计算机科学
COMPUTER SCIENCE

PSwin:基于Swin Transformer的边缘检测算法

胡名扬, 郭燕, 金杨爽

引用本文

胡名扬, 郭燕, 金杨爽.PSwin:基于Swin Transformer的边缘检测算法[J]. 计算机科学, 2023, 50(6): 194-199.

HU Mingyang, GUO Yan, JIN Yangshuang. PSwin:Edge Detection Algorithm Based on Swin Transformer [J]. Computer Science, 2023, 50(6): 194-199.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700145

PSwin:基于 Swin Transformer 的边缘检测算法

胡名扬[1,2]　郭　燕[1,2]　金杨爽[2]
1 中国科学技术大学苏州高等研究院　江苏 苏州 215123
2 中国科学技术大学软件学院　江苏 苏州 215123
(myoung@mail.ustc.edu.cn)

摘　要　边缘检测作为一种传统的计算机视觉算法,已经被广泛应用于车牌识别、光学字符识别等现实场景。当边缘检测作为更高层级算法的基础时,比如目标检测、语义分割等算法,又可以应用于城市安防、自动驾驶等领域。好的边缘检测算法能够有效提升上述计算机视觉任务的效率和准确度。边缘提取任务的难点在于目标的大小以及边缘细节的差异性,因此边缘提取算法需能够有效处理不同尺度的边缘。PSwin 首次将 Transformer 应用于边缘提取任务,并提出了一种新型特征金字塔网络,以充分利用骨干网络多尺度和多层次的特征。PSwin 使用自注意力机制,相比卷积神经网络架构,可以更有效地提取图像中的全局结构信息。在 BSDS500 数据集上进行评估时,PSwin 边缘检测算法达到了最佳水平,ODS F-measure 为 0.826,OIS 为 0.841。

关键词:边缘检测;特征金字塔;视觉注意力;迁移学习;BSDS500
中图法分类号　TP391

PSwin:Edge Detection Algorithm Based on Swin Transformer

HU Mingyang[1,2],GUO Yan[1,2] and JIN Yangshuang[2]
1 Suzhou Institute for Advanced Research,University of Science and Technology of China,Suzhou,Jiangsu 215123,China
2 School of Software Engineering,University of Science and Technology of China,Suzhou,Jiangsu 215123,China

Abstract　As a traditional computer vision algorithm,edge detection has been widely used in real-world scenarios such as license plate recognition and optical character recognition. When edge detection is used as the basis for higher-level algorithms,such as target detection,semantic segmentation and other algorithms. Edge detection can also be applied to urban security,autonomous driving and other fields. A good edge detection algorithm can effectively improve the efficiency and accuracy of the above computer vision tasks. The difficulty of the edge extraction task lies in the size of the target and the difference of edge details,so the edge extraction algorithm needs to be able to effectively deal with edges of different scales. In this paper,the Transformer is applied to the edge extraction task for the first time,and a novel feature pyramid network is proposed to make full use of the multi-scale and multi-level features of the backbone network. PSwin uses a self-attention mechanism,which can extract global structural information in images more efficiently than convolutional neural network architectures. When evaluated on the BSDS500 dataset,the proposed PSwin edge detection algorithm achieves the best performance,with an ODS F-measure of 0.826 and an OIS of 0.841.

Keywords　Edge detection,Feature pyramid network,Visual attention,Transfer learning,BSDS500

1　引言

边缘检测,即从自然图像中提取具有视觉意义的边界和边缘,其在图像分割、场景识别、物体检测等计算机视觉的许多领域中发挥着重要作用。四十年来,随着深度卷积神经网络的发展,边缘检测精度有了显著提高,一些方法在标准数据集如 BSDS500[1] 上具有突出的表现。然而,如何提取图像中的多尺度边缘,依旧是一个非常有挑战性的问题。由于自然图像中不同的物体形态各异,即使是类型相同的物体,它们的大小也可能不同,边缘特征也会发生改变。因此,准确识别不同尺度的边缘对于边缘检测算法至关重要。

长期以来,计算机视觉算法都是以卷积神经网络(CNN[2])为主导,文献[3-5]中的骨干网络都基于 ResNet[6]。卷积神经网络能够理解的上下文范围取决于网络自身的感受野小,然而根据文献[7]所述,卷积操作难以提取全局信息,导致卷积神经网络的实际感受野比理论上小得多。例如,卷积操作很难判断一张图片中的某一部分是否发生了错位,因为卷积层专注于提取局部特征。解决此问题有两种常见的方法,第一种直观的解决方案是扩大网络的感受野,即增加卷积核的大小,或者进行多次池化操作,但这会导致局部特征,以及其他有用信息如相对位置信息的丢失;另一种方法是增加网络的深度,但这样会显著增加训练的难度。因此,如何提高

到稿日期:2022-07-15　返修日期:2022-10-27
通信作者:郭燕(guoyan@ustc.edu.cn)

神经网络对全局信息的提取能力是非常重要的工作。

近两年，Transformer[8] 在图像处理中的应用越来越多，效果也不断提升。Transformer 的优点是在自注意力机制的帮助下，可以有效提取长距离的特征依赖关系。基于 Transformer 的主干网络[9-10] 通过将图像切分为不同的小块，在多种视觉任务如图像分类、目标识别方面都取得了 SOTA 的效果。其中，Swin Transformer[11] 提出了一个通用的视觉骨干网络，在物体检测和语义分割等任务中大幅刷新了此前的记录，被广泛应用于众多视觉任务中。

特征金字塔网络（Feature Pyramid Networks，FPN[12]）是一种层级结构，在边缘提取任务中，可以帮助骨干网络处理输入图片的多尺度变化问题。特征金字塔能够将神经网络顶层中语义上较强的特征与网络底层特征相整合。相比传统的图像金字塔方式，特征金字塔具有计算量少的优点。

基于以上分析，本文将 Swin 作为边缘检测任务的主干网络，并设计了适合视觉 Transformer 的特征金字塔结构。这种设计一方面能够充分利用骨干网络不同层级的输出；另一方面，加入了金字塔池化模块，实现了对高层语义信息的有效利用。通过将两种方法相结合，PSwin 在标准数据集中的 ODS F-measure 和 QIS 分别达到了 0.826和 0.841。

本文的贡献包括：1）首次基于视觉 Transformer 构建了一个端到端的边缘检测系统，证明了视觉 Transformer[13] 网络可以有效应用于边缘检测任务，超越了一系列基于 ResNet[14] 骨干网络算法的效果；2）提出了新型的特征金字塔结构，有效地利用了骨干网络提取的多尺度和多层次的特征，可以生成高质量且细致的对象边缘。

2 相关工作

本节主要从边缘检测算法、特征金字塔以及视觉 Transformer 网络这 3 方面介绍相关工作。

边缘检测已有 40 年左右的发展历史，边缘检测算法大致可以分为 3 种：1）传统的基于梯度计算的边缘算子，其原理是通过检测颜色、纹理等的突变来确定图像边缘；2）基于人工设计的模型来检测边缘的特征；3）基于卷积神经网络的方法，可以端到端自动识别图像边缘。基于神经网络的算法显著提高了边缘检测算法的性能，甚至优于人类对于边缘的感知，因此本节重点介绍基于深度学习的算法。

DeepEdge[15] 首先使用 Canny[16] 检测器在可能的边缘候选点周围提取多个图片块，然后将这些图片块输入到 CNN，最后使用分类网络来确定这些图片块是否是边缘。DeepContour[17] 也基于图片块，首先将图像分成多个块，然后使用 CNN 进行分类。N4-Fields[18] 结合了 CNN 和最近邻搜索，首先使用 CNN 提取图像块，然后在预设的字典中搜索与当前块相似的特征，最后融合块生成输出。HED[4] 是第一个端到端的边缘检测方法，该方法使用预训练的 VGG16[15] 作为主干网络，并将 VGG16 的每个阶段输出的特征连接到最后的卷积层，因此能够自动学习图像的层次表示。RCF[3] 尝试同时使用 VGG 和 ResNet 作为主干网络，并使用 5 个卷积层输出，最后将输出的所有特征进行逐像素预测，以获得更准确的图像表示。BDCN[19] 提出了一种双向模型来融合多尺度信息，它与 HED 和 RCF 的区别在于，BDCN 的网络会使用特定尺度的标记边缘，每一层的输出都会进行监督，避免了使用相同大小的标签。因此，网络输出的多尺度特征更为准确，整体性能得到了实质性的提升。

除了上述研究，我们的工作还受到 FPN，EfficientDet[17] 和 PSPNet[20] 的启发。FPN 开发了一种具有横向连接的自上而下的架构，使高层语义特征可以传播到底层，因此它可以有效地融合不同尺度的高级语义特征。EfficientDet 提出了一个加权双向特征金字塔网络（BiFPN），通过双向的方式，使网络达到了更高的准确性和效率。PSPNet 提出了空间金字塔池化模块，金字塔池化可以提取并融合不同尺度的特征，从而提高网络获取全局信息的能力。

相比图像处理，自然语言处理近几年取得了巨大的进展，主要得益于对模型容量的大幅扩展以及 Transformer 架构的使用。视觉 Transformer 的出现为 Transformer 如何处理图像提供了思路。同时，多项工作[21] 尝试将 CNN 与自注意力相结合。DeiT[22] 提出使用蒸馏图像块的输入特征，将基于卷积神经网络提取的特征转移到视觉 Transformer；T2TViT[23] 提出使用图像块的输入特征模块，将图像递归地重组切分，处理时考虑相邻像素的特征；DETR[24] 方法将 ResNet 抽取的特征再次输入 Transformer 中，使用 CNN 处理过的特征来帮助 Transformer 对全局关系进行建模。而 ViT 第一个成功地将 Transformer 应用到图像分类任务中。Swin Transformer 在 ViT 的基础上构建，借鉴卷积神经网络，采用分层的设计，整体包括 4 个阶段，每个阶段都同卷积神经网络一样逐层增加感受野。因此，Swin 可以充分利用 Transformer 的长距离依赖性，以及 CNN 归纳偏置的能力，如层次性、局部性和平移不变性，在物体检测和语义分割任务中大幅刷新了纪录。同时，Swin 设计了移位的不重叠窗口，大幅降低计算复杂度，使得计算复杂度随着输入图像的增大呈线性增长。

综上，将自注意力机制引入视觉领域可以获得性能上的较大提升，而 Swin 一方面具有超越其他视觉 Transformer 的效果，另一方面具有类似 CNN 的层级结构，可以较为方便地结合特征金字塔提取准确的边缘。因此本文将 Swin Transformer 作为骨干网络应用在边缘检测任务中。

3 网络架构

本节将具体介绍基于 Swin 的边缘检测算法，包括：1）视觉 Transformer 应用于图像边缘检测的基本原理，以及利用 Swin Transformer 进行边缘检测算法的整体网络结构；2）详细介绍了特征金字塔，包括特征金字塔模块和池化金字塔模块。

3.1 基于特征金字塔的网络结构

如图 1 所示，算法第一步使用 Swin Transformer 作为骨干网络，提取输入图像的基本特征。

Swin Transformer 网络首先将输入图片切分成不同的小块，利用注意力机制学习不同图片块之间的特征关系。Swin 网络在不同的网络层级中对特征维度进行了压缩，这种压缩操作能够扩大骨干网络的整体感受野。同时，维度压缩的过程也是去除背景噪声的过程。在去噪的过程中，骨干网络逐渐提取

出图片的高层语义信息，即图片的大致边缘轮廓。本文通过新型的特征金字塔结构，充分利用了 Swin 骨干网络中包含的输入边缘细节信息和高层语义信息，最终得到准确的边缘。

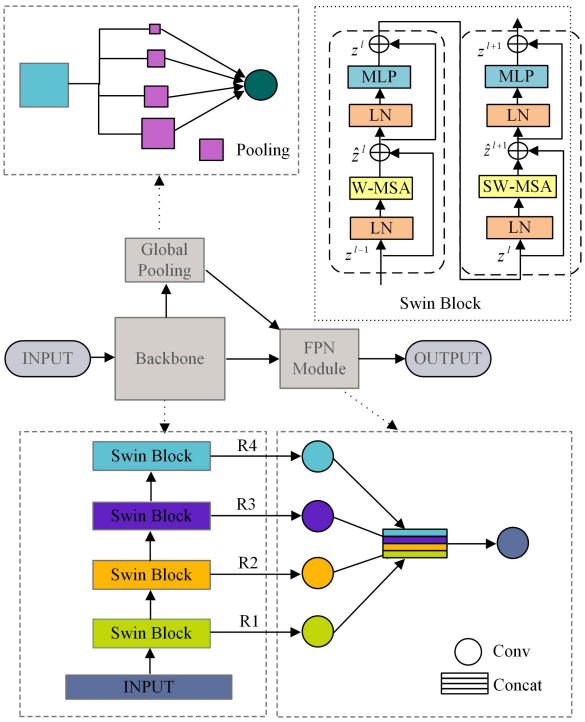

图 1 基于 Swin 的边缘提取网络结构

Fig. 1 Edge extraction network architecture based on Swin

具体来说，网络首先将图像输入到 Swin 骨干网络。输入图像的大小为 $H\times W\times 3$，其中 H 和 W 分别是图像的高和宽，3 表示图像的 3 个通道。在骨干网络中首先将输入图像分割成不重叠的切块，每个切块大小为 $4\times 4\times 3$，因此，一张图片的切块总数是 $\frac{H}{4}\times\frac{W}{4}$。Swin 的第一个阶段是将图片切块的特征维度进行线性嵌入，之后每个阶段都会对图片的相邻切块进行合并，并对特征维度进行压缩，以减少计算量。因此从第二阶段开始，每个阶段的输出维度会缩小一半。同时，为了降低计算复杂度，Swin 使用了窗口注意力机制，即将图片分成多个窗口，每个窗口包含若干个切块，只计算窗口内部的注意力。但这样会失去获取全局特征的能力。

为了获得全局特征，Swin 使用了 Shift-Window 方法来对图片中的窗口进行重新划分，这种设计可以让网络在不增加复杂度的同时，扩大网络的感受野。在 Swin 结构中，每个阶段都包括偶数个 Transformer 块，相邻的两个块分别对输入图片进行规则的窗口划分和 Shift-Window 划分，分别对应于 W-MSA 和 SW-MSA 操作，用这种方式在不增加计算量的同时，增大网络的感受野。MSA 是多头自注意力机制，W-MSA 是窗口多头自注意力机制，SW-MSA 是在 W-MSA 的基础上进行了移轴操作。相邻的两个 Transformer 块的计算如下所示：

$$\hat{z}^{l}=W\text{-}MSA(LN(z^{l-1}))+z^{l-1} \tag{1}$$

$$z^{l}=MLP(LN(\hat{z}^{l}))+\hat{z}^{l} \tag{2}$$

$$\hat{z}^{l+1}=SW\text{-}MSA(LN(z^{l}))+z^{l} \tag{3}$$

$$z^{l+1}=MLP(LN(\hat{z}^{l+1}))+\hat{z}^{l+1} \tag{4}$$

其中，$\hat{z}^{l}$ 和 z^{l} 分别代表第 l 个 Transformer 块的 W-MSA 的输出和 LayerNorm 的输出。

PSwin 从下向上的 4 个阶段的输出维度分别为[1 024，512，256，128]，将这 4 个特征分别输出到特征金字塔，以将不同分辨率、不同语义信息的特征融合。这种融合的方式旨在有效地捕获并处理骨干网络提取的分层特征。

本文设计的特征金字塔，首先对每个阶段的输出进行卷积操作，主要目的是将维度降维成 64。之后通过采样操作，将不同阶段输出的特征大小采样为 256。最后通过连接操作，将不同层级输出的特征拼接起来，并通过卷积操作降维成 1。

相比 RCF 中使用特征金字塔将 ResNet 每层提取出的信息进行两次降维操作，本文方法更高效且易于训练，同时节省了计算时间。如实验部分所示，最终的融合层能够更有效地利用网络提取出来的特征，得到更好的输出结果。

3.2 基于特征池化金字塔的网络结构

相比卷积神经网络，基于 Swin Transformer 骨干网络的边缘检测算法在全局特征的获取能力上有了较大的提升。但是 Swin 为了降低计算复杂度，仅在局部窗口内计算自注意力，Shift-Window 操作也只是在一定程度上增大了网络感受野，并没有真正地对整张图像进行注意力计算。因此我们使用池化金字塔来辅助骨干网络提取全局特征。Swin 的第四阶段特征中包括较多的全局信息，因此针对第四阶段的输出，我们设计了金字塔池化模块，网络结构如图 1(Global Pooling)所示。

相比以往的边缘检测网络，本文的网络结构增加了对 Swin 阶段四的输出的进一步处理，特征池化金字塔能够有效减少输出特征中的无关信息。如图 1 所示，池化模块共分为 4 层，每层分别使用不同大小的池化核[1，3，5，7] 对骨干网络最后一层的输出特征进行池化，不同层次分别对应不同尺度的特征映射。

然后使用双线性插值算法，对池化金字塔不同层次输出的结果进行上采样，得到和输入大小相同的特征图。将所有层次的输出与原始输入相连接，作为最终的金字塔池化模块的输出。最后，将该全局特征与特征金字塔的结果相连，作为图片的最终特征。整体结构如图 2 所示。

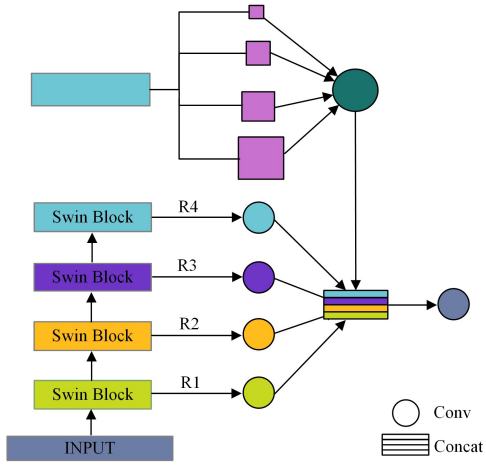

图 2 PSwin 具体网络架构

Fig. 2 PSwin network architecture

3.3 网络参数

本文的骨干网络使用的是 Swin-B，窗口大小设置为 $M=7$。对于所有实验，每个 Head 的查询维度 $d=32$，每个 MLP 的扩展层 为 4。C 是第一阶段隐藏层的通道数。4 个阶段的 Transformer Block 的层数分别是 2，2，18，2。因此，每一层次的输出维度分别是 128，256，512，1024 。

4 实验设置

本节首先介绍实验所使用的公开数据集，然后分别介绍算法实现细节、评估指标、消融实验等。

4.1 数据集

BSDS500，即 Berkeley Segmentation Dataset and Benchmark，是边缘检测中广泛使用的数据集。BSDS 500 包含 200 张训练图像、200 张验证图像和 200 张测试图像。BSDS 中的每个图像都由 4～9 位标注人员进行边缘标注；大部分图像的边缘结果高度一致，但也存在一些有争议的像素。训练时真值采用平均值，评测代码中会依次对这 5 个真值进行对比。

原始 BSDS500 数据集数据量过小，不利于网络的训练，因此本文使用以下几种机制对该数据集进行扩充。首先，对数据集中的图像进行多尺度处理，即调整图像大小以构建图像金字塔，并将每一个图像输入到网络中。在网络处理之后，所有生成的边缘概率图都使用双线性插值，把图片调整为原始图像大小，以便对它们进行平均以获得最终的边缘图。其次，训练数据集还包括翻转的 PASCAL VOC 上下文数据集，它被用作 BSDS500 的增强数据。PASCAL VOC 是一个常用于图像分割任务的数据集，也用于边缘检测任务的训练和测试，如 RCF。再次，通过将图像旋转到 16 个不同的角度并裁剪最大的矩形来增强数据。此外，图像会以不同角度进行翻转，因此，生成的数据集是原始数据集大小的 32 倍。

4.2 实验说明

本文代码基于 Pytorch，并建立在公开可用的 Swin Transformer 之上。Swin Transformer 采用 Swin-Base 默认预训练参数，本文的采用 AdamW 优化器训练网络，mini-batch 在每次迭代中随机对 5 个图像进行采样。全局学习率为 $1\times10^{-6}/4$，使用线性衰减对学习率进行动态设置。每个图像的输入大小为 1024×1024。

在结果上本文考虑了 3 个被广泛用于边缘检测领域的评估指标：固定轮廓阈值（Optimal Dataset Scale，ODS）、每幅图像最佳阈值（Optimal Image Scale，OIS），以及平均准确率（Average Precision，AP）。其中，ODS 即选取一个固定的阈值应用于所有图片，使得整个数据集上的 F 值最大；OIS，也被称为单图最佳，在每一张图片上均选取不同阈值使得该图 F 值最大；AP 指平均准确率，由于模型输出的结果是 0～1 之间的一个值，而某个像素是否为边缘标记为 0 或 1，由多个标记者分别进行标记，因此在不同的阈值情况下，网络输出结果的精确度（网络预测为边缘的像素数量之和/标签中至少有一个像素标记为边缘的像素数量之和）与召回度（网络预测为边缘的位置中，标签被标记为边缘的数量之和/标签中总的标记为边缘的个数）是不同的。由此可以绘制一条曲线，曲线积分的结果即为算法输出结果的 AP。

F-measure 值的计算方式如下：

$$F=\frac{2*Precision*Recall}{Precision+Recall} \tag{5}$$

为了处理数据集中有争议的像素，本文仿照 RCF 对其进行处理。对于所有的边缘图片，平均 ground-truth 上的每个像素点以生成概率图。借助超参数 $\eta(0<\eta<1)$，如果某个像素的概率值小于 η，那么它就被认为是有争议的像素，并且该像素的损失应取 0。对于典型的图像，负（非边缘）和正（边缘）像素的比例高度失衡，90% 的 ground-truth 是负的。因此，我们使用加权交叉熵来平衡这两个类别，每个像素的损失函数如式(6)所示：

$$L(X_i;W)=\begin{cases}a*\log(1-P(X_i;W)), & \text{if } y_i=0\\ 0, & \text{if } 0<y_i<n\\ b*\log(P(X_i;W)), & \text{otherwise}\end{cases} \tag{6}$$

$$a=\lambda*|Y_+|/(|Y_+|+|Y_-|) \tag{7}$$

$$b=|Y_-|/(|Y_+|+|Y_-|) \tag{8}$$

$$Y_+=y_j,y_j>\eta,Y_-=y_j,y_j=0 \tag{9}$$

其中，超参数 λ 用来更好地平衡正负像素；y_i 是一个像素的真实标签，X_i 是网络预测值的标签；$P(X)$ 是标准的 Sigmoid 函数；W 表示网络中所有参数的集合。网络最终的损失函数为图片中每个像素的损失和，计算式为：

$$L(W)=\sum_{i=1}^{|n|}L(X_i^{\text{fuse}};W) \tag{10}$$

4.3 与现有算法的性能对比

本文在 BSDS500 上将 PSwin 与 Canny，gPb，SE，DeepContour，DeepEdge，HED，RCF，Dexined 等各种边缘检测算法进行性能比较。

为了评估，将标准的非最大抑制（NMS）应用于检测到的精细边缘后，结果如表 1 所列。

表 1 BSDS500 数据集上与各种边缘检测算法的比较

Table 1 Comparison with various edge detection algorithms on BSDS500 dataset

Method	*ODS*	*OIS*	*AP*
Canny	0.611	0.676	0.751
gPb	0.729	0.755	0.892
SE	0.743	0.763	0.925
DeepContour	0.757	0.776	0.916
DeepEdge	0.753	0.772	0.915
HFL	0.767	0.788	0.892
HED	0.788	0.808	0.923
CEDN	0.788	0.804	—
MIL+G-DSN+MS+NCuts	0.813	0.831	—
RCF	0.811	0.83	0.913
RCF-ResNet50	0.814	0.833	0.910
RCF-ResNet101	0.819	0.836	0.909
CED	0.794	0.811	0.871
BDCN	0.820	0.838	0.888
Dexined	0.729	0.745	0.583
PSwin	0.826	0.841	0.908

从表 1 可以看出，PSwin 边缘检测算法在 ODS 和 OIS 上都达到了最佳水平，分别达到了 0.826 和 0.841。由于 PSwin 可以在抑制噪声的情况下输出准确的边缘，在高准确度的情况下降低了召回度，因此，从整体来看最终影响了 AP 值的结果。但通过 ODS 和 OIS 或者图 3 均可以证明 PSwin 输出了最准确的边缘。

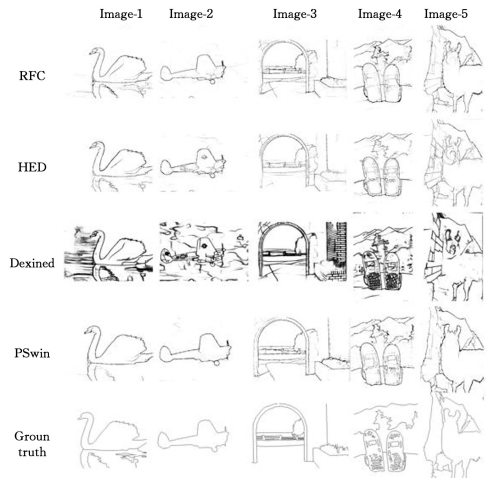

图 3 与其他算法输出结果对比

Fig. 3 Algorithm output results comparison

图 3 给出了典型边缘检测算法输出的边缘结果，图中比较了目前比较强大的基于深度学习的边缘算法如 RCF，HED，Dexined 和本文算法 PSwin 在 BSDS500 上的最终输出。HED 是第一个端到端的边缘检测算法；RCF 对 HED 做了改进，能够更好地利用骨干网络的特征；Dexined 能够得到准确且薄的边缘。因此本文选择将本文算法与这些算法进行对比。从图 3 可以看出，整体而言，PSwin 输出的背景噪声更少，边缘更薄。相比其他算法，PSwin 网络通过池化金字塔得到了更为准确的高层物体级别语义信息，并通过特征金字塔提取了骨干网络中多层级的边缘信息，通过简单且高效的融合方式，将高层的语义信息与低层的边缘信息融合并输出。在准确输出图片主体边缘的同时，保留了背景中的主要边缘；在降低背景噪声的同时输出的整体轮廓更为清晰，并且抑制了局部非边缘的生成，因此得到了更为准确的边缘。

4.4 消融实验

为了充分验证本文提出的模块的有效性，本节设计了消融实验，目的是探究池化金字塔模块对 PSwin 的影响，结果如表 2 所列。从表中的结果可以看出，空间池化金字塔对算法有很大的帮助，无论是 ODS 还是 OIS 均得到了提升。

表 2 消融实验

Table 2 Ablation experiment

Method	*ODS*	*OIS*	*AP*
PSwin without pooling	0.823	0.839	0.858
PSwin	0.826	0.841	0.834

图 4 给出了池化金字塔对算法的影响，可以观察到在池化金字塔的帮助下，背景噪声得到了更有效地抑制，也因此得到了更准确的边缘。

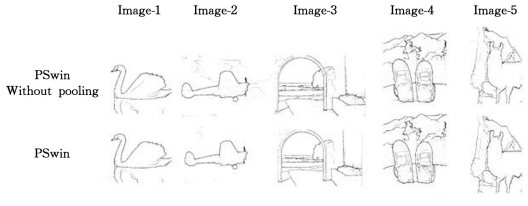

图 4 消融实验输出结果对比

Fig. 4 Ablation experiment results comparison

结束语 卷积神经网络的边缘提取算法由于自身的限制，无法准确获取输入图片的全局特征。本文将视觉注意力网络应用于边缘提取算法，并设计了特征金字塔结构，进一步利用了全局信息。在公开数据集 BSDS 500 上的实验证明，基于 Transformer 的主干网络同样适用于边缘提取任务。PSwin 充分利用了自注意力机制的优点，准确地提取了图像边缘，并抑制了无关噪声的产生。目前的边缘检测算法参数量较为庞大，未来可以尝试使用模型蒸馏等方式，在减少模型参数的同时输出较为准确的边缘。

参 考 文 献

[1] ARBELÁEZ P, MAIRE M, FOWLKES C, et al. Contour Detection and Hierarchical Image Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 898-916.

[2] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[3] LIU Y, CHENG M M, HU X, et al. Richer convolutional features for edge detection[C]// Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. 2017: 5872-5881.

[4] XIE S, TU Z. Holistically-Nested Edge Detection[J]. International Journal of Computer Vision, 2017, 125(1/2/3): 3-18.

[5] BERTASIUS G, SHI J, TORRESANI L. Deepedge: A multi-scale bifurcated deep network for top-down contour detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4380-4389.

[6] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[7] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Object Detectors Emerge in Deep Scene CNNs[J]. arXiv: 1412.6856, 2014.

[8] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.

[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. arXiv: 2010.11929, 2020.

[10] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[11] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

[12] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015: 3431-3440.

[13] CANNY J. A Computational Approach to Edge Detection[J].

IEEE Transactions on Pattern Analysis and Machine Intelligence，1986，PAMI-8(6)：679-698.

[14] TAN M，PANG R，LE Q. EfficientDet：Scalable and Efficient Object Detection[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：10778-10787.

[15] LONG J，SHELHAMER E，DARRELL T. Fully convolutional networks for semantic segmentation[C]/ IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015：3431-3440.

[16] CANNY J. A Computational Approach to Edge Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，1986，PAMI-8(6)：679-698.

[17] TAN M，PANG R，LE Q. EfficientDet：Scalable and Efficient Object Detection[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：10778-10787.

[18] GANIN Y，LEMPITSKY V. N^4-Fields：Neural Network Nearest Neighbor Fields for Image Transforms[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015：536-551.

[19] HE J，ZHANG S，YANG M，et al. Bi-directional cascade network for perceptual edge detection[C] // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019：3823-3832.

[20] ZHAO H，SHI J，QI X，et al. Pyramid Scene Parsing Network [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2017：6230-6239.

[21] PENG Z，HUANG W，GU S，et al. Conformer：Local Features Coupling Global Representations for Visual Recognition[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021.

[22] TOUVRON H，CORD M，DOUZE M，et al. Training data-efficient image transformers &amp；distillation through attention [C]// Proceedings of the 38th International Conference on Machine Learning. 2021：10347-10357.

[23] YUAN L，CHEN Y，WANG T，et al. Tokens-to-Token ViT：Training Vision Transformers from Scratch on ImageNet[C]// ICCV2021. 2021：558-567.

[24] CARION N，MASSA F，SYNNAEVE G，et al. End-to-End Object Detection with Transformers[C]// ECCV 2020. 2020：213-229.

HU Mingyang，born in 1997，master. His main research interests include computer vision and natural language processing.

GUO Yan，born in 1981，lecturer. Her main research interests include information security，blockchain and NLP.

（责任编辑：何杨）