《智能机器人设计》

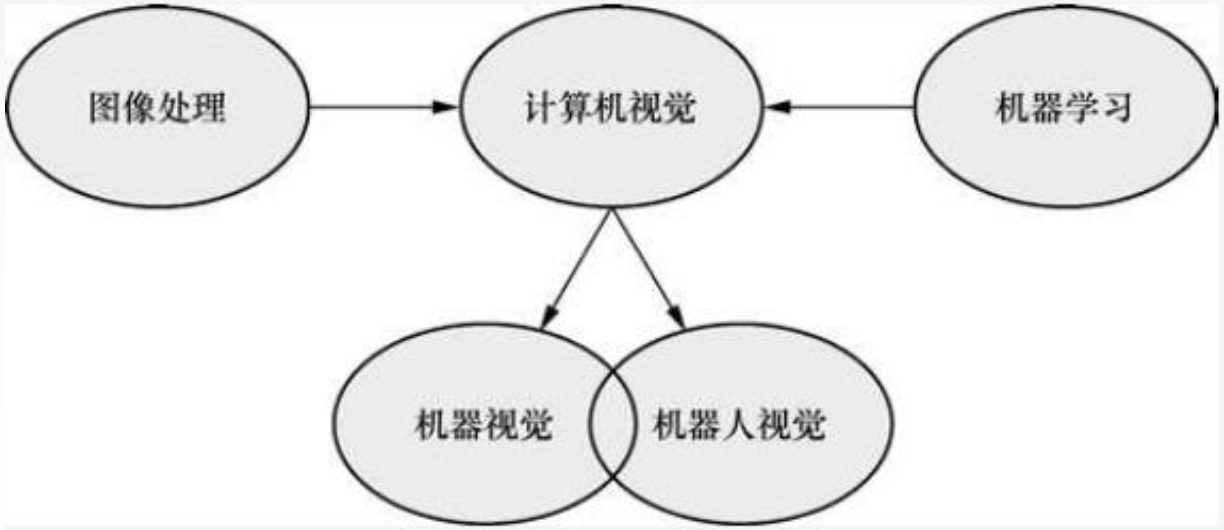
智能机器人视觉

机器人视觉系统

★ 机器人视觉含义

近年来，随着计算机软硬件的发展，以及传感器技术、人工智能等学科的发展，越来越多的研究者关注如何赋予机器人感知世界的能力。智能机器人视觉的研究旨在使机器人能通过视觉感知这个世界，这是机器人路径规划、运动控制等的前提。智能机器人视觉的研究与应用涉及多方面的技术：目标检测、视觉SLAM算法、语义分割、实际应用等

为了更好地理解机器人视觉的含义，必须理清机器人视觉和其他研究领域之间的区别与联系。下图是一个知识图谱，展示了机器人视觉与其他领域之间的联系与区别。



机器人视觉系统

★ 机器人视觉含义

图像处理旨在改善原始图片的质量，或者将图片转换成另一种格式（如直方图等），或者对原始图片进行某些其他改变。

计算机视觉旨在从图片中提取有意义的信息。因此，在实际的应用过程中，可以先使用图像处理技术对原始图片进行处理，然后使用计算机视觉技术从图片中获取需要的信息。

机器学习对于计算机视觉来说非常重要，它可以帮助计算机实现模式识别等任务。在实际应用中，通常可以将计算机视觉和机器学习相结合，使用计算机视觉技术从图片中检测特征和信息，然后将其用作机器学习算法的输入。。例如，在一个缺陷检测系统中，计算机视觉技术会检测传送带上零件的尺寸和颜色，然后机器学习算法会根据已学到的外观知识，判断这些零件是否有故障。

机器人视觉系统

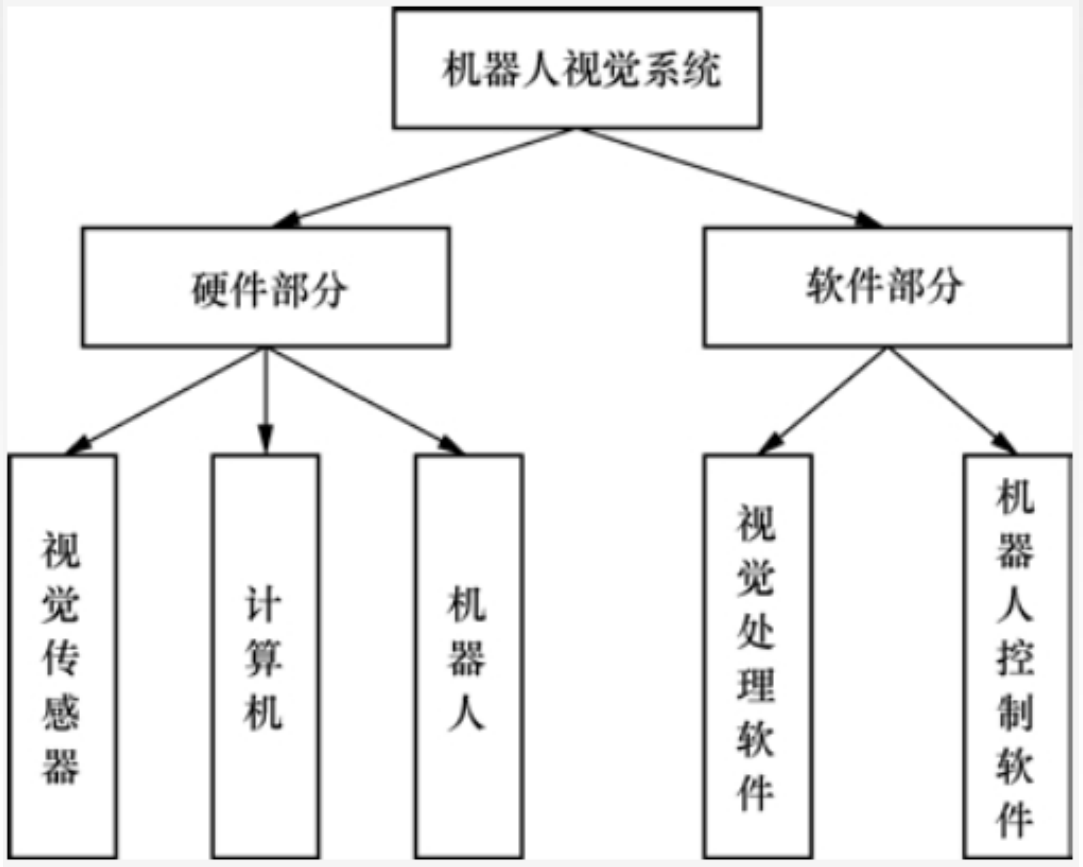
★ 机器人视觉含义

机器视觉和**机器人视觉**都属于计算机视觉的范畴，都是把视觉信息作为输入，获取周围环境的有用信息。在很多情况下，机器视觉和机器人视觉的概念可以互换使用。但是，两者有一些细微的差异。机器视觉是指视觉的工业应用，某些机器视觉的应用（如零件的缺陷检测）与机器人技术无关。而机器人视觉不一定必须是工业应用，且必须将相关的机器人技术纳入其技术和算法中，如机器人运动学、参考系校准及环境的物理感知等。没有机器人视觉，机器人将会失去基本的判断力。

机器人视觉系统

★ 机器人视觉系统的组成

机器人视觉系统可以分为硬件部分和软件部分。硬件部分主要包括以下3个部分：视觉传感器、计算机和机器人。软件部分包含以下部分：视觉处理软件、机器人控制软件。



机器人视觉系统

★ 机器人视觉系统组成

- **视觉传感器**：通常是指相机，相机可按照不同的标准分为标准分辨率数字相机和模拟相机等，要根据不同的实际应用场合选择不同的相机。
- **计算机硬件**：主要由图像采集卡、输入输出单元、模拟-数字转换器、帧存储器和控制装置等构成。
- **机器人**：负责机械的运动和控制。
- **视觉处理软件**：通过对视觉传感器获取环境的图像进行分析和解释，进而让机器人能够辨识物体，并确定其位置。
- **机器人控制软件**：负责输出指令，实现对机器人运动的控制

机器人视觉系统的软件设计是一个复杂的课题，不仅要考虑到程序设计的最优化，还要考虑到算法的有效性，在软件设计的过程中要考虑可能出现的各种问题。机器人视觉系统的软件设计完成后，还要对其鲁棒性进行检测和提高，以适应复杂的外部环境。

机器人视觉系统

★ 单目机器人视觉系统

单目机器人视觉系统只使用一个单目相机。单目视觉系统在成像过程中由于从三维客观世界投影到二维图像上，从而损失了深度信息，这是此类视觉系统的主要缺点。尽管如此，单目机器人视觉系统由于结构简单、算法成熟且计算量较小，在自主移动机器人中已得到广泛应用，如用于目标跟踪、基于单目视觉特征的室内定位导航等。

同时，单目视觉系统是其他类型视觉系统的基础，如双目视觉系统、RGB-D视觉系统等都是在单目视觉系统的基础上，通过附加其他手段和措施而实现的。

机器人视觉系统

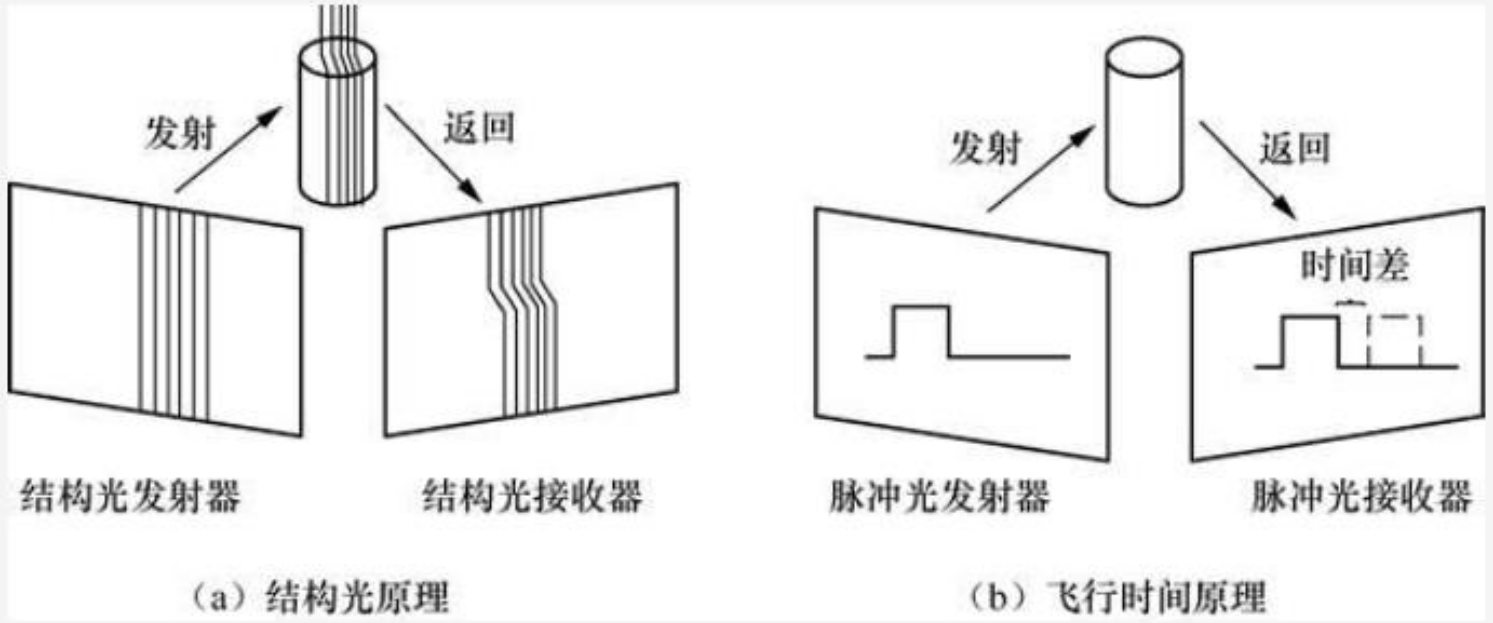
★ 双目机器人视觉系统

双目机器人视觉系统由双目相机组成，利用三角测量原理获得场景的深度信息，并且可以重建周围景物的三维形状和位置，类似人眼的体视功能，原理简单。双目机器人视觉系统需要精确地知道两个摄像机之间的空间位置关系，而且场景环境的3D信息需要两个摄像机从不同角度，同时拍摄同一场景的两幅图像，并进行复杂的匹配，才能准确得到立体视觉系统，能够比较准确地恢复视觉场景的三维信息，在移动机器人定位导航、避障和地图构建等方面得到了广泛的应用。然而，双目机器人视觉系统的难点是对应点匹配的问题，该问题在很大程度上制约着立体视觉在机器人领域的应用前景。

机器人视觉系统

★ RGB-D机器人视觉系统

相比于单目相机和双目相机，RGB-D相机获取深度的方式更为直接，它能够主动测量每个像素的深度。如图所示，RGB-D相机按照原理结构可以分为两种：①通过红外结构光测量像素距离的，如微软的Kinect1代、Project Tango1代、Intel RealSense等；②通过飞行时间（Time-of-Flight，ToF）原理测量像素距离的，如Kinect2代和一些现有的ToF传感器等。



RGB-D相机原理示意图

★ RGB-D机器人视觉系统

一个RGB-D相机不仅包含一个普通的摄像头，还包含一个发射器和一个接收器。因为无论RGB-D相机是哪种结构，它都需要向外界发射一种红外光线。在**基于红外结构光**的RGB-D相机中，相机会根据返回的结构光图案，计算物体离自身的距离。在**基于飞行时间**的RGB-D相机中，会根据光线在目标和相机之间的往返飞行时间确定目标的距离。通过这种方式，深度相机可以获得整幅图像中每个像素点的深度值。不仅如此，RGB-D相机还会自动完成彩色图和深度图之间的匹配，得到像素——对应的彩色图和深度图。

★ RGB-D机器人视觉系统

RGB-D相机的出现，增强了机器人视觉系统对场景深度的感知能力，避免了以往算法对场景深度进行估计的依赖，简化了计算过程。此外，由于RGB-D相机在主动测量方面具有的受光照和纹理影响小的优势，便于RGB-D机器人视觉系统完成构建稠密地图、目标检测与定位和巡检等任务。但是，RGB-D相机成本高，体积大，有效探测距离太短，因此，RGB-D机器人视觉系统普遍在室外表现效果不佳，更多用于室内环境。

视觉目标检测方法

视觉目标检测将目标定位和目标分类结合起来，目标检测是机器人视觉中的一个重要问题。现有的目标检测方法分为**传统的目标检测方法**和**基于深度学习的目标检测方法**。

传统的目标检测方法分为3个步骤：首先，使用不同大小的滑动窗口对待检测图像进行遍历，选择候选区域；其次，从这些区域中提取视觉特征；最后，使用训练好的分类器进行分类。然而，使用滑动窗口法进行区域选择的方法复杂度高且存在大量冗余。另外，手工设计的特征没有很好的鲁棒性。近年来，传统的目标检测方法的性能已经难以满足人们的要求。随着深度学习在图像分类任务上取得巨大进展，基于深度学习的目标检测方法逐渐成为主流。卷积神经网络不仅能够提取更高层、表达能力更好的特征，还能在同一个模型中完成对于特征的提取、选择和分类。目前，基于深度学习的目标检测方法主要有两类：一类是结合region proposal(候选区域)的，基于分类的R-CNN系列目标检测框架；另一类是将目标检测转换为回归问题的算法。

视觉目标检测方法

★ R-CNN

R-CNN模型是由R.吉尔西克等人在2014年提出的。R-CNN模型的结构示意图如图所示。首先，R-CNN模型利用选择性搜索(selective search)算法从输入图像中获得候选区域（约2000个）。然后，对每个候选区域的大小进行归一化，用作CNN网络的标准输入。再使用卷积网络AlexNet提取候选区域中的特征。最后，接一个分类器预测这个区域包含一个感兴趣对象的置信度，也就是说，将目标检测问题转换成了一个图像分类问题。通常这个分类器是独立训练的SVM,当然也可以是简单的Softmax分类器。

视觉目标检测方法

★ R-CNN

R-CNN利用显著性检测方法提取区域。显著性检测方法是一种语义分割中常使用的方法，它通过在像素级的标注把颜色、边界、纹理等信息作为合并条件，多尺度地综合采样方法，划分出一系列的区域，这些区域要远远少于传统的滑动窗口的穷举法产生的候选区域。显著性检测用到了多尺度的思想，可以在不同层级下找到不同的物体。这里的多尺度不是通过缩放图片或者使用多尺度的窗口。显著性检测，通过图像分割的方式将图片分成很多个区域，并用合并的方法将区域聚合成大的区域，重复这样的过程直到整幅图片变成一个最大的区域，这个过程就能够生成多尺度的区域。使用一种随机计分的方式对每个区域进行打分，并按照分数进行排序，取出分值最高的K个子集作为输出。

R-CNN模型有严重的速度瓶颈，原因也很明显，R-CNN模型对所有候选区域分别提取特征时会有重复计算。

视觉目标检测方法

★ R-CNN

R-CNN模型的检测效果相比传统的目标检测方法有较大的提升。R-CNN模型在ILSVRC2013数据集上的准确率达到31.4%，在VOC2007数据集上的准确率达到58.5%。但是，R-CNN模型的实时性不强。它需要对约2000个候选区域分别作特征提取，而候选区域之间存在大量的重复区域，导致大量重复的运算，运行缓慢，每幅图片的平均处理时间高达34s。同时，对每一步的数据进行存储，极为损耗存储空间，实际测试过程中5000幅图像需要数百个GB大小的特征文件。另外，对候选区域进行归一化操作，会对最终结果产生影响。

视觉目标检测方法

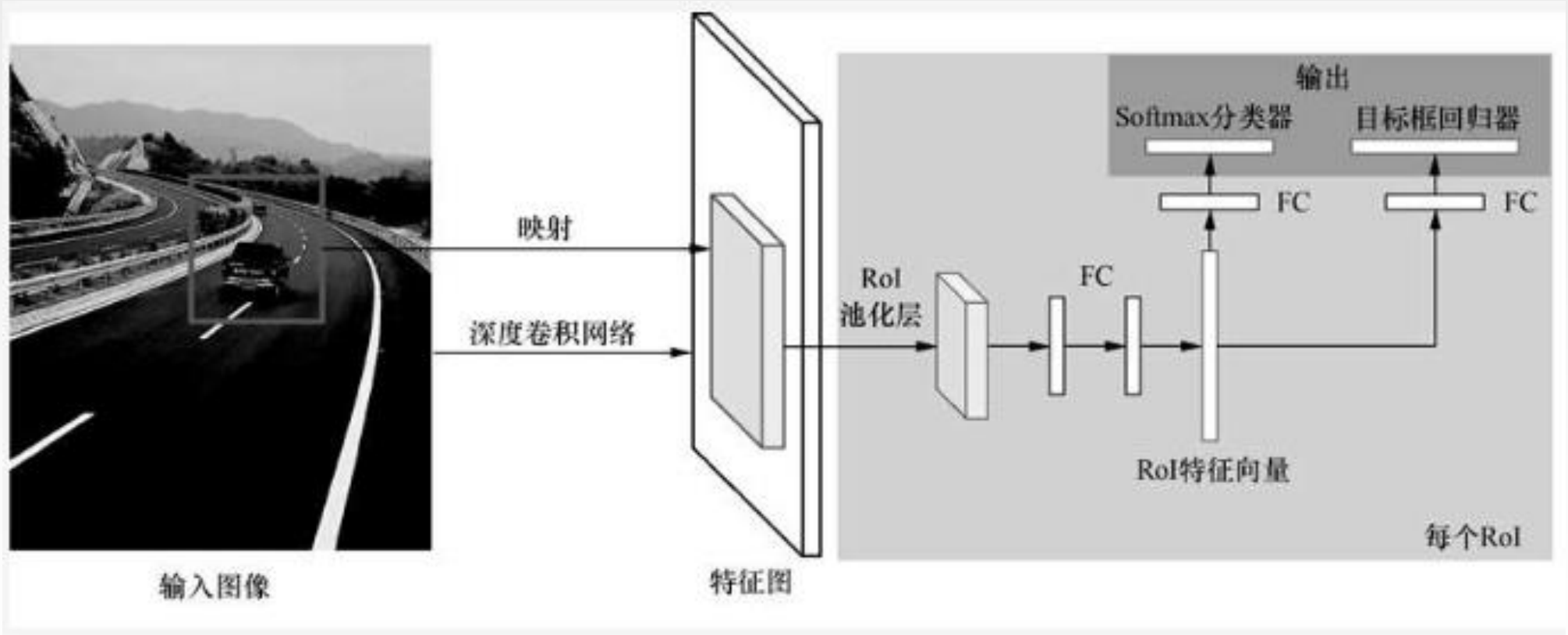
★ Fast R-CNN

Fast R-CNN是一种端到端的目标检测方法，它通过引入空间金字塔池化（spatial pyramid pooling,SPP),避免R-CNN算法对同一区域多次提取特征的情况，从而提高算法的运行速度。Fast R-CNN提出了感兴趣区域（region of interest， RoI）层，这个网络层可以把不同大小的输入映射到一个固定尺度的特征向量，再通过Softmax进行类型识别和通过窗口回归算法进行定位。另外，之前R-CNN的处理流程是先获取区域候选框，然后使用CNN提取特征，之后用SVM分类，最后再做窗口回归。Fast R-CNN把窗口回归放进了神经网络内部，与区域分类合并成一个多任务模型，实际实验也证明，这两个任务能够共享卷积特征，并相互促进。

视觉目标检测方法

★ Fast R-CNN

Fast R-CNN模型的结构示意图如图所示，



视觉目标检测方法

★ Fast R-CNN

以AlexNet（5个卷积层和3个全连接层)作为特征提取网络为例，大致的过程可以理解为：

(1)使用显著性检测在一幅图片中得到约2000个感兴趣区域。

(2)缩放图片的尺寸得到图片金字塔，通过前向传播得到第5层卷积层的特征金字塔。

(3)对于每个尺度的每个感兴趣区域，求取图像到特征图中的映射，在第5层卷积层中截取对应的特征小块。并用一个单层的SPP层统一到一样的尺度，并与全连接层相连，得到同一尺寸的感兴趣区域特征向量（RoI feature vector）。

(4)将感兴趣区域特征向量作奇异值分解（singular value decomposition,SVD）操作，得到两个输出向量：softmax分类器和目标框回归器，得到当前感兴趣区域的类别及目标框。

(5)对所得到的目标框进行非极大值抑制（non-maximum suppression,NMS),得到最终的目标检测结果。

★ Fast R-CNN

感兴趣区域采样层的作用主要有两个，一个是将图片中的感兴趣区域定位到特征图对应的小块(patch)中，另一个是用一个单层的SPP层将这个特征图中的对应区域池化为大小固定的特征再传入全连接层。因为不是固定尺寸的输入，因此每次的池化网格大小需要手动计算，比如，某个感兴趣区域坐标为 x_1, x_2, y_1, y_2，那么输入尺寸为 $(y_2-y_1)\times(x_2-x_1)$，如果池化的输出尺寸为 pooled_height x pooled_width，那么每个网格的尺寸如式所示：

$$\frac{y_2 - y_1}{\text{pooled\_height}} \times \frac{x_2 - x_1}{\text{pooled\_width}}$$

★ Fast R-CNN

在Fast R-CNN中，有两个输出层：第一个是针对每个感兴趣区域的分类概率预测，$p=(p_0, p_1 \cdots\cdots p_K)$；第二个是针对每个感兴趣区域坐标的偏移优化，$t^k=(t^k_x, t^k_y, t^k_w, t^k_h)$，$0 \leqslant k \leqslant K$多类检测的类别序号。第二个输出层，即坐标的偏移优化。假设对于类别K*在图片中标注了一个正样本的坐标：$t^*=(t^*_x, t^*_y, t^*_w, t^*_h)$，而预测值的坐标为$t^*=(t_x, t_y, t_w, t_h)$，二者理论上越接近越好，下式定义了窗口回归的损失函数：

$$L_{loc}(t,t^*)=\sum_{i\in\{x,y,w,h\}} \text{smooth}_{L1}(t_i-t_i^*)$$

$$\text{smooth}_{L1}(x)=\begin{cases}0.5x^2 & ,|x|<=1\\ |x|-0.5, & \text{其他}\end{cases}$$

这里$\text{smooth}_{L1}(x)$中的x即为$t_i-t_i^*$，即对应坐标的差距。这样设置的目的是想让损失函数对于离群点更加鲁棒，从而增强模型对异常数据的鲁棒性。

视觉目标检测方法

★ Fast R-CNN

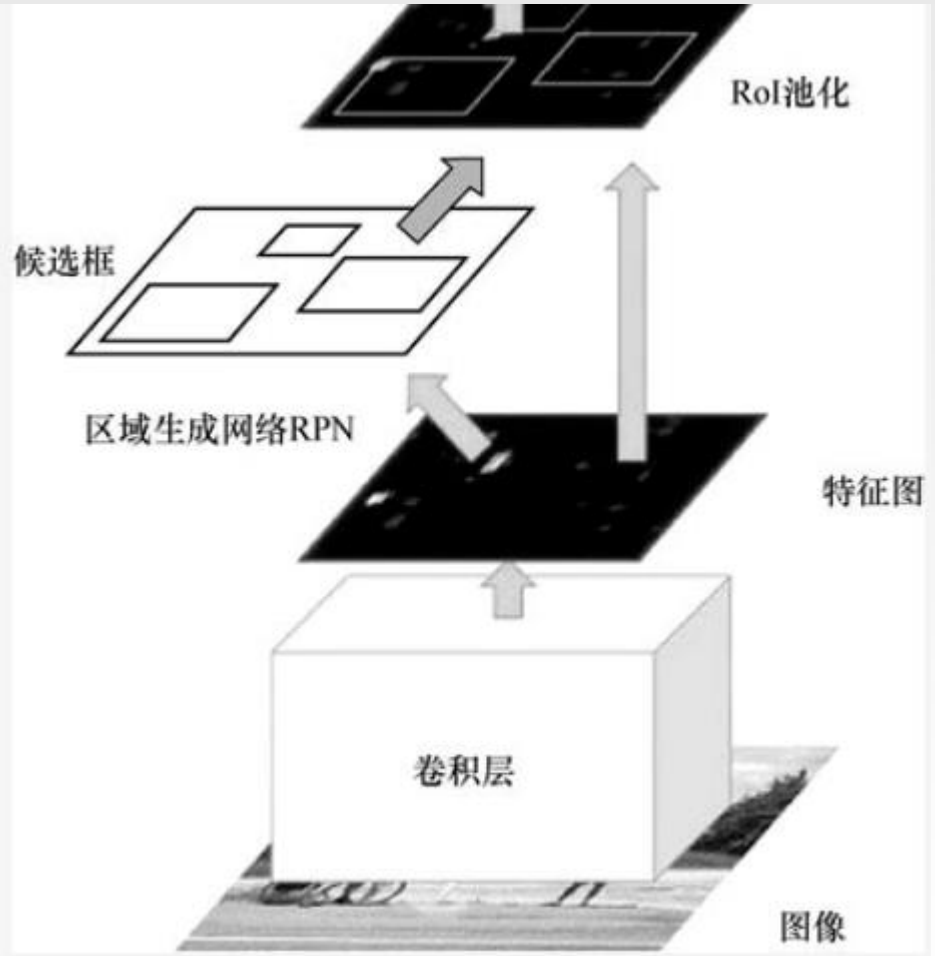
R-CNN的缺点是速度慢，Fast R-CNN提供了一种SPP的方法，使得任意尺寸的输入图像可以学习得到等长的特征，Fast R-CNN是R-CNN和SPP的融合，利用SPP进行加速，整幅图片只需要通过一次卷积神经网络，大大减少了目标检测过程中需要的时间。

训练过程中，FastR-CNN使用多任务的损失函数，成功将分类问题和目标框回归问题进行合并，用Softmax代替SVM,分类和窗口回归在统一的框架中端到端地训练，不需要分开训练，大大减少了训练中的运算开销。同时，SVD操作在保证检测精度的同时，大大加快了检测速度。Fast R-CNN在VOC2007数据集上的平均准确率达到70.0%,且训练速度较R-CNN提升了9倍，每幅图片的检测速度为0.3s(除去获取候选区域阶段)。然而，Fast R-CNN仍然使用选择性搜索算法获取感兴趣区域，这一过程包含大量运算。在CPU上，获取每张图片的候选区域平均需要2s。因此，改进选择性搜索算法是提升Fast R-CNN速度的关键。

视觉目标检测方法

★ Faster R-CNN

R-CNN和Fast R-CNN都存在候选区域生成速度慢的问题。虽然关于候选区域的选择也出现了很多优化的算法，但是始终是在CPU上运行。Faster R-CNN的最大贡献是把区域候选框提取的部分从网络外嵌入网络里，从而一个网络模型即可完成端到端的检测识别任务，不需要手动先执行一遍候选区域提取的搜索算法。如图所示，在得到卷积特征后增加两个额外的层，构造区域生成网络（region proposal network，RPN）。第一层把每个卷积特征编码为一个256维的特征向量，第二层输出这个位置上多种尺度和长宽比的k个区域建议的目标得分和回归边界。



Faster R-CNN框架图

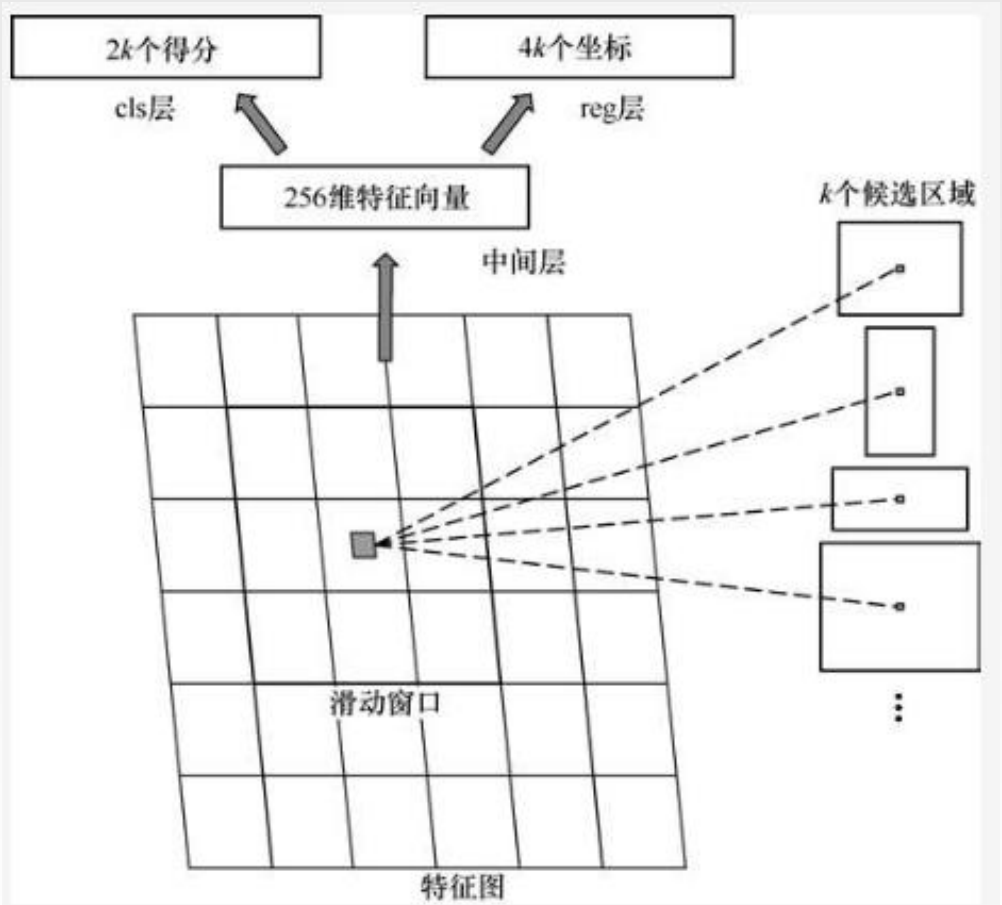
★ Faster R-CNN

具体来说，为了生成区域建议框，在最后一个共享的卷积层输出的卷积特征映射上滑动小网络，这个网络全连接到输入卷积特征映射的n×n的空间窗口上。每个滑动窗口映射到一个低维向量上。这个向量输出给两个同级的全连接层：候选区域回归层（reg层)和候选区域分类层(cls层)。

视觉目标检测方法

★ Faster R-CNN

右图图以这个小网络在某个位置的情况举例说明。在每一个滑动窗口的位置，我们同时预测k个区域建议，所以reg层有4k个输出，即k个目标包围框的坐标编码。cls层输出2k个得分，即对每个候选区域是目标/非目标的估计概率。K个候选区域被相应的k个anchor参数化。每个anchor以当前滑动窗口中心为中心，并对应一种尺度和长宽比。假设使用3种尺度和3种长宽比，这样在每一个滑动位置就有k=9个anchor。对于大小为WH(典型值约2400)的卷积特征映射，总共有W×H×k个anchor。该方法有一个重要特性，就是平移不变性，对anchor和对计算anchor相应的候选区域的函数而言都具有这样的特性。



区域候选网络的结构示意图

视觉目标检测方法

★ Faster R-CNN

RPN使得Faster R-CNN的候选区域生成时间大大加快，获取每幅图片的候选区域平均只需要10ms。目标检测的速度达到每秒5帧。检测精度也大大提升，在VOC2007数据集上的平均准确率达到73.2%。总的来说，从R-CNN、Fast R-CNN、Faster R-CNN一路走来，基于深度学习目标检测的流程变得越来越精简，精度越来越高，速度也越来越快。可以说基于region proposal的R-CNN系列目标检测方法是当前目标检测技术领域最主要的一个分支。

视觉目标检测方法

★ YOLO

从R-CNN到Faster-RCNN，目标检测始终遵循“候选区域+分类”的思路，训练两个模型必然导致参数、训练量的增加，影响训练和检测的速度。而YOLO(you only look once)算法使用端到端的设计思路，将目标物体定位和分类两个任务合并，从图像的像素数据直接获取目标物体坐标和分类概率，目标检测速度达到实时性的要求。

YOLO算法采用针对目标检测任务设计的CNN进行特征提取，随后采用全连接层对识别出来的目标进行分类和位置的检测。YOLO的网络结构中含有24个卷积层和2个全连接层。

视觉目标检测方法

★ YOLO

从R-CNN到Faster-RCNN，目标检测始终遵循“候选区域+分类”的思路，训练两个模型必然导致参数、训练量的增加，影响训练和检测的速度。而YOLO(you only look once)算法使用端到端的设计思路，将目标物体定位和分类两个任务合并，从图像的像素数据直接获取目标物体坐标和分类概率，目标检测速度达到实时性的要求。

YOLO算法采用针对目标检测任务设计的CNN进行特征提取，随后采用全连接层对识别出来的目标进行分类和位置的检测。YOLO的网络结构中含有24个卷积层和2个全连接层。

视觉目标检测方法

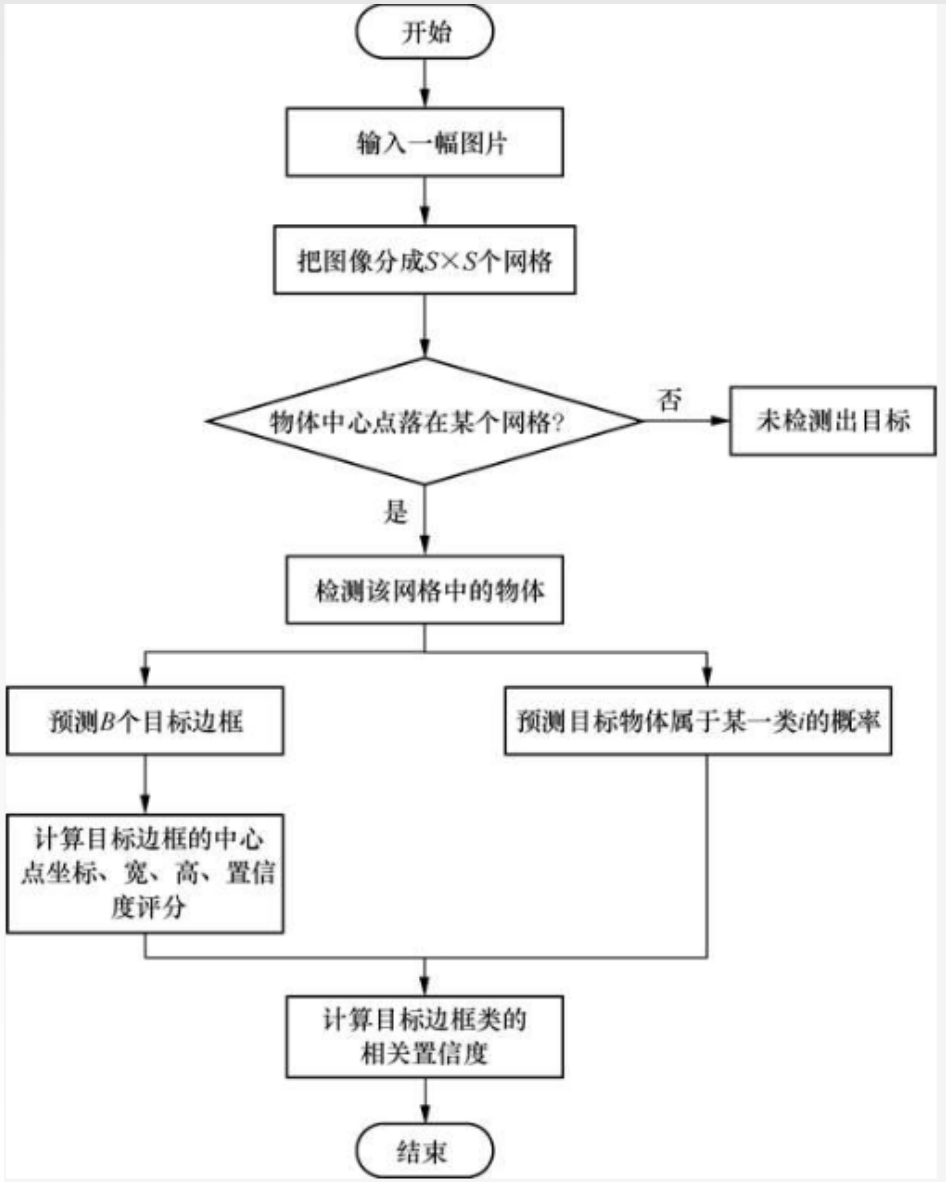
★ YOLO

如图所示。YOLO将输入图像划分为S×S个网格，每个网格负责检测中心点落在其中的目标物体。其中，每个网格中存在B个检测目标，每个检测目标由一个五维度的预测参数(x, y, w, h, s)组成，分别代表目标框的中心点坐标、宽、高和置信度评分。

置信度评分s由式计算得到 $s_i = \Pr(O) \times \mathrm{IoU}$

其中，Pr(0)表示当前网格目标框中存在物体的可能性，0表示目标对象。IoU（intersection over union，交并比）展示了预测边框的准确性。假设预测的目标边框为p，真实的目标边框为t，box_t表示真实的目标边框，box_p表示预测的目标边框

$$\mathrm{IoU}_p^t = \frac{box_p \cap box_t}{box_p \cup box_t}$$



YOLO算法的检测流程

VSLAM方案

★ VSLAM方案

由于视觉传感器在诸多方面具有显著的优势，VSLAM已经成为一个非常重要的研究方向，主要可分为基于特征的VSLAM方法和直接的VSLAM方法。

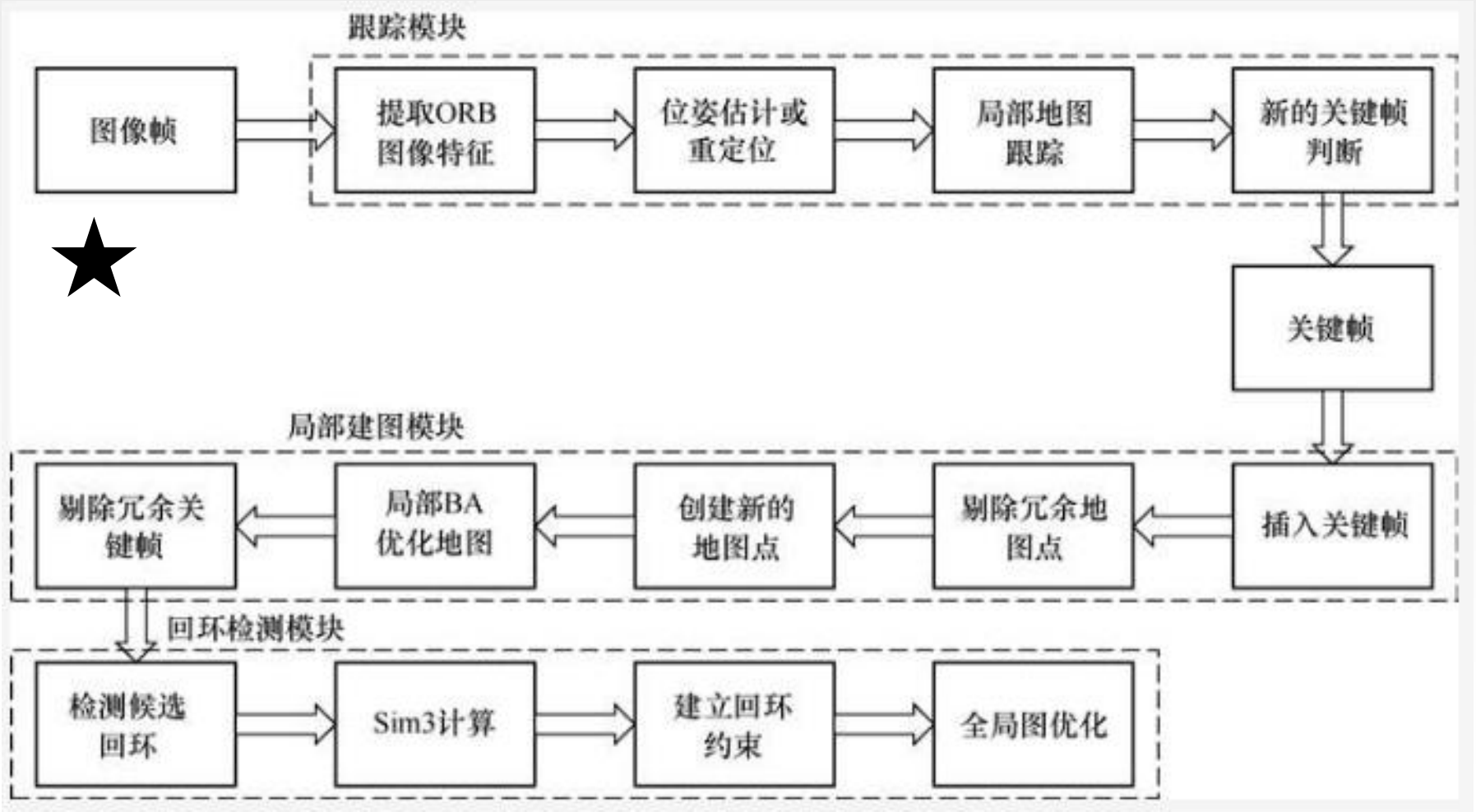
基于特征的VSLAM方法指的是对输入的图像进行特征点检测及提取，并基于2D或3D的特征匹配计算相机位姿及对环境进行建图。如果对整幅图像进行处理，则计算复杂度太高，由于特征在保存图像重要信息的同时有效减少了计算量，因此得到广泛使用。

直接的VSLAM方法指的是直接对像素点的强度进行操作，避免了特征点的提取，该方法能够使用图像的所有信息。此外，提供更多的环境几何信息，有助于对地图的后续使用，且对特征较少的环境有更高的准确性和鲁棒性。

VSLAM方案

★ ORB-SLAM2

ORB-SLAM2是基于特征的VSLAM方案，是当前性能最出色的VSLAM系统之一。ORB-SLAM2的系统框架如图所示。它主要包含3个并行线程：跟踪（tracking）、局部建图（local mapping）、回环(loop closure)。



VSLAM方案

★ ORB-SLAM2

跟踪线程的主要任务是对输入的每一帧图像提取ORB图像特征并估计相机位姿，其跟踪状态随环境变化或相机运动等因素而变化。其跟踪模型分为运动模型、参考帧模型、重定位模型。虽然不同跟踪模型的输入数据存在差异，但其目标都是求解初始相机位姿。ORB-SLAM2利用了非线性优化的思想，跟踪线程首先通过EPnP（efficient perspective n-point）算法估计初始相机位姿，然后构建最小二乘优化问题对初值进行优化，这种优化问题称为BA（bundle adjustment）问题，在EPnP中，BA将空间点与相机位姿同时看作优化变量，其优化目标是最小化重投影误差（reprojection error）。

VSLAM方案

★ ORB-SLAM2

局部建图线程主要负责接收处理新的关键帧，增加新的地图点，维护局部地图的精度和关键帧集合的质量与规模。具体步骤如下。

(1)处理新的关键帧。首先处理当前关键帧的词袋向量；然后更新当前关键帧的地图点观测值，并将这些地图点添加到当前新增地图点列表中；最后更新共视图和本质图，并将当前关键帧加入地图中。

(2)地图点的筛选。通过检查当前新增地图点列表，按规则剔除冗余点，剔除规则为：①该地图点被标记为坏点；②地图点能够被观测到的关键帧数量不超过25%；③能够观测到地图点的关键帧不超过3个，单目情况下为2个。

(3)根据当前关键帧恢复新的地图点。首先，从共视图中选取当前关键帧附近的关键帧；然后，对当前关键帧和选取出的关键帧进行特征匹配，获得匹配特征点的归一化坐标并构建对极约束，通过对极几何计算出当前关键帧的位姿；之后，通过三角化恢复特征点方法计算获得特征点的深度。最后，根据所获得的特征点深度计算恢复出新地图点的重投影误差，并根据误差与给定阈值的关系确定地图点是否被剔除。

★ ORB-SLAM2

(4)局部BA。当新的关键帧被增加到共视图中时，通过执行2次迭代优化与外点剔除，完成局部地图点与位姿的优化。

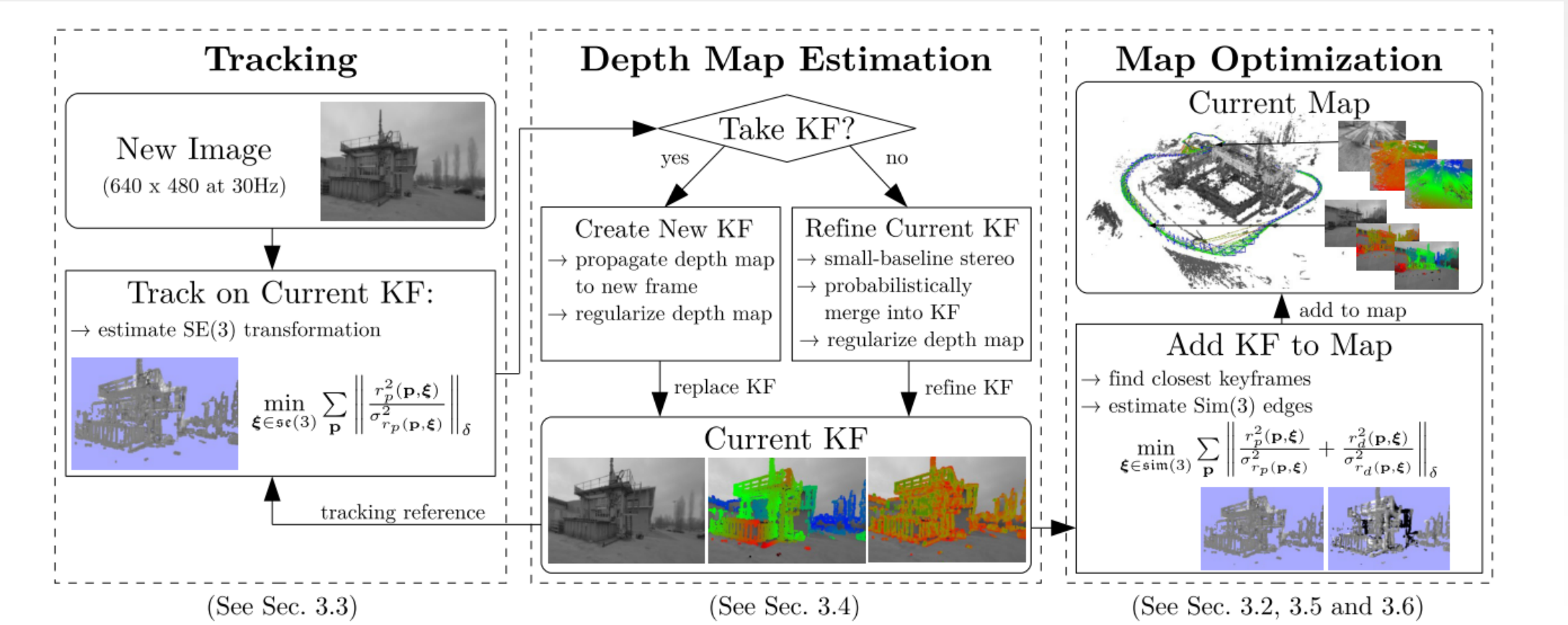
(5)局部关键帧筛选。ORB-SLAM2筛选冗余关键帧的标准为：若关键帧能够看到的90%的地图点能够被其他3个以上关键帧观察到，则剔除该关键帧。

回环线程包括回环检测和后端优化两部分。回环检测负责筛选并确认回环，首先计算当前关键帧与相连关键帧的BoW分值，并以最低分为阈值选取回环候选帧，然后统计共有单词数量和聚类得分，剔除质量不高的独立关键帧，并对留存的候选关键帧进行连续性检测；检测到回环后利用RANSAC（random sample consensus）框架求解相似变换Sim3,然后通过再匹配和g2o优化Sim3,校正当前关键帧的位姿。后端优化部分负责消除全局的累积误差，首先利用传播法调整与当前关键帧相连的关键帧位姿，并更新对应的地图点，最后根据调整的地图点更新关键帧的链接关系；在完成地图融合之后，通过本质图进行位姿图优化。

VSLAM方案

★ LSD-SLAM

LSD-SLAM是单目VSLAM中功能完备的优秀算法。此算法采用直接法，只利用梯度比较明显的像素点，就可以完成高精度的位姿估计、跟踪和回环检测等任务，并能构建大规模且全局一致的半稠密地图。整个LSD-SLAM分为3个部分：跟踪、深度图估计和全局地图优化。



VSLAM方案

★ LSD-SLAM

（1）Tracking线程：首先将当前图像构造为新的普通帧（当前帧），如果它的参考关键帧不是最近的关键帧的话就先更新参考关键帧（就说我反正要跟踪离我最近的关键帧），把上一帧与参考关键帧的位姿当做初始位姿，然后构建归一化方差的光度残差的代价函数进行优化求解当前帧与参考关键帧之间的位姿变换，然后进行跟踪是否失败的判断和关键帧筛选。

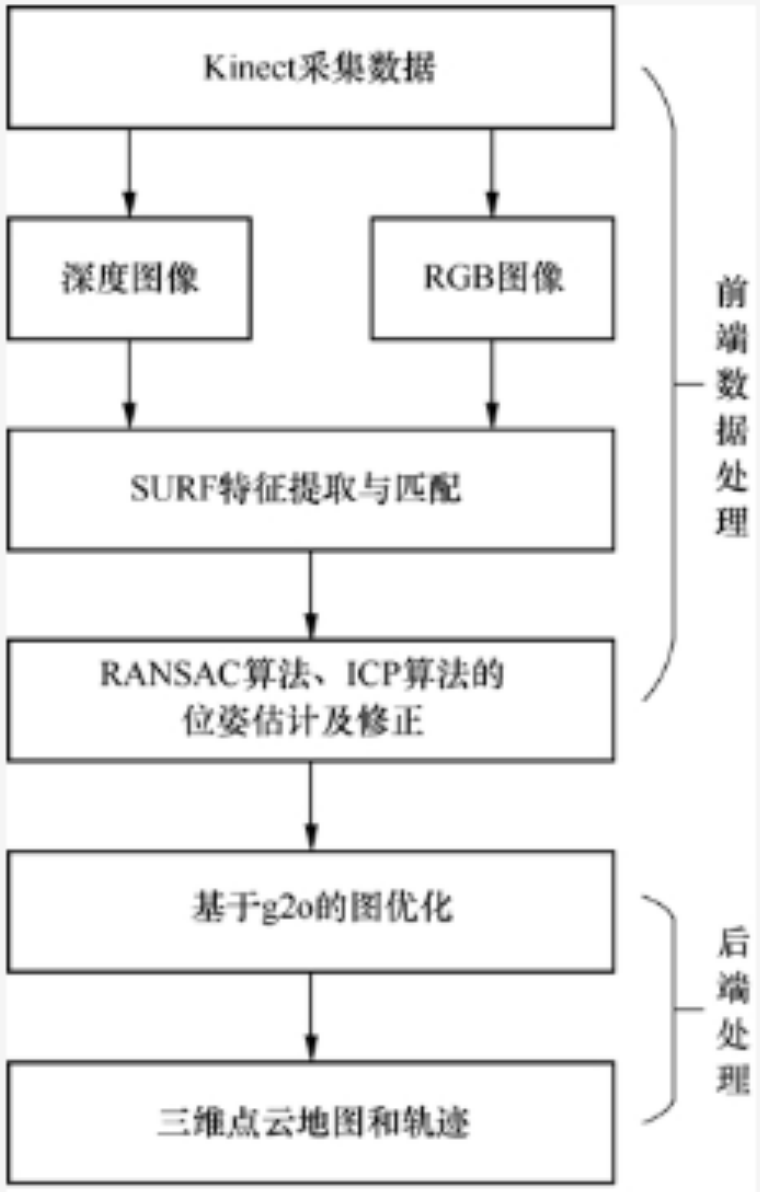
（2）Depth Map Estimation线程：首先判断其是否为关键帧，如果是关键帧，则根据它的参考关键帧构建新的深度图，如果不是关键帧，则用来更新它的参考关键帧的深度图。

（3）Map Optimization线程：如果新关键帧队列为空则在优化过的帧中随机选取图像帧，如果不为空则从新关键帧的第一帧开始选取，主要是根据视差、关键帧连接关系，判断选取的图像帧是否为候选帧，然后对每个候选帧和测试的闭环关键帧之间进行双向Sim3跟踪，如果求解出的两个李代数满足马氏距离在一定范围内，则认为是闭环成功，并且在位姿图中添加边的约束，最后进行全局的图优化。

VSLAM方案

★ RGB-D SLAM

在众多基于深度相机的VSLAM研究中，弗莱堡大学Endres等提出的RGB-DSLAM算法是最早的方法之一。它具有精度较高、鲁棒性好的优点，可以实时获取移动机器人的当前位置和姿态，已得到了广泛的认可。RGB-DSLAM算法的流程图如图所示。



VSLAM方案

★ RGB-D SLAM

算法分为图像前端处理和后端位姿优化两部分。在算法中充分利用了环境RGB信息和深度信息。

首先根据Kinect传感器实时获取的RGB信息提取图像特征，与前期获取的彩色图像进行图像特征匹配。通过对提取的特征所在点的深度信息进行评估，得到任意两帧图像之间的一系列相互对应的3D点对。利用迭代最近邻(ICP)算法得到当前帧与历史帧的平移、旋转向量，进行移动机器人的位姿估计。以此为基础，使用随机采样一致性(RANSAC)算法对不同帧之间的对应点进行优化估计。由于不同帧之间对应的点对位姿估计不一定是全局一致的，因此使用g2o对RANSAC算法得到的不同帧之间的对应关系进行优化，得到RGB-D传感器相对于初始位姿的当前帧位姿关系。同时，融合不同帧之间的数据，得到融合后的3D环境点云数据。由于点云数据量过大，需要进行像素化，对数据进行压缩，便于运算以及存储，最终获得3D点云地图以及移动机器人运动的轨迹。

深度学习在机器人视觉中的应用

近年来，深度学习已经在机器人视觉中崭露头角，深度学习的发展不仅突破了很多难以解决的机器人视觉难题，提升了对于图像认知的水平，加速了机器人视觉相关技术的进步，更重要的贡献是改变了处理机器人视觉问题的传统思想。本节总结了这些年将深度学习引入回环检测、语义地图、三维重建、人脸识别等相关领域，以及遇到的挑战和技术难点。

深度学习在机器人视觉中的应用

★ 回环检测

早期的回环检测方法大多基于场景不变的假设，这些方法在稳定的室内环境下尚能正常运行，但面对复杂的场景变化，如光照变化、季节变化、视角变化、动态场景等时，检测的准确率和回召率会大大降低。近年来，随着深度学习的快速发展，基于深度学习的回环检测受到了国内外研究者的广泛关注，并取得优异的性能。现在基于深度学习的回环检测研究主要集中在场景描述上。

场景描述方法主要包括：**全局特征描述子；局部描述子；局部区域的全局描述子；结合深度信息的场景描述；场景的时变描述。**

深度学习在机器人视觉中的应用

★ 回环检测

在全局特征描述子的方法中，Chen等人首先使用深度神经网络进行回环检测研究，通过预训练的CNN模型提取图片特征，然后用于相似性比较，证明CNN提取出的特征相较于传统的手工提取的特征能更好地应对环境变化。紧接着，Chen等人又在一个大规模场景数据集SPED上训练了两个用于场景分类的CNN模型：AMOSNet和HybridNet，取得了更好的回环检测性能。Sunderhauf等人仔细分析了AlexNet的各层特征，证明了AlexNet的中层特征能够应对环境外观变化，AlexNet的高层特征能够应对环境的视角变化。虽然使用现有的CNN生成全局图像描述子的方法获得了很多应用，但这种方法仍然存在一些问题，如提取到的特征描述子的维度较高，不够有区分度，对动态区域敏感，不能有效应对复杂的场景变化等。其根本原因在于这种方法是专门用于图像分类设计和训练的，因而它并不具备专门用于闭环检测任务的网络该有的特点。为了更好地应对回环检测任务，通常需要在特定场景下重新设计或者微调这些网络模型。

★ 回环检测

在局部描述子的方法中，Gao等人将每幅图片分成许多的图片块，然后离线训练了一个堆栈自编码器（stacked denoising auto-encoder,SDA)提取图片的局部特征，用于相似性比较。Li等人将图片分成图片块，使用CNN模型提取图片的局部特征并构造相似性矩阵，通过一种自适应加权方案确定图片相似度。但是，这些方法的特征提取和相似性比较的过程都非常耗时。

深度学习在机器人视觉中的应用

★ 回环检测

在局部区域的全局描述子方法中，使用全局描述的方法对图像的局部区域生成描述子。局部区域可使用各种局部区域探测器生成，各个局部区域的描述子合在一起形成对当前场景的描述。这种方法的关键在于如何生成稳定的局部区域，使其在环境条件发生变化时也能保证一定的可重复性。近年来，随着目标识别领域的发展，出现了很多更加优秀的物体提案方法，如RPN网络和EdgeBoxes算法等。相比于RPN通过学习的方法获得特定目标的潜在区域，EdgeBoxes算法通过方框内部轮廓信息量的大小判断其是否包含物体，因而它具有通用性，并不局限于特定目标的物体提案生成。N.Sunderhauf和S.Cascianelli等人的文章中描述采用这种方法生成局部区域，然后使用CNN生成局部区域的全局描述子。但是，这种方法的缺点是很难实时运行，一方面，因为现有的算法都很耗时；另一方面，要提高算法的稳定性，需要增加局部区域的数量，而每个区域都需要CNN前向传播提取特征，这比单纯地使用全局描述要更加复杂。

深度学习在机器人视觉中的应用

★ 回环检测

在结合深度信息的场景描述的方法中，深度信息可结合语义分割，从而生成更高级的语义特征描述场景，从而增强对环境的认知能力。对回环检测而言，由深度信息结合图像信息建立的语义特征，不仅增强了对外观变化和视角变化的适应能力，而且简化了地图描述，节省了存储空间，因为语义地图只需要存储特征的语义标签即可，而不是整个三维信息。此外，深度信息还可结合多视图几何对生成的回环进行验证，剔除错误的回环。

★ 语义地图

然后是基于超体元的三维目标物体点云分割，以进一步分割出前述基于图像划分得到的物体所对应的点云；最后是基于最近邻方法的物体数据关联，以确定当前物体和地图中物体之间的对应性，进而添加或更新地图中目标物体的点云信息和从属类别置信度等数据。

Salas-Morenoe等人提出的SLAM++系统将环境语义信息结合到3D地图中，构建了场景的3D语义地图。目前VSLAM系统大多使用点、线等低级视觉特征进行数据关联，其在长时、大尺度的环境中受到很大限制，利用更高层次的语义信息在一定程度上可以提高系统的鲁棒性。实际上，从长远来看，几何地图构建的过程和地图的语义标注过程可以是相互促进的两部分。目前国内外对地图构建与深度学习技术的研究尚处于起步阶段，缺乏广泛的探索及深层次的研究。

深度学习在机器人视觉中的应用

★ 三维重建

近年来，人工智能技术飞速发展，三维重建作为环境感知的关键技术之一，可用于自动驾驶、虚拟现实等。如何基于深度学习对场景进行准确的三维重建，使机器人具有一定的视觉感知能力呢？

D.Eigen等人改进传统单一尺度的CNN,提出多尺度CNN,并针对深度预测提出尺度不变损失函数，实现对单幅图像的深度估计。H.Jung等人使用条件生成对抗网络实现单张图像深度估计，采用基于编码器-解码器与精炼网络相结合的生成器网络，在客观数据集上达到了较好的实验结果。I.Laina等人使用残差结构设计网络，并提出快速上卷积网络，在NYUD v2数据集上取得了优异的表现。S.Zagoruyko等人提出基于CNN的匹配相似度计算网络。该网络输入为两幅图像块（由双目相机拍摄），网络直接输出两幅图像块的相似度。解决基线距离较大时的立体匹配问题。H.Fan等人提出的点集生成网络，研究了如何通过单幅图像实现三维重建，该网络的输入为单幅图像，输出为点集的三维坐标。Y.LeCun等人提出用于计算立体匹配中匹配代价的CNN，该网络解决了基线距离较小时的匹配代价计算问题，且输出的图像块较小。B.Ummenhofer等人提出DeMoN网络，实现从非约束的图像对中获取深度信息和相机的运动参数。将深度学习方法用于三维重建，相较于传统算法均有提升，但仍存在一些问题，如网络模型的泛化能力、准确度等。

深度学习在机器人视觉中的应用

★ 人脸识别

基于深度学习的人脸识别技术受到机器人视觉发展的推动而得到提升，是机器人视觉领域的一个重要分支。

现在人脸识别的主要技术路线的步骤为：

第一步，使用相机拍摄有效的人脸图片，可以为静态图像也可以为动态形式的视频帧；

第二步，选择有效算法对人脸图像提取所需要的人脸特征，建立特征模型库；

第三步，判别待分类的人脸图像在系统人脸特征库中有无该类模板，根据相似度大小判别需识别图像对应的对象身份信息。

深度学习在机器人视觉中的应用

★ 人脸识别

Taigman等通过3D模型对400万幅的人脸图像进行对齐处理，仅利用CNN模型获取人脸的表征信息。后来，随着CNN在网络层次上的不断加深，结构的不断复杂，出现了FractalNet、ResNeXt等，将网络进行融合，它们都对人脸识别技术的发展提供了强有力的推动。在2015年，谷歌的FaceNet在LWF数据集上的识别率大大提高，达到了99.63%的准确率。在2016年，由石世光领导的中国科学院计算所研究团队提出了一种基于C++代码的SeetaFace人脸识别引擎，该识别引擎包括实现全自动人脸识别系统的全部模块，不依赖第三方库。该引擎奠定了整个人脸识别社区的基准，能够实现对面部特征点的自动检测定位，获得相应的人脸特征信息，得到对比模块。同时，在商业方面，人脸识别技术得到广泛应用。在2018年，中国著名的芯片研发公司瑞芯微发布AI人脸识别一站式的解决方案，使得人脸识别应用的场景化、商业化更丰富。腾讯和百度等公司的人脸识别App产品在市场上占有优势，离不开对深度学习相关最新技术的研究与应用。