Website: ycyk.brit.com.cn Email: ycyk704@163.com

# 基于深度学习的目标检测框架组件研究

乔腾飞，张 超，熊建林，刘 斌，胡剑平
（北京遥测技术研究所 北京 100076）

**摘要：**深度学习与计算机视觉的结合给目标检测研究领域带来了全新的检测模式，通过对基于深度学习的目标检测网络分析研究，目标检测网络框架可模块化地拆分为特征提取网络、多尺度融合和预测网络三个部分。从组成目标检测网络模块化的角度对各个模块进行了详细的分析综述，并给出了如何根据实际需求来构建适合的模型框架建议，为基于深度学习的目标检测方法研究提供参考。

**关键词：**深度学习；目标检测；计算机视觉；模块化
**中图分类号：**TP183 **文献标识码：**A **文章编号：**CN11-1780(2022)06-0108-09
DOI：10.12347/j.ycyk.20220105001
**引用格式：**乔腾飞, 张超, 熊建林, 等. 基于深度学习的目标检测框架组件研究[J]. 遥测遥控, 2022, 43(6): 108–116.

# Research components of object detection framework based on deep learning

QIAO Tengfei, ZHANG Chao, XIONG Jianlin, LIU Bin, HU Jianping
（Beijing Research Institute of Telemetry, Beijing 100076, China）

**Abstract:** The combination of deep learning and computer vision has brought a new detection mode in the field of object detection. Through the analysis of deep learning-based object detection network, the object detection network framework can be modularized and divided into three parts: feature extraction network, multi-scale fusion network and prediction network. This paper analyzes and summarizes each module from the modularized perspective of detection network, and gives suggestions on how to build a suitable model framework according to actual demand, which provides a reference for the research of target detection method based on deep learning.

**Key words:** Deep learning; Object detection; Computer vision; Modularization
**DOI：**10.12347/j.ycyk.20220105001
**Citation:** QIAO Tengfei, ZHANG Chao, XIONG Jianlin, et al. Research components of object detection framework based on deep learning[J]. Journal of Telemetry, Tracking and Command, 2022, 43(6): 108–116.

## 引 言

随着深度学习技术的发展，其应用场景也在不断地扩展，从最初的图像分类网络到后续的目标检测、实例分割等应用场景，深度学习都展现出优异的性能，取得了不错的效果。2012 年 AlexNet[1]网络模型的提出，以巨大的优势在图像分类赛道上取得了远超第二名的优异成绩，成功将基于卷积神经网络的深度学习模型引入到人们的视野中，后续又有 VGGNet[2]、GoogLeNet[3]等优秀的网络陆续出现，均表现优异。2014 年，Girshick R 通过将卷积神经网络引入目标检测领域，形成了整个 R-CNN（Region-based Convolution Neural Networks）[4]系列的两阶段检测网络，并取得了不错的效果，之后学术界又陆续出现了 SSD[5]、YOLO[6]、CenterNet[7]等优秀的一阶段检测网络。

目标检测框架的流程要比分类网络模型更为复杂，并且不同的检测框架会带来不同的检测效果，为了获得一个高性能的检测网络，需要将网络框架中的不同模块进行适当的修改和组合。但随着网络模块的层出不穷，各种网络模块容易让初学者眼花缭乱，因此有必要对组成检测网络的不同模块进行分析综述。

收稿日期：2022-01-05 收修改稿日期：2022-02-20

## 1　目标检测网络框架

目标检测一直是图像处理领域的一个重要话题，早在上个世纪就已经开展了很多的研究，并且也形成了一套非常成熟的框架体系，一般会先对图像进行滤波降噪增强处理，提取出候选区域，然后通过手动设计方式进行特征提取，之后会对目标进行分类，判断属于哪一个类别。但是在实际图像中，检测目标的形态是多种多样的，光照条件多变，背景复杂，传统特征提取算子很难适应不同条件下的目标检测需求，为此相关学者采用卷积神经网络来提取特征，结合深度学习技术，形成了目前主流的目标检测方法。

R-CNN 是早期经典的基于深度学习的目标检测方法，该方法采用卷积神经网络来作为特征提取模块，并通过训练学习的方式来自动提取特征，然后将提取的特征送入传统 SVM（Support Vector Machine）分类器进行判断类别，取得了较好的实验效果。以该方法为基础，除了特征提取模块以外，后续相关学者又继续将多尺度特征融合、分类定位预测等功能都替换为相应的神经网络模块，从而形成了基于深度学习的目标检测网络框架，该框架主要包括特征提取网络模块、多尺度特征融合模块和预测网络模块，如图 1 所示。

### 1.1　特征提取网络

特征提取网络一般是网络模型的基础部分，是输入数据的接口，其作用主要是从输入的图像数据中提取出一些高维度的特征信息，如人脸中的眼鼻口耳、动物的毛发纹理等特征，这些特征很难用传统的特征描述算子直接表示，卷积网络通过大量的参数构建了一个高维度的空间，将输入的数据映射到这个空间中进行表示。著名的 ImageNet 竞赛中每年都会出现一些优秀的网络模型，2012 年 Hinton 和他的学生 Alex Krizhevsky 设计出的 AlexNet 以远超第二名的成绩夺冠，证明了卷积神经网络 CNN（Convolution Neutral Network）具有优秀的特征提取能力；2015 年何凯明等人提出了 ResNet[8]网络，该网络成为了至今最常用的特征提取网络。除了基于 CNN 的特征提取网络之外，来自 NLP 领域的 Transformer 也开始在计算机视觉领域大放异彩，尤其是 2021 年的 ViT[9]和 Swin Transformer[10]在图像领域取得巨大成功。除此之外，还有一些其他的优秀网络，如 DLA[11]、Hourglass[12]网络等。

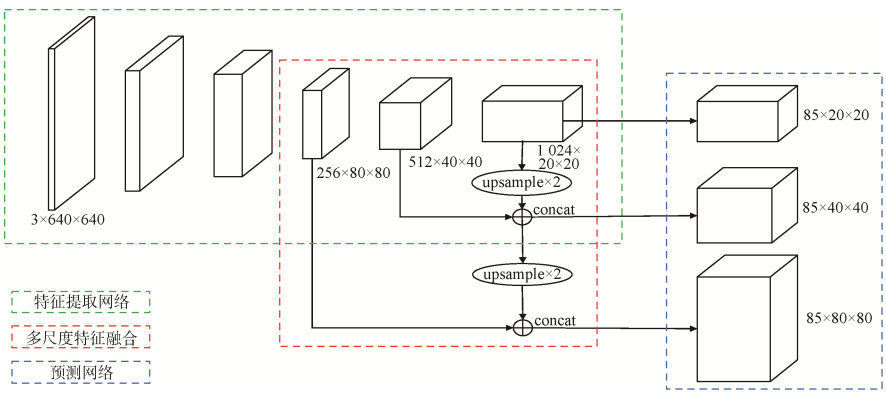


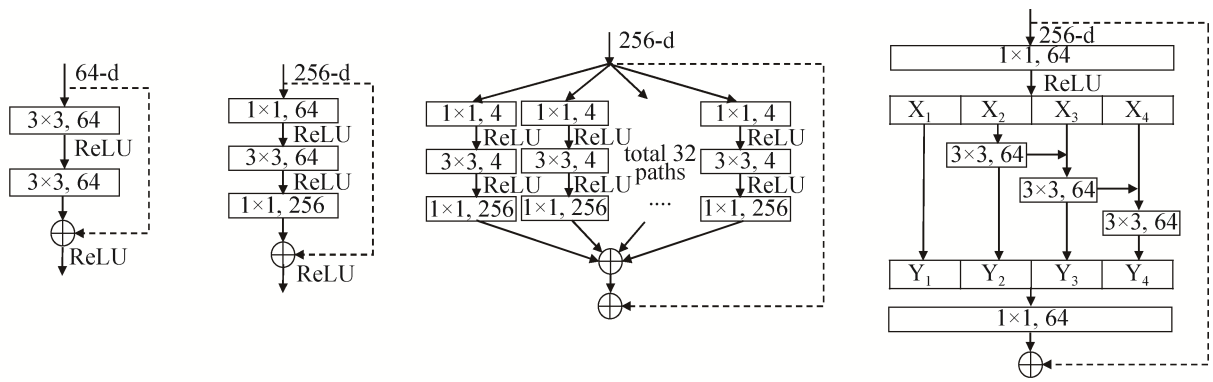

图 1　基于深度学习的目标检测框架

Fig. 1　Object detection framework based on deep learning

ResNet 是 2015 年何凯明等人提出的十分经典的特征提取网络，在此之前大部分的网络模型都是在 30 层以内，如 VGGNet 网络是由 19 层组成，GoogLeNet 是 22 层。从原理上来说，增加网络的深度可以提高模型的特征提取能力，有了更多的参数，模型就可以提取更多的信息，但是直接暴力的加深网络并不一定会带来一个更好的效果，甚至会导致网络变差。这是因为网络深度的增加带来的不仅有计算量的负担，网络模型训练过程中还会出现梯度爆炸和梯度消失，使得深层模型的训练过程无法成功收敛。ResNet 在一定程度上解决了这个问题，其引入的残差网络可以轻易使得网络的深度达到上百层，极大地增强了网络的特征提取能力。通过直接将输入信息绕道传到输出，保护信息的完整性，整个网络只需要学习输入输出的差别部分即可，简化了学习的难度。论文中提出了两种残差模块，图 2（a）是在常规的两层卷积之间加入一个高速连接，将靠近输入的数据逐元素相加；图 2（b）模块称为瓶颈（bottle neck）模块，首先将输入的特征图经过 1×1 的卷积降低通道的维数，再进行 3×3 的卷积操作，最后再次经过 1×1 卷积来恢复通道数，与短连接相加。

2017 年何凯明等[13]又提出了 ResNeXt 网络结构，继承了 ResNet 的残差策略，也是通过多个模块来堆叠网络结构。不同之处在于，ResNeXt 模块在内部模块执行了一组拆分-转换-合并的策略，其中转换

操作是在低维中进行，最后输出还是通过求和进行汇总，并且每一条都具有相同的拓扑结构，如图 2（c）所示。这样做的好处有两点：第一，可以降低模型的参数量，虽然模块中存在多路分支的结构，但是整体的卷积核参数量是有所减少的，可以增加更多的维度来提高语义特征；第二是将特征空间划分为多个子空间，在每个子空间内进行卷积计算，这样可以更加有效、更有针对性地进行特征提取。随后 2017 年，高尚华等[14]提出了新的 Res2Net 模块，与经典的 ResNet 模块相比，作者将原来的3×3 卷积分成了四组，并且每组之间又有一层高速连接，即在残差单元结构中又增加了小的残差块，如图 2（d）所示。




（a）残差模块　（b）残差瓶颈模块　（c）ResNeXt模块　（d）Res2Net 模块

（a）ResNet block　（b）bottleneck block　（c）ResNeXt module　（d）Res2Net module

图 2　残差模块系列

Fig. 2　Series of residual modules

稠密网络 DenseNet[15]于 2018 年提出，不同于 ResNet 的残差结构，DenseNet 认为残差破坏了原有的特征表示，于是将模块的输出相加结构改成了输出按通道合并的方式，这样可以直接复用网络前面的特征层。由于按通道合并的策略会导致网络的通道数越来越多，为了减少模型的参数量，使用1×1进行适当的通道降维。CSP（Cross Stage Partial）[16]结构于 2019 年提出，该网络在 2020 年被 YOLOv4 用作特征提取网络，CSPNet 在 DenseNet 网络的基础之上进行了改造，其初衷是减少计算量并且增强梯度的表现。如图 3 所示，在输入 DenseBlock 之前，将输入数据分为两个部分，一部分输入 DenseBlock 进行计算，另一部分直接通过一个短接通道绕到 DenseBlock 的输出端进行拼接。该结构与 DenseBlock 的主要区别在于多了一个分支，仅有一半的特征层用来参与卷积，另一半保留参与后面的信息传递。作者在论文中阐述了 CSP 结构的优点：可加强 CNN 的学习能力；能减少计算瓶颈，现在的网络大多计算代价昂贵，不利于工业的落地；能够减少内存消耗。CSP 结构与 DenseNet 并不属于耦合的关系，其中的 DenseBlock 也可以替换成 ResNet 模块或者任何一个特征提取网络的模块。

Transformer 是谷歌于 2017 年首次提出的一种应用在自然语言处理领域的深度学习框架，之后两年席卷了自然语言处理的大部分方向，表现优异。2021 年，微软亚洲研究院发布了 Swin Transformer，成功将 Transformer 应用在目标检测、语义分割等领域，并取得了不错的成绩。

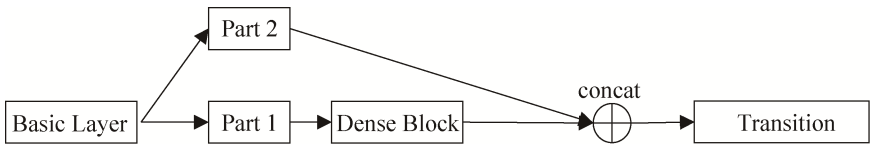



图 3　CSPNet 模块系列

Fig. 3　Series of CSPNet modules

与已有的网络不同，Transformer 不同于卷积神经网络，其将图像分割为一系列的小切片进行序列化再输入网络，为了与现有的视觉领域的检测框架相融，其必须显式的构建出不同尺度的层次结构。随着网络层次的加深，节点的感受也在不断扩大，CNN 这一特征在 Swin Transformer 中也是满足的。Swin Transformer 的这种层次结构，也赋予了它可以像 FPN，U-Net 等结构实现可以进行分割或者检测的任务。

此外，还有一些轻量级的特征提取网络，如 MobileNet[17]、SqueezeNet[18]、ShuffleNet[19]等系列，这

些网络与上面介绍的相比，具有更小的权重模型，如 MobileNet 中引入的深度可分离卷积可以实现计算量上的优化，进一步提高了模型的计算速度，为工业部署提供了可能。SqueezeNet 提出将 3×3 卷积替换成1×1卷积来减少卷积运算的参数量来降低模型的复杂度。

目前的特征提取网络的优化主要集中在模型的深度（卷积层数），通道方面的分离（CSP Block）和注意力机制（SENet），以及分组卷积（Res2Net）和深度可分离卷积（MobileNet）相关方面，选择一个好的特征提取网络对后续的多尺度融合和预测网络至关重要，在选择的时候不仅要考虑精度，也要根据实际的需求，选择适当大小的模型以实现速度上的要求。

### 1.2　多尺度特征融合

多尺度特征融合的作用是将特征提取网络中提取到的不同尺度特征进行融合，可以进一步提高网络的特征提取能力，并且引入了多尺度检测之后还可以帮助网络更好地检测到不同尺寸的目标。图像金字塔是图像多尺度表达的一种，是一种以多分辨率来解释图像的有效但简单的结构，特征图金字塔网络 FPN（Feature Pyramid Networks）[20]是 2017 年提出的一种网络，FPN 主要解决的是物体检测中的多尺度问题，在基本不增加原有模型计算量的情况下，大幅度提升了小目标的检测性能。

早期的检测网络，如 Faster RCNN、YOLOv1 等均采用特征提取网络的最后一层作为最终的特征层，在其上进行分类和定位预测。由于卷积池化的过程会构建出一个多尺度的特征层，因此可以通过使用不同层的特征图来进行不同尺度目标的预测。如下采样倍数为 8 的特征图适合用来预测小目标，下采样倍数为 16 的特征图适合用来预测中等大小目标，而下采样倍数为 32 的特征图适合用来预测大目标，根据实际需求，如果有特别大的目标也可以引入 64 倍下采样的特征层来。

在特征提取网络中，浅层的特征语义信息较少，但是目标位置准确；高层的特征语义信息比较丰富，但是目标位置等细节信息比较粗略。虽然也有部分算法采用多尺度特征融合的方式，但通常采用融合后的特征做预测，而 FPN 不一样的地方在于预测是在不同特征层独立进行的，并且增加了一条自上而下的路径，用来进行预测的每一层特征图不仅来自对应尺度的主干网络的输出，还有来自于上一层特征图的下采样，两者共同组成了当前层的特征图。自上而下的过程是对更抽象、语义更强的高层特征图进行上采样，而横向连接则是将上采样的结果和自底向上生成的相同大小的特征图进行融合。横向连接的两层特征在空间上尺寸相同，这样做可以利用底层定位细节信息，这个过程是迭代的，直到生成一系列的多分辨率图。如图 4 所示，其中绿色模块为特征提取网络，浅蓝色模块为 FPN 输出层。

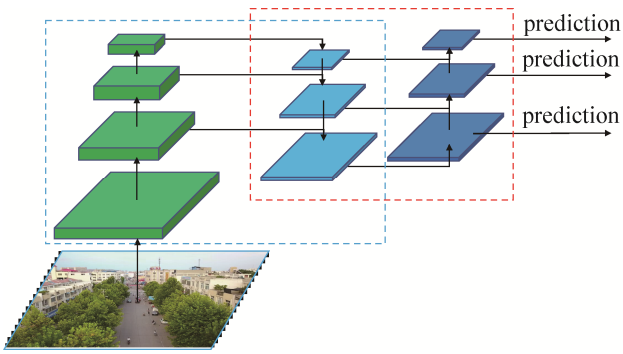



图 4　多尺度特征融合

Fig. 4　Multi-scale feature fusion

FPN 网络获得的多层特征图搭配相应的目标检测网络可以使得模型同时具备多个尺度目标检测能力，这种方式也成为了后续的多尺度目标检测的基石，后续的工作主要都是围绕着 FPN 的改进工作进行。PANet[21]是 2018 年提出来的，作者认为 FPN 网络的特征提取还不够充分，FPN 做的是将高层的语义信息反向传递，利用高层语义信息来提升低层的特征表达。网络的低层特征中含有更多的位置信息，高层特征中含有更多的语义信息。但是主干网络中的卷积层数太多，在卷积运算的过程中位置信息不能充分地上传到高层的特征图中，因此 PANet 在 FPN 下采样结束后，再返回进行上采样，并通过横向连接获取同级下采样的信息，提高底层信息的利用率。如图 4 所示，通过主干网络浅层信息传递到上层往往需要 100 多层的卷积层，引入 PANet 之后，经过的路径通常少于 10 层，也就是说底层的信息能更快速地传递到上层网络中，使回归网络更好地定位目标位置。

除了 PANet 之外，谷歌团队在 2019 年提出的 EfficientDet[22]中首次提到了 BiFPN。以往的特征融合是平等地对待不同尺度特征，BiFPN 引入了权重，能更好地平衡不同尺度的特征信息。BiFPN 将图 4 中

深蓝色的输入除了同一层的浅蓝色特征图和上采样特征图之外，还把同一层的绿色特征图也进行了融合，三个特征图分别赋予不同的权重系数，这些系数可以通过网络的学习获得。此外，还有 2020 年的 Recursive-FPN[23]，通过将特征提取网络中的特征图和 FPN 输出的特征图进行迭代增强，进一步加强了网络的特征表示能力，但是要求的算力很高，需要权衡。ASFF[24]研究了将 FPN 的不同尺度的输出层再一次融合，融合的时候也是带权重融合。

FPN 引入了一条自顶向下的通道来融合特征，PANet 在 FPN 的基础上增加了一条自底向上的通道，而 BiFPN 是在 PANet 的基础上增加了一条额外的边，每层的特征图输出有三个来源，每个来源都有对应的权重，充分利用了已有的特征图，丰富了 BiFPN 网络的最终输出特征层信息。整体的思路围绕着如何复用已有的特征图信息来增强网络的特征提取能力。

### 1.3 预测网络

预测网络即分类定位网络，是一个网络的末端部分，负责预测类别和回归边界框的位置坐标。根据网络的任务目标不同，会针对这些特征进行进一步的处理，比如分类网络会将这些特征信息通过全连接层映射为类别的预测概率信息，定位网络会通过卷积提取出目标的位置信息，分割网络会通过反卷积来恢复像素的类别信息等。之前所介绍的一些特征提取网络、多尺度特征融合、注意力机制等都是所有网络中通用的技巧，是用来提高卷积模型的特征提取能力，而检测头的设计则根据不同任务不同模型会有所不同。根据是否需要提供候选框分支网络来区分，目前的目标检测网络可以分为一阶段和两阶段检测器，比较经典的模型有三种，分别以两阶段 R-CNN、一阶段 YOLO 和 Anchor-Free 对应的预测网络模块为代表。

R-CNN 系列是深度学习应用在目标检测领域最早的一个模型之一，该系列是属于两阶段的检测器，网络中存在一个候选框预测分支网络，负责提供图像中潜在的感兴趣区域，将感兴趣区域送入预测网络进行类别预测和位置定位，这也是两阶段的检测器普遍的一个框架。第二阶段提取到的是目标的候选框，意味着每一个候选框对应一个潜在目标。一般来说，R-CNN 的检测头都会包含有全连接层，全连接层的输入维度必须是提前计算好的，为了能应对不同大小的图片输入，网络必须保证在全连接层之前的特征图大小是固定的，ROI Pooling 操作将不同大小的特征图归一化为统一的、固定的尺寸。在分类网络中 SoftMax 层替换了传统的 SVM，使用卷积、全连接层作为边框回归，这一操作也基本成为了后续目标检测流程中的模板。

后续又出现了 Cascade R-CNN 网络[25]，该网络是通过级联 R-CNN 检测头的方式来加强检测的精度。该网络解决了两个问题，首先以往基于锚框的检测头算法在计算正负样本的时候都是通过比较 IoU 的阈值来判定，一般这个阈值设为 0.5，为了避免漏检，这个阈值的设定就比较小，但问题也就随之而来，低阈值会产生大量的误检框，给模型的训练引入了一定的噪声。提高阈值可以减少候选框的数量，但是随之带来的是模型精度的下降。Cascade R-CNN 通过设置三个 R-CNN 检测头级联的方式，三个检测头的 IoU 阈值分别是[0.5, 0.6, 0.7]，靠前的检测头的阈值较低，因此可以减少漏检，靠后的阈值较高，可以提高检测的精度，通过这种组合的形式从而提高了模型整体的检测精度。

Redmon J 等人于 2015 年提出了一个基于单阶段的目标检测模型 YOLOv1，与之前的 R-CNN 系列不同，YOLO 利用单一的卷积神经网络将目标检测问题转变成一个简单的回归问题，把整张图片作为检测网络的输入，直接在网络输出层就能得到目标边界框的位置坐标以及目标种类，实现了端到端的优化，避免了冗长的处理流程。YOLO 将物体检测作为回归问题求解，假设提取到的特征层的尺寸是 $w\times h\times c$，其中 $c$ 是通道维度，$w$ 和 $h$ 是特征图的大小。利用卷积的感受野原理，将特征图映射到原图上，将图像离散划分成 $w\times h$ 个网格，特征图上的每个特征点都是 $c$ 维的特征向量，每个特征向量都可以预测一个或多个框，每个边界框要预测 $(x,y,w,h)$ 和置信度 5 个值。$(x,y)$ 表示边界框相对于网络单元边界框的中心，宽度 $w$ 和高度 $h$ 是相对于整个图片预测的，最后还要预测 $c$ 个类别概率。这些预测被编码为 $s\times s\times b\times(5+c)$ 的张量。目前的 YOLO 系列最新的框架是 YOLOv4 和 YOLOv5，都是非常优秀的目标检测网络模型，其在特征提取部分和特征融合部分做出了一些改动，但是其预测网络部分却一直沿用了 YOLO 的设计。

目前的检测算法主要思路还是设置大量 anchor+正负样本分配+训练的一个思路，anchor 的本质是目标的候选框，目的是帮助网络更好地收敛到真实目标框。但是因为目标的形状和位置的多种可能性，anchor 的数量往往非常庞大，否则会出现遗漏的情况，这种情况对于一阶段的检测算法更加突出。anchor 有两个缺点：① 通常会产生大量的 anchor，但只有少部分和真实框的重合比较大，可以作为正样本训练，其它都是负样本，这样就带来了正负例 anchor 的比例不均衡，也降低了网络的训练速度。② anchor boxes 的引入带来了许多的超参数，并且需要进行细致设计，包括 anchor boxes 的数量、尺寸、长宽比例。特别是单一网络在多尺度进行预测的情况下会变得复杂，每个尺度都需要独立设计。

在 YOLO 系列中，网络最后输出的每个尺度的特征图只有一个主干分支，在这个特征图上同时预测目标的类别、中心坐标、长宽信息。有一些网络则采用多分支预测的方法，比如 CenterNet 采用了三个分支网络，分别预测三个目标信息。这种方法和 Anchor-Based 的方法关系十分密切，因为 feature map 中的每一个像素点都可以看作是一个 anchor，只不过这种 anchor 只和位置有关。此外，由于 Anchor Free 算法计算量小、速度快的优点，这类算法一般也可以直接作为二阶段检测算法中的候选区域提取算法，来替换如 Faster R-CNN 网络中的 RPN 网络，如最新的 CenterNet2[26]模型中，就是使用了 CenterNet 网络来作为一个候选框的提取，如图 5 所示。

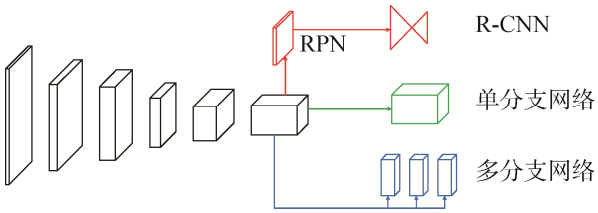



图 5　预测网络

Fig. 5　Prediction network

预测网络从功能上来说还是为了解决目标框的类别和位置信息的预测，不同算法采用了不同的预测方式。其中两阶段算法 R-CNN 系列是使用了两个分支来进行预测，全连接层用来预测类别，卷积层用来预测目标框的位置信息四元组 $(x,y,w,h)$。一阶段的检测器如 SSD、YOLO 系列，则直接用一个卷积分支来预测类别和位置信息 $(c,x,y,w,h)$，其中 $c$ 表示类别，与位置信息放在了一个卷积分支中进行预测。CenterNet 则采用了三个分支，分别是类别、框中心位置、框长宽，使用了三个卷积层分支来进行拟合。目前大部分的算法都是采用以上三种类型的回归网络，采用一阶段单分支的又较多，主要是 R-CNN 中的全连接层计算较慢，参数量较大，而三分支在网络设计上又较为复杂冗余。

## 2　框架搭建分析

从 2014 年开始，研究人员对深度学习在目标检测领域的研究不断深入，不断有新的优秀模型被提出，如 R-CNN 系列、YOLO 系列、Anchor-Free 系列等。同时，也有很多根据特定的场景需求，在经典模型上加入一些新的技巧进行改造出来的模型。

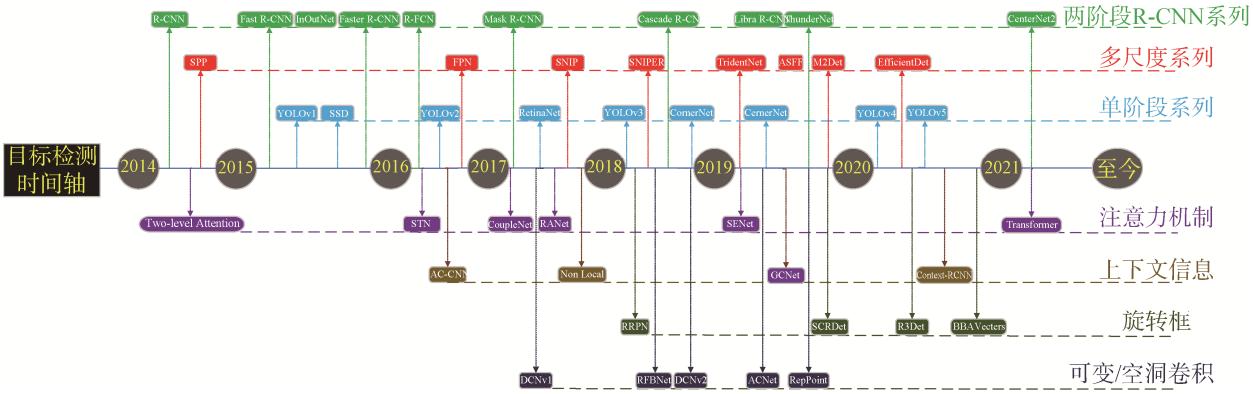



图 6　目标检测模型时间轴

Fig. 6　Timeline of object detection algorithm

本文对目标检测网络的三个主要组件进行简单的分析，根据实际需求构建出一个具体的检测网络，

一般步骤如下。

① 选取特征提取网络

特征提取网络通常是一个模型最关键的部分，对最终的检测精度有很大的影响；也是模型最大的部分，模型的卷积层数从十几层到几百层，参数量从几兆到几百兆。实际选取时，需要注意速度和精度两个因素。速度和精度是互相制约的，速度的极限意味着精度的损失，反之亦然。速度的快慢通常会受到卷积层数（网络深度）、通道数、卷积核大小、特征图大小等因素的影响，比如 ResNet50 和 ResNet151 两种特征提取网络，前者只有 50 层，后者多达 150 多层，前者的速度要快于后者，但是后者的精度要优于前者。

在设计自己的模块时，需要注意当通道数快速增加时，如 DenseBlock 的特征层拼接、Swin Tranformer 的池化操作都伴随着通道维数的增加，随后需要适当地使用1×1卷积来降低通道维度，目的是降低网络的计算量的同时减少模型的过多冗余参数。同时由于1×1卷积可以改变特征层的维度的特点，该方法也经常出现在一些轻量化模型设计中，如 bottleneck 模块先进行1×1卷积降维，再进行3×3卷积，最后通过卷积恢复维度，这样可以减少模型的参数量，加快网络的速度。

除了深度和通道数，也可以选择不同的特征增强模块加入网络主干中，如注意力机制模块，有可变形卷积等技巧来增强卷积网络的特征提取能力。注意力机制模块一般可以分为通道注意力和空间注意力，典型的模块有 SE-block[27]、CBAM[28]、SKNet[29]等，这些特征增强模块的特点是即插即用，与之前介绍的网络模型完全兼容，可以适当地加入模型中进行验证。

目前大多数的检测模型都会提供不同大小的模型，如 YOLOv5 就有 5s、5x、5m、5l 四个类型选择，其主要区别就是特征提取网络中的模块重复次数的区别，参数小的网络检测精度会差一些，但是速度会更快。

② 选取多尺度融合策略

多尺度特征金字塔是目标检测中普遍采用的策略，可以针对不同大小的目标进行调整。常规的目标检测网络如 YOLO 系列会使用三层特征图信息，其下采样倍数分别为 8、16、32，下采样的倍数越大，其感受野就越大，对应的可以检测更大尺寸的目标。而 SSD 采用了六层特征图信息，CenterNet 仅使用了 4 倍下采样特征层，实际建模时采用什么样的特征图需要根据实际的需求来确定。除了选取多尺度尺寸外，在多尺度融合策略上还可以采用不同的结构，如上文介绍的 PANet 和 BiFPN，这两个结构都是从特征层信息复用的角度出发，为了更好地使用特征提取网络获得的特征图信息。

③ 选取预测网络

预测网络选取一般是各个模型最大的区别，不同的模型对应不同的需求。一般来说选择的原则是追求精度可以选择两阶段的预测网络，如 Cascade R-CNN，通过级联方式进一步增强网络精度；追求速度可以选择一阶段 YOLO 系列的预测网络。根据 YOLOX[30]中提出的实验数据，分类适用全连接层，定位适用卷积操作，在预测时选择将分类和定位分开，这样的预测效果更佳。

## 3 未来发展趋势及展望

经过近十年的不断探索，基于深度学习的目标检测模型在公开数据集上的表现已经十分优异，人们逐渐开始将深度学习目标检测落地到具体的工业商用领域，如遥感图像、自动驾驶、人脸识别检测等领域，这些领域中的实际场景不同，给模型提出了更艰巨的挑战。

遥感领域中，图像的特点是尺寸非常大，一般是上万的分辨率，这与常规图片中几百的分辨率相差巨大，并且遥感图像中的目标具有整体稀疏、局部集中的特点，如港口船舶、机场飞机等目标对象，在一幅图中几乎 95%以上的区域都是没有目标的，甚至没有价值信息。因此避免不必要的运算量在遥感图像目标检测中是至关重要的一点，否则会因为巨大的计算量导致模型无法使用。类似的还有一些航空图像，远距离成像的图片都具备这一特点。

自动驾驶领域中，图像是车辆最主要的信息来源。由于车辆的行驶速度很快，对实时性要求非常高，只有快速的检测才能让车辆反应更快。且大多数车辆终端的处理器计算能力较弱，为了兼顾速度上的需求，精度上就无法做到很高，如何解决类似终端芯片上的目标检测应用也是一个很有价值的研究方向。

一些密集目标检测的场景，如行人检测，对模型的要求非常高。在模型的检测过程中，由于很多目标框之间的互相重叠，会导致模型无法识别到被遮挡的目标，从而漏检甚至误检。针对小目标检测，也是一个非常难解决的问题，小目标由于尺寸小，导致信息量较少，模型很难提取到充分的信息，但小目标又是实际应用中很常见的类型，这方面还需进一步的研究来提高检测效果。

此外，目前大部分的深度学习模型都是数据驱动型，必须要有大量标注好的数据来供模型进行训练，数据集制作的好坏将直接影响到模型的检测性能，网络模型对数据的依赖程度非常高。但很多情况下，由于标注成本或者采集难度等问题，无法获得足够量的数据或者获得的数据无法标注，导致模型就很难进行足够的训练，减少模型对大量数据的过度依赖也是一个亟待解决的问题。这方面的研究有半监督学习、弱监督学习和主动学习等，其目的都是为了减少模型对数据的依赖，从而减少人为的参与，让模型更加智能。

尽管近几年基于深度学习的目标检测模型研究十分成熟，尤其是卷积神经网络相关的模型已经形成了一系列成熟的框架，但在实际的应用过程中也面临着诸多困难。总体来说，目前关于检测模型方面的研究趋势主要有两个：

① 在学术界，基于深度学习的目标检测模型主要针对的是检测框架中的组件，致力于如何提高这些组件的性能，使用的多是常规数据集如 COCO、VOC 等，大多是即插即用的组件，如注意力机制、上下文信息等一些特征增强的模块，或者跨学科、跨领域的知识迁移，如将自然语言处理中的 Transformer 机制引入到视觉中。

② 在工业界，目标检测模型研究以应用为主，根据实际的背景需求来搭建一个合适的网络模型，这需要对检测框架中的不同模块进行一些验证测试，选择最佳的网络组件。航拍图像中如停车场、港口船舶会引入旋转框、遥感图像中的稀疏目标会进行粗筛选等。同时，考虑到部署相关的需求还会针对模型的剪枝量化做出一些改进工作。

影响一个模型检测性能的因素有很多，不同的组件在组合的过程中，单个最优并不一定组合最优，还要进行一定的速度和精度的权衡，实际的应用也是主要围绕速度和精度这两点展开。由于深度学习中有过多的超参数以及各种模块的搭建都需要人工测试，需要很大的精力，目前也有一部分围绕着训练超参数的工作展开，使用基于强化学习的策略主动去学习设计网络。

## 4 结束语

随着深度学习技术的发展，目标检测模型的性能上限在不断刷新，应用领域也在不断扩展，基于深度学习的目标检测应用出现在了不同的行业，给各个专业领域都带来了新的研究模式。目前已经出现了越来越多的深度学习模型，但是其中经典的模型数量却有限，大多还是根据已有的一些技巧进行多方面的组合或者针对其中一个技巧进行针对性的改善。但是针对具体场景、具体需求时还是需要研究工作者多去尝试不同技巧的组合，这样才能找到最适合的网络模型。同时，也可以根据相关的研究趋势进行特定模块的优化。

## 参考文献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[C]// Conference and Workshop on Neural Information Processing Systems(NIPS), 2012.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint:1409.1556, 2014.

[3] SZEGEDY C, LIU W, JIA Y Q. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2014.

[4] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440–1448.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector [C]//European Conference on Computer Vision (ECCV), 2015.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016.

[7] ZHOU X Y, WANG D Q, PHILIPP K. Objects as points[J]. arXiv preprint: 1904.07850, 2019.
[8] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016.
[9] ALEXEY D, LUCAS B, ALEXANDER K, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]//In International Conference on Learning Representations, 2021.
[10] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2021.
[11] YU F, WANG D Q, SHELHAMER E, et al. Deep layer aggregation[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017.
[12] NEWELL A, YANG K Y, DENG J. Stacked hourglass networks for human pose estimation[C]//European Conference on Computer Vision (ECCV), 2016.
[13] XIE S, TU Z W, GIRSHICK R, et al. Aggregated residual transformations for deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017.
[14] GAO S H, CHENG M M, ZHAO K, et al. Res2net: A new multi-scale backbone architecture[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.
[15] HUANG G, LIU Z, KILIAN Q, et al.Densely connected convolutional networks[C]//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 4700–4708.
[16] WANG C Y, YUAN H, WU Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPR Workshop), 2020.
[17] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017.
[18] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016.
[19] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017.
[20] LIU T Y, DOLLAR P, GRISHICK R, et al. Feature Pyramid for Object Detection[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017.
[21] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018.
[22] TAN M X, PANG R M, QUOC V. Efficientdet: Scalable and efficient object detection[J]. arXiv preprint:1911.09070, 2019.
[23] QIAO S Y, CHEN L C, YUILLE A. DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2020.
[24] LIU S T, HUANG D, WANG Y H. Learning spatial fusion for single-shot object detection[J]. arXiv preprint: 1911.09516v2, 2019.
[25] CAI Z W, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018.
[26] ZHOU X Y, KOLTUN V, KRAHENBUHL P. Probabilistic two-stage detection[J]. arXiv preprint:2103.07461, 2021.
[27] HU J, SHEN L, SAMUEL A, et al. Squeeze-and-Excitation Networks[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018.
[28] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//In Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3–19.
[29] LI X, WANG W H, HU X L, et al. Selective kernel networks[C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018.
[30] GE Z, LIU S T, WANG F, et al. YOLOX: Exceeding YOLO series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

[作者简介]

乔腾飞 1996 年生，硕士研究生，主要研究方向为智能图像处理。
张 超 1986 年生，博士，工程师，主要研究方向为图像处理与模式识别。
熊建林 1965 年生，硕士，研究员，主要研究方向为航天测控技术。
刘 斌 1978 年生，博士，研究员，主要研究方向为空间电子信息技术。
胡剑平 1970 年生，本科，研究员，主要研究方向为卫星数据存储与处理技术。

(本文编辑：杨秀丽)