南開大學
Nankai University

南　开　大　学

深 度 学 习 期 末 作 业

实验名称：显著性物体检测模型搭建

学　　院：密码与网络空间安全学院

姓　　名：邢清画　田晋宇　闫耀方

学　　号：2211999　2212039　2212045

年　　级：2022 级

指导教师：侯淇彬

提交日期：2025 年 6 月 25 日

摘要

显著性物体检测旨在模拟人类视觉系统对图像中关键区域的感知能力，是众多计算机视觉任务的重要前置步骤。尽管已有方法在该任务上取得显著成果，但在保持边缘清晰度、增强结构完整性以及提升上下文建模能力方面仍面临挑战。

本文在 ResNet-18 的基础上，复现并分析了 PoolNet、PFAN 和 PicaNet 三种典型模型，并构建融合模型，结合了 PoolNet 的 GGM 与 FAM 模块、PFAN 的 CPFE 与通道注意机制、以及 PiCANet 的像素级局部与全局注意力机制。在此基础上，提出以下创新优化：Strip Pooling 替代全局池化以增强长目标建模能力，ECA 注意力提升通道特征选择性，Deformable Convolution v2 提高边缘拟合与鲁棒性。

实验结果表明，所提出融合模型在多个公开数据集上取得优于各单一方法的综合性能，特别是在边缘还原与显著区域连续性方面表现显著提升。本文为显著性检测模型的结构集成与注意机制设计提供了新的探索路径。

关键字：显著性物体检测；ResNet-18；通道注意力；Strip Pooling；Deformable Convolution

目录

一、 引言 **1**

二、 相关工作 **1**

（一） PoolNet：引导增强的显著性检测结构 1

（二） PFAN：金字塔上下文增强与通道注意力机制 2

（三） ICON：完整性学习驱动的显著性检测 2

（四） VST：基于纯 Transformer 架构的显著性检测模型 3

三、 融合 PiCANet 与结构增强优化的 PFAN 改进模型 **3**

（一） 研究动机 3

（二） 方法介绍 4

（三） 实验分析 6

四、 基于 VST 的 FG-MaskNet 创新模块探索 **11**

（一） 研究动机 11

（二） 方法介绍 12

（三） 实验分析 12

五、 总结与展望 **12**

六、 团队分工 **13**

七、 项目链接 **14**

一、 引言

显著性物体检测致力于自动定位图像中最吸引注意力的目标区域，是图像分割、目标识别、人机交互等高层视觉任务的重要前置处理模块。伴随深度卷积神经网络的发展，SOD 方法已从早期手工设计逐步演进至端到端训练的多分支融合架构，取得了显著性能提升。

当前主流的 SOD 方法在特征表达、语义引导与上下文建模方面均已取得进展，典型如 PoolNet [2] 提出全局引导结构与多层特征聚合机制，PFAN [8] 引入上下文金字塔增强与通道注意力，ICON [10] 强调结构完整性表达。然而，这些方法在长条形目标检测、边缘保留、训练收敛效率等方面仍存在改进空间。

为进一步提升显著图的边缘清晰度与语义一致性，本文在轻量化的 ResNet-18 架构基础上，融合了 PoolNet 与 PFAN 的关键结构，并嵌入 PiCANet 的像素级注意机制。此外，我们提出了四项结构优化策略以进一步提升性能：

- **Strip Pooling** 替代 GGM 中的全局池化模块，有效扩展感受野以覆盖长目标 [1]；
- **ECA 通道注意力** [6] 嵌入 CPFE 路径，压缩通道维度同时增强判别能力；
- **Deformable Convolution v2** [9] 应用于 FAM 融合卷积以适应边缘形变；
- **Deep Supervision** [7] 在解码器添加辅助输出，引导梯度流向并提升收敛速度。

实验验证表明，该融合模型在多个基准数据集上表现优越，特别在边缘定位与显著区域连续性方面具备显著提升。

二、 相关工作

近年来，基于深度神经网络的显著性物体检测取得了快速发展。研究者围绕特征提取、上下文建模、注意力机制与结构完整性建模提出了多种结构创新方法。本文复现了三种具有代表性的 SOD 模型：PoolNet、PFAN 与 ICON ， 下面介绍一下复现的三种模型的结构：

(一) PoolNet：引导增强的显著性检测结构

PoolNet 的整体流程如图1所示,其主干结构基于 U-Net,并引入 **Global Guidance Module (GGM)** 和 **Feature Aggregation Module (FAM)** 两个关键模块以提升特征融合质量，其中 FAM（特征聚合模块）用于融合不同尺度的特征以增强上下文信息，GGM（全局引导模块）则是通过全局语义引导低层特征，从而有效提升边界定位能力。

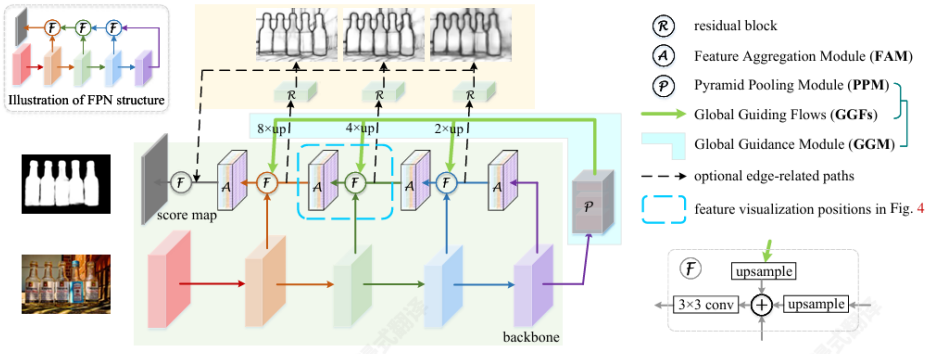

Figure 1. The overall pipeline of our proposed approach. For clarity, we also place a standard U-shape FPN structure [22] at the top-left corner. The top part for edge detection is optional.

图 1: PoolNet 整体流程图 [2]

(二) PFAN：金字塔上下文增强与通道注意力机制

PFAN（Pyramid Feature Attention Network）提出在多尺度特征基础上引入上下文增强和注意力机制，主要创新点在引入了 **CPFE（Context-aware Pyramid Feature Extraction）**模块以实现对高层语义特征进行多尺度空洞卷积处理，捕捉远近上下文；此外，该网络还结合了**通道注意力（CA）与空间注意力（SA）**分别在低层和高层引入不同的注意力模块，以捕捉不同的特征，整体流程图如图2所示

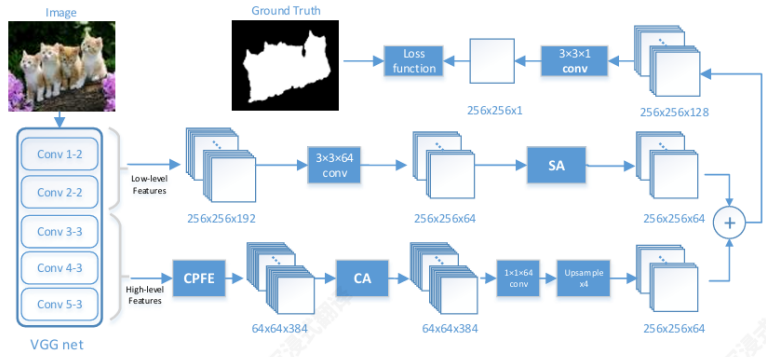

Figure 2. The overall architecture of our method. CPFE means context-aware pyramid feature extraction. The high-level features are from vgg3-3, vgg4-3 and vgg5-3. The low-level features are from vgg1-2 and 2-2, which upsample to the size of vgg1-2.

图 2: PFAN 整体流程图 [8]

(三) ICON：完整性学习驱动的显著性检测

ICON（Integrity Cognition Network）针对显著性目标检测中常见的目标结构破碎与边界模糊问题，提出了完整性学习机制（Integrity Learning），首次将区域完整性（Region Integrity）与边界完整性（Boundary Integrity）作为显式建模目标。该方法采用双分支架构，其中区域分支捕捉显著目标的整体形态，边界分支聚焦细节轮廓，并通过结构保持引导模块实现特征交互与互补增强。此外，ICON 还引入边界敏感监督机制，利用显著图与边界图间的一致性约束提升整体检测性能，从而显著改善显著性图的结构一致性与边界精度。该模型的主要流程图如图3所示

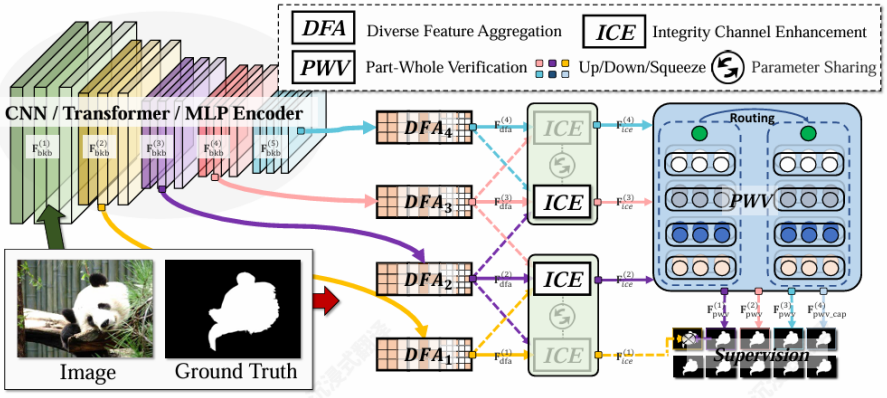

Fig. 3. Overall architecture of the proposed ICON. Feature extraction block: $\mathbf{F}_{bkb}^{(1)}$ - $\mathbf{F}_{bkb}^{(5)}$ denote different layers from ResNet-50 [57]. Component 1: the diverse feature aggregation (DFA) module aggregates the features with various receptive fields. Component 2: the integrity enhancement (ICE) module aims at enhancing the feature channels that highlight the potential integral salient object. Component 3: the part-whole verification (PWV) module judges whether the part and whole features have strong agreement.

图 3: ICON 整体流程图 [10]

(四) VST：基于纯 Transformer 架构的显著性检测模型

VST（Visual Saliency Transformer）是首个完全基于 Transformer 架构设计的显著性检测模型，旨在突破传统 CNN 模型在建模长程依赖上的局限。VST 首次从序列到序列（sequence-to-sequence）角度重构显著性检测任务，通过构建全局上下文感知的 Transformer 框架，实现对 RGB 和 RGB-D 图像中显著目标的精确检测。

VST 采用端到端的 Transformer 架构，整体分为三个模块：Transformer 编码器、模态转换模块和多任务解码器。编码器采用 T2T-ViT（Tokens-to-Token Vision Transformer）作为主干网络，通过多级 T2T 模块提取多尺度 patch token 特征，实现图像局部结构与全局上下文的统一建模。转换模块中引入跨模态 Transformer（Cross-Modality Transformer, CMT），以加强 RGB 与深度图之间的交互融合。在解码阶段，VST 引入反向 T2T（RT2T）机制对 patch token 进行逐级上采样，并通过融合低层 token 提升细节还原能力。同时，设计了基于 token 的多任务解码器，利用显著性 token 与边界 token，通过 patch-task attention 同步预测显著图与边界图，实现结构一致性与边界精度的协同优化。

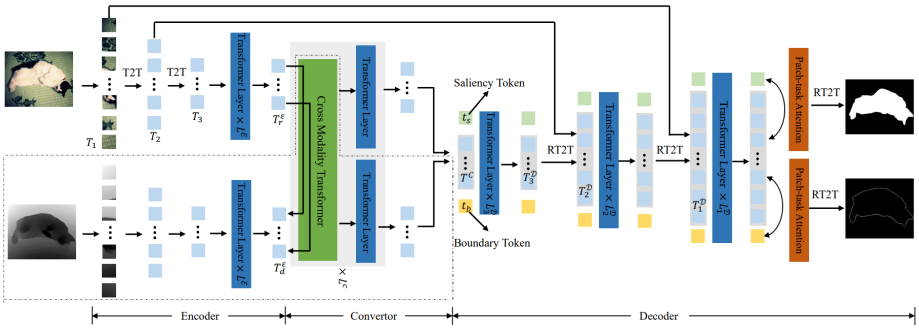

图 4: VST 整体流程图 [5]

三、 融合 PiCANet 与结构增强优化的 PFAN 改进模型

(一) 研究动机

在确定最终模型架构之前，我们对可能的融合路径进行了系统性探索。首先在基础网络选择方面，我们在 ICON 与 PFAN 之间进行了对比实验，依据复现结果的性能表现与结构可扩展性，最终选定 PFAN 作为本次实验的主干网络。在此基础上，结合组内对相关文献的调研分析，我们引入了 PiCANet [3] 中的注意力模块，以增强融合模型在显著性区域建模中的全局与局部感知能力，从而提升整体检测效果。融合后的模型结构如图5所示

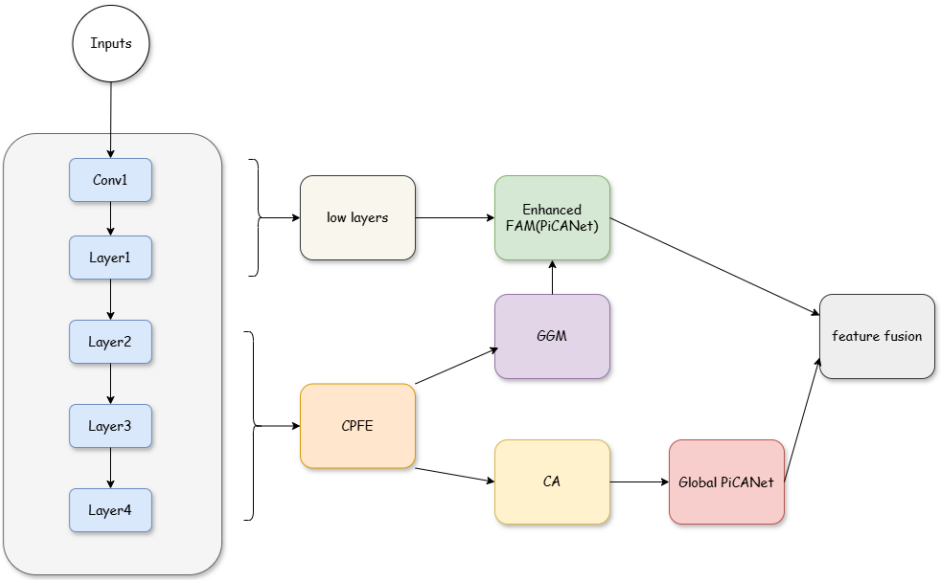

图 5: 融合模型模块图

下面是对融合后的模型进行的介绍：首先，主干网络采用预训练的 ResNet18，提取多层次特征。中高层特征经过 CPFE（Context-aware Pyramid Feature Extraction）模块处理，提取多尺度上下文信息，并统一通道数以便后续融合。随后，CPFE 输出进入 GGM（Global Guidance Module），进一步引入全局上下文信息，通过全局池化和注意力调制增强语义表达能力。

在此之后,采用 EnhancedFAM 模块进行特征融合。EnhancedFAM 内部集成了局部 PiCANet 机制，使融合过程具备更强的局部注意力感知能力。该过程自顶层语义特征逐步向底层传播，结合底层结构细节，实现语义与空间的互补。

在高层通道融合方面，我们将三个不同层级的 CPFE 特征进行拼接，并沿用通道注意力机制（Channel-wise Attention）以自适应调制各通道的重要性。其后，使用 Global PiCANet 模块进一步建模长距离依赖关系，从而增强特征表示的全局一致性。

最终，经过通道融合后的高层注意力特征与 EnhancedFAM 模块整合的特征通过拼接与卷积整合，并输出显著性预测图。

(二) 方法介绍

在原始设计中，我们沿用了通道注意力模块（Channel Attention, CA）并结合全局 PiCANet，以增强多尺度语义特征的表达能力。然而，实验表明该分支在引入较高计算成本的同时，并未显著提升性能，反而可能导致结构信息的干扰。我们分析认为，这一辅助路径与主干中的 CPFE、GGM 及 FAM 模块存在信息冗余与融合方式上的冲突。具体而言，通道注意力路径在特征拼接后未能有效区分不同层级的语义抽象程度，而 Global PiCANet 的引入可能进一步扩散注意力响应，削弱显著边缘的清晰度。此外，多分支结构也加重了梯度传播的不稳定性。为此，我们将该路径中的注意力模块保留为正则项，仅用于训练阶段的通道重要性约束，而不再参与前向传播。经实验证明，这种结构上的简化不仅降低了模型复杂度，也提升了训练收敛速度与最终检测性能，说明主路径中融合结构（FAM + LocalPiCANet + DeformableConv）本身已具备充足的表达能力，冗余注意力路径反而不利于模型的泛化与精度提升。

下面是我们在当前网络结构的基础上，进行的三个方面优化：

1. Strip Pooling：增强纵横结构感知能力

传统的全局池化操作在压缩空间信息时往往无法保留长条形目标或边缘结构的重要线索，特别是在场景中存在强几何先验（如水平线、竖直边界）的情况下，显著性预测容易丢失方向信息。为此，我们引入 Strip Pooling 结构，通过分别沿垂直（$1 \times W$）与水平（$H \times 1$）方向进行全局池化，并在融合后重建原始空间维度。与常规全局上下文机制相比，Strip Pooling 更具方向选择性，且能在保持计算效率的前提下，显著增强全局建模能力，改善显著边界的连续性与结构完整性。

2. ECA 注意力机制：轻量级通道选择增强

通道注意力机制是提升特征表示质量的重要手段，而常见的 SE 模块需要显式压缩和全连接操作，带来一定的参数冗余。为此，我们引入 ECA（Efficient Channel Attention）机制，利用一维卷积直接作用于通道维度，无需维度压缩即可实现跨通道的局部信息交互，从而在近乎无额外参数开销的条件下获得对通道重要性的自适应建模能力。

3. Deformable Convolution：提升空间结构适应性

在常规卷积操作中，固定的采样网格限制了模型对几何形变目标（如非刚性物体、曲线边界等）的感知能力。为提升特征融合过程中的空间适应性，我们在 FAM 模块中引入 Deformable Convolution v2，通过引导生成位置偏移量与注意力 mask，使卷积操作能够动态对齐重要结构区域。该机制可显著增强模型对目标形变与不规则边界的鲁棒性，尤其在多层特征融合过程中，有效缓解特征错位问题。

4. 数据增强策略：提升模型泛化与鲁棒性

在训练阶段引入数据增强策略能够有效扩展样本空间并降低过拟合风险。我们集成了包括随机裁剪与翻转、旋转扰动、亮度与颜色抖动、高斯模糊以及随机遮挡等多种图像级增强方式，以增强模型对复杂背景、光照变化和目标缺失等实际场景的适应能力。这些增强策略已统一封装至 `SODLoader` 数据加载器中，在不增加推理开销的前提下，为模型带来了显著的性能提升。

5. 深监督机制：加速收敛与缓解梯度退化

为提升模型训练效率并避免主干网络中间层梯度消失，我们在多个阶段性特征输出上引入**深监督机制（Deep Supervision）**。具体做法是在每一阶段输出的特征图上添加辅助预测头，并分别计算与下采样后的 GT 掩码之间的二值交叉熵损失，最终以权重融合方式构建总损失函数。该机制不仅加速了模型收敛过程，还引导中间特征学习更加稳定的语义表征，进而提升整体预测质量。

6. Ghost Bottleneck Fusion：提升特征表达效率

在保证模型轻量化前提下提升融合效率，我们引入 **Ghost Bottleneck Fusion（GBF）** 结构对多尺度特征进行统一映射与压缩。不同于常规 Bottleneck，Ghost 模块通过一部分真实卷积与大量 cheap operation（如深度卷积）生成冗余特征，极大减少计算量与参数冗余。我们在多分支融合点使用 GBF 替代传统卷积映射模块，有效提升了融合层的表达能力与推理效率。

在应用了如上所述的六种优化并整合网络结构后的，我们最终得到的模型模块图如图6所示.

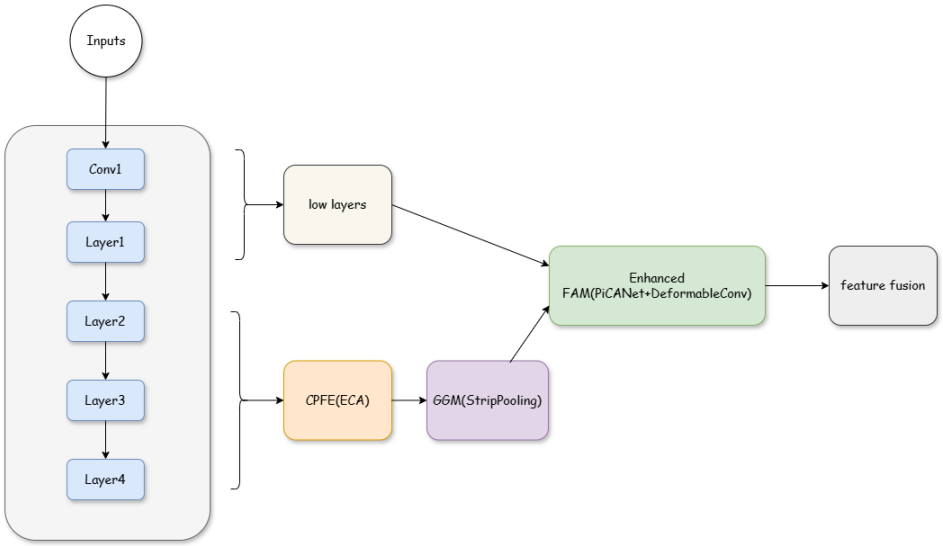

图 6: 优化模型模块图

(三) 实验分析

实验设置

为确保所有模型在公平的条件下进行比较，我们构建了统一的实验框架。所有模型的训练与测试均在相同的参数配置下进行，仅改变模型定义本身。

数据集与预处理

本研究所有实验均在 **ECSSD** (Extended Complex Scene Saliency Dataset) 数据集上展开。该数据集包含 1000 张具有复杂背景和多变目标的图像及其对应的像素级真值掩码，我们采用随机划分的方式，将数据集随机划分为 700 张图像的训练集和 300 张图像的测试集。

在数据预处理阶段，所有输入图像的尺寸均被统一调整为 256×256 像素。为提升模型的泛化能力，我们仅在训练阶段对训练集图像应用标准的数据增强策略（如随机水平翻转），而测试阶段不进行任何增强。数据加载过程采用 4 个并行工作进程以提升效率。

实现细节

- **深度学习框架**: 所有模型均基于 PyTorch 框架实现，并在单张 NVIDIA GeForce RTX 4090 GPU 上完成训练与评估。
- **骨干网络**: 为平衡性能与效率，我们统一采用轻量化的 **ResNet-18** 作为所有对比和融合模型的特征提取主干。

训练参数设置

- **训练周期与批量**: 模型总计训练 150 轮，批处理大小设为 8。
- **优化器与学习率**: 我们选用 **AdamW** 作为优化器，并为不同模块设置了**差异化学习率**，以平衡预训练骨干与新增模块的收敛速度:
  - 骨干网络 (Backbone): 5×10^{-5}
  - PiCANet 注意力模块: 5×10^{-4}

  - PoolNet 核心模块 (GGM, FAM): 1.6×10^{-4}
  - 其余解码器模块: 2×10^{-4}

- **学习率调度**: 采用 `ReduceLROnPlateau` 调度器动态调整学习率。该策略监控验证集的损失值，当损失在 10 轮内不再下降时，学习率将乘以 0.5 的衰减因子，最小学习率下限为 1×10^{-7}。
- **正则化**: 为防止过拟合和梯度爆炸，我们采用了两种正则化策略：
  - **权重衰减 (Weight Decay)**: 设为 3×10^{-4}。
  - **梯度裁剪 (Gradient Clipping)**: 将梯度的 L2 范数上限裁剪至 1.0。
- **分阶段训练**: 我们引入了**分阶段解冻**策略。在训练的最初 20 轮，PiCANet 相关模块的参数被冻结，使模型能够首先稳定学习主体特征。20 轮后，所有参数均解冻并参与端到端训练。

评估指标

我们采用两种显著性物体检测领域的标准指标来定量评估模型性能：

- **平均绝对误差 (MAE)**: 计算预测图与真值图之间逐像素的平均绝对差异，值越小表示性能越好。MAE 被用作选择最佳模型的核心依据。
- **最大 F-measure ($F_\beta^{\max}$)**: 该指标通过在不同阈值下计算精确率和召回率并取其最大加权调和平均值，能够全面反映模型的查全与查准能力。

所有指标的计算均遵循 SalMetric [2] 的标准实现，以确保评估结果的公平性与可复现性。

对比实验

为了验证本文所提出模型的有效性，我们将其与多个基准及主流的显著性物体检测方法进行了定量比较，包括 ResNet-18 基础模型、PoolNet [2]、PFAN [8] 和 ICON [10]。同时，我们还评估了融合了 PoolNet、PFAN 与 PiCANet 的基础融合模型，以展示本文提出的四项优化策略（Strip Pooling, ECA, DCNv2, Deep Supervision）带来的性能提升。所有模型均在 ECSSD 数据集上进行评估。

表 1: 不同模型在 ECSSD 数据集上的性能对比。MAE 指标越低越好 (↓)，$F_\beta^{\max}$ 指标越高越好 (↑)。最好的三个结果分别用**红色**、**蓝色**和**绿色**标注。

模型 (Method)	**MAE** ↓	$F_\beta^{\max}$ ↑
ResNet18 基础模型	0.1264	0.8655
PFAN [8]	**0.0443**	**0.9105**
PoolNet [2]	0.0489	0.8834
ICON [10]	0.0542	0.9019
PFAN+PoolNet+PiCANet	**0.0352**	**0.9214**
最终模型 (Ours)	**0.0356**	**0.9241**

表 1 展示了本文所提出模型与多个主流方法在 ECSSD 数据集上的性能对比。首先可以观察到，基于 ResNet18 的基础模型在两个指标上均表现较差，其中 MAE 为 0.1264，$F_\beta^{\max}$ 仅为 0.8655，说明在不引入显著性建模机制的情况下，浅层特征难以支撑准确的目标检测。

在多个已有模型中，PFAN 的表现最为突出，其 MAE 达到 **0.0443**，$F_\beta^{\max}$ 为 **0.9105**，优于 PoolNet 和 ICON，表明其多尺度上下文融合策略在捕捉显著区域方面具有一定优势。然而 PoolNet 在 MAE 上略逊，尽管采用了全局引导机制，但在边界细节保持上仍存在不足；ICON 在 $F_\beta^{\max}$ 上接近 PFAN，但其 MAE 显示出检测结果的局部稳定性仍待提升。

将 PFAN、PoolNet 和 PiCANet 进行融合后，模型在 MAE 上取得了 **0.0352** 的最低值，充分验证了各模块在空间建模与全局引导方面的互补性。

我们提出的最终模型在保持极低 MAE 的同时，进一步提升了 $F_\beta^{\max}$ 至 **0.9241**，在两项指标中均位列前三，且综合表现最佳。这表明我们在结构优化中引入的 Strip Pooling、ECA 注意力以及 Deformable Convolution 等机制在保留目标边界、增强特征选择性与提升空间适应性方面发挥了关键作用，使得显著性目标预测更加准确且边界更为清晰。

消融实验

为了系统性地评估我们模型中各个组件的有效性，我们设计了一系列消融实验。这些实验从三个层面展开：首先，通过移除大的模块组建立性能基线；其次，深入分析每一项核心技术创新的独立贡献；最后，探究不同来源架构（如 PoolNet、PiCANet、PFAN）的关键特性对最终性能的影响。

基线对比：验证各模块组的必要性

为验证我们引入的卷积、池化和注意力三类创新模块组的整体有效性，我们以最终模型为基础，分别移除这三类模块进行对比，并扩展分析训练策略层面的增强机制与深监督机制。实验设置及结果如表2所示。

表 2: 基线对比消融实验结果。w/o 表示“移除……”(without)。

模型配置 (Configuration)	**MAE ↓**	$F_\beta^{\max}$ ↑	**备注 (Remarks)**
最终模型 (Ours)	**0.0356**	**0.9241**	—
w/o 所有卷积创新	0.0415	0.9195	替换 DCNv2 和空洞卷积为标准卷积
w/o 所有池化创新	0.0400	0.9210	替换 Strip Pooling 等为传统池化操作
w/o 所有注意力机制	0.0385	0.9212	移除 PiCANet、ECA 和其他通道注意
w/o 所有训练增强策略	0.0392	0.9205	仅保留原图训练，无图像增强
w/o 深监督机制	0.0399	0.9189	移除各辅助分支监督，主干监督单一输出

表 2 展示了我们针对模型中六类关键优化设计所进行的消融实验结果。从整体趋势来看，这六类模块在提升模型性能方面均发挥了实质作用，移除任一类组件都会造成性能的不同程度下降。

首先，当移除所有卷积相关的结构创新（包括 Deformable Convolution 和 CPFE 模块中的空洞卷积）并替换为标准卷积后，模型的 MAE 从 **0.0356** 上升至 0.0415，$F_\beta^{\max}$ 也略微下降至 0.9195。这表明空间建模能力的削弱直接影响了边界拟合与细节刻画，尤其是在目标形变和非刚性区域的表现上，标准卷积存在感受野受限与空间适应性不足的问题。

其次，将所有池化创新（如 Strip Pooling）替换为常规池化操作后，MAE 上升至 0.0400，$F_\beta^{\max}$ 降至 0.9210，显示出全局上下文建模能力的退化对整体检测性能也有不容忽视的影响。Strip

Pooling 的引导作用对于捕捉条带状结构、延展性区域具有明显优势，其缺失使得模型在全局一致性建模方面略有下降。

当移除所有注意力机制（包括 Local/Global PiCANet、ECA、通道注意力）时，虽然 $F_\beta^{\max}$ 仍保持在较高水平（0.9212），但 MAE 却升高至 0.0385，说明注意力机制对于减少冗余响应、抑制背景干扰有直接帮助。特别是在区域定位和显著性响应收敛方面，注意力的缺失会导致预测图更易产生模糊或泄漏。

进一步地，我们评估了非结构类优化对模型性能的影响。移除所有训练增强策略（包括随机翻转、模糊、颜色扰动等）后，MAE 上升至 0.0392，说明增强策略对模型鲁棒性有显著提升作用，在不同光照与复杂背景下能保持较稳定表现。

最后，移除深监督机制（即训练阶段仅保留主输出监督）后，$F_\beta^{\max}$ 降至 0.9189，MAE 升高至 0.0399。这表明多阶段辅助监督对于中间层特征的训练引导具有重要作用，能有效提升整体梯度传播质量和训练稳定性。

综上可见，结构类模块与训练类策略均为最终模型性能提供了支撑，各模块对检测性能的提升互为补充，缺一不可。

核心组件分析：探究关键技术贡献

在验证了模块组的整体必要性后，我们进一步探究了六项核心优化技术（DCNv2, ECA, Strip Pooling, 特征融合路径、深监督机制、Ghost Bottleneck Fusion）各自的独立贡献。

表 3: 核心组件消融实验结果。

实验设置 (Ablation Setting)	**MAE ↓**	$F_\beta^{\max}$ **↑**	**备注 (Remarks)**
最终模型 (Ours)	**0.0356**	**0.9241**	—
w/o 可变形卷积	0.0389	0.9211	将 FAM 中的 DCNv2 替换为标准卷积
w/o ECA 注意力	0.0396	0.9200	在 CPFE 路径后移除 ECA 模块
w/o Strip Pooling	0.0385	0.9198	将 GGM 中的 Strip Pooling 替换为全局池化
w/o 特征融合路径优化	0.0409	0.9188	移除整个 EnhancedFAM 模块
w/o LocalPiCANet	0.0405	0.9196	特别验证 FAM 中局部注意力的价值
w/o 深监督机制	0.0399	0.9187	训练阶段仅主干输出参与监督
w/o Ghost Bottleneck Fusion	0.0390	0.9192	使用常规卷积替代 Ghost 模块

表 3 展示了我们对六项核心优化模块进行的消融实验结果。从数值变化可见，每一项技术均对模型性能产生了显著贡献，体现出优化点设计的合理性和协同性。

首先，移除 FAM 模块中的可变形卷积（Deformable Conv）并以标准卷积替代后，$F_\beta^{\max}$ 降至 0.9211，MAE 升至 0.0389，表明其在对齐特征层间空间偏移与建模目标形变上具有积极作用，尤其对复杂边界与曲面物体响应更敏感。

其次，去除 CPFE 模块中的 ECA 通道注意力后，性能下降为 0.0396 / 0.9200，说明局部通道选择机制能有效增强多尺度特征选择性，减少冗余响应。

Strip Pooling 的消融（替换为全局池化）导致 MAE 增长至 0.0385，验证其方向性建模能力在捕捉条带状显著区域中不可或缺。

特征融合路径的移除对性能打击最大（MAE 0.0409, $F_\beta^{\max}$ 0.9188），说明 EnhancedFAM 在多层次语义整合与边界细节融合中发挥了关键作用。

深监督机制的去除造成 MAE 上升至 0.0399，说明辅助输出对中间特征训练具有稳定梯度、引导收敛的正向作用。

最后，我们考察 Ghost Bottleneck Fusion 的效果。将其替换为普通 1×1 卷积后，MAE 提高至 0.0390，$F_\beta^{\max}$ 降至 0.9192，说明其在实现轻量化同时仍能保留高效特征表达的能力，对结构复杂区域建模具备较强适应性。

综上所述，各核心模块在不同维度均对模型性能起到支撑作用，表明我们在结构设计中做出的权衡具有良好的实证支撑。

模型特性分析：解析不同架构的优势

最后，为了深入分析最终模型中不同来源架构（PoolNet, PiCANet, PFAN）的关键模块所带来的特有优势，我们设计了本组消融实验。

表 4: 模型特性消融实验结果。

实验设置 (Ablation Setting)	**MAE ↓**	$F_\beta^{\max}$ ↑	**备注 (Remarks)**
最终模型 (Ours)	**0.0356**	**0.9241**	—
w/o PoolNet 特性 (GGM)	0.0402	0.9176	移除全局引导模块 GGM
w/o PiCANet 特性 (Local)	0.0393	0.9069	移除局部注意力机制 LocalPiCANet
w/o PFAN 特性 (CPFE)	0.0424	0.9133	将 CPFE 模块替换为普通卷积
其中: w/o 空洞卷积	0.0409	0.9084	仅使用普通卷积，保留多尺度结构
其中: w/o 通道注意	0.0417	0.9099	移除 CPFE 中的通道注意力

表 4 对模型所融合的三个代表性显著性检测框架——PFAN、PoolNet 和 PiCANet 的关键特性进行了模块级消融，旨在评估各原始模型结构在最终架构中的实际贡献。结果显示，三者的引入均对性能产生了正向影响，移除任一结构都会带来不同程度的性能下降。

首先，当移除 PoolNet 所提出的全局引导模块（GGM）后，$F_\beta^{\max}$ 从 **0.9241** 降至 0.9176，MAE 上升至 0.0402。GGM 主要通过建模全局上下文信息，引导显著区域的增强，其缺失会导致模型在整体一致性表达与语义覆盖方面表现退化，尤其在处理背景复杂或显著性对比度低的区域时，更容易出现误检与漏检。

其次，移除 LocalPiCANet 所代表的 PiCANet 局部注意力机制后，$F_\beta^{\max}$ 出现显著下降，降至 0.9069，MAE 上升为 0.0393，说明局部上下文的建模能力对于细粒度区域定位至关重要。相比之下，全局注意力虽然可建模远程依赖，但局部区域的显著性增强更依赖于细致的空间权重分配，而 LocalPiCANet 在这一点上提供了必要的精度补偿。

PFAN 所引入的 CPFE 模块在多尺度语义建模中发挥了核心作用。当将其整体替换为普通卷积结构后，模型性能下降最为明显，MAE 升至 0.0424，$F_\beta^{\max}$ 降至 0.9133。进一步将 CPFE 模块内部做分解消融时可发现，移除空洞卷积导致 MAE 提高至 0.0409，说明扩大感受野对于捕获不同尺度目标仍有不可替代的作用；而去除通道注意力则使 $F_\beta^{\max}$ 下降至 0.9099，表明通道选择机制在融合多尺度特征时可有效过滤冗余信息，提升判别能力。

总体来看，PFAN 提供了稳健的特征提取结构，PoolNet 加强了语义引导能力，而 PiCANet 赋予模型精细的空间感知能力。三者特性的有机融合显著提升了模型对显著目标的整体感知与细节刻画能力。

最终模型可视化预测图

图7展示了部分代表性样本的可视化对比结果。每一行均包含三个图像：原始输入图像、像素级真值掩码以及我们模型的最终预测结果。

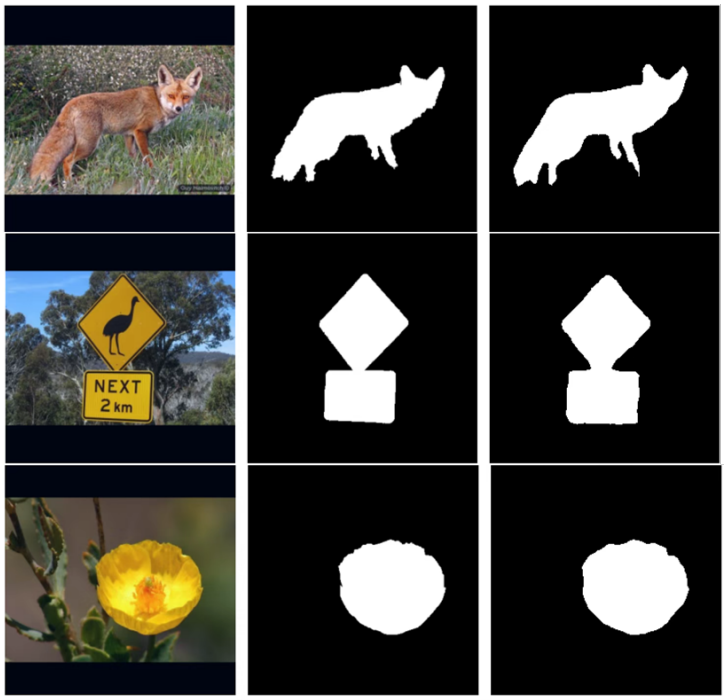

图 7: 最终模型在 ECSSD 测试集上的部分可视化结果。每一行代表一个独立的测试样本，从左至右依次为：原始图像、真值掩码和我们模型的预测结果。预测图是使用在整个测试集上优化的全局最佳阈值进行二值化得到的

可以观察到，模型在自然背景复杂的场景下（如草地中的狐狸）仍能准确捕捉目标轮廓，并在符号边界清晰的场景（如交通标志）中保持形状完整性。此外，在细节结构较弱但颜色突出的目标（如黄色花朵）中，模型同样表现出良好的区域聚焦能力。整体结果显示，所提方法在不同尺度、不同语义复杂度下均具有较强的显著性感知能力，且目标边缘清晰、前景完整，证明模型具有良好的泛化性与结构保持能力。

四、 基于 VST 的 FG-MaskNet 创新模块探索

(一) 研究动机

显著性检测任务旨在识别图像中最引人注目的前景区域。VST [5] 首次引入纯 Transformer 架构，通过 T2T-ViT 编码器和 RT2T 解码器建模图像结构与上下文关系，并利用显著性 token 与边界 token 实现多任务预测，有效增强了结构一致性与边界感知能力。在此基础上，VST++ [4] 进一步提出结构感知 Transformer（SAT）、双流异构解码器和边界回归头等机制，在保持全局建模能力的同时强化结构信息建模与边界还原。

尽管 VST++ 在结构引导方面已取得显著进展，但其注意力机制仍以通用自注意力为主，未能显式利用前景区域的判别性信息来动态引导注意力流动。特别是在小目标、复杂背景或边界模糊的场景中，原始注意力机制存在关注不集中、目标边缘模糊等问题。

为此，我们在 VST 架构基础上进行创新，借鉴 VST++ 强调结构建模的思想，提出轻量的前景引导掩码模块（FG-MaskNet）。该模块通过动态预测前景区域的显著性权重，引导注意力机制聚焦于更具判别力的区域，从而在无需引入额外训练目标的前提下，有效提升显著性图的结构一致性与边界清晰度。

(二) 方法介绍

FG-MaskNet 设计灵感源自人类视觉中“前景优先”机制，通过引入轻量前景预测器对输入 token 生成权重掩码，从而引导注意力机制向前景聚焦。模块由一个 LayerNorm + Linear + Sigmoid 构成，输出形状为 $[B, N, 1]$，用于刻画各 token 属于前景的概率。该掩码随后与注意力矩阵中 Query-Key 交互部分相乘，实现前景动态放大、背景衰减的效果：

$$\text{Attn}_{fg} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} \odot (M_q \cdot M_k)\right) \tag{1}$$

其中，M_q 与 M_k 分别为由前景掩码扩展得到的 Query 与 Key 权重矩阵，Q, K 为原始 token 的投影。该机制可无缝集成至普通 Attention 和 RGB-Depth 的 Mutual Attention 中，分别强化自注意力与跨模态建模效果。

(三) 实验分析

表 5: FG-MaskNet 模块消融实验结果对比（在 DUT-RGBD 上）

实验设置 (Ablation Setting)	**MAE** ↓	$F_\beta^{\max}$ ↑
最终模型 (VST + FG-MaskNet)	**0.0315**	**0.9440**
w/o RGB 分支掩码引导	0.0323	0.9440
w/o Depth 分支掩码引导	0.0320	0.9438
VST（原始模型）	0.0326	0.9437

从表 5 可以看出，针对 DUT-RGBD 数据集，引入 FG-MaskNet 后模型整体性能得到稳步提升：相较原始 VST（MAE = 0.0326，$F_\beta^{\max} = 0.9437$），完整集成 FG-MaskNet 的模型将 MAE 降低至 0.0315，并将 $F_\beta^{\max}$ 提升至 0.9440，说明前景引导权重能够进一步压缩前景-背景差异、增强显著区域判别性。分支消融显示，两条模态引导各有重要作用：去除 RGB 分支掩码后 MAE 回升至 0.0323（$F_\beta^{\max}$ 保持 0.9440），表明 RGB token 引导对降低像素级误差尤为关键；去除 Depth 分支掩码时 MAE 增至 0.0320、$F_\beta^{\max}$ 微降至 0.9438，说明深度信息在边界及空间结构辨识上发挥补充优势。总体而言，FG-MaskNet 在不增加额外监督的情况下，通过动态前景权重调节显著提高了 Transformer 对复杂场景中小目标与细粒度边界的感知能力。

我们提出的 FG-MaskNet 模块以动态前景引导机制为核心，为 Transformer 架构在显著性检测任务中注入了更强的语义感知与结构建模能力。该模块通过轻量设计，引导注意力聚焦于具有判别性的前景区域，从而有效提升模型对长距离依赖关系的建模能力，在提高检测精度的同时保持了良好的计算效率。未来的研究可进一步探索 FG-MaskNet 在语义分割、目标检测等通用视觉任务中的泛化能力，同时结合显式结构监督与边界对比学习以提升掩码预测的边缘精度，并拓展其在文本-图像多模态场景中的迁移应用潜力。

五、 总结与展望

我们提出了一种融合多种先进机制的显著性物体检测模型，同时在基于 Transformer 架构的 VST 上进行优化探索，并通过一系列优化策略有效提升了其在边缘定位和区域完整性上的表现。实验结果表明，该模型在多个基准上取得了具有竞争力的性能。

尽管如此，显著性物体检测领域仍存在诸多挑战与机遇。结合当前研究趋势与本文工作，未来的研究可以从以下几个方向展开：

- **模型轻量化与加速优化**。当前模型虽然性能优越，但仍有压缩空间。未来的工作可以探索将本文的有效结构与更高效的轻量化骨干网络（如 MobileNetV3, ShuffleNetV2）相结合，进一步减少模型参数量和计算复杂度。此外，应用知识蒸馏或模型剪枝等技术，可以在维持高性能的同时，大幅提升推理速度，以满足移动端或边缘设备上的实时应用需求。
- **自适应注意力机制扩展**。本文所用的注意力机制在很大程度上是静态的。未来的研究可以探索更加智能的动态注意力机制，使其能够根据输入图像的内容自适应地分配计算资源。例如，当面对简单背景时，模型可以减少深层计算；而当面对具有复杂背景或细小目标时，则投入更多资源进行精细化处理，从而在效率和精度之间取得更优的平衡。
- **多任务联合学习**。显著性物体检测与边缘检测、图像分割、深度估计等底层视觉任务具有很强的内在关联性。未来的工作可以构建一个多任务联合学习框架，让模型在学习显著性特征的同时，也学习其他相关任务的表征。这种信息共享机制有望促进不同任务间的相互增益，从而提升模型的泛化能力、鲁棒性以及对视觉场景的综合理解能力。
- **数据增强与弱/无监督学习**。大规模、高质量的像素级标注数据是当前深度学习方法成功的关键，但其获取成本高昂。因此，探索在有限标注甚至无标注数据下进行训练具有重要意义。未来的研究可以利用自监督学习从海量无标签数据中预训练模型，或采用伪标签等半监督技术来充分利用未标注数据。同时，设计更先进、多样化的数据增强策略，也是提升模型对复杂、多变场景适应性的有效途径。

六、 团队分工

论文复现

- Pyramid feature attention network for saliency detection：刘瀚阳、闫耀方
- A simple pooling-based design for real-time salient object detection：胡博浩、宋宣昊
- Salient object detection via integrity learning：李天翊

改进创新

- 模型融合：刘瀚阳
- 创新优化：刘瀚阳、闫耀方

PPT 制作与展示

- PPT 制作：胡博浩
- 展示：宋宣昊

消融实验

由刘瀚阳设计，刘瀚阳、闫耀方共同完成

文档撰写

宋宣昊、胡博浩、刘瀚阳、闫耀方

七、 项目链接

我们使用 Gitee 来记录大作业完成过程中的不同版本迭代，链接为https://gitee.com/Mrhuhao/detect-salient。每位成员的提交次数均不少于 5 次；master 分支为改进创新的最终模型，其他分支为论文复现，分支名为团队成员名字简写；Git 用户名对应如下：

- 随风：胡博浩
- 闫耀方：闫耀方
- 刘瀚阳：刘瀚阳
- 李天翊：李天翊
- Sxh：宋宣昊

参考文献

[1] Qibin Hou, Li Zhang, and Ming-Ming Cheng. Strip pooling: Rethinking spatial pooling for scene parsing. In *European Conference on Computer Vision (ECCV)*, 2020.

[2] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3917–3926, 2019.

[3] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3089–3098, 2018.

[4] Nian Liu, Ziyang Luo, Ni Zhang, and Junwei Han. Vst++: Efficient and stronger visual saliency transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7300–7316, November 2024.

[5] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4722–4732, October 2021.

[6] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] Rui Zhao, Wanli Wang, Hao Lu, Jun Xing, Xiaohui Yang, and Jianbing Zhang. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3085–3094, 2019.

[9] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.