深度学习大作业报告

许宸　张玉硕　魏靖轩
Nankai University

1. 复现模型介绍

1.1. 张玉硕

复现的方法为 F3net[2]。网络结构：F3net 整体框架（如图1）所示

- Cross Feature Module：低层特征由于感受野受限，保留了丰富的细节信息和背景噪声，有清晰的边界。高层特征由于多次下采样，边界模糊，损失了很多细节信息，但仍然有一致的语义和清晰的背景。这两种特征之间存在较大的统计差异。CFM 执行特征交叉来缓解这种差异。具体操作是高层和低层特征经过卷积、BN 和 ReLU 后，元素相乘进行特征融合，来提取特征的公共部分，然后分别和原来的特征进行元素加法进行特征细化。即作者认为元素乘是提取公共部分，再加到原来分支起到补全信息以及压制噪声的效果，单独的乘法或者加法会污染原来的特征。
- Cascaded Feedback Decoder：这一部分就是一个级联不断 refine 的过程，每次将最后一个卷积层的特征传播回前面几层，从而对其进行修正和细化。每个 decoder 包含两个过程，自底而上和自顶而下。自底而上就是正常的解码器操作，生成一张粗糙的显著性图；自顶而下就是将最后一个卷积层前的特征进行下采样元素加到之前的特征层中进行细化。所有的特征传递到下一个解码器，进行同样的操作。
- Pixel Position Aware Loss：PPA 损失由加权 BCE 损失 [4]和加权 IoU 损失 [6]组成。

$$\mathrm{L}^{\mathrm{s}}_{ppa} = \mathrm{L}^{\mathrm{s}}_{\mathrm{wbce}} + \mathrm{L}^{\mathrm{s}}_{\mathrm{w}iou} \tag{1}$$

加权 BCE 损失定义如下：

$$\mathrm{L}^{\mathrm{s}}_{\mathrm{wbce}} = -\frac{\sum_{\mathrm{i}=1}^{\mathrm{H}}\sum_{\mathrm{j}=1}^{\mathrm{W}}(1+\gamma \mathrm{a}_{\mathrm{ij}})\sum_{\mathrm{l}=0}^{1}1(\mathrm{g}^{\mathrm{s}}_{\mathrm{ij}}=\mathrm{l})\log \mathrm{Pr}(\mathrm{p}^{\mathrm{s}}_{\mathrm{ij}}=\mathrm{l}|\,)}{\sum_{\mathrm{i}=1}^{\mathrm{H}}\sum_{\mathrm{j}=1}^{\mathrm{W}}\gamma \mathrm{a}_{\mathrm{ij}}} \tag{2}$$

这里每个像素被赋予一个权重 α，更难预测的像素应该对应于更大的权重，反之亦然。α 定义如下：

$$\mathrm{a}^{\mathrm{s}}_{\mathrm{ij}} = \left| \frac{\sum_{\mathrm{m,n}\in \mathrm{A}_{\mathrm{ij}}} \mathrm{gt}^{\mathrm{s}}_{\mathrm{mn}}}{\sum_{\mathrm{m,n}\in \mathrm{A}_{\mathrm{ij}}} 1} - \mathrm{gt}^{\mathrm{s}}_{\mathrm{ij}} \right| \tag{3}$$

和 BCE 相比，加权 BCE 更关注于难像素，同时加入了局部结构信息。
同时为了加入全局结构信息，引入了加权 IoU 损失

1.2. 许宸

所选的方法为 PoolNet[5]

1.2.1 核心机制

FPN（Feature Pyramid Network）结构：PoolNet 采用了 FPN 结构，这种结构能够在不同尺度上提取和融合特征。具体来说，FPN 通过在多个不同尺度的特征图上进行处理，并将这些特征图进行融合，来捕获图像中的不同尺度信息。这种多尺度特征提取和融合的能力，使得 PoolNet 能够更加精准地识别和定位显著性区域，不论这些区域是大是小。

全局引导流（Global Guiding Flows, GGF）和全局引导模块（Global Guidance Module, GGM）：GGF 和 GGM 是 PoolNet 的重要组成部分。GGF 通过在各个特征层之间传递全局信息，增强了特征提取的效果。这意味着每一层的特征不仅包含局部

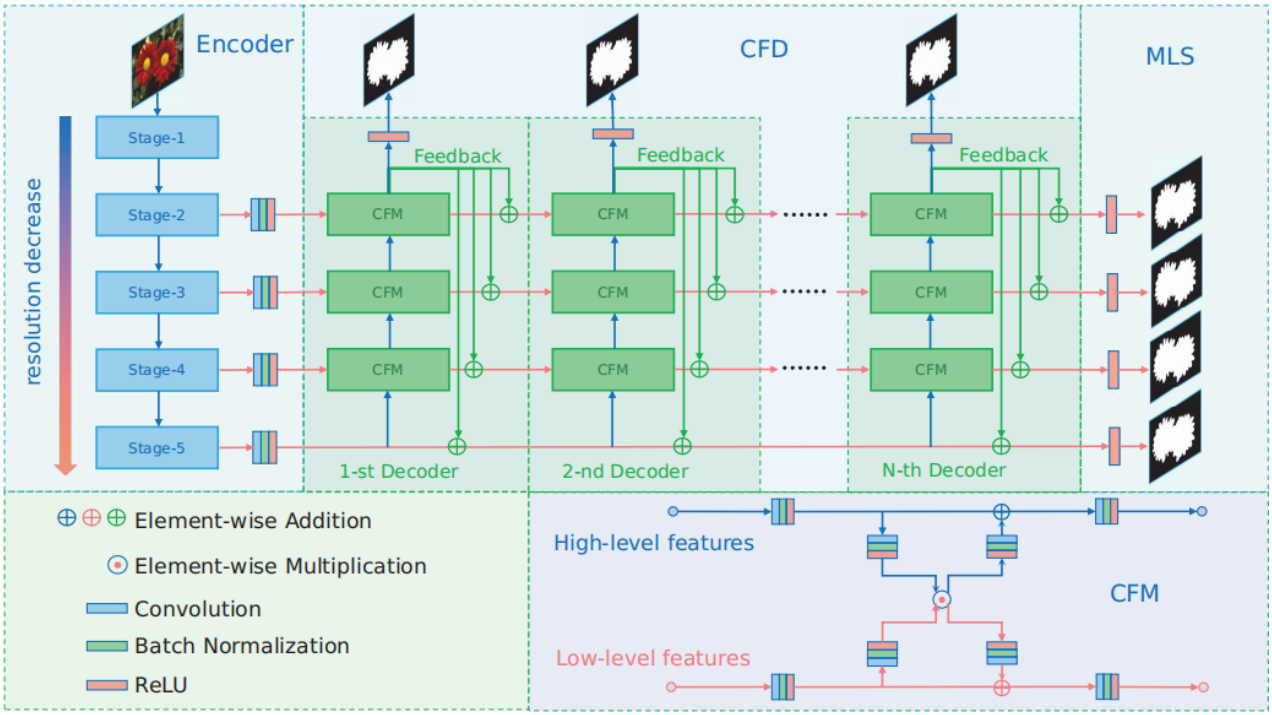

图 1. F3Net[2]结构概述。ResNet-50 用作主干编码器。使用交叉特征模块（CFM）作为融合不同层特征的基本模块。级联反馈解码器（CFD）包含多个子解码器，以反馈和优化多级功能。。

信息，还被注入了全局上下文的信息，从而提升了模型的整体感知能力。GGM 则进一步融合这些全局上下文信息，使得显著性检测的准确性得到了提高。GGM 的设计确保了模型在复杂场景中仍能准确定位显著性区域。

残差模块（Residual Block）：残差模块在特征提取过程中起到了关键作用。它通过引入残差连接（skip connection），解决了深层神经网络训练中的梯度消失问题，并提高了特征表示能力。残差模块使得信息可以在网络中更有效地传递，从而在深层网络中保持高效的特征提取和融合。

特征聚合模块（Feature Aggregation Module, FAM）：FAM 用于在不同层次的特征图之间进行特征聚合。通过融合来自不同层次的特征，FAM 能够增强对显著性区域的检测能力。这种多层次特征的聚合使得模型可以综合不同层次的信息，从而更好地理解和识别显著性区域。

金字塔池化模块（Pyramid Pooling Module, PPM）：PPM 通过多尺度池化操作，提取全局上下文信息。这些全局信息对于显著性检测任务至关重要，因为它们帮助模型理解图像的整体结构和背景。PPM 的多尺度池化策略确保了不同尺度的全局信息都能被有效提取和利用。

1.2.2 创新之处

多尺度特征融合：PoolNet 通过 FPN 和 FAM 实现了多尺度特征的融合，能够有效地捕获不同尺度的显著性信息。FPN 结构使得模型能够在不同尺度上提取特征，而 FAM 则将这些特征进行有效的聚合，从而提升显著性检测的准确性。

全局上下文信息的引入：通过 GGF 和 GGM，PoolNet 在特征提取过程中引入了全局上下文信息。全局信息的引入使得模型在复杂场景中也能准确识别显著性区域，因为模型不仅依赖局部特征，还考虑到了整体的上下文信息。

优化的残差结构：使用残差模块，PoolNet 在保持高效训练的同时，提升了模型的表达能力和准确性。残差结构使得深层网络的训练更加稳定和高效，

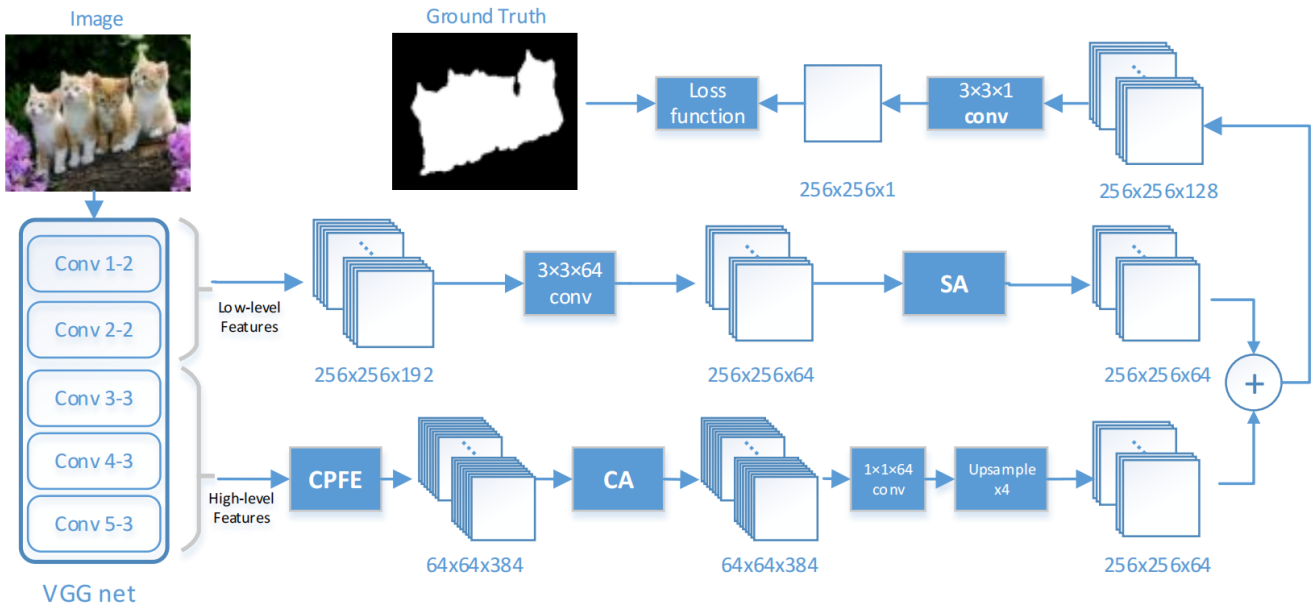

图 2. 复现方法 PFAN 的总体架构：CPFE 为上下文感知的金字塔特征提取模块，其中高层特征来自 vgg3-3、vgg4-3 和 vgg5-3，低层低特征来自 vgg1-2 和 2-2，这两个特征会将样本放大到 vgg1-2 的大小 [7]。

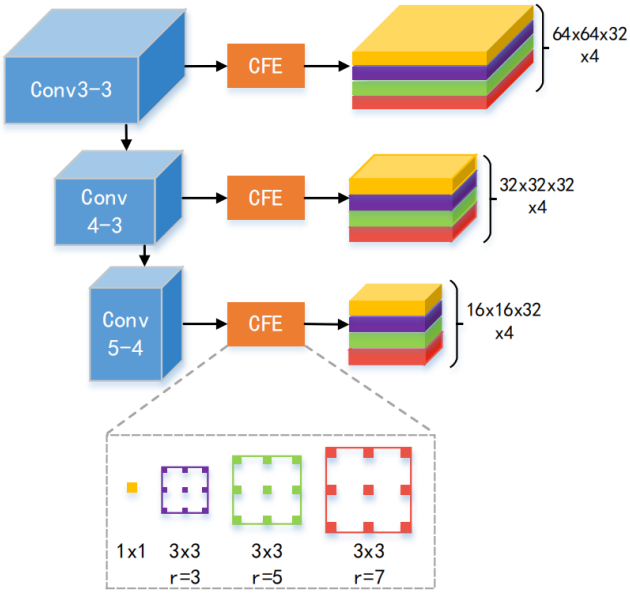

图 3. 上下文感知金字塔特征提取模块的结构：上下文感知特征提取模块以网络侧输出的特征作为输入，它包含三个 3×3 具有不同膨胀率的卷积层和 1×1 的卷积层，每个卷积层的输出通道为 32[7]。

确保了特征可以在网络中有效传递。

有效的特征聚合策略：FAM 和 PPM 的设计使得不同层次和尺度的特征能够被有效融合和利用。通过 FAM 的多层次特征聚合和 PPM 的多尺度池化，PoolNet 能够综合不同层次和尺度的信息，从而显著提升模型的整体性能。

1.3. 魏靖轩

本次实验，我选择复现的为 PFAN[7]。

PFAN（如图2）的主要结构为通过使用不同扩张率的空洞卷积来获取具有多重感受野的高层特征的一个上下文感知的金字塔特征提取模块（CPFE）（如图3）。

PFAN 采用了通道注意力机制和空间注意力机制。通道注意力机制主要用于提取包含抽象语义的高层特征，生成粗略的显著图；而空间注意力机制主要用于提取保留了更多的空间细节的低层特征，以此重建对象边界。

$$L_S = -\sum_{i=0}^{\text{size}(Y)} \Big(\alpha_s Y_i \log(P_i) + (1-\alpha_s)(1-Y_i)\log(1-P_i) \Big), \tag{4}$$

$$L_B = -\sum_{i=0}^{\text{size}(Y)} \Big(\Delta Y_i \log(\Delta P_i) + (1-\Delta Y_i)\log(1-\Delta P_i) \Big), \tag{5}$$

网络的边缘保留损失函数通过定义式4：确保边界清晰的边界损失 L_S[7]，以及式5：用于确保内部的一致性的内部损失 L_B[7]来增强网络在边界定位方面的能力和整体效果。

2. 我们的方法

2.1. 简介

作为一个经过超过 1100 万张图像预训练的大型视觉模型，Segment-Anything Model（SAM）[3]备受瞩目。然而，最近的研究表明，SAM 在下游任务中获得令人满意的性能方面遇到了挑战，包括显著目标检测（SOD）。当将大型预训练模型应用于各种下游任务时，一个重要的问题是如何有效地调整它们。我们的目标是通过设计一个统一的微调策略 [1]来提高 SAM 在不同应用场景中的有效性，以此来解决 SAM 在各种下游任务中的次优性能的挑战。我们设计了一个轻量级的卷积侧适配器（Tuned Blocks），以帮助 SAM 处理各种具有挑战性的场景。此外，为了满足分割任务的特点，我们提出了多尺度细化模块（Multi-scale Refiner）来提取更精细的图像位置特征，以进行更细粒度的分割。在解码过程中，我们设计了一个轻量化的特征结合解码器（Feature Aggregation Decoder），在解码过程中整合不同尺度的特征，得到了精细化的分割结果。我们在 ECSSD 数据集上对方法进行了评估。实验结果表明，该方法显著优于复现的 baseline 的结果。

2.2. 方法介绍

2.2.1 第一阶段：重新训练 SAM Mask Decoder

Segment Anything Model(SAM)：视觉大模型，强大的零样本泛化能力，利用 prompt（提示）进行分割。如图4所示缺少了 prompt 的输入，mask decoder

图 4. 结果对比图

并不能很好地处理特征但可以看出 encoder 端对显著特征有一定的感知。

2.2.2 第二阶段：加入多尺度模块

在分割任务中，要想获取更精细的分割结果，需要模型具有强大的描述物体边缘等细节特征的能力。但 SAM 图像编码器在 patch embed 阶段对图像进行 16 倍下采样可能会导致目标位置信息难以提取。因此，为了充分提取目标位置信息，我们提出了多尺度精化模块 (Multi-scale Refiner)，以获取更高分辨率，具有更多细节的特征。如图图6（c）是 Multi-scale Refiner 的结构，对于给定的第 j-1 层 SAM 图像编码器的输出特征 $F_{vit}^{j} \in \mathbb{R}^{C*H_1*W_1}$ 和第 i-1 层 Multi-scale Refiner 输出的层次特征 $\{F_{msr_i}^{k}\}_{k=1}^{2}$，我们首先通过 1*1 卷积模块压缩 F_{vit}^{j} 的特征维度，得到 $\widehat{F}_{vit}^{j} \in \mathbb{R}^{C^{'}*H_1*W_1}$，这可以在 SAM 图像编码器的特征维度较高时有效控制 Multi-scale Refiner 的参数量。然后 $\widehat{F}_{vit}^{j}$ 通过反卷积模块得到更高分辨率的层次特征表示 $\{\widehat{F}_{msr_i}^{k}\}_{k=1}^{2}$。上述过程可以被形式化表述为：

$$\widehat{F}_{msr_{i+1}}^{k} = deconv_k(conv_{1*1}(F_{vit}^{j})) \quad k = 1,2,3 \tag{6}$$

其中 deconv 表示反卷积模块，它对 $\widehat{F}_{vit}^{j}$ 分别做 2 倍和 4 倍的上采样，得到两种不同尺度的高分辨率特征。进一步地，需要使获得的高分辨率特征与上一层 Multi-scale Refiner 输出的层次特征融合。为了控制特征融合的程度，避免引入无效特征，我们使用了轻量级的门控单元分别作用于不同尺度的高分辨率特征。通过线性层和激活操作获取像素级别的权重，从而细粒度的控制特征融合的程度。门控

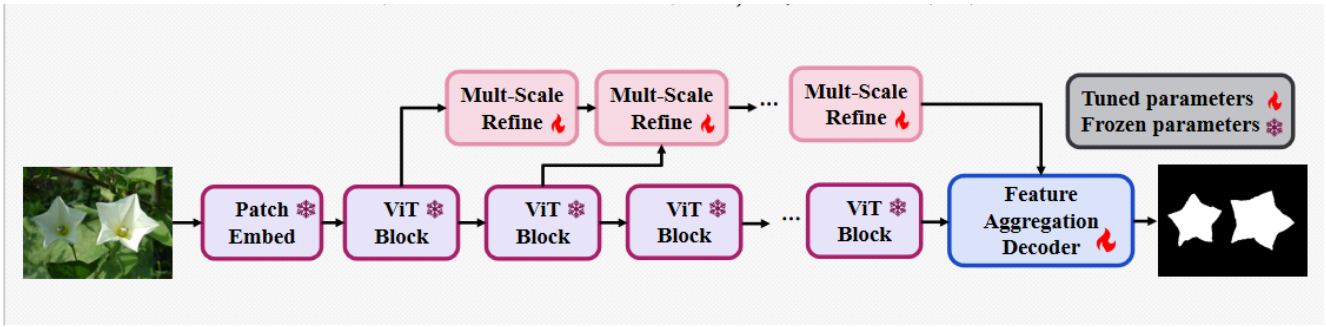

图 5. multi-scale refiner 模块架构

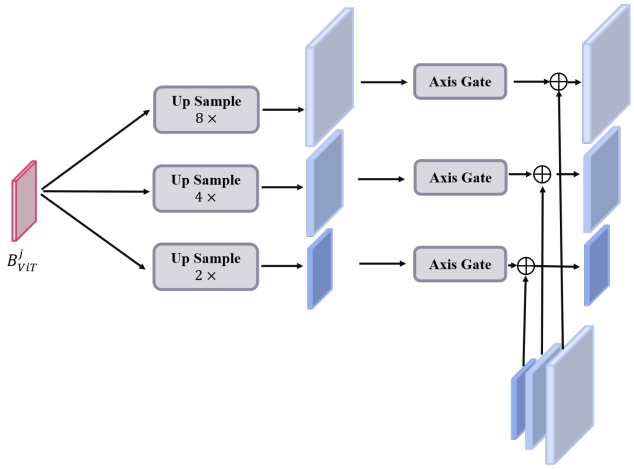

图 6. multi-scale refiner 模块

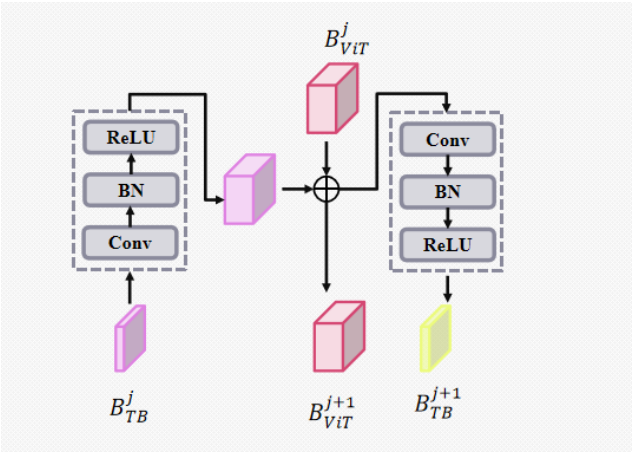

图 7. tuned block 模块

单元的操作可以被形式化表述为：

$$\widetilde{F}^{k}_{msr_{i+1}} = Tanh\left(Linear\left(ReLU\left(Linear(\widehat{F}^{k}_{msr_{i+1}})\right) \otimes \widehat{F}^{k}_{msr_{i+1}} \quad k=1,2,3\right.\right. \tag{7}$$

最后，将多个特征简单相加，实现特征融合：

$$F^{k}_{msr_{i+1}} = \widetilde{F}^{k}_{msr_{i+1}} + F^{k}_{msr_i} \quad k=1,2,3 \tag{8}$$

2.2.3 第三阶段：微调 SAM Image Encoder

对 SAM Image Encoder 进行高效参数微调。加入 Tuned Blocks 如图7所示与原始的适配器不同，Tuned Blocks 包含两个 1*1 卷积模块。第一个 1*1 卷积将压缩的特征展开到 SAM 图像编码器特征维度，然后通过加法操作融合 SAM 图像编码器的输出特征嵌入。第二个 1*1 卷积将融合特征压缩到输入特征维度，作为下一层 Tuned Blocks 的输入。给定第 i-1 个 Tuned Blocks 的输出特征 $F^{i}_{TunedBlocks} \in \mathbb{R}^{C_1*H_1*W_1}$ 和第 j-1 层 SAM 图像编码器的输出特征 $F^{j}_{vit} \in \mathbb{R}^{C*H_1*W_1}$，则第 i 个 Tuned Blocks 的操作可以被表示为：

$$F^{j+1}_{vit} = F^{j}_{vit} + conv_{1*1}(F^{i}_{TunedBlocks}) \tag{9}$$

$$F^{i+1}_{TunedBlocks} = conv_{1*1}(F^{j+1}_{vit}) \tag{10}$$

其中表示 $conv_{1*1}$ 卷积模块，包含 1*1 卷积，批归一化和激活操作。F^{j+1}_{vit} 是第 j 层 SAM 图像编码器的输入，$F^{i+1}_{TunedBlocks}$ 是第 i 个 Tuned Blocks 的输入。Tuned Blocks 是轻量级的，维持了适配器的简单性。

3. 实验结果

3.1. 实验设置

实验数据集为 ECSSD 数据集，分为 700 训练 +300 验证/测试，训练 50 轮。训练使用一块 RTX3080 16G，占用显存 9G，训练总时间为 6h。

Components		Tunable	ECSSD			
Tuned Block	Feature Aggregation Decoder	Param	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
-	-	4.0M	0.900	0.918	0.828	0.050
✓	-	12.3M	0.920	0.941	0.874	0.039
-	✓	1.3M	0.919	0.933	0.857	0.041
✓	✓	9.3M	0.928	0.951	0.896	0.031

(a) 针对不同模块的消融结果

Tuned	ECSSD			
Blocks	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
6	0.918	0.937	0.877	0.039
9	0.922	0.946	0.883	0.034
12	0.928	0.951	0.896	0.031

(b) 针对不同模块数的消融结果

表 1.

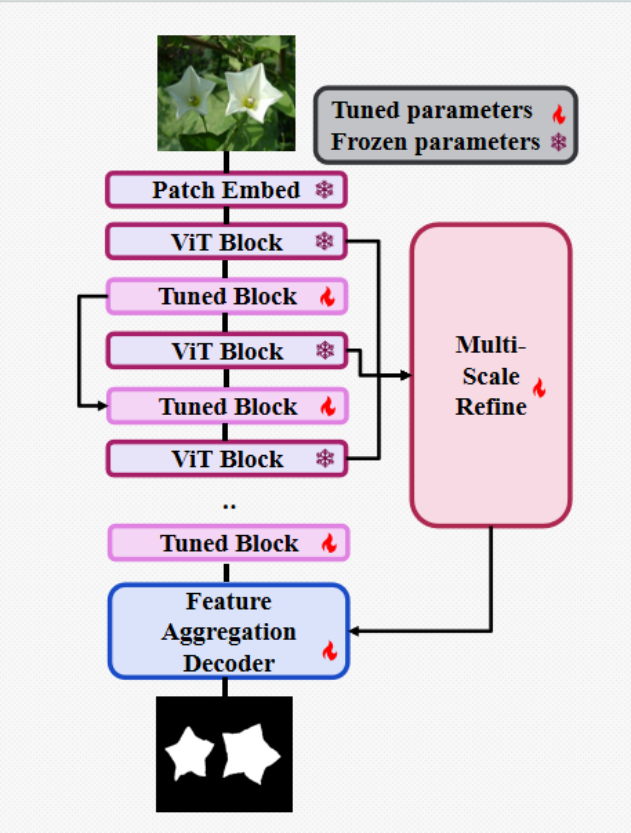

图 8. 模型总架构

Methods	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
PFAN[7]	0.889	0.873	0.861	0.0057
F3Net[2]	0.909	0.917	0.889	0.041
PoolNet[5]	0.913	0.919	0.885	0.050
SAM[3]	0.900	0.918	0.828	0.050
Tuned SAM	0.928	0.951	0.896	0.031

表 2. BASELINE 与我们的方法，在 ECSSD 数据集上取得的性能指标。

3.2. BASELINE 与我们的方法的指标结果

我们针对所复现的三个网络：PFAN[7]，F3Net[2]，PoolNet，在 ECSSD 数据集上进行了训练和测试，并用 S_α、E_ϕ、F_β^ω 和 MAE 进行评价，得到的结果如 表2所示。

同时我们也对原始的 SAM[3]模型和我们的模型在 ECSSD 数据集上进行了评估，得到的结果同样已在表2中列出。

3.3. 消融实验与结果

针对我们所提出的方法，我们对模型进行了一定的消融实验：

- 针对有无 Tuned Block 和 Feature Aggregation Decoder 进行消融实验。
- 针对 Tuned Block 的数量进行消融实验。

我们测量得到的结果如表1所示。

表1(a) 为不同模块的消融结果，可以看到，在两个模块均不存在的情况下，我们可调优参数为 4.0M，在 ECSSD 数据集上的结果是最差的；在添加了 Tuned Block 之后，我们可以进行调优的参数就变为了 12.3M，模型取得的结果有较大的提升，但是参数量的增大使得我们的训练需要耗费更多的资源；在添加了 Feature Aggregation Decoder 之后，我们模型的可调优的参数大幅度减少，这有效的降低了我们训练所需要的资源，同时模型的效果也取得了提升。最后我们将两个模块都采用，我们取得了比只采用 Tuned Block 更好的表现，且参数量在相比 Tuned Block 项有了明显的减少，证明我们的两个模块可以很好的相互发挥长处提高我们的模型的表现。

表1(b) 为不同 Tuned Block 数的消融结果，可

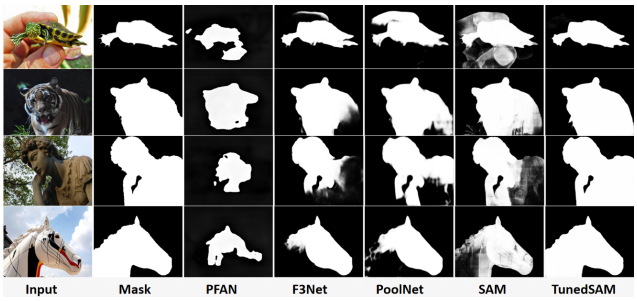

图 9. BASELINE 和我们的方法在输入图像上的结果

以看到，模块数越多，在测试集上的表现就会更好，四项指标随着模块数的变多都有一定的提升，但是考虑到模块数的增多会使得参数量变大，因此选择一个合适的模块数是最佳的策略。

3.4. 我们的方法与 **BASELINE** 的对比提升

首先表2已经指出，在 ECSSD 数据集上我们的模型各项指标都取得了非常大的提升。

最后，我们将我们所复现的 PFAN[7]，F3Net[2]，PoolNet[5]和 SAM[3]，以及我们的方法 TunedSAM 在测试集上进行了输出图片的直观比较，可以看出，我们的模型对显著性物体检测取得了最接近 Mask 的结果。

4. 总结

4.1. 人员分工

- 许宸: 复现 PoolNet，IDEA 的提出与模型的整体构建，多尺度模块，部分消融实验。
- 张玉硕: 复现 F3Net，模型的 Tuned Blocks 模块构建与调整，部分消融实验，文档撰写。
- 魏靖轩: 复现 PFAN，模型的 Tuned Blocks 模块构建与调整，部分消融实验，文档撰写。

4.2. 代码仓库

gitee 仓库

4.3. 总结与展望

我们将侧向网络引入到 SAM 的微调中，形成一个能有效地从 SAM 编码器中提取特征的双流侧网络，同时参考 [5]提出了针对分割任务而定制的多尺度细化模块（Multi-scale Refiner）和轻量化解码器（Feature Aggregation Decoder）。这些模块通过高分辨率的层次特征获取细化的目标位置信息，并在解码过程中进行充分整合，以获得详细的分割结果。

经过一系列的实验和测试，我们的模型在测试集上取得了非常大的提升，同时在对输入图像的显著性物体检测方面也有明显的改善，对比各项输出图像，我们的模型的表现是最接近 Mask 的，这也为后续的科研以及工程应用提供了一种新的思路和方法。

参考文献

[1] T. Chen, L. Zhu, C. Ding, R. Cao, Y. Wang, S. Zhang, Z. Li, L. Sun, Y. Zang, and P. Mao. Sam-adapter: Adapting segment anything in underperformed scenes. pages 3359–3367, 10 2023. 4

[2] Q. H. Jun Wei, Shuhui Wang. F3net: Fusion, feedback and focus for salient object detection. In AAAI Conference on Artificial Intelligence (AAAI), 2020. 1, 2, 6, 7

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. 04 2023. 4, 6, 7

[4] Q. Li, X. Jia, J. Zhou, L. Shen, and J. Duan. Rediscovering bce loss for uniform classification. 03 2024. 1

[5] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang. A simple pooling-based design for real-time salient object detection. In IEEE CVPR, 2019. 1, 6, 7

[6] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan. Focal and efficient iou loss for accurate bounding box regression. Neurocomputing, 506, 07 2022. 1

[7] T. Zhao and X. Wu. Pyramid feature attention network for saliency detection. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3080–3089, 2019. 3, 4, 6, 7