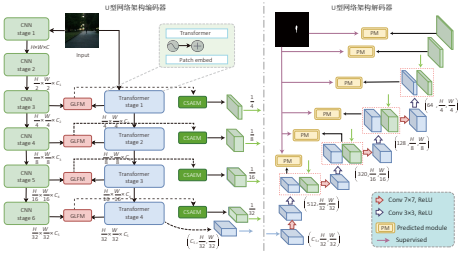
OEE | 光电工程 Opto-Electronic Engineering

Article

2026 年 3 月, 第 53 卷, 第 3 期

一种低光图像显著性目标检测的 CNN-Transformer 混合模型



张　航, 魏英姿*, 于聚壮

沈阳理工大学信息科学与工程学院, 辽宁 沈阳 110159

摘要：针对低光照环境下显著性目标检测面临的图像质量退化、暗区噪声及低对比度等挑战，提出一种 CNN-Transformer 双流混合编码器-解码器模型 (dual-stream hybrid network, DSHNet)。采用非对称双分支架构，以融合 CNN 多尺度局部特征提取能力与 Transformer 全局建模优势。首先，采用多层次特征提取策略分别从 CNN 和 Transformer 模块中提取特征，通过全局-局部特征协同建模，将同尺度不同分支特征两两交叉递阶完成双向融合。其次，设计通道与空间特征增强模块，经过特征变换和通道压缩优化，再做多分支特征融合。最后，采用 Focal Loss+Dice Loss 联合损失函数聚焦难样本和优化区域重叠，解决边界模糊的问题。在自建复杂目标数据集 LISOD 及标准数据集上进行实验验证，DSHNet 较同类混合模型 ABiUNet 参数量降低 23%，在 LISOD 数据集上 F-measure、S-measure 和加权 F-measure 分别达到 91.2%、83.7% 和 87.2%，可有效应用于低光图像显著性检测任务。

关键词：显著性检测；低光图像；CNN；Transformer 模型；混合架构

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oee.2026.250269 | CSTR: 32245.14.oee.2026.250269 | 中图分类号：TP391.4 | 文献标识码：A

张航，魏英姿，于聚壮. 一种低光图像显著性目标检测的 CNN-Transformer 混合模型 [J]. 光电工程，2026，**53**(3): 250269

Zhang H, Wei Y Z, Yu J Z. A low-light image saliency detection method based on a CNN-Transformer hybrid architecture[J]. *Opto-Electron Eng*, 2026, **53**(3): 250269

1　引　言

低光场景下的显著目标检测 (salient object detection, SOD) 在自动驾驶、安防监控、夜间摄影等领域中具有广泛应用价值[1-2]。现有的单模态显著性目标检测算法主要关注模型设计中的特征集成，未能综合分析图像本身所蕴含的特征信息[3-4]。双模态 (RGB-D 或 RGB-T) SOD 模型融合热成像数据和深度信息，在低光图像中的检测性能有所增强，但引入新的模态数据往往会增加软、硬件成本[5-6]。因此，低光场景单模态显著性检测研究仍具有广泛应用价值。

在低光照条件下实现 SOD 高精度检测，不仅需要捕捉暗区中极易丢失的微弱边界信息，还需建立全局光照分布的时空关联性。文献 [7] 提出了一种无源领域自适应 (source-free domain adaptation, SFDA) 方法，将正常光照图像源域下训练的模型知识迁移至低光照条件图像目标域。文献 [8] 提出了一种自适应前景聚焦算法，采用分离-学习-融合策略，加强对前景焦点特征的提取能力和对背景细节信息的保留能力，改善图像前景目标尺度差异大导致的漏检问题。文献 [9] 为丰富图像边缘细节信息，利用图像的像素值分布进行自适应分段线性变换，提升图像细节信息。文献 [10] 利用图像自适应增强网络与传统目标检测网络相结合，减少低光目标检测过程中的漏检现象。文献 [11] 利用特征残差块 (feature

收稿日期：2025-09-08；修回日期：2025-11-22；录用日期：2025-11-26；网络出版日期：2026-03-19

基金项目：辽宁省自然科学基金计划机器人学国家重点实验室联合开放基金 (2022-KF-12-08)

*通信作者：魏英姿，wings_syit@126.com

版权所有©2026 中国科学院光电技术研究所

residual block, FRB) 和特征金字塔结构提取语义特征，同时利用联合序列和空间的自注意力机制实现对焦堆栈图像序列特征的增强与融合。文献 [12] 使用聚焦模块和定位模块来准确定位和增强不同层次的特征表达式，通过边缘细化模型进一步生成清晰的边界线索和最终的精细显著性图。卷积神经网络 (Convolutional neural network, CNN) 固有的局部感受野特性只能提取局部邻域特征，导致无法准确区分暗区中的物体轮廓。Transformer 在全局关系建模方面的优势，有利于低光图像中显著目标的精确定位。文献 [13] 构建了一个层次式的自适应交互超像素引导的 Transformer 去噪网络，通过超像素分割引导的方式，提升窗口内局部特征的表达力，在通道维度上增强了空间特征的关联性。文献 [14] 通过自适应特征融合级联 Transformer 增强模型对细节的捕捉能力。文献 [15] 提出 LFSepNet，利用 Transformer 的自注意力机制进行全局建模，更全面地整合低光图像中的信息。文献 [16] 将 Transformer 和 DenseNet 相结合组成双分支网络，提高模型对细小目标的感知能力。文献 [17] 通过将 Transformers 集成到 CNN 架构中，来模拟人类视觉特征，有效缓解 CNN 在低光条件下易受局部噪声干扰的问题。文献 [18] 提出了一种双分支多尺度融合的 Encoder-Decoder 网络结构，利用多尺度融合模块实现多目标语义分割，解决目标遮挡和重叠问题。文献 [19] 用混合级联 CNN-Transformer 架构作为特征提取的编码器，充分利用了两者在捕获局部细节和建模全局依赖关系方面的优势，但仍存在特征融合效率低的问题。

在低光图像中，目标边界因噪声和低对比度变得模糊，传统 CNN 的局部感受野难以捕捉微弱边缘信息，而全局建模 Transformer 可弥补这一缺陷，突破局部感受野限制是提升低光检测性能的关键[20]。将 CNN 和 Transformer 模型的双分支特征进行融合后，仍存在较多问题，如全局建模能力与计算效率相互影响，图像全局信息与局部细节特征难以兼顾，多尺度特征融合与暗光噪声相互重叠。因此，本文利用 Transformer 全局建模能力和 CNN 表示学习能力，设计用于低光图像检测的双边非对称编码器-解码器的轻量化结构。具体思路如下：

1) 研究采用 Transformer 作为主干网络进行全局语义上下文信息提取，同时采用轻量级 CNN 作为局部编码器，利用全局-局部特征融合策略实现局部特征和全局信息的结合，解决低光条件下目标模糊、边界不清的问题。

2) 重新构建编码器，采用 Swin Transformer 滑动窗口设计，移除固定大小的位置编码，并在每级 Transformer 特征上添加空间压缩注意力模块进行特征变换和空间缩减，以此降低模型的参数量和计算复杂度。

3) 设计通道与空间注意力增强模块，利用自适应通道加权，在光照不均或图像较暗时，抑制背景噪声，增强对有效特征的关注，以提高低光检测性能鲁棒性。

4) 在解码阶段，采用两阶段层次式设计，利用多尺度特征进行特征提取和解码，通过逐层融合与插值上采样，缓解低光图像下边缘模糊的问题，以增强结构细节还原与语义一致性。

2 CNN-Transformer 混合模型轻量化网络设计

在融合 CNN 和 Transformer 特征方面，现有研究通常采用编码器-解码器架构，引入 CNN 编码路径作为 Transformer 编码路径的补充，可以缓解 Transformer 过度关注全局特征的缺陷，为显著性检测任务带来了新的生机，但这种方法并不适用于复杂低光场景。低照度环境图片多个目标显著性具有复杂性和多变性，图像全局信息与局部细节特征难以兼顾，多尺度特征融合与暗光噪声相互重叠，全局建模能力与计算效率相互影响仍是待解决的问题。

对照非对称双边 U-net 模型 (ABiU-Net) 网络结构[20]，本文提出双流混合网络 DSHNet 模型双边非对称编码器-解码器轻量化结构。与已有编码器-解码器架构对比，图 1 为不同编码器-解码器架构。其中，图 1(a) 为基本的 Hyper column 架构，其通过多尺度特征拼接增强细节，但计算冗余大[21]；图 1(b) 为 U-shape 架构，其通过跳跃连接保留空间信息，但全局建模能力弱[22]；图 1(c) 为深度监督显著性检测网络 (deeply supervised salient object detection Network，DSS) 架构[23]，其在局部特征学习和细节保留上表现突出，但在计算效率与全局上下文建模上难以平衡性能与实用性；图 1(d) 为 ABiU-Net 双边特征融合编码器-解码器U-Net 架构，其在处理图像时采用固定到小的位置编码，导致处理多尺度图像时不灵活；图 1(e) 为本文提出的 DSHNet 双边非对称编码器-解码器架构，与 ABiU-Net 相比，采用轻量化的 Transformer 编码架构，使不同分支特征递阶执行双向融合，解决了 ABiU-Net 处理多尺度图像时不灵活的问题，适应多种尺度的图像。

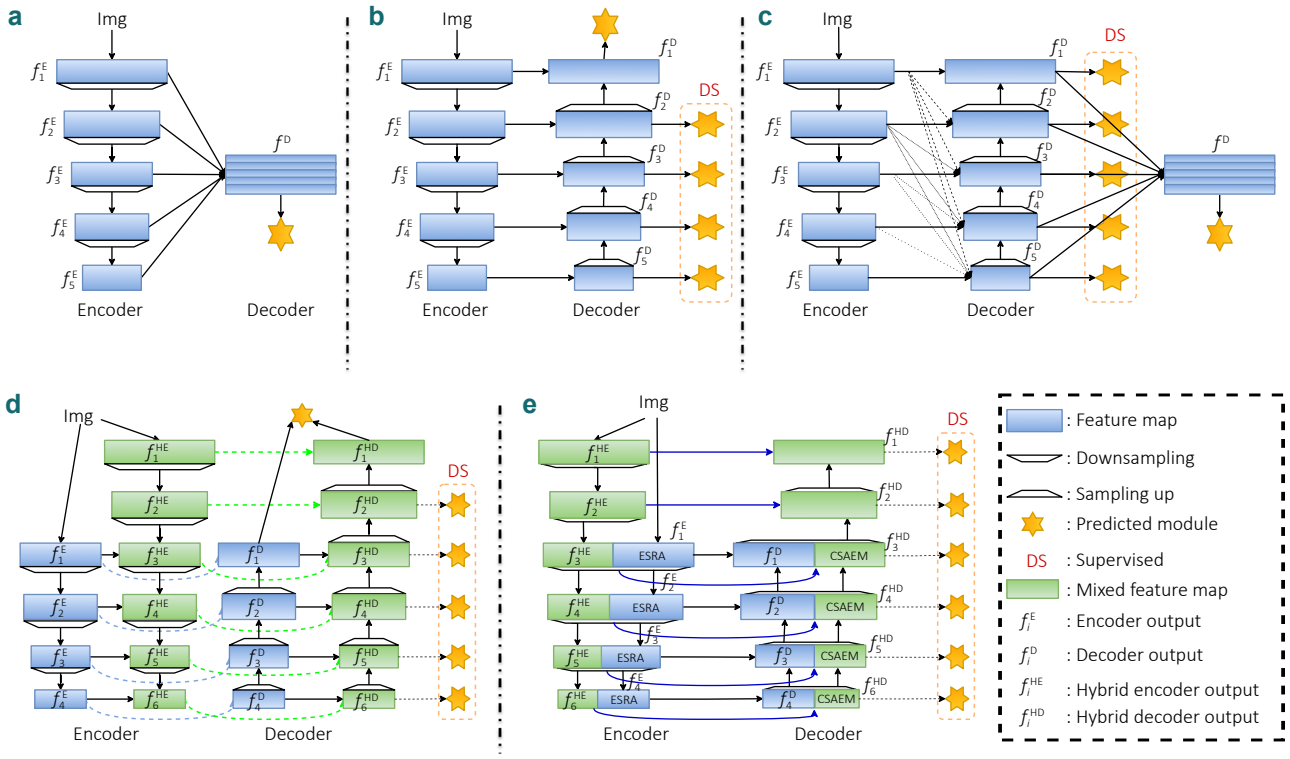


图 1 不同编码器-解码器架构示意图. (a) Hypercolumn; (b) U-Shape; (c) DSS; (d) Encoder; (e) Ours

Fig. 1 Illustration of various encoder-decoder architectures. (a) Hypercolumn; (b) U-Shape; (c) DSS; (d) Encoder; (e) Ours

3 低光图像 SOD 检测的双流混合网络 DSHNet

在低光图像显著性检测任务中，由于特征退化，使图像特征上下文相关性很难被感知，极易丢失微弱边界信息。因此，本文在经典骨干网络 Transformer 模型中，通过自注意力机制建模全局特征关联，并利用堆叠式多头注意力突破低光图像中目标像素间长距离依赖的建模瓶颈，解决光照不足导致的全局语义断裂问题。采用 CNN 多种层次的卷积核，提取低光图像中的局部特征。通过浅层卷积运算抑制光照不均匀引起的噪声干扰，深层卷积多尺度感受野则用于捕获细粒度纹理特征。在 CNN-Transformer 混合模型双分支特征提取框架下，构造双流混合网络 DSHNet，主要由混合编码器、特征增强模块、解码器、显著性预测头四部分组成，DSHNet 网络模块如图 2 所示，其中：W、H、C 分别表示特征图的宽度、高度以及通道数。为解决多尺度特征融合过程中信息损失问题，使用全局-局部特征融合策略 (global-local fusion module, GLFM)[24] 接收不同分支的特征作为输入，通过其内部卷积、归一化层和激活函数，实现高效特征融合。同时高效空间压缩注意力模块 (efficient spatial reduction attention, ESRA)[25] 被逐级应用于 Transformer 特征，通过特征变换与空间降维操作提升计算效率。在解码阶段，通道与空间注意力增强模块 (channel spatial attention enhancement module, CSAEM) 强化暗区边缘特征后，将处理后的全局特征和局部特征融合送入显著性预测头 (predictor module, PM)，再利用深度监督策略进行最后预测。

3.1 构建混合编码器结构

利用 CNN 和 Transformer 相结合的方法，重新构建编码器。在编码阶段，首先利用四级金字塔视觉转换器 (pyramid vision transformer, PVT)[26] 作为特征提取器，通过多个自注意力层探索目标空间关系。其次，引入 6 级轻量化 CNN 编码路径，作为局部特征编码器，利用深度可分离卷积与空间注意力构建特征提取单元，通过动态通道校准增强模块，使低照度区域的特征产生响应，避免牺牲局部相关性。最后，在编码器的四个特征尺度

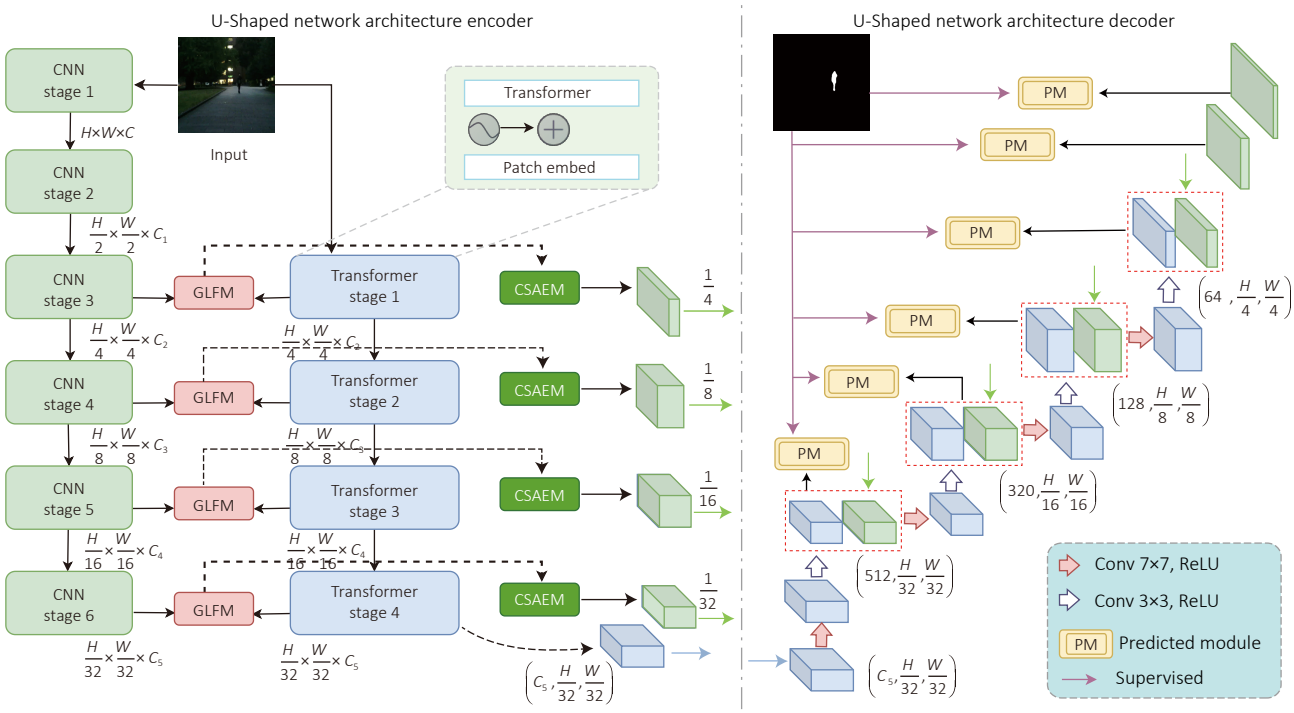


图 2 DSHNet 网络结构

Fig. 2 Structure of DSHNet network

上建立双向融合机制，使信息充分交流，将同尺度 Transformer 特征与 CNN 特征进行双流融合，图 3 为全局-局部特征融合结构。

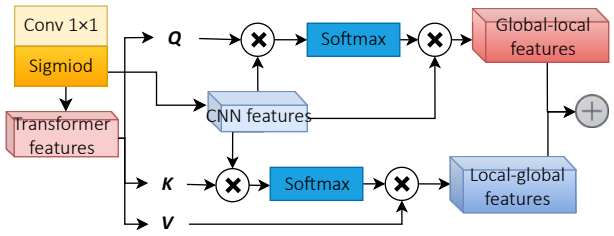


图 3 全局-局部特征融合结构

Fig. 3 Global-local feature fusion module

在预处理时，对双分支特征，各自进行 1×1 卷积和 Sigmoid 归一化，生成标准化特征。然后，Transformer 特征经卷积生成查询 (query, $\boldsymbol{Q}$)，将 CNN 特征同时作为键 (key, $\boldsymbol{K}$) 和值 (value, $\boldsymbol{V}$)，然后作点积运算 $\boldsymbol{Q}\cdot\boldsymbol{K}$，计算注意力权重，再对 $\boldsymbol{V}$ 进行加权融合，实现全局信息对局部特征的调制。类似的，将 CNN 特征作为查询，将 Transformer 经卷积生成的特征作为键和值，实现局部信息到全局信息的交流。最后，同步执行上述双路径交互后，通过 1×1 卷积整合特征，最终输出融合全局-局部信息的增强表征。

3.1.1 空间压缩注意力模块 ESRA

由于低光场景存在低光噪声引发的特征污染，暗光区域常常出现漏检以及背景混淆等问题。在低光条件下，如果直接堆叠轻量级卷积进行局部特征提取，会导致高分辨率特征映射出现严重的计算冗余与几何失真。如果沿通道维度拼接 CNN 与 Transformer 的输出特征，会因特征流差异引发特征空间错位与语义冗余。为解决上述问题，利用 ESRA 模块在每级 Transformer 特征上执行特征变换和空间缩减，图 4 为加入 ESRA 后的 PVT 网络的各阶段 stage i 结构。ESRA 通过使用卷积操作压缩多头自注意力 (multi-head self-attention, MHSA) 中的键和值，可大幅降低模型参数量和计算复杂度，缓解过拟合问题。ESRA 与自注意力的区别在于，ESRA 增加了键值压缩运算和 DropKey 操作。

输入特征图 $\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$ 经过卷积操作，可将 MHSA 中的键和值压缩到更小的维度，从而减少参数量，实现轻量化计算。DropKey 机制对注意力权重进行随机丢弃，从而约束注意力分布，避免过度关注局部区域。查询 Query 保持原始分辨率，即 $\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_Q$，其中 $\boldsymbol{W}_Q\in\mathbb{R}^{C\times d}$，$d$ 代表 $\boldsymbol{Q}$ 的通道维度。键和值做增强降采样，$\boldsymbol{K}=\boldsymbol{X}'\boldsymbol{W}_K$，$\boldsymbol{V}=$

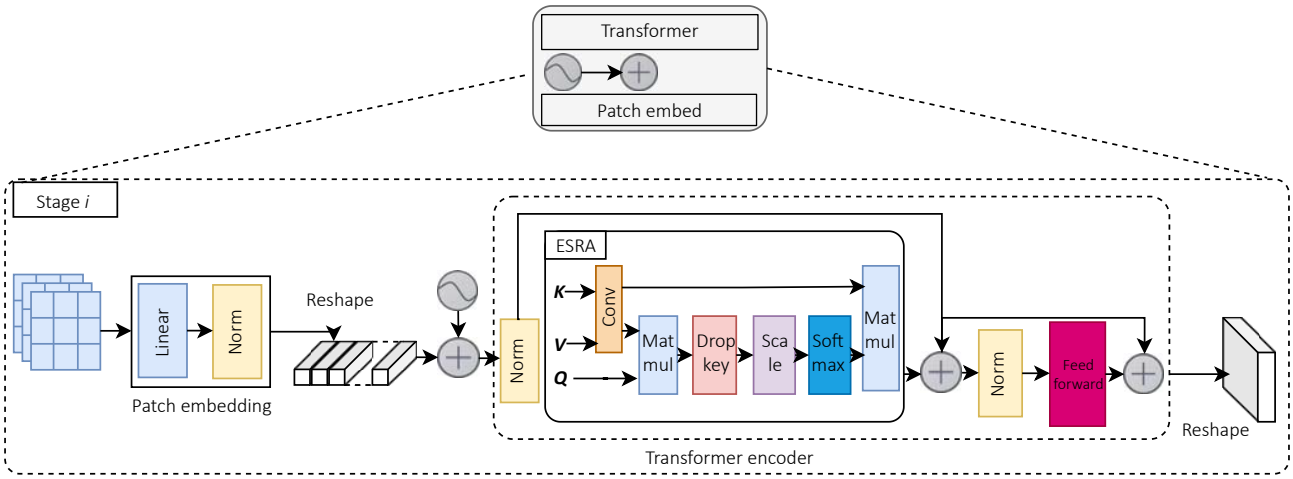


图 4 加入 ESRA 的 PVT 网络的各阶段结构

Fig. 4 The stage *i* structure of the PVT network by incorporating ESRA

$\boldsymbol{X}'\boldsymbol{W}_V$，将 $\boldsymbol{X}$ 经过一次卷积后，降采样后的特征表示为 $\boldsymbol{X}' \in \mathbb{R}^{B\times N'\times C}$，其中 B 表示批次大小，C 代表特征图的通道数，$N'=N/s^2$，s 为降采样因子。计算注意力后，再通过线性变换恢复输出的通道维度，其过程可表示为

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}, \tag{1}$$

$$Output = Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})\boldsymbol{W}_{\mathrm{O}}. \tag{2}$$

3.1.2 重叠补丁和卷积嵌入前馈网络

图像数据属于非重叠块序列，如果使用不同模型对图像子块直接处理会在一定程度上破坏图像的局部连续性，影响暗区微弱信号的连续性检测。低光图像中光照分布不均匀，固定位置编码可能引入噪声。需要利用不同模型分支的功能优势，同时维持性能和计算复杂度之间的平衡，参考 Swin Transformer[27]，采用重叠补丁嵌入 (overlapping patch embedding) 策略对图像进行序列化，按 50% 重叠率进行窗口划分，利用零填充位置编码再作卷积运算，摒弃固定大小的位置编码，增强局部特征连续性，避免固定编码对暗区特征的稀释效应。如图 5 所示，在前馈网络中的第一个全连接层和激活函数之间添加填充大小为 1 的 3×3 深度卷积 DwConv，为重叠补丁嵌入模块示意图。在保持全局优秀建模的同时，提升低光场景下的局部特征保真度、特征融合质量和计算效率，以实现性能-复杂度的优化平衡。

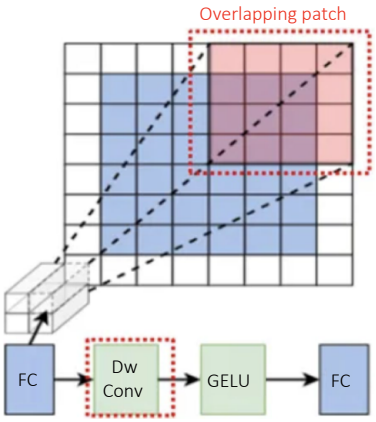


图 5 重叠补丁和卷积运算嵌入前馈网络

Fig. 5 Feedforward network embedded by overlapping patch and DwConv

3.2 双边解码器结构

采用多层级监督与噪声抑制策略，设计双边解码器架构，其中 Transformer 解码器分支利用全局语义关联特征解码显著物体的位置信息，CNN 解码器则专注于精细化复原暗部纹理与局部对比度。粗粒度解码分支 (Decoder1) 以 Transformer 编码特征为引导划分为四个阶段，通过特征金字塔 (feature pyramid network，FPN) 上采样逐步恢复空间分辨率。每阶段通过 (双线性上采样+跳跃连接+混合注意力融合) 逐步恢复空间细节，最终生成粗粒度显著图作为中间监督。细粒度解码分支 (Decoder2) 以 CNN 编码特征为输入，采用 6 级渐进式解码架构，融合当前层解码特征与 Decoder1 对应阶段的中间特征实现多尺度特征聚合。

引入深度监督策略，在解码器的 6 个特征阶段部署显著预测头 (PM)，如图 6 所示为 PM 显著目标预测模块，该模块主要由一系列 1×1、3×3 和 1×1 的卷积层构成，

并引入 Sigmoid 激活函数以增强特征表示。在经过双分支网络 (包含全局和局部特征) 处理后，得到由多个通道与空间注意力增强模块 (channel and spatial attention enhanced module, CSAEM) 输出特征强化的特征图，以此实现从粗到细的梯度反向传播，强化暗区目标的边界约束。

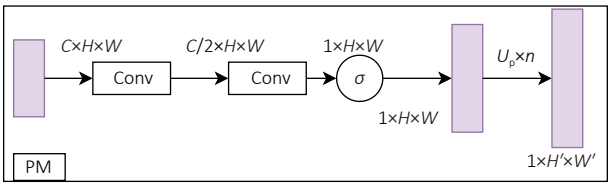


图 6 显著目标预测模块

Fig. 6 Salient target prediction module

3.2.1 通道与空间注意力增强模块 CSAEM

传统卷积神经网络在处理光照不均匀图像时，受不同尺度、形状和方向特征影响，容易陷入过度关注局部信息，导致特征辨别力降低。考虑到低光图像的亮度降低发生在原始 RGB 空间中，为提升特征表达能力，设计 CSAEM 双分支模块进行增强，图 7 为双分支增强模块。通过动态加权融合通道-空间注意力 (channel-spatial attention, CSA) 和通道重标定 (squeeze-and-excitation, SE)[28]，利用 SE 增强高频特征通道，然后再结合 CSA 实现对边缘结构的显式感知，加权系数 α 通过 1×1 卷积层学习，使模型自适应平衡通道与空间注意力。

设输入特征图为 $\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$，形状为 (B,C,H,W)，先通过卷积层提取特征，可表达为

$$\boldsymbol{F}=\sigma(P*\boldsymbol{X}+b), \tag{3}$$

式中：P 和 b 分别表示卷积核权重和偏置；$*$表示卷积操作；$\sigma(\cdot)$表示 ReLU 激活函数。输入特征图同时流经两个并联分支，其中 CSA 分支结合平均池化和最大池化特征进行通道与空间注意力建模，使用注意力加权抑制无关通道，通过 Sigmoid 函数得到每个空间位置权重系数。过程可表示为

$$\boldsymbol{M}_{\mathrm{c}}=\sigma(MLP(GAP(\boldsymbol{F}))+MLP(GMP(\boldsymbol{F}))), \tag{4}$$

$$\boldsymbol{M}_{\mathrm{s}}=\sigma(C(Cat[Avg(\boldsymbol{F}_{\mathrm{c}}),Max(\boldsymbol{F}_{\mathrm{c}})])), \tag{5}$$

$$\boldsymbol{F}_{\mathrm{csa}}=\boldsymbol{M}_{\mathrm{s}}\otimes(\boldsymbol{M}_{\mathrm{c}}\otimes\boldsymbol{F}), \tag{6}$$

式中：σ为 Sigmoid 函数；$\otimes$表示按通道逐元素乘法；

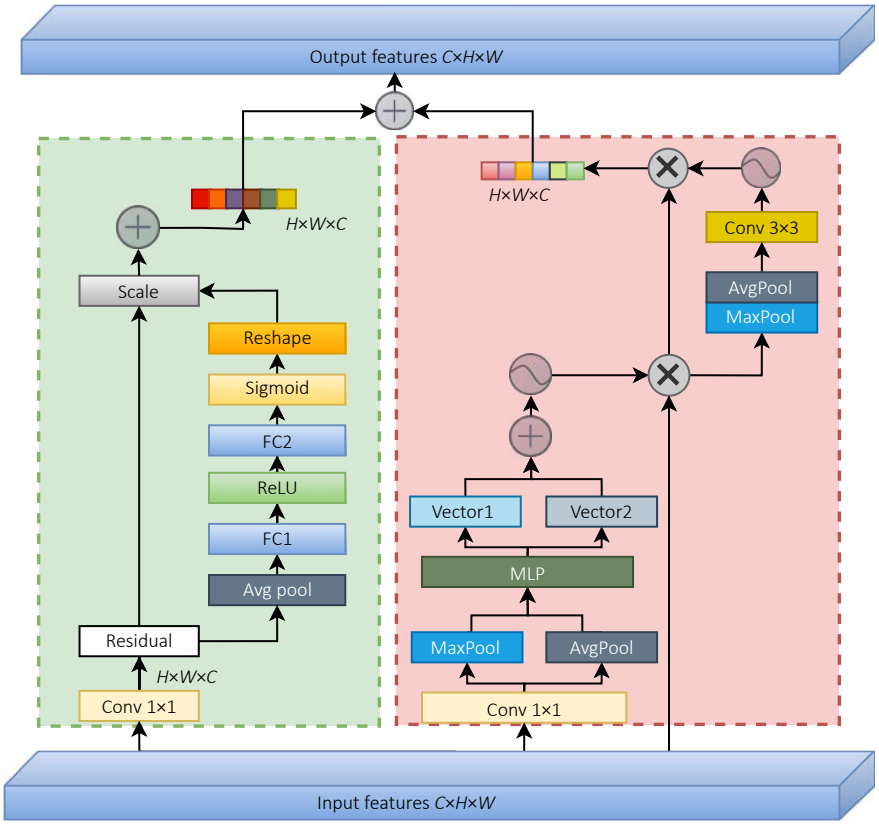


图 7 双分支增强模块

Fig. 7 Double-branched enhanced module

M_c、M_s分别为通道和空间注意力输出。SE 分支对每个通道进行压缩，后经一个带激活的双层连接网络 MLP，实现通道维度的自适应特征重标定，此过程可表示为

$$\boldsymbol{M}_{se}=\sigma(\boldsymbol{W}_2\cdot\delta(\boldsymbol{W}_1\cdot GAP(\boldsymbol{F}))), \tag{7}$$

$$\boldsymbol{F}_{se}=\boldsymbol{M}_{se}\otimes\boldsymbol{F}, \tag{8}$$

式中：$\boldsymbol{W}_1\in\mathbb{R}^{\frac{C}{r}\times C},\boldsymbol{W}_2\in\mathbb{R}^{C\times\frac{C}{r}}$ 是两个全连接层的权重，r 为通道压缩率；δ表示 MLP 层中 ReLU 激活函数；权重向量$\boldsymbol{M}_{se}$表示每个通道的重要性。SE 模块通过显示通道依赖建模增强前景目标相关通道的相应强度，CSA 通过双池化策略保证通道敏感性。随后利用融合平均池化与最大池化后的空间权重图突出显著性目标的高响应区域，与 SE 的通道优化互补，实现对边缘结构的显式感知，可表示为

$$\boldsymbol{F}_{attn}=\alpha\cdot\boldsymbol{F}_{se}+(1-\alpha)\cdot\boldsymbol{F}_{csa}, \tag{9}$$

式中：α代表加权融合系数。空间-通道注意力通过动态加权强化低光目标区域的特征响应，在低对比度条件下增强显著目标的特征辨别能力，通道重标定通过抑制噪声通道、增强高频特征通道，提升边缘区域的显著性响应。

3.3 联合损失函数

由于低光场景中前景与背景类别严重不平衡，使用标准的二分类交叉熵损失函数会因负样本背景部分主导梯度更新，模型预测偏向背景区域，使分割效果较差。为此，使用 Focal Loss 和 Dice Loss 联合损失函数。Focal Loss 通过聚焦于难分类样本 (如边缘、弱对比区域) 来增强模型对关键区域的关注，Dice Loss 则通过优化预测结果与真实标签的重叠度，提升整体显著区域的结构完整性。计算联合损失函数L_{loss}、L_F、L_{FL}、L_{Dice}分别为

$$L_F=(y_n\cdot\ln(x_n)+(1-y_n)\cdot\ln(1-x_n)), \tag{10}$$

$$L_{FL}=-\frac{1}{N}\sum_{n=1}^{N}[(1-p_t)^{\gamma}\cdot L_F], \tag{11}$$

$$L_{Dice}=1-\frac{2\times|Pred\cap Label|}{|Pred\cup Label|}, \tag{12}$$

$$L_{loss}=\alpha\cdot L_{FL}+(1-\alpha)\cdot L_{Dice}, \tag{13}$$

式中：y_n为真实值，$y_n=\{0,1\}$；x_n表示预测结果，$x_n=\{0,1\}$；p_t是模型对真实类别的预测概率；γ为调节因子；*Pred*为 DSHNet 网络的预测图像；*Label* 为真值图像。

在低光场景中，显著目标边缘因噪声和对比度低成为难样本，Focal Loss 通过调节因子$(1-p_t)^{\gamma}$对难易样本赋予不同权重，提升边界分割精度。Dice Loss 对孤立噪声点不敏感，鼓励模型学习全局结构而非局部像素，减少空洞和离散预测，改善交叉熵损失导致的预测结果碎片化。

4 实验与分析

4.1 构造复杂场景混合数据集

为了验证方法的有效性，构建低光场景混合数据集。数据集构建以场景真实性、视角关联性、应用导向性为核心原则，在核心场景选取上，聚焦实际生产生活中的复杂低光环境。根据对各类数据集的整理和评估，本文决定以 LLVIP[29] 为基础来构建数据集。LLVIP 包含 15488 对可见-红外图像，使用海康威视 ds-2td8166bjzfy-75h2f/v2 双目相机平台，由真实的低光环境拍摄的视频截取所得，涵盖 24 个不同的夜晚场景和 2 个白天场景。去除缺乏显著性目标与大量连续帧后，使用自然图像质量评估器 (natural image quality evaluator，NIQE) 进行筛选，最终选择 NIQE 大于 6.78 的图片 1500 张。为保证数据的多样性，将 LLVIP 的筛选图片与 YLLSOD[30] 数据集中的数据混合，选用领域内通用图像标注工具 Labelme，秉持准确性、样本多样性等原则进行标注，构建 LlSOD 数据集，包含 3960 张图片与其对应的二值掩码，数据样例如图 8 所示。

4.2 实验数据集及配置

在自建 LlSOD 数据集以及标准测试数据集 RGB-252、RGBD-385、RGBT-621、VI-789 等多个数据集做测试，包括多种场景与不同光照条件的图像。实验硬件条件为 AMD Ryzen9 7845HX CPU 处理器、NVIDIA GeForce RTX 4060 8 G GPU 显卡。软件环境为 CUDA 11.8、torch 2.0.1、python 3.8。设置训练 epoch 为 50，参考同类模型 40~60 轮常规范围，适配 DSHNet 轻量化结构，平衡收敛效果与训练成本，规避过拟合。初始学习率设置为 5e-5，适配 Transformer 模块对学习率的敏感性与 Adam 优化器自适应特性，稳定更新参数。poly 学习率策略 (*power*=0.9)，实现学习率缓慢非线性衰减，保留后期微调能力，避免梯度震荡，提升边界分割精度与特征融合稳定性。权重衰减为 1e-4，可抑制模型过拟合。批次大小 4，小批次训练可提升梯度更新的随机性，并避免因混合架构的特征图计算导致的显存溢出。



图 8 LISOD 自建数据集样例

Fig. 8 Samples of LISOD dataset we constructed

4.3 实验评价指标

为评估 DSHNet 性能，采用以下评价指标：F-measure (F_β)、平均绝对误差 (MAE)、加权 F-measure (ωF_β)、S-measure (S_m)、参数量 (Params)、浮点运算数 (Flops)。Params 指的是模型中需要学习的参数的数量，能反映模型参数的内存占用量，Flops 指模型在进行推理时执行的浮点运算量。计算表达式为

$$F_\beta = \frac{(1+\beta^2)\times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{14}$$

$$\omega F_\beta = \frac{(1+\beta^2)\times Precision^\alpha \times Recall^\alpha}{\beta^2 \times Precision^\alpha + Recall^\alpha}, \tag{15}$$

$$MAE = \frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w}\left|S_{ij}-G_{ij}\right|, \tag{16}$$

$$S_\mathrm{m} = \alpha\cdot S_\mathrm{o} + (1-\alpha)\cdot S_r, \tag{17}$$

式中：F_β为准确率和召回率加权和计算，β^2设置为 0.3，强调准确率的精度值；S_{ij}为模型预测的显著图中点(i,j)的概率值；G_{ij}为真实标注中像素点(i,j)的值；S_o为原始显著性图；h 和 w 分别为图像的高、宽。MAE 用来预测真值图和显著图之间的像素级平均绝对误差。S_m用来评估显著性检测中评估结果与真实标注之间的结构相似性。

4.4 实验结果与分析

4.4.1 消融实验

为量化验证 DSHNet 中双分支结构和 CSAEM 增强模块在网络中的实际作用，以分析不同改进方法的独立作用与交互增益，在 LISOD 数据集上设计消融实验。消融实验结果如图 9 所示。以仅 Transformer 单分支作为基线模型，通过逐次添加目标模块，对比各模块对模型精度与效率的影响。第 1 组实验仅采用 Transformer 单一路径作为基线模型。在第 2 组实验中，加入轻量化的 CNN 卷积路径作为局部编码器。通过表 1 可以看到，在仅增加 0.3 M 参数的前提下显著提升了模型的特征提取能力，MAE降低了 1.4%，S_m提升了 1.3%，这说明引入 CNN 分支后能明显改善在 Transformer 低光场景下因缺乏局部细节建模的问题。在第 3、4 组实验中，分别在 CNN 路径部分引入通道重标定 SE 和通道-空间注意

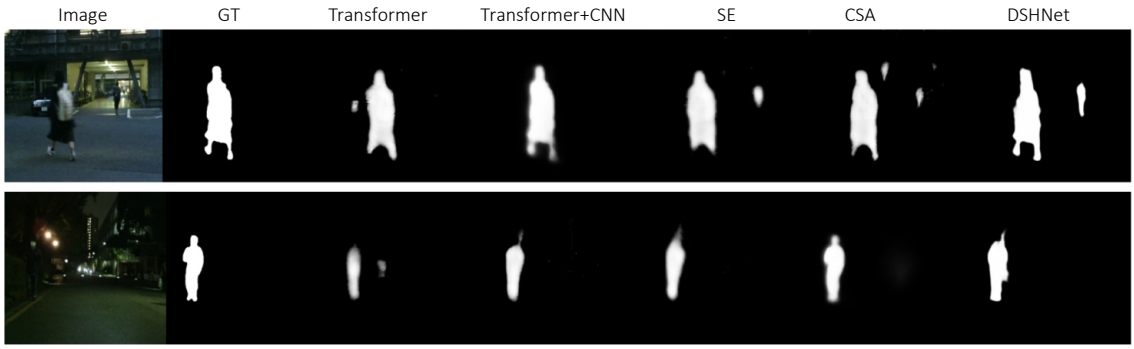


图 9 复杂低光场景消融试验 SOD 检测结果

Fig. 9 SOD detection results of immersion experiments in complex low-light scenarios

表 1 Transformer 模型消融实验结果对比

Table 1 Comparative results of Transformer model immersion experiments

Experiment	Strategy				Model error, accuracy and other performance indicators				Computational volume	
	CNN	SE	CSA	Loss	*MAE*	F_β	S_m	ωF_β	Params/MB	GFLOPs
1					0.125	0.834	0.801	0.813	25.4	26.7
2	√				0.101	0.886	0.814	0.852	25.7	27.1
3	√	√			0.098	0.892	0.818	0.860	26.0	27.3
4	√		√		0.097	0.894	0.825	0.867	26.0	27.5
5	√	√	√		0.093	0.901	0.832	0.870	26.1	27.5
6	√	√	√	√	0.091	0.912	0.837	0.872	26.1	27.8

力 CSA 机制。可以看到，针对低光环境中存在的特征不明显和边界不清晰的问题，单独使用 SE 能抑制噪声通道，S_m仅提升 0.4%，边界模糊无明显改善；单独使用 CSA 能增强空间敏感区域，F_β提升 0.8%，S_m提升 0.7%，但对噪声的抑制弱于 SE，边缘模糊问题仍然存在，说明单一通道注意力无法解决空间维度的边缘丢失问题。第五组实验中，采用 CSAEM 双分支增强模块，MAE 系数为 0.093，降低了 0.8%，F_β达到了 0.901，涨幅 1.5%。这表明双分支注意力能协同增强语义与空间特征，对低光噪声和边缘的适应性更强，验证了通道-空间协同优于单一注意力。在第 5 组试验引入联合损失函数，将所有改进策略整合为 DSHNet 算法，整体性能达到最优，相较于结合 CNN 与 Transformer 的基线模型，F_β提升 2.2%，S_m提升 2.3%。引入不同策略可在合理控制计算量的前提下，显著提升模型在低光复杂场景下的检测能力。

4.4.2 对比实验

将本文 DSHNet 方法与当下主流的模型进行对比，如 U2Net、 MVGNet、 YLLSOD、 CTDNet、 VST、 PGNet、CTIFNet、ABiUNet，单模态检测方法与本文方法实验结果对比如表 2 所示。在精度方面，DSHNet 在 LISOD 数据集上的F_β与S_m分别达到了 91.2%、83.7%，相较于同样用于低光图像显著性检测的 YLLSOD 模型分别提升了 5.4%、5.7%，比其他模型表现更优。在计算参数量方面，本文方法相较于同样使用 CNN+Transformer 混合架构的 ABiUNet，参数量降低了 23%。在计算复杂度方面，本文方法仅比 MVGNet 混合架构略高，计算资源消耗仍在可接受范围内。

为了验证模型的泛化能力，采用多个模型在 RGB-252、RGBD-385、RGBT-621、VI-789 等多个数据集 (RGBD-385、RGBT-621 为双模态显著性检测数据集) 上进行性能测试。表 3 为不同模型在多个数据集上的实验结果对比，表 4 为双模态显著性检测模型与本文方法实验结果对比。可以看出，与双模态数据检测模型对比，即使没有额外设备，本文方法仍然表现出较好的性能，这是由于 DSHNet 通过全局-局部特征融合模拟了多模态信息交互。因此，DSHNe 对 RGB-D 多模态数据集的 SOD 检测也具有较好的鲁棒性。

为了更直观地展示所提方法的优点，直观比较了不同方法的显著性检测结果，如图 10 所示。

表 2 单模态方法与本文方法实验结果对比列表

Table 2 Quantitative evaluation results with single-modal methods

Model	*MAE*	F_β	S_m	ωF_β	Params/MB	GFLOPs
U2Net	0.191	0.782	0.779	0.765	44.0	82.7
MVGNet	0.093	0.876	0.815	0.853	27.5	25.6
YLLSOD	0.110	0.858	0.780	0.824	71.4	78.2
CTDNet	0.172	0.784	0.762	0.776	11.8	7.4
VST	0.124	0.829	0.806	0.817	44.5	57.3
PGNet	0.130	0.806	0.812	0.787	72.1	76.4
CTIFNet	0.098	0.879	0.814	0.867	–	–
ABiUNet	0.100	0.889	0.806	0.867	35.8	35.2
Ours	**0.091**	**0.912**	**0.837**	**0.872**	**26.1**	**27.8**

表 3 不同模型在多个数据集上的实验结果对比

Table 3 Comparison of generalization experimental results

Model	RGB-252				VI-789			
	MAE	F_β	S_m	ωF_β	*MAE*	F_β	S_m	ωF_β
U2Net	0.042	0.677	0.716	0.686	0.039	0.856	0.792	0.810
MVGNet	0.021	0.843	0.828	0.802	0.034	0.901	0.877	0.870
YLLSOD	0.020	0.801	0.784	0.781	0.034	0.902	0.874	0.836
CTDNet	0.038	0.667	0.733	0.652	0.041	0.816	0.814	0.805
VST	0.031	0.726	0.751	0.682	0.036	0.847	0.833	0.817
PGNet	0.035	0.707	0.730	0.673	0.049	0.838	0.846	0.776
CTIFNet	0.028	0.815	0.793	0.806	0.036	0.861	0.844	0.830
ABiUNet	0.023	0.834	0.822	0.809	0.038	0.875	0.873	0.827
Ours	**0.021**	**0.852**	**0.824**	**0.817**	**0.037**	**0.907**	**0.887**	**0.889**

表 4 双模态显著性检测模型与本文方法实验结果对比列表

Table 4 Comparison with the experimental results of the two-modal saliency detection model

Model	RGBD-385				Model	RGBT-621			
	MAE	F_β	S_m	ωF_β		*MAE*	F_β	S_m	ωF_β
VST	0.094	0.751	0.746	0.723	MIDD	0.043	0.880	0.880	0.836
C2DFNet	0.123	0.729	0.698	0.670	MIA_DPD	0.043	0.871	0.885	0.824
SwinNet	0.076	0.828	0.826	0.768	CRSNet	0.042	0.878	0.879	0.843
Ours	**0.071**	**0.850**	**0.835**	**0.806**	**Ours**	**0.063**	**0.854**	**0.851**	**0.829**

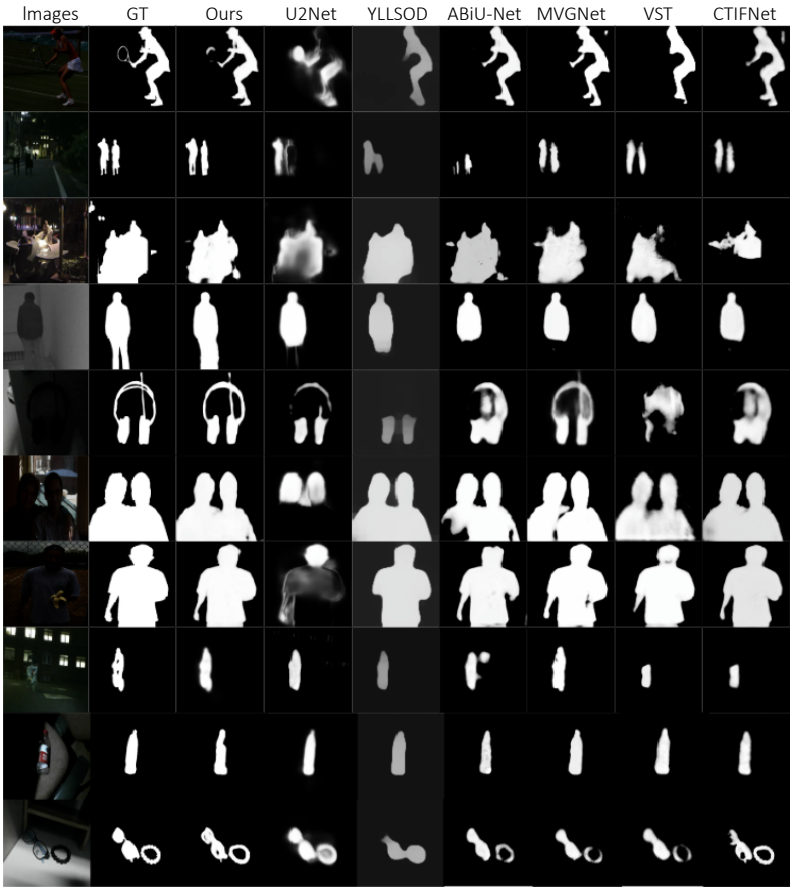


图 10 模型检测结果对比图

Fig. 10 Comparison of model detection results

图 10 包括逆光、室内夜间和室外夜间等各种场景，包含不同类型的显著目标 (大、小和多个显著目标)。大多数方法受弱光的影响，难以从背景分割前景和不准确的物体定位，导致检测精度差。相比之下，所提方法表现出更好的检测性能。总体而言，在各种弱光条件下，所提出的方法实现了稳定的物体定位和分割性能，使其比其他方法更具竞争力。

5 结 论

低光图像显著性目标的全局信息与局部细节特征很难被同时检测到，低光目标的纹理特征被弱化关注，因此，将混合模型中同尺度不同分支的全局-局部特征数据进行渐进式融合，可充分发掘显著性特征数据相关性，基于上述思想提出双流混合网络 (DSHNet) 模型。重新构建编码器，有效解决全局建模能力与计算效率之间的矛盾问题，可实现性能-复杂度的平衡。在混合模型解码阶段，采取双线性上采样、跳跃连接和混合注意力融合多种策略，以逐步恢复空间细节，通过设计 CSAEM 特征增强模块，抑制噪声通道、增强高频特征，有效解决多尺度特征融合与暗光噪声相互重叠问题。DSHNet 在不引入新模态数据并保持轻量化结构的前提下，体现出较好的模型性能，为本文成果的工程应用提供理论依据。

作者贡献

张航：实验数据的收集和分析，论文构思和撰写；魏英姿：负责方法的改善及设计、论文审核与最终修订；于聚壮：实验的设计及数据整理和分析。

利益冲突

所有作者声明无利益冲突

参考文献

1. Sun Y Z, Zhao H W, Zhou J H. DKETFormer: salient object detection in optical remote sensing images based on discriminative knowledge extraction and transfer[J]. *Neurocomputing*, 2025, **625**: 129558.
2. Wang F F, Li Y M, Wang L J et al. Asymmetric deep interaction network for RGB-D salient object detection[J]. *Expert Syst Appl*, 2025, **266**: 126083.
3. Xie C X, Xia C Q, Ma M C et al. Pyramid grafting network for one-stage high resolution saliency detection[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 2022: 11717–11726. https://doi.org/10.1109/CVPR52688.2022.01142.
4. Yuan J B, Zhu A Q, Xu Q Z et al. CTIF-Net: a CNN-transformer iterative fusion network for salient object detection[J]. *IEEE Trans Circuits Syst Video Technol*, 2024, **34**(5): 3795−3805.
5. Zhang M, Yao S Y, Hu BQ et al. C^{2}DFNet: criss-cross dynamic filter network for RGB-D salient object detection[J]. *IEEE Trans Multimed*, 2023, **25**: 5142−5154.
6. Bi H B, Wu R W, Liu Z Q et al. Cross-modal hierarchical interaction network for RGB-D salient object detection[J]. *Pattern Recognit*, 2023, **136**: 109194.
7. Li S Y, Huang Z X, Hu Y et al. Low-light salient object detection based on source-free domain adaptation[J]. *Comput Eng*, 2025, **51**(4): 75−84.
   李书玮, 黄正翔, 胡云, 等. 基于无源领域自适应的低光照显著性目标检测[J]. 计算机工程, 2025, **51**(4): 75−84.
8. Xiao Z J, Wu Z W, Zhang J H et al. Adaptive foreground focusing for target detection in UAV aerial images[J]. *Opto-Electron Eng*, 2024, **51**(9): 240149.
   肖振久, 吴正伟, 张杰浩, 等. 自适应前景聚焦无人机航拍图像目标检测[J]. 光电工程, 2024, **51**(9): 240149.
9. Wei Y P. Infrared image enhancement by fusion of linear transformation and local equalization[J]. *Laser Technol*, 2024, **48**(5): 705−710.
   魏艳平. 线性变换与局部均衡融合的红外图像增强[J]. 激光技术, 2024, **48**(5): 705−710.
10. Wang F F, Chen X A, Ren W H et al. Low-illumination object-detection algorithm based on image adaptive enhancement[J]. *Comput Eng*, 2024, **50**(10): 352−361.
    王非凡, 陈希爱, 任卫红, 等. 基于图像自适应增强的低照度目标检测算法[J]. 计算机工程, 2024, **50**(10): 352−361.
11. Jiang W H, Cheng Y B, Fang Y M et al. Light field saliency detection based on joint sequence and spatial attention mechanism[J]. *Chin J Comput*, 2023, **46**(9): 1977−1990.
    姜文晖, 程一波, 方玉明, 等. 联合序列和空间注意力机制的光场显著性检测算法[J]. 计算机学报, 2023, **46**(9): 1977−1990.
12. Wan B, Zhou X F, Zhang B L et al. LFRNet: localizing, focus, and refinement network for salient object detection of surface defects[J]. *IEEE Trans Instrum Meas*, 2023, **72**: 2508212.
13. Song Q Z, Chen Z J, Qin P L et al. Superpixel guide for transformer low-light image denoising method[J]. *Comput Eng*, 2025. https://doi.org/10.19678/j.issn.1000-3428.0070426.
    宋泉臻, 陈作钧, 秦品乐, 等. 基于超像素引导的 Transformer 低光图像去噪方法[J]. 计算机工程, 2025. https://doi.org/10.19678/j.issn.1000-3428.0070426.
14. Liang L M, Lu B H, Long P W et al. Adaptive feature fusion cascade Transformer retinal vessel segmentation algorithm[J]. *Opto-Electron Eng*, 2023, **50**(10): 230161.
    梁礼明, 卢宝贺, 龙鹏威, 等. 自适应特征融合级联 Transformer 视网膜血管分割算法[J]. 光电工程, 2023, **50**(10): 230161.
15. Li K X, Gao Z J. LFSepNet: a face recognition method with illumination and facial feature decoupling based on transformer[J]. *Comput Eng Appl*, 2026, **62**(4): 201−209.
    黎克迅, 高治军. LFSepNet: 融合 Transformer 的照明和面部特征解耦人脸识别方法[J]. 计算机工程与应用, 2026, **62**(4): 201−209.
16. Liang L M, Zhong Y, Chen K Q et al. Fusion dual-attention retinal disease grading algorithm with PVTv2 and DenseNet121[J]. *Opto-*

Electron Eng, 2025, **52**(4): 240273.
梁礼明, 钟奕, 陈康泉, 等. 融合 PVTv2 和 DenseNet121 的双注意力视网膜病变分级算法[J]. 光电工程, 2025, **52**(4): 240273.

17. He C Y, Cui X Y. Low-light image enhancement method based on CNN-transformer lightweight network[J]. *Light Light*, 2025, **49**(1): 118−120.
贺晨亚, 崔学英. 基于 CNN-transformer 轻量级网络的低光图像增强方法[J]. 灯与照明, 2025, **49**(1): 118−120.
18. Ding J H, Yuan M H. A multi-target semantic segmentation method for millimetre wave SAR images based on a dual-branch multi-scale fusion network[J]. *Opto-Electron Eng*, 2023, **50**(12): 230242.
丁俊华, 袁明辉. 基于双分支多尺度融合网络的毫米波 SAR 图像多目标语义分割方法[J]. 光电工程, 2023, **50**(12): 230242.
19. Wang L J, Jiang G Q, Wang J et al. MVGNet: multiview graph network with interactive shared fusion for fault diagnosis of wind turbines[J]. *IEEE Sens J*, 2023, **23**(21): 26804−26819.
20. Qiu Y, Liu Y, Zhang L et al. Boosting salient object detection with transformer-based asymmetric bilateral U-net[J]. *IEEE Trans Circuits Syst Video Technol*, 2024, **34**(4): 2332−2345.
21. Liu J J, Hou Q B, Liu Z A et al. PoolNet+: exploring the potential of pooling for salient object detection[J]. *IEEE Trans Pattern Anal Mach Intell*, 2023, **45**(1): 887−904.
22. Hou Q B, Cheng M M, Hu X W et al. Deeply supervised salient object detection with short connections[J]. *IEEE Trans Pattern Anal Mach Intell*, 2019, **41**(4): 815−828.
23. Mao Y X, Zhang J, Wan Z X et al. Transformer transforms salient object detection and camouflaged object detection[Z]. arXiv: 2104.10127, 2021.
https://doi.org/10.48550/arXiv.2104.10127.
24. Wang Y F, Huang X, Wang J L et al. Dual-branch remote sensing shadow detection with the fusion of Transformer and CNN[J]. *Natl Remote Sens Bull*, 2025, **29**(3): 741−751.
王一帆, 黄鲜, 王建林, 等. Transformer 与 CNN 融合的双分支遥感图像阴影检测[J]. 遥感学报, 2025, **29**(3): 741−751.
25. Li X S, Qin X, Huang C L et al. SUnet: a multi-organ segmentation network based on multiple attention[J]. *Comput Biol Med*, 2023, **167**: 107596.
26. Wang W H, Xie E Z, Li X et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]// *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, Montreal, 2021: 568–578.
https://doi.org/10.1109/ICCV48922.2021.00061.
27. Liu Z, Lin Y T, Cao Y et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, Montreal, 2021: 10012–10022.
https://doi.org/10.1109/ICCV48922.2021.00986.
28. Hu J, Shen L, Albanie S et al. Squeeze-and-excitation networks[J]. *IEEE Trans Pattern Anal Mach Intell*, 2020, **42**(8): 2011−2023.
29. Jia X Y, Zhu C, Li M Z et al. LLVIP: a visible-infrared paired dataset for low-light vision[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops*, Montreal, 2021: 3489–3497.
https://doi.org/10.1109/ICCVW54120.2021.00389.
30. Yu N N, Wang J, Shi H et al. Degradation-removed multiscale fusion for low-light salient object detection[J]. *Pattern Recognit*, 2024, **155**: 110650.

作者简介



张航 (2001-)，男，硕士研究生，主要研究方向为图像处理。
E-mail：h19862129328@126.com



【通信作者】魏英姿 (1973-)，女，硕士研究生导师，教授，主要研究计算机视觉、数据挖掘、智能系统等。
E-mail：wings_syit@126.com

A low-light image saliency detection method based on a CNN-Transformer hybrid architecture

Zhang Hang, Wei Yingzi*, Yu Juzhuang

School of Information Science and Engineering, Shenyang Ligong University, Shenyang, Liaoning 110159, China

Abstract

Objective Salient Object Detection (SOD) is difficult in low-light conditions. The images often suffer from poor quality, heavy noise in dark areas, and low contrast. These issues make it hard for models to detect objects accurately and reliably. Current methods mostly rely on single-mode features. These methods cannot capture complex information in low-light images. This is especially true when light is uneven or local details are lost. Therefore, this study proposes an efficient and robust hybrid model. We combine the strengths of Convolutional Neural Networks (CNNs) and Transformers. Our goal is to improve detection performance and overcome the limits of traditional methods in the dark.

Methods This study proposes a Dual-Stream Hybrid Network (DSHNet) based on a hybrid CNN-Transformer architecture, which adopts an asymmetric dual-branch encoder-decoder lightweight structure to make up for the defects of the existing hybrid models in low-light image processing, such as inflexible multi-scale image processing and low feature fusion efficiency. DSHNet combines two main strengths: first, CNNs extract local features at multiple scales with shallow convolution suppressing noise interference caused by uneven illumination and deep convolution capturing fine-grained texture features; second, Transformers provide powerful global modeling with self-attention mechanism breaking the bottleneck of modeling long-distance dependence between target pixels in low-light images. The model uses a multi-level strategy to extract features from both the CNN and Transformer modules, then adopts a global-local synergy method to realize cross-hierarchical bidirectional fusion of features from different branches at the same scale, solving the problem of information loss in multi-scale feature fusion. We also designed a Channel Spatial Attention Enhancement Module (CSAEM), which dynamically fuses channel-spatial attention and channel recalibration, and uses a 1×1 convolution layer to learn the weighted coefficient to adaptively balance channel and spatial attention, realizing explicit perception of edge structures and effectively suppressing background noise in low-light images. In the encoding stage, the Efficient Spatial Reduction Attention (ESRA) module is applied step by step to Transformer features, compressing the keys and values in multi-head self-attention through convolution operations and introducing the DropKey mechanism to reduce the number of model parameters and computational complexity while avoiding over-focus on local areas. We also adopt an overlapping patch embedding strategy and embed depth convolution in the feedforward network to enhance the continuity of local features and avoid the dilution effect of fixed position encoding on dark area features. In the decoding stage, we use a two-stage hierarchical design with a coarse-grained decoding branch restoring the spatial resolution of global semantic features and a fine-grained decoding branch refining the dark texture and local contrast, and deploy saliency prediction heads at six feature stages with a deep supervision strategy to realize gradient backpropagation from coarse to fine and strengthen the boundary constraints of dark area targets. Finally, we use a combined loss function of Focal Loss and Dice Loss. This helps the model focus on difficult samples and improves area overlap, effectively solving the issue of unclear boundaries caused by severe imbalance between foreground and background in low-light scenes.

Results and Discussions To verify the effectiveness of DSHNet, we constructed a complex low-light scene mixed dataset LlSOD, which includes 3960 images and corresponding binary masks by screening and mixing LLVIP and YLLSOD datasets with NIQE greater than 6.78 and labeling with Labelme. Experimental validation was conducted on LlSOD and standard datasets including RGB252, RGBD-385, RGBT-621 and VI-789, with experimental configuration set as 50 training epochs, initial learning rate of 5e-5, poly learning rate strategy with *power*=0.9, weight decay of 1e-4 and batch size of 4. Ablation experiments on the LlSOD dataset show that the introduction of the lightweight CNN branch only increases 0.3M parameters, while the F-measure is increased by 5.2% and MAE is reduced by 2.4% compared with the single Transformer branch; the CSAEM dual-branch enhancement module further reduces MAE to 0.093 and increases F-measure to 0.901, which verifies that the synergy of channel and spatial attention is superior to a single attention mechanism. Comparative experiments show that DSHNet is much lighter than the similar hybrid model ABiUNet, reducing the number of parameters by 23% (only 26.1MB). On the LlSOD dataset, it achieved high scores: 91.2% for F-measure, 83.7% for S-measure, and 87.2% for weighted F-measure, which are 5.4%, 5.7% and 4.8% higher than the YLLSOD model dedicated to low-light image SOD respectively. In the generalization performance test on standard datasets, DSHNet achieved 0.021 MAE and 0.852 F-measure on RGB252, 0.037 MAE and 0.907 F-measure on VI-789, and even showed competitive performance on dual-modal datasets RGBD-385 and RGBT-621 without introducing additional modal data, with 0.850 F-measure on RGBD-385 and 0.854 F-measure on RGBT-621, which proves that the global-

Foundation item: Liaoning Provincial Natural Science Foundation Robotics National Key Laboratory Joint Open Fund (2022-KF-12-08)

*Correspondence: Wei Y Z, E-mail: wings_syit@126.com

local feature fusion of DSHNet can simulate multi-modal information interaction and has strong robustness. Visual comparison results show that DSHNet can stably realize object positioning and segmentation in various low-light scenes such as backlight, indoor night and outdoor night, effectively solving the problems of foreground and background segmentation difficulty and inaccurate object positioning faced by most methods in low-light environments.

Conclusions In conclusion, DSHNet successfully solves key problems in low-light object detection such as difficult balance between global modeling ability and computational efficiency, difficulty in considering global information and local detail features, and overlap between multi-scale feature fusion and low-light noise. It combines the advantages of both CNNs and Transformers, and realizes the lightweight design of the model on the premise of improving detection performance. The results prove that our model is both lightweight and powerful, and provides a new effective approach for single-modal low-light image saliency detection. In the future, we will continue to optimize the model structure, further reduce the computational complexity of the model while maintaining detection accuracy, and plan to explore more real-world applications such as autonomous driving and security monitoring to promote the engineering transformation and application of this technology.

Keywords: saliency detection; low-light images; CNN; Transformer; hybrid architecture

Zhang H, Wei Y Z, Yu J Z. A low-light image saliency detection method based on a CNN-Transformer hybrid architecture[J]. *Opto-Electron Eng*, 2026, **53**(3): 250269; DOI: 10.12086/oee.2026.250269

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文