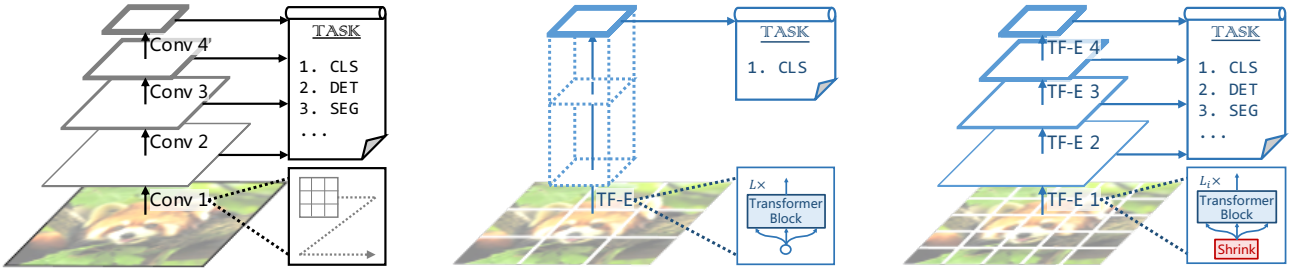
视觉金字塔 Transformer：一个用于稠密预测且无需卷积的多功能主干网络

王文海 [1], 谢恩泽 [2], 李翔 [3], 范登平 [4]✉, 宋恺涛 [3], 梁鼎 [5], 路通 [1]✉, 罗平 [2], 邵岭 [4]
[1] 南京大学 [2] 香港大学 [3] 南京理工大学 [4] 起源人工智能研究院 [5] 商汤科技

https://github.com/whai362/PVT



(a) CNNs: VGG [1], ResNet [2], 等. **(b)** Vision Transformer [3] **(c)** 视觉金字塔 Transformer （本文）

图 1. 不同架构的比较, 其中 “Conv” 和 “TF-E” 分别代表 “卷积” 和 “Transformer 编码器”。(a) 许多 CNN 主干网络使用金字塔结构进行稠密预测，如：目标检测 (DET)，实例和语义分割 (SEG)。(b) 最近提出的 Vision Transformer (ViT) [3] 是专门为图像分类 (CLS) 设计的 “柱状” 结构。(c) 通过整合 CNN 的金字塔结构，本文提出了视觉金字塔 Transformer (PVT)，它可以作为许多计算机视觉任务的多功能主干网络，拓宽了 ViT 的范围和影响。此外，本文的实验还表明，PVT 和 DETR [4] 可轻易结合，构建一个不含卷积的端到端目标检测系统。

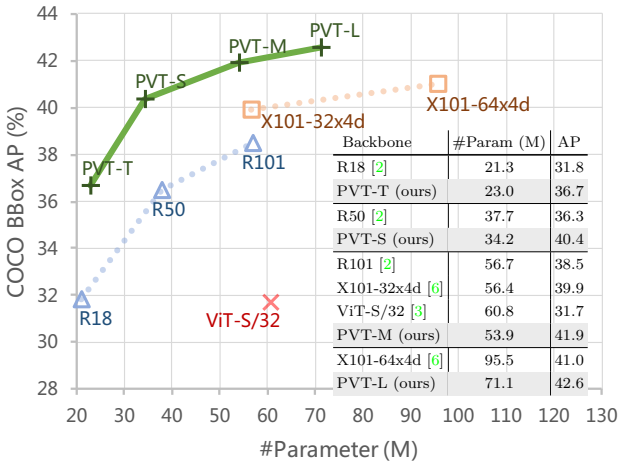
摘要

*尽管卷积神经网络 (CNN) 在计算机视觉方面已经取得了巨大的成功，但本文提出了一种简单且无卷积的主干网络，可以应用于许多稠密预测任务。与最近提出的专门为图像分类而设计的 Vision Transformer (ViT) 不同，本文提出的视觉金字塔 Transformer (Pyramid Vision Transformer, **PVT**)，克服了将 Transformer 移植到各种稠密预测任务中时遇到的困难。与现有的技术相比，PVT 具有以下几个优点。* ***(1)*** *与通常会产生低分辨率输出并且导致高计算和内存成本的 ViT 不同，PVT 不仅可以在密集分块的图像上进行训练以及实现高分辨率的输出，这对于稠密预测任务很重要，而且还可以使用渐进收缩的金字塔结构来减少高分辨率特征图的计算量。* ***(2)*** *PVT 继承了 CNN 和 Transformer 的优点，使其可以成为各种视觉任务的统一主干网络，且无需卷积，能够直接替代 CNN 主干网络。* ***(3)*** *本文通过大量的实验对 PVT 的有效性进行验证，结果表明它能提高许多下游任务的性能，包括目标检测、实例分割和语义分割。具体来说，在参数量相当的情况下，PVT+RetinaNet 在 COCO 数据集上的 AP 达到 40.4，超过了 ResNet50+RetinNet (36.3 AP) 4.1(见图 2)。我们希望 PVT 可以成为像素级稠密预测任务中可替代且有用的主干网络，并促进未来的研究。*

1. 引言

卷积神经网络 (CNN) 在计算机视觉领域已经取得了显著的成功，这使得它成为了几乎适用于所有任务的主流方法 [1, 2, 6, 7, 8, 9, 10, 11]。尽管如此，本文的工作旨在探索可以替代 CNN 应用于除了图像分类 [12]

✉ 通信作者：范登平 (dengpfan@gmail.com)；路通 (lutong@nju.edu.cn)。 本文为 *ICCV*2021 [5] 的中文翻译版，由陈佳辉翻译，范登平、邢浩哲和王文海校稿。



Backbone	#Param (M)	AP
R18 [2]	21.3	31.8
PVT-T (ours)	23.0	36.7
R50 [2]	37.7	36.3
PVT-S (ours)	34.2	40.4
R101 [2]	56.7	38.5
X101-32x4d [6]	56.4	39.9
ViT-S/32 [3]	60.8	31.7
PVT-M (ours)	53.9	41.9
X101-64x4d [6]	95.5	41.0
PVT-L (ours)	71.1	42.6

图 2. 使用 RetinaNet 进行目标检测时，不同主干网络在 COCO `val2017` 上的性能比较，其中“T”、“S”、“M”和“L”表示本文的 PVT 模型具有微小、小、中和大的尺寸。可以看到，当不同模型之间的参数数量相近时，PVT 的变体模型明显地优于其他相应的模型，如：ResNets (R) [2]、ResNeXts (X) [6] 和 ViT [3]。

以外，如：目标检测 [13, 14]、语义分割 [15] 和实例分割 [13] 等稠密预测任务的主干网络。

由于受到了 Transformer [16] 在自然语言处理方面优异表现的启发，研究者们便对 Transformer 在计算机视觉中的应用进行了探索。例如，一些工作 [4, 17, 18, 19, 20, 21] 将视觉任务建模为带有可学习查询参数的字典查找问题，并使用 Transformer 解码器作为 CNN 主干上处理特定任务的头部网络 (Task-Specific Head)。虽然一些现有技术也将注意力模块 [22, 23, 24] 引入到 CNN 中，但据我们所知，*很少有研究人员开发无卷积的纯 Transformer 主干网络，来解决计算机视觉中的稠密预测任务。*

最近，Dosovitskiy 等人 [3] 将视觉 Transformer (ViT) 用于图像分类。他们用一个无卷积模型来代替 CNN 主干网络，这是一个有趣且非常有意义的尝试。如图 1 (b) 所示，ViT 具有柱状结构，它以粗粒度图像块作为输入[1]。虽然 ViT 适用于图像分类，但是直接将它应用于像素级稠密预测是具有挑战性的，例如，目标检测和分割。因为,(1) 其输出的特征图尺度单一，分辨率低；(2) 即使是对于常见的输入尺寸，它的计算和内存成本也相对较高 (如：在 COCO 基准 [13] 中，较短的边的像素值为 800)。

为了解决上述限制，本文提出了一个纯 Transformer 主干网络，称为“视觉金字塔 Transformer (PVT)”，在图像级预测以及像素级稠密预测等多个下游任务中，它可以替代 CNN。具体来说，如图 1 (c) 所示，本文的 PVT 通过以下的方法克服了传统 Transformer 的困难：(1) 以细粒度图像块作为输入 (每个图像块为 4×4 像素)，来学习高分辨率表示，这对稠密预测任务至关重要；(2) 引入渐进式收缩金字塔，随着网络的深化，来减少 Transformer 的序列长度，显著地降低计算成本；(3) 采用空间降维注意力 (spatial-reduction attention SRA) 层，进一步降低学习高分辨率特征时的资源消耗。

总的来说，PVT 有以下优点。(1) 与传统的 CNN 主干网络相比 (见图 1 (a))，传统的主干网络的局部感受野会随着网络深度的增加而变大，本文的 PVT 会产生一个全局的感受野，更适合检测和分割。(2) 相对于 ViT(如图 1 (b))，本文的方法拥有先进的金字塔结构，因此更容易插入到许多具有代表性的稠密预测流程中，如：RetinaNet [9] 和 Mask R-CNN [8]。(3) 可以将 PVT 与其他用于特定任务的 Transformer 解码器 (如 PVT+DETR [4]) 结合起来，构建一个可用在目标检测上的无卷积模型。*据我们所知，这是第一个完全无卷积的目标检测模型。*

本文的主要贡献如下：

(1) 本文提出的视觉金字塔 Transformer (PVT)，是第一个为各种像素级稠密预测任务设计的纯 Transformer 主干网络。结合 PVT 和 DETR，可构建一个不包含卷积和手工组件 (如：密集锚框 (Dense Anchors) 和非最大值抑制 (NMS)) 的端到端目标检测系统。

(2) 通过设计渐进式收缩金字塔和空间降维注意力机制 (Spatial Reduction Attention, SRA)，本文克服了将 Transformer 移植到稠密预测任务中时遇到的许多困难。这些都能够减少 Transformer 的资源消耗，使 PVT 能够灵活地学习多尺度和高分辨率特征。

(3) 本文在多个不同的任务上评估了 PVT，包括图像分类、目标检测、实例和语义分割，并将它与 ResNets [2] 和 ResNeXts [6] 进行了比较。如图 2 所

[1]由于资源限制，ViT 不能使用细粒度的图像块作为输入 (如：每个图像块的大小为 4×4 像素)，而只能使用粗粒度的图像块作为输入 (如：每个图像块的大小为 32×32 像素)，这导致它只能输出低分辨率的图像 (如：步长为 32 像素)

示，与现有技术相比，因为 PVT 具有不同的参数尺度，所以可以持续提高性能。例如，在参数的数量相近时，使用 RetinaNet [9] 进行目标检测，PVT-Small 在 COCO val2017 上，AP 达到了 40.4，比 ResNet50 高出 4.1 (40.4 vs. 36.3)。PVT-Large 的 AP 达到了 42.6，比 ResNeXt101-64x4d 提高 1.6，且参数数量减少 30%。

2. 相关工作

2.1. CNN 主干网络

在视觉识别中，CNN 是深度神经网络的常用框架。标准的 CNN 最初是在 [25] 中提出的，用于区分手写数字。该模型包含具有一定感受野的卷积核，可以捕获有利的视觉上下文。为了保持平移不变性，卷积核的权重在整个图像空间上共享。最近，随着计算资源 (如 GPU) 的快速发展，在大规模图像分类数据集 (如 ImageNet [26]) 上成功训练堆叠卷积块 [27, 1] 已成为可能。例如，GoogLeNet [28] 证明了包含多个内核路径的卷积算子可以实现非常有竞争力的性能。在 Inception 系列 [29, 30]、ResNeXt [6]、DPN [31]、MixNet [32] 和 SKNet [33] 中进一步验证了多路径卷积块的有效性。此外，ResNet [2] 在卷积块中引入了跳跃连接，使得创建/训练非常深的网络成为可能，并在计算机视觉领域获得了令人瞩目的成果。DenseNet [34] 引入了一种密集连接的拓扑结构，它将每个卷积块连接到之前所有的块。最新的进展可以在最近的综述或者调研论文 [35, 36] 中看到。

与成熟的 CNN 不同的是，视觉 Transformer 主干网络仍然处于发展的起步阶段。在这项工作中，本文尝试设计一种适用于大多数视觉任务的新型多功能 Transformer 主干网络，用来扩展视觉 Transformer 的应用范围。

2.2. 稠密预测的任务

预备知识：稠密预测任务的目的是在特征图上进行像素级分类或回归。目标检测和语义分割是两个具有代表性的稠密预测任务。

目标检测：在深度学习时代，CNN [25] 已成为目标检测的主要框架，其中包括单阶段检测器 (如：SSD [37]、RetinaNet [9]、FCOS [38]、GFL [39, 40] 、PolarMask [41] 和 OneNet [42]) 和多阶段检测器 (Faster R-CNN [7]、Mask R-CNN [8]、Cascade R-CNN [43] 和 Sparse R-CNN [44])。大多数流行的目标检测器都是建立在高分辨率图或多尺度特征图上的，因此可以获得良好的检测性能。最近，DETR [4] 和 Deformable DETR [17] 将 CNN 主干网络和 Transformer 解码器结合起来，构建了一个端到端的目标检测器。同样，它们还需要高分辨率或多尺度特征图进行准确的目标检测。

语义分割：CNN 在语义分割中也发挥着重要作用。在早期阶段，FCN [45] 引入了全卷积框架，可以为任意大小的给定图像生成空间分割图。之后，Noh 等人 [46] 引入了反卷积操作，并在 PASCAL VOC 2012 数据集 [47] 上取得了令人印象深刻的表现。受 FCN 的启发，U-Net [48] 被提出并应用于医学图像分割领域，它连接了相同空间尺寸中相应低级和高级特征图之间的信息流。为了探索更丰富的全局上下文表示，Zhao 等人 [49] 在各种池化尺度上设计了一个金字塔池化模块，Kirillov 等人 [11] 基于 FPN [50] 开发了一种称为语义 FPN 的轻量级分割头。最后，DeepLab 系列 [51, 52] 使用膨胀卷积来扩大感受野，同时保持特征图的分辨率。与目标检测方法类似，语义分割模型也依赖于高分辨率图或多尺度特征图。

2.3. 视觉中的自注意力机制和 Transformer

在训练之后，卷积滤波器的权值通常是固定的，因此它们无法根据不同输入来进行动态地调整。基于动态过滤器 [53] 或自注意机制 [16] 的许多方法，已经被用来缓解这个问题。非局部块 [22] 尝试在空间和时间上对远程依赖关系进行建模，这已经被证明有利于准确的视频分类。虽然取得了成功，但非局部操作仍然面临着高计算和内存成本的问题。Criss-cross [54] 通过 criss-cross 路径生成稀疏注意图，进一步降低了复杂性。Ramachandran 等人 [23] 提出了独立自注意力，用局部自注意力单元来代替卷积层。AANet [55] 结合了自注意力机制和卷积运算，实现了具有竞争力的结果。LambdaNetworks [56] 使用了 lambda 层，一种高效的自注意力机制来替换 CNN 中的卷积。DETR [4] 利用 Transformer 解码器，将目标检测建模为具有可学习查询的端到端字典查找问题，成功消除了对手工流程的需要，如 NMS。Deformable DETR [17] 在 DETR 的基础上进一步采用了可变形的注意层，将注意力集中在稀疏的感兴趣区域上，收敛速度更快，性能更好。最近，视觉 Transformer (ViT) [3] 采用纯 Transformer [16] 模

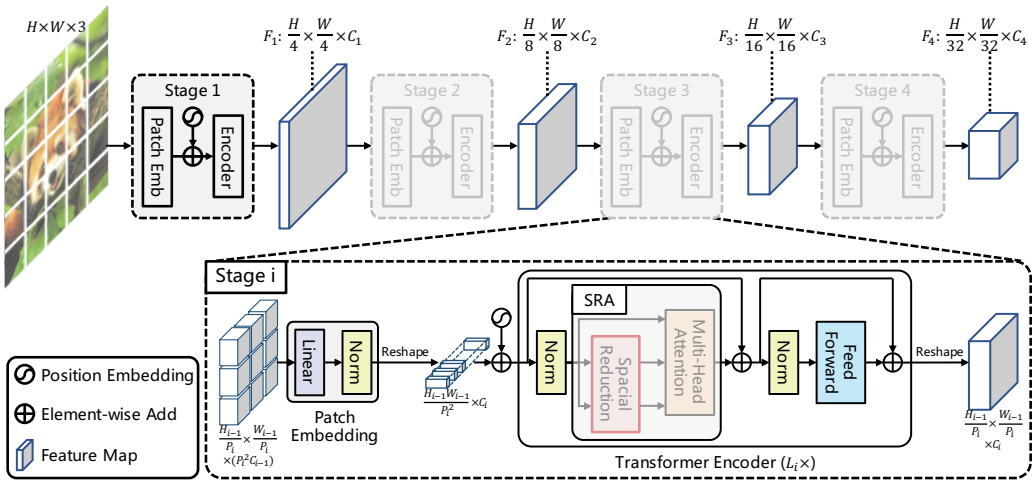


图 3. 视觉金字塔 Transformer (PVT) 的整体架构。整个模型分为四个阶段，每个阶段由一个块嵌入层（Patch Embedding）和一个 L_i 层 Transformer 编码器组成。遵循金字塔结构，四个阶段的输出分辨率从高 (步长为 4 像素) 逐渐缩小到低 (步长为 32 像素)。

型，将图像作为块序列进行分类。DeiT [57] 使用一种新颖的蒸馏方法进一步扩展了 ViT。与以往模型不同的是，本文将金字塔结构引入 Transformer，为密集的预测任务提供了一个纯 Transformer 主干网络，而不是针对特定任务的头部网络或图像分类模型。

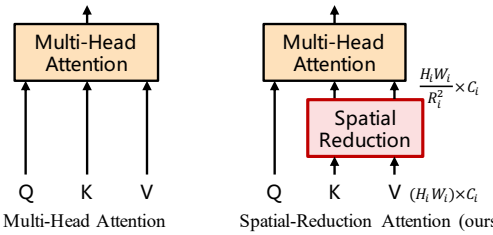
3. 视觉金字塔 Transformer (PVT)

3.1. 总体框架

本文的目标是将金字塔结构引入 Transformer 中，以便它可以为稠密预测任务 (如：目标检测和语义分割) 生成多尺度特征图。PVT 的结构如图 3 所示。与 CNN 主干网络 [2] 类似，本文的方法也有四个阶段，可以生成不同尺度的特征图。所有阶段共享一个相似的结构，这个结构由一个块嵌入层（Patch Embedding）和一个 L_i Transformer 编码器层组成。

在第一阶段，给定一个大小为 $H \times W \times 3$ 的输入图像，本文首先将其划分为 $\frac{HW}{4^2}$ 个图像块[2]，每个图像块的大小为 $4 \times 4 \times 3$。然后，将展平后的图像块进行线性投影，获得大小为 $\frac{HW}{4^2} \times C_1$ 的嵌入式图像块。之后，嵌入的图像块与位置嵌入一起通过具有 L_1 层的 Transformer 编码器，并将输出重新整理成尺寸为 $\frac{H}{4} \times \frac{W}{4} \times C_1$ 的特征图 F_1。同样，使用前一阶段的特征图作为输入，可以得到以下特征图：F_2、F_3 和 F_4，与输入图像相比，其步长分别为 8、16 和 32 个像素。借助特征金字塔 $\{F_1, F_2, F_3, F_4\}$，本文的方法可以直接应用在大多数的下游任务中，如：图像分类、目标检测和语义分割。

[2]同 ResNet，本文保持最高分辨率的输出特征图的步长为 4。



.7

图 4. 多头注意力 (MHA) *vs.* 空间降维注意力 (SRA)。通过空间降维操作，本文的 SRA 的计算/内存成本远低于 MHA。

3.2. Transformer 的特征金字塔

CNN 主干网络 [1, 2] 使用不同的卷积步长来获取多尺度的特征地图，与它不同的是，本文的 PVT 使用一种渐进式收缩策略，通过图像块嵌入层来控制特征图的大小。

这里，本文使用 P_i 表示第 i 阶段的图像块的尺寸。在第 i 阶段开始时，本文首先将输入特征图 $F_{i-1} \in \mathbb{R}^{H_{i-1} \times W_{i-1} \times C_{i-1}}$ 均匀地划分成 $\frac{H_{i-1}W_{i-1}}{P_i^2}$ 个图像块，然后将每个图像块展平并线性投影到一个 C_i 维嵌入向量。在线性投影之后，嵌入的图像块的形状为 $\frac{H_{i-1}}{P_i} \times \frac{W_{i-1}}{P_i} \times$

C_i。其中高度和宽度均缩小为输入的 $\frac{1}{P_i}$ 倍。

这样，本文就可以灵活地调整每个阶段特征图的尺寸，使得为 Transformer 构建特征金字塔成为可能。

3.3. Transformer 编码器

在第 i 阶段的 Transformer 编码器中有一个 L_i 编码层，它由一个注意层和一个前馈层组成 [16]。由于 PVT 需要处理高分辨率 (如：步长为 4) 特征图，本文提出了一个空间降维注意 (SRA) 层，来取代编码器中传统的多头注意 (Multi-Head Attention, MHA) 层 [16]。

与 MHA 类似，本文的 SRA 的输入是：查询 (Query) Q、键 (Key) K、值 (Value) V，输出是优化后的查询特征。不同之处在于，本文的 SRA 在注意力操作之前，减少了 K 和 V 的空间维度 (见图 4)，这大大降低了计算/内存开销。第 i 阶段中的 SRA 流程详细介绍如下：

$$\text{SRA}(Q, K, V) = \text{Concat}(\text{head}_0, ..., \text{head}_{N_i})W^O, \quad (1)$$

$$\text{head}_j = \text{Attention}(QW_j^Q, \text{SR}(K)W_j^K, \text{SR}(V)W_j^V), \quad (2)$$

其中 Concat$(\cdot)$ 是与论文 [16] 一致的连接操作。$W_j^Q \in \mathbb{R}^{C_i \times d_{\text{head}}}$，$W_j^K \in \mathbb{R}^{C_i \times d_{\text{head}}}$，$W_j^V \in \mathbb{R}^{C_i \times d_{\text{head}}}$，和 $W^O \in \mathbb{R}^{C_i \times C_i}$ 是线性投影参数。N_i 是第 i 阶段注意力层的头部编号。因此，每个头部的尺寸 (即 d_{head}) 等于 $\frac{C_i}{N_i}$。SR$(\cdot)$ 是降低输入序列 (K 或 V) 空间维数的操作：

$$\text{SR}(\mathbf{x}) = \text{Norm}(\text{Reshape}(\mathbf{x}, R_i)W^S). \quad (3)$$

其中 $\mathbf{x} \in \mathbb{R}^{(H_i W_i) \times C_i}$ 表示输入序列，R_i 表示第 i 阶段的注意层的缩减比。Reshape$(\mathbf{x}, R_i)$ 负责将输入序列 $\mathbf{x}$ 调整成大小为 $\frac{H_i W_i}{R_i^2} \times (R_i^2 C_i)$ 的序列。$W_S \in \mathbb{R}^{(R_i^2 C_i) \times C_i}$ 是将输入序列的维数降为 C_i 的线性投影。Norm$(\cdot)$ 表示层归一化 [58]。与原始的 Transformer [16] 一样，我们的注意力操作 Attention$(\cdot)$ 的计算公式为：

$$\text{Attention}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{Softmax}(\frac{\mathbf{q}\mathbf{k}^\mathsf{T}}{\sqrt{d_{\text{head}}}})\mathbf{v}. \quad (4)$$

通过这些公式，可以发现本文的注意力操作的计算/内存成本比 MHA 低 R_i^2 倍，所以本文的 SRA 可以在有限的资源下处理较大的输入特征图和特征序列。

3.4. 讨论

与本文的模型最相关的工作是 ViT [3]。在此，本文将讨论两者之间的关系和区别。首先，PVT 和 ViT 都是不包含卷积的纯 Transformer 模型。它们之间的主要区别是金字塔结构。与传统的 Transformer [16] 类似，ViT 的输出序列与输入序列的尺寸相同，这意味着 ViT 的输出是单尺度的 (见图 1 (b))。此外，由于资源有限，ViT 的输入是粗粒度的 (如：图像块的大小为 16 或 32 像素)，因此其输出分辨率相对较低 (如：步长为 16 或 32 像素)。因此，很难将 ViT 直接应用于需要高分辨率或多尺度特征图的稠密预测任务。

本文的 PVT 引入了渐进式收缩金字塔，这打破了 Transformer 的常规。它可以像传统的 CNN 主干网络一样生成多尺度特征图。此外，本文还设计了一个简单但有效的注意力层—SRA，用于处理高分辨率特征图并降低计算/内存成本。通过以上设计，本文的方法与 ViT 相比具有以下优点：1) 更灵活，可在不同阶段生成不同比例/通道的特征图；2) 通用性更强，可以在大多数下游任务模型中轻松插入和应用；3) 对计算/内存更友好，可以处理更高分辨率的特征图或更长的序列。

4. 下游任务中的应用

4.1. 图像级预测

图像分类是图像级预测中最经典的任务。为了提供讨论实例，本文设计了一系列不同规格的 PVT 模型，分别是 PVT-Tiny、-Small、-Medium 和-Large，它们的参数数量分别与 ResNet18、50、101 和 152 相近。补充材料提供了详细的 PVT 系列超参数设置。

对于图像分类，本文按照 ViT [3] 和 DeiT [57] 将可学习的分类令牌 (Classification Token) 附加到最后阶段的输入中，然后使用全连接 (FC) 层在令牌之上进行分类。

4.2. 像素级稠密预测

除了图像级别的预测之外，下游任务还包括稠密预测，它需要在特征图上进行像素级别的分类或回归。在这里，本文讨论了两个典型的任务，即目标检测和语义分割。

本文将 PVT 模型应用于三种具有代表性的稠密预测方法中，即 RetinaNet [9]、Mask R-CNN [8] 和 Semantic FPN [11]。RetinaNet 是一种应用广泛的单阶段检测器，Mask R-CNN 是目前最流行的两阶段实例分割框架，Semantic FPN 是一种普通的语义分割方法，不需要特殊的操作 (如：膨胀卷积)。这些方法作为

Method	#Param (M)	GFLOPs	Top-1 Err (%)
ResNet18* [2]	11.7	1.8	30.2
ResNet18 [2]	11.7	1.8	31.5
DeiT-Tiny/16 [57]	5.7	1.3	27.8
PVT-Tiny (ours)	13.2	1.9	24.9
ResNet50* [2]	25.6	4.1	23.9
ResNet50 [2]	25.6	4.1	21.5
ResNeXt50-32x4d* [6]	25.0	4.3	22.4
ResNeXt50-32x4d [6]	25.0	4.3	20.5
T2T-ViT$_t$-14 [59]	22.0	6.1	19.3
TNT-S [60]	23.8	5.2	18.7
DeiT-Small/16 [57]	22.1	4.6	20.1
PVT-Small (ours)	24.5	3.8	20.2
ResNet101* [2]	44.7	7.9	22.6
ResNet101 [2]	44.7	7.9	20.2
ResNeXt101-32x4d* [6]	44.2	8.0	21.2
ResNeXt101-32x4d [6]	44.2	8.0	19.4
T2T-ViT$_t$-19 [59]	39.0	9.8	18.6
ViT-Small/16 [3]	48.8	9.9	19.2
PVT-Medium (ours)	44.2	6.7	18.8
ResNeXt101-64x4d* [6]	83.5	15.6	20.4
ResNeXt101-64x4d [6]	83.5	15.6	18.5
ViT-Base/16 [3]	86.6	17.6	18.2
T2T-ViT$_t$-24 [59]	64.0	15.0	17.8
TNT-B [60]	66.0	14.1	17.2
DeiT-Base/16 [57]	86.6	17.6	18.2
PVT-Large (ours)	61.4	9.8	18.3

表 1. 图像分类在 ImageNet 验证集上的性能。“Top-1” 表示 Top-1 误差。“#Param” 是指参数的数量。“GFLOPs” 是在输入图片尺寸为 224×224 的情况下计算的。“*” 表示在原始论文的策略下进行训练的方法的性能。

基线，有利于充分地检查不同主干网络的有效性。

实现细节如下: (1) 与 ResNet 一样，本文使用在 ImageNet 上预训练的权值来初始化 PVT 主干网络; (2) 本文使用特征金字塔的输出 $\{F_1, F_2, F_3, F_4\}$ 作为 FPN [50] 的输入，然后将优化后的特征图输入到后续的检测/分割头部网络; (3) 当训练检测/分割模型时，PVT 中没有任何层会被冻结; (4) 因为检测/分割的输入可以是任意分辨率的，所以在 ImageNet 上预训练得到的位置嵌入向量将会失效。因此，本文根据输入的分辨率对预训练得到的位置嵌入向量进行双线性插值。

5. 实验

本文将 PVT 与两个最具代表性的 CNN 主干网络进行比较，即 ResNet [2] 和 ResNeXt [6]。这两个网络被广泛应用在许多下游任务的基准测试中。

5.1. 图像分类

设置: 图像分类实验在 ImageNet 2012 数据集 [26] 上进行，该数据集包括来自 1000 个类别的 128 万张训练图像和 5 万张验证图像。为了公平比较，所有模型都在训练集上进行训练，并报告验证集上的 Top-1 误差。本文遵循 DeiT [57] 的设置，并使用随机裁剪、随机水平翻转 [28]、标签平滑正则化 [29]、mixup [61]、CutMix [62] 和随机擦除 [63] 来进行数据增强。在训练期间，本文设置动量为 0.9、批量大小为 128、AdamW [64] 的权重衰减为 5×10^{-2}，来优化模型。初始学习率设置为 1×10^{-3}，并按照余弦退火策略 [65] 递减。所有模型都是在 8 个 V100 GPU 上从零开始训练 300 轮。为了进行基准对比，在验证集上本文使用中心裁剪，裁剪了一个 224×224 图像块来评估分类的精度。

结果: 从表 1 中可以看出，在相近的参数数量和计算预算下，本文的 PVT 模型优于传统的 CNN 主干网络。例如，当 GFLOPs 大致相同时，PVT-Small 的 Top-1 误差达到 20.2，比 ResNet50 [2] 低 1.3 (20.2 *vs.* 21.5)。同时，在类似或更低的复杂度下，PVT 模型的性能可以与最近提出的基于 Transformer 的模型相媲美，如 ViT [3] 和 DeiT [57] (PVT-Large: 18.3 *vs.* ViT(DeiT)-Base/16: 18.2)。这些结果都是在我们的预料之内的，虽然金字塔结构有利于稠密预测任务，但是给图像分类带来的帮助却不大。

注意，ViT 和 DeiT 均有局限性，因为它们是专门为分类任务设计的模型，所以不适合稠密预测任务，稠密预测任务通常需要有效的特征金字塔。

5.2. 目标检测

设置: 目标检测实验在具有挑战性的 COCO 数据集 [13] 上进行。所有模型在 COCO `train2017` (11.8 万张图像) 上训练，在 `val2017` (5000 张图像) 上评估。本文在两个标准检测器 (即 RetinaNet [9] 和 Mask R-CNN [8]) 上验证 PVT 主干的有效性。在训练之前，本文使用在 ImageNet 上预训练的权重值，来初始化主干网络，使用 Xavier [66] 来初始化新添加的层。本文的模型在 8 个 V100 GPU 上以批量大小为 16 来进行训练，并使用初始学习率为 1×10^{-4} 的 AdamW [64] 进行优化。按照惯例 [9, 8, 67]，本文采用 1× 或 3× 训练策略 (即 12 或 36 轮) 对所有检测模型进行训练。训练图像的短边为 800 像素，长边不超过 1333 像素。当使用 3× 训练策略时，本文将输入图像的较短边随机地调整在 [640, 800] 这个范围内。在测试阶段，输入图像的较短一边固定为 800 像素。

结果: 如表 2 所示，在使用 RetinaNet 进行目标检测时，本文发现在参数数量相近的情况下，基于 PVT 的模型

Backbone	#Param (M)	RetinaNet 1x						RetinaNet 3x + MS					
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet18 [2]	21.3	31.8	49.6	33.6	16.3	34.3	43.2	35.4	53.9	37.6	19.5	38.2	46.8
PVT-Tiny (ours)	23.0	36.7(+4.9)	56.9	38.9	22.6	38.8	50.0	39.4(+4.0)	59.8	42.0	25.5	42.0	52.1
ResNet50 [2]	37.7	36.3	55.3	38.6	19.3	40.0	48.8	39.0	58.4	41.8	22.4	42.8	51.6
PVT-Small (ours)	34.2	40.4(+4.1)	61.3	43.0	25.0	42.9	55.7	42.2(+3.2)	62.7	45.0	26.2	45.2	57.2
ResNet101 [2]	56.7	38.5	57.8	41.2	21.4	42.6	51.1	40.9	60.1	44.0	23.7	45.0	53.8
ResNeXt101-32x4d [6]	56.4	39.9(+1.4)	59.6	42.7	22.3	44.2	52.5	41.4(+0.5)	61.0	44.3	23.9	45.5	53.7
PVT-Medium (ours)	53.9	41.9(+3.4)	63.1	44.3	25.0	44.9	57.6	43.2(+2.3)	63.8	46.1	27.3	46.3	58.9
ResNeXt101-64x4d [6]	95.5	41.0	60.9	44.0	23.9	45.2	54.0	41.8	61.5	44.4	25.2	45.4	54.6
PVT-Large (ours)	71.1	42.6(+1.6)	63.7	45.4	25.8	46.0	58.4	43.4(+1.6)	63.6	46.1	26.1	46.0	59.5

表 2. 目标检测在 COCO `val2017` 上的性能。“MS” 表示使用多尺度训练 [9, 8]。

Backbone	#Param (M)	Mask R-CNN 1x						Mask R-CNN 3x + MS					
		AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet18 [2]	31.2	34.0	54.0	36.7	31.2	51.0	32.7	36.9	57.1	40.0	33.6	53.9	35.7
PVT-Tiny (ours)	32.9	36.7(+2.7)	59.2	39.3	35.1(+3.9)	56.7	37.3	39.8(+2.9)	62.2	43.0	37.4(+3.8)	59.3	39.9
ResNet50 [2]	44.2	38.0	58.6	41.4	34.4	55.1	36.7	41.0	61.7	44.9	37.1	58.4	40.1
PVT-Small (ours)	44.1	40.4(+2.4)	62.9	43.8	37.8(+3.4)	60.1	40.3	43.0(+2.0)	65.3	46.9	39.9(+2.8)	62.5	42.8
ResNet101 [2]	63.2	40.4	61.1	44.2	36.4	57.7	38.8	42.8	63.2	47.1	38.5	60.1	41.3
ResNeXt101-32x4d [6]	62.8	41.9(+1.5)	62.5	45.9	37.5(+1.1)	59.4	40.2	44.0(+1.2)	64.4	48.0	39.2(+0.7)	61.4	41.9
PVT-Medium (ours)	63.9	42.0(+1.6)	64.4	45.6	39.0(+2.6)	61.6	42.1	44.2(+1.4)	66.0	48.2	40.5(+2.0)	63.1	43.5
ResNeXt101-64x4d [6]	101.9	42.8	63.8	47.3	38.4	60.6	41.3	44.4	64.9	48.8	39.7	61.9	42.6
PVT-Large (ours)	81.0	42.9(+0.1)	65.0	46.6	39.5(+1.1)	61.9	42.5	44.5(+0.1)	66.0	48.3	40.7(+1.0)	63.4	43.7

表 3. 目标检测和实例分割在 COCO `val2017` 上的性能。AP^b 和 AP^m 分别表示 box AP 和 mask AP。

.

Backbone	Semantic FPN		
	#Param (M)	GFLOPs	mIoU (%)
ResNet18 [2]	15.5	32.2	32.9
PVT-Tiny (ours)	17.0	33.2	35.7(+2.8)
ResNet50 [2]	28.5	45.6	36.7
PVT-Small (ours)	28.2	44.5	39.8(+3.1)
ResNet101 [2]	47.5	65.1	38.8
ResNeXt101-32x4d [6]	47.1	64.7	39.7(+0.9)
PVT-Medium (ours)	48.0	61.0	41.6(+2.8)
ResNeXt101-64x4d [6]	86.4	103.9	40.2
PVT-Large (ours)	65.1	79.6	42.1(+1.9)
PVT-Large* (ours)	65.1	79.6	44.8

表 4. ADE20K 验证集上不同主干的语义分割性能。“GFLOPs” 是在输入图片尺寸为 512×512 的情况下计算的。“*” 表示 320K 次迭代训练和多尺度翻转测试。

明显优于同类模型。例如，在 1× 训练计划下，PVT-Tiny 的 AP 比 ResNet18 高出 4.9 (36.7 *vs.* 31.8)。此外，PVT-Large 通过 3× 训练计划和多尺度训练，获得了 43.4 的最佳 AP，超过了 ResNeXt101-64x4d (43.4 *vs.* 41.8)，并且本文的参数数量减少了 30%。这些结果表明，本文的 PVT 可以很好地替代 CNN 主干网络，来完成目标检测任务。在基于 Mask R-CNN 的实例分割实验中也得到了类似的结果，如表 3 所示。PVT-Tiny 在 1× 训练策略下实现了 35.1 mask AP (AP^m)，比 ResNet18 (35.1 *vs.* 31.2) 高出 3.9，甚至比 ResNet50 (35.1 *vs.* 34.4) 高出 0.7。PVT-Large 获得的最佳 AP^m 为 40.7，比 ResNeXt101-64x4d 高出 1.0 (40.7 *vs.* 39.7)，并且参数数量减少了 20%。

Method	DETR (50 Epochs)					
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet50 [2]	32.3	53.9	32.3	10.7	33.8	53.0
PVT-Small (ours)	34.7(+2.4)	55.7	35.4	12.0	36.4	56.7

表 5. 纯 Transformer 目标检测的性能。本文通过结合 PVT 和 DETR 构建了一个纯 Transformer 检测器，其 AP 比基于 ResNet50 [2] 的原始 DETR [4] 高 2.4。

5.3. 语义分割

设置。本文选择 ADE20K [15]，一个具有挑战性的场景解析数据集，来对语义分割的性能进行基准测试。ADE20K 包含 150 个细粒度语义类别，分别有 20210、2000 和 3,352 张图像用于训练、验证和测试。本文在语义 FPN [11] 的基础上对 PVT 主干网络进行评估，这是一种不包含膨胀卷积的简单分割方法 [68]。在训练阶段，使用在 ImageNet [12] 上预训练的权重，来初始化主干网络，对于其他新添加的层，使用在 Xavier [66] 上预训练的权重，来进行初始化。本文使用初始学习率为 1e-4 的 AdamW [64] 来优化模型。按照惯例 [11, 51]，本文在 4 个 V100GPU 上对模型进行 80k 次迭代训练，其中批量大小为 16。学习率按照幂为 0.9 的多项式衰减策略进行衰减。为了适合训练，本文将图片随机裁剪为 512 × 512 像素，在测试时，将图片较短的一边调整为 512 像素。

结果。如表 4 所示，在使用 Semantic FPN [11] 进行语义分割时，基于 PVT 的模型的性能，始终优于

Method	#Param (M)	RetinaNet 1x					
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ViT-Small/4 [3]	60.9	Out of Memory					
ViT-Small/32 [3]	60.8	31.7	51.3	32.3	14.8	33.7	47.9
PVT-Small (ours)	34.2	40.4	61.3	43.0	25.0	42.9	55.7

表 6. ViT 和使用 RetinaNet 的 PVT 在目标检测任务上的性能比较。因为是小的图像块 (即每个块的尺寸为 4×4 像素)，所以 ViT-Small/4 耗尽了 GPU 的内存。在 COCO`val2017` 上，ViT-Small/32 的 AP 为 31.7 ，比本文的 PVT-Small 低 8.7。

基于 ResNet [2] 或 ResNeXt [6] 的模型。例如，在参数和 GFLOPs 的数量几乎相同的情况下，本文的 PVT-Tiny/Small/Medium 的 mIoU 至少比 ResNet-18/50/101 高出 2.8。此外，虽然本文的 PVT-Large 的参数数量和 GFLOPs 比 ResNeXt101-64x4d 低 20%，但 mIoU 高出 1.9 (42.1 *vs.* 40.2)。通过较长的训练策略和多尺度测试，PVT-Large+Semantic FPN 的最佳 mIoU 为 44.8，非常接近 ADE20K 基准的最好性能。注意，语义 FPN 只是一个简单的分割头。这些结果表明，与 CNN 主干网络相比，本文的 PVT 主干网络能够更好地提取语义分割特征，这得益于全局注意力机制。

5.4. 纯 Transformer 目标检测

为实现无卷积，本文将 PVT 与基于 Transformer 的检测—DETR [4] 简单地结合起来，构建了一个用于目标检测的纯 Transformer 流程。以 1×10^{-4} 的初始学习率，在 COCO `train2017` 上对模型进行了 50 轮的训练。在第 33 轮的时候，学习速率除以 10。本文使用随机翻转和多尺度训练来进行数据增强。所有其他的实验设置与第 5.2 节相同。如表 5 所示，基于 PVT 的 DETR 在 COCO `val2017` 上，AP 达到了 34.7，比原来基于 ResNet50 的 DETR 高 2.4 (34.7 *vs.* 32.3)。结果表明，纯 *Transformer* 检测器也能很好地完成目标检测任务。补充材料中，本文也尝试将一个纯 Transformer 模型 PVT+Trans2Seg [18] 用于语义分割。

5.5. 消融实验

设置：本文在 ImageNet [12] 和 COCO [13] 数据集上进行消融研究。ImageNet 上的实验设置与第 5.1 节中的设置相同。对于 COCO 数据集，所有模型均采用 1× 训练策略 (即 12 轮) 进行训练，不采用多尺度训练，其他设置参照章节 5.2 。

金字塔结构：当将 Transformer 应用于稠密预测任务时，金字塔结构是至关重要的。ViT(参见图 1 (b)) 是一个柱状框架，其输出是单尺度的。这导致当使用粗图像块 (如：32×32 像素的图像块) 作为输入时，输出的特征图分辨率较低，检测性能较差 (在 COCO `val2017` 上 AP 为 31.7)[3]，如表 6 所示。当像 PVT 一样使用细粒度的图像块作为输入时，ViT 将耗尽 GPU 内存 (32G)。本文使用一个渐进收缩金字塔来避免这个问题。具体来说，本文的模型可以处理浅层中的高分辨率特征图和深层中的低分辨率特征图。因此，它在 COCO `val2017` 上 AP 为 40.4，比 ViT-Small/32 高出 8.7(40.4 *vs.* 31.7)。

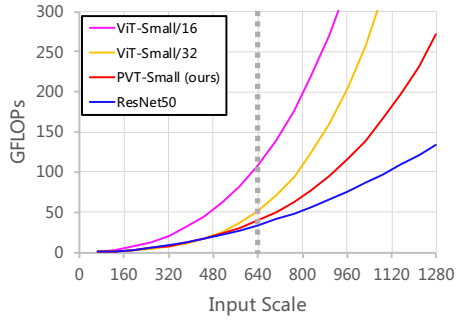


图 5. 在不同的输入尺寸下，模型的 GFLOPs。GFLOPs 的增长率:ViT-Small/16 [3]>ViT-Small/32 [3]>PVT-Small (ours)>ResNet50 [2].

计算开销：随着输入尺寸的增加，PVT 的 GFLOPs 增长率大于 ResNet [2]，但低于 ViT [3]，如图 5 所示。然而，当输入的尺寸不超过 640×640 像素时，PVT-Small 和 ResNet50 的 GFLOPs 是相近的。这意味着本文的 PVT 更适合输入为中等分辨率的任务。

在 COCO 数据集上，输入图像的较短一边为 800 像素。在这种情况下，基于 PVT-Small 的 RetinaNet 的推理速度比基于 ResNet50 的模型要慢。(1) 解决这个问题的直接方法是降低输入的尺寸。当将输入图像的短边缩小到 640 像素时，基于 PVT-Small 的模型比基于 ResNet50 的模型运行得更快 (51.7ms *vs.* 55.9ms)，且 AP 高出 2.4(38.7 *vs.* 36.3)。(2) 另一种解决方案是开发一个计算复杂度较低的自注意力层。这是一个值得探索的方向，本文最近提出了一个解决方案 PVTv2 [69]。

在补充材料中，本文对 PVT 的其他特征进行更多的定性或定量分析，并提供了稠密预测任务的可视化。

[3]为了将 ViT 应用于 RetinaNet，本文从 ViT-Small/32 的第 2、4、6 和 8 层提取特征，并将它们插值到不同的尺度。

6. 结论与未来工作

本文提出的 PVT 是一个纯 Transformer 主干网络，它被用于稠密预测任务，如目标检测和语义分割。本文设计渐进式收缩金字塔和空间降维注意力层，是为了在有限的计算/存储资源下，获得高分辨率图像和多尺度的特征图。目标检测和语义分割基准上的大量实验证实，在参数数量相近的情况下，本文的 PVT 比精心设计的 CNN 主干网络更强。

虽然 PVT 可以替代 CNN 主干网络 (如 ResNet、ResNeXt)，但仍有一些专门为 CNN 设计的模块和操作在本文中没有被考虑到，如 SE [70]、SK [33]、Dilated Convolution [68]、Model Pruning [71] 和 NAS [72]。此外，经过多年的快速发展，已经出现了许多设计精良的 CNN 主干网络，如 Res2Net [73]、EfficientNet [72]、ResNeSt [74]。相比之下，基于 Transformer 的计算机视觉模型仍处于发展的早期阶段。因此，我们认为未来会有许多潜在的技术和应用将会被探索 (如 OCR [75, 76, 77]，3D [78, 79, 80]，医学图像分析 [81, 82, 83])，希望 PVT 可以作为一个好的开端。

References

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *Proc. Int. Conf. Learn. Representations*, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *Proc. Int. Conf. Learn. Representations*, 2021.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. Eur. Conf. Comp. Vis.*, 2020.

这项工作得到了国家自然科学基金 61672273 和 61832008, 江苏省杰出青年科学基金 BK20160021, 中国博士后创新人才支持计划 BX20200168,2020M681608, 以及香港综合研究基金 No.27208720 的支持。

[5] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *international conference on computer vision*, 2021.

[6] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. Advances in Neural Inf. Process. Syst.*, 2015.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. Eur. Conf. Comp. Vis.*, 2018.

[11] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2009.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comp. Vis.*, 2014.

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, 2010.

[15] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Advances in Neural Inf. Process. Syst.*, 2017.

[17] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *Proc. Int. Conf. Learn. Representations*, 2021.

[18] Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer. In *Proc. Int. Joint Conf. Artificial Intell.*, 2021.

[19] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

[20] Ronghang Hu and Amanpreet Singh. Transformer is all you need: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2211.

[21] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2021.

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[23] Niki Parmar, Prajit Ramachandran, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Proc. Advances in Neural Inf. Process. Syst.*, 2019.

[24] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020.

[25] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. 1998.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 2015.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Proc. Advances in Neural Inf. Process. Syst.*, 2012.

[28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015.

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[30] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proc. AAAI Conf. Artificial Intell.*, 2017.

[31] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. *Proc. Advances in Neural Inf. Process. Syst.*, 2017.

[32] Wenhai Wang, Xiang Li, Jian Yang, and Tong Lu. Mixed link networks. *Proc. Int. Joint Conf. Artificial Intell.*, 2018.

[33] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019.

[34] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[35] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516, 2020.

[36] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[37] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and

Alexander C Berg. Ssd: Single shot multibox detector. In *Proc. Eur. Conf. Comp. Vis.*, 2016.

[38] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019.

[39] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Proc. Advances in Neural Inf. Process. Syst.*, 2020.

[40] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021.

[41] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020.

[42] Peize Sun, Yi Jiang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Onenet: Towards end-to-end one-stage object detection. *arXiv preprint arXiv:2012.05780*, 2020.

[43] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[44] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021.

[45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015.

[46] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015.

[47] Suyash Shetty. Application of convolutional neural network for image classification on pascal voc challenge 2012 dataset. *arXiv preprint arXiv:1607.03785*, 2016.

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[49] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[50] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[51] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[52] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019.

[53] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Proc. Advances in Neural Inf. Process. Syst.*, 2016.

[54] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019.

[55] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019.

[56] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention. In *Proc. Int. Conf. Learn. Representations*, 2021.

[57] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proc. Int. Conf. Mach. Learn.*, 2021.

[58] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[59] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.

[60] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021.

[61] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. Int. Conf. Learn. Representations*, 2018.

[62] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 6023–6032, 2019.

[63] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proc. AAAI Conf. Artificial Intell.*, 2020.

[64] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. Int. Conf. Learn. Representations*, 2019.

[65] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *Proc. Int. Conf. Learn. Representations*, 2017.

[66] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. Int. Conf. Artificial Intell. & Stat.*, 2010.

[67] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[68] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In Yoshua Bengio and Yann LeCun, editors, *Proc. Int. Conf. Learn. Representations*, 2016.

[69] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021.

[70] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[71] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[72] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. Int. Conf. Mach. Learn.*, 2019.

[73] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[74] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[75] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019.

[76] Wenhai Wang, Xuebo Liu, Xiaozhong Ji, Enze Xie, Ding Liang, ZhiBo Yang, Tong Lu, Chunhua Shen, and Ping Luo. Ae textspotter: Learning visual and linguistic representation for ambiguous text spotting. In *Proc. Eur. Conf. Comp. Vis.*, 2020.

[77] Wenhai Wang, Enze Xie, Xiang Li, Xuebo Liu, Ding Liang, Yang Zhibo, Tong Lu, and Chunhua Shen. Pan++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[78] Le Hui, Rui Xu, Jin Xie, Jianjun Qian, and Jian Yang. Progressive point cloud deconvolution generation network. In *Proc. Eur. Conf. Comp. Vis.*, 2020.

[79] Mingmei Cheng, Le Hui, Jin Xie, and Jian Yang. SSPC-Net: Semi-supervised semantic 3D point cloud segmentation network. In *Proc. AAAI Conf. Artificial Intell.*, 2021.

[80] Le Hui, Mingmei Cheng, Jin Xie, and Jian Yang. Efficient 3D point cloud feature learning for large-scale place recognition. *arXiv preprint arXiv:2101.02374*, 2021.

[81] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[82] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020.

[83] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.