

基于重组性高斯自注意力的视觉 Transformer

赵亮^{1,2} 周继开¹

摘要 在目前视觉 Transformer 的局部自注意力中, 现有的策略无法建立所有窗口之间的信息流动, 导致上下文语境建模能力不足. 针对这个问题, 基于混合高斯权重重组 (Gaussian weight recombination, GWR) 的策略, 提出一种新的局部自注意力机制 SGW-MSA (Shuffled and Gaussian window-multi-head self-attention), 它融合了 3 种不同的局部自注意力, 并通过 GWR 策略对特征图进行重建, 在重建的特征图上提取图像特征, 建立了所有窗口的交互以捕获更加丰富的上下文信息. 基于 SGW-MSA 设计了 SGWin Transformer 整体架构. 实验结果表明, 该算法在 mini-imagenet 图像分类数据集上的准确率比 Swin Transformer 提升了 5.1%, 在 CIFAR10 图像分类实验中的准确率比 Swin Transformer 提升了 5.2%, 在 MS COCO 数据集上分别使用 Mask R-CNN 和 Cascade R-CNN 目标检测框架的 mAP 比 Swin Transformer 分别提升了 5.5% 和 5.1%, 相比于其他基于局部自注意力的模型在参数量相似的情况下具有较强的竞争力.

关键词 Transformer, 局部自注意力, 混合高斯权重重组, 图像分类, 目标检测

引用格式 赵亮, 周继开. 基于重组性高斯自注意力的视觉 Transformer. 自动化学报, 2023, 49(9): 1976–1988

DOI 10.16383/j.aas.c220715

Vision Transformer Based on Reconfigurable Gaussian Self-attention

ZHAO Liang^{1,2} ZHOU Ji-Kai¹

Abstract In the current vision Transformer's local self-attention, the existing strategy cannot establish the information flow between all windows, resulting in the lack of context modeling ability. To solve this problem, this paper proposes a new local self-attention mechanism shuffled and Gaussian window-multi-head self-attention (SGW-MSA) based on the strategy of Gaussian weight recombination (GWR), which combines three different local self-attention forces, and reconstructs the feature map through GWR strategy, and extracts image features from the reconstructed feature map. The interaction of all windows is established to capture richer context information. This paper designs the overall architecture of SGWin Transformer based on SGW-MSA. The experimental results show that the accuracy of this algorithm in the mini-imagenet image classification dataset is 5.1% higher than that in the Swin Transformer, the accuracy in the CIFAR10 image classification experiment is 5.2% higher than that in the Swin Transformer, and the mAP using the Mask R-CNN and Cascade R-CNN object detection frameworks on the MS COCO dataset are 5.5% and 5.1% higher than that in the Swin Transformer, respectively. Compared with other models based on local self-attention, it has stronger competitiveness in the case of similar parameters.

Key words Transformer, local self-attention, Gaussian weight recombination (GWR), image classification, objection detection

Citation Zhao Liang, Zhou Ji-Kai. Vision Transformer based on reconfigurable Gaussian self-attention. *Acta Automatica Sinica*, 2023, 49(9): 1976–1988

收稿日期 2022-09-10 录用日期 2023-01-13

Manuscript received September 10, 2022; accepted January 13, 2023

国家自然科学基金 (51209167, 12002251), 陕西省自然科学基金 (2019JM-474), 陕西省岩土与地下空间工程重点实验室开放基金 (YT202004), 陕西省教育厅服务地方专项计划 (22JC043) 资助

Supported by National Natural Science Foundation of China (51209167, 12002251), Natural Science Foundation of Shaanxi Province (2019JM-474), Open Fund Project of Key Laboratory of Geotechnical and Underground Space Engineering in Shaanxi Province (YT202004), and Shaanxi Provincial Department of Education Service Local Special Plan Project (22JC043)

本文责任编辑 黄华

Recommended by Associate Editor HUANG Hua

1. 西安建筑科技大学信息与控制工程学院 西安 710055 2. 陕西省岩土与地下空间工程重点实验室 西安 710055

1. College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055 2. Shaanxi Provincial Key Laboratory of Geotechnical and Underground Space Engineering, Xi'an 710055

目前计算机视觉领域使用的方法有两大类, 分别是卷积神经网络 (Convolutional neural networks, CNN) 和 Transformer. 其中 CNN 是图像分类^[1]、目标检测^[2]和语义分割^[3]等计算机视觉任务的主流方法, 自 AlexNet^[4] 诞生并在 ImageNet 图像分类挑战中获得冠军以后, 研究者们开始通过各种方法设计卷积神经网络, 使得网络变得更深、更密集、更复杂^[5-8], 在随后的几年内出现了很多经典的卷积神经网络. VGGNet^[5] 探索了 CNN 的深度及性能之间的关系, 通过使用很小的卷积叠加增加网络的深度达到提升网络精度的效果; DenseNet^[6] 通过从特征图的角度入手, 为每一个模块加入密集连接达到了更好的性能和更少的参数量; ResNet^[7] 通过引

入残差结构解决了随着网络层数的加深出现梯度消失的问题; GoogLeNet^[9] 使用密集成分来近似最优的稀疏结构, 在提升性能的同时不增加计算量; EfficientNet^[10] 提出了一种多维度混合的模型缩放方法, 可以同时兼顾模型的精度以及速度. 在 CNN 模型性能越来越强的同时, 另一类视觉 Transformer 的方法横空出世. Transformer 由于其自注意力模块具有捕捉长距离依赖^[11] 的能力广泛被应用于自然语言处理的任务中, 而后被用到了计算机视觉任务中并取得了比 CNN 方法更优的效果. 在文献 [12–15] 中将自注意力模块嵌入到 CNN 中并应用于图像分类、目标检测和语义分割等计算机视觉任务中. Vision Transformer (ViT)^[16] 不使用卷积神经网络而是通过将图像序列化的方法首次将 Transformer 架构应用到图像领域中, 并且在 ImageNet 数据集上取得了比 ResNet 更好的效果, 而后在短时间内被引入改进^[17–20] 并应用于各种图像领域的各种下游任务^[21–24]. 但是 Transformer 的复杂度成为了其性能最大的瓶颈, 为了减小因全局自注意力引起的二次复杂度, 现有的方法较多使用局部自注意力机制. 目前现有的局部自注意力机制主要有 7 类(如图 1 所示).

1) 目前几乎所有的基于局部自注意力的 Transformer 模型都会使用常规窗口自注意力 (Window-multi-head self-attention, W-MSA), 通过 W-MSA 与其他类型的局部自注意力交替使用来建立窗口之间的通信, 如图 1(a) 所示.

2) HaloNet^[25] 通过对窗口进行缩放的操作来收集窗口之外的信息并建立跨窗口的信息交互, 如图 1(b) 所示.

3) Swin Transformer 通过在连续的局部注意力层之间移动窗口的分区建立跨窗口之间的信息通信缓解感受野受限的问题, 如图 1(c) 所示.

4) CrossFormer^[26] 提出了跨尺度嵌入层和长短注意力, 有效地建立了长距离的跨窗口的连接.

5) Shuffle Transformer^[27] 在连续的局部自注意力层之间加入空间 shuffle 的操作, 以提供长距离窗口之间的连接并增强建模能力.

6) GG Transformer^[28] 受到了人类在自然场景中识别物体的 Glance 和 Gaze 行为的启发, 能够有效地对远程依赖性和局部上下文进行建模, 4) ~ 6) 这 3 种局部注意力可统一归为图 1(d) 的形式.

7) Axial-DeepLab^[29] 将二维自注意力分解为横向和纵向两个一维的自注意力, 如图 1(e) 所示.

8) CSWin Transformer^[30] 提出了一种在“十”字等宽窗口内计算自注意力的方式 (Cross-shaped window self-attention), 通过横条和纵条窗口自注意力并行实现, 如图 1(f) 所示.

9) Pale Transformer^[31] 提出了“十”字等间隔窗口自注意力 (Pale-shaped-attention, PS-Attention), 如图 1(g) 所示.

图 1 展示了现有的局部自注意力方法. 不同的颜色表示不同的窗口, 在每个窗口内执行计算自注意力, 并通过引入各种策略来建立跨窗口之间的连接. 这些工作虽然取得了优异的性能, 甚至优于一些最新的 CNN 的方法, 但是每个自注意力层中的依赖性仍然具有局限性, 具体表现在当特征图很大时, 通过有间隔的采样点组成的窗口无法建立所有窗口之间的信息流动导致了模型捕获的上下文语义

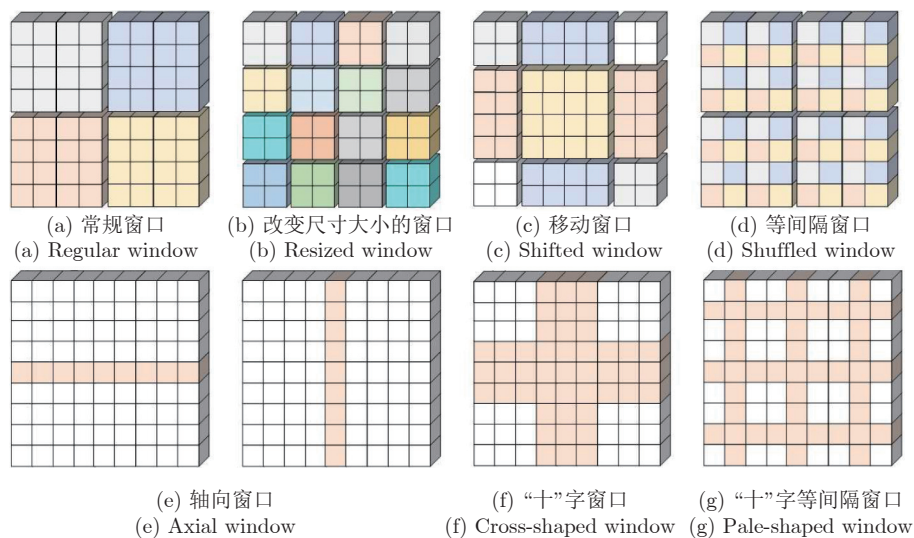


图 1 现有局部自注意力方法

Fig.1 Existing local self-attention methods

信息的能力不足. 针对上述问题, 本文提出了一种高斯窗口自注意力机制 (Gaussian window-multi-head self-attention, GW-MSA), 它包括纵向高斯窗口自注意力 (Vertical Gaussian window-multi-head self-attention, VGW-MSA) 和横向高斯窗口自注意力 (Horizontal Gaussian window-MSA, HGW-MSA) 两种类型的局部自注意力. GW-MSA 与图 1(d) 中的 Shuffled W-MSA 联合组成了 SGW-MSA, 有效地捕捉更丰富的上下文依赖, 如图 2 所示, 不同颜色的点代表不同的窗口组成, 在 GW-MSA 中, 通过混合高斯权重重组 GWR 策略重构特征图, 并在重构后的特征图上计算局部自注意力. 本文在 Swin Transformer 结构的基础上, 引入 SGW-MSA 设计了 SGWin Transformer 模型, 在公开数据集 CIFAR10、mini-imagenet、KITTI、PASCAL VOC 和 MS COCO 上进行了实验, 实验结果表明 SGWin Transformer 在图像分类和目标检测的任务上优于其他同等参数量的基于局部自注意力的 Transformer 网络.

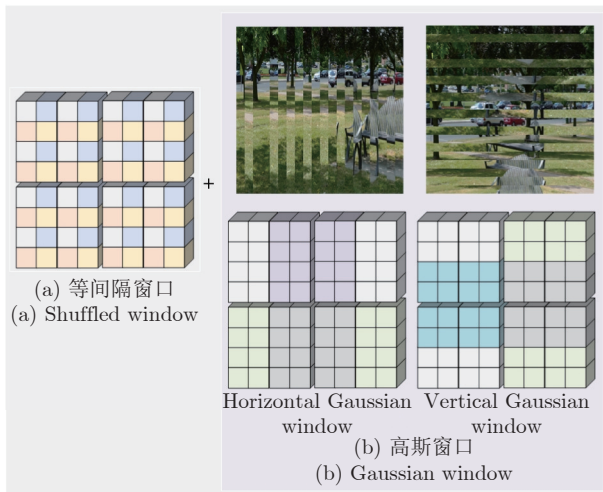


图 2 局部自注意力组合
Fig.2 Local self-attention combination

1 模型框架

1.1 Swin Transformer 算法

Swin Transformer 提出了一种新的基于 Transformer 的视觉主干网络, 自注意力的计算在局部非重叠窗口内进行. 一方面可以将复杂度从之前的和图像大小成平方的关系变成线性关系, 并且采用非重叠局部窗口, 大大减小了计算量; 另一方面在不同的注意力层之间采用移动窗口的操作, 使得不同窗口之间的信息可以交换. 并且由于性能超越了参

数量相似的 CNN 主干, 推动了 Transformer 成为了视觉主干网络的新主流, 在近两年出现了越来越多基于局部自注意力机制的视觉 Transformer 方法, 然而目前的各种局部自注意力建立远距离跨窗口连接策略具有一定的局限性. 当特征图很大时, 现有的窗口连接策略无法建立所有窗口之间的信息流动导致无法捕捉足够的上下文信息. 假设特征图的高和宽分别为 h 和 w , 局部窗口的高和宽分别为 W_h 和 W_w , 对于特征图上划分的某一个局部窗口, 该窗口在纵向和横向可以建立最近窗口连接的距离分别为:

$$d_{\min}^h = \max\left(0, \frac{h}{W_h} - W_h + 1\right) \quad (1)$$

$$d_{\min}^w = \max\left(0, \frac{w}{W_w} - W_w + 1\right) \quad (2)$$

在纵向和横向可以建立最远窗口连接的距离分别为:

$$d_{\max}^h = \frac{h}{W_h}(W_h - 1) + W_h \quad (3)$$

$$d_{\max}^w = \frac{w}{W_w}(W_w - 1) + W_w \quad (4)$$

所以具有 4 种不能建立窗口连接的情况: 1) $d_{\min}^h > W_h$; 2) $d_{\min}^w > W_w$; 3) $d_{\max}^h < h - W_h$; 4) $d_{\max}^w < w - W_w$. 当 h, w, W_h, W_w 之间的关系满足式 (5) ~ 式 (7) 中的一种情况时就会出现特征图过大导致无法建立所有窗口之间信息交互的情况. 当满足式 (5) 或式 (6) 时, 窗口之间的纵向距离或者横向距离分别大于 d_{\max}^h 和 d_{\max}^w 时无法建立连接, 当满足式 (7) 中的情况时, 窗口之间的纵(横)向距离小于 d_{\min}^h (d_{\min}^w) 或者大于 d_{\max}^h (d_{\max}^w) 都无法建立连接.

$$2W_h^2 - W_h < h \leq 2W_h^2 - 1 \quad (5)$$

$$2W_w^2 - W_w < w \leq 2W_w^2 - 1 \quad (6)$$

$$h > 2W_h^2 - 1 \text{ 或 } w > 2W_w^2 - 1 \quad (7)$$

1.2 SGWin Transformer 的整体结构

为了解决当特征图过大时现有的局部自注意力机制无法建立所有窗口之间的信息交互的问题, 本文提出了一种新的局部自注意力机制 SGW-MSA, 并在 Swin Transformer 的基础上将所有的移动窗口自注意力 SW-MSA 替换为 SGW-MSA 得到一种新的 SGWin Transformer 模型, 模型的整体架构如图 3(a) 所示. 主干网络符合标准的视觉分层 Transformer 的 PVT^[32] 的结构, 该设计包含了 4 个阶段的金字塔结构, 每个阶段由 Patch embed 或 Patch merging 和多个 SGWin Transformer block 串联组

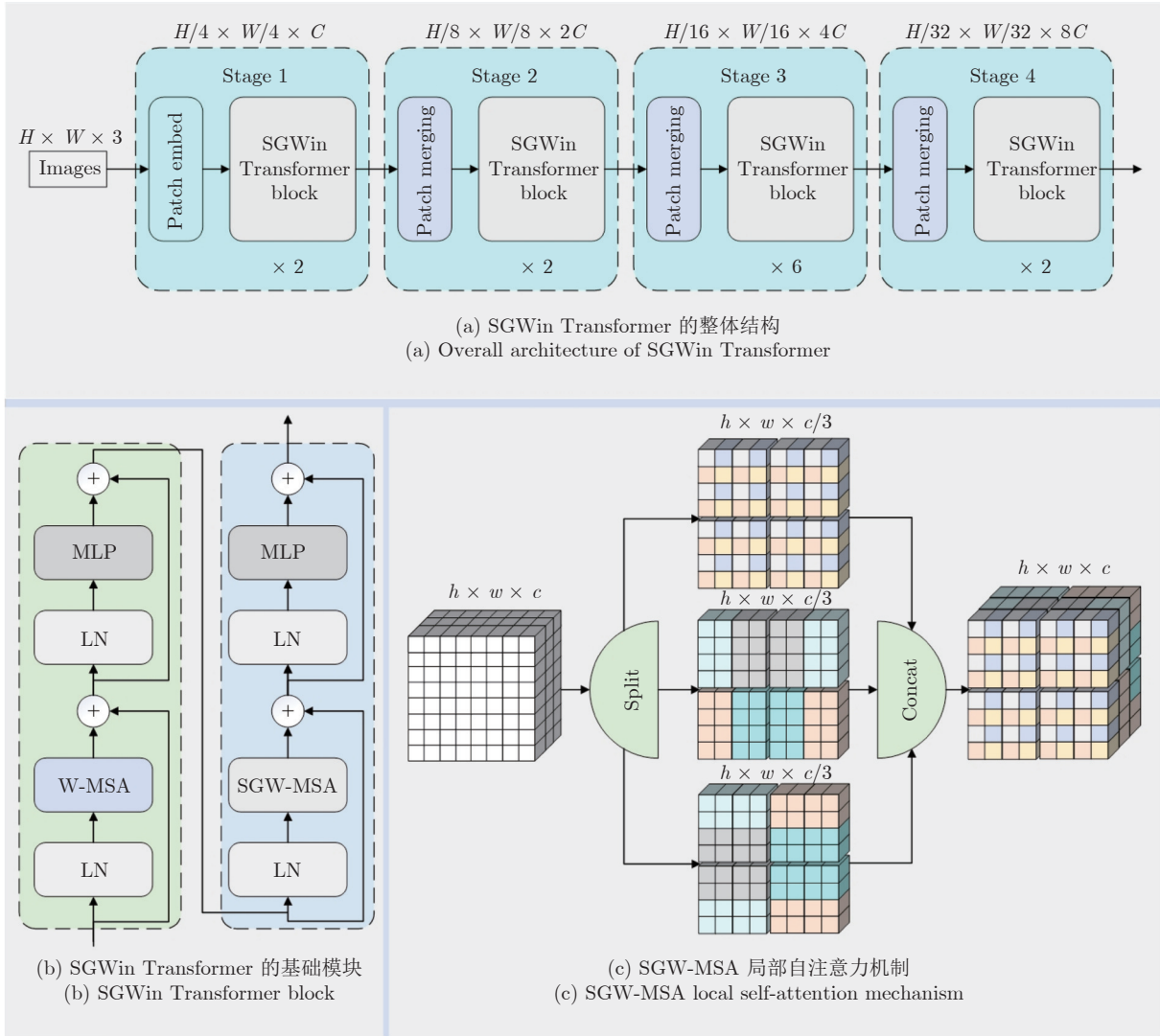


图 3 SGWin Transformer 整体架构

Fig.3 Overall architecture of SGWin Transformer

合而成. 如图 3(b) 所示, 每个 SGWin Transformer block 由两组结构串联组成, 第一组结构包括一个 W-MSA 模块和一个 MLP, 第二组结构由一个 SGW-MSA 模块和一个 MLP 模块组成, MLP 对输入特征图进行非线性化的映射得到新的特征图, SGW-MSA 局部自注意力机制的示意图如图 3(c) 所示. 整个模型的计算过程: 输入图片通过 Patch embed 将输入图像下采样 4 倍, 并得到指定通道数的特征图, 特征图会被送入 Stage 1 的 SGWin Transformer block 中, 通过 W-MSA、SGW-MSA 模块提取局部特征和图像中的上下文信息并建立所有窗口之间的信息流通, Stage 1 最后一个 SGWin Transformer block 的输出会被送入 Stage 2 中, 除 Stage 1 之外的所有 Stage 会通过一个 Patch merging 将上一个阶段输出的特征图尺寸降采样两倍 (宽和高

变为原来的二分之一), 通道维度变为原来的两倍. 整个网络之后可以接一个 Softmax 层和一个全连接层用于图像分类任务, 并且每个阶段的特征图可输入到目标检测的 FPN^[33] 部分中进行多尺度目标检测.

1.3 SGW-MSA 局部自注意力机制

当出现式 (5) 或式 (6) 中的情况时, 两个窗口之间的纵 (横) 向距离大于一定值时就无法建立连接. 当出现式 (7) 中的情况时, 两个窗口之间的纵 (横) 向距离大于或小于一定值时都无法建立连接. 因此式 (7) 中的问题包含式 (5) 和式 (6) 存在的问题. 仅考虑式 (7) 中的情况, 将纵向无法建立窗口连接的两个距离分别记为 d_{\min}^h 和 d_{\max}^h , 将横向无法建立窗口连接的两个距离分别记为 d_{\min}^w 和 d_{\max}^w . 如图 4

所示, 为了能够建立所有窗口之间的信息交互, SGW-MSA 将输入特征图在通道上均匀拆分成 3 组, 对第一组特征图使用现有的 Shuffled W-MSA 等间隔采样点组成窗口用于纵(横)向距离大于 d_{\min}^h (d_{\min}^w) 且小于 d_{\max}^h (d_{\max}^w) 窗口之间的联系; 后两份特征图分别使用横向高斯窗口自注意力 HGW-MSA 和纵向高斯窗口自注意力 VGW-MSA 计算局部自注意力, 建立 Shuffled W-MSA 未能建立的窗口的联系. 最后将 3 个部分的局部自注意力计算结果在通道上进行合并得到最终的输出结果.

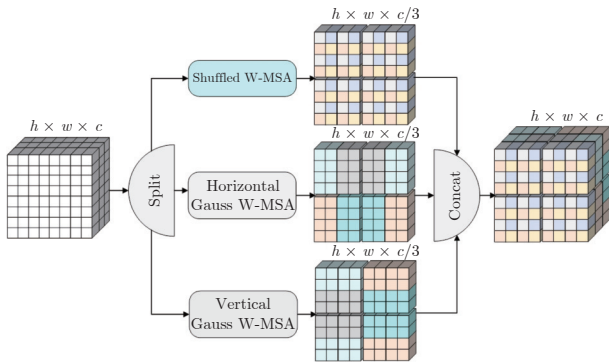


图 4 SGW-MSA 局部自注意力示意图
Fig.4 SGW-MSA local self-attention diagram

1.3.1 GW-MSA 局部自注意力机制

GW-MSA 可用于建立 Shuffled W-MSA 未能建立的窗口连接, 分为 VGW-MSA 和 HGW-MSA 两种不同的形式. 如图 5 所示, 每个形式的 GW-MSA 由混合高斯权重重组 GWR 模块、常规局部自注意力 W-MSA 和逆混合高斯权重重组 (re Gaussian weight recombination, reGWR) 模块 3 部分组成, 其中 GWR 是本文为了建立纵(横)向距离小于 d_{\min}^h (d_{\min}^w) 或者大于 d_{\max}^h (d_{\max}^w) 窗口之间的信息交互提出的一种特征图重组的策略.

假设特征图的高和宽分别为 h 和 w , 局部窗口的高和宽分别为 W_h 和 W_w . GWR 会将输入特征图划分成多个长条形状的基础元素块 (Basic element block, BEB), 计算纵向的 VGW-MSA 时将特征图按高切分成若干份高宽分别为 W_b ($W_b < W_h$) 和 w 的横条基础元素块, 如图 6(a) 所示. 计算横向的 HGW-MSA 时将特征图按宽切分成高宽分别为 h 和 W_b ($W_b < W_w$) 的竖条基础元素块, 如图 6(b) 所示. 当 h 或 w 不能整除 w_b 时, 取最大可以整除 W_b 的长度作为重组区域.

为所有的基础元素块建立高斯权重分布表, 结合高斯权重分布表尽可能使距离小于 d_{\min}^h (d_{\min}^w) 或者大于 d_{\max}^h (d_{\max}^w) 的基础元素块放在一起用于重组

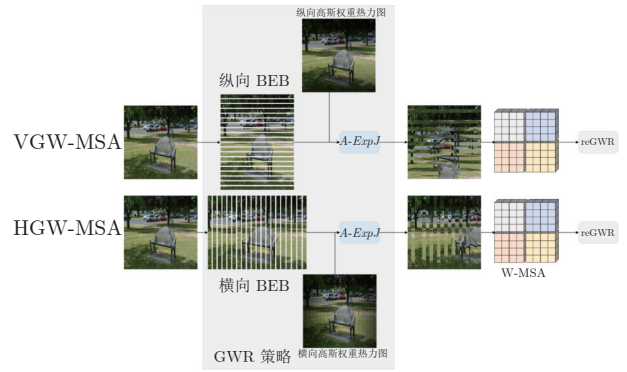
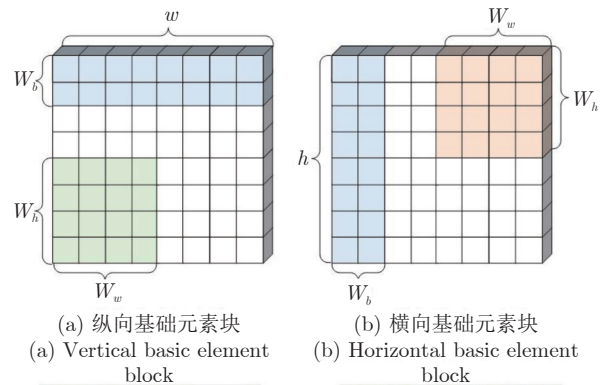


图 5 GW-MSA 局部自注意力示意图
Fig.5 GW-MSA local self-attention diagram



(a) 纵向基础元素块
(a) Vertical basic element block
(b) 横向基础元素块
(b) Horizontal basic element block
(c) 纵向基础元素块高斯权重
(c) Vertical basic element block Gaussian weights
(d) 横向基础元素块高斯权重
(d) Horizontal basic element block Gaussian weights

图 6 纵横向基础元素块示意图
Fig.6 Schematic diagram of vertical and horizontal basic element block

特征图. 然后在重组后的特征图上使用 W-MSA 计算局部自注意力. 高斯权重分布表由一维高斯分布公式得到:

$$f(x) = A \cdot \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (8)$$

式中 A 代表权重的幅值, μ 表示均值, σ^2 为方差. GWR 策略的思想就是根据高斯分布的特性. 如图 6(c) 和图 6(d) 所示, 纵向基础元素块越靠近图像上边缘或下边缘, 权重越小; 横向基础元素块越靠近左边缘和右边缘, 权重越小. 权重越高的基础元素块对应图像中的位置颜色越亮, 代表对应位置的权重越

高; 权重越低的基础元素块对应图像中的位置颜色越暗, 代表对应位置的权重越低. 将特征图上的每一个基础元素块看作一个点, 以特征图中心的基础元素块为原点建立坐标系, 依据每个基础元素块在坐标系中的位置可以被赋予一个对应的高斯分布权重, 纵向和横向的中心基础元素块的位置坐标记为 $cx = (h/2W_b$ 或 $w/2W_b)$, 对于任意 x 位置下的基础元素块对应的高斯权重分布遵循下式:

$$\text{Weight}(x) = \exp\left(-\left(\frac{x-cx}{cx}\right)^2 \frac{1}{2\sigma^2}\right) \quad (9)$$

式中的 σ 取值为 1.5, 分子部分除以 cx 是为了控制权重不会过小而约等于 0. 为了尽可能将权重近似的基础元素块放在一起, 本文采用了带权重的随机抽样 $A\text{-ExpJ}$ ^[34] 依据每一个基础元素块的索引以及对应的权重进行随机抽样, 最后将所有基础元素块的索引按照抽样的顺序进行排列得到新的重组后的特征图. 假设将特征图划分成基础元素块的序列索引为 $idx = [1, 2, \dots, n]$, 其中 $n = h/W_b$ 或 w/W_b ; 基础元素块的高斯权重分布表为 $W = [W_1, W_2, \dots, W_n]$, 其中 $n = h/W_b$ 或 w/W_b ; 重组的索引 idx_{new} 可以由式 (10) 得到, 其中 n 表示通过权重抽样的个数. $A\text{-ExpJ}$ 表示带权重的随机抽样函数. 最后按照新的基础元素块的索引对特征图进行重组得到 GWR 策略的输出结果.

$$idx_{\text{new}} = A\text{-ExpJ}(idx, W, n) \quad (10)$$

1.3.2 SGW-MSA 的计算过程

假设输入特征图为 $X \in \mathbf{R}^{h \times w \times c}$, SGW-MSA 首先将输入特征图 X 在通道上切分成 3 个部分, 第一个部分的特征图记为 $X_S \in \mathbf{R}^{h \times w \times \frac{c}{3}}$, 第二个部分的特征图记为 $X_V \in \mathbf{R}^{h \times w \times \frac{c}{3}}$, 第三个部分的特征图记为 $X_H \in \mathbf{R}^{h \times w \times \frac{c}{3}}$. 对 X_S 使用 Shuffled W-MSA 在特征图上使用等间隔采样点组成窗口, 并在所有的窗口内部计算自注意力. 对 X_V 和 X_H 分别使用纵向和横向的 GWR 策略对特征图进行重组, 并在重组的特征图上使用 W-MSA 计算局部自注意力. 具体计算过程如下.

首先在 X_S 上通过等间隔采样特征点形成多个具有相同尺寸 (W_h, W_w) 的窗口:

$$X_S = [X_S^1, X_S^2, \dots, X_S^N] \quad (11)$$

其中 $X_S^i \in \mathbf{R}^{h \times w \times \frac{c}{3}}$, $i \in [1, 2, \dots, N]$, 窗口的总数 $N = h \cdot w / (W_h \cdot W_w)$. 然后使用 GWR 策略对 X_V 和 X_H 进行重组, 将重组后的 X_V 和 X_H 拆分成多个具有相同尺寸 (W_h, W_w) 的窗口:

$$X_V = [X_V^1, X_V^2, \dots, X_V^N] \quad (12)$$

$$X_H = [X_H^1, X_H^2, \dots, X_H^N] \quad (13)$$

其中 $X_V^i \in \mathbf{R}^{W_h \times W_w \times \frac{c}{3}}$, $X_H^i \in \mathbf{R}^{W_h \times W_w \times \frac{c}{3}}$, $i \in [1, 2, \dots, N]$, 窗口的总数 $N = h \cdot w / (W_h \cdot W_w)$. 当 $h \cdot w$ 不能被 $W_h \cdot W_w$ 整除时, 可以对特征图进行填充或者插值的方法确保 $h \cdot w$ 可以被 $W_h \cdot W_w$ 整除.

每一个窗口内部单独计算局部自注意力. 在计算局部自注意力时, 使用 3 个全连接层 ℓ_Q, ℓ_K, ℓ_V 计算得到 Q (Query), K (Key), V (Value), 计算式如下:

$$Y_S^i = \text{MSA}(\ell_Q(X_S^i), \ell_K(X_S^i), \ell_V(X_S^i)) \quad (14)$$

$$Y_V^i = \text{MSA}(\ell_Q(X_V^i), \ell_K(X_V^i), \ell_V(X_V^i)) \quad (15)$$

$$Y_H^i = \text{MSA}(\ell_Q(X_H^i), \ell_K(X_H^i), \ell_V(X_H^i)) \quad (16)$$

其中 $i \in [1, 2, \dots, N]$, MSA 表示 Multi-head self-attention^[33]. 最后将所有的局部自注意力的计算结果在空间上进行合并得到新的特征图:

$$Y_S = [Y_S^1, Y_S^2, \dots, Y_S^N] \quad (17)$$

$$Y_V = [Y_V^1, Y_V^2, \dots, Y_V^N] \quad (18)$$

$$Y_H = [Y_H^1, Y_H^2, \dots, Y_H^N] \quad (19)$$

因为 GWR 策略将原有的特征图根据新的基础元素块的顺序进行了重组, 所以需要将 Y_V 和 Y_H 依据原先的基础元素块的顺序进行还原. 将两个部分的局部自注意力计算结果在通道上进行合并, 得到最终的输出结果, 如式 (20) 所示, 其中 Concat 表示在通道上进行合并.

$$Y = \text{Concat}(Y_S, Y_V, Y_H) \quad (20)$$

1.3.3 计算复杂度分析

对于给定的尺寸为 $\mathbf{R}^{h \times w \times c}$ 的特征图, 局部窗口的尺寸为 ($W_h \times W_w$), 用 \mathcal{O} 表示复杂度. 标准的全局自注意力 (Global self-attention) 的计算复杂度如式 (21) 所示:

$$\mathcal{O}_{\text{Global}} = 4hwc^2 + 2c(hw)^2 \quad (21)$$

SGW-MSA 的计算复杂度如式 (22) 所示

$$\mathcal{O}_{\text{SGW}} = 4hwc^2 + 2W_w W_h hwc \quad (22)$$

其中 W_w, W_h 分别为局部窗口的宽和高. 对比式 (9) 和式 (10), 因为 $S_w S_h \ll hw$, 所以 $\mathcal{O}_{\text{SGW}} \ll \mathcal{O}_{\text{Global}}$, 即 SGW-MSA 的计算复杂度远小于全局自注意力的计算复杂度.

1.4 SGWin Transformer block

SGWin Transformer block 由两组结构串联组成. 如图 7 所示, 第一组结构包括一个 W-MSA 模

块和一个多层感知机模块 MLP, 第二组结构由一个 SGW-MSA 模块和一个 MLP 模块组成, MLP 对输入特征图进行非线性化的映射得到新的特征图, W-MSA 用于捕捉特征图的局部自注意力, SGW-MSA 用于捕捉局部自注意力并建立所有窗口之间的信息流通. 整个 SGWin Transformer block 的向前传播式如下:

$$\hat{x}^l = \text{W-MSA}(\text{LN}(x^{l-1})) + x^{l-1} \quad (23)$$

$$x^l = \text{MLP}(\text{LN}(\hat{x}^l) + \hat{x}^l) \quad (24)$$

$$\hat{x}^{l+1} = \text{SGW-MSA}(\text{LN}(x^l)) + x^l \quad (25)$$

$$x^{l+1} = \text{MLP}(\text{LN}(\hat{x}^{l+1})) + \hat{x}^{l+1} \quad (26)$$

其中 x^{l-1} 为前一个 Patch embed 或者 Patch merging 或者 SGWin Transformer block 的输出, \hat{x}^l 和 x^l 分别代表 (SG)W-MSA 模块和 MLP 模块的输出, LN 代表 LayerNorm.

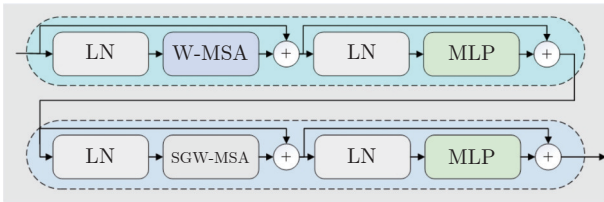


图 7 SGWin Transformer block 结构示意图

Fig.7 Structure diagram of SGWin Transformer block

1.5 SGWin Transformer 的超参数配置

SGWin Transformer 的超参数配置与 Swin Transformer 相同, 如表 1 所示. 其中 Stage = i 表示 SGWin Transformer 模型的第 i 个阶段. Stride 表示 SGWin Transformer 模型在每个阶段下采样的倍数. Layer 表示当前阶段的模块名字, 一个阶段包含两个模块, Patch embed 和 Patch merging 负责对特征图进行下采样, 下采样的倍数分别为 4 和 2, Patch embed 和 Patch merging 的输出会被送入后续的 Transformer block 中提取局部自注意力并进行特征的映射, 在最后一个 Transformer block 后接一个平均池化层和全连接层可用于图像分类任务, 或者将每一层的特征图输出可用于目标检测任务. 模型的第 i 个 Stage 的模型的超参数定义如下:

1) P_i . 第 i 个 Stage 的输入特征图下采样的倍数, 第一个 Stage 下采样的倍数是 4, 其余 3 个 Stage 的下采样倍数为 2;

2) C_i . 第 i 个 Stage 的输入特征图下采样后新特征图的通道数;

表 1 SGWin Transformer 的超参数配置表
Table 1 Super parameter configuration table of SGWin Transformer

Stage	Stride	Layer	Parameter
1	4	Patch embed	$P_1 = 4$ $C_1 = 96$
		Transformer block	$\begin{bmatrix} S_1 = 7 \\ H_1 = 3 \\ R_1 = 4 \end{bmatrix} \times 2$
2	8	Patch merging	$P_2 = 2$ $C_2 = 192$
		Transformer block	$\begin{bmatrix} S_2 = 7 \\ H_2 = 6 \\ R_2 = 4 \end{bmatrix} \times 2$
3	16	Patch merging	$P_3 = 2$ $C_3 = 384$
		Transformer block	$\begin{bmatrix} S_3 = 7 \\ H_3 = 12 \\ R_3 = 4 \end{bmatrix} \times 2$
4	32	Patch merging	$p_4 = 2$ $C_4 = 768$
		Transformer block	$\begin{bmatrix} S_4 = 7 \\ H_4 = 24 \\ R_4 = 4 \end{bmatrix} \times 2$

3) S_i . 第 i 个 Stage 的 Transformer block 中计算局部自注意力的窗口大小;

4) H_i . 第 i 个 Stage 的 Transformer block 中多头自注意力机制的 Head 数量;

5) R_i . 第 i 个 Stage 的 Transformer block 中 MLP 模块的通道扩展比.

2 实验结果

本文分别在图像分类数据集 CIFAR10^[35] 以及目标检测数据集 KITTI^[36]、PASCAL VOC^[37]、MS COCO^[38] 上进行了实验, 与其他参数量相似且具有代表性的基于局部自注意力的 Transformer 的模型进行了对比, 并通过消融实验分析验证了本文提出的局部自注意力机制 SGW-MSA 模块的有效性.

2.1 热力图对比实验分析

热力图通常是对类别进行可视化的图像, 表示着模型特征提取的能力. 图 8 展示了本文算法与基线算法 Swin Transformer 的热力图对比, 第一行是原图, 第二行是 Swin Transformer 的热力图, 第三行是 SGWin Transformer 的热力图. (a)、(b)、(c) 列的对比可以看出 SGWin Transformer 比 Swin

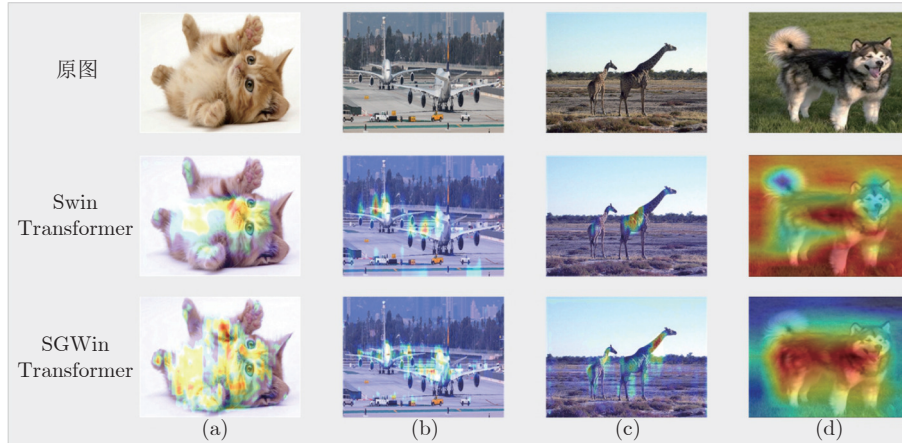


图 8 本文算法与 Swin Transformer 的热力图对比

Fig. 8 Comparison between the algorithm in this paper and the thermal diagram of Swin Transformer

Transformer 热力图覆盖的目标范围更全面; (d) 列的对比可以看出 SGWin Transformer 比 Swin Transformer 的定位更准确且小目标检测能力更强. 所以 SGWin Transformer 算法比 Swin Transformer 算法的目标定位更加准确, 也验证了本文提出的 SGW-MSA 局部自注意力机制的有效性. 此外 SGWin Transformer 对小目标检测的性能也有一定提升.

2.2 消融实验分析

为了验证 SGW-MSA 模块的有效性, 首先在 MS COCO 数据集上进行了消融实验分析. 实验使用 mmdetection^[39] 目标检测库以及 Mask R-CNN^[40] 目标检测框架, 将主干网络替换为 Swin Transformer, 然后依次将本文改进的策略加入到 Swin Transformer 中进行实验, 优化器采用对超参数不敏感的 AdamW^[41] 优化算法更新参数, 训练 Epoch 为 12, 初始学习率为 1×10^{-4} , 在第 8 Epoch 和第 11 Epoch 结束时分别衰减 10 倍, 评价指标采用目标检测平均精度 AP^b 以及实例分割平均精度 AP^m .

2.2.1 GWR 策略超参数消融实验分析

GWR 策略通过横条和竖条状的基础元素块重组特征图来建立距离小于 d_{\min}^h (d_{\min}^w) 或者大于 d_{\max}^h (d_{\max}^w) 的窗口的连接, 对于基础元素块的宽度 W_b 的设置会直接影响重组后的特征图的结果, 也会对网络的性能造成影响. 为了验证 W_b (W_b 小于局部窗口的宽和高) 的最佳取值, 本文在默认窗口大小为 7×7 的情况下, W_b 的值从 1 到 6 取值进行对比实验, 在不使用预训练模型的情况下, 实验结果如表 2 所示.

从表 2 中可以看出当基础元素块的宽度 W_b 从 1 到 6 改变的过程中, 在 1 到 4 的区间内精度呈现上升趋势, 在 4 到 6 区间内精度呈现下降趋势, 在

表 2 基础元素块宽度消融实验对比

Table 2 Comparison of ablation experiments of basic element block width

W_b	AP^b (%)	AP^m (%)
1	34.2	31.9
2	34.9	32.5
3	35.8	33.2
4	36.3	33.7
5	35.5	32.4
6	34.7	32.0

取值为 4 时模型的精度达到了最高, 达到了最好的效果, 所以本文的 GWR 策略中基础元素块的宽度确定为 4.

2.2.2 纵向 VGW-MSA 与横向 HGW-MSA 的消融实验分析

在验证 GW-MSA 局部自注意力中包含的纵向 VGW-MSA 和横向 HGW-MSA 的有效性时, 本文依次将基线算法 Swin Transformer 的 SW-MSA 替换为 Shuffled W-MSA、Shuffled W-MSA+VGW-MSA、Shuffled W-MSA+VGW-MSA+HGW-MSA, 逐步验证每个模块的有效性, 在不使用预训练模型的情况下, 实验结果如表 3 所示.

从表 3 中可以看出本文算法的基线模型 Swin Transformer 使用 SW-MSA 局部自注意力的目标检测和实例分割的平均精度分别为 30.8% 和 29.5%; 将 SW-MSA 替换为 Shuffled W-MSA 后精度分别提升了 2.8% 和 2.1%; 将 SW-MSA 替换为 Shuffled W-MSA 与纵向高斯窗口自注意力 VGW-MSA 的结合后精度分别提升了 1.3% 和 1.1%; 将 SW-MSA 替换为 SGW-MSA (Shuffled W-MSA+VGW-MSA+HGW-MSA) 后精度分别提升了 1.4% 和

表 3 SGW-MSA 消融实验结果

Table 3 SGW-MSA ablation experimental results

序号	方法	AP^b (%)	AP^m (%)
A	SW-MSA (baseline)	30.8	29.5
B	Shuffled W-MSA	33.6 (+2.8)	31.6 (+2.1)
C	B+VGW-MSA	34.9 (+1.3)	32.7 (+1.1)
D	C+HGW-MSA	36.3 (+1.4)	33.7 (+1.0)

1.0%。这些消融实验的数据进一步验证了本文提出的 SGW-MSA 局部自注意力机制的有效性。

2.2.3 三种局部自注意力特征图融合的消融实验与分析

为了更直观地感受到 SGW-MSA 联合 3 种自注意力机制的优势, 选用 ImageNet 中的图像分别可视化 3 种局部自注意力机制的注意力热力图。输入图像采用 224×224 像素的尺寸, 每一个 stage 中特征图的尺寸分别为 56×56 , 28×28 , 14×14 , 7×7 , 越靠后的 stage 可视化出的热力图覆盖的物体范围越大、效果越好, 但是考虑到最后一个 stage 特征图的尺寸为 7×7 等于局部自注意力机制的窗口大小, 此时的三个局部自注意力全部退化为全局自注意力。因此选取第 3 个 stage 中最后一个 SGWin Transformer block 中 SGW-MSA 的 3 个自注意力的热力图进行可视化对比。融合效果示意图如图 9 所示。

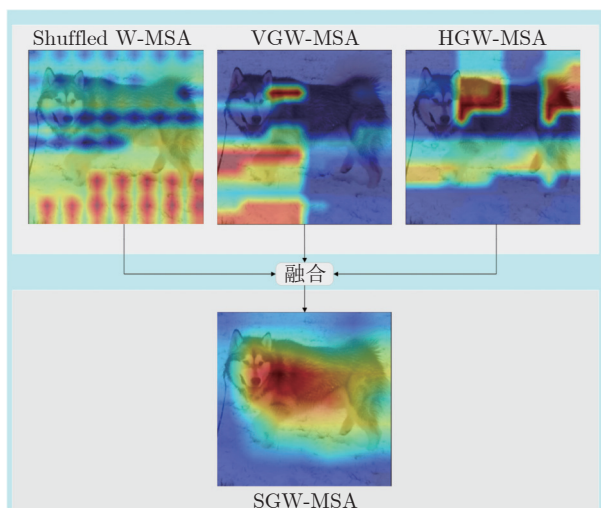


图 9 融合效果示意图

Fig.9 Schematic diagram of fusion effect

图 9 展示了各部分注意力机制的输出结果。可以看到每一种注意力的关注部分都有所不同。Shuffled W-MSA 建立固定距离的窗口连接, 对跳跃的关注目标和周围信息的联系比较敏感。VGW-MSA 建立纵轴上任意距离的窗口连接, 对目标和纵向背景之间的联系比较敏感。HGW-MSA 建立横

轴上任意距离的窗口连接, 更关注目标和横向背景之间的联系。因此, 相比于单一的局部自注意力机制, SGW-MSA 通过融合 3 种自注意力机制的方式, 具有更优秀的上下文信息提取能力。

2.3 图像分类实验

2.3.1 CIFAR10 图像分类实验

CIFAR10 数据集包含 60 000 张尺寸为 32 的彩色图片, 分为 10 个类别, 每一个类别有 6 000 张图像。分为训练集 50 000 张, 测试集 10 000 张。本文在训练集上训练模型, 并用测试集测试输出的 Top1 准确率 (排名第一的类别与实际结果相符的准确率)。在训练模型时, 采用 PyTorch 深度学习框架和 Timm 图像分类库, 优化器采用了对超参数不敏感的 AdamW^[42], 学习率采用余弦退火^[43]的方式, 初始的学习率设置为 1×10^{-3} , 最小学习率为 1×10^{-6} , warmup 学习率为 1×10^{-4} , warmup Epoch 设置为 3, 权重衰减率为 2×10^{-5} , 动量为 0.9, 数据增强采用随机裁剪和水平随机翻转。训练总轮数为 130 Epoch, 在 120 个 Epoch 之后保持最低学习率继续训练 10 Epoch。损失函数采用标准的交叉熵分类损失函数。在不使用预训练模型的情况下, 所有的模型均在一张 RTX2070 的 GPU 上训练, 基础配置采用表 1 中的配置。因为 CIFAR10 数据集中的图像较小, 所以配置中的窗口大小 S_i 设置为 3; 4 个阶段的通道数 C_i 分别对应 [32, 64, 128, 256]; 4 个阶段 Transformer block 的 Head 数量 H_i 分别设置为 [2, 4, 8, 16]; SGWin Transformer 的基础元素块的宽度 W_b 设置为 1。表 4 展示了参与对比的模型在 CIFAR10 数据集上的实验结果。可以看出本文所设计的 SGWin Transformer 在参数量相当的情况下的性能明显优于现有具有代表性的其他基于局部自注意力的 Transformer 模型。Top1 准确率比目前最先进的 Pale Transformer 提升 0.41%, 相比于基线算法 Swin Transformer, SGWin Transformer 在参数量相同的情况下, 仅仅通过替换 SW-MSA 为 SGW-MSA 就达到了 5.2% 的提升, 验证了本文设计的 SGW-MSA 的有效性。

2.3.2 mini-imagenet 数据集上的实验

本文还在 mini-imagenet 数据集上进行了实验。mini-imagenet 数据集包含 60 000 张图像, 分为 100 个类别, 每张图像的宽高中的长边均为 500 个像素, 每个类别的图像大约有 6 000 张。将 50 000 张图像作为训练集, 10 000 张图像作为验证集, 训练模型的设置基本与第 2.3.1 节中的 CIFAR10 数据集相同, 不同的是模型的超参数配置采用表 1 中

表 4 CIFAR10 数据集上的 Top1 精度对比
Table 4 Top1 accuracy comparison on CIFAR10 dataset

算法	Top1 准确率 (%)	Parameter (MB)
Swin Transformer	85.44	7.1
CSWin Transformer	90.20	7.0
CrossFormer	88.64	7.0
GG Transformer	87.75	7.1
Shuffle Transformer	89.32	7.1
Pale Transformer	90.23	7.0
SGWin Transformer	90.64	7.1

的配置, 训练的 Epoch 数为 100. SGWin Transformer 的基础元素块的宽度 W_b 设置为 4. 表 5 展示了参与对比的模型在 mini-imagenet 数据集上的实验结果. 从表 5 中的结果可以看出本文算法相比于基线 Swin Transformer 提升了 5.1%, 同时比最先进的 Pale Transformer 提升了 0.67%. 证明了 SGW-MSA 的有效性.

2.4 目标检测实验

2.4.1 MS COCO 数据集上的实验结果

本文使用 mmdetection 库以及 Mask R-CNN 目标检测框架, 将主干网络替换为所有具有代表性的基于局部窗口自注意力的 Transformer 模型, 并与本文的方法进行了对比, 采用 AdamW 优化器更新网络参数, 训练周期为 36 Epoch, 设置初始学习率为 1×10^{-4} , 在第 27 Epoch 和 33 Epoch 结束之后分别衰减 10 倍. 所有的模型均不使用预训练模型. 实验结果如表 6 所示. 其中 Params (M) 代表模型的参数量, FLOPs (G) 代表模型的计算复杂度. 可以看出本文提出的 SGWin Transformer 算法达到了 45.1% 的 mAP, 相比于目前最先进的 Pale Transformer 模型提升 1.8%, 并且在参数量不变的情况下比基线算法 Swin Transformer 提升了 5.5%. 此外, SGWin Transformer 在实例分割上也具有一定

表 5 mini-imagenet 数据集上的 Top1 精度对比
Table 5 Top1 accuracy comparison on mini-imagenet dataset

算法	Top1 准确率 (%)	Parameter (MB)
Swin Transformer	67.51	28
CSWin Transformer	71.68	23
CrossFormer	70.43	28
GG Transformer	69.85	28
Shuffle Transformer	71.26	28
Pale Transformer	71.96	23
SGWin Transformer	72.63	28

的提升, 比最先进的 Pale Transformer 提升了 1.3%, 比基线算法 Swin Transformer 提升了 4.2%, 也验证了本文提出的 SGW-MSA 的有效性. 此外使用 mmdetection 库以及 Cascade R-CNN^[44] 目标检测框架, 除训练周期外实验配置如同上述的 Mask R-CNN, 训练周期设置为 11 Epoch, 初始学习率为 1×10^{-4} , 在第 8 Epoch 和 11 Epoch 结束后分别衰减 10 倍. 实验结果如表 7 所示. 本文提出的 SGWin Transformer 算法达到 42.9% (AP^b) 和 37.8% (AP^m), 相比于 Pale Transformer 模型分别提升了 1.4% 和 1.7%, 并且在参数量不变的情况下比基线算法 Swin Transformer 分别提升了 5.1% 和 4.4%. 证明了 SGW-MSA 的有效性.

为了更直观地展示 SGWin Transformer 的有效性, 本文选取 MS COCO 测试集的图像进行检测并将结果进行可视化, 如图 10 所示. 以 Cascade R-CNN 为目标检测框架, 分别将 Swin Transformer 以及 SGWin Transformer 作为主干网络进行检测. 从图中可以看出, SGWin Transformer 相比于基线算法检测到了更多的小目标 (如图 10(a) 中心的人和车, 如图 10(b) 中心处的绵羊) 和遮挡目标 (图 10(c) 最下边的游艇, 图 10(d) 泳池中的人). 证明了 SGW-MSA 能够通过提取更多的上下信息来提高遮挡目标和目标的检测效果.

表 6 以 Mask R-CNN 为目标检测框架在 MS COCO 数据集上的实验结果
Table 6 Experimental results on MS COCO dataset based on Mask R-CNN

Backbone	Params (M)	FLOPs (G)	AP^b (%)	AP_{50}^b (%)	AP_{75}^b (%)	AP^m (%)	AP_{50}^m (%)	AP_{75}^m (%)
Swin	48	264	39.6	61.3	43.2	36.6	58.2	39.3
CSWin	42	279	42.6	63.3	46.9	39.0	60.5	42.0
Cross	50	301	41.3	62.7	45.3	38.2	59.7	41.2
GG	48	265	40.0	61.4	43.9	36.7	58.2	39.0
Shuffle	48	268	42.7	63.6	47.1	39.1	60.9	42.2
Focal	49	291	40.7	62.4	44.8	37.8	59.6	40.8
Pale	41	306	43.3	64.1	47.9	39.5	61.2	42.8
SGWin	48	265	45.1	66.0	49.9	40.8	63.5	44.2

2.4.2 在其他目标检测数据集上的实验结果

本文还在 KITTI 数据集和 PASCAL VOC 数据集上进行了对比实验, 使用 PyTorch 深度学习框架以及 YOLOv5^[45] 目标检测架构, 采用 SGD^[46] 优化器, 学习率采用余弦退火的方式, 初始学习率设置为 0.01, 最小学习率为 1×10^{-6} , warmup 学习率为 0.1, warmup 学习率为 0.1, warmup Epoch 为 3, 权重衰减为 5×10^{-4} , 动量为 0.937, 数据增强采用 Mosaic^[47]、水平翻转和色调变换. 在 3 张 RTX3090 的 GPU 上训练模型, 超参数采用表 1 中的配置. 采用上述的训练策略, 所有的算法均不使用预训练模型, 在 PASCAL VOC 数据集上训练 100 Epoch, 在 KITTI 数据集上训练 300 Epoch, 训练 Batch size 数为 64, 实验结果如表 8 所示. 可以看出在模

型参数量相当的情况下, 本文提出的 SGWin Transformer 模型在 KITTI 数据集和 PASCAL VOC 数据集的精度比最先进的 Pale Transformer 分别提升了 0.3 和 0.6, 比基线算法 Swin Transformer 分别提升了 1.9 和 4.5. 在检测速度方面, SGWin Transformer 的 FPS 达到了 56, 超出最先进的 Pale Transformer 算法 16%, 相比于基线算法 Swin Transformer 提升了 12%. 所以本文设计的 SGWin Transformer 在速度和精度上都优于其他 Transformer, 整体性能最好.

3 结论

本文针对现有的基于局部自注意力机制的 Transformer 模型不能建立所有窗口之间信息流通的

表 7 以 Cascade R-CNN 为目标检测框架在 MS COCO 数据集上的实验结果
Table 7 Experimental results on MS COCO dataset based on Cascade R-CNN

Backbone	Params(M)	FLOPs(G)	AP^b (%)	AP_{50}^b (%)	AP_{75}^b (%)	AP^m (%)	AP_{50}^m (%)	AP_{75}^m (%)
Swin	86	754	47.8	55.5	40.9	33.4	52.8	35.8
CSWin	80	757	40.7	57.1	44.5	35.5	55.0	38.3
Cross	88	770	39.5	56.9	43.0	34.7	53.7	37.2
GG	86	756	38.1	55.4	41.5	33.2	51.9	35.1
Shuffle	86	758	40.7	57.0	44.4	35.8	55.1	38.0
Focal	87	770	38.6	55.6	42.2	34.5	53.7	39.0
Pale	79	770	41.5	57.8	45.3	36.1	55.2	39.0
SGWin	86	756	42.9	60.9	46.3	37.8	57.2	40.5

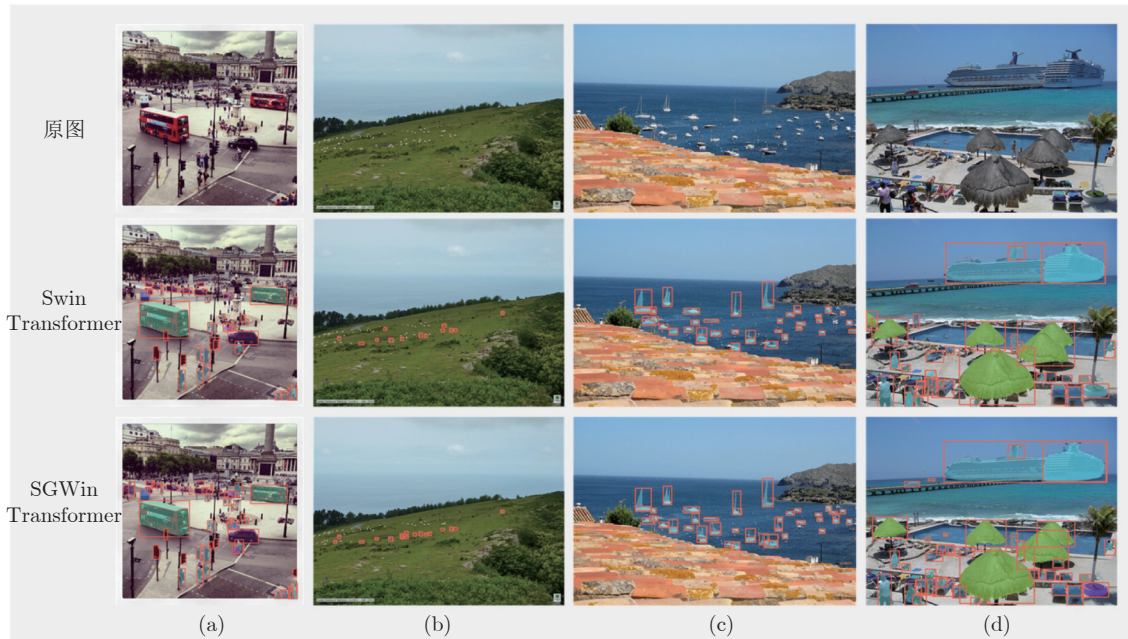


图 10 MS COCO 检测结果或可视化

Fig.10 MS COCO test results or visualization

表 8 KITTI 和 PASCAL VOC 数据集上的实验结果

Table 8 Experimental results on KITTI and PASCAL VOC dataset

Backbone	KITTI mAP@0.5:0.95	VOC mAP@0.5	Params (M)	FPS
Swin	57.3	59.6	14.4	50
CSWin	58.7	64.1	14.2	48
Cross	58.1	62.8	13.8	20
Shuffle	58.7	64.6	14.4	53
GG	57.8	62.4	14.4	46
Pale	58.9	64.5	14.2	48
SGWin	59.2	65.1	14.4	56

问题, 提出了一种 SGW-MSA 局部自注意力以及 SGWin Transformer 模型, 在 SGW-MSA 中结合 3 种不同的局部自注意力机制的特点, 有效地建立所有窗口之间的信息交互. 实验结果表明在参数量和计算量相当的情况下, 本文提出的算法比现有的基于局部自注意力的 Transformer 模型更具有优势, 证明了本文提出的 SGW-MSA 通过高斯随机窗口策略建立所有窗口之间的信息流动能够捕捉更多的特征图语义信息并且具有更强大的上下文建模能力.

References

- Jiang Hong-Yi, Wang Yong-Juan, Kang Jin-Yu. A survey of object detection models and its optimization methods. *Acta Automatica Sinica*, 2021, **47**(6): 1232–1255
(蒋弘毅, 王永娟, 康锦煜. 目标检测模型及其优化方法综述. 自动化学报, 2021, **47**(6): 1232–1255)
- Yin Hong-Peng, Chen Bo, Chai Yi, Liu Zhao-Dong. Vision-based object detection and tracking: A review. *Acta Automatica Sinica*, 2016, **42**(10): 1466–1489
(尹宏鹏, 陈波, 柴毅, 刘兆栋. 基于视觉的目标检测与跟踪综述. 自动化学报, 2016, **42**(10): 1466–1489)
- Xu Peng-Bin, Zhai An-Guo, Wang Kun-Feng, Li Da-Zi. A survey of panoptic segmentation methods. *Acta Automatica Sinica*, 2021, **47**(3): 549–568
(徐鹏斌, 翟安国, 王坤峰, 李大字. 全景分割研究综述. 自动化学报, 2021, **47**(3): 549–568)
- Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 2017, **60**(6): 84–90
- Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014.
- Huang G, Liu Z, Laurens V D M. Densely connected convolutional networks. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 4700–4708
- He K, Zhang X, Ren S. Deep residual learning for image recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770–778
- Xie S, Girshick R, Dollár P. Aggregated residual transformations for deep neural networks. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 1492–1500
- Szegedy C, Liu W, Jia Y. Going deeper with convolutions. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 1–9
- Tan M, Le Q V. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning. New York, USA: JM-LR, 2019. 6105–6114
- Tomar G S, Duque T, Tckstrm O. Neural paraphrase identification of questions with noisy pretraining. In: Proceedings of the First Workshop on Subword and Character Level Models in NLP. Copenhagen, Denmark: Association for Computational Linguistics, 2017. 142–147
- Wang C, Bai X, Zhou L. Hyperspectral image classification based on non-local neural networks. In: Proceedings of the International Geoscience and Remote Sensing Symposium. Yokohama, Japan: IEEE, 2019. 584–587
- Zhao H, Jia J, Koltun V. Exploring self-attention for image recognition. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 10073–10082
- Ramachandran P, Parmar N, Vaswani A. Stand-alone self-attention in vision models. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2019.
- Carion N, Massa F, Synnaeve G. End-to-end object detection with transformers. In: Proceedings of the 16th European Conference. Glasgow, UK: ECCV, 2020. 213–229
- Dosovitskiy A, Beyer L, Kolesnikov A. An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations. Virtual Event: ICLR, 2021.
- Chu X, Tian Z, Zhang B. Conditional positional encodings for vision transformers. In: Proceedings of the International Conference on Learning Representations. Virtual Event: ICLR, 2021.
- Han K, Xiao A, Wu E. Transformer in transformer. *Advances in Neural Information Processing Systems*, 2021, **34**: 15908–15919
- Touvron H, Cord M, Douze M. Training data-efficient image transformers distillation through attention. In: Proceedings of the International Conference on Machine Learning. Jeju Island, South Korea: PMLR, 2021. 10347–10357
- Yuan L, Chen Y, Wang T. Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet. In: Proceedings of the International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 558–567
- Henaff O. Data-efficient image recognition with contrastive predictive coding. In: Proceedings of International Conference on Machine Learning. Berlin, Germany: PMLR, 2020. 4182–4192
- Liu Z, Lin Y, Cao Y. Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 10012–10022
- Rao Y, Zhao W, Liu B. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in Neural Information Processing Systems*, 2021, **34**: 13937–13949
- Lin H, Cheng X, Wu X. CAT: Cross attention in vision transformer. In: Proceedings of the International Conference on Multimedia and Expo. Taipei, China: IEEE, 2022. 1–6
- Vaswani A, Ramachandran P, Srinivas A. Scaling local self-attention for parameter efficient visual backbones. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 12894–12904
- Wang W, Chen W, Qiu Q. Crossformer++: A versatile vision transformer hinging on cross-scale attention. arXiv preprint arXiv: 2303.06908, 2023.
- Huang Z, Ben Y, Luo G. Shuffle transformer: Rethinking spatial shuffle for vision transformer. arXiv preprint arXiv: 2106.03650, 2021.

- 28 Yu Q, Xia Y, Bai Y. Glance-and-gaze vision transformer. *Advances in Neural Information Processing Systems*, 2021, **34**: 12992–13003
- 29 Wang H, Zhu Y, Green B. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: Proceedings of the 16th European Conference. Glasgow, UK: ECCV, 2020. 108–126
- 30 Dong X, Bao J, Chen D. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2022. 12124–12134
- 31 Wu S, Wu T, Tan H. Pale transformer: A general vision transformer backbone with pale-shaped attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA: 2022. 2731–2739
- 32 Wang W, Xie E, Li X. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 568–578
- 33 Ren S, He K, Girshick R. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 2015: 28
- 34 Efraimidis P S, Spirakis P G. Weighted random sampling with a reservoir. *Information Processing Letters*, 2006, **97**(5): 181–185
- 35 Krizhevsky A, Hinton G. Convolutional beep belief networks on CIFAR-10. *Unpublished Manuscript*, 2010, **40**(7): 1–9
- 36 Geiger A, Lenz P, Stiller C. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*, 2013
- 37 Everingham M, Eslami S M A, Van Gool L. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 2015, **111**: 98–136
- 38 Veit A, Matera T, Neumann L. Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv: 1601.07140, 2016.
- 39 Selvaraju R R, Cogswell M, Das A. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 618–626
- 40 Chen K, Wang J, Pang J. MMDetection: Open MMLab detection toolbox and benchmark. arXiv preprint arXiv: 1906.07155, 2019.
- 41 He K, Gkioxari G, Dollár P. Mask R-CNN. In: Proceedings of the International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2961–2969
- 42 Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv: 1711.05101, 2017.
- 43 You Y, Li J, Reddi S. Large batch optimization for deep learning: Training bert in 76 minutes. arXiv preprint arXiv: 1904.00962, 2019.
- 44 Cai Z, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 6154–6162
- 45 Wu W, Liu H, Li L. Application of local fully convolutional neural network combined with YOLO v5 algorithm in small target detection of remote sensing image. *PloS One*, 2021, **16**(10): 1–10
- 46 Bottou L. Stochastic gradient descent tricks. *Journal of Machine Learning Research*, 2017, **18**: 1–15
- 47 Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv: 2004.10934, 2020.



赵亮 西安建筑科技大学信息与控制工程学院教授。主要研究方向为智能建筑检测, 计算机视觉和模式识别。本文通信作者。

E-mail: zhaoliang@xauat.edu.cn

(ZHAO Liang Professor at College of Information and Control Engineering, Xi'an University of Architecture and Technology. His research interest covers intelligent building detection, computer vision and pattern recognition. Corresponding author of this paper.)



周继开 西安建筑科技大学信息与控制工程学院硕士研究生。主要研究方向为图像处理和目标检测。

E-mail: m18706793699@163.com

(ZHOU Ji-Kai Master student at College of Information and Control Engineering, Xi'an University of Architecture and Technology. His research interest covers image processing and object detection.)