摘要

语义分割是计算机视觉领域中的一个十分重要的分支，传统方法多是直接提取图像的RGB信息实现对场景的像素级分类，这些方法在物体的表面、内部等连续区域能较为准确的预测类别，但是在物体的边缘部分，仅使用RGB信息的方法不能表现出了足够的鲁棒性，所以场景深度信息被引入，深度信息可为RGB图像提供对应的几何关系，其提供的额外结构信息能够有效辅助复杂和困难场景下的分割，能够有效修正物体边缘像素处出现的分类错误。而光场深度提取是深度提取领域中精确度领先的一个分支，光场图像中包含来自场景的多视角信息，这使得深度估计成为可能。相对于传统的多视角深度估计问题，光场视角更加丰富且均匀整齐地排列在规则网格上，视角间场景几何信息变换规律更加清晰，使深度提取的难度大大降低。从光场中提取出的深度信息与RGB信息融合后的特征被应用于场景语义分割，有效提升了像素的分类精度。

本文提出了基于光场的语义分割网络，在光场图像中获取RGB信息以及精确的场景深度信息并融合成特征信息从而获取语义分割结果。

关键词：语义分割，深度提取，光场

语义分割一直是计算机视觉中非常重要的领域。从传统语义分割的发展到深度学习的普及，语义分割的任务和方法也得到了广泛的扩展和改进。在本章中，我们首先解释语义分割的背景和定义，然后从本文开始概述近年来的各种语义分割解决方案，包括这些算法提出的背景，设计思路和需要解决的问题，同时我们仍将介绍语义分割领域中当前使用的主流数据集。本章简要介绍了深度学习上下文中从浅层模型到深层模型的各种语义分割技术，包括这些技术的背景，方法和问题。如今，语义分割已成为计算机视觉的关键问题之一。通过语义分割生成的分割结果被应用到静态2D图像，视频甚至3D数据和体数据中，被广泛用于计算机视觉的各种高级应用中。从宏观意义上讲，语义分割是一个高级任务，为场景理解铺平了道路。作为计算机视觉的核心问题，场景理解的重要性越来越突出，因为现实中越来越多的应用场景需要从图像中推断出相关的知识或语义（即从具体到抽象的过程）。这些应用包括自动驾驶，人机交互，计算摄影，图像搜索引擎，增强现实等。使用各种传统的计算机视觉和机器学习技术，已经解决了这些问题。尽管这些方法非常流行，但是深度学习革命带来了相关领域的翻天覆地的变化，不仅大大提高了语义分割的准确性，而且加快了场景分割结果的生成速度。因此，包括语义分割，视觉问题在内的许多计算机开始使用深度网络架构来解决，通常是卷积神经网络CNN，而CNN的准确性甚至效率远远超过传统方法。但是，与固有的计算机视觉和机器学习分支相比，深度学习的发展还远远没有成熟。因此，语义分割的发展空间仍然很大。这个领域的快速发展使得对初学者的教育更加困难。此外，由于已经陆续提出了大量工作，因此跟上发展的步伐也非常耗时。因此，很难跟上与语义分割有关的工作，合理地解释它们的论据，过滤掉底层工作，并验证相关的实验结果是非常困难的。

语义分割是在像素级别对图像进行分类。语义图像分割的目的是用相应的类标记图像的每个像素。因为在此任务中我们正在预测图像中每个像素的分类，所以语义分割任务通常称为密集预测。

在完全卷积网络出现之前，早期的研究经常使用图像块方法进行语义分割，即将图像切成块并输入到深度模型中，然后对图像块中的像素进行分类。这些分类网络通常包含一个全连接层，并且全连接层的输入需要固定大小的图像块，因此图像被切成块。构造神经网络模型的一种相对基本的方法是简单地堆叠几个卷积层（使用相同的填充以确保维数不变）并输出最终的分割图。该模型通过一系列特征映射转换直接学习，以获得从输入图像到输出分割结果的映射。但是，由于图像尺寸在整个过程中不会改变，因此整个网络都以全分辨率参与计算，并且该过程带来的计算量非常大。为了降低网络学习的计算复杂度，需要提出新的网络分段结构。这些结构中最流行的图像分割模型之一是Encoder-Decoder（encoder-decoder）结构。编码器部分通过下采样降低输入的空间分辨率，从而生成低分辨率特征图（有效计算并有效区分不同类别）；解码器对这些特征描述进行上采样，并将其恢复为全分辨率分割图，这种结构不仅可以保持语义分割的准确性，而且可以大大减少深度网络对场景特征的提取。2014年，提出了全卷积网络[1]（Fully Convolutional Networks，FCN），这促进了先前分类网络的CNN结构。本文提出用卷积层代替全连接层以实现像素级密集预测。 FCN该结构可以接受任何大小的输入图像，并且去卷积层用于对最后一个卷积层的特征图进行升采样，以将其恢复为输入图像的相同大小，从而为每个像素生成一个预测，逐像素实现像素分类。 FCN已成为语义分割网络研究的基本结构，并在此基础上进行了许多研究。编码器-解码器结构也首先引入了FCN体系结构。编码器使用池化层逐渐减小空间尺寸以获得场景特征图，而解码器使用上采样方法恢复空间尺寸和详细信息，最后获得与输入尺寸一致的特征图。通过上层网络和下层网络之间的跳转，在保留特征提取效率的同时保留足够的像素级细节，除了FCN，经典语义分割网络SegNet和UNet是此类结构的典型应用。SegNet [2]的建议概括了以前的Encoder-Decoder结构，其在编码网络中的连续下采样导致输出特征图上的每个像素与输入图像中的大量空间信息重叠。最大池化层在解码网络中被重用。此应用程序改进了边界划分，并减少了端到端训练参数的数量。与以前的编码和解码网络相比，SegNet在速度和效率上有了显着的提高。随后，Ronneberger等。通过扩展解码器模块的容量，改进了上述全卷积架构。具体来说，此U-Net [3]结构“包括用于捕获上下文的收缩路径和用于精确定位的对称扩展路径”。到目前为止，这种结构非常流行，并且已经扩展到各种分割问题。在UNet的基础上，Drozdzal [4]和其他人用堆栈中的剩余模块替换了基本卷积模块。该剩余模块在模块内部包含跳过连接，同时在编码器和解码器对应的特征图之间保留与U-Net相同的跳过连接。他们声称这种方法可以使网络收敛更快，也可以应用于更深的网络结构。同时，基于这些工作，Jegou [5]等。提出在网络结构中使用密集块，这也符合U-Net结构。他们声称，密集块的属性使它们更好地适应语义分割任务，因为它们自然携带跳跃。连接和多尺度监督。这些密集块之所以有效，是因为它们携带了从上一层获得的低级特征，并且还包含了从下一层获得的较高级特征，从而实现了更有效的特征重用。后来，有人[6]提出了使用圆环/膨胀卷积来应用完全连接的CRF的ztrous空间金字塔合并（ASPP）。扩张卷积在不增加参数数量的情况下实现了接收场的增加。多尺度处理可以通过将原始图像的不同尺度同时传递到并行CNN分支（图像金字塔）中，或者通过（或同时使用）具有不同采样率的多个并行多孔卷积层（ASPP）来实现。通过完全连接的CRF实现结构化的预测。 CRF作为独立的后处理步骤进行了单独培训。然而，应用膨胀/无穷卷积设计方法的局限性在于，膨胀卷积的计算量很大，并且要处理大量高分辨率特征图需要大量内存。该点以高分辨率输入到预测计算中。例如，DeepLab的预测是原始输入大小的1/8。因此，RefineNet [7]论文建议使用编码器-解码器结构。编码部分是ResNet-101模块，解码器包括RefineNet模块，该模块将编码器的高分辨率功能与以前的RefineNet模块的低分辨率功能进行级联/融合。每个RefineNet模块包括通过对低分辨率功能进行上采样来融合多分辨率功能的组件，以及基于5x5步长为1的池层捕获内容的组件，上述两个组件中的每个组件均基于以下组件应用残余连接设计：在身份地图思维方式上。但是，编码和解码网络的结构不能支持从输入图像中提取足够的特征信息。在后来的发展中，为了弥补这一不足，语义分割领域出现了越来越多的网络，例如PSPNet [8]，PSPNet。网络结构中有一个复杂的金字塔池层，用于提取不同规模的信息。在功能图中。与UNet相比，PSPNet具有更复杂的网络结构，并且在多类分类任务中通常更强大。太棒了全局场景类别的意义在于，它为细分类别的分布提供了线索。金字塔池模型通过应用大型核的池层来捕获此信息。如本文中所述，使用了扩展卷积来修改Resnet，并添加了金字塔池模型。该模型在上采样后集成了ResNet的特征图。除了主分支上的损失外，辅助损失还应用于ResNet的第四阶段。这个概念在其他地方也称为中间监督。

后来[9]提出具有很大核心的编码器-解码器结构的语义分割需要同时对分割对象进行分割和分类。考虑到全连接层不能完成分割任务，本文用大尺寸内核的卷积运算代替了全连接层。

应用大尺寸核心的另一种意愿是，尽管像ResNet这样的更深的网络结构可以具有更大的接受域，但研究表明，该网络实际上是从较小的区域（有效的接受域）收集信息的。大尺寸内核具有很高的计算成本和众多参数，因此可以将kxk卷积与1xk + kx1和kx1和1xk卷积。在本文中，该模型称为全球卷积网络（GCN）。回到结构本身，ResNet（没有任何扩展的卷积）构成编码层，GCN和解卷积构成解码层。该算法还使用简单的残差模块边界细化（BR）。

除了在固定的二维场景中进行语义分割外，[10]还提出了空间路径和顶点路径来实现实时语义分割。 BiSeNet与U形和Dilation结构不同，请在维持空间上下文和空间细节的同时尝试一种新方法。因此，本文设计了空间路径和上下文路径的两个部分。顾名思义，空间路径使用更多的通道和较浅的网络来保留丰富的空间信息，以生成高分辨率特征。上下文路径使用更少的通道和更深的网络来快速下采样以获得足够的上下文。基于这两个网络的输出，还设计了功能融合模块（FFM）以融合这两个功能。 TDNet [11]本文提出了一种基于时间分布式网络的视频语义分割算法TDNet（时间分布式网络）。 ，其设计思想来自这样的观察：可以通过组合一系列浅网络输出的特征来获得较深网络输出的特征。在视频分割任务中，视频具有定时属性，视频分割也具有定时属性。因此，可以在一定的时间范围内将浅层网络用于特征提取，然后组合以实现深层网络输出特性。影响。这种思维迁移也很简单，那么如何实现本文中提到的多个浅层特征的组合？本文中的解决方案是使用新的APM（注意力传播模块）来实现这一目标，从而减少细分目标移动对细分性能的影响。另外，引入了知识提纯的概念以使知识能够在浅层网络输出特征和组合特征的两个级别上进行传递，从而提高学生网络的性能。

[12]本文提出了一种轻型版本的RefineNet，即轻型RefineNet，它将实时分割任务的速度从20FPS提高到55FPS。整个网络结构分为下采样编码器部分和上采样解码器部分。该网络主要包含4个模块，RCU，CRP，FUSION和CLF。为了减轻网络负担，分别使用了RCU-LW和CRP-LW。

FUSION-LW取代了原始网络的RCU，CRP和FUSION。通过随后的实验，作者发现RCU对网络精度的影响较弱，因此也删除了RCU模块。为什么要卸下RCU模块，对网络精度影响不大，因为（1）尽管RCU模块中的3 \* 3卷积使网络具有较大的接收范围，但通过快捷方式结构，低级功能和高级功能功能也可以共享。 （2）CRP模块还可以获取上下文信息。

语义分割在深度学习时代取得了长足的进步。但是，从以上评论可以看出，语义分割中仍有许多问题需要克服。目前，还远远没有解决。更准确的分割边界，小对象分割，实时语义分割和其他问题仍然是一个挑战，它需要学术界和业界的不断努力。

在对语义分割的结果做评价时，一般会使用到以下的定量计算标准：

1.像素精度（Pixel accuracy）：每一类像素正确分类的个数/ 每一类像素的实际个数。

2.均像素精度（Mean pixel accuracy）：所有类像素精度的均值。

3.图像交并比（Intersection-over-Union，IoU）：是检测结果与实际结果的交集和并集的比值，最理想情况是完全重叠，即比值为1。

我们已经讨论过许多基于RGB信息的分割网络了，今天就来谈谈RGB-D分割。RGB-D分割中的D指的是“Depth”，即“深度”，也就是相机到物体在实际空间中的距离。那么既然RGB信息已经能够实现分割，为什么要用到深度信息呢？原因很简单：引入深度信息后，其提供的额外结构信息能够有效辅助复杂和困难场景下的分割。比如，与室外场景相比，由于语义类别繁杂、遮挡严重、目标外观差异较大等原因，室内场景的分割任务要更难实现。此时，在结合深度信息的情况下，能够有效降低分割的难度虽然道理容易，但是如何实现RGB-D分割还是有些问题需要解决的：

如何有效地表述和融合共存的深度和光度（RGB）数据

如何在特征学习过程中有效获取全局场景上下文

LSTM-CF [13]的网络结构详细描述了如何在RGB-D下实现语义分割。

基于ReNet [14]，LSTM-CF使用长时记忆和短时记忆以及DeepLab来实现分段。 LSTM-DF主要包括四个部分：用于垂直深度上下文提取的层，用于垂直光度上下文提取的层，用于将光度和深度上下文集成到2D全局上下文中的内存融合层以及像素级场景分割Floor。输入深度信息后，LSTM-CF使用HHA描述将深度信息转换为视差，表面法线和高度的三个通道。随后，使用ReNet提取不同方向的上下文信息，并在两个方向上双向传播。同时，对于RGB通道信息，网络使用上图所示的卷积结构来提取特征，并使用插值将所有级别的特征恢复为相同的分辨率，并将它们级联。之后，ReNet还用于获取上下文信息。结果表明，分割的主要精度仍来自于RGB信息，但是随着深度信息和上下文信息的引入，网络的精度可以得到一定程度的提高。

而【15】提出了RGB-D信息就是标准的RGB图像信息中引入了深度信息，而深度信息可为RGB图像提供对应的几何关系。现有的大多数工作只是简单地假设深度测量是准确的，且与RGB像素能够良好地对齐，由此将该问题建模为交叉模式特征融合以获得更好的特征表示从而实现更准确的分割，但是，通常传感器无法获得令人满意的精准的深度结果，实际的深度数据通常比较嘈杂，这可能会随着网络的深入而降低分割的准确性。提出了一个统一而有效的跨模态引导的编码器，该编码器不仅可以有效地重新校准RGB特征响应，而且还可以通过多个阶段来提取准确的深度信息，并将两者交替合并在一起重新校准的表示形式。提出的体系结构的关键是新颖的“分离与聚合gate”操作，该操作在交叉模态聚合之前共同过滤和重新校准两种表示形式。同时，一方面引入了双向多步传播策略，以帮助在两种形态之间传播和融合信息，另一方面，在长期传播过程中保持它们的特异性。此外，提出的编码器可以轻松地注入到先前的编码器-解码器结构中，以提高其在RGB-D语义分割任务上的性能。在室内和室外具有挑战性的数据集上，本文的模型均始终优于最新技术。

同时在RGBD的语义分割领域上论文【16】提出了一种基于三维点云的三维图神经网络(3DGNN)，它在点云的基础上构造了一个k近邻图。图模型中的每个节点对应一组点，并与一个隐藏的表示向量相关联，该向量初始化时使用一元CNN从2D图像中提的外观特征。每个节点都依赖于递归函数，根据当前状态和邻居传入的消息动态更新其隐藏表示。该传播模型展开一定数量的时间步长，最后的每个节点表示用于预测每个像素的语义类。通过深度信息将2D像素转化为3D点云，同时每个点与一个一元特征向量关联起来（比如2D CNN网络的分割输出）然后通过3D点构建图模型，边通过最近邻的方式获得。使用图像特征向量对点特征表达进行初始化。然后依据点的历史状态和邻居传递的信息（包括外观和三维几何信息）实现特征的动态更新。将节点的最终状态作为每个节点的分类输出。根据任务的不同，最终的预测可以在节点级，也可以在图级。即可以实现语义分割也可以实现图像分类。例如，可以将隐藏的表示(或其聚合)提供给另一个神经网络来执行节点(或图)分类。

与RGB语义分割相比，RGBD语义分割可以通过考虑深度信息来实现更好的性能。然而，由于RGB和深度（D）图像的特征分布在不同场景中显着变化，因此当代分割器有效地利用RGBD信息仍然存在问题。在论文【17】中，我们提出了一个注意补充网络（ACNet），它选择性地从RGB和深度分支中收集特征。主要贡献在于注意补充模块（ACM）和具有三个并行分支的架构。更确切地说，ACM是一个基于通道注意的模块，可从RGB和深度分支中提取加权特征。该架构提供原始RGB和深度分支的推断，并同时启用融合分支。基于上述结构，ACNet能够利用来自不同渠道的更高质量的功能.

而【18】提出了一种自底向上的反馈式弱监督分割算法框架。该框架采用了随机森林进行初标注,然后通过改进的SLIC算法对图像预分割得到超像素,在此基础上提取超像素特征,使用条件随机场对初标注结果进行全局优化。最后,通过设计的反馈式机制解决弱监督带来的问题,从而构成整体算法模型。(2)探究并优化了用于随机森林的区域纹元特征组合模式。在使用随机森林初标注的同时,我们探究了多种不同的区域纹元特征组合模式,并融入了改进的深度归一化处理。最终,得出了最优的随机森林区域纹元特征的计算方案。(3)提出了多种新颖的用于条件随机场的超像素特征。在使用条件随机场过程中,我们针对超像素不规则的特点提出了三类特征,分别是外观特征、几何特征和纹理梯度特征。同时,还将深度信息融入到超像素分割当中。(4)提出了一种基于图像评价的反馈式分割机制。本文使用综合了边缘评价和区域评价的图像筛选器选出好的分割图像反馈回训练样本中,形成递归结构。通过实验,我们证明了提出的整体算法框架在弱监督语义分割上可以取得很好的效果,甚至优于部分全监督算法。

前人的研究已经充分证明了在RGB图像的语义分割上引入场景的深度信息对于语义分割的精度是有一定提升的，而提升的效果除了在于如何引入深度信息之外，更重要的事在于如何获取精确的场景深度信息，而传统的深度提取方法都存在着精确性和鲁棒性的缺陷，因此在本文中，我们引入了在深度估计领域上前沿的算法，也就是光场深度估计，引入光场信息的深度估计在精确度跟鲁棒性上相较于传统方法都有着十分明显的提升。本章节将具体介绍光场深度提取的相关概念与内容。

传统的相机由主镜头和传感器平面组成。由于传统相机通过主镜头记录了空间中不同方向的光的聚集，因此传统相机无法区分光方向信息，而只能以该视角保存场景图像。全光摄像机的概念由Adelson和Wang [19]在1992年提出。通过在主透镜和传感器平面之间插入微透镜阵列，全光摄像机可以将场景中不同方向的光记录在传感器上的不同像素中。因此，全光相机可以通过一次拍摄来捕获场景中光线的空间和角度信息，并获得不同视角的场景图像信息。与传统图像不同，光场相机的特殊成像方法使光场图像包含各种不同的表示形式，因此本节将介绍光场图像的不同表示形式和数字重新聚焦的原理，并在此基础上进行分析。每种图像形式的成像一致性性能。对于子孔径图像（即透视图像），大多数研究都是基于场景点的成像一致性，并且使用立体匹配算法来估计深度。传统的多视点立体匹配算法已经成熟发展。文献中有相关的摘要和介绍[20，21]。其中，常用的匹配度计算方法包括绝对差之和（Ab-of-Absolute-Differences，SAD），Sum-of-Squared-Differences（Sum-of-Squared-Differences，SSD），Sum-of-Gradient -差异（GRAD）和正交交叉相关（NCC）等。由于不同的匹配方法各有优缺点，因此许多研究[22,23]选择使用单个或多个匹配项来计算不同类型的光场图像根据特征的不同而匹配。特别是Jeon等。 [24]设计了一种基于学习的匹配项算法，并使用随机森林算法从31种不同的匹配项中选择了最可靠的匹配成本组合方法和计算方法，以用于不同场景下的光场图像。进一步获得准确的深度估计结果。另外，根据光场图像的结构特点，提出了一些专门为光场图像设计的匹配算法。 Heber等。 [25]引入了基于稳健主成分分析（RPCA）的匹配项。该方法首先根据不同的深度标签使透视图像失真，并提出使用正确的深度标签失真形成的透视图像等级较低；然后通过最小化矩阵的秩，该方法等效于同时计算所有视角之间的匹配度。另外，该方法提出的优化问题等同于在立体重建的同时对视点图像进行去噪，从而可以进一步获得超分辨率的视点图像。 Johannsen等。 [26]提出了一种基于稀疏编码的深度估计框架。通过移动中心透视图图像对应点的位置，模拟不同深度标签下的光场图像，形成字典元素，并使用字典比较真实的光场图像。

像稀疏表示。可以使用与稀疏编码对应的系数来计算深度标签，并且可以对系数分布进行统计分析。

用于判断场景中是否存在由透明度或反射引起的多个深度，并进一步通过高斯混合

该模型获得不同表面的深度信息。威利姆等。 [27]提出了角熵矩阵（Angular Entropy）

公制）。由于光场图像提供了足够的视角图像匹配点以形成角度采样图像，因此该方法建议

对于未被遮挡的区域，角点采样图像的颜色直方图具有明显的峰值，因此您可以计算角点采样

样本图像的熵用于测量一致性。极坐标图像的斜率估计算法在传统的多视点图像中并未得到广泛应用，主要是因为统一的多视点图像的基线更宽，并且匹配点的位置在相邻的视点图像中。大的跨度使极性图像中直线的斜率太大而难以检测。光场图像受到主透镜的拍摄范围的限制，并且提取的透视图像之间的基线很窄，并且可以通过使用不同方法测量极性图像的线性斜率来获得深度信息。

Wanner等人针对光场极性图像的斜率估计。 [28]提出了一种基于结构张量的测量直线方向的方法，并通过建立能量函数获得了全局优化的深度图。与传统的点对点匹配计算不同，基于结构张量的方法计算速度更快，精度不受标签的限制。但是，由于结构张量要求角空间中极性图像的采样密集，因此对于太大的视差点（相邻视图之间的位置距离大于1），该方法无法获得相应的坡度信息，从而限制了申请方法。随着近年来卷积神经网络（CNN）的发展，Heber等人（2016） [29]建议

使用极坐标图像块进行训练并预测极坐标图像中直线的斜率，最后通过测量来估计斜率

为克服无纹理区域的预测误差，构造了高阶正则项的凸优化以克服2017年的可靠性

等等。 [30]提出了一种新的U形神经网络架构，该架构使用极性图像的堆叠而不是二维极性图像的堆叠。

深度估计。 U形网络体系结构由两个对称部分组成。编码部分组合三维图像信息（二维空间

信息和一维方向信息）抽象成高维特征图，解码部分将得到特征图进行计算

获得所需的输出。可以通过聚焦在不同位置的一系列图像来计算光场图像，即聚焦

焦点顺序。因此，传统的离焦深度（DfD）和离焦深度

（来自焦点的深度，DfF）可用于估计焦点序列不同图像中的焦点区域，然后获得该区域

域的位置信息。为了计算聚焦度，文献[31]使用拉普拉斯算子计算聚焦序列。

对于散焦度，文献[32]使用角度采样图像的平均值作为聚焦线索。此外，林等。 [33]

着眼于真实深度两侧的偏移具有相同颜色的原理，提出聚焦序列位于真实深度值处

根据双方的对称性原理，设计了基于聚焦序列对称性的匹配项，以实现深度鲁棒

估计。该文档还提出了一种基于仿真分析的方法，该方法比较了深度图计算出的仿真焦点

序列与真实聚焦序列之间的差异，设计一个新的聚焦序列匹配项，将方差替换为角度采样图像的平均值

减少噪音的影响。威利姆等。 [34]提出了一种约束自适应散焦匹配成本，它降低了原始焦点

该序列被划分为不同的图像块，选择未遮挡的部分以计算散焦度以消除遮挡的阴影。

环。文献比较并总结了不同焦点匹配项在不同光场图像和不同场景下的性能。

结