基于注意力模块改进的 YOLOv5 通用目标检测器

吴靖 23150224

1 引言

目标检测是计算机视觉中的一项核心任务，旨在识别图像中的所有物体，并为其分配相应的类别标签和边界框位置。近年来，深度学习方法特别是卷积神经网络（CNN）在目标检测领域取得了显著的进展，推动了目标检测技术向更高的精度和效率发展。

1.1 Fast-RCNN

通过改进 R-CNN 提高了计算效率和精度。与 R-CNN 需要为每个候选区域单独计算特征不同，Fast R-CNN 只对整个图像进行一次卷积神经网络（CNN）前向传播，生成共享的特征图。然后，利用 ROI Pooling 将每个候选区域的特征映射为固定大小的特征向量，进行分类和边界框回归 [1]。此外，Fast R-CNN 支持端到端训练，优化了分类和回归任务。

1.2 SSD

SSD 属于单阶段检测器，通过在多个尺度的特征图上进行边界框预测来处理不同大小的目标，该网络结合了来自具有不同分辨率的多个特征图的预测 [2]。其核心创新是利用卷积神经网络提取多层特征，并在每个特征图上进行多个默认框（Default Boxes）预测，以处理不同尺寸的物体。SSD 不仅提高了检测速度，还能够较好地平衡精度和效率，尤其是在多尺度物体检测中表现出色。

1.3 RetinaNet

RetinaNet 是一种单阶段目标检测算法，旨在解决类别不平衡问题。与其他单阶段方法不同，RetinaNet 引入了创新的 Focal Loss 损失函数，能够聚焦于难以检测的目标，减少对背景类的过度关注，从而显著提升了对小物体和难检测物体的精度 [3]。RetinaNet 采用类似于 SSD 的多尺度特征图，通过这些特征图进行目标检测，并在每个位置预测多个边界框。该算法在保持较高检测速度的同时，能够在精度上超过许多传统单阶段检测器，尤其在处理类别不平衡和复杂场景时表现优越。

1.4 YOLO

YOLO 是一种实时目标检测算法，其核心思想是将目标检测问题转化为回归问题，通过在整个图像上一次性预测所有物体的类别和边界框。YOLO 将图像划分为网格，每个网格负责预测物体的类别概率和多个边界框坐标，从而实现端到端的检测 [4]。该算法的最大优势在于速度极快，能够实时处理视频流和动态场景，适用于嵌入式设备和实时监控等应用。

1.5 DETR

DETR 是一种基于 Transformer 架构的目标检测模型，旨在通过端到端的方式简化传统目标检测流程。它首次将 Transformer 应用于目标检测任务，利用自注意力机制有效建模图像中物体之间的关系。DETR 通过将输入图像编码为一系列特征向量，并与一组可学习的 query 结合，生成目标的边界框和类别预测。与传统检测方法相比，DETR 消除了复杂的区域提议生成和后处理步骤，实现了更高效的检测性能。此外，DETR 在多种数据集上表现出色，显示出其在目标检测任务中的潜力和优势。[5]

2 实验目标

2.1 实验目的

探究注意力模块对模型性能的影响，通过添加注意力模块提升模型精度

2.2 网络输入输出

输入：[48,3,540,640] [batch,channel,height,width]

输出：[48,(class_num+5)*3,20,20]、[48,(class_num+5)*3,40,40]、[48,(class_num+5)*3,80,80]

3 YOLOv5s 网络架构

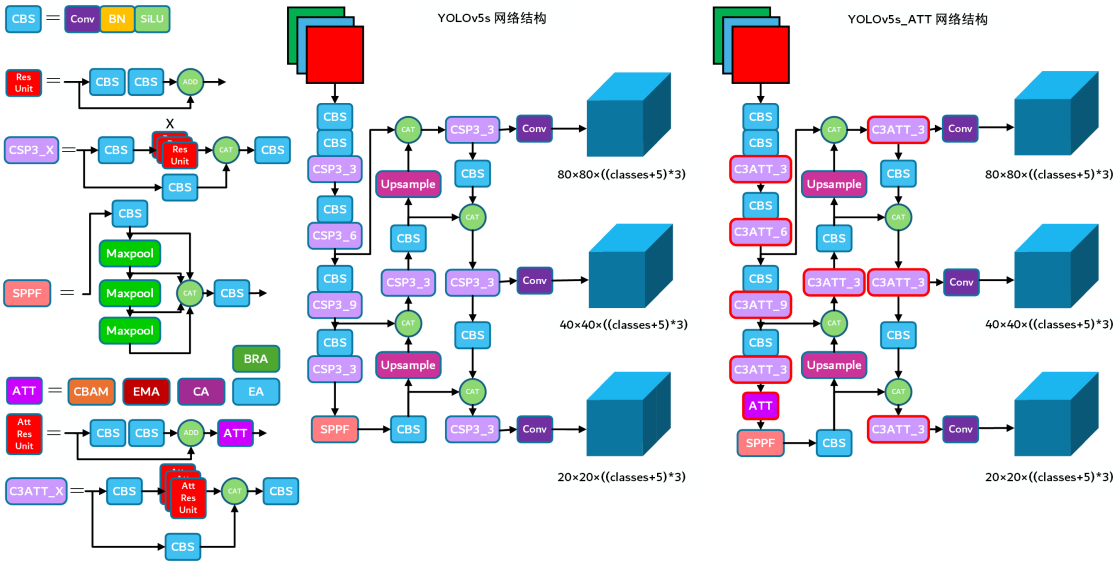


图 1: YOLOv5s 与 YOLOv5s_att

3.1 模块添加

1. 在 backbone 与 neck 之间添加注意力模块
2. 在 C3 模块的残差单元中添加注意力模块
3. 其他添加方法

3.2 实验处理

1. 所有模型不进行微调
2. epoch=100，平均耗时 10h
3. 使用种子固定随机初始化
4. 超参数配置大体一致

3.3 添加示例

```yaml
backbone:
  # [from, number, module, args]
  [[-1, 1, Conv, [64, 6, 2, 2]],
   [-1, 1, Conv, [128, 3, 2]],  #
   [-1, 3, C3, [128]],
   [-1, 1, Conv, [256, 3, 2]],  #
   [-1, 6, C3, [256]],
   [-1, 1, Conv, [512, 3, 2]],  #
   [-1, 9, C3, [512]],
   [-1, 1, Conv, [1024, 3, 2]],  #
   [-1, 3, C3, [1024]],
   [-1, 1, SPPF, [1024, 5]],  # 9
  ]
```

(a) YOLOv5s

```yaml
backbone:
  # [from, number, module, args]
  [[-1, 1, Conv, [64, 6, 2, 2]],
   [-1, 1, Conv, [128, 3, 2]],  #
   [-1, 3, C3, [128]],
   [-1, 1, Conv, [256, 3, 2]],  #
   [-1, 6, C3, [256]],
   [-1, 1, Conv, [512, 3, 2]],  #
   [-1, 9, C3, [512]],
   [-1, 1, Conv, [1024, 3, 2]],  #
   [-1, 3, C3, [1024]],
   [-1, 1, EMA, [1024]],
   [-1, 1, SPPF, [1024, 5]],  # 9
  ]
```

(b) YOLOv5s_EMA

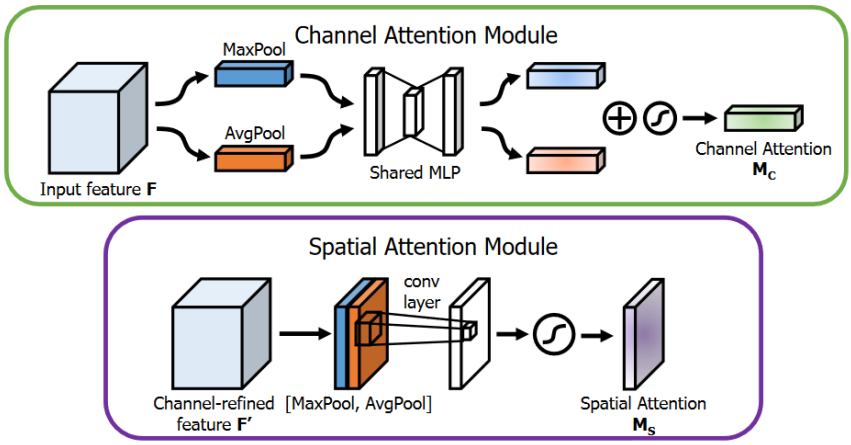
```yaml
backbone:
  # [from, number, module, args]
  [[-1, 1, Conv, [64, 6, 2, 2]],
   [-1, 1, Conv, [128, 3, 2]],  #
   [-1, 3, C3EMA, [128]],
   [-1, 1, Conv, [256, 3, 2]],  #
   [-1, 6, C3EMA, [256]],
   [-1, 1, Conv, [512, 3, 2]],  #
   [-1, 9, C3EMA, [512]],
   [-1, 1, Conv, [1024, 3, 2]],  #
   [-1, 3, C3EMA, [1024]],
   [-1, 1, EMA, [1024]],
   [-1, 1, SPPF, [1024, 5]],  # 9
  ]
```

(c) YOLOv5s_C3EMA

4 注意力模块

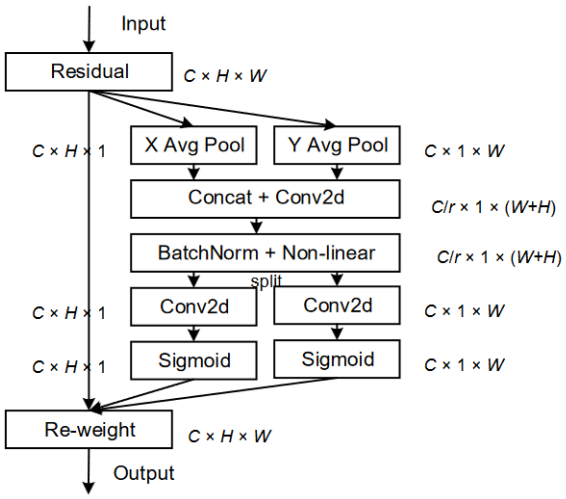
模块	类型	来源	时间
CBAM	通道、空间	ECCV	2018
CA	坐标	CVPR	2021
EA	外部注意力	TPAMI	2022
EMA	通道	ICASSP(CCF-B)	2023
BRA	自注意力	CVPR	2023

4.1 CBAM(Convolutional Block Attention)



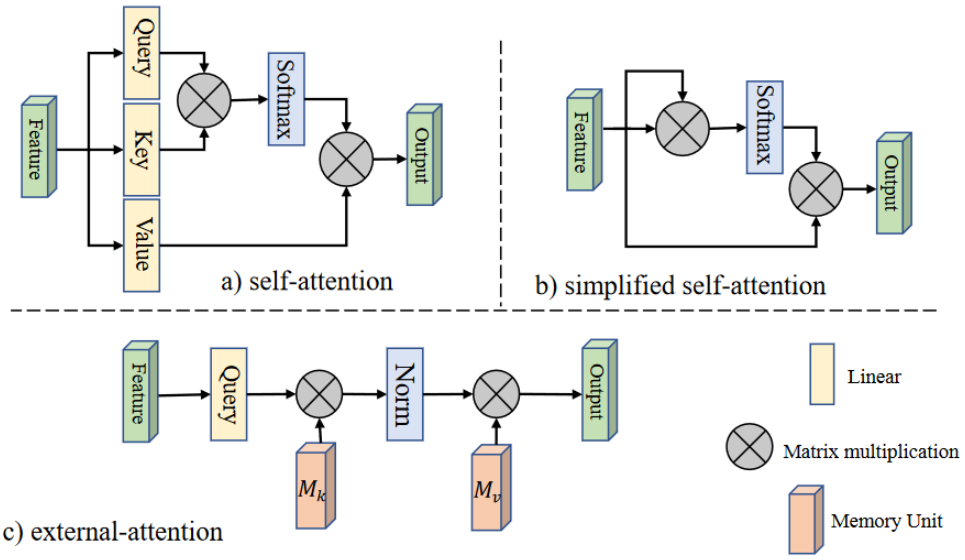
CBAM 是一种轻量级的注意力机制模块，旨在增强卷积神经网络在图像识别和目标检测等任务中的表现。CBAM 通过两种关键的注意力机制——通道注意力和空间注意力——来提升特征图的表达能力。通道注意力模块通过全局平均池化和最大池化生成通道描述符，并利用全连接层对各个通道进行加权，从而突出对任务最重要的通道。空间注意力模块则通过压缩通道信息，生成一个二维的空间注意力图，并通过卷积操作对特征图进行加权，强调图像中的关键区域 [6]。CBAM 在提高模型精度的同时，保持了较低的计算开销，能够有效地集成到现有的 CNN 架构中，提升图像分类、目标检测和语义分割等多种视觉任务的性能。

4.2 CA(Coordinate Attention)



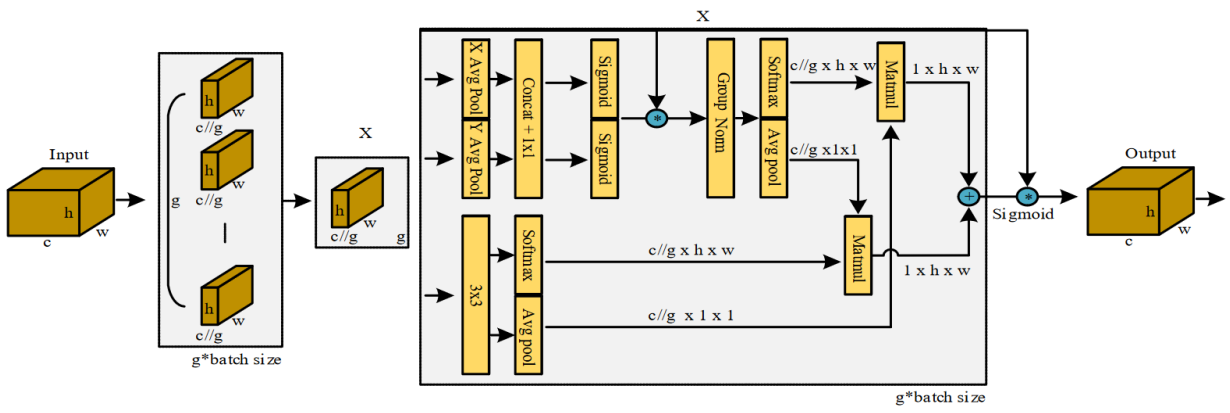
CA 模块通过引入坐标注意力机制，提升了卷积神经网络对空间信息的建模能力。与传统注意力机制不同，CA 模块将空间信息分解为水平和垂直两个方向，通过分别编码这两个方向的坐标信息，生成空间注意力图，并与特征图进行逐元素相乘，从而增强关键区域的特征。该方法不仅减少了计算复杂度，还提高了模型的表达能力。实验表明，CA 模块在图像分类、目标检测等任务中表现优异 [7]，并且在计算效率上具有明显优势，特别适用于资源受限的应用场景。

4.3 EA(External Attention)



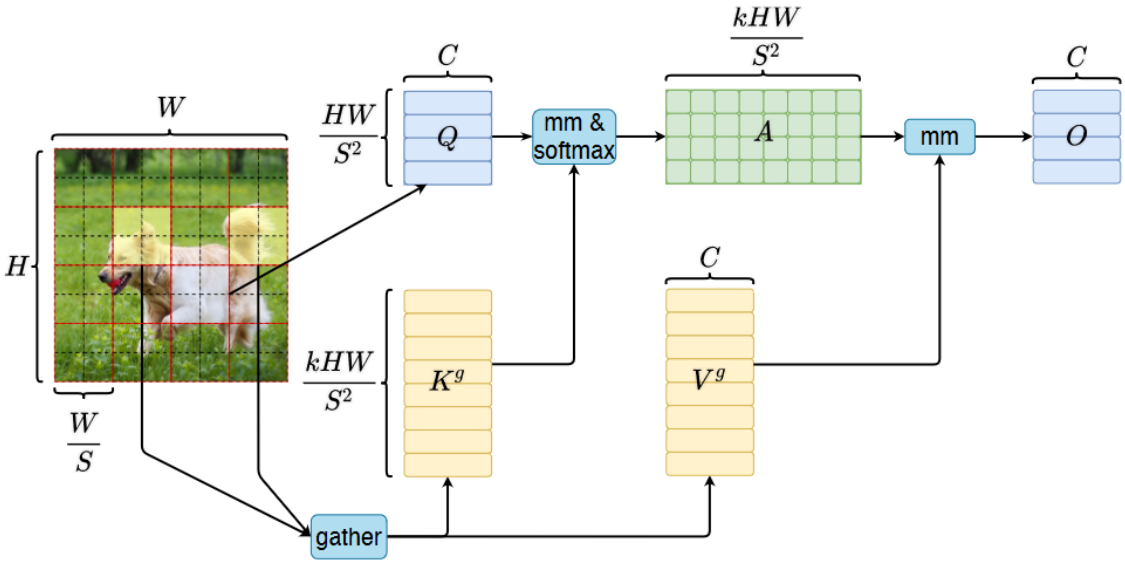
EA 模块是一种新型的注意力机制，旨在提高卷积神经网络或视觉模型在特征提取过程中的表达能力。与传统的注意力机制不同，EA 模块通过引入外部信息来优化模型的特征表示。具体来说，EA 通过利用外部信号（如全局或上下文信息）来强化特征图中有用的信息，从而提升网络对重要特征的关注能力 [8]。

4.4 EMA(Efficient Multi-Scale Attention)



提出了一个高效的多尺度注意力模块，通过跨空间学习（Cross-Spatial Learning）来增强特征的表示能力。EMA 模块在不同尺度上聚焦于重要区域，利用跨空间信息进行特征融合，从而有效捕捉多尺度的上下文信息。与传统的单一尺度注意力机制不同，EMA 通过在空间域和尺度之间的交互，能够自适应地选择关键特征，减少冗余信息，同时提高计算效率。该模块在图像分类、目标检测等任务中表现出色，显著提升了模型的性能和泛化能力 [9]。

4.5 BRA(BiLevel Routing Attention)



BRA 模块是一种新型的双层路由注意力机制，旨在提升神经网络对特征的自适应选择与加权能力。与传统的单向注意力机制不同，BRA 模块通过双向信息流动，使得输入特征能够通过双层路由机制进行有效的交互和调整。具体而言，BRA 模块通过上下文信息与特征之间的双向反馈，增强了特征选择的精准性和信息传递的效率，从而改善了网络的表示能力和推理性能。该模块的关键特点在于其双向路由机制和多尺度特征融合能力。通过双向交互，BRA 模块能够自动聚焦于重要的特征，并对冗余信息进行抑制。此外，BRA 模块通过跨层信息传递，能够提高特征的表达能力，使网络在处理复杂视觉任务时表现更加优异。实验结果表明，BRA 模块在图像分类、目标检测和语义分割等视觉任务中取得了显著的性能提升，尤其在处理多尺度信息和复杂依赖关系时具有明显优势 [10]。

5 实验过程

先按照第一种添加方式在 backbone 和 neck 之间分别添加了五个注意力模块，结果只有 CA 和 BRA 模块有精度提升，CBAM、EA、EMA 分别有不同程度的精度下降。再尝试第二种添加方法，在 C3 模块的残差单元中添加注意力模块，尝试了 CBAM、CA、EMA 模块，结果 C3CBAM 模块仍然使精度下降了，但 C3EMA 模块使 mAP@.5、mAP@.5:.95 分别上升了 0.9，1.3，C3CA 模块也使 mAP 有较大提升。于是再次尝试 EMA 模块，添加在每个 C3 模块的输出后，一共添加 4 层，结果和 C3EMA 模块性能相仿。在第一种添加方式中 BRA 模块展现出较强的性能，在 backbone 末尾叠加三层 BRA 模块，结果模型性能下降较大。再使用 C3CA 模块和 C3EMA 模块与其搭配，这样的添加方法对模型性能有较大提升。

6 实验结果

Model	Params	mAP@.5	mAP@.5:.95
YOLOv5s	7.073M	74.8	46.7
YOLOv5s+CBAM	7.096M	74.3(-0.5)	46.5(-0.2)
YOLOv5s+CA	7.089M	75.3(+0.5)	47.5(+0.8)
YOLOv5s+EA	7.664M	74.3(-0.5)	46.4(-0.3)
YOLOv5s+EMA	7.114M	74.7(-0.1)	47.0(+0.3)
YOLOv5s+BRA	8.129M	75.4(+0.6)	47.8(+1.1)
YOLOv5s+ODConv	7.179M	73.5(-1.3)	45.0(-1.7)
YOLOv5s+C3CBAM	7.135M	74.2(-0.6)	46.5(-0.2)
YOLOv5s+C3CA	7.135M	75.1(+0.3)	47.6(+0.9)
YOLOv5s+C3EMA	7.150M	75.7(+0.9)	48.0(+1.3)
YOLOv5s+3BRA	10.240M	73.5(-1.3)	46.5(-0.2)
YOLOv5s+4EMA	7.128M	75.8(+1.0)	47.5(+0.8)
YOLOv5s+BRA+C3CA	8.165M	75.3(+0.5)	47.8(+1.1)
YOLOv5s+BRA+C3EMA	8.165M	75.6(+0.8)	48.1(+1.4)
YOLOv5s+4EC	7.864M	75.8(+1.0)	48.1(+1.4)
YOLOv5s+4EC+C3EMA	7.900M	**76.0**(+1.2)	**48.7**(+2.0)
YOLOv5s+ODConv+4EC+C3EMA	8.006M	75.0(+0.2)	47.2(+0.5)
YOLOv8s	11.143M	81.6	60.9

6.1 结果分析

从实验结果来看，EMA 模块单独加在 backbone 最后会使模型性能下降，但如果加在 C3 中的残差模块里会使得模型性能有较大提升，而在每个 C3 后加 EMA 模块，也能取得相仿的效果。可见寻找一个合适的位置添加注意力模块对其效果的影响很大，合适的位置能很好地提升模型性能，位置选的不合适反而会降低模型性能。考虑到不同的注意力关注的侧重点不同，尝试“并联”了 EMA 与 CA 模块组成 EC 模块，在每个 C3 模块后加 EC 模块，mAP@.5:.95 相较于 EMA 增加 0.6。于是组合 C3EMA 模块与 4EC，取得实验中最高的 mAP，可见各注意力模块之间可以优势互补，达到更好的效果。CBAM 模块在两种添加方式上都让模型性能有所下降，可能是添加位置不当，也可能使该模块不适合本模型。BRA 模块单层的效果不错，但是在相同位置叠加三层却让模型性能下降很大，可见不光是位置，数量的添加也很重要，如果只是暴力叠加，会让模型变得臃肿，性能反而下降。最后额外尝试了 ODConv 卷积模块，将网络中所有普通卷积替换，但指标大幅下降，可见不同的模块有不同的适用场所，使用不当反而降点。

6.2 总结

通过注意力机制改进后的 YOLO 模型，特别是在 EMA 模块引入后，显著提升了目标检测的精度和效率。这一技术不仅在传统的视觉任务（如智能监控、自动驾驶、安防检测等）中具有广泛应用潜力，

还可以扩展到更广泛的领域，推动中国现代科技事业的发展。未来，改进后的 YOLO 算法可在智能医疗、工业自动化、环境监测、无人机巡检等场景中发挥重要作用。例如，在智能医疗中，它能够提高对医疗影像中细微病变的检测精度，助力早期疾病诊断；在工业自动化中，可以应用于生产线的智能检测与缺陷识别，有效提升生产效率与质量控制。这些应用不仅推动了人工智能技术在各行业的普及，也将助力中国在全球科技创新竞争中占据领先地位。

参考文献

[1] R. Girshick, "Fast r-cnn," *arXiv preprint arXiv:1504.08083*, 2015.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[3] T. Lin, "Focal loss for dense object detection," *arXiv preprint arXiv:1708.02002*, 2017.

[4] J. Redmon, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[6] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[7] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.

[8] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5436–5447, 2022.

[9] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau, "Biformer: Vision transformer with bi-level routing attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 323–10 333.