第 5 章

工业机器视觉目标检测与跟踪

本章深入探讨了工业机器视觉中的目标检测与跟踪技术，旨在为读者提供一个全面理解这些技术如何在现代工业生产中应用的基础。目标检测作为计算机视觉的关键领域，其核心在于准确地识别和定位图像或视频中的特定物体。在工业应用中，这一技术不仅提升了生产效率和质量控制水平，还增强了工业安全性和稳定性，促进了智能制造的发展。

本章围绕以下内容展开。

5.1 介绍了目标检测的定义，以及其在工业中的重要性和应用价值。

5.2 概述了传统的目标检测方法，包括基于模板匹配、基于特征点的方法和其他经典算法。

5.3 探讨了基于深度学习的目标检测方法，重点介绍了 YOLO 系列、R-CNN 系列和其他目标检测方法。

5.4 介绍了目标跟踪的基本概念和技术方法，包括传统方法、基于相关滤波的跟踪算法、基于孪生网络的跟踪方法以及其他前沿模型。

5.5 通过具体案例展示了上述技术在工业中的应用，如医药自动化生产线中的药液微弱异物检测、可见光与红外图像融合的电力热故障判别、电力自动化巡检中的异物检测等。

5.6 总结了本章的主要内容，回顾了目标检测与跟踪的基本概念、技术方法和工业应用。

5.7 提供了思考题，帮助读者巩固和思考所学知识的实际应用与未来趋势。

5.1 目标检测的基本概念

5.1.1 目标检测的定义与意义

目标检测是计算机视觉的核心领域之一，其目的是在图像或视频中准确地识别并定位一个或多个特定类别的物体。从技术角度来说，目标检测任务通常包括两个基本子任务：分类和回归。如图 5.1 所示，分类是指确定图像中是否存在特定类别的物体；而回归则要求精确地标出这些物体在图像中的位置，通常使用边界框来表示。

目标检测在工业领域中具有广泛而深远的意义，涵盖了从提高生产效率到推动产业智能化转型的多方面如图 5.1 所示。

图 5.1　目标检测的定义

1. 显著提升生产效率与质量控制水平

目标检测技术在工业生产中的应用，如同为生产线上的产品质检工作注入了“智慧之眼”；它能够实时、精准地识别生产线上的产品，并对产品的外观质量进行快速检测。相比传统的人工质检方式，目标检测技术不仅速度更快，而且检测精度更高，能够识别出人工难以察觉的微小缺陷。在高度自动化的工业生产环境中，目标检测技术能够与生产线上的其他自动化设备实现无缝对接，形成完整的自动化生产流程。例如，在汽车制造业中，目标检测技术可以应用于车身焊装线的质量检测环节，对车身的各个部位进行精确检测，确保焊接质量符合标准。同时，它还能够实时监测生产线的运行状态，及时发现并解决潜在的质量问题，从而避免了因产品质量不合格而导致的返工和材料浪费。此外，目标检测技术还能够通过数据分析和挖掘，为生产者提供有价值的生产信息。例如，在半导体制造业中，目标检测技术可以收集并分析生产过程中的各种数据，帮助生产者了解生产线的运行状态、产品质量的分布情况等信息，从而为生产优化提供决策支持。

2. 增强工业安全性与稳定性

目标检测技术在工业安全领域的应用，对于保障工业生产的安全性和稳定性具有重要意义；它能够实时识别操作环境中的障碍物和目标物体，为工业机器人等自动化设备提供精准的导航和避障信息，从而避免了因设备碰撞而导致的安全事故。在涉及重型机械或危险作业的工业场景中，目标检测技术的应用更是不可或缺。例如，在石油化工厂中，目标检测技术可以实时监测生产过程中的异常情况，如泄漏、火灾等，及时发出警报并采取措施，有效防止事故的发生。同时，它还能够对生产环境中的各种安全隐患进行预警和监控，为生产者提供可靠的安全保障。此外，目标检测技术还能够与工业互联网等先进技术相结合，构建智能化的安全生产管理系统。通过实时监测和分析生产数据，系统能够及时发现潜在的安全隐患，并采取相应的措施进行防范和应对。这种智能化的安全管理方式不仅提高了工业生产的安全性，还降低了因安全事故而导致的经济损失和社会影响。

3. 推动智能制造与工业转型升级

目标检测技术是实现智能制造的重要技术支撑之一；它能够提供精确的物体识别与定位信息，为工业系统实现自动化的检测、分拣、装配等任务提供有力的技术支持。随着工业4.0的推进和智能制造的兴起，目标检测技术在工业领域的应用越来越广泛。在智能工厂中，目标检测技术可以应用于生产线的智能化改造和升级。通过实时监测生产线的运行状态和产品质量信息，系统能够自动调整生产参数和工艺流程，从而实现了生产过程的智能化控制。同时，目标检测技术还能够与其他智能设备实现互联互通，形成智能化的生产网

络。这种智能化的生产方式不仅提高了生产效率和质量水平，还降低了人力成本和能耗水平。此外，目标检测技术还能够推动制造业向更高层次、更高水平的智能化、自动化方向发展。通过与大数据、云计算等先进技术相结合，目标检测技术能够实现对生产数据的深度挖掘和分析，为生产者提供精准的决策支持和预测服务。这种智能化的生产模式不仅提高了制造业的竞争力水平，还推动了整个产业链的协同发展。

总的来说，目标检测技术作为工业机器视觉的重要组成部分，正在不断推动工业生产向智能化、自动化方向发展。它不仅提高了生产效率和产品质量，还为实现更高水平的工业自动化和智能化奠定了坚实的技术基础。随着技术的进一步发展，目标检测将在更多工业场景中发挥关键作用，成为未来智能制造的核心驱动力之一。

5.1.2 目标检测的挑战与技术难点

尽管目标检测技术在工业领域已取得了显著的进展，但在实际应用中，仍面临着诸多挑战和技术难点。这些问题不仅影响了检测的准确性和效率，也决定了目标检测技术在复杂工业环境中的适用性。

1. 多样化的目标物体

在工业检测领域，目标物体的多样化是目标检测算法面临的一大挑战。这种多样化不仅体现在物体的形态、尺寸、颜色、材质等物理特性上，还体现在物体的种类、功能、用途等方面。首先，从物理特性上看，工业环境中的目标物体往往具有高度的多样性。例如，在汽车制造过程中，不同型号的零部件外观差异巨大，即使是同一类零部件，也可能因产品型号不同而存在差异，如图 5.2 所示。这种差异可能导致传统的目标检测算法难以准确识别和定位这些物体。为了应对这一挑战，算法需要采用更复杂的特征提取方法和更强大的模型结构，以学习到目标物体的本质特征。其次，从种类、功能、用途等方面看，工业环境中的目标物体也呈现出多样化的特点。例如，在电子制造行业中，需要检测的目标物体可能包括电路板、芯片、电容器等多种不同类型的元器件。这些元器件在形态、尺寸、颜色等方面都存在差异，且每种元器件都有其特定的检测要求和标准。因此，算法需要具备高度的灵活性和可扩展性，以适应不同种类、功能、用途的目标物体的检测需求。为了应对多样化的目标物体带来的挑战，目标检测算法需要不断优化和改进。例如，可以采用深度学习中的卷积神经网络(CNN)等先进技术来提取更丰富的特征信息；同时，结合领域知识和专家经验来构建更高效的检测模型和算法。此外，还可以采用数据增强等技术手段来提高模型的泛化能力，以应对不同批次和型号的零部件检测任务。

图 5.2 多样化的目标物体

2. 复杂背景与环境干扰

在工业检测中，复杂背景与环境干扰是另一个重要的挑战。工业场景中的背景往往复

杂多变，包含多种颜色、纹理、反光表面等干扰因素；此外，工厂内的光照条件也可能随时变化，如自然光与人工照明的切换，以及遮挡现象等条件变化。这些干扰因素都会严重影响目标检测的稳定性和准确性。首先，复杂背景中的颜色、纹理等特征可能与目标物体相似，导致算法难以准确区分目标和背景。为了解决这个问题，算法需要采用更精细的特征提取方法和更强大的分类器来区分目标和背景。例如，可以利用深度学习中的注意力机制等技术手段来使算法更准确地聚焦于目标物体本身，而忽略背景中的干扰因素。其次，环境干扰如光照变化、反光表面等也会对目标检测造成严重影响。光照变化可能导致图像中的目标物体明暗不均，反光表面则可能产生强烈的反射光，这些都会使目标物体的特征变得模糊或难以识别。为了应对这些挑战，算法需要采用自适应阈值、图像增强、去噪等预处理技术来降低背景干扰；同时，结合深度学习中的鲁棒性训练等技术手段来提高模型对光照变化和反光表面的鲁棒性；为了进一步提高目标检测算法在复杂背景与环境干扰下的性能表现，还可以采用多模态融合等技术手段。例如，可以结合视觉和雷达等多种传感器信息来构建更全面的检测系统；同时，利用深度学习中的多模态学习等技术手段来融合不同模态的信息，以提高检测的准确性和稳定性。

3. 多场景的跨域适应性

在实际应用中，目标检测模型往往需要在不同的工业场景中使用，而每个场景可能存在不同的光照条件、背景噪声、相机参数等。这种跨域应用的需求要求目标检测模型具备良好的适应性，能够在新的环境下仍然保持高效的检测性能。首先，不同工业场景中的光照条件可能存在显著差异。例如，在室外环境中进行目标检测时，可能会受到阳光直射或阴影的影响；而在室内环境中，则可能受到人工照明或自然光的影响。这些光照条件的变化可能导致图像中的目标物体明暗不均或产生反光现象，从而影响检测的准确性。为了应对这一挑战，算法需要采用自适应光照调整等技术手段来降低光照变化对检测的影响。其次，不同工业场景中的背景噪声也可能存在差异。例如，在自动化生产线上进行目标检测时，可能会受到机械振动、电磁干扰等噪声的影响；而在仓储系统中进行目标检测时，则可能受到灰尘、污渍等噪声的干扰。这些噪声可能导致图像数据的质量下降，进而影响目标检测的准确性。为了应对这些挑战，算法需要采用去噪、图像增强等预处理技术来提高图像质量；同时，结合深度学习中的鲁棒性训练等技术手段来提高模型对背景噪声的鲁棒性。为了进一步提高目标检测模型在不同工业场景中的适应性，可以采用域适应技术和迁移学习方法。域适应技术旨在使模型能够适应不同域的数据分布，从而提高模型在新环境下的性能表现。迁移学习方法则可以利用已有的相关任务数据来提高模型在新场景下的性能表现。例如，可以利用对抗性训练等技术来使模型对不同的光照条件和背景噪声具有更强的鲁棒性；同时，结合迁移学习方法利用已有的相关任务数据来提高模型在新场景下的泛化能力。通过这些技术手段的应用，可以进一步提高目标检测模型在不同工业场景中的适应性和性能表现。

总之，目标检测技术在工业环境中的应用面临多重挑战，从目标的多样性、复杂的背景干扰，到实时性的严格要求以及硬件的协同优化，每一个环节都对技术提出了新的要求。尽管如此，随着技术的发展和不断的创新，这些挑战正在逐步被克服，目标检测在工业领域的应用将越来越广泛和深入。

5.2 传统目标检测方法

传统目标检测方法是在深度学习普及之前广泛应用的技术手段，这些方法通过基于特征和模型的方式来识别和定位图像中的目标。虽然在某些复杂场景下，这些方法可能不及现代深度学习方法的表现，但它们在特定应用中仍然具有独特的优势，特别是在计算资源有限或实时性要求较高的工业环境中。

5.2.1 基于模板匹配的目标检测

基于模板匹配的目标检测是计算机视觉领域中一种简单直观的方法，广泛应用于各种目标定位和识别任务中；该方法的核心思想是在原始图像中寻找与预设模板相匹配的区域，通过相似度度量来确定目标的存在与否。

1. 算法原理

1）基本假设

模板匹配算法的基本假设包括以下几点。

（1）目标形状固定：假设目标在图像中的形状、大小和方向是固定的或变化很小，可以通过一个特定的模板来表示目标；

（2）背景相对简单：目标周围的背景应尽量简单，以减少误检的可能性；

（3）光照条件一致：目标区域与模板之间的光照条件应尽可能一致，以保证匹配的准确性。

2）模板匹配过程

（1）模板准备。

从已知包含目标的图像中提取出一个或多个模板；这些模板通常选择目标清晰、特征明显的部分；模板的选择对最终的检测结果至关重要，一个好的模板应该能够反映目标的主要特征，同时尽量减少背景干扰。

（2）滑动窗口搜索。

在待检测图像上使用一个与模板大小相同的窗口进行滑动搜索；这个窗口在图像上的每一个可能的位置上都会停下来，与模板进行比较；滑动步长可以根据需要调整，较小的步长可以提高精度，但会增加计算量；较大的步长可以加快搜索速度，但可能会错过某些潜在的匹配位置。

（3）相似度度量。

在每个停下的位置，计算窗口内的子图像与模板之间的相似度。根据相似度的高低来判断该位置是否为目标所在；常用的相似度度量方法包括平方差之和（Sum of Squared Differences，SSD）和归一化交叉相关（Normalized Cross-Correlation，NCC），示例计算结果如图 5.3 所示。

平方差之和：SSD 是一种简单的度量方式，用于计算模板与图像局部区域之间的差异。计算公式为

$$M_{\mathrm{SSD}}=\sum_{x,y}(I(x,y)-T(x,y))^2 \tag{5.1}$$

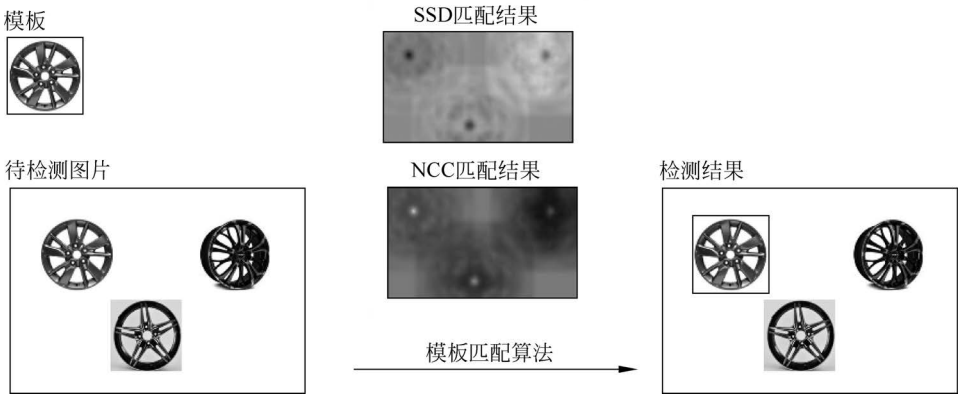

图 5.3　基于模板匹配的目标检测

其中，$I(x,y)$表示图像在位置(x,y)处的像素值，$T(x,y)$表示模板图像在相同位置的像素值；SSD 值越小，表明匹配度越高。

归一化交叉相关：NCC 是一种更为复杂的度量方法，它考虑了模板和图像局部区域之间的线性相关性。计算公式为

$$M_{\mathrm{NCC}}=\frac{\sum_{x,y}(I(x,y)-\bar{I})(T(x,y)-\bar{T})}{\sqrt{\sum_{x,y}(I(x,y)-\bar{I})^2\sum_{x,y}(T(x,y)-\bar{T})^2}} \tag{5.2}$$

其中，$\bar{I}$ 和 $\bar{T}$ 分别为图像窗口和模板的平均灰度值。M_{NCC} 的取值在$-1\sim1$，值越接近 1，表示匹配度越高。

(4) 结果分析。

在所有位置计算完相似度后，可以选择具有最高相似度(对于 NCC 而言)或最低差异(对于 SSD 而言)的位置作为目标的最佳匹配位置；通常，可以设置一个阈值来过滤掉那些相似度较低的候选位置，以减少误检。

2. 优点与缺点

模板匹配一般不需要复杂的模型训练过程，算法逻辑简单，易于编程实现；此外，在模板和图像较小的情况下，计算速度较快，适用于对实时性要求较高的应用。然而，模板匹配对光照、遮挡和噪声敏感，这些因素可能严重影响匹配结果。常规模板匹配算法通常难以处理目标的姿态变化和尺度变化，需要对每种可能的变化都预备一个模板。基于上述优缺点，基于模板匹配的目标检测算法适合在印制电路板制造的视觉检查系统、条形码或 QR 码识别和机器人姿态估计和对齐等场景应用。

5.2.2　基于特征点的目标检测

基于特征点的目标检测方法利用图像中的显著特征点，如角点、边缘和兴趣点来识别和定位目标。这些特征点是局部图像区域中具有显著变化的点，通常对旋转、尺度变化和光照变化具有一定的鲁棒性。

1. 算法原理

特征点检测方法的核心思想是在图像中寻找具有特殊属性的点，这些属性可以是局部

最大梯度、特定形状的边缘交叉点或者是某种特殊的纹理模式。通过提取这些特征点及其描述符，可以在不同图像之间进行匹配，从而识别出目标物体，示例检测效果如图 5.4 所示。算法步骤如下。

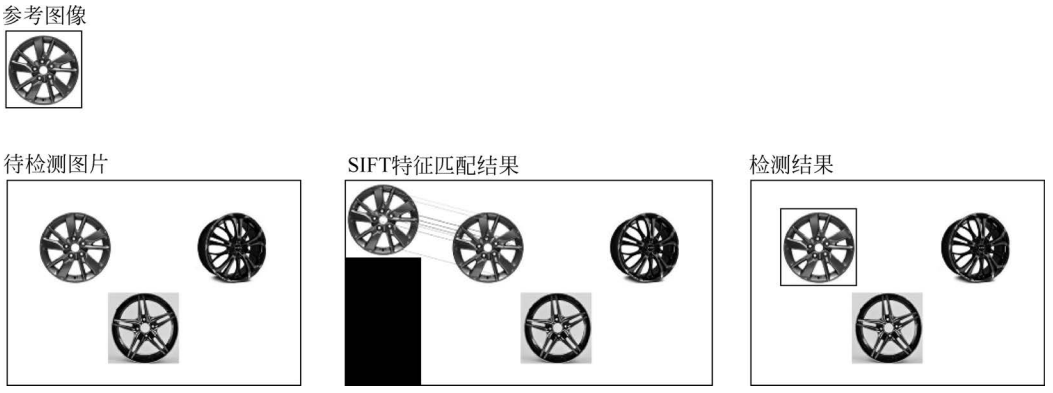

图 5.4　基于 SIFT 特征点的目标检测

1）特征点检测

使用前文提到的 Harris 角点检测器、SIFT（尺度不变特征变换）、SURF（加速稳健特征）等特征算法来自动检测图像中的特征点；这些算法通常通过计算局部梯度、二阶导数或局部对比度来识别特征点。

2）描述符计算

为每个检测到的特征点计算一个描述符，该描述符编码了特征点周围的局部图像结构信息。例如，SIFT 描述符使用梯度直方图来描述特征点的邻域，而 SURF 描述符则使用 Haar 小波响应来编码特征点周围的局部信息。描述符的设计目的是使特征点在不同的视角、尺度和光照条件下仍然保持一致性和可识别性。

3）特征点匹配

在参考图像和待检测图像之间进行特征点的匹配，通常使用欧几里得距离或其他相似度度量（如余弦相似度）来确定潜在的匹配对；为了提高匹配的准确性，可以使用双向匹配策略，即不仅要求参考图像中的特征点与待检测图像中的特征点匹配，还要求待检测图像中的特征点也能反向匹配到参考图像中的特征点。

4）匹配对筛选

应用 RANSAC（随机抽样一致性）或其他鲁棒性方法来剔除错误匹配，确保只有正确的匹配对被用于后续处理；其中 RANSAC 通过随机选择一组匹配对，计算一个变换模型（如单应性矩阵），然后评估该模型对所有匹配对的适用性。多次迭代后，选择最符合大多数匹配对的模型。

5）目标定位与识别

根据匹配的特征点计算单应性矩阵或其他变换模型，实现目标的精确定位和识别。单应性矩阵的计算满足式（5.3）关系：

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \boldsymbol{H} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{5.3}$$

其中，(x,y)和(x',y')分别为原图像和变换后图像中对应点的坐标，$\boldsymbol{H}$ 为单应性矩阵。单应性矩阵描述了两个平面之间的几何变换关系，可以用于校正图像的透视变形，从而实现目标的精确定位。

2. 优点与缺点

基于特征点的方法对图像的旋转、尺度变化和一定程度的视角变化具有较好的鲁棒性。特征点描述符能够捕获丰富的局部结构信息，有助于区分不同的目标物体；此外，该方法适用于多种不同的目标检测任务，特别是在无法获得大量标注数据的情况下。

值得注意的是，特征点的检测和描述符计算可能较为耗时，尤其是在高分辨率图像中。在低纹理或重复纹理的区域，特征点的匹配可能变得不可靠。这种算法的性能在很大程度上依赖参数的选择和优化。

5.2.3　其他经典目标检测算法

除了模板匹配和基于特征点的目标检测方法外，传统目标检测领域还有许多经典的算法。这些方法主要依赖于人工设计的特征和统计模型，在特定应用场景中具有独特的优势。以下介绍几种具有代表性的经典目标检测算法。

1. 基于边缘检测的目标检测

边缘检测是图像处理中的基本方法之一，通过识别图像中像素值变化剧烈的区域来检测目标的边缘。边缘通常是图像中灰度值发生剧烈变化的区域，代表了物体的轮廓和形状特征。这种方法适用于具有明确边界和对比度较高的目标物体的检测。

常用算法有以下两种。

(1) Canny 边缘检测：Canny 算法是一种经典的边缘检测方法，它通过对图像进行高斯平滑、计算梯度、非极大值抑制和双阈值处理，最终提取出图像中的边缘，检测效果如图 5.5 所示。

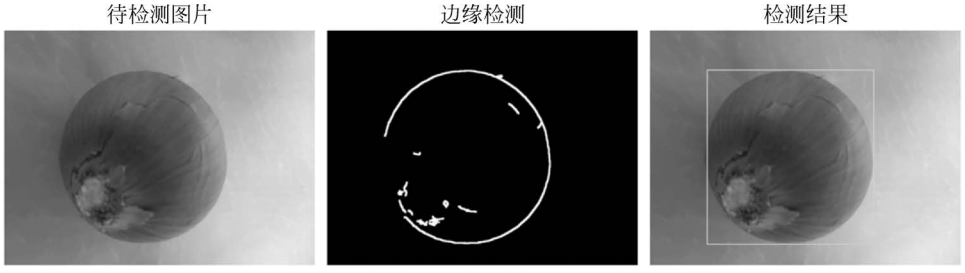

图 5.5　使用 Canny 算法做目标检测

(2) Sobel 算子：Sobel 算子是一种简单的边缘检测方法，通过在水平方向和垂直方向上对图像进行卷积操作，计算出图像梯度，进而提取边缘信息。

基于边缘检测的目标检测算法实现简单，计算量相对较小，特别适合实时处理和资源有限的应用场景；此外，边缘检测方法具有较强的普适性，能在多种视觉任务中使用，如物体轮廓检测、形状分析等；然而，边缘检测方法对噪声敏感，容易受光照变化和物体表面纹理的影响，导致误检或漏检。因此，在实际应用中，通常会结合其他特征（如颜色、纹理）或进行多尺度检测，以增强算法的鲁棒性和精度。在工业领域，基于边缘检测的目标检测广泛应用于产品质量检测、零部件识别和装配检测等任务。通过准确提取物体的边缘信息，

可以实现高精度的目标定位和尺寸测量，从而提高生产过程的自动化和精确度。

2. 基于颜色特征的目标检测

基于颜色特征的目标检测方法利用图像中物体的颜色信息来识别和定位目标。这种方法常用于颜色分明、光照条件稳定的场景。

常用算法有以下两种。

(1) 颜色直方图：基于颜色特征的目标检测中常用的工具，通过统计图像中各个颜色通道的像素分布，形成一个直方图来表示图像的颜色特征。目标检测时，将待检测区域的颜色直方图与模板的颜色直方图进行比较，以判断目标是否存在。

(2) 颜色阈值分割：通过设定颜色的上下阈值，将图像中符合特定颜色范围的像素提取出来，形成二值图像，以此进行目标检测和定位；这种方法在背景单一、目标颜色明显的场景中效果显著，如图 5.6 所示，通过胡萝卜的显著颜色特征进行检测。

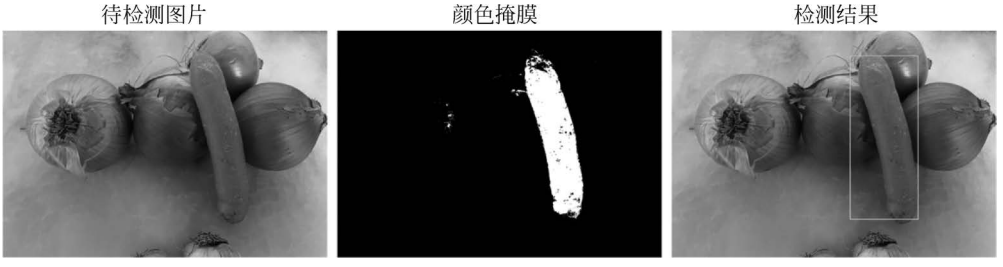

图 5.6　基于颜色阈值分割的胡萝卜检测

基于颜色特征的检测算法实现简单，计算量较小，适合实时处理，因此在工业环境中得到了广泛应用；然而，颜色特征易受到光照变化的影响，导致检测精度下降。因此，在实际应用中，通常需要在光照稳定的环境中使用该类算法，或结合其他特征(如形状、纹理)来增强鲁棒性。例如，在自动化生产线中，基于颜色的检测算法常用于产品分类和分拣，如根据颜色分拣不同类型的果蔬或商品。这种方法不仅能提高分拣效率，还能保证分类结果的一致性。

3. 基于形状模型的目标检测

基于形状模型的目标检测方法依赖于目标的几何形状信息，通过匹配图像中检测到的形状与预定义的模型来识别目标。这类方法特别适用于具有明显几何形状特征的目标物体。

常用算法有以下两种。

(1) 霍夫变换(Hough Transform)：霍夫变换是一种用于检测特定几何形状(如直线、圆形、椭圆等)的经典算法。通过在参数空间中对图像中的边缘点进行投票，霍夫变换可以有效地检测图像中的直线、圆形和其他规则形状。图 5.7 所示为通过圆形检测洋葱目标。

(2) 主动轮廓模型(Active Contour Models)：也称为 Snake 模型，通过让轮廓线在图像中沿着目标物体的边缘收缩或扩展，最终与目标物体的边缘对齐，从而实现目标检测。

基于形状模型的方法可以准确检测出具有特定几何形状的目标，适合于结构明确的物体检测，它在一定程度上应对噪声和部分遮挡，具有较好的鲁棒性；但霍夫变换等方法在高维参数空间中的计算复杂度较高，可能影响实时性。此外，基于形状模型的方法对目标的几何形状有较强的依赖性，不适合形状变化大或形状不规则的目标。

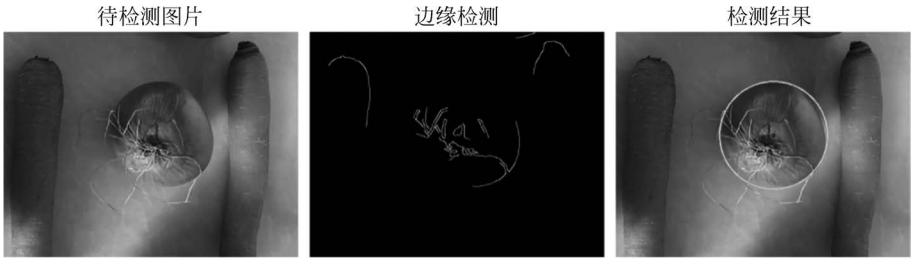

图 5.7　基于圆形的目标检测

在制造业中，基于形状模型的方法可用于检测特定形状的零件或装配件的完整性，如检测圆形零件的缺损；也可用于检测产品是否符合预定的形状规格，如瓶盖是否密封、产品表面是否光滑。

虽然传统目标检测算法在特定任务和应用中表现出色，但它们通常依赖手工设计的特征和启发式规则。随着深度学习技术的发展，基于数据驱动的方法因其自适应性和强大的表征能力，在许多复杂场景中逐渐取代了传统方法；然而，传统算法由于其简洁性、可解释性和在某些特定任务中的高效性，仍然值得学习和研究，尤其是在资源受限或实时性要求高的应用中。

5.3　基于深度学习的目标检测

随着深度学习技术的发展，基于深度学习的目标检测方法已成为计算机视觉领域的主流技术。相比传统的目标检测方法，深度学习算法能够自动学习图像中的复杂特征，实现更加准确和鲁棒的目标检测。基于深度学习的目标检测方法不仅在精度上有了显著提高，还在处理多类目标、复杂背景和尺度变化等挑战上表现出了极大的优势。

深度学习目标检测算法通常基于 CNN，通过端到端的方式进行训练和预测。这类算法可以分为两大类：单阶段检测器（如 YOLO 系列）和两阶段检测器（如 R-CNN 系列）。单阶段检测器以速度见长，适用于实时检测任务；而两阶段检测器则在精度上有优势，适用于要求高精度的场景。此外，随着研究的深入，许多改进的深度学习模型不断涌现，如 DETR、Swin Transformer 等，进一步提升了目标检测的性能。

在工业应用中，基于深度学习的目标检测方法已被广泛应用于产品缺陷检测、智能制造、自动驾驶等领域。例如，在产品质量检测中，深度学习模型可以有效识别细微的缺陷，并对多种类型的缺陷进行分类；在自动驾驶中，目标检测算法用于识别和跟踪行人、车辆、交通标志等，保证行车安全。深度学习的强大特征提取能力使其在处理复杂场景时尤为有效，特别是在应对光照变化、视角变换和遮挡等挑战时表现出色。

接下来将详细介绍几种经典的深度学习目标检测算法系列。

5.3.1　YOLO 系列介绍

YOLO（You Only Look Once）系列是近年来目标检测领域中的一项突破性进展，实现了高效的实时目标检测，并在准确性和速度之间取得了极佳的平衡。YOLO 的出现改变了传统目标检测算法的检测思路，使得基于深度学习的目标检测技术得到了广泛应用。

1. YOLO 的基本结构

YOLO 的核心思想是将目标检测问题视为一个单一的回归问题。与传统的滑动窗口或区域建议方法不同，YOLO 将整个图像一次性输入神经网络中，直接输出图像中各目标的位置和类别。具体来说，如图 5.8 所示，YOLO 的检测流程可以分为以下几个步骤。

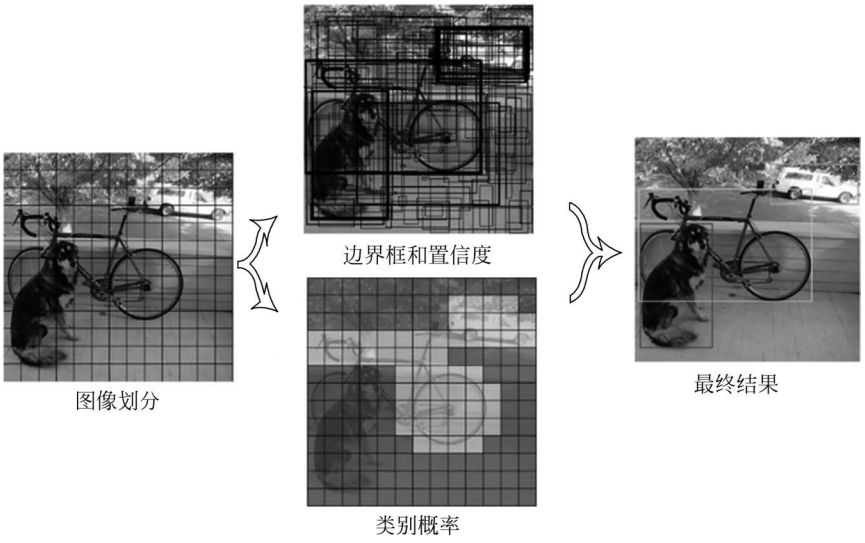

图 5.8　YOLO 的检测流程

1) 图像划分

YOLO 将输入图像划分为 $S\times S$ 的网格，每个网格负责检测其中心落在该网格中的目标；常见的网格大小有 13×13、26×26 和 52×52。这种划分使每个网格能够专注于检测其内部的目标，并减少背景干扰。

2) 边界框预测

(1) 边界框坐标：对于每个网格，YOLO 预测一个或多个边界框的坐标，包括边界框的中心位置(x,y)、宽度 w 和高度 h。这些坐标是相对于整个网格进行归一化处理的，以提高模型的泛化能力。式(5.4)～式(5.7)为具体预测公式

$$b_x=\sigma(t_x)+c_x \tag{5.4}$$

$$b_y=\sigma(t_y)+c_y \tag{5.5}$$

$$b_w=p_w e^{t_w} \tag{5.6}$$

$$b_h=p_h e^{t_h} \tag{5.7}$$

其中，(c_x,c_y)是网格单元的左上角坐标，(t_x,t_y)是模型预测的偏移量，σ 是 sigmoid 函数，(p_w,p_h)是先验边界框的宽度和高度，(t_w,t_h)是模型预测的尺度变化。

(2) 置信度：每个网格还预测一个置信度值，表示该网格中是否包含目标及其边界框的准确性。置信度值是边界框和实际目标的重叠度(IoU)的预测：

$$\text{Confidence}=P(\text{Object})\cdot \text{IoU}_{\text{pred,true}} \tag{5.8}$$

其中，$P(\text{Object})$是目标存在的概率，$\text{IoU}_{\text{pred,true}}$ 是预测边界框与真实边界框的重叠度。

(3) 类别概率：YOLO 还预测每个边界框对应的目标类别概率分布。假设有 C 个类别，每个网格会预测一个 C 维向量，表示该网格内目标属于每个类别的概率：

$$P(C_1),P(C_2),\cdots,P(C_C) \tag{5.9}$$

其中，$P(C_j)$表示边界框属于第 j 类的概率。

3) 最终检测

在模型的输出中，对于每个网格，YOLO 生成的最终检测框是结合边界框坐标、置信度和类别概率进行的。式(5.10)为最终检测框的格式。

$$(b_x,b_y,b_w,b_h,\text{Confidence},P(C_1),P(C_2),\cdots,P(C_C)) \tag{5.10}$$

为了去除重叠的边界框，YOLO 会应用非极大值抑制算法。具体步骤包括：根据置信度值对所有检测框进行降序排序；选择置信度最高的检测框作为当前最佳检测框；计算当前最佳检测框与其他检测框的 IoU，如果 IoU 大于某个阈值(如 0.5)，则抑制(即移除)其他检测框；重复上述步骤，直到所有检测框都被处理完毕。

YOLO 的优点在于其检测速度非常快，适用于实时应用；然而，由于 YOLO 模型将目标检测问题简化为回归任务，可能会在处理小目标或密集目标时遇到挑战。因此，YOLO 系列在不断演进中，后续版本如 YOLOv2、YOLOv3、YOLOv4、YOLOv5 和 YOLOv8 等不断改进网络结构、增加特征提取能力，提升了检测精度和鲁棒性，使其在多种实际应用场景中得到了广泛的应用。

2. YOLO 版本演进

YOLO 系列自 2016 年首次提出以来，经过了多次改进与优化，每个版本都在性能上有显著提升，如图 5.9 所示。以下是 YOLO 系列的主要版本及其特性。

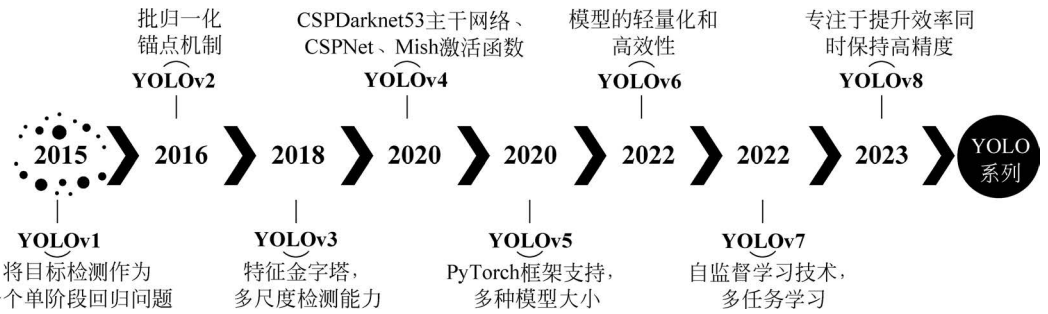

图 5.9　YOLO 系列的发展历程

1) YOLOv2

YOLOv2 在 YOLOv1 的基础上进行了多项改进，显著提高了模型的精度和稳定性。首先，YOLOv2 引入了批归一化(Batch Normalization)，加速了模型的收敛速度并提高了稳定性；其次，YOLOv2 引入了锚点机制(Anchor Boxes)，每个网格预测多个预定义的边界框，提高了对不同形状和大小目标的检测能力，锚点的尺寸和比例是根据训练数据中的真实边界框统计得到的；此外，YOLOv2 采用了多尺度训练(Multi-Scale Training)，提高了模型对不同尺寸图像的适应能力；接着，YOLOv2 还提出了 YOLO9000，能够在训练时结合 COCO 和 ImageNet 数据集，实现大规模检测任务，支持超过 9000 个类别的检测；数据增强方面，YOLOv2 引入了随机缩放和平移的方法，增强了模型的泛化能力。

2）YOLOv3

YOLOv3 在 YOLOv2 的基础上进一步优化了模型结构和训练方法，显著提高了对小目标的检测能力。YOLOv3 引入了特征金字塔网络(FPN)，在不同尺度的特征图上进行预测，提高了对小目标的检测能力；YOLOv3 在三个不同尺度的特征图上进行预测，分别为 13×13、26×26 和 52×52，分别对应大、中、小目标，提高了多尺度检测能力；数据增强方面，YOLOv3 引入了 Mosaic 数据增强方法，将四幅图像拼接成一张，增加了对复杂场景的适应能力。YOLOv3 还使用了改进的 Focal Loss，解决了正负样本不平衡的问题，提高了检测精度。

3）YOLOv4

YOLOv4 在 YOLOv3 的基础上进行了大量优化，引入了许多先进的技术和组件，进一步提高了模型的检测精度和速度。YOLOv4 引入了 CSPDarknet53 作为主干网络，通过跨阶段局部网络(CSPNet)减少了计算量，提高了模型的效率；YOLOv4 使用了 Mish 激活函数，提高了模型的非线性表达能力，增强了泛化能力；数据增强方面，YOLOv4 引入了 CutMix 和 Mosaic 数据增强方法，增加了对复杂场景的适应能力；YOLOv4 还使用了改进的 CIoU 损失函数，提高了边界框回归的精度。

4）YOLOv5

YOLOv5 虽然并非由 YOLO 系列的原作者团队发布，但因其简单易用、支持多种框架(如 PyTorch)以及出色的性能而广受欢迎。YOLOv5 简化了模型结构和训练流程，提供了多种模型大小(如 Small、Medium、Large)，适应了不同计算资源和应用场景的需求；YOLOv5 使用了改进的 CSPDarknet53 作为主干网络，结合了 PANet 和 SPP 模块，提高了特征提取能力；数据增强方面，YOLOv5 进一步优化了 Mosaic 和 MixUp 方法，增加了对复杂场景的适应能力；YOLOv5 还引入了自动混合精度训练(AMP)和模型蒸馏等技术，提高了训练效率和模型性能。

5）YOLOv6

YOLOv6 专注于模型的轻量化和高效性，适合在边缘设备和嵌入式系统上部署。YOLOv6 引入了模型剪枝和量化技术，减少了模型的参数量和计算量，使模型能够在资源受限的环境中高效运行；YOLOv6 使用了更轻量的主干网络，如 EfficientNet 或 MobileNet，结合了轻量级的颈部和头部模块，提高了特征提取能力；数据增强方面，YOLOv6 引入了更高效的数据增强方法，如 RandAugment，增加了对复杂场景的适应能力；YOLOv6 还优化了卷积层和池化层的设计，减少了计算冗余，提高了模型的效率。

6）YOLOv7

YOLOv7 在 YOLOv6 的基础上进一步优化了模型结构和训练方法，引入了自监督学习技术，提高了模型的泛化能力和鲁棒性。YOLOv7 支持多任务学习，可以在同一模型中同时进行目标检测、实例分割和关键点检测等任务，这使得模型在多任务场景下表现出色；YOLOv7 使用了更复杂的主干网络，如 ResNet 或 DenseNet，结合了多尺度特征融合模块，提高了特征提取能力；数据增强方面，YOLOv7 引入了更高级的数据增强方法，如 StyleTransfer 和 ColorJitter，增加了对复杂场景的适应能力；YOLOv7 还引入了动态锚点生成和自适应锚点匹配等技术，提高了训练效率和模型性能。

7）YOLOv8

YOLOv8 对整个 YOLO 体系进行了重大重构，解决了之前版本中的一些根本性问题，如模型复杂度和训练难度。YOLOv8 引入了更高效的训练策略，如渐进式学习和自适应学习率调度，提高了模型的训练效率和精度；YOLOv8 支持多模态输入，如 RGB-D 图像和 LiDAR 点云，这使得模型在复杂环境下的适应能力更强；数据增强方面，YOLOv8 引入了更高级的数据增强方法，如 StyleTransfer 和 ColorJitter，增加了对复杂场景的适应能力；YOLOv8 还优化了卷积层和池化层的设计，减少了计算冗余，提高了模型的效率。

3. YOLO 的优点与缺点

YOLO 的单次前向传递设计使其检测速度非常快，适合实时性要求高的应用场景，如自动驾驶、安防监控和工业检测等；此外，YOLO 的检测过程是端到端的，简化了训练和推理流程，不需要复杂的后处理步骤，易于部署和优化。由于 YOLO 在检测过程中考虑了全局上下文信息，这有助于减少背景误检和目标重叠检测的情况。

尽管 YOLO 具有较高的检测速度，但由于其基于全局回归的方法，定位精度可能不如基于区域的检测方法（如 Faster R-CNN）；尤其在处理小目标和目标密集的场景时，YOLO 的表现可能逊色。YOLO 的网格划分限制了其对小目标的检测能力，虽然后续版本通过多尺度预测进行了改进，但在极端情况下仍可能出现漏检。

YOLO 系列模型在实际应用中具有广泛的适用性和灵活性，特别是在实时检测需求较高的场景中表现优异。例如，在工业生产线中，YOLO 可以用于自动化的产品检测和分类，实时识别和定位产品缺陷，从而提高生产效率和质量控制。尽管存在一些局限性，YOLO 系列模型的快速发展和优化仍使其成为目标检测领域的重要工具。

5.3.2　R-CNN 系列

R-CNN（Region-based Convolutional Neural Networks）系列是目标检测领域中的经典方法，它通过结合区域建议与 CNN 的强大特征提取能力，开创了基于深度学习的目标检测的新篇章。R-CNN 系列包括多种改进版本，如 R-CNN、Fast R-CNN、Faster R-CNN 及 Mask R-CNN 等，每个版本都在前一个版本的基础上优化了检测速度和精度。

1. R-CNN 的基本结构

R-CNN 模型将目标检测问题分为两个主要步骤：首先生成一组潜在的目标区域（即区域建议），然后对这些区域进行分类和边界框回归。如图 5.10 所示，R-CNN 模型的主要工作流程包括以下几个步骤。

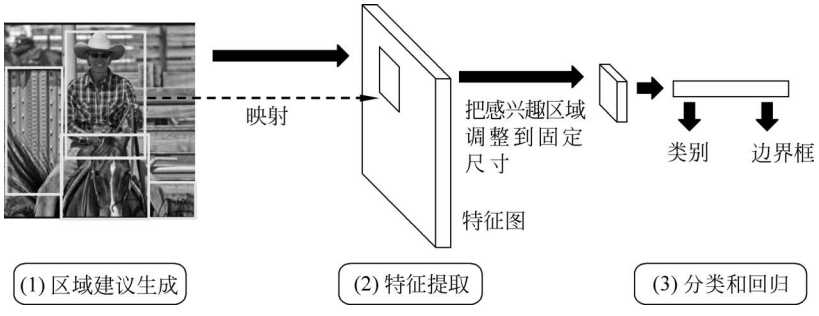

图 5.10　R-CNN 的主要工作流程

1）区域建议生成

R-CNN 使用选择性搜索(Selective Search)算法生成一组候选区域，这些区域可能包含目标物体。选择性搜索通过多尺度分割图像并合并相邻区域，最终得到数千个候选框。选择性搜索具体包括以下步骤。

(1) 多尺度分割：将图像分割成多个小区域，每个区域可能包含不同的对象或背景；

(2) 层次合并：根据颜色、纹理、大小和形状等特征，逐步合并相似的区域，形成更大的候选区域；

(3) 候选区域生成：最终生成数千个候选区域，这些区域可能覆盖图像中的所有目标，假设输入图像中生成的候选区域为(R_1，R_2，R_3，…，R_n)，这些区域可能覆盖图像中的所有目标。

2）特征提取

对每个候选区域，R-CNN 通过预训练的卷积神经网络(如 AlexNet 或 VGG)提取特征。首先从原始图像中裁剪出每个候选区域，再将裁剪出的候选区域调整为卷积神经网络所需的固定尺寸(例如 227×227 像素，适用于 AlexNet)，然后将调整后的候选区域输入预训练的卷积神经网络，提取特征。假设第 i 个候选区域对应的特征表示为 $f(R_i)$，这表示从候选区域中提取出的卷积特征向量。这个过程将每个候选区域转换为固定大小的特征表示，使得后续的分类和回归任务更加简单。

3）目标分类与边界框回归

(1) 目标分类：对于每个候选区域提取的特征，R-CNN 通过一组全连接层进行目标分类。具体步骤如下。

全连接层：将提取的特征向量 $f(R_i)$ 输入到全连接层，进行特征的进一步处理。

分类输出：全连接层的输出是一个类别概率分布，表示该候选区域属于各个类别的概率，假设共有 C 个类别，输出的类别概率分布为 $P(C_1)$，$P(C_2)$，…，$P(C_C)$。

(2) 边界框回归：同时，R-CNN 还预测该区域的边界框位置。具体步骤如下。

全连接层：将提取的特征向量 $f(R_i)$ 输入另一组全连接层，进行边界框的回归。

回归输出：全连接层的输出是 4 个值(Δx，Δy，Δw，Δh)，表示预测边界框相对于候选区域的偏移量。最终的边界框计算公式如式(5.11)～式(5.14)所示。

$$\hat{x} = x_{\text{proposal}} + w_{\text{proposal}} \cdot \Delta x \tag{5.11}$$

$$\hat{y} = y_{\text{proposal}} + h_{\text{proposal}} \cdot \Delta y \tag{5.12}$$

$$\hat{w} = w_{\text{proposal}} \cdot \mathrm{e}^{\Delta w} \tag{5.13}$$

$$\hat{h} = h_{\text{proposal}} \cdot \mathrm{e}^{\Delta h} \tag{5.14}$$

其中，x_{proposal}、y_{proposal}、w_{proposal}、h_{proposal} 是候选区域的坐标和尺寸，$\hat{x}$、$\hat{y}$、$\hat{w}$、$\hat{h}$ 是最终预测的边界框坐标和尺寸。

为了去除重叠的边界框，R-CNN 类似 YOLO 也会应用非极大值抑制算法。

2. R-CNN 系列的改进

1）Fast R-CNN

Fast R-CNN 是 R-CNN 的第一个主要改进，通过共享卷积特征图显著提高了检测速度。与原始 R-CNN 相比，Fast R-CNN 不再对每个候选区域单独计算特征，而是在整幅图

像上执行一次卷积操作，生成一个共享的特征图；这不仅减少了计算冗余，还降低了存储需求。Fast R-CNN 的结构如图 5.11 所示。生成的特征图用于所有后续的计算，从而大幅提高了效率。Fast R-CNN 引入了 RoI(Region of Interest，RoI)池化层，对每个候选区域进行固定大小的特征提取，无论候选区域的大小如何。通过 RoI 池化层，Fast R-CNN 可以将不同大小的候选区域转换为固定大小的特征图，确保输入全连接层的特征图大小一致；此外，Fast R-CNN 采用了多任务损失函数，将分类和边界框回归联合训练。这种联合训练方式提高了检测精度，使模型能够更准确地预测目标的位置和类别。

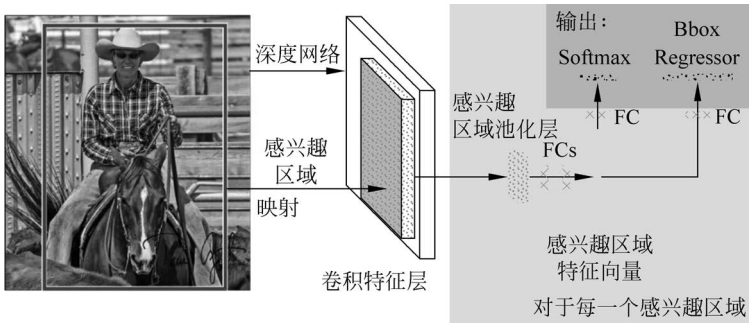

图 5.11　Fast R-CNN 结构

2) Faster R-CNN

Faster R-CNN 的出现进一步提升了目标检测的速度与精度，结构如图 5.12 所示。Faster R-CNN 的核心创新是引入了区域建议网络(Region Proposal Network，RPN)，用于直接在共享的卷积特征图上生成候选区域。RPN 通过滑动窗口机制，对特征图进行遍历，生成一组锚点(Anchors)，并对这些锚点进行分类和回归，输出可能包含目标的候选区域。这种方法不仅消除了对外部区域建议算法(如选择性搜索)的依赖，还将区域建议的生成与目标检测整合为一个端到端的训练过程。RPN 和目标检测网络共享同一个卷积特征图，减少了重复计算，提高了效率。RPN 生成的候选区域直接用于后续的目标检测任务，避免了外部区域建议算法的计算开销。Faster R-CNN 的多任务损失函数包括 RPN 的分类和回归损失，以及目标检测网络的分类和边界框回归损失，这些任务在训练过程中同时优化；Faster R-CNN 的这些改进使得它在大规模数据集上表现出色，适用于需要处理大量图像和视频的场景，同时也适用于实时检测任务。

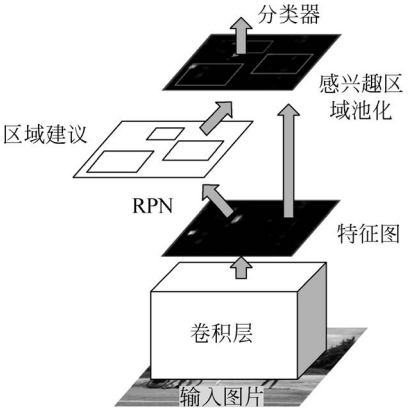

图 5.12　Faster R-CNN 结构

3) Cascade R-CNN

Cascade R-CNN 是一种多阶段的检测框架，通过逐级提高 IoU(Intersection over Union)的阈值来增强检测精度，结构如图 5.13 所示。传统的目标检测模型通常使用固定的 IoU 阈值来区分正样本和负样本，而 Cascade R-CNN 则通过逐步提高这个阈值，使模型在每一阶段都更加专注于难以检测的目标。每个阶段的检测器都使用前一阶段的输出作

为输入，并逐步提高检测精度。这种方法特别适合处理那些边界模糊、尺寸变化大的目标。Cascade R-CNN 由多个检测器组成，每个检测器负责一个阶段的检测任务；每个阶段的检测器都使用更高的 IoU 阈值，逐步提高检测精度；每个阶段的检测器都对前一阶段的输出进行优化，逐步提高边界框的精度和分类的准确性；每个阶段的检测器都使用多任务损失函数，优化分类和边界框回归。通过这种多阶段的设计，Cascade R-CNN 能够在保持较高检测速度的同时，显著提高检测精度。

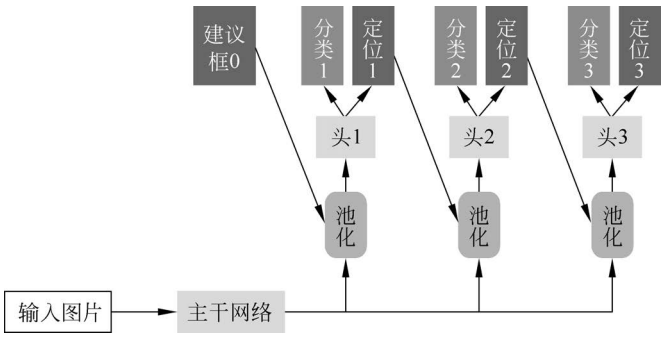

图 5.13　Cascade R-CNN 结构

3. R-CNN 的优点与缺点

R-CNN 系列方法通过区域建议和深度特征提取，能够在复杂背景和拥挤场景中实现高精度的目标检测。尤其是 Faster R-CNN 和 Mask R-CNN，能够在多个目标检测基准数据集上取得领先的性能。此外，R-CNN 系列模型可以很容易地扩展到其他任务，如实例分割、关键点检测等，这使得它们在不同的计算机视觉任务中具有广泛的应用。

由于原始 R-CNN 需要对每个候选区域独立提取特征，计算代价非常高。尽管 Fast R-CNN 和 Faster R-CNN 通过共享卷积特征图显著降低了计算复杂度，但其检测速度仍然不及 YOLO 和 SSD 等单阶段检测器。此外，由于计算复杂度较高，R-CNN 系列模型在实时检测任务中可能表现不佳，尤其是在需要处理高分辨率图像或视频的应用场景中。

R-CNN 系列模型在实际应用中同样具有广泛的适用性，尤其是在高精度检测需求较高的场景中表现出色。例如，在工业检测中，R-CNN 可以用于复杂背景下的目标识别和定位，精确检测产品的缺陷和异常，从而提高产品质量和生产线的可靠性。尽管 R-CNN 系列模型的计算复杂度较高，可能不适合实时检测任务，但其在精度和灵活性方面的优势使其成为工业视觉中高要求检测任务的重要工具。

5.3.3　其他目标检测方法

除了较为经典的 YOLO 和 R-CNN，之后也涌现了一些其他有效的目标检测方法。这些方法通过不同的网络结构、训练策略和优化手段，进一步提升了检测的速度、精度和鲁棒性。以下介绍几种具有代表性的目标检测方法。

1. SSD

SSD(Single Shot MultiBox Detector)是一种单阶段目标检测方法，它结合了 YOLO 的快速检测能力和 Faster R-CNN 的精确定位能力。SSD 的核心思想是将目标检测任务划分为多个尺度上的检测，以同时处理大目标和小目标，从而提高检测的精度。

1）模型架构

SSD使用一个主干网络(如VGG16)来提取输入图像的特征。主干网络通常是一个预训练的卷积神经网络,能够提取图像的高层次特征,如图5.14所示。SSD在多个尺度的特征图上进行目标检测,通过在不同层次的特征图上应用卷积核,SSD能够检测不同大小的目标。这些特征图通常是从主干网络的不同层提取的,每一层的特征图负责检测不同大小的目标。

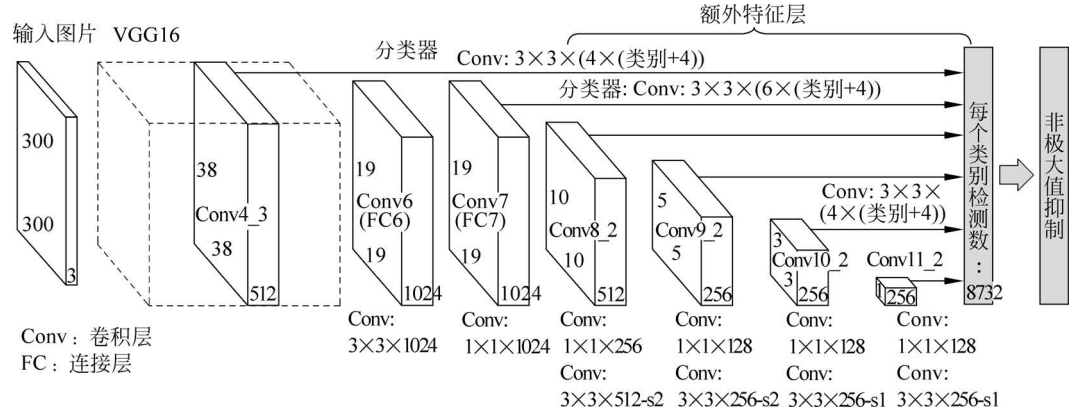

图5.14　SSD网络结构

2）目标检测

(1) 候选边界框生成。

在每个尺度的特征图上,SSD为每一像素位置生成多个默认框(Default Boxes),也称为先验框(Prior Boxes)。这些默认框有不同的比例以适应不同大小和形状的目标。假设在特征图上的每一像素位置生成 k 个默认框。

(2) 边界框和类别预测。

边界框预测：对于每个默认框,SSD类似R-CNN预测其相对于默认框的偏移量,包括边界框的中心位置(x,y)、宽度 w 和高度 h。

类别预测：对于每个默认框,SSD也还预测其属于各个类别的概率。

SSD通过在多尺度的特征图上直接预测目标类别和边界框,实现了高效的单阶段目标检测。其主要优势在于结合了YOLO的快速检测能力和Faster R-CNN的精确定位能力,能够在保持高检测速度的同时,实现较为准确的目标定位和分类；通过多尺度检测、默认框和多任务损失函数等机制,SSD能够有效地处理不同大小的目标,提高了检测的鲁棒性和精度。

2. DETR

DETR(Detection Transformer)是将Transformer模型引入目标检测的创新方法。其核心思想是将目标检测问题转化为一个序列预测问题,通过Transformer的自注意力机制捕捉图像中的全局信息,从而实现高效且准确的目标检测。具体来说,如图5.15所示,DETR的工作流程如下。

输入图像首先通过一个卷积神经网络(如ResNet)提取特征,生成特征图 $\boldsymbol{F}$。特征图的形状为 $H\times W\times C$,其中 H 和 W 是特征图的高度和宽度,C 是通道数；然后特征图 $\boldsymbol{F}$ 被展平为一维向量序列 $\boldsymbol{X}$,每个向量表示特征图的一个位置；这些向量通过Transformer编码器生成高维特征表示 $\boldsymbol{E}$；Transformer编码器通过自注意力机制(Self-Attention)捕捉特征

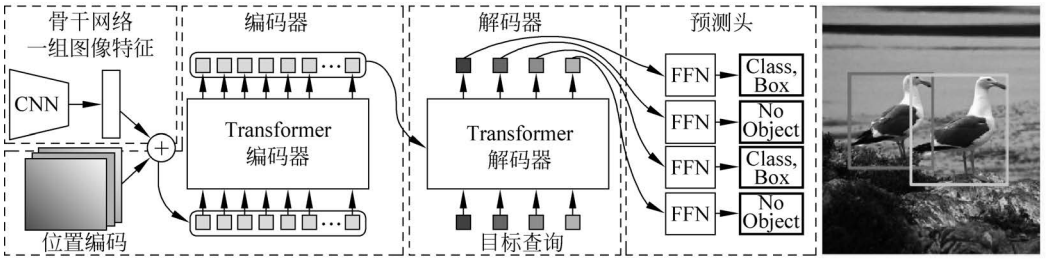

图 5.15 DETR 的工作流程

图中的全局依赖关系，生成的高维特征表示 $\boldsymbol{E}$ 为后续的解码器提供了丰富的上下文信息。

DETR 还引入了一组固定数量的目标查询向量 $\boldsymbol{Q}$，每个查询向量表示一个潜在的目标。这些查询向量通过 Transformer 解码器与编码器生成的特征表示 $\boldsymbol{E}$ 进行交互，生成最终的预测结果；每个目标查询向量 $\boldsymbol{Q}_i$ 预测一个目标的类别概率 $\boldsymbol{p}_c$ 和边界框坐标 $\boldsymbol{b}$。

为了优化模型，DETR 通过二元匹配损失(Bipartite Matching Loss)将预测结果与真实标签进行匹配。具体是使用匈牙利算法(Hungarian Algorithm)将预测的目标与真实的目标进行最优匹配，匹配的目标是最小化预测边界框与真实边界框之间的总代价，通常使用边界框回归损失(如 L1 损失)和分类损失(如交叉熵损失)的加权和。

DETR 的优势在于通过 Transformer 的自注意力机制，能够捕捉图像中的全局信息，提高检测的准确性和鲁棒性；此外，模型结构相对简单，不需要复杂的后处理步骤，如非极大值抑制(NMS)，并且可以端到端地进行训练，简化了模型的训练和推理过程。然而，由于 Transformer 的自注意力机制计算复杂度较高，DETR 的训练时间相对较长，且需要较多的计算资源，尤其是在处理大规模数据集时。

3. Swin Transformer

Swin Transformer 是基于视觉 Transformer 的一种高效目标检测方法。它通过引入层次化的特征表示和滑动窗口机制，显著提高了视觉 Transformer 的效率和精度。Swin Transformer 的设计旨在克服标准 Transformer 在处理高分辨率图像时的计算复杂度问题，同时保持对局部和全局信息的有效捕捉，整体结构如图 5.16 所示。

首先，Swin Transformer 将输入图像划分为不重叠的小块(Patches)，每个小块的大小通常为 4×4 像素或 8×8 像素；这些小块被展平并线性投影到一个固定维度的向量，形成初始的特征表示。接下来，Swin Transformer 采用分层结构，每层特征图的分辨率逐渐降低。这种分层结构类似于传统的卷积神经网络，但使用 Transformer 的自注意力机制来提取特征；通过这种方式，Swin Transformer 能够生成多尺度的特征表示，适用于目标检测、语义分割等任务。此外，与标准 Transformer 不同，Swin Transformer 通过滑动窗口(Shifted Window)机制在跨块区域之间引入自注意力计算，捕获图像的局部和全局信息。具体来说，滑动窗口机制允许在相邻的小块之间共享信息，从而提高模型的表达能力。标准窗口内，自注意力机制仅在当前窗口内的小块之间进行计算；而在滑动窗口机制中，窗口在特征图上滑动，使得相邻窗口之间的信息能够相互作用。通过这种方式，Swin Transformer 能够在不同尺度上捕捉局部和全局的依赖关系。最终，通过分层结构和滑动窗口机制，Swin Transformer 生成多尺度的特征表示；这些特征表示在不同尺度上捕捉了图像的局部和全局信息，适用于各种计算机视觉任务，如目标检测和语义分割。

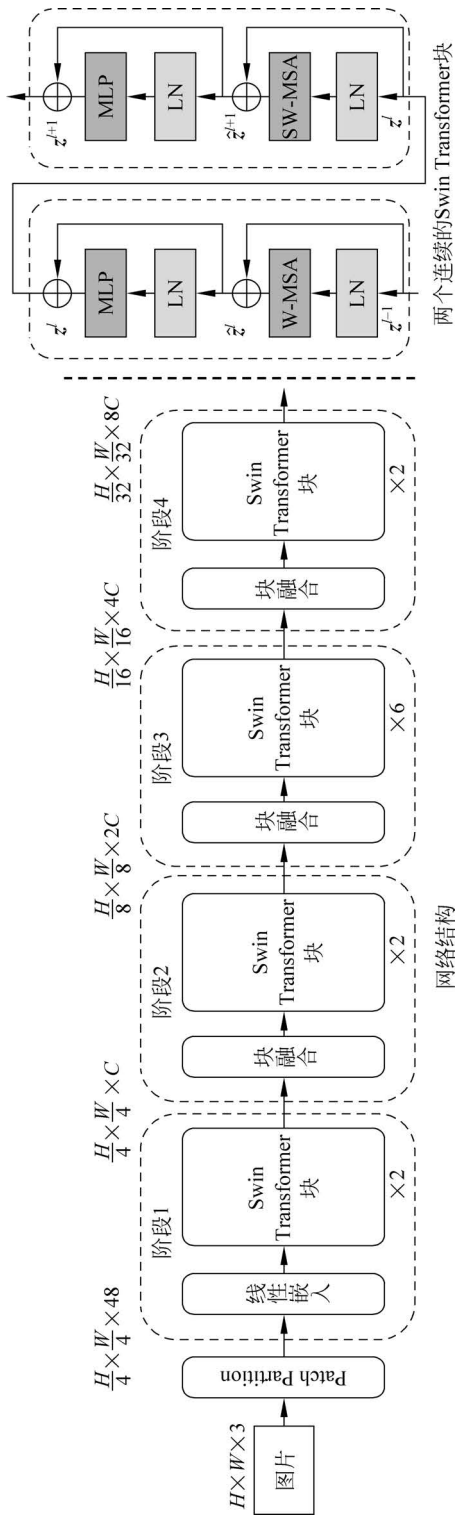

图 5.16　Swin Transformer 网络结构

在生成多尺度的特征表示后，Swin Transformer 通过一系列的卷积层和全连接层来输出检测框。Swin Transformer 生成的多尺度特征图被传递到一个检测头（Detection Head）；检测头通常包含几个卷积层和全连接层，用于提取更高级的特征；在每个尺度的特征图上，检测头生成一组候选框（Anchor Boxes）；这些候选框是预先定义的边界框，具有不同的尺度和长宽比，用于覆盖图像中的潜在目标；对于每个候选框，检测头预测其相对于候选框的偏移量；对于每个候选框，检测头还预测其属于各个类别的概率。这与 5.3.1 节和 5.3.2 节提到的基本一致。

总体来说，Swin Transformer 通过引入层次化的特征表示和滑动窗口机制，显著提高了视觉 Transformer 的效率和精度。Swin Transformer 通过检测头生成候选框，进行边界框回归和类别预测，并通过非极大值抑制（NMS）去除重叠的边界框，最终输出检测结果。Swin Transformer 在处理高分辨率图像时表现出色，适用于目标检测、语义分割等任务，具有计算效率高、强泛化能力和局部与全局信息捕捉能力强等优势。

上述这些目标检测方法各有优劣，如 DETR 之类的结合了经典的卷积神经网络和创新的 Transformer 结构的方法，为工业视觉中的目标检测任务提供了多样化的解决方案。在工业生产中，这些方法能够有效地识别和定位产品缺陷、检测生产线上的异常情况，以及分类和分拣不同类型的产品，展现出了卓越的实用性和可靠性。

5.4 工业目标跟踪技术

目标跟踪技术在工业机器视觉中同样占据着重要位置，广泛应用于自动化生产线、工业机器人、智能监控系统等多个领域；目标跟踪的主要任务是实时地识别并跟踪运动中的物体，以确保对物体位置的准确掌握和实时更新。下面将目标跟踪技术分为传统方法和基于深度学习的方法进行介绍。

5.4.1 传统目标跟踪方法

传统的目标跟踪方法主要依赖于图像处理和计算机视觉的经典算法。这些方法通过分析图像序列中的目标特征，实现对目标的跟踪。经典的跟踪方法包括光流法、均值漂移（Mean Shift）和粒子滤波（Particle Filter）等。

1. 光流法

光流法是一种基于运动估计的跟踪技术，用于估计图像序列中像素点的运动矢量，从而跟踪目标。光流法假设目标在相邻帧之间的运动是平滑且小范围的，通过求解图像亮度恒定假设方程来估计像素的位移。

1）基本原理

光流法的基本假设是图像的亮度恒定，即在时间 t 时刻的像素点在时间 $t+\Delta t$ 时刻，其亮度值保持不变。这个假设可以用以下方程表示

$$\boldsymbol{I}(x,y,t)=\boldsymbol{I}(x+u\Delta t,y+v\Delta t,t+\Delta t) \tag{5.15}$$

其中，$\boldsymbol{I}(x,y,t)$ 表示在时间 t 时刻位置 (x,y) 处的像素亮度，u 和 v 分别表示像素在 x 和 y 方向上的速度分量。然后，将上述方程在时间和空间上进行泰勒展开，并忽略高阶项，可以得到光流基本方程

$$\frac{\partial \boldsymbol{I}}{\partial x}u+\frac{\partial \boldsymbol{I}}{\partial y}v+\frac{\partial \boldsymbol{I}}{\partial t}=0 \tag{5.16}$$

其中，$\frac{\partial \boldsymbol{I}}{\partial x}$和$\frac{\partial \boldsymbol{I}}{\partial y}$分别是图像在 x 和 y 方向上的梯度，$\frac{\partial \boldsymbol{I}}{\partial t}$是图像在时间上的变化率，$u$ 和 v 分别表示像素在 x 和 y 方向上的速度分量。

方程求解：由于一个方程不足以求解两个未知数 u 和 v，光流法通常需要在一个小区域内应用平滑约束来求解该方程组。一种著名的方法是 Lucas-Kanade 方法。该方法在小窗口内假设运动是一致的，通过最小二乘法求解。

在小窗口内，假设所有像素的运动是一致的。对于窗口内的每一像素，可以写出光流基本方程

$$\frac{\partial \boldsymbol{I}}{\partial x}u+\frac{\partial \boldsymbol{I}}{\partial y}v+\frac{\partial \boldsymbol{I}}{\partial t}=0 \tag{5.17}$$

对于窗口内的 N 像素，可以构建以下方程组

$$\begin{bmatrix} \boldsymbol{I}_x^1 & \boldsymbol{I}_y^1 \\ \boldsymbol{I}_x^2 & \boldsymbol{I}_y^2 \\ \vdots & \vdots \\ \boldsymbol{I}_x^N & \boldsymbol{I}_y^N \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \boldsymbol{I}_t^1 \\ \boldsymbol{I}_t^2 \\ \vdots \\ \boldsymbol{I}_t^N \end{bmatrix} \tag{5.18}$$

其中，$\boldsymbol{I}_x^i$、$\boldsymbol{I}_y^i$ 和 $\boldsymbol{I}_t^i$ 分别是第 i 像素在 x、y 和时间方向上的梯度。

通过最小二乘法求解上述方程组，得到 u 和 v 的最优解

$$\begin{bmatrix} u \\ v \end{bmatrix} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b} \tag{5.19}$$

其中，$\boldsymbol{A}$ 是梯度矩阵，$\boldsymbol{b}$ 是时间变化率向量

$$\boldsymbol{A}=\begin{bmatrix} \boldsymbol{I}_x^1 & \boldsymbol{I}_y^1 \\ \boldsymbol{I}_x^2 & \boldsymbol{I}_y^2 \\ \vdots & \vdots \\ \boldsymbol{I}_x^N & \boldsymbol{I}_y^N \end{bmatrix}, \quad \boldsymbol{b}=\begin{bmatrix} \boldsymbol{I}_t^1 \\ \boldsymbol{I}_t^2 \\ \vdots \\ \boldsymbol{I}_t^N \end{bmatrix} \tag{5.20}$$

2）方法优点与缺点

光流法适用于目标运动平稳且背景较为简单的场景，如图 5.17 所示效果。由于光流法假设目标在相邻帧之间的运动是平滑且小范围的，因此在目标运动缓慢且连续的场景中表现较好。在背景较为简单且变化不大的场景中，光流法也能够更准确地估计像素的运动矢量。

图 5.17　光流法效果

然而尽管光流法在许多场景中表现出色，但也存在一些局限性。当目标快速运动时，亮度恒定假设可能不再成立，导致光流法的估计误差增大。此外，当目标部分被遮挡时，光流法无法准确估计被遮挡区域的运动矢量；当场景中存在显著的光照变化时，亮度恒定假设失效，也影响光流法的准确性。

2. 均值漂移法

均值漂移(Mean Shift)是一种基于直方图匹配的非参数统计方法，用于在多维空间中寻找数据集的模态。在目标跟踪中，均值漂移通过计算目标区域的颜色直方图，并在当前帧中搜索与目标直方图最相似的区域，实现目标的定位。

1) 基本原理

均值漂移法的基本思想是通过迭代更新目标位置，使目标区域的密度函数(通常为颜色直方图的核密度估计)达到局部最大值。具体来说，均值漂移法通过不断移动目标的位置，使目标区域的密度函数值逐渐增加，最终收敛到密度的模态。

假设给定目标的核密度估计为

$$\hat{f}(x)=\frac{1}{nh^{d}}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \tag{5.21}$$

其中，x_i 是样本点，K 是核函数，h 是带宽参数，n 是样本点的数量，d 是数据的维度。核密度估计用于估计目标区域的密度分布。常用的核函数包括高斯核、均匀核等。

然后通过迭代更新目标的位置，直至收敛到密度的模态。更新公式为

$$m(x)=\frac{\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)x_i}{\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)} \tag{5.22}$$

在每次迭代中，将目标的位置移动到密度最大的方向，即

$$x_{\text{new}}=m(x_{\text{old}}) \tag{5.23}$$

在目标跟踪中，均值漂移法通过计算目标区域的颜色直方图，并在当前帧中搜索与目标直方图最相似的区域，实现目标的定位，跟踪示例如图 5.18 所示。具体流程包括：在第一帧中手动选择目标区域，并计算该区域的颜色直方图；在当前帧中，使用核密度估计计算目标区域的密度分布；通过均值漂移法迭代更新目标的位置，直到收敛到密度的模态；在当前帧中找到与目标直方图最相似的区域，实现目标的定位。

图 5.18　均值漂移法结果示例

2) 方法优点与缺点

均值漂移法具有计算简单、收敛快的优点，适用于目标外观稳定、背景简单的场景。然而，当目标的外观发生变化(如形状、颜色等)时，均值漂移法的效果会下降。此外，在背景

复杂或存在相似颜色的干扰时，方法容易陷入局部最优解，导致跟踪失败。

3. 粒子滤波

粒子滤波(Particle Filter)是一种基于蒙特卡洛方法的递推贝叶斯估计，用于在噪声和非线性条件下估计动态系统的状态。在目标跟踪中，粒子滤波通过一组随机样本(粒子)表示目标的可能位置，并在每个时间步更新粒子分布，从而实现对目标的跟踪。

1) 基本原理

粒子滤波的基本思想是通过一组随机样本(粒子)来近似表示目标的状态分布。每个粒子代表目标的一个可能状态，通过迭代更新粒子的分布，逐步逼近目标的真实状态。粒子滤波在处理非线性和非高斯分布的目标跟踪问题上具有很强的适应性，特别适用于多模态分布的复杂场景。

粒子滤波的基本步骤包括预测、更新和重采样，具体如下。

(1) 预测。

根据系统的运动模型对粒子进行预测。假设在时间 $t-1$ 时刻，有 N 个粒子 $\{x_{t-1}^{(i)}\}_{i=1}^{N}$，每个粒子表示目标的一个可能状态。在时间 t 时刻，根据状态转移函数 $f(\cdot)$ 和过程噪声 $w_t^{(i)}$，预测每个粒子的新状态

$$x_t^{(i)} = f(x_{t-1}^{(i)}, w_t^{(i)}) \tag{5.24}$$

其中，$x_t^{(i)}$ 是第 i 个粒子的预测状态。

(2) 更新。

根据观测模型更新粒子的权重。假设在时间 t 时刻，观察到的数据为 y_t，每个粒子的权重 $w_t^{(i)}$ 通过观测模型的似然函数 $p(y_t \mid x_t^{(i)})$ 计算

$$w_t^{(i)} \propto p(y_t \mid x_t^{(i)}) \tag{5.25}$$

其中，$p(y_t \mid x_t^{(i)})$ 是观测模型的似然函数，表示在给定粒子状态 $x_t^{(i)}$ 下观察到数据 y_t 的概率。

(3) 重采样。

通过重采样步骤从新的粒子分布中采样，得到当前的目标状态估计。重采样的目的是消除权重较低的粒子，保留权重较高的粒子，从而避免粒子退化。常见的重采样方法包括有放回的随机采样(Multinomial Resampling)、系统重采样(Systematic Resampling)和剩余重采样(Residual Resampling)等。重采样后，所有粒子的权重重新归一化为 $1/N$。

粒子滤波跟踪效果示例如图 5.19 所示。

图 5.19　粒子滤波跟踪效果示例

2）方法优点与缺点

粒子滤波在处理非线性和非高斯分布的目标跟踪问题上具有很强的适应性，特别适用于多模态分布的复杂场景；粒子滤波可以灵活地选择不同的状态转移函数和观测模型，适用于多种目标跟踪任务；通过重采样步骤，粒子滤波能够有效地处理噪声和不确定性，提高跟踪的鲁棒性。然而，粒子滤波的计算复杂度较高，特别是当粒子数量较大时，计算开销会显著增加；在某些情况下，粒子可能会集中在少数几个高权重的粒子上，导致粒子退化现象，影响跟踪性能。

5.4.2 基于相关滤波的跟踪算法

相关滤波(Correlation Filter)是一类用于目标跟踪的高效算法，因其计算速度快、实现简单而受到广泛关注。相关滤波的核心思想是通过在训练阶段学习一个滤波器模型，该滤波器在目标区域与其周围背景之间进行区分。然后在测试阶段，通过将滤波器应用于新帧图像来进行目标的位置更新。相关滤波算法具有显著的计算效率，特别适合实时性要求较高的应用场景。

1. 最小输出平方误差滤波器

最小输出平方误差滤波器(Minimum Output Sum of Squared Error，MOSSE)是一种基于相关滤波的跟踪算法，通过最小化滤波器在训练样本上的输出误差来学习一个理想的滤波器。该滤波器可以用于后续帧的目标检测与跟踪。

1）基本原理

MOSSE 滤波器的目标是通过优化滤波器 h 使其对每个训练样本 x_i 的卷积输出接近理想的高斯响应 g_i。具体来说，优化目标可以表示为

$$\min_h \sum_{i=1}^{N} | h * x_i - g_i |^2 \tag{5.26}$$

其中，$*$ 表示卷积操作。然后，为了简化计算，MOSSE 在频域中求解该优化问题；根据卷积定理，空间域中的卷积操作可以通过频域中的乘法实现。因此，优化目标可以表示为

$$\min_H \sum_{i=1}^{N} | HX_i - G_i |^2 \tag{5.27}$$

其中，H 是滤波器 h 的傅里叶变换，X_i 是训练样本 x_i 的傅里叶变换，G_i 是期望输出 g_i 的傅里叶变换。通过傅里叶变换，MOSSE 滤波器的优化问题可以转化为

$$H = \frac{\sum_{i=1}^{N} G_i X_i^*}{\sum_{i=1}^{N} X_i X_i^* + \lambda} \tag{5.28}$$

其中，X_i^* 是 X_i 的复共轭，λ 是正则化参数，用于防止过拟合和数值不稳定。

2）跟踪流程

(1) 初始化：在第一帧中手动选择目标区域，并计算该区域的傅里叶变换 X_1；然后生成理想的高斯响应 g_1，该响应通常是一个在目标中心处为峰值的高斯分布，并计算高斯响应的傅里叶变换 G_1。

(2) 训练滤波器：使用初始目标区域的傅里叶变换 X_1 和理想的高斯响应 G_1，训练

MOSSE 滤波器 H：

$$H=\frac{G_1X_1^*}{X_1X_1^*+\lambda} \tag{5.29}$$

(3) 后续帧的目标检测：在后续帧中，对每个候选区域进行傅里叶变换 X_t；再通过滤波器 H 计算卷积输出 Y_t

$$Y_t=HX_t \tag{5.30}$$

然后逆傅里叶变换 Y_t 回到空间域，得到卷积输出 y_t；选择卷积输出 y_t 最大值对应的区域作为目标位置，如图 5.20 所示。

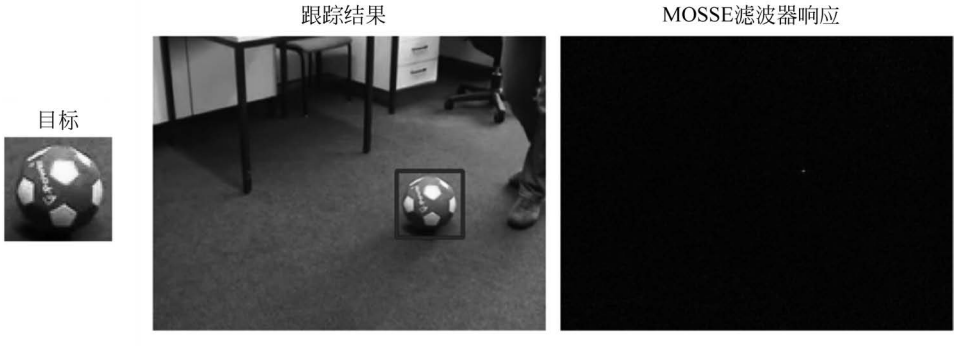

图 5.20　MOSSE 滤波器跟踪示例

(4) 更新滤波器：根据当前帧的检测结果，更新滤波器 H 以适应目标的变化。更新公式为

$$H_{\text{new}}=(1-\alpha)H_{\text{old}}+\alpha\frac{G_tX_t^*}{X_tX_t^*+\lambda} \tag{5.31}$$

其中，λ 是学习率，控制新旧滤波器的混合比例。

3) 方法的优点与缺点

MOSSE 滤波器通过频域计算，大大减少了计算复杂度，可以在资源受限的设备上实现实时跟踪。由于计算效率高，MOSSE 滤波器适用于实时目标跟踪任务。然而，MOSSE 滤波器对目标的旋转和尺度变化的鲁棒性较差，容易在这些情况下失去目标。此外，MOSSE 滤波器对光照变化的敏感性较高，当光照条件发生显著变化时，跟踪效果会下降。在背景复杂或存在相似目标的场景中，MOSSE 滤波器容易受到干扰，导致跟踪失败。

2. 核相关滤波器

核相关滤波器(Kernelized Correlation Filter，KCF)是相关滤波的一种扩展，它将核技巧引入到相关滤波框架中，通过在高维特征空间中进行线性分离，提升了滤波器的表达能力，从而增强了对目标的精确跟踪。

1) 基本原理

KCF 的核心思想是利用核技巧将原始特征映射到一个高维特征空间，在这个高维空间中进行线性相关滤波。通过这种方法，KCF 能够在低维特征空间中处理非线性问题，从而提升滤波器的表达能力和跟踪性能。

核技巧通过一个核函数 K 将原始特征 x 映射到一个高维特征空间 $\phi(x)$，而无须显式

地计算高维特征向量。常见的核函数包括线性核、多项式核和高斯核等，如高斯核函数定义为

$$K(x,y)=\exp\left(-\frac{|x-y|^2}{2\sigma^2}\right) \tag{5.32}$$

其优化目标可以表示为

$$\min_f \sum_{i=1}^{N} |f*\phi(x_i)-g_i|^2 \tag{5.33}$$

此外，为了简化计算，KCF 在频域中求解该优化问题；根据卷积定理，空间域中的卷积操作可以通过频域中的乘法实现。因此，优化目标可以表示为

$$\min_F \sum_{i=1}^{N} |F\Phi(\boldsymbol{X}_i)-\boldsymbol{G}_i|^2 \tag{5.34}$$

其中，F 是滤波器 f 的傅里叶变换，$\Phi(\boldsymbol{X}_i)$是训练样本 $\boldsymbol{x}_i$ 的高维特征的傅里叶变换，$\boldsymbol{G}_i$ 是期望输出 $\boldsymbol{g}_i$ 的傅里叶变换。通过傅里叶变换，KCF 滤波器的优化问题可以转换为：

$$F=\frac{\sum_{i=1}^{N}\boldsymbol{G}_i\Phi(\boldsymbol{X}_i)^*}{\sum_{i=1}^{N}\Phi(\boldsymbol{X}_i)\Phi(\boldsymbol{X}_i)^*+\lambda} \tag{5.35}$$

其中，$\Phi(\boldsymbol{X}_i)^*$ 是 $\Phi(\boldsymbol{X}_i)$的复共轭，λ 是正则化参数，用于防止过拟合和数值不稳定。

在 KCF 中，为了简化计算，使用了循环矩阵(Circulant Matrix)；循环矩阵的性质使得在频域中的计算更加高效。如图 5.21 所示，具体来说，假设 $\boldsymbol{x}$ 是一个 n 维向量，循环矩阵 $\boldsymbol{C}_x$ 可以通过将 $\boldsymbol{x}$ 的元素循环排列得到。循环矩阵的一个重要性质是它可以表示为傅里叶变换的对角矩阵乘法

$$\boldsymbol{C}_x=\boldsymbol{F}^{-1}\boldsymbol{\Lambda}_x\boldsymbol{F} \tag{5.36}$$

其中，$\boldsymbol{F}$ 是离散傅里叶变换矩阵，$\boldsymbol{\Lambda}_x$ 是 $\boldsymbol{x}$ 的傅里叶变换后的对角矩阵。

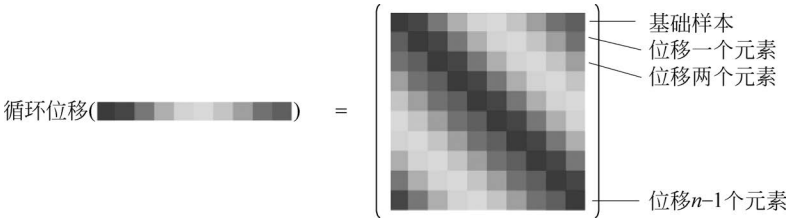

图 5.21　KCF 的循环矩阵

通过这一性质，KCF 可以将空间域中的卷积操作转换为频域中的乘法操作，从而显著提高计算效率。对于一个循环矩阵 $\boldsymbol{C}_x$ 和一个向量 $\boldsymbol{y}$，卷积 $\boldsymbol{C}_x\boldsymbol{y}$ 可以通过以下步骤计算

$$\boldsymbol{C}_x\boldsymbol{y}=\boldsymbol{F}^{-1}\boldsymbol{\Lambda}_x(\boldsymbol{F}\boldsymbol{y}) \tag{5.37}$$

其中，$\boldsymbol{F}\boldsymbol{y}$ 是 $\boldsymbol{y}$ 的傅里叶变换，$\boldsymbol{\Lambda}_x(\boldsymbol{F}\boldsymbol{y})$是$\boldsymbol{\Lambda}_x$ 与 $\boldsymbol{F}\boldsymbol{y}$ 的逐元素乘法，$\boldsymbol{F}^{-1}$ 是逆傅里叶变换。

2) 跟踪流程

其跟踪流程与上述 MOSSE 的基本一致，区别在于利用了核函数将特征映射到高维特征空间，并计算高维特征的傅里叶变换 $\Phi(\boldsymbol{X}_1)$。其 KCF 滤波器的训练为

$$\boldsymbol{F}=\frac{\boldsymbol{G}_1\Phi(\boldsymbol{X}_1)^*}{\Phi(\boldsymbol{X}_1)\Phi(\boldsymbol{X}_1)^*+\boldsymbol{\lambda}} \tag{5.38}$$

滤波器的更新公式为

$$\boldsymbol{F}_{\text{new}}=(1-\alpha)\boldsymbol{F}_{\text{old}}+\alpha\frac{\boldsymbol{G}_t\Phi(\boldsymbol{X}_t)^*}{\Phi(\boldsymbol{X}_t)\Phi(\boldsymbol{X}_t)^*+\boldsymbol{\lambda}} \tag{5.39}$$

其中，λ 是学习率，控制新旧滤波器的混合比例，$\boldsymbol{G}_t$ 是当前帧中目标区域的高斯响应的傅里叶变换，$\Phi(\boldsymbol{X}_t)$是当前帧中目标区域的高维特征的傅里叶变换。

3）方法的优点与缺点

KCF 通过频域计算和循环矩阵的性质，大大减少了计算复杂度，可以在资源受限的设备上实现实时跟踪。通过核技巧，KCF 能够在高维特征空间中处理非线性问题，提升了滤波器的表达能力和跟踪性能；由于计算效率高，KCF 适用于实时目标跟踪任务。

尽管 KCF 在许多场景中表现出色，但也存在一些局限性。KCF 对目标的旋转和尺度变化的鲁棒性较差，容易在这些情况下失去目标；KCF 对光照变化的敏感性较高，当光照条件发生显著变化时，跟踪效果会下降；在背景复杂或存在相似目标的场景中，KCF 容易受到干扰，导致跟踪失败。

3. 多通道相关滤波

多通道相关滤波（Discriminative Scale Space Tracker，DSST）是对 KCF 的进一步扩展，针对目标的尺度变化问题进行了优化。DSST 通过在尺度空间中引入多通道滤波器，实现了对目标尺度的自适应跟踪。

1）基本原理

DSST 的核心思想是在不同的尺度下同时应用 KCF 滤波器，并通过多尺度响应图的最大值来确定目标的最佳尺度。通过这种方法，DSST 能够在目标尺寸变化较大的场景中保持较高的跟踪精度和鲁棒性。如图 5.22 所示，具体来说，DSST 在不同的尺度下同时应用 KCF 滤波器，并通过多尺度响应图的最大值来确定目标的最佳尺度。

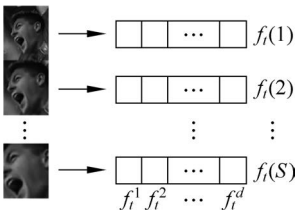

图 5.22　DSST 的尺度滤波器

假设当前帧的目标位置已知，则在尺度空间中，DSST 通过以下公式计算尺度响应图

$$\boldsymbol{R}(s)=F^{-1}(\boldsymbol{H}_s\odot\boldsymbol{X}_s) \tag{5.40}$$

其中，s 表示尺度因子，$\boldsymbol{H}_s$ 是当前尺度下的滤波器的傅里叶变换，$\boldsymbol{X}_s$ 是当前尺度下的输入图像的傅里叶变换，$\boldsymbol{R}(s)$是尺度响应图，F^{-1} 表示傅里叶逆变换，$\odot$表示逐元素乘法。

通过对所有尺度响应图进行最大值搜索，DSST 可以实时地调整目标的尺度，从而增强了跟踪的精确度和鲁棒性。

2）跟踪流程

其 DSST 的跟踪流程与 KCF 的区别在于根据不同尺度因子得到对应的滤波器。具体来说，首先初始化尺度空间，选择一系列尺度因子 $s_1,s_2,\cdots,s_N$；其次对每个尺度 s_i，使用初始目标区域的高维特征的傅里叶变换和理想的高斯响应 $\boldsymbol{H}_{s_i}$；最后在后续帧

的检测中，也是对每个尺度 s_i，通过滤波器 $\boldsymbol{H}_{s_i}$ 计算卷积输出 $\boldsymbol{Y}_{t,s_i}$，并选择所有尺度响应图空间域中卷积输出响应图 $\boldsymbol{Y}_{t,s_i}$ 中最大值对应的区域和尺度作为目标位置和最佳尺度。

3）方法的优点与缺点

DSST 同时通过频域计算和循环矩阵的性质，大大减少了计算复杂度，可以在资源受限的设备上实现实时跟踪；更重要的是，它通过在尺度空间中引入多通道滤波器，DSST 能够处理目标的尺度变化，提升了跟踪的精确度和鲁棒性。

然而，虽然 DSST 在处理目标尺度变化方面表现出色，但其计算复杂度相比 KCF 有所增加，尤其是在多尺度空间中进行滤波器训练和检测时；同样在背景复杂或存在相似目标的场景中，DSST 仍然容易受到干扰，导致跟踪失败。

5.4.3 基于孪生网络的目标跟踪

孪生网络(Siamese Network)是近年来在目标跟踪领域中广泛应用的一类深度学习模型。与传统方法不同，孪生网络通过对比学习的方式来判断目标是否匹配，而不依赖于目标的显式建模。孪生网络的基本思想是将目标跟踪问题转化为一个相似性度量问题，即通过对比目标图像与搜索区域图像的特征表示，确定目标在新图像中的位置。

1. 孪生网络的基本原理

孪生网络由两个共享权重的卷积神经网络组成，分别用于处理模板图像(包含目标的图像)和搜索区域图像。通过对这两幅图像的特征进行比较，网络可以生成一个相似性响应图，表示目标在搜索区域中的可能位置。

假设模板图像的特征表示为 z，搜索区域图像的特征表示为 x，网络计算的相似性响应图为

$$R = z * x \tag{5.41}$$

其中，$*$ 表示卷积操作。响应图的峰值位置表示目标在搜索区域中的最可能位置。

2. SiamFC 模型

完全卷积孪生网络(Siamese Fully-Convolutional Network，SiamFC)是孪生网络中最早提出的模型之一。SiamFC 摒弃了全连接层，采用全卷积网络架构，以保持输入图像的空间信息。如图 5.23 所示，其工作流程如下。

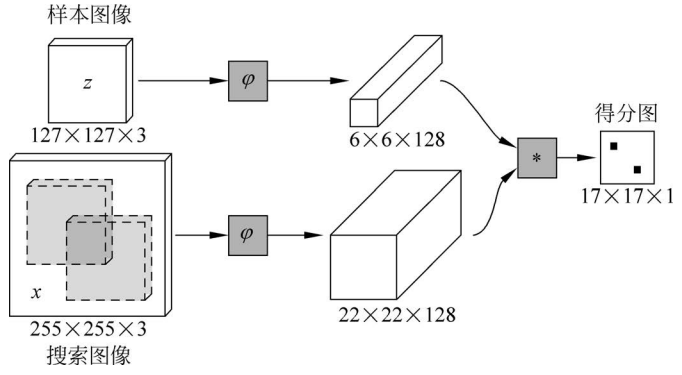

图 5.23 SiamFC 结构

（1）特征提取：模板图像 z 通过共享的卷积神经网络进行特征提取，得到特征图 z；搜索区域图像 x 通过相同的卷积神经网络进行特征提取，得到特征图 x。

（2）相似度计算：然后，对两个特征图进行卷积操作，计算出相似性响应图 $R(i,j)$，表示搜索区域中各位置的相似度

$$R(i,j)=(z*x)(i,j) \tag{5.42}$$

其中，(i,j)表示搜索区域中的位置坐标。

（3）目标定位：相似性响应图 R 的最大值位置对应于目标在搜索区域中的位置。具体来说，找到 R 中的最大值 $R_{\max}$ 及其对应的坐标$(i_{\max},j_{\max})$，这个坐标即为目标在搜索区域中的最可能位置

$$(i_{\max},j_{\max})=\underset{i,j}{\operatorname{argmax}}\,R(i,j) \tag{5.43}$$

3. SiamRPN 模型

区域建议孪生网络（Siamese Region Proposal Network，SiamRPN）是在 SiamFC 的基础上引入区域建议网络（RPN）的改进模型，通过联合预测目标位置和尺度，显著提升了跟踪精度。如图 5.24 所示，其工作流程如下：

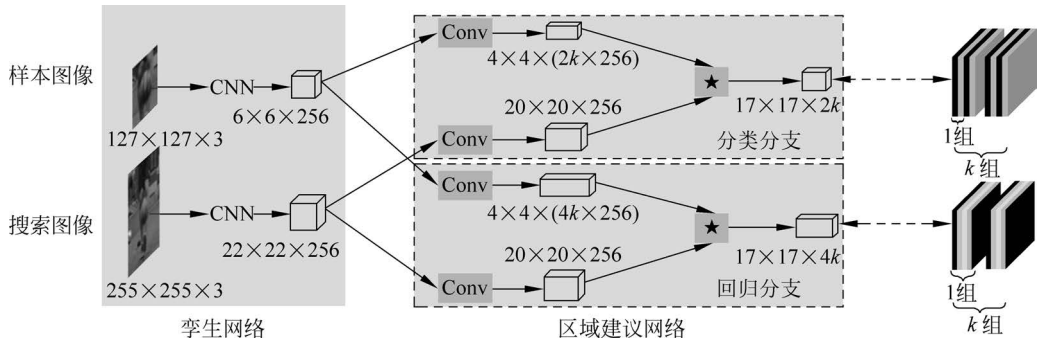

图 5.24　SiamRPN 网络结构与工作流程

（1）特征提取：模板图像 z 通过共享的卷积神经网络进行特征提取，得到特征图 z；搜索区域图像 x 通过相同的卷积神经网络进行特征提取，得到特征图 x。

（2）区域建议：SiamRPN 通过 RPN 生成一组候选区域，并对每个候选区域进行分类和边界框回归。具体地，SiamRPN 为每个候选区域生成一个类别预测 $p(i,j)$和一个边界框回归参数 $t(i,j)$

$$p(i,j)=\sigma(W_c*z*x) \tag{5.44}$$

$$t(i,j)=W_r*z*x \tag{5.45}$$

其中，$\sigma(\cdot)$为 sigmoid 函数，分别表示分类和回归的权重。

（3）目标定位与尺度调整：SiamRPN 根据 RPN 输出的边界框参数调整目标的位置信息，并根据回归结果对目标的尺度进行自适应调整；通过分类得分最高的候选区域确定目标的位置，并通过回归参数调整目标的尺度和位置。

4. 基于孪生网络的目标跟踪的优点与缺点

孪生网络的端到端设计使得它在计算效率和实时性方面表现出色，尤其适合实时跟踪任务；孪生网络结构简单明了，也易于实现和训练，同时可以有效地处理跟踪任务中的各种

复杂场景。此外，孪生网络对各种目标跟踪任务表现良好，特别是对具有强外观变化的目标有很好的适应能力。

然而，原始 SiamFC 模型对目标的尺度变化适应性较差，尽管后续的 SiamRPN 有所改进，但在某些场景中仍存在不足。孪生网络的跟踪精度可能在复杂背景或多目标场景中受到影响，尤其是在目标发生剧烈变化时。

5.4.4 其他前沿目标跟踪模型

随着深度学习的快速发展，除了孪生网络外，许多前沿的目标跟踪模型不断涌现。这些模型通过引入新的网络结构、学习策略和优化方法，大幅提升了目标跟踪的精度、鲁棒性和实时性。以下介绍几种代表性的前沿目标跟踪模型。

1. 基于 Transformer 的目标跟踪

基于 Transformer 的目标跟踪(TransT)是一种创新的目标跟踪模型，它巧妙地融合了卷积神经网络(CNN)与 Transformer 架构的优势，特别适用于处理复杂场景下的目标跟踪任务。其主要特点在于通过 Transformer 的自注意力机制来捕捉目标与周围环境之间的全局依赖关系，从而提高跟踪的准确性和鲁棒性。

1) 基本原理

TransT 首先使用共享的 CNN 从模板图像和搜索区域图像中提取基础特征；模板图像是包含目标对象的初始图像，而搜索区域图像是在视频序列的后续帧中截取的，用于寻找目标的新位置。这两个过程产生的特征图分别记为 z(模板特征)和 x(搜索区域特征)。然后，提取出的特征图被送入 Transformer 模型中，利用自注意力机制来增强特征间的关联。这一机制允许模型关注到图像中的关键部分，并且能够有效地捕捉长距离依赖关系，这对于理解目标与其周围环境的关系至关重要。自注意力机制的基本计算公式如下

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{5.46}$$

其中，Q、K 和 V 分别代表查询、键和值向量，而 d_k 则是键向量的维度。通过这种方式，模型能够更好地理解目标与背景之间的关系。经过自注意力机制处理后的特征会进一步通过位置编码层和输出层，最终生成一个相似性响应图，该图用于确定目标在搜索区域中的具体位置。位置编码帮助保持了空间信息，确保模型能准确地识别目标的位置。如图 5.25 所示，TransT 还采用了自我上下文增强(ECA)和交叉特征增强(CFA)两种技术来进一步提升模型性能。ECA 通过增加局部上下文信息来增强目标检测的准确性；CFA 则促进了模板特征和搜索区域特征之间的有效交互，有助于模型更好地理解目标与环境的动态变化。

2) 方法的优点与缺点

得益于 Transformer 的自注意力机制，TransT 能够在跟踪过程中维持对目标及周围环境的全局感知，即使在目标出现大范围移动或形变时也能保持良好的跟踪效果。此外，无论是面对目标尺度的变化、部分遮挡还是背景中的干扰因素，TransT 都能展现出较强的鲁棒性。

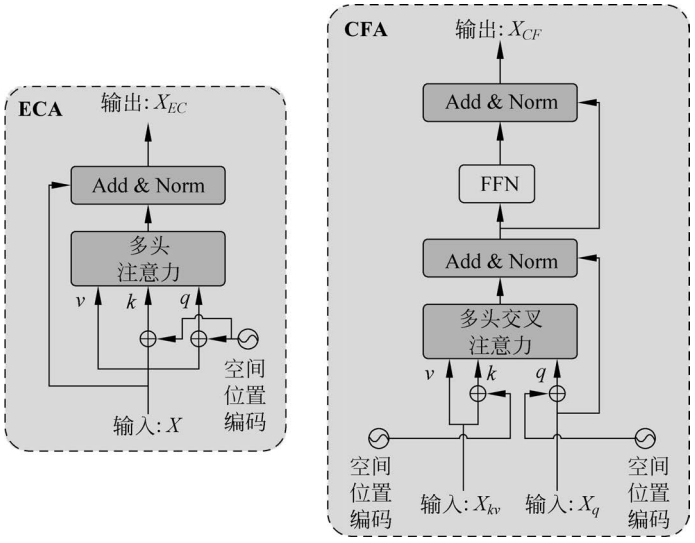

图 5.25　TransT 中的注意力机制：自我上下文增强(ECA)和交叉特征增强(CFA)

然而，Transformer 的自注意力机制涉及大量的矩阵运算，尤其是当特征图的分辨率较高时，计算量会显著增加。这可能导致模型在实时跟踪任务中的计算延迟，影响实时性能。此外，Transformer 模型通常需要较大的内存来存储中间特征和权重，特别是在处理高分辨率图像时，内存消耗会更高。

2. 基于强化学习的目标跟踪

强化学习(Reinforcement Learning，RL)近年来在目标跟踪领域开始崭露头角。RL-Track 将目标跟踪问题视为一个序列决策问题，通过学习智能体(Agent)的跟踪策略来动态调整目标的跟踪路径，如图 5.26 所示。其核心思想是通过试错学习，让跟踪器逐步优化其跟踪策略，适应复杂场景中的目标运动和环境变化。

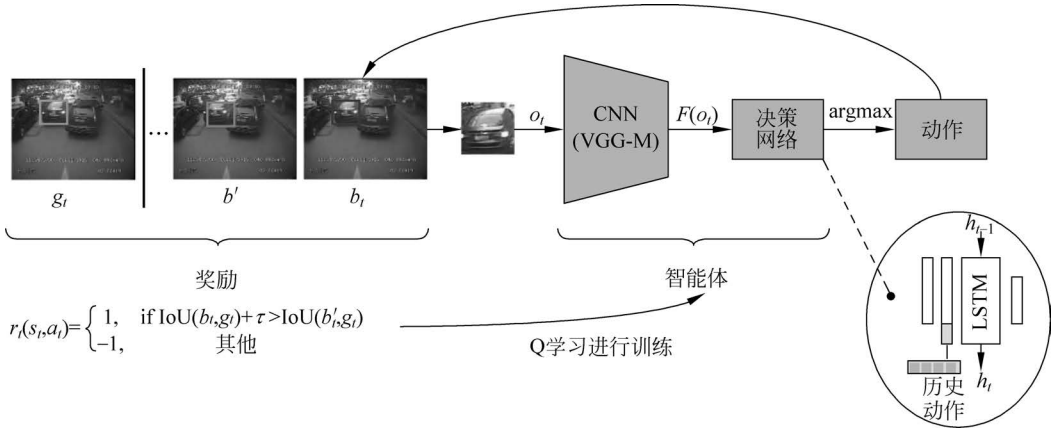

图 5.26　基于强化学习的目标跟踪框架

1) 基本原理

(1) 状态表示：在每个时间步 t，跟踪器从环境中获取一个状态 s_t。状态 s_t 通常包括以下信息：目标在当前帧中的位置和尺寸；当前帧的背景特征，如图像的局部和全局特征；

前几帧中目标的位置和运动信息。状态 s_t 可以通过多种方式表示，如使用特征图、边界框坐标、光流信息等。

(2) 动作选择：根据当前状态 s_t，跟踪器选择一个动作 a_t。动作可以是几种类型：目标边界框的位置、尺寸、比例调整；切换不同候选跟踪器等。动作选择通常通过策略函数 $\pi(a_t|s_t)$来决定，该函数表示在给定状态 s_t 下选择动作 a_t 的概率分布。

(3) 奖励函数：执行动作 a_t 后，跟踪器根据目标的跟踪精度和边界框的重叠度(IoU)得到一个即时奖励 r_t。奖励函数通常设计为关于 IoU 的函数，例如

$$r_t=\begin{cases}1 & \text{IoU}>\theta\\ -1 & \text{IoU}<\phi\\ 0 & \text{其他}\end{cases} \tag{5.47}$$

其中，θ 和 ϕ 是预设的阈值，用于区分高精度和低精度的跟踪结果。跟踪器的目标是最大化累积奖励 R，即在多个时间步中获得的总奖励

$$R=\sum_{t=0}^{T}\gamma^t r_t \tag{5.48}$$

其中，γ 是折扣因子，用于平衡近期和远期奖励的重要性。

(4) 策略优化：跟踪器通过强化学习算法不断更新其策略 $\pi(a_t|s_t)$，以最大化累积奖励。常用的强化学习算法包括 Q-Learning、Deep Q-Network 等。通过与环境的交互，收集状态、动作和奖励的样本，逐步优化策略函数。训练过程中通常需要大量的试错和迭代，以找到最优的跟踪策略。

2) 方法的优点与缺点

RL-Track 能够通过试错学习，逐步优化跟踪策略，适应复杂场景中的目标运动和环境变化。它因可以灵活地定义状态、动作和奖励函数，适用于多种跟踪任务和应用场景；此外，通过不断学习和调整，RL-Track 在目标尺度变化、遮挡和背景干扰等复杂情况下表现出较好的鲁棒性。

然而，强化学习算法通常需要大量的计算资源和时间来进行训练，尤其是在深度强化学习中；RL-Track 同样需要大量的训练数据和环境交互来学习有效的策略，数据收集和标注成本较高。强化学习的训练过程可能存在不稳定性和收敛问题，需要精心设计奖励函数和算法参数。

5.5 工业目标检测与跟踪的应用案例

随着工业 4.0 的推进，工业目标检测与跟踪技术在现代制造业中发挥着至关重要的作用。通过引入人工智能和深度学习算法，目标检测与跟踪技术能够极大地提高生产效率、保障产品质量、优化生产流程，并实现智能化生产。在工业自动化、工业机器人和交通领域，这些技术得到了广泛应用。

5.5.1 医药自动化生产线中的药液微弱异物检测

1. 面临挑战

医药产品的质量直接影响公众健康，质量不达标不仅会削弱药效，还可能对用药者造

成严重身体伤害；因此，药品生产中的质量检测显得尤为重要，尤其是药液中的异物检测。面对突发情况下药品需求激增的挑战，中国正在加速推动医药生产的自动化和智能化进程。为提高药液产线的检测效率，业界开始尝试使用机器视觉技术替代传统的人工检测；然而，由于异物的尺寸微小、形态多样，加之检测环境中的多种干扰因素，现有检测方法常常难以达到高精度和低漏检率的要求。

2. 系统设计

在药品生产过程中，诸如清洗、配药、灌装和封盖等操作均可能导致异物混入药液中。因此，质量检测设备作为生产环节最后的保障，能够在线检测输入药瓶的瓶身缺陷、瓶口缺陷、封盖缺陷以及瓶内异物。本文所采用的医药异物检测系统包括药液入口、次品出口和合格品出口，如图 5.27 所示，主要结构由回转轮盘、进瓶螺杆、检测光源、工业相机、伺服电机、旋转轮组及上位机组成。系统内部设有多个质量检测工位，包括暗场异物检测、瓶身裂纹检测、偏振纤毛检测及药瓶封口检测，主要采用暗场异物检测工位进行药液异物图像采集。

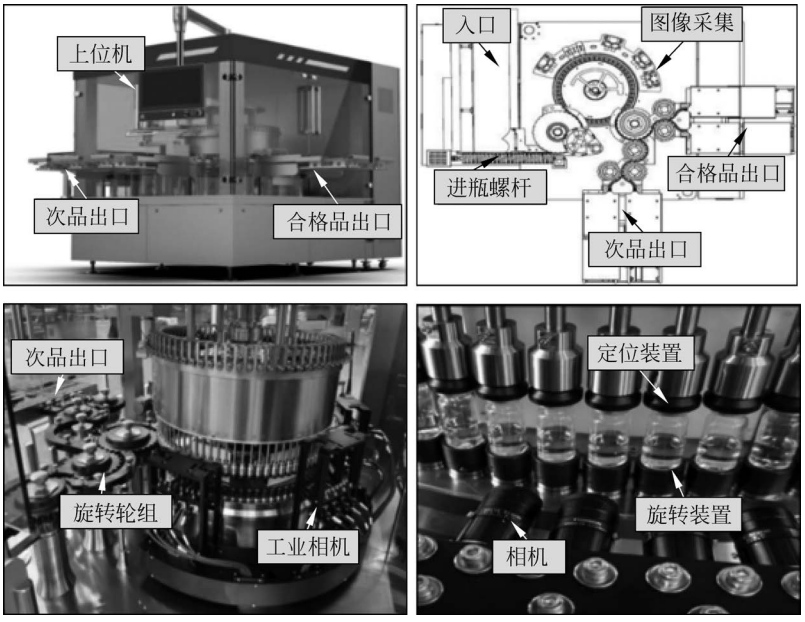

图 5.27　医药质量检测装备的结构图与实物图

3. 研究方法

(1) 异物聚集与图像采集：该设备首先通过定位装置测量药瓶的高度，并施加一定压力，以确保药液在检测过程中固定不动；然后，使用旋转装置以 4000r/min 的高速旋转药瓶，借此将异物汇集至瓶身中部。该旋转方法有效避免了微小气泡对后续检测的干扰，确保异物能被准确定位。旋转数秒后，系统将迅速制动，并在药液平稳后，使用工业相机连续采集多帧药液图像；每个药瓶会被采集 30 帧图像，这些图像通过数据线传输至计算设备进行后续处理。为提升拍摄效率及图像质量，系统采用多组超高速、高分辨率相机，每个相机配备高倍数长焦镜头，将被检测物放大 60 倍，显著提高了图像的数量和细节，确保微小异物的轨迹细节被保留。采集数据的示例如图 5.28 所示。

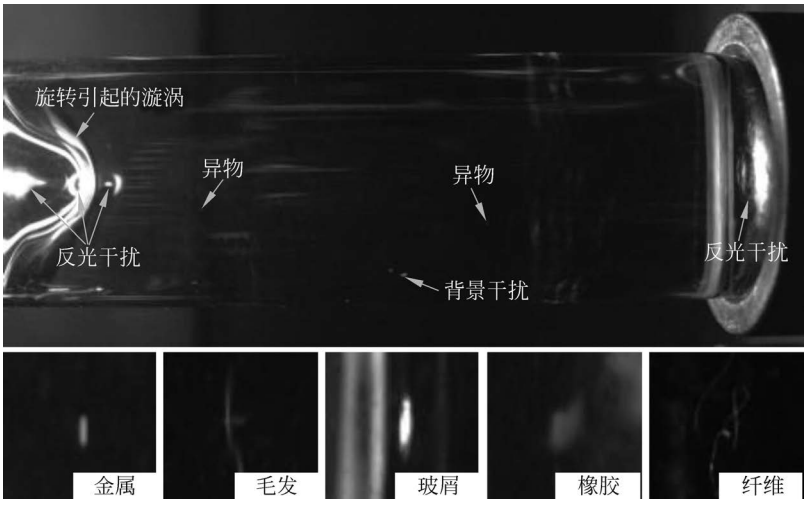

图 5.28 采集数据示例

(2) 检测结果的后处理：为了最大程度减少误检和漏检的影响，本文设计了一种基于视觉融合的轨迹检测网络。该网络首先对单帧图像的检测结果进行统计，获取每帧图像中高于检测阈值的异物位置。随后，生成一张与药液图像序列帧大小一致的纯黑灰度图，根据每帧图像的检测结果调整对应区域的灰度值，保持原始图像的时间顺序。较早的帧赋予较低的灰度值，而较晚的帧赋予较高的灰度值，使得如果药液中存在异物，它们将在灰度图中形成明显的轨迹，最终效果如图 5.29 所示。最后，通过视觉检测算法对这张处理后的灰度图进行轨迹检测，判断药瓶中是否存在异物。

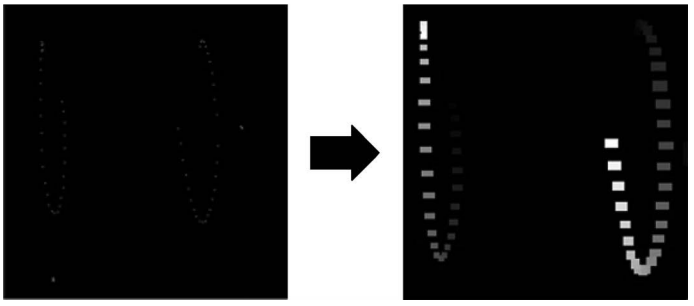

图 5.29 用检测框代替点集表示单帧异物检测结果

(3) 级联结构：为确保检测模型与轨迹检测网络的有效结合，设计了一种级联策略，如图 5.30 所示。具体而言，利用蒸馏方法生成的学生网络对每瓶药液的连续帧图像进行逐一检测；这些序列图像及其检测结果依次送入基于视觉融合的轨迹检测网络中，进行结果的融合与可视化处理；视觉检测网络输出轨迹检测结果，依据此结果判断药瓶中是否存在异物。

最终，通过测试 100 瓶药液的连续帧图像，结果显示：该方法成功检测出 98 瓶药液中的异物，漏检率仅为 2%，检测速度达到了 0.5 秒/瓶；相比人工灯检的 15 秒/瓶有显著提升，有效降低了误检率和漏检率，证明了其在医药质量检测中的有效性和实用性。

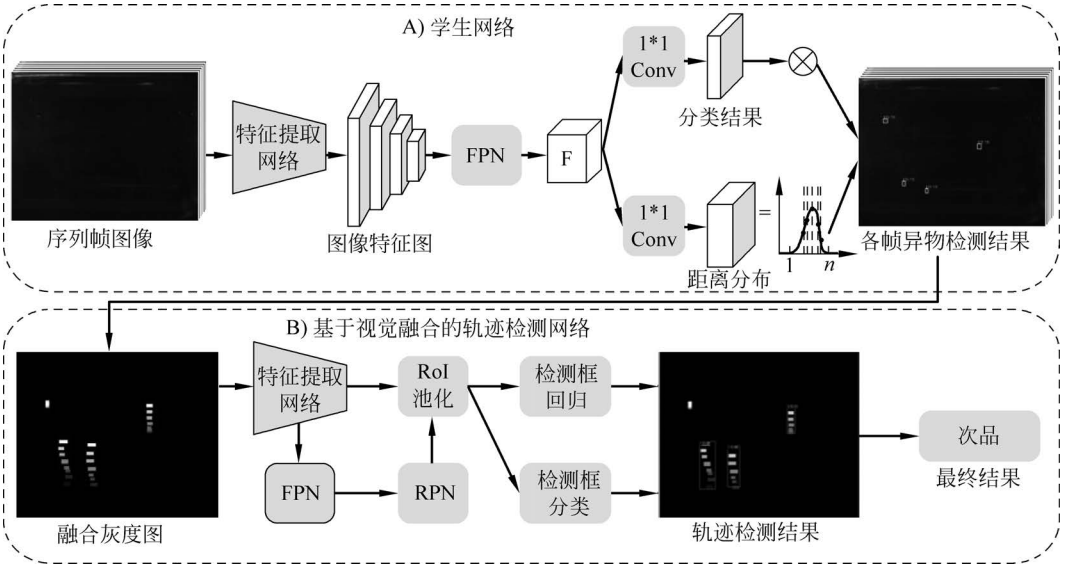

图 5.30　多模型级联的检测网络

5.5.2　可见光与红外图像融合的电力热故障判别

1. 面临挑战

当前热故障判别巡检作业大多基于单一红外图像来进行。但是现有的无人机红外图像空间分辨率较低，难以准确感知配网线路绝缘子、配电变压器等典型设备的三相温度和相间温差；而可见光图像空间分辨率高，能较好地反映线路的空间外观，却不能反映温度信息。因此通过红外与可见光图像的双光融合得到高质量的红外图像是解决这一问题的常用方法；通过双光融合能够使可见光高分辨率的优点与红外图像的温度信息相结合，一定程度上提高红外图像的分辨率。然而双光融合方法仍面临着两个问题：(1)因可见光相机与红外相机的成像差异和无人机外部参数问题，导致可见光图像与红外图像在图像中的空间信息不对齐；这种不对齐的情况在实际场景中难以避免，且直接影响了融合的效果。(2)进行热故障判别时需要使用具有精准温度信息的红外图像，而融合后的红外图像必然会损失一定的信息从而导致判别不准确。

2. 研究方法

鉴于实际任务场景对配准方法的需求，需要具备实时性和高精度的特点，而且由于红外图像与可见光图像之间存在显著的模态差异，现有方法难以取得理想的效果。在对配电网部件进行热故障判别时，通常需要使用具有温度信息的红外图像。然而，现有的解决方法大多集中在对红外图像中的部件进行目标检测。由于红外图像的分辨率较低，直接对红外图像中的配电网部件进行检测往往难以满足后续热故障判别任务的精度要求。相比其他方法，自适应配准方法避免了使用低分辨率的红外图像进行部件的目标检测，而是利用配准后的可见光图像代替，如图 5.31 所示。因此通过适用于实际场景的自适应配准方法，基于手工特征点选取，使得可见光图像中的部件位置信息与红外图像中的位置已经对齐，能够满足实际应用中的需求，检测效果如图 5.32 所示。

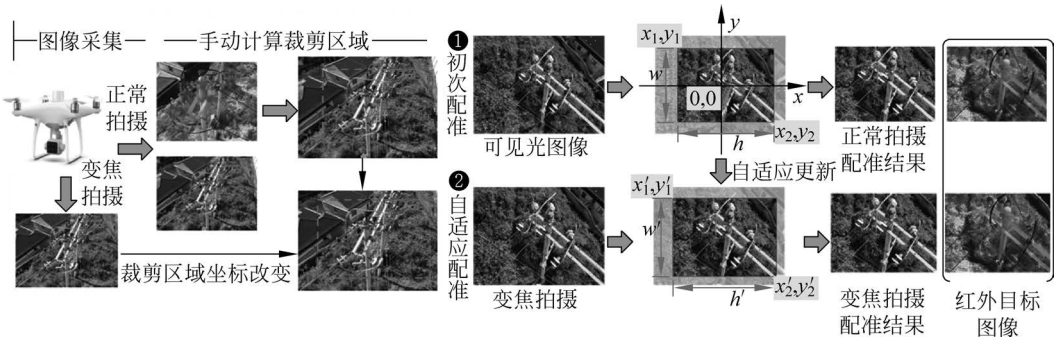

图 5.31 多模态融合自适应配准检测方法

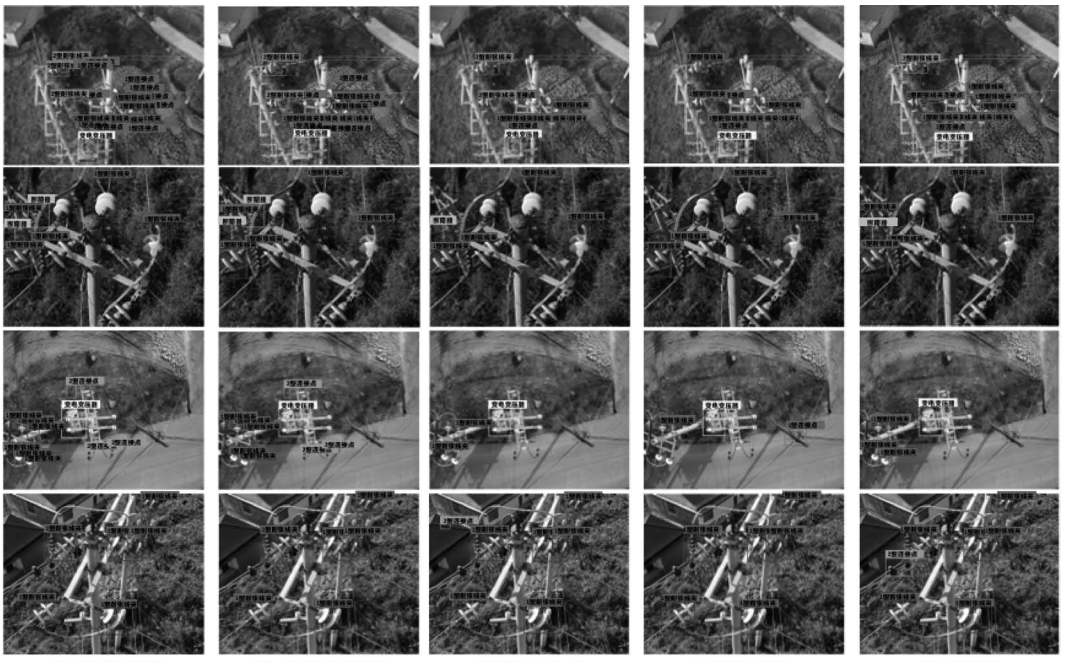

图 5.32 自适应配准多模态检测结果

3. 研究意义

这种多模态图像自适应检测方法，解决了配网热部件故障判别任务对红外图像目标检测精度低的问题；包括自适应的配准和预测信息迁移两步骤。自适应配准可以用于解决无人机搭载的相机直接因成像差异产生的空间内容不对齐情况。相比于其他方法，自适应配准方法能够忽略模态间巨大的差异，完成高精度的配准。预测信息迁移方法在自适应配准基础上通过对高空间分辨率的可见光图像进行训练以及预测，并将预测结果精准的迁移到红外图像中，间接完成了对红外图像的部件检测。这种方法实现简单、易部署且实时性高。在配网部件多模态巡检中具有高效性，在未来研究中可以结合其他工程应用问题展开进一步研究。

5.5.3　电力自动化巡检中小样本情况下的异物检测

随着电力系统规模的不断扩展和无人机巡检技术的普及，电力线路异物检测逐渐成为电力系统安全维护的关键任务，例如，鸟巢、塑料袋或树枝等可能会缠绕在电力线路上，导致短路、线路损坏甚至火灾等严重安全事故等。由于这些异物的体积通常较小，且多发生在复杂的户外环境中，传统的人工巡检方式难以及时发现并处理这些潜在的安全隐患。因此，基于工业视觉的自动化异物检测技术应运而生，为电力系统的智能化巡检提供了新的解决方案。

1. 面临挑战

当前的电力线路异物检测面临多重挑战。首先，异物种类繁多、形态各异，且在图像中往往只占据很小的像素区域，这使得传统的目标检测算法在处理小目标时容易出现漏检或误检现象；其次，由于无人机通常以鸟瞰角度采集图像，异物与线路、背景物体的重叠较多，导致目标边界模糊不清，进一步增加了检测难度；此外，电力场景中数据获取相对困难，尤其是包含异物的样本数据稀缺，这对模型的训练和泛化能力提出了更高要求。

2. 检测方法

针对这些问题，近年来，基于深度学习的小样本目标检测技术被引入电力线路异物检测中，通过迁移学习和注意力机制的结合，这些方法在小样本条件下也能够取得较好的检测效果。例如，某些方法通过在线难度样本选择技术，对复杂样本进行优先训练，进一步提升了模型对小目标异物的检测精度。此外，内卷积等新型网络结构的引入增强了模型对小目标异物的聚合能力，显著提高了异物的识别率。一种典型的针对电力线路异物检测的小样本方法结构如图 5.33 所示。该方法包括两阶段训练：第一阶段，采用新的锚框方案在具有大量标注信息的基类数据集（电力部件）上训练检测器主体部分；第二阶段，冻结训练检测器绝大部分参数，将训练好的模型迁移至小数据集（异物）进行微调。

该方法在多个类别样本和所有实例测试中都取得了显著效果，充分证明了该方法的优势性。在以 15 个实例进行训练时，该方法在异物类别上的检测 AP 值可达 98.6%，高出其他方法至少 4.4%，提升了小样本电力线路异物检测性能。可视化的效果如图 5.34 所示。

3. 研究意义

引入基于工业视觉的自动化异物检测技术，能够显著提高巡检效率，减少人工巡检的局限性，并保证在复杂天气、恶劣环境下的检测可靠性。深度学习技术与无人机巡检相结合，能够覆盖大面积的线路，准确检测到体积小、形态多样的异物，弥补了传统方法在漏检、误检等方面的不足。这不仅提升了电网的维护和运营效率，也降低了由于异物引发的安全事故概率，保障了电力系统的稳定性。此外，智能化的异物检测技术还能为电力巡检的全面自动化铺平道路，减少对人力资源的依赖，降低巡检成本，提升电力系统的安全防护水平，为未来智能电网的建设提供坚实的技术基础。通过预防性维护，电力系统能够更具韧性，减少停电和设备损坏的风险，进一步确保社会生活和经济生产的持续稳定。

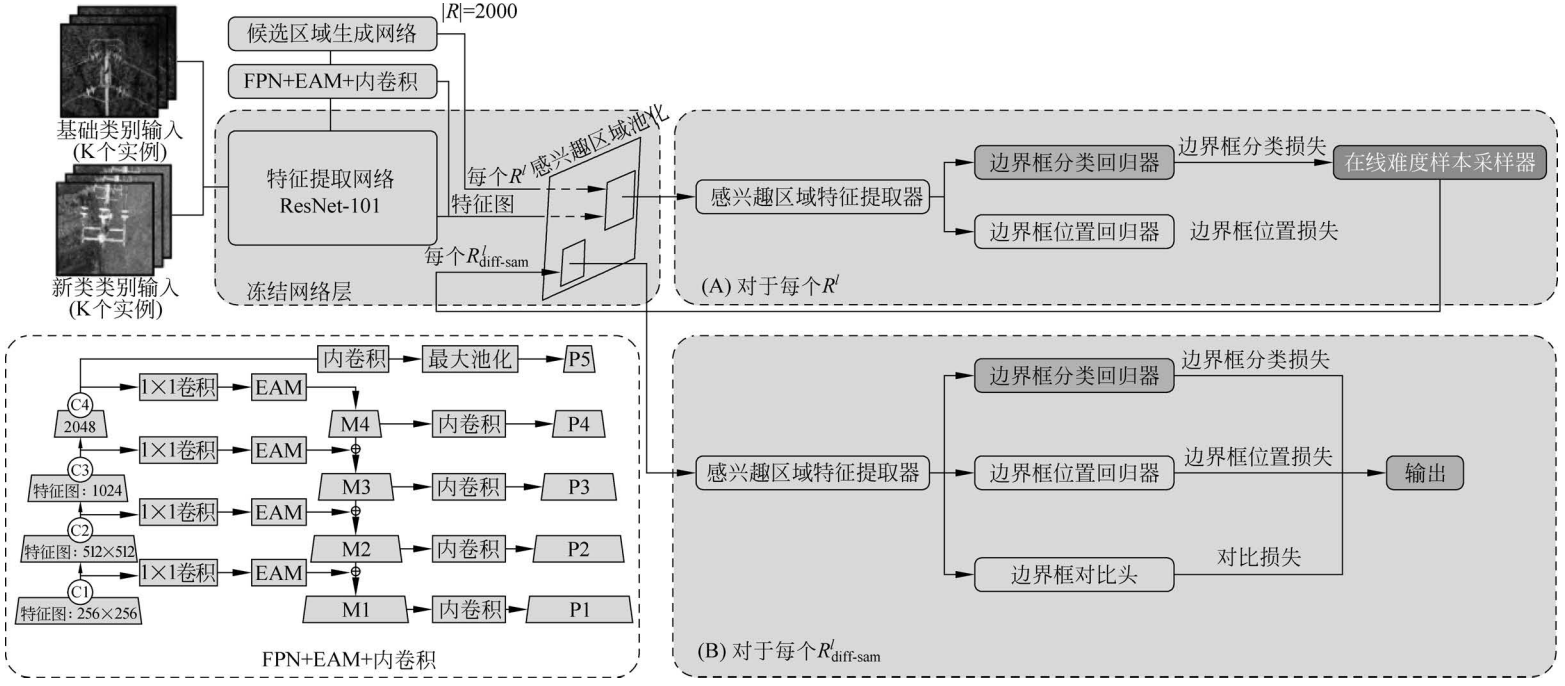

图 5.33　一种典型的针对电力线路异物检测的小样本方法

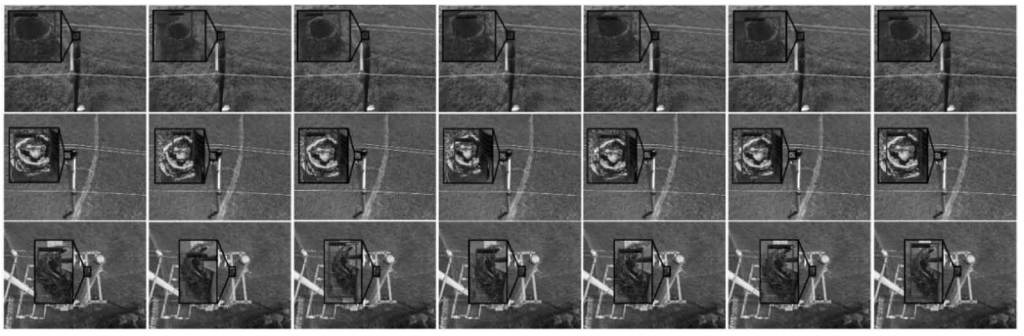

图 5.34　检测效果对比图

5.6　本章小结

本章系统地探讨了工业机器视觉中目标检测与目标跟踪的关键技术及其在工业应用中的重要性。

首先，从目标检测的基本概念出发，介绍了目标检测的定义、意义以及在实际应用中面临的挑战。

接着，详细讲解了几种经典的传统目标检测方法。基于模板匹配的方法依赖于物体的形状特征，在特定环境中具有较好的应用效果；基于特征点的目标检测方法利用图像中的特征点进行匹配，适用于多种目标检测任务。随着深度学习技术的发展，基于深度学习的目标检测算法成为了工业机器视觉中的主流。我们深入分析了 YOLO、Faster R-CNN 等代表性算法的工作原理、结构特点以及各自的优势与不足。通过这些算法的对比，读者可以理解如何在实际工业应用中选择合适的算法，平衡检测速度与精度，并满足工业生产中对实时性和可靠性的要求。

最后，在目标跟踪部分，从传统的跟踪算法出发，讲解了如相关滤波、光流法等经典算法的基本原理和应用场景。这些方法在运动目标的跟踪中，尤其是简单场景下具有较好的表现，但随着工业应用场景的复杂化，传统方法在准确性和鲁棒性方面显得力不从心。基于深度学习的目标跟踪方法，尤其是孪生网络(Siamese Network)的引入，极大提升了跟踪的精度和抗干扰能力，这为自动化生产线、机器人导航、物流系统等场景中的目标跟踪应用提供了更加高效的解决方案。

通过本章的学习，读者不仅能够了解从传统到现代的目标检测与跟踪技术演变过程，还可以掌握如何将这些技术应用到具体的工业场景中。传统方法在特定条件下依然发挥着重要作用，而深度学习方法则通过提高检测精度和跟踪鲁棒性，使得工业机器视觉系统能够适应更加复杂的任务需求。未来，随着技术的进一步发展，如何更好地融合传统与现代技术，解决实际工业应用中的挑战，将是读者需要思考的重要问题。

5.7　思考与习题

1. 在工业环境中，遮挡、光照变化、背景复杂等问题时常出现。请结合所学，讨论这些问题对目标检测与跟踪算法的影响，并提出可能的解决方案。

2. 随着深度学习在工业机器视觉中的广泛应用，传统的目标检测与跟踪方法是否还有其应用价值？请讨论这些方法在某些特定工业场景中的优势。

3. 请比较 YOLO、SSD 与 Faster R-CNN 三种目标检测算法的原理、性能和适用场景。在实际工业应用中，如何选择适合的算法？

4. 在深度学习目标检测模型中，如何平衡检测速度与检测精度？在实时性要求较高的工业应用中，哪些方法可以优化检测过程？

5. 传统目标检测方法如模板匹配、基于特征点的方法在何种场景中仍有应用价值？请结合实际工业应用，分析这些方法的优势与局限性。

6. 请解释基于相关滤波的目标跟踪算法的原理，并讨论它在工业机器视觉中的应用场景及优势。与其他跟踪方法相比，相关滤波算法的优缺点是什么？

7. 在交通场景中，目标检测与跟踪如何结合应用于智能交通系统？例如，如何通过视觉系统实现对车辆或行人的检测与跟踪？

8. 电力场景中的设备巡检常常需要对目标进行检测与跟踪。请结合电力场景的特殊需求，讨论如何设计合适的目标检测与跟踪方案。