第1章

视频目标检测

1.1 视频目标检测概述

目标检测是一种经典的计算机视觉任务，其以定位并识别视觉图像或视频中的感兴趣的对象为目的，需要在图像中用矩形包围框、关键点位置或者二值掩膜等定位目标的位置，并对目标所属的具体类别进行识别。目标检测与人的视觉感知机理密切关联，是高级场景内容理解应用的基础和前提，可为高阶计算机视觉任务，例如图像描述、行人再识别、姿态识别等，奠定重要的基础，因此受到研究者的广泛关注。进一步，视频目标检测需要在视频中的每帧图像中定位并识别感兴趣目标。相比常见的面向图像的静态目标检测，视频目标检测的特点是对运动信息加以应用。考虑到运动信息是人类视觉感知的重要组成部分，因此视频目标检测的相关研究对特征表征的综合性更强，具备更重要的理论意义。此外，视频目标检测具备重要的实际应用价值，在智能视频监控、智能机器人等多场景中有着重要的应用。

深度学习的发展为计算机视觉领域带来了显著进展。基于卷积神经网络[1]、Transformer[2]等深度网络结构的静态目标检测方法大幅提升了检测准确率。简单而言，基于深度神经网络的目标检测算法摒弃了手工设计特征，从大量的训练数据中自动学习目标的特征，其检测效果也远优于传统的目标检测算法。基于卷积神经网络的静态目标检测方法主要分为单阶段目标检测（如 YOLO[3]、SSD[4]）和双阶段目标检测（如 Faster R-CNN[5]）这两种不同的结构；而基于 Transformer 的 DETR[6]等方法通过集合预测的建模思路避免非极大值抑制的常用后处理步骤。相对于基于图像的智能识别和检测而言，视频目标检测不仅要保证每帧图像检测的准确性，还要保证检测结果具有一致性和连续性，即视频目标检测的时序一致性。因此，如何结合静态目标检测的优势，融合多帧图像中的时序信息和特征表征，实现高精度且高效率的视频目标检测是智能视觉识别的重要课题。

相比静态目标检测，视频目标检测除了受到目标类内差异大、复杂背景中易混

淆等常见挑战外，视频中特有的运动模糊和视频散焦等都给目标检测带来了挑战。具体而言，视频目标检测需检测并识别序列内出现的帧图像中的目标。视频拍摄过程中，目标与成像设备之间存在相对运动，会因运动模糊和视频散焦而导致低成像质量；在视频的某些帧中，目标的成像视角特殊，罕见视角目标的存在会增加检测难度；此外，目标的尺寸变化、遮挡与视角切换，也会增加某些帧的检测难度，如图 1-1 所示。与静态图像相比，视频数据中涵盖丰富的时间信息，有助于理解运动场景的内容。在视频目标检测过程中，面临的主要挑战是如何有效利用时序信息，保持视频中出现目标的时空一致性（时间和空间位置发生变化的目标还是属于同一个目标）。现有大多数视频目标检测算法主要通过时序信息和时序一致性约束增强单帧的静态检测性能，常见思路是利用其他帧的特征增强关键帧的特征与目标预测。在视频目标检测方面，尽管目前做了一些针对性的改进研究，但还有很大的提升空间，特别是在快速移动的目标和运动信息的帧间传输方面还有很大的研究空间。

图 1-1　视频目标检测的难点

除类内差异、视角变化、遮挡和截断等常见的静态目标检测面临的常见挑战外，视频目标检测还受限于视频中的运动模糊和散焦等问题。

（1）运动模糊指的是在视频拍摄过程中，由于目标与成像设备相对位置变化，造成视频中的某些帧画面模糊，导致目标难以分辨。

（2）视频散焦指的是视频内的中心场景发生变化时，焦距不准确或者焦距调整不及时，散焦同样会导致视频帧模糊。

（3）遮挡变形指的是在一段视频中，某个目标可能在某个时间范围内被其他障碍物遮挡，导致信息缺失。

（4）罕见视角指的是由于目标的特殊动作，目标在不同的帧中外观呈现剧烈变化，与静态图像中的目标视角有明显差异。

（5）目标尺度变化指的是运动导致视频中摄像头距离变化，从而导致帧间常见的目标尺度变化。

1.1.1　视频目标检测的数据集

现有视频目标检测的常用数据集是 ImageNet VID 数据集[7]，其包含 30 个基本类别，是 ImageNet DET 目标检测任务数据集所包含的 200 个基本类别的子集。ImageNet VID 数据集主要关注交通工具和动物类别的目标检测，重点考虑目标运动类型、视频内容的复杂程度、平均目标数量等因素，其中动物类别涵盖的相对多一些，具体类别如表 1-1 所示。整个数据集包含训练集、验证集和测试集 3 个部分，其中训练集包含 3862 个视频，验证集和测试集分别包括 555 个和 937 个视频片段。进行实验时，首先使用带真实标签的训练集进行模型训练，然后利用验证集测试模型性能。训练集包含超过 112 万幅图像（视频帧），每个类别平均有约 3.7 万幅图像样本，大规模数据有利于深度网络模型的参数学习，以完成视频目标检测任务。

表 1-1　视频目标检测数据集 ImageNet VID 中的目标类别

交通工具(7 类)	Airplane, Bicycle, Bus, Car, Motorcycle, Train, Watercraft
动物(23 类)	Antelope, Bear, Bird, Cattle, Domestic Cat, Dog, Elephant, Fox, Giant Panda, Hamster, Horse, Lion, Lizard, Monkey, Red Panda, Rabbit, Sheep, Snake, Squirrel, Tiger, Turtle, Whale, Zebra

ImageNet VID 数据集提供了每帧图像共 30 类目标的实例标注。对于每个视频片段，模型需输出形式为$\{f_i, c_i, s_i, b_i\}$的检测结果，其中 f_i 代表帧位置的索引，c_i、s_i、b_i 分别为检测框的预测类别、置信度及目标框位置。视频目标检测的评价标准与静态图像中的目标检测类似，即当检测结果与人工标注框的帧位置、目标类别相符，且目标框的交并比满足一定阈值时，可认为正确检测。

1.1.2　视频目标检测的研究思路

目标检测是计算机视觉领域的经典研究方向。一个直接的思路是将视频中的每帧看作图像并使用静态目标检测的方法独立输出每帧的结果。传统基于统计学习的静态目标检测方法的解决思路是将目标检测问题转化为分类问题，对图像区域进行提取和特征描述并将其作为分类算法的输入，构建目标表征模型，进一步在待检测图像中获取目标检测位置并进行类别预测。传统方法可大致分为基于滑动窗判别的方法和基于目标区域提议的方法。随着深度学习的发展，目标检测网络模型又可分为单阶段目标检测模型和双阶段目标检测模型。

视频应用需求的急剧增加使视频目标检测引起研究者的广泛关注。与静态目标检测相比，视频目标检测的关键在于时序运动信息的应用，即如何结合时空一致性的判别信息并实现多帧表观信息的融合。由于视频编码的差异，以及目标在视频中的成像差异，视频目标检测可建立在关键帧和非关键帧区分的基础上。考虑

到视觉跟踪领域的蓬勃发展，联合关键帧检测和非关键帧跟踪，建立视频目标检测模型是一种直接思路。此外，光流作为视频任务中运动信息的常见表示方式[8]，可引入视觉目标检测任务。考虑到帧间存在表观互补，通过多帧信息融合，可以有效提升目标在运动模糊、视频散焦和遮挡等信息缺失条件下的精度。综上而言，如何有效利用视频的时空信息，平衡目标检测的精度和视频任务的效率，实现精度高且实时的视频目标检测，是当前计算机视觉理论和应用领域的一个重要研究课题。

1.1.3 视频目标检测的应用场景

视频目标检测研究在多领域都具有重要的应用前景，它是机器人视觉应用的基本前提。机器人需要对环境场景进行全面的感知理解才能做出决策，因此高效、高精度的视频目标检测是机器人应用的基本前提。例如，在机器人导航领域，视频目标检测可以告知场景中存在的目标类别及位置，在此基础上生成目标关系图谱，实现场景理解与感知。在对话型机器人领域，视频目标检测可以实现与环境的交互，获取更加真实的交互式体验，提升机器人智能化程度。在无人自动驾驶领域，机器人可以在视频目标检测的基础上，实现有效避障，并结合交通场景的变化做出决策。此外，在国防军事等领域，视频目标检测也具有重大的需求和重要的应用价值。

1.2 基于深度学习的视频目标检测

1.2.1 静态目标检测方法

随着深度学习的兴起和发展，基于深度神经网络的智能识别与检测取得了显著突破。深度卷积神经网络由多层首尾相接的卷积神经元组成，通过多层的线性卷积、非线性变换及空间池化实现由图像像素到描述特征向量的转化，并首先在图像分类领域得到应用。典型的网络结构包括 AlexNet[9]、VGGNet[10]、GoogLeNet[11]、ResNet[12]、SENet[13]等。通过大规模的数据学习深度神经网络的参数，并利用深度神经网络产生层次化的图像特征表征。神经网络浅层的特征图对应基元的特征表征（如图像边缘和纹理等），深层的特征图对应中高层的视觉特征表征（如目标部件等）。静态目标检测方法先以分类网络框架为基础，再进一步添加分类和回归分支，对目标位置和类别进行预测判别。以视频帧为处理对象，利用静态目标检测方法，是实现视频目标检测的一种直接思路。

静态目标检测方法主要可以分为两大类，单阶段图像目标检测方法和双阶段图像目标检测方法。单阶段图像目标检测方法可直接视作在用于分类的深度网络中引入滑窗判别实现目标检测，通常先对图像进行网格化划分或锚点设置，并对网格或锚点进行包围盒初始设置，进一步结合深度网络对目标位置和类别进行预测。

双阶段图像目标检测方法则增加了区域提议阶段以产生目标候选区域，对目标包围框的分类和回归利用了预选框内的区域特征。此外，以 DETR[6] 为代表的端到端图像目标检测方法通过集合预测的建模方法，避免了非极大值抑制的后处理步骤，2020 年之后得到了研究者的广泛关注。

1. 单阶段图像目标检测方法

YOLO(you only look once)[3] 是一种影响广泛的代表性单阶段图像目标检测方法。YOLO 的核心思想是对整幅图像进行网格化分区并作为网络的输入，直接在输出层预测网格区域是否为目标的得分、所属目标类别及边界框的位置。具体的实现过程如下：首先将图像划分为 $S\times S$(如 7×7)的网格，每个网格对应图像的不同区域，根据网格与目标位置及覆盖情况进行分配预测，每个网格负责预测一定数量的目标并输出置信度及对应物体各类别的概率；测试时，通过网络模型输出对应网格的分类置信度筛选目标框，并结合类别与目标框坐标信息分别预测每个锚框的位置及该网格对应目标的类别；最后将所有网格对应的预测框进行非极大值抑制，得到最终的检测结果。算法示意图如图 1-2 所示。YOLO 的精度低于一般的双阶段图像目标检测模型，但是运算速度可以达到 25 fps，是一种实时的目标检测方法。由于每个网格可预测的目标数设置为定值(如 2)，YOLO 难以应对分布较为稠密的物体，且定位误差相对较大。在此基础上进一步发展出的 YOLO-9000[14] 及 YOLO-v3[15] 分别结合单词树和多尺度特征实现了广泛类别的目标检测，并提升了检测精度。目前 YOLO 系列检测算法已成为性能最优的通用目标检测算法之一，且已发展出一系列目标检测方法。YOLO 系列具有较少的参数量，便于部署至各种嵌入式平台，具有优异的实时性，被广泛应用于工程实践。

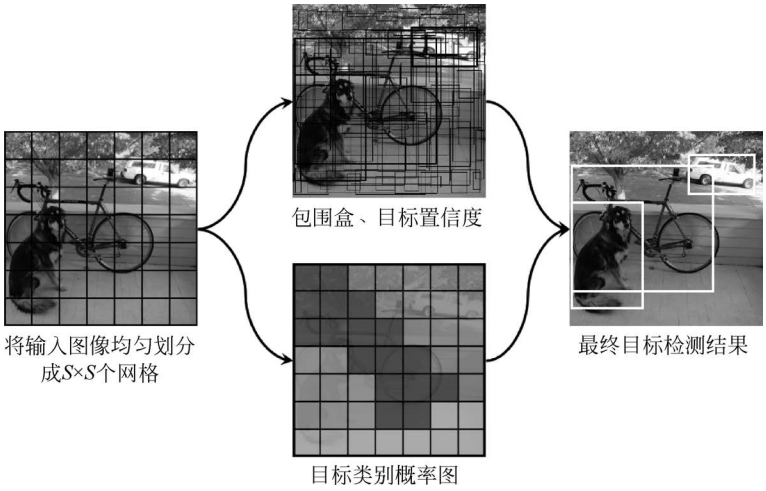

图 1-2　YOLO 目标检测算法示意图[3]

SSD(single shot multibox detector，单次多框检测器)[4] 通过引入多层特征并

将不同尺度的目标分配到对应不同尺度的特征提取网络分支上进行多尺度检测。简单来说，利用高层语义强的特征检测大目标，利用底层细节多、分辨率高的特征检测小目标。作为单阶段目标检测方法，SSD 可在提升目标检测效率的同时保持检测精度。输入图像通过卷积网络提取特征图，多层网络可以获取不同分辨率的特征图并组成特征金字塔。特征金字塔上每个像素点均可预测多个锚框的类别，并对目标相对于锚框的位置进行回归修正。通过密集分布在整张图像上的预测框密集进行多尺度的目标类别和位置预测，即可进行目标检测。同时产生大量的冗余预测，因此，最后需采用非极大值抑制过滤重复框。因为利用了不同层次的特征图，SSD 可以有效提取不同尺度的目标特征，进而提高目标检测效果。

RetinaNet[16]是一种基于特征金字塔(FPN)[17]结构的单阶段目标检测方法。在结构方面，RetinaNet 使用了特征金字塔结构，如图 1-3 所示，并在多层次特征图的每个像素点上设置预设形状和大小的锚框；对应每个锚框，直接在卷积神经网络的特征图上预测目标类别和相对于锚框的位置修订。考虑到多尺度特征金字塔的特征图包含的像素数，可以发现绝大部分的像素位置对应原图的背景区域，仅有极少部分的像素位置对应原图的目标。因此，训练过程中存在严重的正负样本不均衡问题，导致网络参数的优化极易被大量的负样本主导，而正样本提供的信息被淹没，进一步导致网络难以有效学习到目标的特征，影响单阶段目标检测性能。针对该问题，引入 Focal Loss 对困难样本，特别是对困难正样本(目标对应锚框)施加更高的权重，以提升单阶段目标检测的精度。Focal Loss 具体计算如式(1-1)所示。

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log p_t \tag{1-1}$$

其中，p_t 为预测目标的置信度，α_t、γ 为调节 Focal Loss 的超参数。由式(1-1)可见，正样本预测置信度越大，说明该样本越简单，所占的权重$(1-p_t)^{\gamma}$越小。训练中大部分的背景样本为简单样本，减少简单样本的权重可使网络训练更多地关注正样本及易错分的背景样本。因此，Focal Loss 通过动态加权的方式调整不同样本所占权重，可有效提高对困难样本的学习效率，极大地缓解正负样本不均衡的问题，进一步提升单阶段目标检测方法的精度。

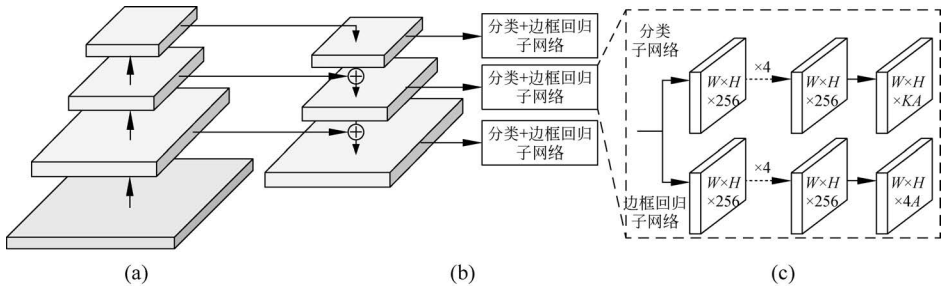

图 1-3 RetinaNet 目标检测算法示意图[16]

(a) ResNet 基线网络；(b) 特征金字塔网络；(c) 分类子网络(上)＋边框回归子网络(下)

上述单阶段目标检测算法对于目标包围框的回归是基于预先设定的一系列不同尺寸和长宽比的矩形框，即锚框。锚框的参数设定需要依赖经验，且与场景中目标的统计尺寸密切相关。由于锚框设定后便作为固定的参数，无法在模型训练和在线检测的过程中进行调整，导致泛化性能在一定程度上受限。因此，越来越多的研究关注无需锚框的目标检测方法。

CornerNet[18]将目标检测建模为关键点检测和匹配问题，并将目标包围框的左上角点和右下角点作为预测对象，通过预测角点坐标并匹配关联不同位置的角点实现目标检测。如图 1-4 所示，输入图像首先经过深度网络提取特征图，分别引入两个分支生成左上和右下角点的热度图，在此基础上预测目标包围框的相应角点。进一步匹配关联至同一目标的角点以确定目标包围框。基于角点的方法可以避免预设锚框，具备更高的灵活度。然而，目标包围框的角点通常位于背景区域，导致角点响应图易受背景干扰，易与背景特征混淆，增加网络学习的难度。为此，CornerNet 引入了角点池化，将目标包围框边缘的信息池化到对角点特征响应进行增强。CornerNet 成功地将目标检测与关键点检测任务相结合，容易推广至三维目标检测、旋转目标检测及视频目标检测。尽管角点池化可增加角点特征，只关注边缘角点而忽略特征较为明显的目标主体部分仍会限制目标检测的精度，同时角点匹配也可能引入误匹配的虚警。

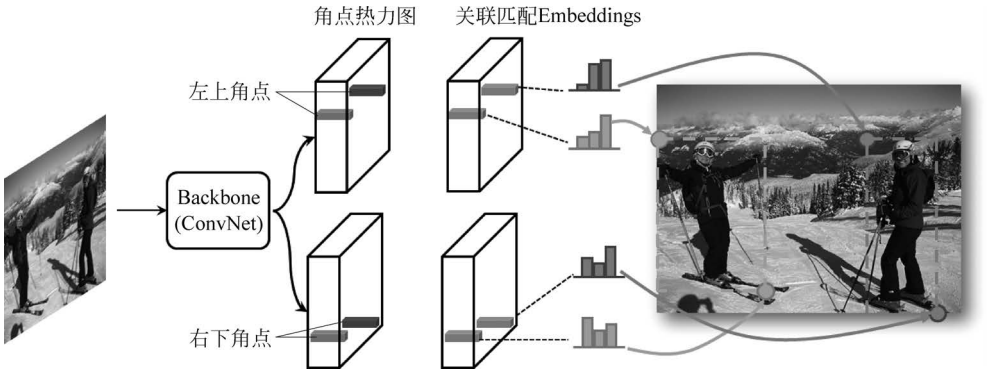

图 1-4　CornerNet 目标检测算法示意图[18]

此外，还有多种基于关键点建模的目标检测方法。例如，Objects as Points[19]对目标中心点位置和目标的长宽进行预测，并获取目标位置。CenterNet[20]则增加了目标包围框中心点热度图的分支，形成左上角点、中心点、右下角点三分支结构。由于目标包围框中心点相比目标角点往往有更大的概率位于目标主体，中心点的热点图可以充分基于目标特征表征获取，从而弥补 CornerNet 中对目标主体特征的缺失，改善检测精度。ExtremeNet[21]将两个角点分解为四个边缘点。具体而言，每个边缘点即目标与包围框边缘的切点，将左上角点分解为左边缘点和上

边缘点，右下角点分解为右边缘点和下边缘点。通过网络模型输出四个边缘点热度图分支和一个中心点热度图分支，匹配时，穷举边缘点的不同组合方式，根据中心点热度图对结果进行筛选。

单阶段图像目标检测通过神经网络生成全幅图像的特征表征图，直接在全局特征图上对目标位置进行回归并对目标类别进行预测。单阶段图像目标检测方法通常有简单直接的流程，具有较小的计算量和较高的检测速度。然而，直接在全幅图像对应的特征图上对目标包围框进行端到端回归预测，容易受背景区域特征影响，这也制约了单阶段图像目标检测方法的精度。

2. 双阶段图像目标检测方法

相对于单阶段图像目标检测方法而言，双阶段图像目标检测方法增加了区域提议阶段。与单阶段图像目标检测方法在全幅图像范围内直接回归预测目标包围框不同，双阶段图像目标检测方法对目标包围框的分类和回归基于区域提议阶段产生的预选框，最终的分类和回归也仅利用了预选框内的特征，而与图像中的其他区域无关。在预选区域框基础上的分类和回归可以看作是对第一阶段预选框进一步的细化和修正，从而可以获得更精确的目标检测结果。也就是说，双阶段图像目标检测网络的特点是通常包含区域预选分支，利用该分支首先从图像中分离出目标区域，然后结合感兴趣区域(ROI)池化操作等提取区域特征，并进一步实现目标位置和类别的预测。

R-CNN(region with CNN features，区域卷积神经网络)[22]是最经典的双阶段检测方法之一，其率先将深度卷积神经网络应用于目标检测的架构设计。R-CNN 在第一阶段采用选择性搜索[23]算法产生大约 2000 个目标候选区域。根据预选区域的包围盒坐标对原图进行裁剪，获得一系列独立区域的子图像，随后进行尺度归一化并将区域子图像缩放至预设的尺寸。在第二阶段，将一系列区域子图像输入至卷积神经网络进行特征提取，然后将这些特征输入至支持向量机(SVM)进行类别预测，同时结合边框回归细化目标包围盒位置。相比传统目标检测模型中使用手工设计的特征(如梯度直方图 HOG)，R-CNN 使用在大规模分类数据集上预训练的 AlexNet[9]进行区域特征提取，并引用微调调整网络结构参数，通过提升特征表征质量，R-CNN 的检测效果远高于传统目标检测方法。然而，R-CNN 的检测速度很慢，其对每个区域都利用深度网络进行特征提取的方式造成了大量的计算耗费冗余，同时分开训练特征提取网络与分类模块，难以达到整体最优。

Fast R-CNN[24]通过引入 ROI 池化，在卷积神经网络的特征图上实现区域特征的提取。ROI 池化操作可将大小不一致的卷积层输出特征图转化为区域相关的固定长度特征向量。通过 ROI 池化操作，Fast R-CNN 仅需要对每幅图像进行单次的神经网络特征提取，避免了区域特征的重复运算。此外，Fast R-CNN 将 ROI 池化提取的特征向量送入由全连接层构成的预测头，进行类别预测和边界框回归，

因此整个网络模型可以通过损失函数的方式进行端到端参数更新，避免独立学习预测模块和特征表征网络造成训练不充分问题。Fast R-CNN 的网络结构如图 1-5 所示。

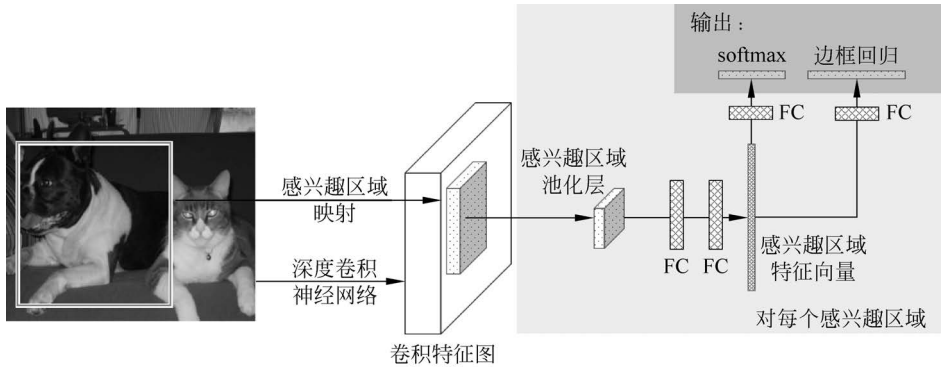

图 1-5　Fast R-CNN 的网络结构[24]

此外，Fast R-CNN 不再使用 SVM 进行分类，而是采用全连接层和 Softmax（式(1-2)）进行具体的目标类别预测和包围框回归。利用 Softmax 函数映射将网络中全连接层的输出转化为多类别目标的分类概率：

$$\mathrm{Softmax}(z_i)=\frac{\exp(z_i)}{\sum_k \exp(z_k)} \tag{1-2}$$

其中，z_i 为第 i 个类别经过全连接层之后计算的值。经过 Softmax 函数，将一个任意取值的实数值映射到(0,1)范围，即可描述对应区域特征匹配目标类别的概率。相比 R-CNN，Fast R-CNN 将包围框的分类和回归部分替换为多个全连接层构成的全连接神经子网络，从而可以通过梯度下降的方式与特征表征的卷积神经网络进行联合训练，同时通过端到端优化提升特征表征和目标检测任务预测的精度。

Faster R-CNN[5] 作为双阶段目标检测的经典算法，得到研究者的广泛青睐，是目前双阶段目标检测重要的基线方法。在 Faster R-CNN 中，引入窗口候选网络（RPN）进行目标/背景区域的预分类，从而实现基于深度网络的目标候选区域提取，如图 1-6 所示。卷积神经网络特征图上的每个像素点对应原图中相应区域的特征，以特征图上的特征点为中心点设置不同大小和长宽比的矩形框，成为锚框。通过引入卷积层生成锚框是不是目标区域的预测，并在预测分数的基础上进行目标区域的筛选，结合锚框的边框回归实现目标候选区域的生成。在获取候选区域后，Faster R-CNN 采用与 Fast R-CNN 一样的 ROI 池化操作、由全连接层组成的预测头部网络，以及基于区域的特征向量进行多目标分类及包围盒回归。Faster R-CNN 是一个端到端的目标检测架构，可打破利用传统方法获取目标候选区域仅能在 CPU 上运算的瓶颈，有效提升目标检测在 GPU 上的运行效率。

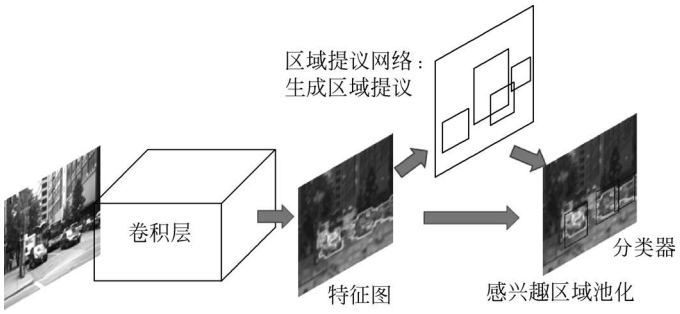

图 1-6　Faster R-CNN 的网络结构[5]

1.2.2　视频目标检测方法

相对于静态目标检测而言，视频目标检测的重点在于视觉信息和运动信息的综合应用。视频目标检测着重于探索三个方面，分别是运动信息的应用、结合时空一致性的信息判别及多帧表观信息的融合。基于深度学习的视频目标检测可以大致分为以下三类：基于跟踪的方法、基于运动信息融合的方法及基于帧间特征聚合的方法。基于跟踪的方法通过联合目标检测和跟踪建立面向视频的检测模型框架，需区分视频中的关键帧和非关键帧。基于运动信息融合的方法通常需提取光流进行运动信息表示，主要不足体现在长时运动表征不足。基于帧间特征聚合的方法是现阶段主要采用的方法，通过帧间信息的聚合实现特征互补。其主要不足是特征融合计算代价高，导致效率受限。如何有效利用视频的时空信息并实现实时且高精度的视频目标检测仍是当前亟须解决的问题。

1. 基于跟踪的视频目标检测

基于跟踪的视频目标检测方法将序列出现的图像帧区分为关键帧和非关键帧，其在关键帧进行目标检测，在非关键帧结合关键帧的检测结果进行目标跟踪，联合检测与跟踪实现完整的视频目标检测框架。早期的视频目标检测方法主要采用基于跟踪方法的思路，其主要依赖于关键帧检测和跟踪算法的性能，但在运动模糊、散焦情况下检测和跟踪可能失效，导致性能受限。

TCN(temporal convolutional network，时域卷积网络)[25]和 T-CNN(tube convolutional neural network，管道卷积神经网络)[26]面对视频目标检测的任务构建了时空基本单元，称为管道(Tubelet)，由单个目标在视频序列中每帧的包围盒按时间顺序连接构成。TCN[27]首先检测视频关键帧中的目标，然后选出置信度高的目标进行跟踪，通过时域卷积网络嵌入时间信息，以提高跨帧的检测结果，如图 1-7 所示。T-CNN 以 Tubelet 的提取与评分为基础，其网络结构包括管道提取模块、管道分类模块及重打分模块。其中管道提取模块通过区域预提取方法实现；