
收稿日期：XXXX-XX-XX；**录用日期**：XXXX-XX-XX

关特征可进一步挖掘判别信息，以获得更具表达能力的判别特征；LIU Jiawei等[7]则将BN层（Batch Normalization，BN）的特征估计建模为动态自我精炼的高斯过程，可获得除已知域外的潜在域的BN估计，提高了泛化能力。针对目标检测任务，WU Aming等[8]构建了循环解耦自蒸馏框架，通过将特征空间划分为域私有子空间与域不变子空间，并利用自蒸馏机制强化跨域共享表征学习；LIN Chuang等[9]从图像级和实例级同时进行解耦，并引入跨层级重建策略，以减轻单一级别解耦带来的细节损失。

除域不变特征学习外，多元专家混合方法和元学习方法也被用于提升模型对未知域的适应能力。多元专家混合方法是采用一种集成技术，通过单独训练多个源域数据的多个专家网络，再通过投票混合机制利用多个领域专家获得多样性特征。YU Shijie等[10]提出多域专家协作学习框架，在训练域特定专家的同时聚合不同专家知识，并在测试阶段仅使用通用域专家。元学习方法则是通过学习多个任务的经验，使模型自己组合不同的算法与超参数，适应不同的任务。LI Da等[11]是首次采用元学习方法解决域泛化问题的团队，在此基础上，ZHAO Yuyang等[12][13]通过记忆非参数识别损失、动态风格生成和一致性约束提升模型的风格鲁棒性与语义稳定性。近年来，基于图像级和特征级增强的源域多样化方法也受到关注，如QI Lei等[14]结合图像级与特征级增强策略，通过颜色扰动和物体—背景风格切换提升源域多样性，FAN Qi等[15]通过扰动低级特征通道统计量生成潜在目标域风格。

深度图匹配方法是基于特征的表示学习构建的，模型训练后参数会被固定，泛化能力受限，为解决该问题，相关研究开始探索如何动态获取卷积参数，以便在深度特征图中进行局部匹配。LIAO Shengcai等[16]提出一种查询图自适应卷积方法，利用查询图像动态构建自适应的卷积核，在候选图片中执行卷积计算和全局最大池化操作，这可实现更为精准的图像对应点匹配，显著提升了模型对未知场景的适应能力，此后LIAO Shengcai等[16]为提升检索结果的质量，提出了一种基于时序共现的相似度加权方法，提升了检索精度。总体来看，上述纯视觉方法能够在一定程度上缓解了域偏移问题，但多数方法侧重通用视觉域泛化，较少结合红外成像特有的频域分布规律进行建模；同时，部分方法依赖复杂模型结构或多源域训练设置，在零样本红外目标检测和轻量化部署场景中的适用性仍然受限。

随着多模态学习的发展，跨模态语义对齐逐渐被应用在跨域目标检测任务中，该类方法通常借助文本提示等语义信息弥补源域样本多样性不足，并增强模型对未见域的理解能力。VIDIT V等[17]提出对比语言—图像预训练模型（Contrastive Language-Image Pre-training，CLIP），充分利用文本模态和视觉模态，增强模型对语义信息的挖掘；DENG Li等[19]利用不同场景的语言描述信息进行以目标为中心的跨模态融合；LI Hao等[19]则通过预设短语提示对齐源域视觉区域与潜在目标域语义区域，从而提升模型对未知域的适配能力。尽管该类方法能够利用视觉与语言模态的互补性增强语义泛化能力，但现有研究多侧重局部特征对齐，对全局语义结构和跨模态频域差异的建模仍然不足。

尽管域泛化目标检测已经取得一定进展，但在零样本红外场景下仍面临三方面挑战。其一，红外成像具有低频信息相对突出、高频纹理相对弱化的模态先验，而现有方法通常将目标域视为完全未知，缺乏针对红外成像频域特性的显式建模，限制了模型向红外域泛化的上限。其二，YOLO系列模型凭借实时推理优势及轻量化优势，已成为红外制导等时效敏感场景和移动端设备的重要选择，但面向单阶段YOLO架构的域泛化研究仍相对不足。其三，红外小目标本身信息量有限，可见光与红外之间的显著模态差异进一步加剧了跨域特征对齐难度，导致小目标漏检问题更加突出。

针对上述问题，本文提出一种基于可见光图像频域特征解耦的零样本红外目标检测方法。首先，设计高频抑制与低频增强（High-Frequency Suppression and Low-Frequency Enhancement，HSLE）模块，在模拟红外成像纹理特征的同时提升源域样本分布的多样性。其次，在主干网络浅层引入特征解耦（Feature Decoupling，FD）模块，通过权重共享的双分支结构显式分离域不变特征与域私有特征，以强化网络对跨波段共性表征的学习。最后，在小目标优化（Small Object Optimization，SO Optimization）模块中引入SPD_Conv[20]与GA_Concat[21]（Guided Attention_Concat，GA_Concat），分别优化主干下采样过程与颈部多尺度特征融合，从而增强模型对红外小目标的感知与定位能力。

1 频域特征解耦目标检测方法

本文基于YOLOv8，提出方法的整体框架如图1所示，采用可见光数据训练模型，在红外数据集上

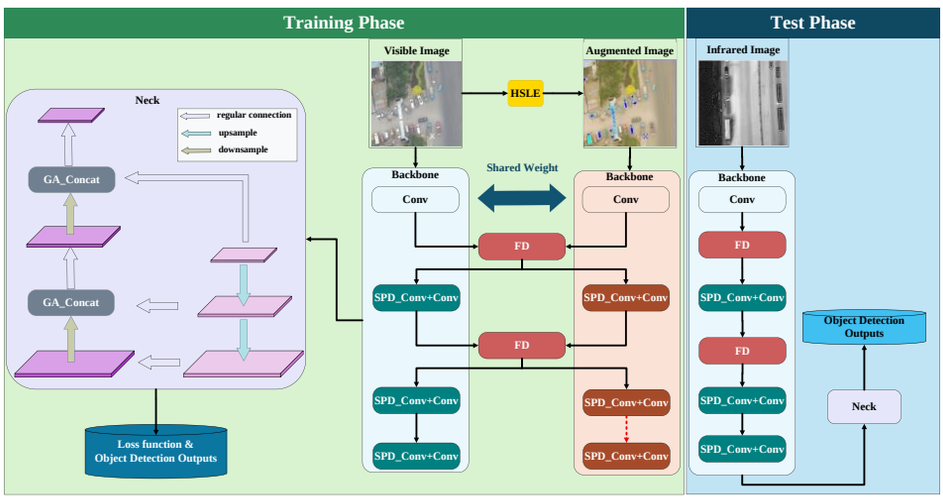


图1 总体架构

Fig. 1 Overall architecture

做验证。为简化表示，图1省略了YOLOv8中的C2f、SPPF等基础模块。在训练阶段，将可见光图像I送入HSLE模块，生成增强图像I_{aug}，随后将I和I_{aug}分别输入结构相同且权重共享的主干网络中，为进一步提取域不变特征，在主干网络浅层嵌入两个FD模块。经过第二个FD模块后，I_{aug}对应的特征图不再参与后续的训练，因此在图1中以红色虚线表示该连接。此外，为提升红外小目标特征的保留能力，将主干网络中的部分传统卷积替换为SPD_Conv，并将SPD_Conv的输出与对应原卷积输出进行逐元素相加，以增强下采样过程中细粒度信息的传递。与此同时，在颈部网络中采用GA_Concat替代原有Concat操作，以浅层特征信息引导深层特征图中对小目标的感知。在测试阶段，红外图像不再经由HSLE处理，直接沿着I的网络路径进行处理，最后得到红外域上的检测结果。

1.1 频域处理模块HSLE

单域泛化通常受限于源域风格单一性，模型容易过拟合源域的纹理统计特征，在可见光到红外的跨模态迁移场景中，该问题更加突出。可见光图像包含丰富的高频纹理细节，而红外图像则主要体现为低频热辐射结构，这种频域上的显著差异会对网络进行跨模态特征建模产生干扰。为此，本文提出HSLE模块，无需额外学习参数，通过抑制高频纹理并增强低频样本多样性，模拟红外成像的部分特性。

首先，利用哈尔小波变换对输入图像在频域上分解。给定一幅图像$I^{B\times C\times H\times W}$，其中$B$是样本数，$C$是通道数，$H$和$W$分别是图像的高度和宽度，哈尔小波变换利用4个卷积核对图像进行处理，即LL^T，LH^T，HL^T，HH^T，其中L和H分别代表低通滤波器和高通滤波器，得到4个分解后的子带分量，分别为I_{ll}，I_{lh}，I_{hl}，I_{hh}，将I_{ll}视为低频部分，剩余三个子带为高频部分[22]。图2展示了经哈尔小波分解后的结果，这5列分别代表$I^{B\times C\times H\times W}$、$I_{ll}$、$I_{lh}$、$I_{hl}$和$I_{hh}$。与原图相比，$I_{ll}$保留了图像的主体结构与语义信息，而剩余3个子带则编码了边缘与纹理等高频细节。

1.1.1 基于低频能量自适应的高频抑制

为模拟红外图像中高频信息较弱的特征，本文提出了一种基于低频能量的自适应高频抑制策略。首先定义低频分量的能量为

$$E_{ll}=\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}|I_{ll}(i,j)|^2 \tag{1}$$

$$E=\frac{E_{ll}-\min(E_{ll})}{\max(E_{ll})-\min(E_{ll})+\varepsilon} \tag{2}$$

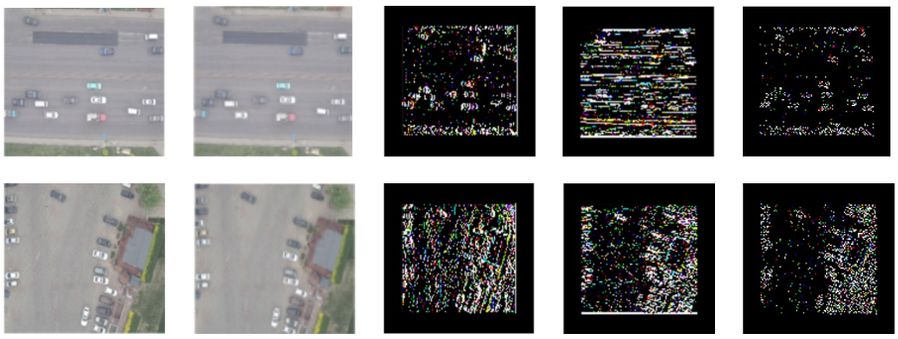


图 2 哈尔小波变换分解可见光图像可视化

Fig. 2 Visualization of visible-image decomposition by Haar wavelet transform

式中$I_{ll}(i,j)$表示在I_{ll}中位置坐标为(i,j)的像素值。由于不同样本的E_{ll}的范围不同,这里对E_{ll}做归一化得到E。基于E,构建指数衰减函数以生成自适应衰减系数$attenaution$为

$$attenaution=(1-base)\times\exp(-\alpha E)+base \tag{3}$$

其中α为高频抑制权重系数,$base$是高频保留系数,$attenaution\in[0,1]$。利用$attenaution$对高频子带进行动态缩放,即

$$I_{lh}=attenaution\times I_{lh} \tag{4}$$

$$I_{hl}=attenaution\times I_{hl} \tag{5}$$

$$I_{hh}=attenaution\times I_{hh} \tag{6}$$

该自适应高频抑制策略通过将高频响应与低频结构强度进行动态耦合,可削弱源域特有纹理对模型学习的干扰,同时保留了必要的边缘线索,从而使重构图像在频域分布上更接近红外图像。

1.1.2 低频风格扰动增强

为进一步提升模型对未知域的泛化能力,需要在保持语义信息稳定的前提下扩展源域样本的风格分布。现有研究多在空间域执行风格扰动[23],但容易引入语义失真;相比之下,频域扰动能够在保持结构信息完整的同时实现风格解耦[24][25]。因此,本文采用NP[14]作为低频扰动策略。

图 3 HSLE 处理前后结果对比

Fig. 3 Comparison of images before and after HSLE processing

具体而言,对低频分量I_{ll}沿通道维度计算均值$\boldsymbol{\mu}_c^{C\times1\times1}$和标准差$\boldsymbol{\sigma}_c^{C\times1\times1}$,并引入随机扰动参数$\beta,\gamma\sim N(0,1)$,生成新的统计量为$\boldsymbol{\mu}_c^*=\beta\boldsymbol{\mu}_c$,$\boldsymbol{\sigma}_c^*=\gamma\boldsymbol{\sigma}_c$,其中$\beta$和$\gamma$分别从0.8~1.2区间中随机采样。构造扰动后的低频分量$\hat{I}_{ll}=NP(I_{ll})$为式(7)所示,其中$\boldsymbol{x}$和$\boldsymbol{y}$分别表示输入和输出的特征向量。

$$\boldsymbol{y}=\boldsymbol{\sigma}_c^*\frac{\boldsymbol{x}-\boldsymbol{\mu}_c}{\boldsymbol{\sigma}_c}+\boldsymbol{\mu}_c^* \tag{7}$$

最后,将处理后的高频分量与低频分量通过哈尔小波逆变换重构,得到风格增强后的图像I_{aug}。经HSLE处理前、后的图像如图3所示。

1.2 特征解耦模块FD

FD模块用于特征解耦,将其设置在主干网络的浅层,即在原YOLOv8主干网络的第2、4层分别插入FD模块,其中编号从第0层开始计数。FD模块结构图如图4所示,包含一套权重独立的三元组特征解耦器,即E_{inv},E_{var}^{rgb},E_{var}^{aug},分别提取域不变特征、I的私有特征、I_{aug}的私有特征。给定原始可见光特征$F_{rgb}^{B\times C\times H\times W}$

和经HSLE处理后的风格增强特征$F_{aug}^{B\times C\times H\times W}$，三元组解耦过程可表示为

$$F_{inv}^{rgb}=E_{inv}(F_{rgb}^{B\times C\times H\times W}) \tag{8}$$

$$F_{inv}^{aug}=E_{inv}(F_{aug}^{B\times C\times H\times W}) \tag{9}$$

$$F_{\mathrm{var}}^{rgb}=E_{\mathrm{var}}^{rgb}(F_{rgb}^{B\times C\times H\times W}) \tag{10}$$

$$F_{\mathrm{var}}^{aug}=E_{\mathrm{var}}^{aug}(F_{aug}^{B\times C\times H\times W}) \tag{11}$$

其中，F_{inv}^{rgb}和F_{inv}^{aug}分别为来自I和I_{aug}的域不变特征，由于二者共享相同的语义空间，故由同一个域不变特征提取器E_{inv}生成，而E_{var}^{rgb}和E_{var}^{aug}则分别捕获各自输入特有的风格噪声。上述三元组特征解耦器的结构完全相同，均为两个卷积神经网络构成，初始化权重相同，并通过损失函数约束以实现特征分离。

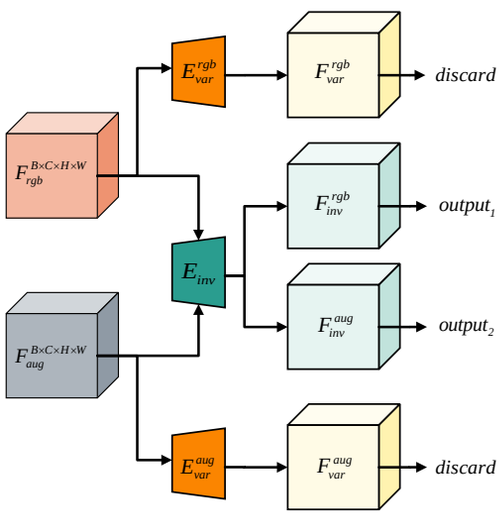


图4　特征解耦模块的结构

Fig. 4　Architecture of the feature decoupling module

(1)　三元组损失

为显式约束解耦后的特征空间，本文要求域不变特征在几何上保持更紧密的分布，而域私有特征与域不变特征之间应具有更大的分离度。故本文采用三元组损失对解耦器进行正则化，如式(12)所示。

$$L_{triplet}=\frac{1}{B}\sum_{i=1}^{B}\left[\begin{matrix}\max(0,d(F_{inv}^{rgb},F_{inv}^{aug})-d(F_{inv}^{rgb},F_{\mathrm{var}}^{rgb})+m)\\+\max(0,d(F_{inv}^{aug},F_{inv}^{rgb})-d(F_{inv}^{aug},F_{\mathrm{var}}^{aug})+m)\end{matrix}\right] \tag{12}$$

其中，B表示样本数量，$d(\cdot,\cdot)$表示欧氏距离，m是一个极小的正数，$L_{triplet}$通过最小化域不变特征的距离，同时最大化域不变特征与私有特征的距离，使网络学习到更具判别性的解耦特征表示。

(2)　熵损失

仅依赖几何距离约束的三元组损失缺乏显式语义监督，容易导致解耦边界模糊。受文献[26]启发，本文进一步引入信息熵约束，具体而言，域不变特征应覆盖更广泛的通用语义，其预测分布趋于均匀，因此应具有较高熵值；而域私有特征主要编码域特定细节，其分布应更集中，因此熵值相对较低。

为此，先构建一个轻量化的熵预测头*EntropyHead*，包含两个全连接层，最后一个全连接层的输出为类别总数*class*。*EntropyHead*对输入的特征做预测，得到类别i的预测类别分数为$\boldsymbol{s}_i$，经过*Soft*max分类器后，得到最终的分类结果$\hat{\boldsymbol{s}}_i$，则香农熵可表示为

$$H(\hat{\boldsymbol{s}}_i)=-\sum_{i=1}^{class}\hat{\boldsymbol{s}}_i\log\hat{\boldsymbol{s}}_i \tag{13}$$

将解耦后的特征F_{inv}^{rgb}、F_{inv}^{aug}、F_{var}^{rgb}、F_{var}^{aug}送入权重共享的*EntropyHead*中，经过*Soft*max后得到香农熵为$H(\hat{\boldsymbol{s}}_{inv}^{rgb})$、$H(\hat{\boldsymbol{s}}_{\mathrm{var}}^{rgb})$、$H(\hat{\boldsymbol{s}}_{inv}^{aug})$、$H(\hat{\boldsymbol{s}}_{\mathrm{var}}^{aug})$，对应的损失可表示为

$$L_{rgb}^{+}=Softplus(H(\hat{\boldsymbol{s}}_{inv}^{rgb})-H(\hat{\boldsymbol{s}}^{rgb})) \tag{14}$$

$$L_{rgb}^{-}=Softplus(H(\hat{\boldsymbol{s}}^{rgb})-H(\hat{\boldsymbol{s}}_{\mathrm{var}}^{rgb})) \tag{15}$$

$$L_{aug}^{+}=Softplus(H(\hat{s}_{inv}^{aug})-H(\hat{s}^{aug})) \tag{16}$$

$$L_{aug}^{-}=Softplus(H(\hat{s}^{aug})-H(\hat{s}_{\mathrm{var}}^{aug})) \tag{17}$$

$H(\hat{s}^{rgb})$和$H(\hat{s}^{aug})$分别是$F_{rgb}^{B\times C\times H\times W}$和$F_{aug}^{B\times C\times H\times W}$对应的香农熵，故熵损失可表示为

$$L_{entropy}=L_{rgb}^{+}+L_{rgb}^{-}+L_{aug}^{+}+L_{aug}^{-} \tag{18}$$

(3) 总损失

综合上述约束，三元组解耦器的总正则化损失可表示为

$$L_{FD}=L_{triplet}+\lambda L_{entropy} \tag{19}$$

其中，λ为$L_{triplet}$和$L_{entropy}$之间的平衡参数。在训练阶段，FD模块同时接收$F_{rgb}^{B\times C\times H\times W}$和$F_{aug}^{B\times C\times H\times W}$，并输出$F_{inv}^{rgb}$和$F_{inv}^{aug}$，在图4中用$output_1$和$output_2$标记，而$F_{\mathrm{var}}^{rgb}$和$F_{\mathrm{var}}^{aug}$不再参与后续特征的传递，因此被丢弃，图4中用*discard*标记。在测试阶段，FD模块仅接收$F_{rgb}^{B\times C\times H\times W}$的输入，故仅有$output_1$输出。

1.3 小目标优化模块SO Optimization

针对跨域检测中小目标表征易丢失的问题，本文提出小目标优化策略SO Optimization，包括SPD_Conv和GA_Concat两个子模块。首先，在主干网络中引入SPD_Conv[20]，将空间维度信息转移到通道维度，以最大程度保留小目标细粒度特征；其次，受文献[22]的启发，在颈部网络中设计GA_Concat模块，替换原Concat结构，GA_Concat利用浅层细节生成引导注意力，对深层语义特征进行自适应重校准，从而增强模型对弱小目标的感知能力。

1.3.1 空间深度卷积块SPD_Conv

SPD_Conv由空间到深度层(Space-to-Depth, SPD)与步长为1的非跨步卷积串联组成，其结构如图5所示。给定一张大小为$S\times S$、通道数为C_1的特征图$X^{S\times S\times C_1}$，SPD层首先沿着空间维度将该特征图切片为子特征图，每张子特征图大小为$(S/scale, S/scale, C_1)$，其中*scale*表示下采样因子。随后，将这些子特

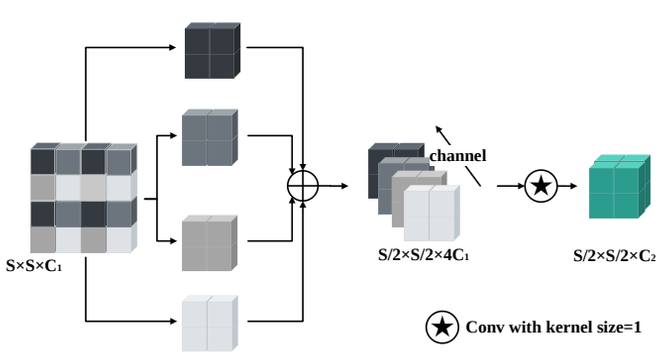


图5 scale=2时SPD_Conv的结构

Fig. 5 Architecture of SPD_Conv with scale=2

征图沿着通道维度拼接后得到新的特征图X'，其大小为$(S/scale, S/scale, scale^2C_1)$。SPD层将空间特征信息转换成通道特征信息，避免因普通卷积下采样过程中带来的信息丢失的问题。紧接着，X'通过一个核大小为1×1的卷积进行特征融合与通道降维，输出最终特征图Y，大小为$(S/scale, S/scale, C_2)$，1×1的卷积核在特征图上滑动的步长为1，这种设计可以使得卷积算子对特征图的每一个空间位置进行密集采样，减少像素的遗漏。

1.3.2 引导注意力拼接模块GA_Concat

尽管SPD_Conv能够有效缓解下采样带来的信息损失，但随着网络层次加深，小目标响应仍可能在语义抽象过程中被弱化。受CoordAttention[27]和SENet[28]的启发，本文设计GA_Concat模块，用于建立浅层空间细节与深层语义信息之间的显式交互，其结构如图6所示。GA_Concat包含两个重要的注意力模块，即通道注意力模块(Channel Attention Module, CAM)和空间注意力模块(Spatial Attention Module, SAM)。浅层特征首先被送入CAM和SAM中，以生成通道掩码和空间掩码，随后，将深层特征沿通道维度划分为两个子分支，分别利用上述两种注意力进行自适应调制，最后，在通道维度上完成拼接。该模块通过双重注意力

的协同作用，可提升模型在跨域场景下的小目标感知能力。

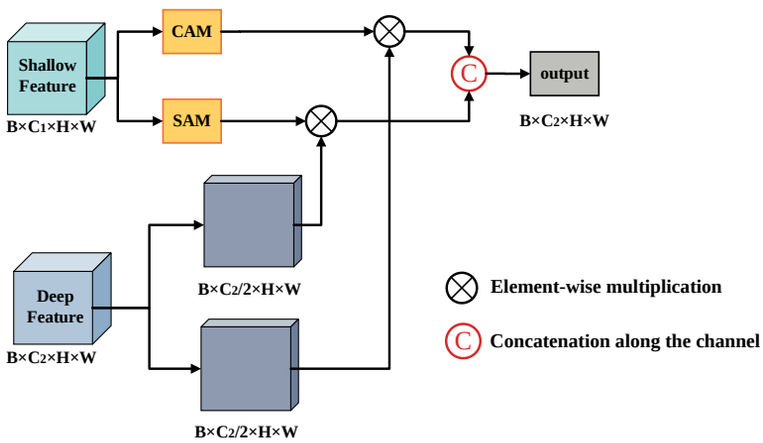


图 6　GA_Concat模块的结构

Fig. 6　Architecture of the GA_Concat module

2　数据集与实验设置

2.1　DroneVehicle数据集

DroneVehicle数据集为本文的主要实验数据集，包含多种光照条件下的无人机俯拍可见光—红外成对图像，目标类别为车辆（vehicle）。为评估模型在跨域场景下的泛化能力，本文在数据划分时保证训练集与验证集在场景类型上不重叠，以降低模型依赖背景偏置进行泛化的可能性。

2.1.1　可见光训练集

训练集以动态行驶车辆场景为主，共选取6351张图像，涵盖5类典型场景：主干道、高架桥、乡村道路、开阔地和交通枢纽。主干道场景道路形态规整、车辆行驶方向一致；高架桥场景受桥梁结构遮挡，车辆部分可见；乡村道路场景车辆稀疏，路侧植被丰富；开阔地场景背景简单，目标清晰无遮挡；交通枢纽场景车流复杂、建筑密集，目标集中分布。典型样例如图7所示。

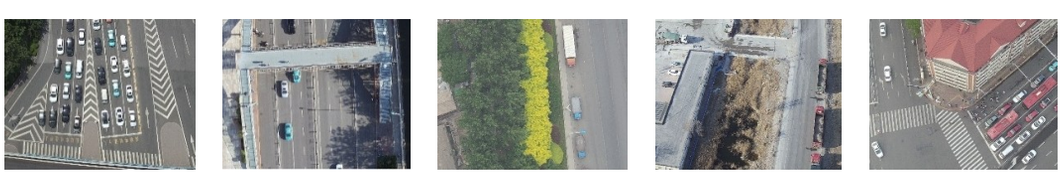


图 7　可见光训练集样例

Fig. 7　Visible-light training samples

2.1.2　红外验证集

验证集主要覆盖静态停车车辆场景，共包含1138张红外图像，分为路侧停车场、商圈停车场和居民区三类场景。路侧停车场景车辆沿道路两侧停放，中央留有通行空间；商圈停车场车辆密度高，方向差异明显且排列紧凑；居民区场景较为复杂，车辆常受到周边设施遮挡。上述三类场景分别包含584、244和310张图像，本文将其对应验证集记为val_1、val_2和val_3，典型样例如图8所示。

2.2　FLIR数据集

由于DroneVehicle数据集为低空无人机俯拍视角，目标外观特征相对单一，本文引入FLIR数据集进行补充论证。FLIR数据集由城市街道场景下的可见光与红外配对图像构成，涵盖车辆前视、后视等多种视角，场景复杂度更高。本文对其原始标注进行预处理，仅保留vehicle类别，共获得9595张标注可见光图像，并按照7∶3比例划分训练集和验证集，其中6717张可见光图像作为源域训练样本，剩余2878张可见光图像对应的配对红外图像作为目标域测试样本。FLIR数据集如图9所示。

图 8 红外验证集样例
Fig. 8 Infrared validation samples

图 9 FLIR 数据集样例
Fig. 9 Samples from the FLIR dataset

2.3 实验环境

实验硬件环境为24GB的NVIDIA GeForce RTX 4090显卡，操作系统为64位Windows 11，软件环境为CUDA 11.1、PyTorch 1.9.1和Python 3.8。实验采用YOLOv8n作为基线模型，最大迭代次数为100，训练批量大小为16，动量为0.937，权重衰减为5×10^{-4}，学习率初始化为0.01，前3个epoch对学习率进行线性预热。

3 实验结果与分析

3.1 模块有效性分析

参照COCO数据集中的目标尺度划分标准，本文将标注框面积范围为$[0,32^2)$的目标定义为小目标，经统计，DroneVehicle数据集中小目标占比仅为4.94%，而FLIR数据集中小目标占比达到28.75%，更适合用于分析SO Optimization模块在跨域小目标检测中的有效性，对于HSLE和FD模块的有效性验证，仍在DroneVehicle数据集上开展。

3.1.1 HSLE模块有效性分析

(1) 参数敏感性分析

HSLE模块中高频抑制权重系数α与高频保留系数*base*的选取，需兼顾高频冗余抑制与图像结构保留。高频信息承载可见光图像的纹理与边缘细节，过度抑制会造成图像平滑失真、结构特征丢失，而适度过滤冗余高频分量，能够有效缩小可见光与红外图像的跨模态域差异，提升模型的跨域特征适配能力。

为定量评估，本文引入结构相似性指标(Structural Similarity Index Measure，SSIM)和梯度幅值相似性指标(Gradient Magnitude Similarity，GMS)，二者的取值范围均为0到1。其中，SSIM用于判定高频衰减是否造成图像失真，数值越大相似性越高，GMS评估图像边缘与轮廓信息的保持情况，数值越大说明衰减图像与原始图像的边缘越接近。GMS的计算公式如式(20)所示，

$$S_{grad}(x)=\frac{2M_1(x)M_2(x)+\varepsilon}{M_1^2(x)+M_2^2(x)+\varepsilon} \tag{20}$$

其中M_1和M_2分别是I和I_{aug}的梯度幅值图，ε是防止分母为零的稳定常数。此外，为衡量参数的选择对最终目标检测结果的影响，还选取mAP@50作为参数选择的参考指标。

实验中，将系数α设置为[0.5，1.0，1.5，2.0，2.5]，*base*设置为[0.3，0.4，0.5，0.6，0.7]，对二者进行全组合实验。为排除其他模块对实验结果的干扰，本文仅在YOLOv8n基线模型中引入HSLE模块，并保持其他实验设置不变，结果如表1所示。分析可知，随着α逐步增大，SSIM、GMS整体呈下降趋势，这表明高频抑制强度越高，图像结构、边缘轮廓信息损失越明显，图像与原图的相似度持续降低；在同一α下，随*base*增大，SSIM、GMS大多小幅回升，说明提升高频保留比例，能够有效挽回部分纹理与边缘特征。

从检测精度来看，当α较小时，例如α为0.5时，模型能够保持较高的SSIM和GMS，但其最高mAP@50

仅为77.9%，这说明过弱的高频抑制虽然能够较好保留可见光图像细节，但仍可能保留较多可见光域特有的纹理干扰；当α增大至2.0或2.5时，虽然部分参数组合仍能保持较高SSIM和GMS，但mAP@50整体明显下降，最高仅分别为76.0%和75.3%，这表明过强的高频抑制会削弱目标检测所依赖的边缘和局部纹理信息，从而影响模型的目标定位和分类性能。

因此，当α为1.5，*base*为0.5时，mAP@50达到78.9%，此时SSIM为97.8%，GMS为94.5%，这组参数在结构保持与高频抑制之间取得了较好的平衡。

表1　HSLE参数敏感性分析

Table 1　Parameter Sensitivity Analysis of HSLE

α	*base*	SSIM/%	GMS/%	mAP@50/%
0.5	0.3	98.1	94.9	77.6
	0.4	98.0	95.1	77.3
	0.5	98.3	95.2	77.9
	0.6	98.1	95.2	77.5
	0.7	98.2	95.4	77.8
1.0	0.3	97.7	94.3	76.7
	0.4	97.8	94.6	77.1
	0.5	98.0	94.7	76.9
	0.6	98.0	95.0	77.4
	0.7	98.1	95.0	77.2
1.5	0.3	97.3	93.6	76.1
	0.4	97.4	93.9	77.0
	0.5	**97.8**	**94.5**	**78.9**
	0.6	98.0	94.7	78.2
	0.7	98.2	95.1	77.7
2.0	0.3	96.9	93.0	75.4
	0.4	97.2	93.6	75.8
	0.5	97.6	93.9	75.2
	0.6	97.8	94.5	76.0
	0.7	98.0	94.9	75.6
2.5	0.3	96.6	92.7	74.8
	0.4	97.0	93.1	74.4
	0.5	97.2	93.7	75.1
	0.6	97.7	94.4	74.7
	0.7	97.9	94.7	75.3

(2)　自适应高频抑制验证

为了验证经自适应高频抑制处理后的可见光图像更接近红外图像，本文选取DroneVehicle验证集中的红外数据及其配对的可见光数据作为实验样本，并引入径向平均功率谱密度（Radial Average Power Spectral Density，RAPSD）作为衡量指标，其计算方法为，对输入图像$I(x,y)$进行傅里叶变换，即

$$f(u,v)=F[I(x,y)] \tag{21}$$

其中$F[\cdot]$表示傅里叶变换，u和v分别表示频域中的水平和垂直坐标，再计算二维功率谱密度为

$$p(u,v)=\left|f(u,v)\right|^2 \tag{22}$$

再以频谱中心(u_0,v_0)为原点，计算每个频率点到(u_0,v_0)的径向距离，即

$$r(u,v)=\sqrt{(u-u_0)^2+(v-v_0)^2} \tag{23}$$

对于第k个径向频率区间，定义其对应的频率点集合为

$$\Omega_k=\{(u,v)|r_k\leqslant r(u,v)<r_{k+1}\} \tag{24}$$

则在该频率区间上的$RAPSD(k)$可定义为式(25),其中$|\Omega_k|$表示第k个径向频率区间内的频率点数量。

$$RAPSD(k)=\frac{1}{|\Omega_k|}\sum_{(u,v)\in\Omega_k}p(u,v) \tag{25}$$

在本文中,对$RAPSD(k)$归一化并取对数后作为纵坐标,横轴表示归一化的径向频率,其取值范围为0到1,频率值越大,表示对应频段越靠近高频区域,纵轴数值越大,表示该频段的能量越强。

图10展示了真实可见光图像(Visible)、真实红外图像(Infrared)以及经自适应高频抑制处理后的可见光图像(Adaption_HS)在频域中的RAPSD分布。其中,实线表示所有样本在对应频率处的平均RAPSD,阴影区域表示该频率处样本间的波动范围。三条曲线随着频率的增加整体呈下降趋势,说明图像能量主要集中在低频区域,而高频区域所占能量相对较少,这与图2中展现的低频是图像结构主体信息的表达一致。相比Infrared,Visible在中高频区域具有更高的功率谱密度,表明其包含更丰富的纹理和边缘信息,经过自适应高频抑制处理后,Adaption_HS曲线在低频区域仍保持与可见光曲线相近的分布趋势,而在中高频区域逐渐向红外曲线靠近。该结果表明,自适应高频抑制策略能够在保留可见光图像主要结构信息的同时,降低其过强的高频成分,使处理后的图像在频域分布上更接近红外图像。

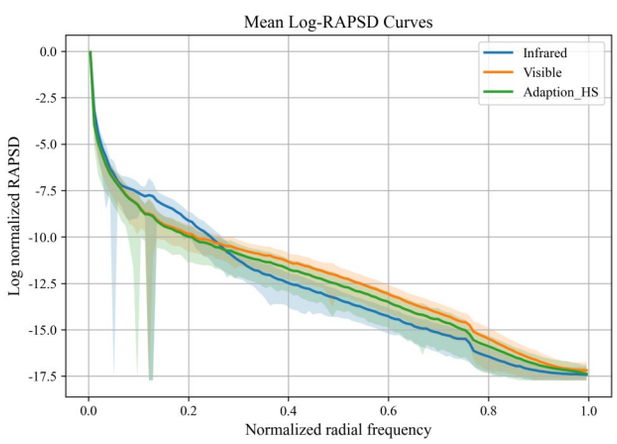


图 10 跨模态频域分布逼近分析图

Figure. 10 Analysis of cross-modal frequency-domain distribution approximation

(3) 不同高频抑制策略的有效性分析

为进一步验证自适应高频抑制策略的有效性,本文首先统计训练集中可见光样本的低频能量分布,并选取低频能量较低、适中和较高的代表性样本进行可视化分析,分别记为$sample_{low}$、$sample_{mid}$、$sample_{high}$,在此基础上,设计三组对比实验:Without_HS(不进行高频抑制)、Fixed_HS(固定高频抑制参数,设置α为1.5、*base*为0.5)以及Adaption_HS(基于低频能量自适应的高频抑制),表2给出了不同策略在红外验证集上的检测结果,图11为可视化结果。

表2 高频自适应抑制对比实验

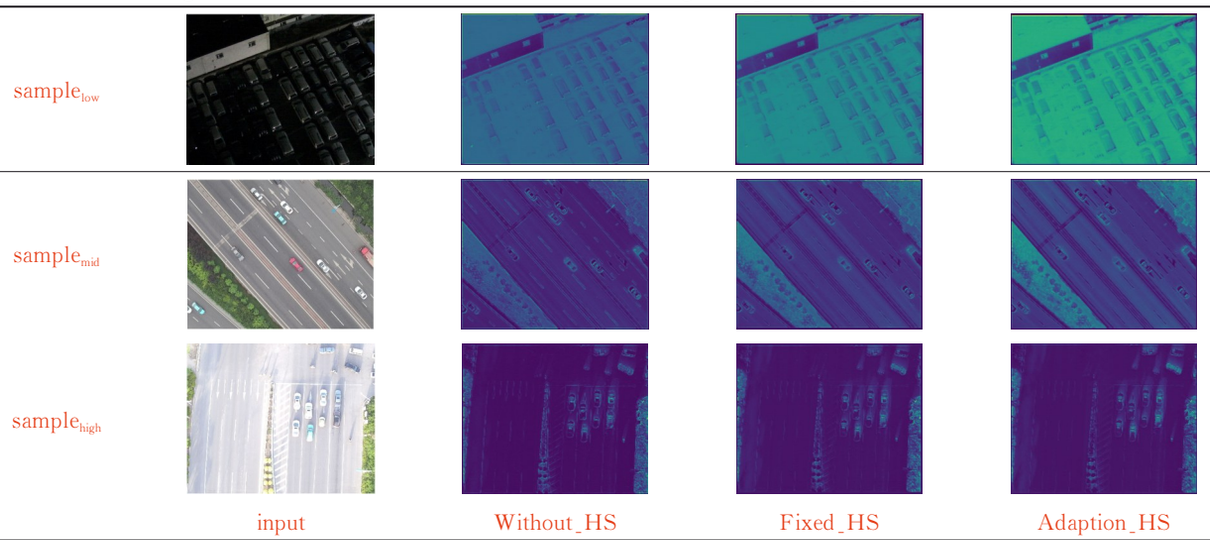
Table 2 Comparative experiments on adaptive high-frequency suppression

Methods	mAP@50/%
Without_HS	72.4
Fixed_HS	76.7
Adaption_HS	79.3

分析可知,Without_HS中图像整体响应较弱,红框内车辆轮廓响应不足,绿框中背景特征干扰明显,相比之下,Fixed_HS和Adaption_HS均能有效抑制冗余高频纹理,使目标轮廓更加清晰,背景响应减弱。表2结果表明,上述三组实验的mAP@50分别为72.4%、76.7%和79.3%,较基线模型YOLOv8n分别提升2.6%、6.9%和9.5%,说明Adaption_HS优于Fixed_HS,更能提升模型在红外域的检测性能。

图 11　不同高频抑制策略下的可视化图

Figure. 11　Visualization results under different high-frequency suppression strategies

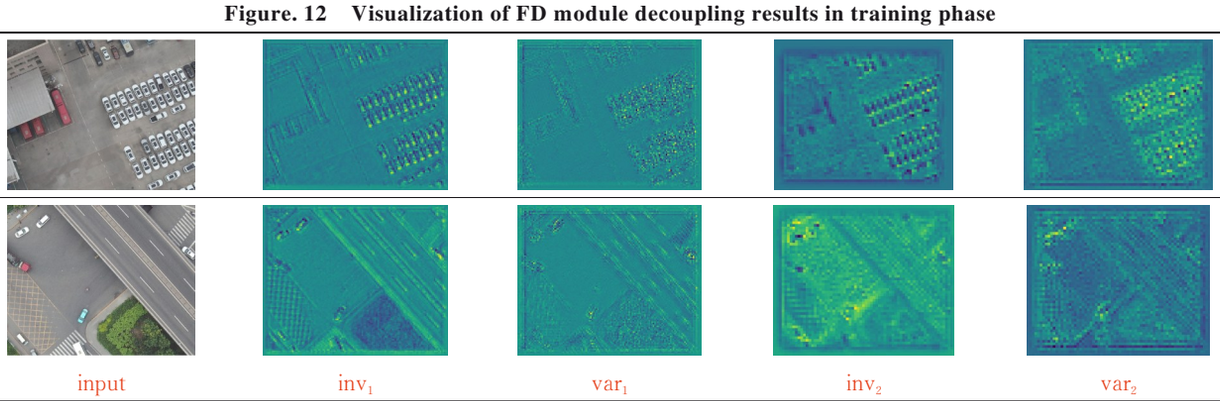


3.1.2　FD模块特征解耦分析

为验证FD模块解耦的有效性，将图1中两个FD模块输出的特征图可视化，结果如图12、13所示，其中input表示原图，inv_1、var_1和inv_2、var_2分别表示经过第1、2个FD模块后得到的域不变和域私有特征图。图12表示使用可见光样本训练时的特征提取，从可视化结果可以看出，域不变特征inv_1和inv_2均较好地保留了目标车辆的轮廓结构及主要语义信息，相比之下，var_1和var_2作为域私有特征，其目标轮廓响应相对较弱，空间结构表征不够清晰。

图 12　训练阶段FD模块解耦结果可视化

Figure. 12　Visualization of FD module decoupling results in training phase



进一步地，本文将红外样本作为输入进行测试，并重点观察FD模块提取的红外域不变特征，结果如图13所示。红框部分是红外目标所在的区域，在相同FD模块层级下，红外样本中提取的域不变特征响应相较于可见光样本略弱，这可能与红外图像纹理细节较少、目标与背景热对比度变化以及跨模态背景差异有关。然而，图13中的inv_1和inv_2仍能在目标区域保持有效响应，说明FD模块能够从红外输入中提取较稳定的共享语义表征。

3.1.3　SO Optimization模块小目标检测性能分析

本文基于FLIR数据集，以YOLOv8n作为基线模型baseline，依次构建仅嵌入SPD_Conv、仅嵌入GA_Concat、同时嵌入SPD_Conv与GA_Concat的三组对比模型，结果如表3所示，目标框可视化结果如图14所示。

与baseline相比，仅添加SPD_Conv、仅添加GA_Concat、添加SPD_Conv和GA_Concat对应的召回率R分别提升6.4%、9.3%、11.2%，mAP@50分别提升2.6%、3.9%、6.2%。从图14可以看出，在baseline条件下，远距离车辆、尺度较小车辆以及受车身遮挡影响的目标更容易出现漏检。SPD_Conv从模型特征保留的

图 13 测试阶段FD模块解耦结果可视化

Figure. 13 Visualization of FD module decoupling results in test phase

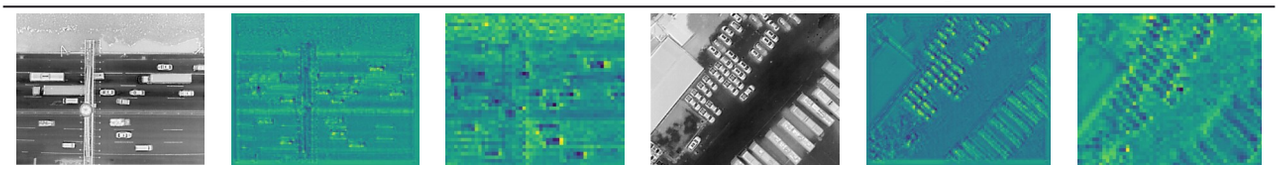


表 3 SO Optimization模块对小目标检测有效性分析

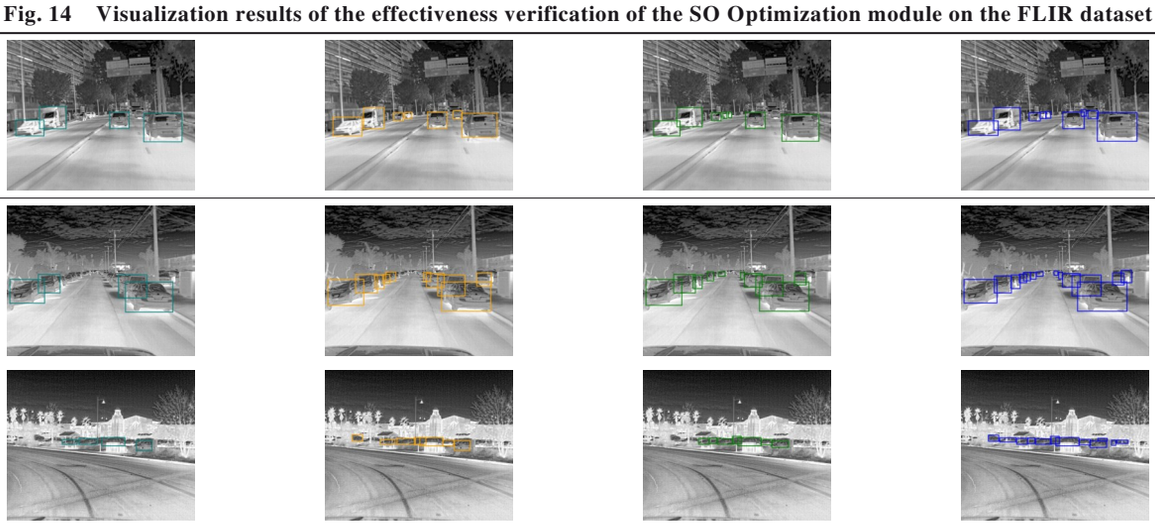
Table 3 Effectiveness analysis of the SO Optimization module for small object detection

Methods	P/%	R/%	mAP@50/%
baseline	50.5	44.2	54.6
baseline+SPD_Conv	51.8	50.6	57.2
baseline+GA_Concat	52.2	53.5	58.3
baseline+SPD_Conv+GA_Concat	56.5	55.4	60.8

角度，将空间信息转化成通道信息，从而避免小目标信息的丢失，而GA_Concat以浅层中丰富的特征引导深层特征对小目标的感知定位，二者结合后，空间细节保留与特征融合增强形成互补，使模型对小目标区域的响应更加充分，漏检现象进一步减少。

图 14 再FLIR数据集上的SO Optimization模块有效性验证可视化结果

Fig. 14 Visualization results of the effectiveness verification of the SO Optimization module on the FLIR dataset



3.2 消融实验

为验证各模块的有效性，本文在DroneVehicle数据集上开展消融实验，结果如表4所示。基线模型Exp0的准确率、召回率和mAP@50分别为71.2%、66.5%和69.8%，加入HSLE模块后，Exp1的准确率、召回率和mAP@50分别提升至81.7%、74.3%和79.3%，较基线模型分别提高10.5%、7.8%和9.5%。仅引入FD模块时，Exp2的准确率、召回率和mAP@50分别为76.4%、71.8%和75.1%，Exp3通过引入SPD_Conv和GA_Concat，取得了78.9%的准确率和76.5%的mAP@50。进一步引入HSLE和FD模块后，Exp4的性能提升至86.3%、78.7%和83.2%，相比Exp1分别提升4.6%、4.4%和3.9%。当同时引入三个改进模块时，Exp7实现最优检测性能，准确率、召回率和mAP@50分别达到88.5%、82.4%和87.2%，较基线模型Exp0提升17.3%、15.9%和17.4%，并较Exp4进一步提高2.2%、3.7%和4.0%。上述结果证明了各个模块的有效性，有利于模型实现跨域目标检测。

表4　基于DroneVehicle数据集的消融实验

Table 4　Ablation experiment on the DroneVehicle dataset

Experiments	HSLE	FD	SO Optimization	P/%	R/%	mAP@50/%
Exp0				71.2	66.5	69.8
Exp1	√			81.7	74.3	79.3
Exp2		√		76.4	71.8	75.1
Exp3			√	78.9	69.2	76.5
Exp4	√	√		86.3	78.7	83.2
Exp5		√	√	82.1	75.4	80.6
Exp6	√		√	84.5	76.9	82.4
Exp7	√	√	√	88.5	82.4	87.2

3.3　对比试验

为验证所提方法的有效性与先进性，本文在统一实验环境下开展对比实验，并选取7种具有代表性的单域泛化对比方法——SW[29]、IBN-Net[30]、ISW[31]和CDSD[8]、G-NAS[32]、OA-DG[33]、PDOC[18]。

3.3.1　场景级泛化性能对比

为评估模型在不同红外场景下的泛化性能，本文分别在DroneVehicle数据集中的val_1、val_2、val_3上进行测试，并以mAP@50作为评价指标，结果见表5。

表5　不同场景下单域泛化目标检测结果

Table 5　Results of single-domain generalized object detection in different scenarios

Methods	val_1/%	val_2/%	val_3/%	average/%
YOLOv8n	72.3	68.7	68.2	69.7
SW[29]	78.2	74.1	73.5	75.3
IBN-Net[30]	76.5	74.9	72.7	74.7
ISW[31]	75.4	70.2	72.3	72.6
CDSD[8]	84.6	81.9	77.2	81.2
G-NAS[32]	80.5	79.1	75.8	78.5
OA-DG[33]	88.6	84.3	81.5	84.8
PDOC[18]	83.2	81.6	75.6	80.1
Ours	90.3	85.4	84.9	86.9

分析可知，本文方法在3类红外验证场景上的平均mAP@50达到86.9%。相较于基线模型YOLOv8n平均提升17.2%，相较于SW、IBN-Net、ISW、CDSD、G-NAS、OA-DG和PDOC等方法，mAP@50分别提升11.6%、12.2%、14.3%、5.7%、8.4%、2.1%和6.8%。从不同场景来看，在路侧停车场景中，车辆目标相对清晰，停放角度差异较小，本文方法取得90.3%的mAP@50，较OA-DG提高1.7%。在商圈停车场景下，各方法的检测结果普遍低于路侧停车场景，本文方法在该场景下取得85.4%的mAP@50，仍高于OA-DG的84.3%。在居民区场景下，PDOC、G-NAS等方法的检测性能下降较明显，而本文方法仍达到84.9%，与商圈停车场景结果接近，说明所提方法在不同场景下具有较好的检测稳定性。

3.3.2　综合检测性能指标对比

为全面评估本文方法与现有主流单域泛化算法的综合性能，本文基于DroneVehicle数据集，在完整红外验证集上进行了对比实验，即取val_1、val_2、val_3的并集，实验结果如表6所示。

从与其他单域泛化方法的比较结果来看，在检测精度上，本文方法相较于SW、IBN-Net、ISW、CDSD、G-NAS、OA-DG和PDOC在mAP@50上分别提升6.8%、10.7%、13.7%、4.6%、8.6%、2.9%和7.2%。在模型复杂度方面，本文方法的参数量仅为3.7M，FLOPs为14.1G。虽然相较于原始YOLOv8n略有增加，但仍显著低于其它对比方法。与此同时，从推理速度来看，本文方法的FPS为36.5，虽然低于YOLOv8n的47.2，但仍明显高于第三名的16.7。这表明，本文方法在引入少量额外计算开销的基础上，实现了检测性能的有效提升，同时依然保持了较快的推理速度，更适合部署于资源受限且对实时性要求较高的边缘设备平台。

表6 不同单域泛化方法的综合性能对比

Table 6 Comprehensive performance comparison of different single-domain generalization methods

Methods	P/%	R/%	mAP@50/%	Params/M	FLOPs/G	FPS
YOLOv8n	71.2	66.5	69.8	3.2	8.7	47.2
SW[29]	77.8	78.4	80.4	41.3	423.7	16.1
IBN-Net[30]	80.5	75.0	76.5	34.9	422.2	16.7
ISW[31]	76.5	69.5	73.5	62.9	393.2	14.4
CDSD[8]	80.1	73.4	82.6	61.1	437.6	13.6
G-NAS[32]	79.7	75.1	78.6	71.0	149.0	13.6
OA-DG[33]	85.1	80.4	84.3	74.6	212.7	10.5
PDOC[18]	84.8	78.2	80.0	87.8	340.6	9.2
Ours	88.5	82.4	87.2	3.7	14.1	36.5

3.3.3 FLIR 数据集泛化性能验证

为验证多视角下的算法泛化性能，在FLIR数据集上执行上述对比实验，结果如表7所示。分析可知，本文方法在FLIR数据集上的准确率、召回率和mAP@50分别为75.3%、67.8%和72.1%，较YOLOv8n分别提升了24.8%、23.6%和17.5%。与其余方法相比，本文方法的mAP@50分别比SW、IBN-Net、ISW、CDSD、G-NAS、OA-DG和PDOC提高16.1%、13.6%、12.5%、11.0%、6.3%、5.0%和5.8%。

表7 不同方法在 FLIR 数据集上的性能对比

Table 7 Performance comparison of different methods on the FLIR dataset

Methods	P/%	R/%	mAP@50/%
YOLOv8n	50.5	44.2	54.6
SW[29]	69.4	45.1	56.0
IBN-Net[30]	65.3	48.4	58.5
ISW[31]	66.2	54.7	59.6
CDSD[8]	69.8	58.5	61.1
G-NAS[32]	70.5	64.7	65.8
OA-DG[33]	72.6	63.5	67.1
PDOC[18]	71.2	60.9	66.3
Ours	75.3	67.8	72.1

3.3.4 检测结果可视化分析

为更加直观地展示所提方法的检测效果，本文在DroneVehicle数据集和FLIR数据集上进行了可视化对比分析，如图15所示。其中，图15的第2列至第4列分别对应DroneVehicle数据集中的路侧停车、商圈停车场和居民区3类红外验证场景，第5列为FLIR数据集上的可视化结果。

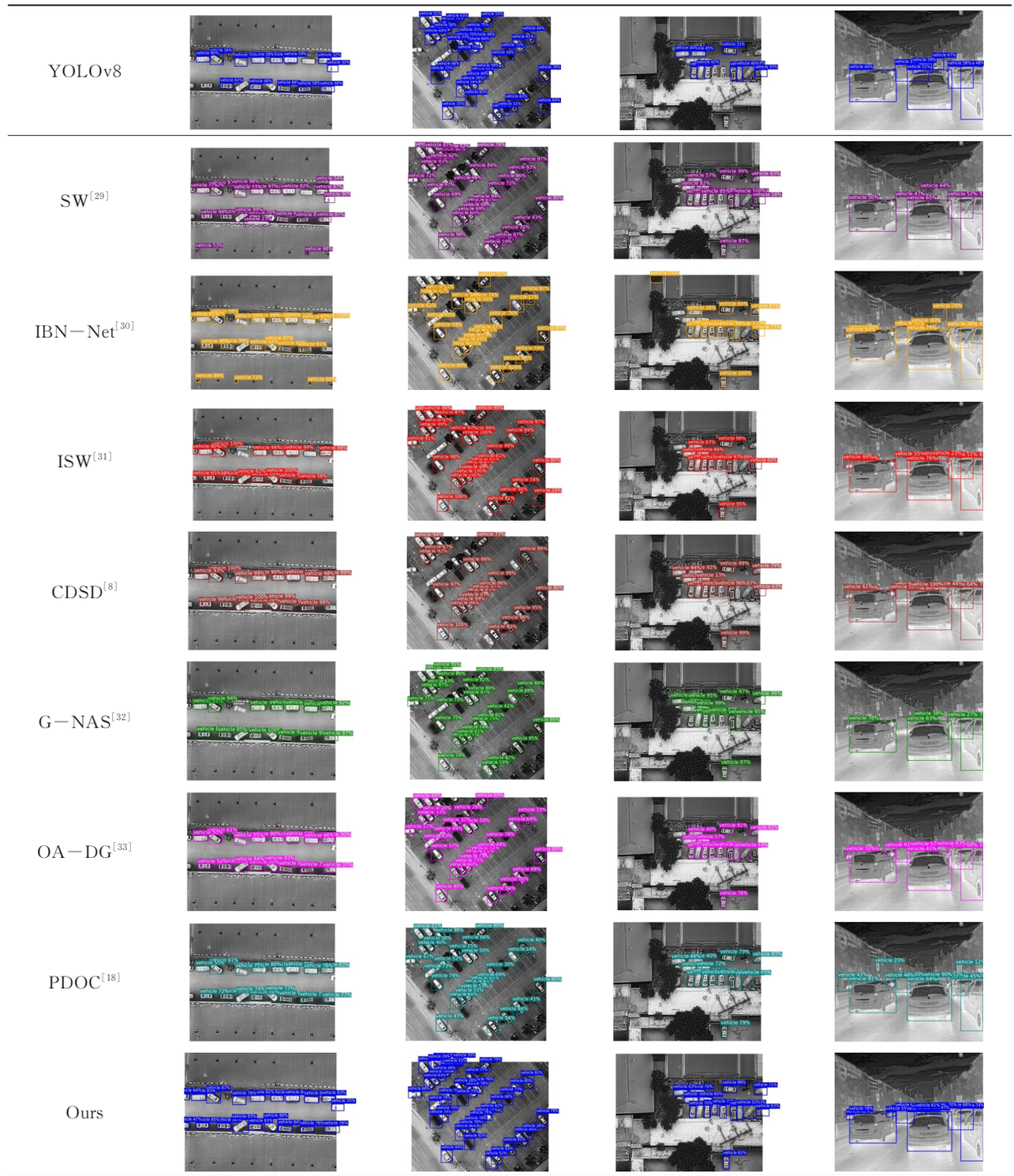
图15显示，现有对比方法虽能在部分显著目标上给出较高检测置信度，但在密集停车、复杂居民区及FLIR数据集的多视角场景中仍存在明显漏检，说明较高的单框置信度并不必然对应更优的整体检测性能。对于跨模态泛化检测任务，若模型仅能检测少量易识别目标，而难以稳定召回远距离小目标、弱纹理目标或复杂背景下的目标，其召回率和mAP@50仍会受到影响。

相比之下，本文方法在部分目标上的置信度虽然低于个别对比方法，但能够检出更多真实目标，尤其是在密集排列、小尺度车辆和复杂背景干扰场景下漏检更少，这表明本文方法并非单纯提高检测框的置信度，而是通过缓解可见光与红外之间的跨模态差异，增强模型对目标区域的稳定感知能力。

故本文方法在可视化中表现为置信度相对平稳但目标召回更充分，这与表6、7中较高的召回率结果一致。进一步观察FLIR数据集可视化结果可以发现，当前视、后视视角变化较大且远距离小目标较多时，对比方法更容易忽略尺度较小的目标，本文方法对这类目标具有更好的检出能力，说明其在多视角红外场景下具有更稳定的泛化性能。

图 15　不同对比方法的可视化结果

Fig. 15　Visualization results of different comparison methods



4　结论

针对跨域目标检测中目标域样本缺失的问题，本文提出了一种基于可见光图像频域特征解耦的零样本红外目标检测方法。该方法通过HSLE模块重构可见光图像，抑制高频纹理干扰并增强低频风格多样性，同时，在主干网络浅层引入FD模块，显式分离域私有特征与域不变特征，最后，利用SPD_Conv与GA_Concat增强小目标特征保留与多尺度特征融合能力，提升模型在跨域红外场景下的检测性能。实验结果表明，本文方法在mAP@50等指标上优于对比算法，展现了良好的跨域泛化能力，同时保持较低的参数量和计算

量,适合移动端部署并满足高时敏目标检测任务需求。

参考文献

[1] ZHENG Guangtao, HUAI Mengdi, ZHANG Aidong, et al. AdvST: revisiting data augmentations for single domain generalization[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2024, 38(19): 21832-21840.

[2] SU Zixian, YAO Kai, YANG Xi, et al. Rethinking data augmentation for single-source domain generalization in medical image segmentation[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2023, 37(2): 2366-2374.

[3] SHI Caijuan, ZHENG Yuanfan, REN Bijuan, et al. Single-domain generalized breast tumor detection in X-ray images[J]. Journal of Image and Graphics, 2024, 29(3): 725-740.
史彩娟, 郑远帆, 任弼娟, 等. 单域泛化X-ray乳腺肿瘤检测[J]. 中国图象图形学报, 2024,29(3):725-740.

[4] ZHANG Chen, LI Yunping, TANG Xin, et al. Review of domain generalization person re-identification based on deep learning[J]. Journal of Kunming University of Science and Technology (Natural Science), 2024, 49(6): 86-99.
张臣, 李云平, 唐鑫, 等. 基于深度学习的域泛化行人重识别综述[J]. 昆明理工大学学报(自然科学版), 2024, 49(6): 86-99.

[5] JIN Xin, LAN Cuiling, ZENG Wenjun, et al. Style normalization and restitution for generalizable person re-identification[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 3143-3152.

[6] ZHENG Kecheng, LIU Jiawei, WU Wei, et al. Calibrated feature decomposition for generalizable person re-identification[J/OL]. arXiv:2111.13945, (2021-11-27)[2026-06-24]. https://doi.org/10.48550/arXiv.2111.13945.

[7] LIU Jiawei, HUANG Zhipeng, LI Liang, et al. Debiased batch normalization via gaussian process for generalizable person re-identification[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022, 36(2): 1729-1737.

[8] WU Aming, DENG Cheng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 847-856.

[9] LIN Chuang, YUAN Zehuan, ZHAO Sicheng, et al. Domain-invariant disentangled network for generalizable object detection[C]. Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2021: 8771-8780.

[10] YU Shijie, ZHU Feng, CHEN Dapeng, et al. Multiple domain experts collaborative learning: multi-source domain generalization for person re-identification[J/OL]. arXiv:2105.12355, (2021-05-26)[2026-06-24]. https://doi.org/10.48550/arXiv.2105.12355.

[11] LI Da, YANG Yongxin, SONG Yizhe, et al. Learning to generalize: meta-learning for domain generalization[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018, 32(1): 3490-3497.

[12] ZHAO Yuyang, ZHONG Zhun, YANG Fengxiang, et al. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 6273-6282.

[13] ZHAO Yuyang, ZHONG Zhun, ZHAO Na, et al. Style-hallucinated dual consistency learning: a unified framework for visual domain generalization[J]. International Journal of Computer Vision, 2024, 132(3): 837-853.

[14] QI Lei, DONG Peng, XIONG Tan, et al. DoubleAUG: single-domain generalized object detector in urban via color perturbation and dual-style memory[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2024, 20(5): 1-20.

[15] FAN Qi, SEGU M, TAI Y W, et al. Towards robust object detection invariant to real-world domain shifts[C]. The Eleventh International Conference on Learning Representations. Kigali: OpenReview, 2023.

[16] LIAO Shengcai, SHAO Ling. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting[J/OL]. arXiv:1904.10424, (2019-04-23)[2026-06-24]. https://doi.org/10.48550/arXiv.1904.10424.

[17] VIDIT V, ENGILBERGE M, SALZMANN M. CLIP the gap: a single domain generalization approach for object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 3219-3229.

[18] DENG Li, WU Aming, WANG Yaowei, et al. Prompt-driven dynamic object-centric learning for single domain generalization[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 17606-17615.

[19] LI Hao, WANG Wei, WANG Cong, et al. Phrase grounding-based style transfer for single-domain generalized object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2026, 36(1): 106-118.

[20] SUNKARA R, LUO Tie. No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects[C]. Machine Learning and Knowledge Discovery in Databases. Cham: Springer Nature Switzerland,

2023: 443-459.
[21] TONG Yunfei, LIU Jing, FU Zhiling, et al. Guided attention and joint loss for infrared dim small target detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-14.
[22] HWANG S, HAN D, JEON M. DG-DETR: toward domain generalized detection transformer[J]. Pattern Recognition Letters, 2026, 199: 128-134.
[23] JEON S, HONG K, LEE P, et al. Feature stylization and domain-aware contrastive learning for domain generalization [C]. Proceedings of the 29th ACM International Conference on Multimedia. New York: Association for Computing Machinery, 2021: 22-31.
[24] GUO Jintao, WANG Na, QI Lei, et al. ALOFT: a lightweight MLP-like architecture with dynamic low-frequency transform for domain generalization[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 24132-24141.
[25] YOO J, UH Y, CHUN S, et al. Photorealistic style transfer via wavelet transforms[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 9035-9044.
[26] YUAN Yuxuan, TANG Luyao, XU Ying, et al. Filling and disentanglement: toward low- and high-order parallel single-domain generalization for SAR ship detection[J]. IEEE Transactions on Aerospace and Electronic Systems, 2025, 61(2): 3668-3682.
[27] HOU Qibin, ZHOU Daquan, FENG Jiashi. Coordinate attention for efficient mobile network design[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 13713-13722.
[28] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.
[29] PAN Xingang, ZHAN Xiaohang, SHI Jianping, et al. Switchable whitening for deep representation learning[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 1863-1871.
[30] PAN Xingang, LUO Ping, SHI Jianping, et al. Two at once: enhancing learning and generalization capacities via IBN-Net[C]. Proceedings of the European Conference on Computer Vision. Cham: Springer, 2018: 464-479.
[31] CHOI S, JUNG S, YUN H, et al. RobustNet: improving domain generalization in urban-scene segmentation via instance selective whitening[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 11575-11585.
[32] WU Fan, GAO Jinling, HONG Lanqing, et al. G-NAS: generalizable neural architecture search for single domain generalization object detection[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2024, 38(6): 5958-5966.
[33] LEE W, HONG D, LIM H, et al. Object-aware domain generalization for object detection[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2024, 38(4): 2947-2955.

Zero-Shot Infrared Object Detection via Frequency-Domain Feature Decoupling of Visible Images

LI Jiale, HUI Bingwei, WANG Kaiyue, ZHANG Jun
(*College of Electronic Science, National University of Defense Technology, Changsha* 410028, *China*)

Abstract: Infrared imaging systems are widely used in battlefield reconnaissance, disaster warning, autonomous driving, and industrial security because of their all-weather sensing capability and robustness to illumination variations. However, in non-cooperative environments, infrared samples and annotations are often difficult or costly to obtain, making zero-shot infrared object detection highly challenging. Traditional domain adaptation and fine-tuning methods usually require target-domain samples, which limits their applicability in infrared scenarios. Single-domain generalization (SDG), which learns from only one source domain and does not rely on target-domain data, provides a feasible solution. Existing SDG methods mainly focus on generic visual feature generalization or cross-modal semantic alignment, but they rarely exploit the frequency characteristics of infrared imaging and are often insufficient for small-target detection under cross-modal domain shifts. Moreover, few SDG methods are specifically designed for the lightweight YOLO architecture required by real-time applications. To address the above challenges, this paper proposes a zero-shot infrared object detection method based on visible-image frequency-domain feature decoupling. First, a High-Frequency Suppression and Low-Frequency Enhancement (HSLE)

module is designed to process visible images in the frequency domain. High-frequency components are adaptively suppressed to reduce source-domain-specific texture interference, while low-frequency statistics are perturbed to increase source-domain diversity and to narrow the visible-to-infrared gap without severely damaging semantic structure. During training, a weight-shared dual-branch backbone is adopted, where the original visible image is used as the primary branch and the HSLE-enhanced image is used as the auxiliary branch. A Feature Decoupling (FD) module is embedded in shallow backbone stages to separate domain-invariant features from domain-private features, so that deeper layers focus more on cross-spectral common semantics rather than source-style cues. To reduce the loss of weak target information, conventional downsampling is replaced by Space-to-Depth Convolution (SPD_Conv), which preserves more fine-grained details during feature transformation. In addition, a Guided Attention Concatenation (GA_Concat) module is introduced in the neck to enhance multi-scale feature fusion by using shallow detail-rich features to guide deep semantic aggregation. During inference, the auxiliary branch is removed, and the detector remains a single-branch structure, thereby preserving deployment efficiency. Experiments are conducted on the DroneVehicle dataset under a strict zero-shot visible-to-infrared protocol, where only labeled visible images are used for training and infrared images are used only for validation. To ensure a rigorous evaluation of scene-level generalization, the visible training set and infrared validation set are constructed with scene-level non-overlap. On the infrared validation set, the proposed method achieves an mAP@50 of 87.2%, which is 17.4 percentage points higher than YOLOv8n. Compared with SW, IBN-Net, ISW, CDSD, G-NAS, OA-DG, and PDOC, the proposed method improves mAP@50 by 6.8, 10.7, 13.7, 4.6, 8.6, 2.9, and 7.2 percentage points, respectively. Meanwhile, the detector remains lightweight, with only 3.7 M parameters and 14.1 GFLOPs. To further evaluate multi-view infrared generalization, experiments are also conducted on the FLIR dataset. The proposed method achieves an mAP@50 of 72.1%, outperforming YOLOv8n by 17.5 percentage points and surpassing SW, IBN-Net, ISW, CDSD, G-NAS, OA-DG, and PDOC by 16.1, 13.6, 12.5, 11.0, 6.3, 5.0, and 5.8 percentage points, respectively. Ablation studies verify the effectiveness of HSLE, FD, and the small-object detection components. Qualitative results further show that the proposed method reduces missed detections and false alarms in cluttered infrared scenes. These results demonstrate that the proposed method improves zero-shot visible-to-infrared object detection and small infrared target detection while maintaining the efficiency of a lightweight one-stage detector.

Key words: Cross-domain generalization; Zero-shot infrared object detection; Frequency-domain enhancement; Feature decoupling; Small-object detection

OCIS Codes: 100.4996; 100.2000; 110.2970; 150.1135

CSTR: 32255.14.gzxb20265508.0810001