轻量化目标检测

Mobile U-ViT

解决问题

资源受限的移动设备上进行高效医疗图像分割的挑战；

为自然图像设计的移动模型在医疗图像上表现不佳，原因在于两种图像之间存在显著的**信息密度差异**

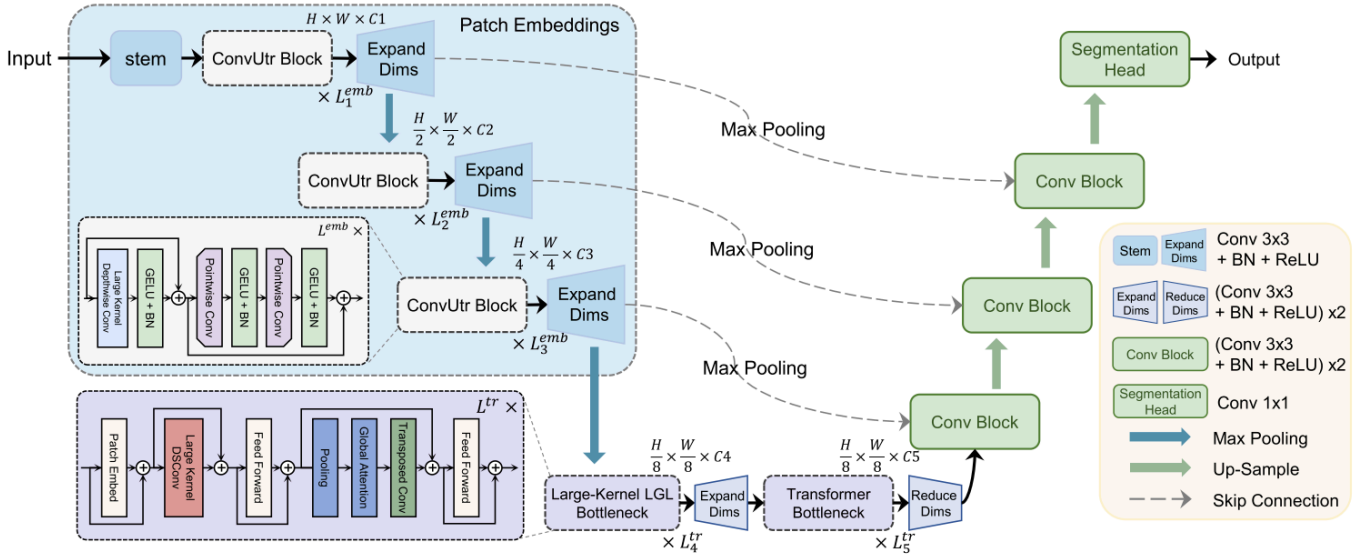
提出Mobile U-ViT,降低参数量和计算量，且保证高精度

创新点

ConvUtr：一个**轻量级、受 Transformer 启发的 CNN 主干网络**，用于高效地将医疗图像的稀疏像素空间压缩成紧凑的潜在表示

LKLGL 模块：一个专门设计的模块，用于有效地进行**局部与全局信息的交互**，以应对医疗图像中模糊的边界和高噪声问题

U形架构和级联解码器：采用 U-Net 类似的架构，并使用**下采样跳跃连接**（downsampled skip-connections），以有效地对齐和融合局部与全局特征，从而实现精确的密集预测



混合编码器-解码器（hybrid encoder-decoder）模型，结合了 CNN 和 Transformer 的优点 。

- **编码器（Encoder）**：由五个阶段组成 。
    - **前三个阶段**使用 **ConvUtr 块**作为特征提取器。ConvUtr 是一种高效的 CNN 模块，它使用**大核深度可分离卷积**（large kernel DepthwiseConv）来扩大感受野，捕获全局特征，同时使用两个点状卷积（PointwiseConv）来模仿 Transformer 的学习模式，以此实现轻量化和高效性 。
    - **第四个阶段**使用 **Large-kernel Local-Global-Local (LKLGL) 模块**。该模块通过**大核卷积**捕捉局部信息，然后使用**池化**（pooling）来聚合 tokens，再通过**注意力机制**（attention mechanism）进行高效的全局信息交互，最后通过**转置卷积**（transposed convolution）将精炼后的全局信息重新分配回局部层面 。

    - **第五个阶段**使用轻量级的 **Transformer 瓶颈**进行长距离建模 。
- **解码器（Decoder）**：采用**级联上采样结构**（cascaded upsampling structure）。它通过

  下采样跳跃连接将编码器的低级 CNN 特征与高级 Transformer 输出特征进行融合，从而过滤冗余信息（如背景噪音）并突出必要的边界信息，这对医疗图像分割至关重要

创新点启发

ConvUtr:

大核是为了适应医学图像分割对于大感受野的需求，深度可分离卷积是轻量化，两次点状卷积+激活函数+BN模仿transformer做特征融合

LKLGL：

- **L**ocal（局部）-**G**lobal（全局）-**L**ocal（局部）
- 第一步局部聚合 (Local)：大核深度可分离卷积
- 第二步全局交互 (Global)：池化+注意力
- 第三步局部精炼 (Local)：转置卷积（高层次信息融合到像素级细节）

U形架构和级联解码器：

- encoder做下采样，提取高层次语义；decoder对语义信息上采样，恢复空间分辨率
- 下采样跳跃连接：
    - 传统的 U-Net 中，跳跃连接直接将编码器的特征图传送到解码器对应阶段，可能会导致低级 CNN 特征与高级 Transformer 特征之间的语义差异，噪声冗余
    - Mobile U-ViT 的跳跃连接在将编码器特征传递给解码器之前，会先对这些特征进行**下采样**处理（**减少背景噪声**并**锐化模糊的边界**）
    - 传递**下采样后的特征**，模型能够在解码器阶段**对齐**（align）来自encoder的低级、高分辨率特征和decoder自身的高级、低分辨率特征

CNN 与 Transformer 的深度融合：

- **CNN (ConvUtr)** 作为编码器的**早期阶段**，高效地提取**局部特征**和**紧凑的潜在表示** 。
- **Transformer (LKLGL)** 作为编码器的**后期阶段**，负责**长程依赖建模**，捕捉全局上下文 。

HS-FPN

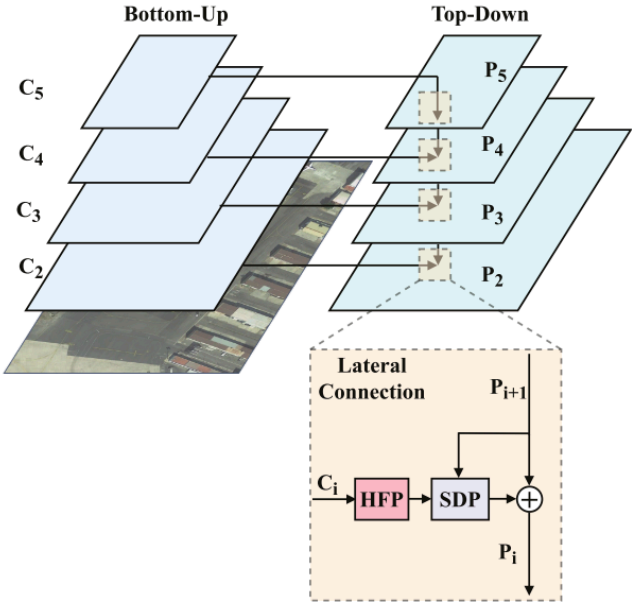
解决问题

- 传统目标检测对于微小目标性能显著下降
- FPN对小目标有3个挑战：
    - 特征稀缺：下采样网络后，特征图仅几个像素
    - 缺乏关注：微小目标的特征响应弱且容易受到干扰，但 FPN 对所有尺度的特征都进行同样的处理，没有专门对其进行增强

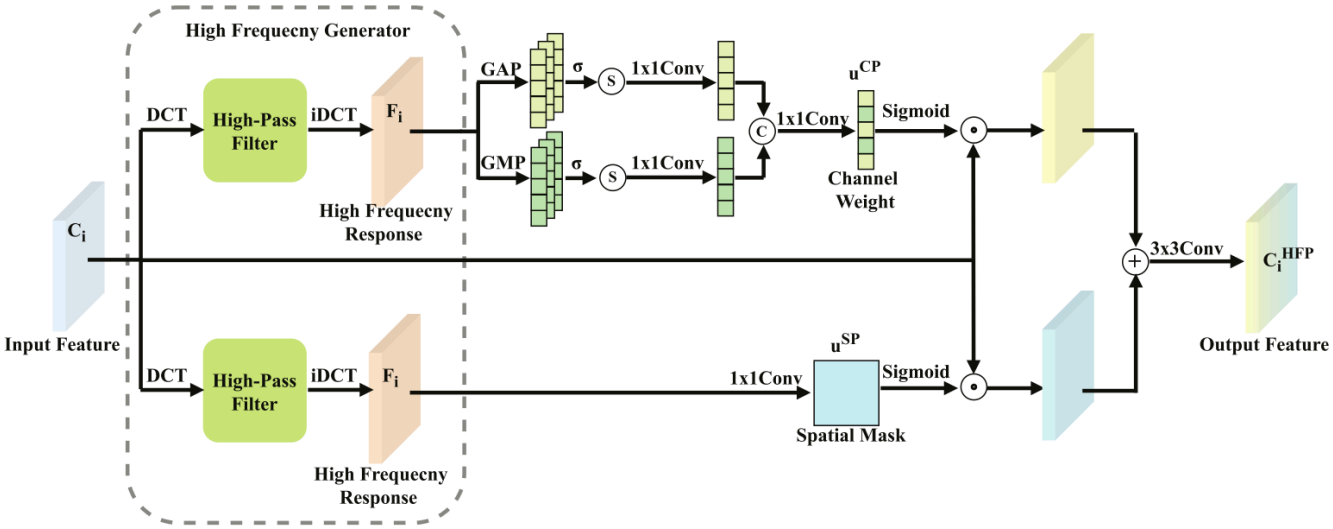
- 空间感知不足

创新点

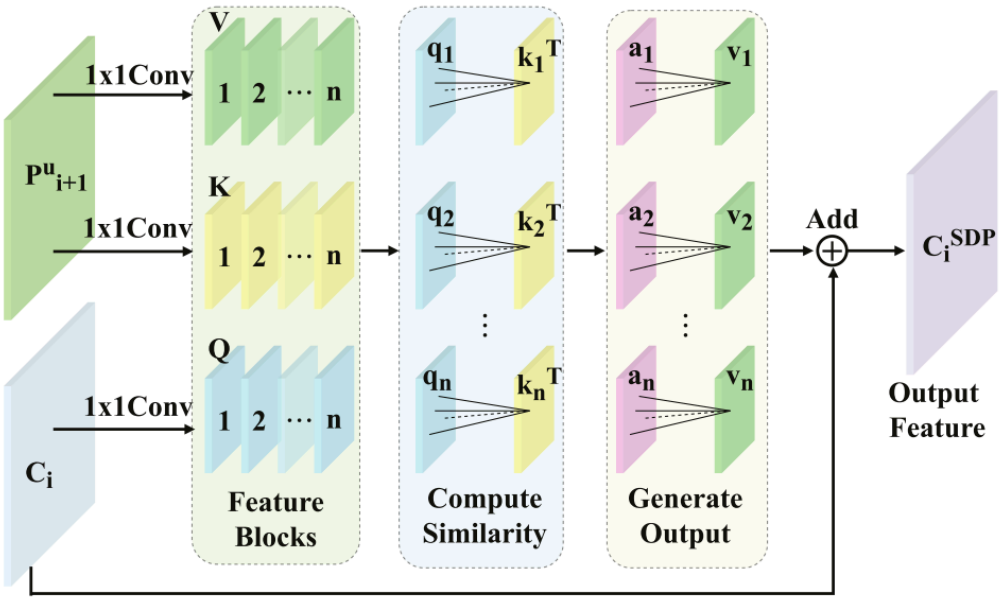
- **高频感知模块（HFP）**：通过高通滤波器突出微小目标的高频特征，并将其作为权重来增强原始特征图
- **空间依赖感知模块（SDP）**：捕获 FPN 所缺乏的像素级空间依赖关系



侧向连接：HFP处理自底向上路径的特征图Ci,生成高频响应，以突出和增强微小目标的特征;SDP接收 HFP 的输出和来自上一层 FPN 的上采样特征图 Pi+1 作为输入,通过像素级注意力学习上下层特征图之间的依赖关系



HFP：通过DCT转成频域，进过high pass再iDCT转回特征图；两条路：（1）两种池化再拼接，和原始特征图融合（2）生成空间掩码，值越高说明越可能是小物体，增强原始特征图上小物体区域特征；两条路相加再卷积



SDP：在特征块内部的像素之间计算交叉注意力（发生在每个特征块内部）：将**Ci 中某个特征块**的像素作为 **Query**，而将 **Pi+1 中对应的特征块**的像素作为 **Key** 和 **Value**；计算这个 **Query** 像素与**所有 Key** 像素之间的相似度；用这些相似度来加权对应的 **Value** 像素，从而融合上层 FPN 信息的像素级新特征

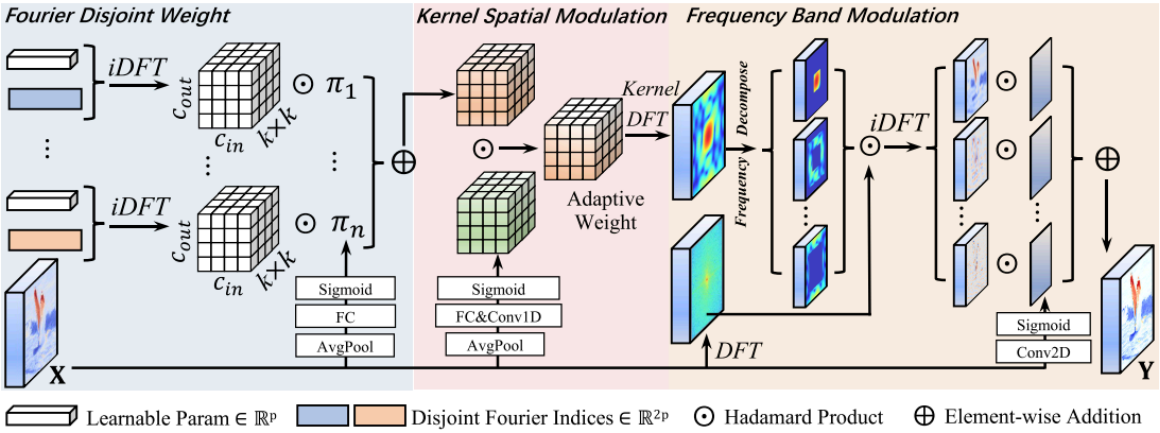
创新点启发

- 改进的U形架构跳跃连接，用下采样精炼特征再传递给decoder
- 小目标呈现高频率特征
- 可以在跳跃连接中加一些轻量化处理模块：pooling/conv/attention，学习融合重要的特征

FDConv

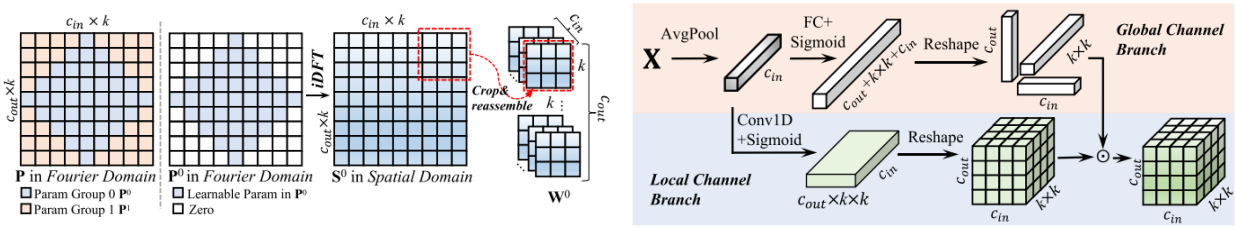
解决问题

- 传统DY-Conv参数多但效果不理想，大多是空间域的核的加权
- 并行卷积和的频率响应高度相似，参数冗余
- 大多数卷积方法是空间不变的，模型不能自适应捕捉不同频率信息（不能根据图像中不同位置的内容动态调整）



核心流程：

- 在频域里构造出一组“互不重叠”的候选卷积核（FDW）
- 对这些卷积核再做更细粒度的逐元素调制（KSM）
- 把特征分成不同频带，对不同空间位置分别调节（FBM）



FDW（一些多种频率但固定的核模版）：

- 解决传统卷积核频率相似的问题
- 在频率域中分配参数，讲整个频率范围切成几段（低中高），每个段单独一个频率核
- 再IDF变回空间域

KSM（让核根据图像微调）：

- FDW保证了核的差异，但是还是比较粗；KSM进一步细化
- 两个分支：
    - **Local 分支**：用轻量的操作（比如一维卷积）来预测一个跟卷积核一样大小的调制矩阵，逐点相乘，实现细粒度调整
    - **Global 分支**：用全局平均池化加全连接层，预测一些全局性的缩放系数，用来补充全局上下文

FBM（分频率加权）：

- 卷积核分频，每个频带单独算一个卷积输出
- 用输入特征再跑一个小的分支（通常是一个卷积+激活函数），得到一张 **门控图**；门控图的大小和特征图一样，但它针对每个频带都生成一份（不同频率的注意力分配）
- 加权融合

创新点启发

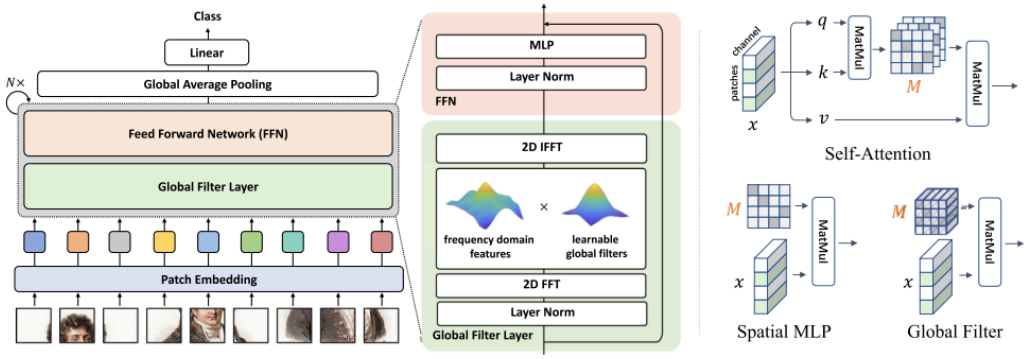
- 频率域构造卷积核，减少功能上重合
- 对于卷积核先设置固定模版，再针对具体图片细粒度+宏观微调
- 结合频域和空间域（小目标会表现为高频）；频率域注意力结合，可以尝试对于更多的“隐含空间”来做注意力（尺度，形状，方向等）

GFNet

解决问题

高计算开销（transformer/mlp都是二次的，高分辨率特征难放大）+长距离空间依赖的高效捕捉效率低下

创新点



在ViT的基础上，把自注意力层换乘了global filter layer

embedding后的patch$(N \times D)$, 2D FFT变成2D频域表示， 与全局滤波器相乘加权，在2D IFFT变回空间域

创新点借鉴

频域上处理优化，全局+局部filter结合，动态权重

可以改进：

多尺度融合可以加，CNN可以和GFNet融合（但是效率会差，加一个高效的卷积？），非线性的全局滤波器

Haar小波变换下采样

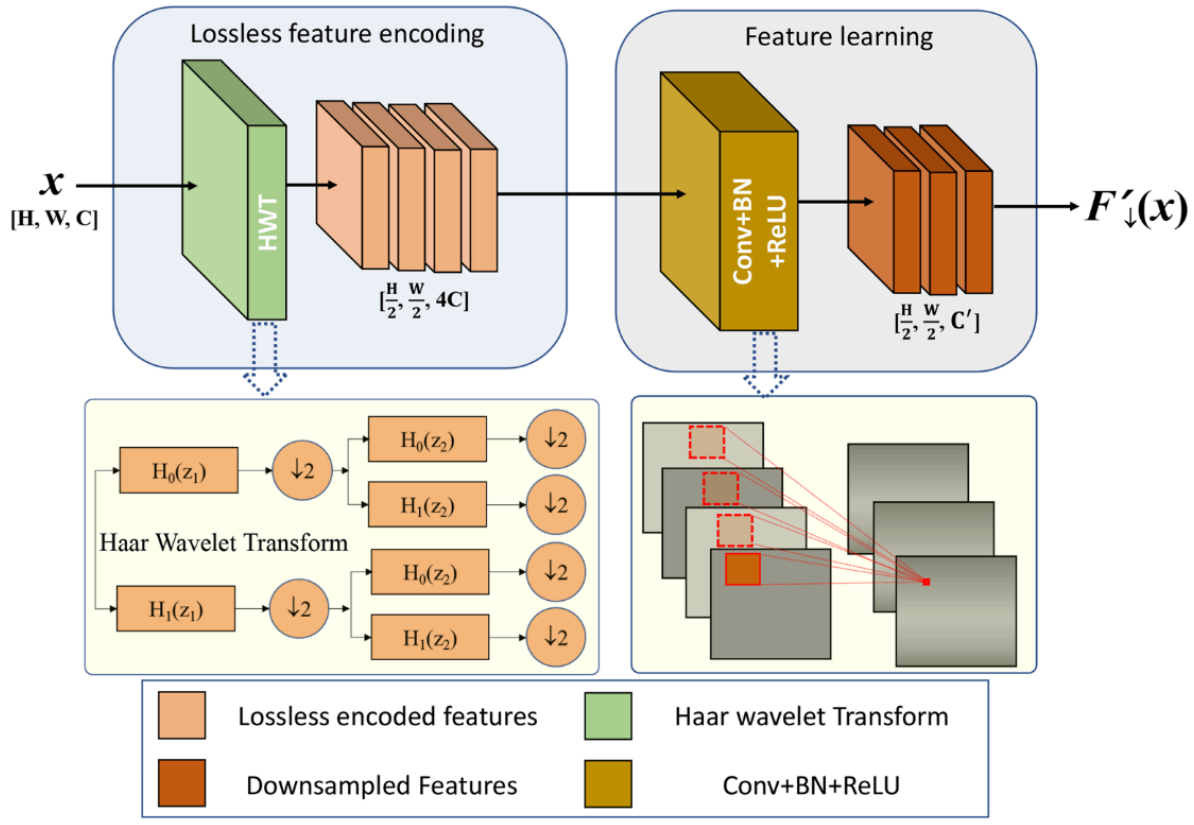
解决问题

传统CNN下采样用最大池化、平均池化或步长卷积等等，会丢失重要信息（边界/纹理/小物体细节）；新方法减小分辨率的同时保持足够的信息

创新点

Haar小波变换：讲特征图分解成高频和低频，对小物体和边界效果显著

提出特征熵指数（FEI）：衡量下采样后特征图的信息不确定性



把图像划分成2×2的块，低频成分$\frac{a+b}{2}, \frac{c+d}{2}$;高频成分$a-b, c-d$

4个通道：低频和3个高频成分

创新点启发

- 下采样方法的替换：Haar小波变换保留更多边缘/小物体细节
- 还是频域特征

可以做的改进尝试：

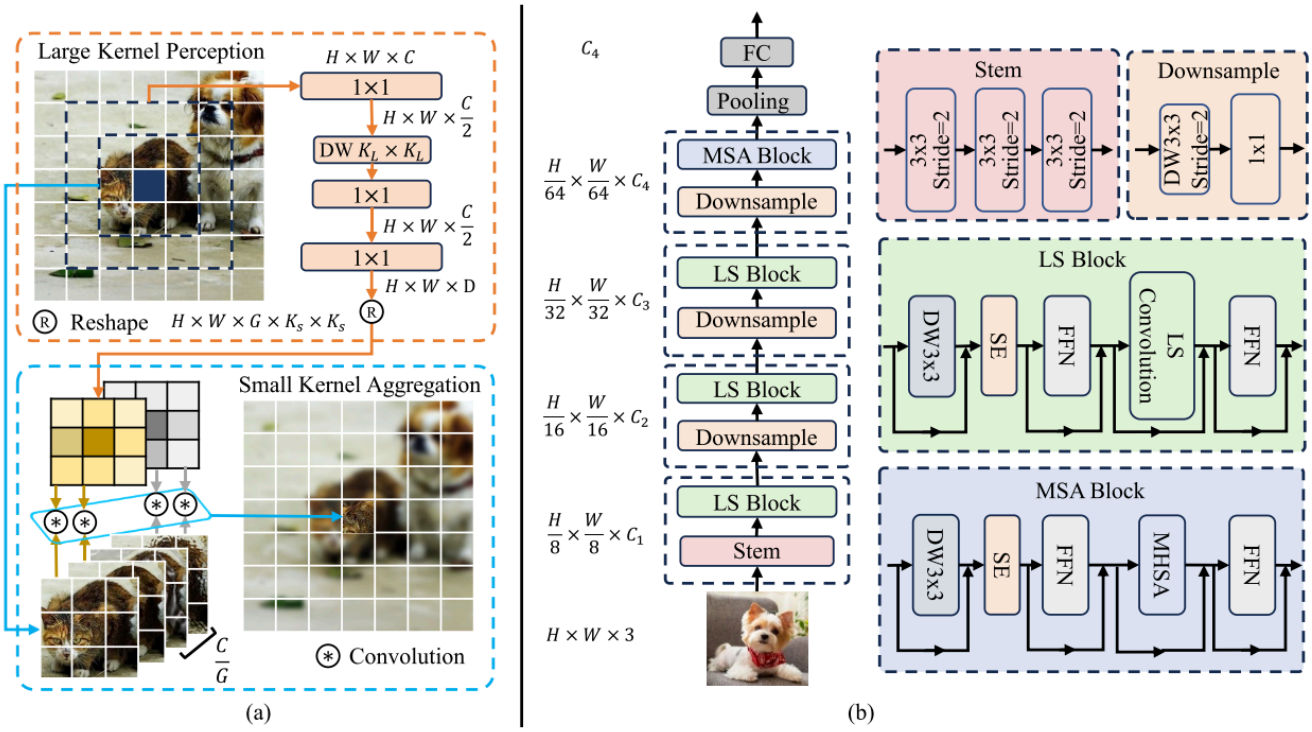
- 其他类型的算子：使用**Daubechies小波**或**Coiflet小波**，这些小波变换比Haar小波在捕捉细节方面更具优势
- 结合**注意力机制**或**金字塔池化（Pyramid Pooling）**来融合特征
- HWD可能在较浅的网络中表现好，更深层的网络得加上残差连接/跳跃连接等等来保证细节信息传递

LSNet

解决问题

自注意力机制和卷积操作效率低；自注意力对于背景等无关区域给予了不必要的注意，特征聚合效率低；固定卷积核权重限制了不同场景下的适应性和灵活度，对于上下文信息的捕捉不够有效。

创新点



大核：1×1卷积+深度卷积得到$H \times W \times D$特征图；

小核：小卷积捕获细粒度的局部特征，图像划分成$G \times G$ patch，每个小块用小卷积生成特征图聚合；

创新点启发

- 大核感知小核聚合结构，符合认知
- Squeeze-and-Excitation模块，自适应地调节通道的权重来增强重要通道的特征，抑制不重要通道的特征
- 这个是建立一个新的架构，并且轻；学习模块化设计

可以在此基础的改进：

- 大核小核的组合可以优化，能不能大中小核，跨尺度融合
- 论文用了MHSA，可不可以加入别的注意力：**非局部注意力**（Non-Local Attention）或**局部自注意力**（Local Attention）
- SE模块换其他自适应模块

WTConv

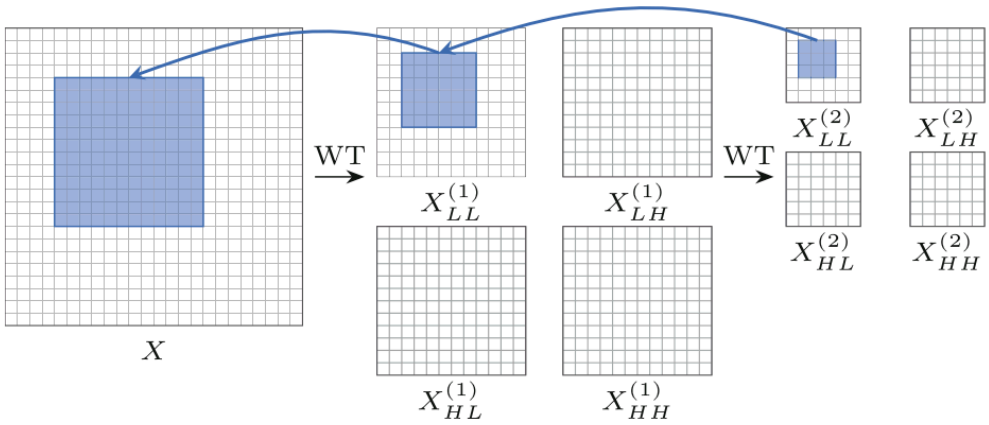
解决问题

感受野大小限制：用大卷积核增加感受野的同时，参数量也成比例增加；和ViT全局特征融合相比，CNN只关注局部

WTConv让感受野扩大的同时，计算量只随着kernal对数增长

创新点

就是把传统conv换成了WTConv



创新点启发

就是根据频域特点（信号知识）来设计卷积

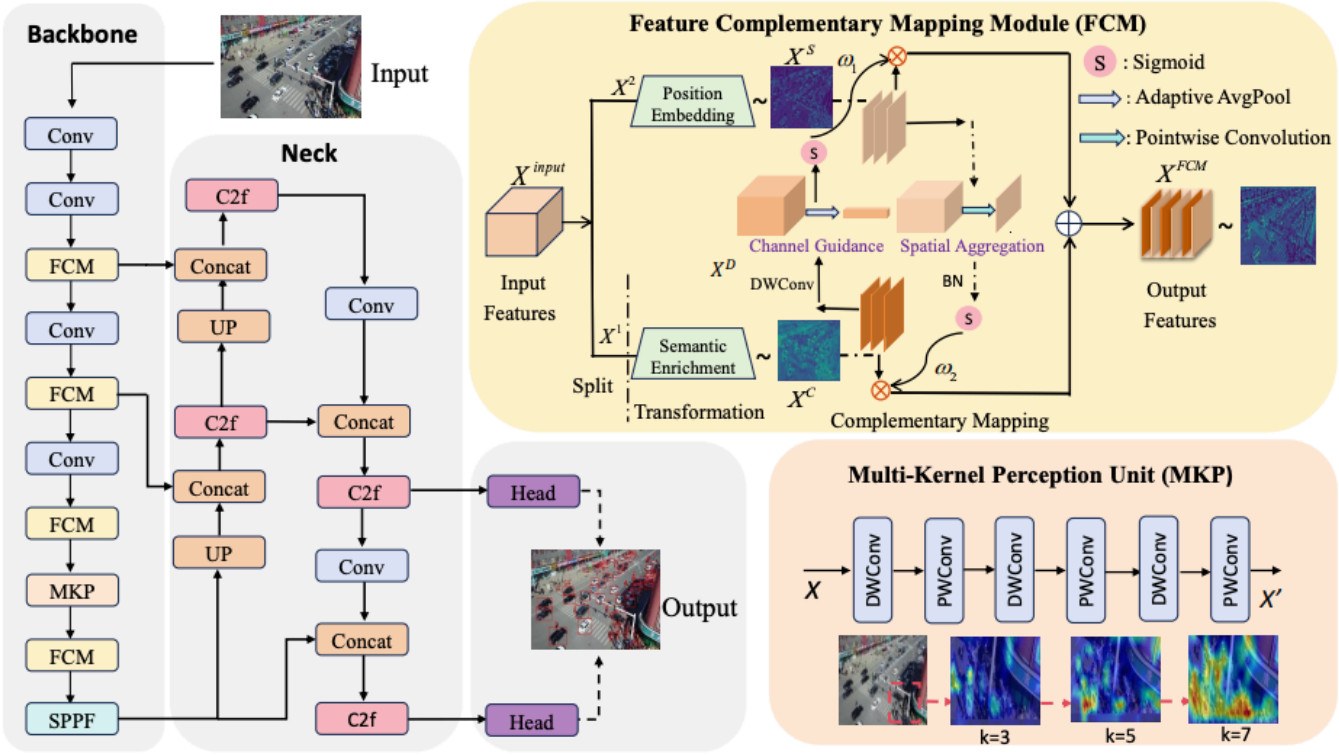
FBRT-YOLO

解决问题

小目标检测难：在深层网络下采样中小目标丢失；

实时性和精度的平衡：边缘设备算力有限

创新点



特征互补映射模块（FCM）：让小目标空间位置传递到深层

- 输入特征$F \in \mathbb{R}^{C\times H\times W}$分成两个分支：

    - 语义分支用标准卷积（3×3或更大）增强上下文
    - 空间分支用1×1卷积保留空间位置关系
- 语义分支引入通道注意力，空间分支引入空间注意力
- 两分支加权融合

多核感知单元（MKP）：不增大计算的前提下扩大/多样化感受野，跨尺度的空间关系

- 用不同kernal大小的深度卷积串联，并在不同核之前加上1×1卷积(point-wise)做通道/尺度信息融合，之后接位置嵌入和检测头

对yolo的backbone优化，减少了不必要的恶部分，用分组卷积+逐点卷积代替标准卷积，减少参数

创新点启发

FCM把浅层的位置信息隐式映射到高维向量，引导深层语义特征补充空间信息，提高对小目标的定位（对经典FPN的补充）

MKP多尺度卷积核同时捕获不同尺寸特征，再融合

下采样阶段用分组卷积+逐点卷积，轻

可以做的改进：

- FCM中用轻量的transformer，浅层空间特征通过全局建模补偿深层语义
- 固定的MKP可以用动态的卷积来替换掉