计算机科学
COMPUTER SCIENCE

基于改进Swin Transformer的中心点目标检测算法

刘家森，黄俊

引用本文

刘家森，黄俊. 基于改进Swin Transformer的中心点目标检测算法[J]. 计算机科学，2024，51(6)：264-271.

LIU Jiasen, HUANG Jun. Center Point Target Detection Algorithm Based on Improved Swin Transformer [J]. Computer Science, 2024, 51(6): 264-271.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300222

基于改进 Swin Transformer 的中心点目标检测算法

刘家森　黄　俊
重庆邮电大学通信与信息工程学院　重庆 400065
(2352556955@qq.com)

摘　要　针对 Swin Transformer 在提取局部特征信息和特征表达能力上存在的不足,提出了一种基于改进 Swin Transformer 的中心点目标检测算法,以提高其在目标检测方面的性能。通过调整网络结构和引入反卷积模块来增强网络对局部特征信息的提取能力,利用自适应二维高斯核和回归头模块检测目标中心点来增强特征表达能力,并在 Swin Transformer block 模块中加入 dropout 激活函数,以缓解网络过拟合问题。在 Pascal VOC 和 MS COCO 2017 数据集上分别对改进后的算法进行验证,实验结果表明,改进后的 Swin Transformer 算法在 Pascal VOC 数据集上的精确度达到了 81.1%,在 MS COCO 数据集上的精确度达到了 37.2%,明显优于其他主流目标检测算法。

关键词:深度学习;图像处理;目标检测;反卷积;Swin Transformer
中图分类号　TP391

Center Point Target Detection Algorithm Based on Improved Swin Transformer

LIU Jiasen and HUANG Jun
School of Communication and Information Engineering,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

Abstract　Aiming at the shortcomings of Swin Transformer in extracting local feature information and expressing features,this paper proposes a center point target detection algorithm based on improved Swin Transformer to improve its performance in target detection. By adjusting the network structure and introducing a deconvolution module to enhance the network's ability to extract local feature information,using an adaptive two-dimensional Gaussian kernel and a regression head module to detect the center point of the target,so as to enhance the feature expression ability,and adding a dropout activation function to the Swin Transformer block module to alleviate the network overfitting problem. The improved algorithm is validated on the Pascal VOC and MS COCO 2017 datasets,respectively. The experimental results show that the improved Swin Transformer algorithm achieves an accuracy of 81.1% on the Pascal VOC dataset and 37.2% on the MS COCO dataset,significantly superior to other mainstream object detection algorithms.

Keywords　Deep learning,Image processing,Object detection,Deconv,Swin Transformer

1　引言

目标检测是计算机视觉和数字图像处理领域的一个热门研究方向,被广泛应用于智慧检测、自动驾驶、航空航天等任务中[1]。同时,目标检测也是许多其他计算机视觉任务的基础部分,如语义分割、实例分割、目标跟踪等。

近年来,基于深度学习的目标检测逐渐成为主流,深度学习的方法可以分为单阶段(one stage)和双阶段(two stage)的方法[2]。以 faster R-CNN[3]为代表的双阶段算法在得到候选区域后,对候选区域做分类和回归,该类算法相较于单阶段算法精度更高但速度更慢。以 YOLOV3[4]系列为代表的单阶段算法不再额外设计候选区域,它们通过端到端的网络设计,直接对目标结果进行回归,在速度上有较大的提高。后续对 YOLOV5[5]锚框、激活函数以及网络层进行了改进。但是,由于该系列 YOLO 都是基于 anchor base,会有大量的冗余锚框,因此对检测速度有一定的影响。

为了避免像大多数基于 anchor base 的单阶段算法一样需设置大量的锚框进行计算,以 FCOS[6]为代表的基于 anchor free 的单阶段算法通过选取关键点获得目标位置以及类别,从而减少了冗余计算和模型参数,提高了检测速度和效率。

以上基于深度学习的目标检测都是基于 CNN 框架,CNN 能够有效提取局部信息,但缺乏整合全局信息的能力。Transformer[7]可以弥补这一缺陷,其通过多头注意力机制能够包含全局特征,在自然语言处理方面有很好的表现。传统 Transformer 主要针对自然语言处理,相较于图像视觉处理存在以下

到稿日期:2023-03-29　返修日期:2023-08-25
基金项目:国家自然科学基金(61771085)
This work was supported by the National Natural Science Foundation of China(61771085).
通信作者:黄俊(huangjun@cqupt.edu.cn)

劣势：视觉实体的规模较小，以及与文本单词的高分辨率图像像素差异巨大。VIT[8]是最早将 Transformer 应用在图像视觉处理任务中的模型，该模型通过简单堆叠 Transformer 块用于提取图像特征，导致网络计算复杂度很高且只能进行图像分类任务。Swin Transformer[9]提出了一种层次化的 Transformer 结构，通过移位窗口方案将自注意力计算限制在不重叠的局部窗口的同时，还允许跨窗口拼接特征图。这种分层架构具有更高的计算时效性和更低的计算复杂度。

Swin Transformer 在计算机视觉领域中表现出了优秀的性能，但在跨窗口拼接时会存在部分局部特征信息丢失的问题，这会导致检测精度下降。目标检测往往需要较为准确的局部特征信息，基于 CNN 目标检测方案的 DSSD[10]使用反卷积模块来增强模型的局部特征提取能力。在 DIT 中，Zhou 等[11]认为反卷积操作可被视为对低分辨率特征信息的“扩张”，有利于提取局部特征信息。此外，配合使用 Transformer 等基于自注意力机制的模块可以进一步提取更丰富的目标特征信息。

在下游目标检测任务中，Swin Transformer 没有采用类似于 CNN 的特征表达方法来检测目标，而是直接通过全连接层展开得到结果。这种做法可能会导致提取到的特征丢失，因此需要采用更加丰富的特征表达方式。基于 CNN 的目标检测算法特征表达方式十分丰富，例如 Centernet[12]等模型通过中心点和高斯分布函数等方式提取特征，且可以取得较好的检测效果。

在深度学习网络模型训练过程中网络模型复杂等原因导致网络过拟合现象十分常见。Hinton 等[13]提出了 dropout 技术，通过随机丢弃神经元来防止模型对某些特征过度依赖而导致过拟合。在各类目标检测网络中，dropout 已成为不可或缺的技术之一，其有效提高了模型的鲁棒性和泛化能力。

本文提出了一种基于改进 Swin Transformer 的中心点目标检测算法。首先引入反卷积模块来增强局部特征提取能力，从而解决了局部特征信息丢失的问题；其次，利用自适应二维高斯核和回归头模块进行目标中心点检测，从而增强模型的特征表达能力；最后，在 Swin Transformer block 模块中增加 dropout 激活函数，以解决网络层数过深导致的过拟合问题。实验结果表明，本文算法将其应用于目标检测，比原算法取得了更好的检测结果。

2 相关工作

2.1 中心点检测

中心点检测是目标检测中的一种重要方法，其主要思想是通过预测目标中心点位置和大小再通过回归来获取目标位置。中心点目标检测只需要检测每个目标的中心点位置，需要预测的参数更少，从而减小了计算量。此外，中心点也可以更准确地估计目标位置和大小，因为它不需要类似锚点进行预测。最后，中心点目标检测方法可以通过在每个中心点处预测每个目标类别的概率来实现多类别检测，这使得它适用于复杂场景中的多类别检测任务。

基于传统锚框和中心点目标检测方法都是将目标从图像背景中分离出来，但是两者采用的方法不同，区别如图 1 所示。

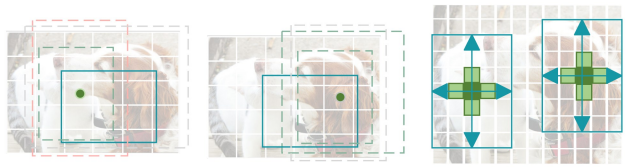


图 1 基于锚框和中心点检测的区别

Fig. 1 Differences between detection based on anchor box and center point

图 1(a)所示，在基于锚框的检测中，通过计算锚框与标注的目标边界框之间的 IoU，可以判断该锚框是否包含前景物体或背景。传统基于锚框的方法中，当 IoU 大于 0.7 时，记该锚框包含的内容为有效目标，反之，当 IoU 小于 0.3 时，则该锚框包含的内容为背景。图 1(b)所示的基于中心点的检测中，利用目标的中心像素来表示检测的目标对象，周围的像素点则为背景，通过回归得到目标对象的大小。这种方法可以避免传统方法中需要先验框的设计和选择，同时也更加精确和灵活。

Centernet 是一种基于中心点单阶段无锚框的目标检测方法。相较于 Cornernet 中的角点检测方式，Centernet 使用高斯分布来表示目标的中心点位置，使其更加灵活，不受尺度、形状、旋转等因素的限制。Centernet 用目标中心点检测方式代替传统目标检测中的锚框，从而降低了计算量以及锚框带来的正负样本不均的影响。

Centernet 通过预测目标中心点的坐标来代表该目标，并利用预测出的目标中心点与宽高的偏移量来确定目标的矩形框。在预测阶段，Centernet 首先对图像进行下采样，并在下采样的特征图上对每个类别预测出中心点的位置。然后，针对每个类别，将下采样的特征图中的每个像素视为该类别的热点，并将其单独提取出来。具体来说，对于每个热点，通过检测当前像素点是否比周围 8 个相邻点的值都大(或相等)，来判断其是否为该类别的中心点。在得到每个类别的中心点位置后，Centernet 再通过预测中心点的偏移量与宽高来获得目标的最终检测框。图 2 展示了 Centernet 网络模型预测出的中心点、中心点偏移量以及该点对应目标的宽高。

图 2 Centernet 中心点检测流程图

Fig. 2 Diagram of Centernet center point detection process

2.2 Swin Transformer

Swin Transformer 作为 Transformer 在计算机视觉领域的一大成果，在目标检测、图像分类、语义分割等下游任务中取得了优秀表现。作为图像视觉处理的一种通用 Transformer 框架，Swin Transformer 主要分为 3 个部分：块嵌入层(Patch Embedding)、移位 Transformer 块(Swin Transformer block)、块合并层(Patch Merging)。

在块嵌入层中，首先将输入的大小为(H,W,3)的图片

分割成大小为 $P\times P\times 3$ 的小块，其中 P 默认为 4，被压缩成固定维度的向量，这个固定维度的长度等于 $P\times P\times 3$，即 48。接着，每个小块的向量会通过一个全连接层(Linear Embedding)映射到任意维度 C，本文 C 为 128，因为采用的是 Swin-B 版本。最后，将所有小块的向量拼接成一个序列，其中序列的长度为 $N=(H/P)\times(W/P)$，每个向量的长度为 C，该序列作为 Transformer 的输入进行特征提取。

在块合并层模块中，将 Patch 按通道数拼接起来，然后使用一个 1×1 的卷积操作对通道数进行压缩处理，从而将 patch 层的长宽分辨率和通道数都缩减一半。长宽分辨率降低这个过程类似于 CNN 中的池化操作，但不同之处在于，块合并层的缩小操作是通过卷积实现的，而不是简单地取最大值或者平均值。图 3 展示了块合并层的工作流程。一个形状为 (H,W,C) 的张量如图 3(a)所示，下采样两倍之后得到图 3(b)所示的 4 个小张量块，然后把小张量块按照通道拼接得到如图 3(c)所示的通道，后续为了减少计算量，通道数通过 1×1 卷积之后减半，这里 Swin Transformer 默认保留前两个通道。该模块的主要作用是将 patch 层中的信息整合并降维，以在减少计算量的同时保留重要位置信息，从而提高模型的效率和泛化能力。

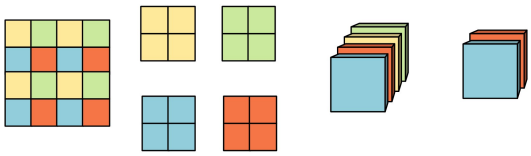


图 3 块合并层流程图

Fig. 3 Diagram of block merge layer process

移位 Transformer 模块是 Swin Transformer 的核心部分，它包含了多个 Swin Transformer 块，其结构如图 4 所示。每个 Swin Transformer 块首先对输入特征进行标准化(Layer Normalization)，然后计算窗口注意力(W-MSA)，通过残差连接将结果加到输入特征上，并再次进行标准化。接下来，经过多层感知机(MLP)进行非线性变换后，得到变换后的特征表示。在后续的网络结构中，只需要将 W-MSA 替换为移位窗口注意力(SW-MSA)，其他步骤保持不变。

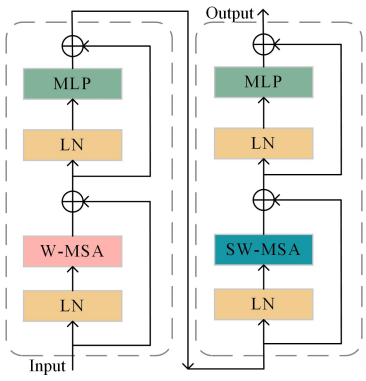


图 4 Swin Transformer Backbone 图

Fig. 4 Diagram of Swin Transformer Backbone

Swin Transformer block 中 W-SMA 和 SW-MSA 的滑窗流程示意图如图 5 所示。输入图片如图 5(a)所示，W-SMA 首先将其划分为 2×2 个大小为 4×4 的不重叠窗口，如图 5(b)所示。为了实现不同窗口之间的注意力计算，WM-MSA 将规则区域分布的窗口沿着垂直向下和水平向右方向都移动 $p/2$ 个距离，p 为窗口的大小，其值为 4，此时得到了 9 个规则不一的 patch，如图 5(c)所示。最后通过 SW-MSA 把不规则的 patch 进行循环移位拼接得到新的 4 个大小一样的 patch，如图 5(d)所示。此外，为了防止不属于一个区域的 patch 之间的干扰，Swin Transformer 加入了掩码以防止不同块之间的注意力干扰。

(a)输入图像 (b)WMSA 窗口分割 (c)窗口滑动操作 (d)SW-MSA 窗口分割

图 5 滑窗操作图

Fig. 5 Diagram of liding window operation

3 整体算法

3.1 算法设计

在 swin Transformer 中加入反卷积 Deconv 模块、回归头 Head 模块和自适应二维高斯核。其中 Deconv 网络结构由一次卷积和一次反卷积组成，通过上采样得到高分辨特征图，并在高分辨特征图上执行卷积操作，从而捕获更多的局部信息。Head 部分网络结构采用两次卷积组成，自适应二维高斯核和 Head 根据高分辨特征图进行计算，对高分辨率特征图实现中心点回归预测，增强模型的特征表达能力。算法网络架构图如图 6 所示。

在结构上，Swin Transformer block 的 LN 层会在每个子层的输入和输出之间添加，这样可以有效地规范化每个子层的输出，但是会导致神经元之间的相关性增加，从而导致过拟合的风险。dropout 可以随机地将某些神经元的输出值设为 0，从而使得每个神经元都不会过于依赖其他神经元的输出。这样可以使得模型更加鲁棒，从而减少过拟合的风险。因此，在 LN 层之前引入 dropout 可以减少神经元之间的相关性，从而缓解过拟合的问题。

在输入部分，网络的输入图像长宽分辨率是 512，为了更好地满足 Swin Transformer 在后续网络中对 patch 的划分，输入图像分辨率经过填充修改为 384。此时输入图像大小为 (384,384,3)。

在 Backbone 部分，根据表 1 中所列模型种类及特点可知，Swin-B 版本的 Backbone 部分输出通道维度和目前主流算法通道数相似，因此采用 Swin Transformer 的 Swin-B 版本，有助于后续对比实验验证。在后续 Deconv 网络模块中进行 3 次上采样。经过块嵌入层之后，输入维度的长度分别变为原来的 1/4，得到大小为 (96,96,48) 的特征图。然后，该特征图经过图 4 中的 4 个 stage，每次长宽尺度缩小一半，通道数增加一倍，其中 C 默认为 128，经过 4 个 stage 之后，长宽为 $384/32=12$，通道数为 $128*8=1\,024$，所以 Backbone 最终得到的输出维度为 (12,12,1 024)。

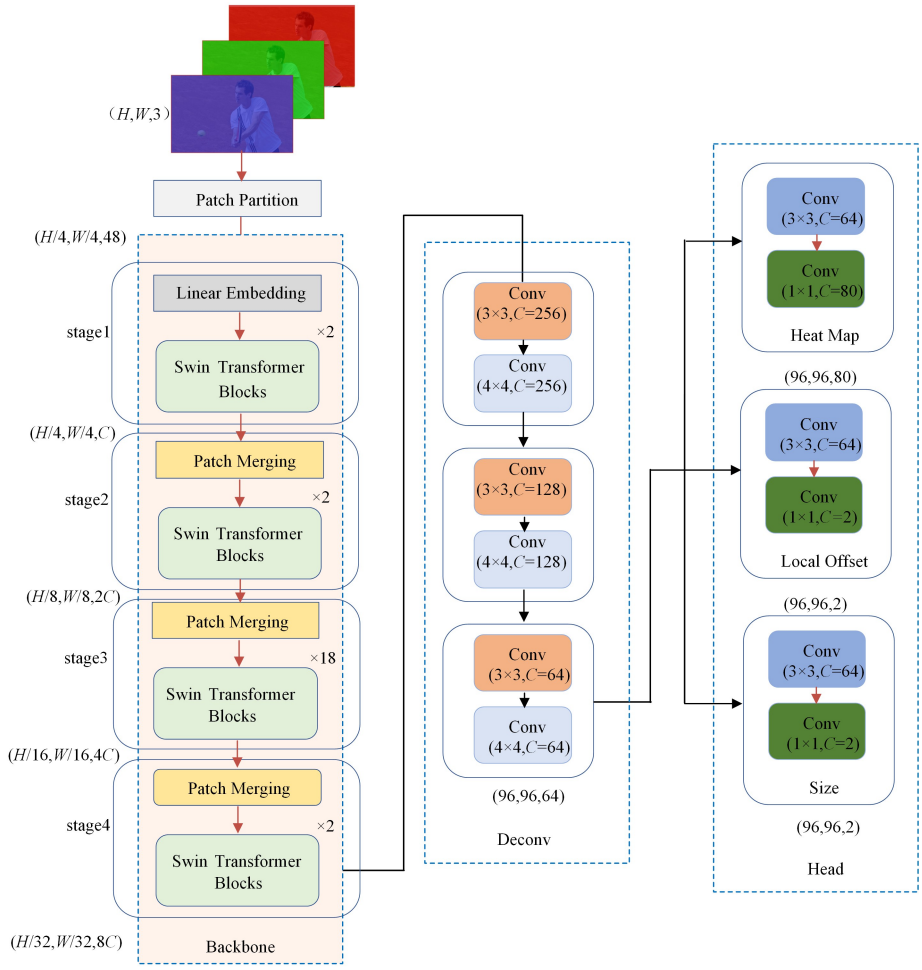


图 6　整体网络架构图

Fig. 6　Overall network architecture

表 1　Swin Transformer

Table 1　Swin Transformer

	Input dim	Head number	Block's number	Outpt dim
Swin-T	96	(3,6,12,24)	(2,2,6,2)	768
Swin-S	96	(3,6,12,24)	(2,2,18,2)	768
Swin-B	128	(4,8,16,32)	(2,2,18,2)	1024
Swin-L	192	(6,12,24,48)	(2,2,18,2)	1536

3.2　Deconv 模块

Swin Transformer 作为具有全局特征信息提取能力的网络，具备优秀的全局交互能力，但在移窗拼接时可能会丢失部分局部信息。这是因为 Swin Transformer 中的注意力机制是基于每个块内的位置编码进行计算的，如果两个感受野存在于不同的块中，则可能无法捕捉到它们之间的局部相关性。

为了增强局部信息的特征提取能力，在 swin Transformer 中加入反卷积 Deconv 模块。反卷积操作可以看作是在输入特征图中间插入一些零值像素，然后使用一个卷积核对这些像素进行卷积操作，从而将特征图的尺寸放大到原来的几倍。如此，网络就能够对更大的感受野进行建模，提取到更多的的局部特征信息。

在 Deconv 反卷积部分，Backbone 得到特征信息经过 Deconv 模块时，将特征信息进行 3 次上采样，每次上采样使得特征信息的长宽各增大一倍，通道数缩小一半，且最后一层的卷积核大小如图 4 所示，为 4×4，$C=64$。因此，经过 Deconv 之后，长宽为 $12*8=96$，得到特征信息为(96,96,64)的高分辨率特征图。

3.3　Head 和高斯核模块

增强 Swin Transformer 的特征表达能力是提升其检测性能的下游重要任务之一，Swin Transformer 在目标检测任务中是通过全连接层直接把 Swin Transformer 得到的特征进行展开、综合，其主要思想是将特征图映射成一个固定大小的向量，然后通过全连接层进行分类和回归。但是全连接层会导致参数过多，以至于特征丢失，检测性能降低。

为了增强模型的特征表达能力，本文设计了中心点检测目标方法。具体做法是加入高斯核和回归头，高斯核可以调整目标大小和中心点位置，从而使得检测器具有更好的空间不变性和鲁棒性。回归头则可以对目标的中心点位置、大小和偏移进行回归。这种方法的优点是可以准确地定位目标，具有较好的鲁棒性。

首先，将尺寸分辨率为 384 的图片输入本文网络中。通过 Backbone 和 Deconv 的处理，得到尺寸分辨率为(96,96,64)的高分辨率特征图。由于目标尺寸缩小为原来的 1/4，因此在高分辨率特征图尺度中，我们也将输入图片中目标的边框缩小为原来的 1/4。接下来，根据边框的宽高计算出目标的中心点坐标。然后，利用中心点坐标和边框信息计算高斯

核半径。最后，根据中心点坐标和高斯半径计算高斯值。这一过程中，以中心点为圆心，沿着半径向外按高斯核衰减，以得到高斯分布的权重，从而准确地确定目标的位置。

相较于一维高斯核只能处理一维信号，二维高斯核可以同时处理图像的水平和垂直方向上的信号，因此可以更好地适应二维图像处理任务。同时，二维高斯核是一个固定大小的卷积核，用于对图像进行滤波操作。而自适应二维高斯核是根据图像局部像素信息自适应生成的卷积核，具有可调整的大小和形状，可以更好地适应不同尺度和方向的特征，能够在不同位置和尺度上对图像进行平滑和边缘保留处理，从而提取更加准确的特征。本文采用自适应二维高斯核，计算公式如下：

$$K_m(x,y)=\exp\left(-\frac{(x-x_0)^2}{2\partial_x^2}-\frac{(y-y_0)^2}{2\partial_y^2}\right)$$
$$\partial_x=\frac{w}{6a},\partial_y=\frac{h}{6\partial},H_m\in R^{1\times\frac{H}{r}\times\frac{W}{r}} \tag{1}$$

其中，(x_0,y_0)为目标的中心点坐标；∂_x 和∂_v 为目标尺寸自适应标准差；α 和为超参数，分别为 2 和 4；(h,w)为目标边界框，(H,W)为输入图片的尺寸。

经过 Deconv 后，自适应二维高斯核和 Head 根据高分辨特征图进行计算，对高分辨率特征图实现中心点回归预测，包括对目标的 heatmap，offset，以及 size 信息。heatmap 卷积的通道数为 *num_class*，结果为(96，96，*num_class*)，其中 *num_class* 为数据集种类数量。本文中的 COCO 数据集，其种类是 80 个，代表每一个热力点是否有物体存在以及物体类别。中心点卷积的通道数为 2，结果为(96，96，2)，代表每个物体中心点距离热力点偏移的情况。宽高卷积的通道数为 2，结果为(96，96，2)，代表每个物体宽高的预测情况。

3.4 损失函数

本文的损失函数 L_{det} 计算式如式(2)所示，其由 3 个回归头损失构成，且分为两种类型，分别是热力图损失 L_k、目标大小损失 L_{size} 和中心偏移量损失 L_{off}。

$$L_{\text{det}}=L_k\lambda_{\text{size}}L_{\text{size}}+\lambda_{\text{off}}L_{\text{off}} \tag{2}$$

其中，调节因子 $\lambda_{\text{size}}=0.1$，$\lambda_{\text{off}}=1$。

热力图损失计算式如(3)所示：

$$L_k=\frac{-1}{N}\sum_{xyc}\begin{cases}(1-y_{xyc})^{\alpha}\ln(y_{xyc}), & y_{\text{xyc}}=1\\(1-y_{xyc})^{\beta}y_{xyc}^{\alpha}\ln(1-y_{xyc}), & y_{xyc}\neq 1\end{cases} \tag{3}$$

其中，N 为输入图像中关键点的数量；α 和 β 为超参数，默认为 2 和 4；y_{xyc} 为关键点的实际位置和类别；y_{xyc} 为关键点位置和类别的预测值。

中心偏移量损失计算式如式(4)所示：

$$L_{\text{off}}=\frac{1}{N}\sum_{p}\left|o_p-\left(\frac{p}{R}-p\right)\right| \tag{4}$$

其中，o_p 为网络预测后的偏移值；p 表示图像中心的坐标值；p 表示缩放后的中心近似坐标值；R 表示缩放因子。

目标大小损失计算式如式(5)所示：

$$L_{\text{size}}=\frac{1}{N}\sum_{K=1}^{N}|S_{PK}-S_K| \tag{5}$$

其中 S_{PK} 表示目标检测大小，S_K 表示目标真实大小。

4 实验设置

4.1 数据集

常用的目标检测数据集包括 PASCAl VOC[14] 和 MS COCO[15]。本文在 COCO 数据集上完成对网络的训练和测试，在 PASCAL VOC 2007 数据集上进行测试。MS COCO 数据集全称是 Common Object in Context。MS COCO 2017 数据集是微软团队于 2017 年提供的一个可用于目标检测的数据集，此外该数据集还提供目标分类、语义分割等下游任务。COCO 目标数据集分为 3 个部分，包括 COCO train-2017 训练集、COCO val-2017 验证集、COCO tets-2017 测试集。数据集分布情况如表 2 所列，共有 80 个生活中常见事物的类别，每张图片平均包含 3.5 个类别，且以小目标居多。

表 2 MS COCO 2017 数据集

Table 2 MS COCO 2017 dataset

Dataset	train	test	val
MS COCO2017	118287	5000	40670

PASCAL VOC 2007 数据集是用于图像分类、目标检测、语义分割的经典数据集之一。该数据集包含 20 种类别，其中人是该数据集中占比最多的类别，占总数的 32%，牛和羊占比较少，分别为 2.9% 和 3.3%。数据集整体分布情况如表 3 所列。

表 3 PASCAL VOC 2007 数据集

Table 3 PASCAL VOC 2007 dataset

Dataset	train	test
PASCAL VOC 2007	5011	4952

4.2 实验环境

本文实验平台为 Ubuntu 20.04，开发语言为 python3.7，深度学习框架为 pytorch1.7，CUDA 版本为 11.1，cpu 为至强 Platinum 8350，主频 2.60 GHz、内存 43 GB、硬盘 1T，显卡为 RTX3090，24 GB 显存。

对搭建的网络模型利用迁移学习和冻结的方式进行训练，加载了 Swin Transformer 模型权重对参数进行初始化，在前 80 个 epoch 设置学习率为 3×10^{-5}，batchsize 设置为 16，droupt 为 0.3。然后进行模型解冻，调整学习率为 5×10^{-5}，batchsize 设置为 8，droupt 为 0.1，训练 30 个 epoch。并设置 loss 变化率提结束训练。

本文在加入 dropout 和没加入 dropout 这两种情况下对 Train Loss 和 Test Loss 进行了对比，损失函数图如图 7 所示。

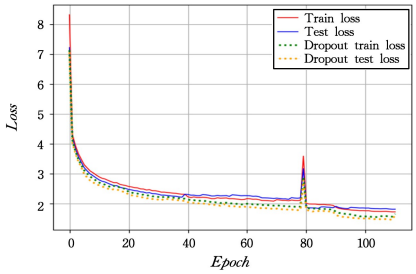


图 7 损失函数图

Fig. 7 Loss function diagram

从图 7 中可以看出，在未加入 dropout 时，网络训练到了 40 轮时，Test Loss 开始超过 Train Loss，出现了典型的网络过拟合现象，这是因为 Swin Transformer 相较于传统 CNN 网络模型复杂度较高，如果训练策略也设置不当，容易导致过拟合。在 80 个 epoch 之后，将模型解冻，各个 loss 在急剧上升后迅速下降，正常训练。但是未加入 dropout 的模型在大约 90 轮左右仍然存在网络过拟合的问题，而加入了 dropout 的网络模型并没有出现过拟合，随着训练到 110 个 epoch 之后，loss 趋于平稳，完成训练。这是因为 dropout 随机地将网络中部分神经元设置为 0，使得在训练中的每个神经元都有一定的概率被忽略，从而迫使网络学习到更加健壮的特征。这种随机性减少了模型对任何一个特定的输入的过度依赖，提高了模型的泛化能力。

模型参数量如图 8 所示。本文模型参数量少于原 Swin-B 模型参数量，这是因为本文模型只保留了原模型中的所有 Transfomer 层，舍弃了后续的检测头卷积层、池化层和全连接层等，此时 Transformer 层模型参数量大约为 81.8×10^6。本文加入的 Deconv 和 Head 层模型参数量大约为 3.9×10^6 和 0.6×10^6。因此本文模型总参数量约为 85.4×10^6。

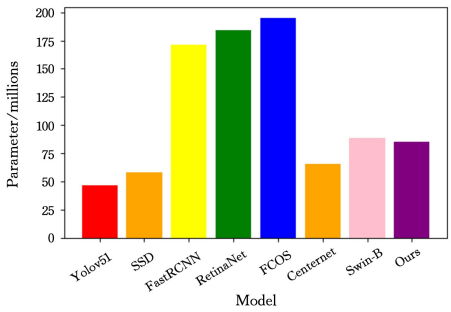


图 8 模型参数图

Fig. 8 Model parameter diagram

4.3 评价指标

本实验采用的评估指标为平均精度均值(mAP)。平均精度均值被定义为数据集中所有类别的平均精度(AP)的均值。平均精度根据 R-P 曲线面积确定，精确率 P 代表模型预测所有目标中，预测正确目标的占比；召回率 R 代表所有真实目标中，预测正确目标的比例。精确度 P 和召回率 R 的计算式如下：

$$P=\frac{TP}{TP+FP} \tag{5}$$

$$R=\frac{TP}{TP+FN} \tag{6}$$

其中，TP 指真实样本且预测为真实样本，FP 指错误样本而预测为真实样本，FN 指真实样本而预测为错误样本。某一种类别的平均精度 AP 的计算式如下：

$$AP=\int_0^1 P(R)\mathrm{d}R \tag{7}$$

mAP 是所有类别的 AP 之和，其计算式如下：

$$mAP=\frac{\sum_{i=1}^{n}AP_i}{n} \tag{8}$$

COCO 数据集在目标检测的评价指标上更为严格，COCO 数据集的 mAP 与常规目标检测固定 IOU 阈值 0.5 不同，它是在 0.5～0.95 之间以 0.05 为步长不断增大，最后将所有 mAP 值求均值的结果。

FPS 是判断目标检测网络检测速度的一个重要指标，它表示在单位时间中，模型能够检测多少帧。其计算式如式(9)所示。

$$FPS=\frac{1}{T_{\mathrm{end}}-T_{\mathrm{start}}} \tag{9}$$

其中，T_{start} 表示检测开始时间，T_{end} 表示检测结束时间。

5 实验结果分析

5.1 Pascal VOC 数据集验证

为了验证本文算法的性能，在 Pascal VOC 数据集上，将所提出的改进算法与原始算法 Swin Transformer 进行对比，结果如图 9 所示。

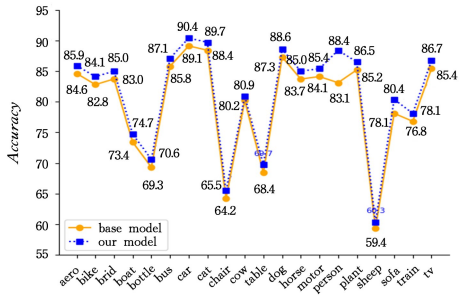


图 9 Pascal VOC 算法对比图

Fig. 9 Comparsion diagram of Pascal VOC algorithms

从图 9 中可以看出，原始算法 base model 的 mAP 为 79.7%，本文算法的 mAP 为 81.1%，提升了 1.4%。且本文算法在类别分布多或少的情况下，其检测精度都高于原算法。这是因为本文算法中的 Swin Transforme 通过将输入图像分块，并对每个块进行 Transformer 计算，从而对图像的全局特征进行建模。同时通过独特的滑动窗口机制，使得低分辨率的特征图可以访问高分辨率的特征图信息，进一步增强了全局特征提取能力。此外，本文算法中还加入了拥有局部特征提取能力的 Deconv 反卷积模块，该模块可以通过卷积层的输出向上采样，从而得到高分辨率的特征图，高分辨率的特征图可以提供更丰富的局部细节信息，从而增强局部特征提取能力。通过这种方式，本文算法可以更好地提取全局和局部特征，从而获得更好的检测效果。此外，本文算法的中心点的特征表达方式可以为模型提供物体位置信息，从而提高模型的识别和定位能力，增强模型的特征表达能力。因此，本文算法在该数据集检测任务上取得了更加优秀的表现。

5.2 MS coco 数据集验证

为了验证本文算法性能，在 MS COCO 2017 数据集上，将所提出的改进算法与原始算法 Swin Transformer 进行对比，结果如表 4 所列。从表 4 中可以看出，本文算法在 MS COCO 2017 数据集上的检测精度达到了 37.2%，超过了基模型 Swin Transformer 网络。各项 mAP 值分别为 37.2%，56.1%和 40.2%，相较于基模型分别提高了 5.2%，6.8%，6.3%，同时，其对于小、中、大目标物体的检测精度分别为 17.4%，41.2%，57.2%，相较于基模型也分别提升了 6.3%，8.4%，5.6%。且检测速度 FPS 也提升了 5.1%。这表明

本文方法在不同尺度目标的检测上，检测精度和速度均有一定提升，这得益于本文算法拥有全局特征和局部提取能力，能够取得更高的目标检查精度。此外，本文模型在检测速度FPS上也有一定提升，这是因为本文方法不再采用全连接层直接把特征展开，而是利用自适应二维高斯核和Head回归头求出目标中心点。回归头的参数量较小，计算量较少，相对于全连接层可以更快地进行前向传播和反向传播。此外，回归头使用卷积层进行特征提取，这些层可以共享参数，减少计算量和参数量，进一步提高了模型的速度。

表4 在MS COCO 2017数据集上与基算法的对比结果

Table 4 Comparison results with base algorithm on MS COCO 2017 dataset

(%)

Model	*FPS*	*mAP*	*mAP*50	*mAP*75	*mAPS*	*mAPM*	*mAPL*
基模型	21.2	32.0	49.3	33.9	11.8	32.8	51.6
Ours	26.3	37.2	56.1	40.2	17.4	41.2	57.2

5.3 对比实验

将本文提出的改进算法和YOLO系列、SSD[16]、RetinaNet[17]等算法在MS COCO数据集上以相同训练参数进行实验对比，对比算法及其结果如表5所列。从表中可以看出，本文改进算法相较于其他目标检测算法，检测精度更高。相较于单阶段算法的基于anchor free系列的FCOS，本文方法的检测速度更快，且精度也有所提升；相较于基于anchor base的YOLO系列算法，本文方法虽然检测速度稍慢，但是精度提升明显；而相较于SSD，RetinaNet等单阶段基于anchor base的算法，本文算法在速度和精度上都有提升。

表5 MS COCO 2017数据集上各算法对比结果

Table 5 Comparison results of each algorithm on MS COCO 2017 dataset

Dector	Backbone	FPS/%	mAP/%
YOLOV4	DarkNet	28.7	30.2
YOLOV5	CSPDarkNet	30.2	33.4
SSD	ResNet101	19.8	20.4
FastRCNN	ResNet101	12.4	35.4
RetinaNet	ResNet101	15.2	35.4
FCOS	ResNet101	17.6	34.3
CenterNet	ResNet101	21.3	26.1
Ours	Swin Transformer	26.3	37.2

5.4 消融实验

同时，为了进一步验证本文算法各个模块的有效性，本文对Deconv模块和Head结合自适应二维高斯核模块在MS COCO数据集上进行消融实验，消融实验总共分为4组进行，结果如表6所列。

表6 消融实验

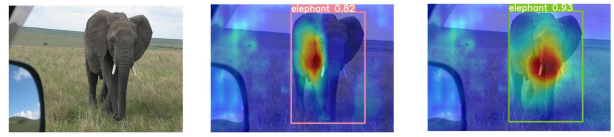
Table 6 Ablation experiment

(%)

Deconv	Head+高斯核	*FPS*/%	*mAP*/%
×	×	21.2	32.0
√	×	19.5	33.4
×	√	23.4	35.1
√	√	26.3	37.2

从表6中可以看出，在只加入了Deconv反卷积模块之后，网络局部特征提取能力所有增强，但在后续网络输出部分仍然存在大量时间开销，这是因为增加了网络模型深度，因此mAP有所提升，但FPS下降。在加入了Head和自适应二维高斯核之后，相较于原算法Swin Transformer直接使用全连接层进行分类和回归检测目标，本文方法不仅增强了模型的特征表达能力，还提高了检测速度。综合上面4组消融实验，验证了本文所提改进算法的有效性。

为了更加直观地验证Swin Transformer在加入Deconv之后特征提取能力的有效性，对输入图片分别进行注意力热图检测，检测结果如图10所示，图10(a)为原始输入图片，图10(b)为没有加入Deconv的特征热力图，图10(c)为加入了Deconv的特征热力图。从图中可以看出，图10(c)的热力图更集中在目标中心处且置信度更高。这是因为Deconv模块可以将底层的特征图进行上采样，使其具有更高的分辨率，从而更好地捕捉目标细节信息，确定目标位置。

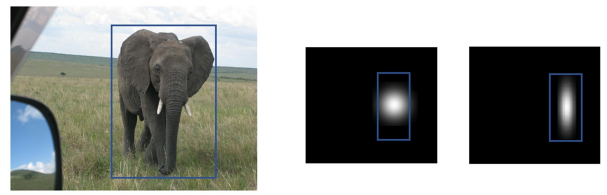


(a)原图 (b)未加入Deconv热力图 (c)加入Deconv热力图

图10 Attention heatmap图

Fig. 10 Attention heatmap

在图11中，图11(a)为原始输入图像，图11(b)和图11(c)分别为使用一维高斯核和自适应二维高斯核进行目标检测，得到的各自的高斯分布热力图。相比一维高斯核得到的热力图，二维自适应高斯核能够更好地适应目标在不同尺度和方向上的特征，因此改进后的椭圆热力图被边界框完全包含，其椭圆形结构更符合当前目标图像的结构。



(a)原图 (b)一维高斯 (c)自适应二维高斯

图11 Gaussian heatmap图

Fig. 11 Gaussian heatmap

此外，为了更加直观地体现本文改进算法在目标检测精度方面的进步，通过本文算法训练好的模型对远距离小目标、暗光环境、复杂环境多目标3种情况进行了测试，检测结果如图12所示。在图12中，第一组为远距离情况下检测的结果，可以看出，对于近距离的人和船，原算法和本文算法都可以识别，但是原算法对远处的船出现了漏检。第二组是暗光环境，本文算法相较于原算法能够更加准确地识别到类似滑板等目标。第三组为在复杂环境多目标情况下，本文算法能够更加准确地识别到类似小刀以及重叠的人等目标，而原算法却出现了大量的漏检。相比之下，本文提出的算法在远距离小目标和复杂环境下多目标的检测中具有更好的表现。在图12的3组测试图片中，对于目标像素特征占比较小的第一组和目标与背景像素特征相似度较大的第二组，以及目标重叠度

较大的第三组，因为本文考虑了全局和局部特征的同时提取，所以模型能够更有效地提取目标特征信息。此外，本文模型的良好表现还得益于中心点的检测方法增强了模型的特征表达能力，能够对重叠目标进行更好的区分。

(a)原图 (b)原算法 (c)本文算法

图 12 目标检测结果对比图

Fig. 12 Comparison of target detection results

结束语 本文针对 Swin Transformer 算法在提取局部特征信息能力上的不足，特征表达能力弱导致目标检测性能较差，以及网络模型过深导致过拟合等问题，提出了一种基于改进 Swin Transformer 的目标检测算法。在 Swin Transformer 的网络中，加入了 Deconv 部分提取局部特征，以及利用回归头和高斯核得到目标中心点，增强模型的特征表达能力，通过 dropout 缓解网络过拟合。在 Pascal VOC 和 MS COCO 2017 数据集上的实验结果表明，与其他先进目标检测算法相比，本文算法检测性能得到了显著提高。但目前本文算法在检测速度上仍然存在不足，因此后续工作将会基于此继续完善。

参 考 文 献

[1] CHEN K Q, ZHU Z L, DENG X M, et al. Deep learning for Multi-Scale Object Detection: A survery[J]. Journal of Software, 2021, 32(4): 1201-1227.

[2] BAO S M, WANG S Q. Overview of Object Detection Algorithms Based on Deep Learning [J]. Transducer and Microsystem Technologies, 2022, 41(4): 5-9.

[3] HAN C, GAO G, ZHANG Y. Real time small traffic sign detection with revised faster-RCNN[J]. Multimedia Tools and Applications, 2018, 7(10): 13263-13278.

[4] REDMON J, FARHADI A. YOLOv3: An inc-remental improvement[J]. arXiv: 1804. 02767, 2018.

[5] LI X J, DENG Y M, CHENG Z H, et al. Improved YOLOv5 algorithm for airport runway foreign object detection [J]. Computer Engineering and Applications, 2023, 59(2): 202-211.

[6] TIAN Z, SHEN C H, CHEN H, et al. FCOS fully convolutional one-stage object detection[C]// Proceedings of IEEE/CVF International Conference on Computer Vision. Washington USA, 2019: 9626-9635.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need [J]. arXiv: 1706. 03762, 2017.

[8] DOSOVITSKIY A, BEYER L, KOESNIKOV A, et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale[C]// International Conference on Learning Representations. Online: ICLR, 2021: 3-7.

[9] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE International Conference on Computer Vision. Montreal. Canada, 2021: 11-18.

[10] FU C Y, LIU W, RANGA A, et al. DSSD: Deconvolutional Single Shot Detector[J]. arXiv: 1701. 06659, 2017.

[11] ZHOU Y, LIU Y, LU J, et al. DIT: A Deformation Invariant Transformer Network for Unsupervised Keypoint Discovery and Description[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 12630-12639.

[12] ZHOU X, WANG D, KRÄHENBÜL P. Objects as Points[J]. arXiv: 1904. 07850, 2019.

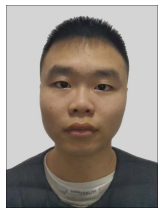
[13] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv: 1207. 0580, 2012.

[14] WANG C, LIU Y J, XIE Q, et al. Anchor Free object detection algorithm based on soft labeland sample weight optimization [J]. Computer Science, 2022, 49(8): 157-164.

[15] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]// Proceedings of Conference on Computer Vision. Berlin, Germany, 2014: 740-755.

[16] LIU W, ANGUELOV D, ERHAND, et al. SSD: Single shot multibox detector[C]// Computer Vision－SCCV 2016. Amsterdam, 2016: 21-37.

[17] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal Loss for Dense Object Detection[C]// Proceedings of Conference on Computer Vision. Venice, 2017: 2980-2988.



LIU Jiasen, born in 1999, postgraduate. His main research interests include object detection and deep learning.

HUANG Jun, born in 1971, Ph.D, professor, master supervisor. His main research interests include object detection and deep learning.

(责任编辑：何杨)