摘 要

近些年，计算机视觉技术发展迅速，视频跟踪技术逐渐成熟，成为智能人机交互、机器人视觉、视频监控、异常事件监测等任务中的基础和核心技术之一。基于视频图像的多目标跟踪技术是目前计算机视觉领域的一个具有挑战性和吸引力的基础研究方向。在现实生活中，多目标跟踪技术可以作为行人、车辆行为识别等视频分析技术的基础，在视频监控、智能交通、运动视频分析、智能机器人等领域有着广泛的应用空间，具有很高的研究意义和实用价值。

目前，国内外学者对多目标跟踪算法进行了大量深入的研究，取得了令人瞩目的研究成果。特别是随着近年来计算机领域对于深度学习的进一步探索，越来越多的研究者提出将深度学习方法用于视频多目标跟踪中。深度学习方法的应用在一定程度上弥补了传统框架的弊端，比如处理拥挤场景下的长期遮挡问题。本文的研究目标是利用短轨迹在多假设跟踪框架中探究深度学习方法来优化视频多目标跟踪中的关联和多假设剪枝问题，

本文先介绍了近年来的多目标跟踪主流算法和深度学习方法。最后对多目标跟踪技术进行了总结。

**关键词**：**多目标跟踪，短轨迹，深度学习，多假设跟踪**

**Abstract**

In recent years, with the rapid development of computer vision technology, video tracking technology has gradually matured, becoming one of the basic and core technologies of intelligent human-computer interaction, robot vision, video monitoring, abnormal event monitoring and other tasks. Multi-target tracking based on video images is a challenging and attractive basic research field in computer vision. In real life, multi-target tracking technology can be used as the basis of video analysis technology such as pedestrian and vehicle behavior recognition. It has wide application space in video monitoring, intelligent transportation, motion video analysis, intelligent robot and other fields, and has high research significance and practical value.

At present, scholars at home and abroad have done a lot of in-depth research on multi-target tracking algorithm, and made remarkable achievements. Especially with the further exploration of deep learning in the computer field in recent years, more and more researchers put forward the deep learning method for video multi-target tracking. To some extent, the application of deep learning method makes up for the disadvantages of the traditional framework, such as dealing with the long-term occlusion problem in crowded scenes. The research objective of this paper is to explore the deep learning method in the framework of multi-hypothesis tracking to optimize the correlation and multi-hypothesis pruning in video multi-object tracking.

This paper first introduces the main algorithms of multi-target tracking and deep learning in recent years. Finally, the paper summarizes the multi-target tracking technology.

**Keywords**: **multi-target tracking, short trajectory, deep learning, multiple hypothesis tracking**

语义分割一直是计算机视觉中十分重要的领域，从传统语义分割的发展到随着深度学习的流行，语义分割任务与方法也得到了大量的拓展和进步。在本章节中我们首先阐述了语义分割的背景和定义，然后再从论文出发概述近些年来多种语义分割的解决方案包括这些算法提出的背景、设计的思路以及解决的问题等，同时我们仍会介绍目前在语义分割领域中使用的主流数据集。本章节在深度学习的背景下由浅层模型到深度模型简要介绍了语义分割各种技术，包括这些技术的背景、方法以及解决的问题。如今，语义分割是计算机视觉的关键问题之一，语义分割产生的应用于静态2D图像、视频甚至3D数据、体数据的分割结果被广泛应用于多种计算机视觉的上层应用中。在宏观意义上来说，语义分割是为场景理解铺平了道路的一种高层任务。而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影像中推理出相关的知识或语义（即由具体到抽象的过程）。这些应用包括自动驾驶，人机交互，计算摄影学，图像搜索引擎，增强现实等。应用各种传统的计算机视觉和机器学习技术，这些问题已经得到了解决。虽然这些方法很流行，但深度学习革命让相关领域发生了翻天覆地的变化，不仅大幅度提升了语义分割的准确性同时也加快了生成场景分割结果的速度，因此，包括语义分割在内的许多计算机视觉问题都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在准确率甚至效率上都远远超过了传统方法。然而，相比于固有的计算机视觉及机器学习分支，深度学习的发展还远不成熟。也因此，语义分割领域的发展仍具有巨大的空间。该领域的飞速发展使得对初学者的启蒙教育比较困难，而且，由于大量的工作相继被提出，要跟上发展的步伐也非常耗时。于是，追随语义分割相关工作、合理地解释它们的论点、过滤掉低水平的工作以及验证相关实验结果等是非常困难的。

语义分割是对图像的像素级分类，语义图像分割的目标是用对应的所表示的类来标记图像的每个像素。因为在该任务中我们是对图像中的每个像素分类进行预测，所以语义分割任务通常也会被称为密集预测。

在全卷积网络出现以前，早期研究常常采用图像块方法进行语义分割，即将图像切成块输入至深度模型，然后对图像块中的像素进行分类。这些分类网络一般会含有全连接层，而全连接层的输入需要固定大小的图像块，所以会将图像进行切块。一个比较初级的方式构造神经网络模型就是单纯堆叠数个卷积层（利用相同的padding以保证维度不变）再输出一个最终的分割图。这种模型通过一系列特征映射的变换直接学习得到一个从输入图像到输出分割结果的映射。然而，因为整个过程中图像大小是没有变化的，所以整个网络都在全分辨率下参与计算，而这个过程所带来的计算量是非常巨大的。为了降低网络学习的计算量，需要提出新的网络分割结构，这些结构里面一种比较流行的图像分割模型是Encoder-Decoder(编码解码体)结构。Encoder部分通过下采样降低输入的空间分辨率，从而生成一个低分辨率的特征映射（计算高效且能够有效区分不同类别）；Decoder则对这些特征描述进行上采样，将其恢复成全分辨率的分割图，这种结构的提出既可以保持着语义分割的准确性同时也大幅降低了深度网络对场景特征的提取。2014年，全卷积网络【1】 (Fully Convolutional Networks,FCN)的提出，推广了过往分类网络的CNN结构，该论文提出以卷积层替换全连接层，从而实现像素级的密集预测，FCN结构可以接受任意尺寸的输入图像，采用反卷积层对最后一个卷积层的特征图进行上采样，使其恢复到输入图像相同的尺寸，从而对每一个像素都产生一个预测，实现逐像素分类。FCN成为了语义分割网络研究的基本架构，众多研究在此基础上展开。Encoder-Decoder结构也被首先引入FCN架构。Encoder使用池化层逐渐减少空间维度从而获取场景特征图，而Decoder使用上采样的方法恢复空间维度和细节信息，最终获取与输入大小一致的特征图，通过高低层网络之间的跃接，在保持着特征提取效率同时保留了足够的像素级细节，除了FCN，经典语义分割网络SegNet与UNet都是此类结构的典型应用。SegNet【2】的提出推广了之前Encoder-Decoder结构,其在编码网络连续的下采样导致了在输出的特征图上的每一个像素都重叠着着大量的输入图像中的空间信息。在解码网络中复用了最大池化层，这种应用改善了边界划分，减少了端到端训练的参数量。相较于之前的编码解码网络，SegNet在速度和效率上都有着明显的改进。随后，Ronneberger等人通过扩张解码器模块的容量的方式改进了上述的全卷积结构（fully convolutional architecture）。具体来讲，这种U-Net【3】结构“包括捕获上下文的收缩路径和实现精确定位的对称扩展路径”。这种结构直至目前也非常流行，已经被拓展至了多种分割问题上。在UNet的基础上，Drozdzal【4】等人在堆叠上使用残差模块替换了基础卷积模块。这种残差模块在模块内部含有skip conncetions，同时保留了与U-Net相同的在编码器与译码器对应特征图之间的skip connections。他们称这种方式能够使得网络更快收敛，也能够适用于更深的网络结构。同时基于这些工作，Jegou【5】等人提出了在网络结构中使用dense blocks，也是遵从于U-Net结构，称dense blocks的属性使得它们能够更好地适应语义分割任务，因为其天然携带skip connections和多尺度监督。这些dense blocks之所以高效是因为它们携带了从前序层中得到的low-level特征，也包含了从后续层中得到的higher level特征，从而实现更高效的特征复用。之后someone【6】使用扩张卷积（atrous/dilated convolutions）提出了ztrous spatial pyramid pooling (ASPP)应用全连接CRF。Dilated卷积在不增加参数个数的基础上实现了感受野的增加。多尺度处理既可以通过同时传递原始图像的不同尺度进入平行CNN分支（图像金字塔）实现，也可以通过（或同时使用）多个不同采样率的平行atrous convolutional layers实现（ASPP）。结构预测（Structured prediction）通过全连接CRF实现。CRF作为一个独立的后处理步骤单独训练。然而应用dilated/atrous卷积设计的方法的局限在于，扩张卷积的计算量巨大，且需要占用大量内存用于处理大量高分辨率的特征图。这一点输入了高分辨率下的预测计算。比如DeepLab’s预测就是原始输入的1/8尺寸。因此，RefineNet【7】论文中建议使用编码器-解码器结构。编码部分是ResNet-101模块，解码器包括RefineNet模块，将编码器下的高分辨率特征与前序RefineNet模块的低分辨率特征级联/融合。每个RefineNet模块都包括通过对低分辨率特征上采样融合多分辨率特征的部件和基于5x5步长为1的pool层进行内容捕获的部件。上述的两个部件每个都应用了基于恒等映射思想（identity map mindset）的残差连接设计。然而编码解码网络的结构不能支持其提取到输入图像中足够多的特征信息，在后来的发展中，语义分割领域出现了更多更复杂的网络来弥补这种不足，如PSPNet【8】、PSPNet的网络结构中存在着一个复杂的金字塔池化层用来提取特征图中不同尺度的信息，相较于UNet，PSPNet具有着更为复杂的网络结构，在多类别分类任务上一般也更具鲁棒性。全局场景类别的意义在于其提供了分割类别的分布线索。金字塔pooling模型通过应用大尺寸核的pooling层以捕获这些信息。如这篇论文，扩张卷积被应用于修改Resnet，同时增加了金字塔pooling模型。这个模型将ResNet的特征图经过上采样后整合。除主分支上的损失外，辅助损失被应用于ResNet的第4个阶段。这个概念在其他地方也被称作intermediate supervision。

之后【9】提出了含有非常大核的编码器-解码器结构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实现分割任务，论文用大尺寸核的卷积运算替代全连接层。

另一个应用大尺寸核的愿意是，尽管类似于ResNet这样更深的网络结构能够拥有比较大的感受野，但是研究显示，网络其实从一个小很多的区域（有效感受野，valid receptive field）收集信息。大尺寸核计算成本高且参数众多，因此一个kxk的卷积可以用1xk+kx1和kx1和1xk卷积。这个模型在论文中被称为Global Convolutional Network (GCN)。回到结构本身，ResNet（不含有任何扩张卷积）组成了编码层，GCNs和反卷积组成了解码层。算法还使用了一个简单的残差模块Boundary Refinement (BR)。

除了在固定二维场景中进行语义分割，【10】提出spatial path 和 contex path实现实时性语义分割。BiSeNet区别于 U-shape 和 Dilation 结构，尝试一种新的方法同时保持 Spatial Context 和 Spatial Detail 。所以，论文中设计了Spatial Path和Context Path两部分。顾名思义，Spatial Path使用较多的 Channel、较浅的网络来保留丰富的空间信息生成高分辨率特征；Context Path使用较少的 Channel、较深的网络快速 downsample来获取充足的 Context。基于这两路网络的输出，文中还设计了一个Feature Fusion Module（FFM）来融合两种特征， TDNet【11】这篇文章提出了一个基于时序分布网络的视频语义分割算法TDNet（Temporally Distributed Network），它的设计思想来自于这么一个观察：较深的网络输出的特征是可以由一系列的浅层网络输出的特征进行组合得到。而在视频分割任务中视频是具有时序属性的，而且视频分割也是有时序属性的，因而就可以在一定的时序范围内使用浅层的网络进行特征抽取，之后在经过组合可以达到深层网络输出特征的效果。这样的思路迁移也是相当简单的，那么怎么来实现文章中说的将多个浅层特征进行组合呢？文章对此的解决办法是使用一个新的APM（attention propagation module）来实现，减少分割目标移动对分割性能的影响。此外，还引入了知识蒸馏的概念，从使其可以在浅层网络输出特征与组合特征的两个层面进行知识迁移，从而提升学生网络的性能。

【12】论文提出了RefineNet 的轻量化版本Light-Weight RefineNet ，针对实时分割任务，将速度从20FPS提升到了55FPS。总体网络结构，分为下采样的encoder部分和上采样的decoder部分。网络主要包含4个模块，RCU，CRP，FUSION，CLF。为了轻量化该网络，分别使用RCU-LW，CRP-LW。

FUSION-LW替换了原始网络的RCU，CRP，FUSION。通过后续的实验作者又发现RCU对于网络的精度提升效果微弱，因此将RCU模块也去掉了。为什么去掉RCU模块，网络精度影响很小，因为(1)虽然RCU模块中的3\*3卷积使得网络具有更大的感受野，但是通过shortcut结构，底层特征和高层特征也可以共享。(2)CRP模块也可以获得上下文的信息。

语义分割在深度学习时代下取得了飞速的进步，然而从上面的回顾也可以看出，语义分割仍然有很多问题需要克服，目前还远称不上已经解决，更准确的分割边界，小物体的分割，实时性语义分割等问题仍然是一个挑战，要因此还需要学术界和工业界的持续努力。

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】的网络结构详细介绍了如何实现RGB-D下的语义分割。

基于ReNet【14】，LSTM-CF利用长短时间记忆和DeepLab实现分割。LSTM-DF主要包括四个部分：用于竖直深度上下文提取的层，用于竖直光度上下文提取的层，用于整合光度和深度上下文成2D全局上下文的记忆融合层，和像素级场景分割层。输入深度信息后，LSTM-CF利用HHA描述，将深度信息转换成视差、表面法线和高这三个通道的信息。随后，利用ReNet提取不同方向上的上下文信息，并在两个方向进行双向传播。与此同时，对于RGB通道信息，网络利用如上图所示的卷积结构提取特征，利用插值将各级特征恢复到相同分辨率下，并级联。之后，同样利用ReNet获取上下文信息。该论文的结果展示了分割的主要精度还是来源于RGB信息的，但是引入了深度信息和上下文信息后，网络的精度能够得到一定的提升。

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。

传统相机由主镜头和传感器平面等部分构成，由于记录的是空间中不同方向的光线通过主镜头的积分，传统相机并不能区分光线方向信息，只保存了该视角下的场景图像。全光相机的概念由 Adelson 和 Wang[19] 在 1992 年提出，通过在主镜头与传感器平面插入微透镜阵列，全光相机可以将场景中来自不同方向的光线记录在传感器上不同的像素中。因此，全光相机可以通过一次的拍摄捕捉到场景中光线的空间和角度信息，得到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式使得光场图像包含了多种不同的表现形式，因此本节将介绍光场图像的不同表示形式以及数字重聚焦原理，并在此基础上分析各个图像形式中成像一致性的表现。针对子孔径图像，也即视角图像，多数研究基于场景点的成像一致性，采用立体匹配算法来进行深度估计。传统多视角立体匹配算法已经得到了成熟的发展，在文献 [20，21] 中有相关的总结和介绍，其中常用的匹配程度计算方法包括有绝对差值和 (Sum-ofAbsolute-Differences，SAD) 、平方差值和 (Sum-of-Squared-Differences，SSD) 、梯度差值(Sum-of-Gradient-Differences，GRAD) 以及正交交叉相关 (Normalized-Cross-Correlation，NCC) 等。由于不同匹配方法具有各自不同的优势和缺点，许多研究[22,23] 根据不同匹配的特点，选择采用单一或多个匹配项结合的方式针对不同种类的光场图像进行计算。特别地，Jeon 等人[24] 设计了基于学习的匹配项算法，针对不同场景下的光场图像采用随机森林算法从 31 个不同的匹配项中选择最为可靠的匹配代价结合方式以及计算方法，进一步得到准确的深度估计结果。此外，基于光场图像的结构特性，一些专为光场图像设计的匹配算法也被陆续提出。Heber 等人[25] 引入了基于鲁棒主成分分析 (Robust Principal Component Analysis，RPCA)的匹配项。该方法首先根据不同的深度标签对视角图像进行扭曲，并提出采用正确深标签扭曲得到的视角图像所构成的矩阵具有较低的秩；然后通过最小化矩阵的秩，该方法相当于同时计算了全部视角之间的匹配度。此外，该方法所提出的最优化问题相当于在立体重建的同时对视角图像进行了去噪，可以进一步得到超分辨的视角图像。Johannsen 等人[26] 提出了基于稀疏编码的深度估计框架，通过移动中心视角图像对应点的位置仿真出不同深度标签下的光场图像构成字典元素，并采用该字典对真实的光场图

像进行稀疏表示。对应稀疏编码的系数可以用于计算深度标签，系数分布的统计分析可

以用来判断场景中是否存在由于透明或者反射造成的多层深度，并进一步通过高斯混合

模型求得不同表面的深度信息。Williem 等人[27] 则提出了角度熵矩阵 (Angular Entropy

Metric)。由于光场图像提供了足够的视角图像匹配点构成角采样图像，该方法提出对于

未被遮挡的区域，其角采样图像的颜色直方图具有明显的峰值，因此可以通过计算角采

样图像的熵来衡量一致性针对极面图像的斜率估计算法在传统多视角图像中的应用并不广泛，主要是因为统多视角图像的基线较宽，匹配点位置在相邻视角图像中的跨度较大，使得极面图像中的直线斜率过大而难以检测。而光场图像受到主镜头拍摄范围的限制，所提取的视角图像之间基线较窄，可以通过不同的方法衡量极面图像的直线斜率得到深度信息。

针对光场极面图像的斜率估计，Wanner 等人[28] 提出了基于结构张量的方法衡量直线方向，并通过建立能量函数得到全局优化的深度图。不同于传统的点与点之间匹配计算，基于结构张量的方法计算速度较快且精度不受标签的限制。然而由于结构张量要求极面图像在角度空间的抽样较为密集，因此对于过大的视差点 (相邻视角中的位置离大于 1)，该方法不能获得相应的斜率信息，也因此限制了该方法的应用。随着近年卷积神经网络 (Convolutional Neural Networks，CNN) 的发展，2016 年，Heber 等人[29] 提出

采用极面图像块进行训练并对极面图像中的直线斜率进行预测，最终通过衡量斜率估计

的可信度，构建高阶正则项的凸优化来克服在无纹理区域的预测错误。2017 年，Heber

等人[30] 提出了新的 U 型神经网络架构，采用极面图像的堆叠代替二维的极面图像进行

深度估计。U 型的网络架构包括对称的两部分，编码部分将三维的图像信息 (二维的空

间信息以及一维的方向信息) 抽象成为高维的特征图，解码部分将得到特征图进行计算

得到期望的输出。光场图像可以通过计算得到一系列聚焦在不同位置图像，也即聚

焦序列。因此传统的从散焦估计深度 (Depth from Defocus，DfD) 以及从聚焦估计深度

(Depth from Focus，DfF) 可以用于估计聚焦序列不同图像中的聚焦区域，进而得到该区

域的位置信息。针对聚焦程度的计算，文献 [31] 采用拉普拉斯算子计算聚焦序列的

散焦度，文献 [32] 则通过计算角采样图像的均值作为聚焦线索，此外，Lin 等人[33] 根据

聚焦在真实深度两侧方向上的偏移量具有一致颜色的原理，提出聚焦序列在真实深度值

两侧的对称性原理，并据此设计了一个基于聚焦序列对称性的匹配项来进行鲁棒的深度

估计。该文献还提出了一个基于仿真分析的方法，通过比较深度图计算得到的仿真聚焦

序列与真实聚焦序列的差异，设计新的聚焦序列匹配项，用角采样图像的均值代替方差

减少噪声的影响。Williem 等人在[34] 提出了约束自适应的离焦匹配代价，将原始的聚焦

序列分为不同的图像块儿选用未被遮挡的部分进行离焦程度的计算，以此排除遮挡的影

响。文献中对不同的聚焦匹配项在不同光场图像以及不同场景中的表现进行了比较和总

结

[1]Valera M, Velastin S A. Intelligent distributed surveillance systems: a review[J]. Vision, Image and Signal Processing, IEE Proceedings -, 2005, 152(2):192-204.

[2]Lipton A, Kanade T, Fujiyoshi H, et al. A system for video surveillance and monitoring[M]. Pittsburg: Carnegie Mellon University, the Robotics Institute, 2000.

[3]Haritaoglu I, Harwood D, Davis L S. W 4: Real-time surveillance of people and their activities[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2000, 22(8): 809-830.

[4] Computer Vision: Detection, recognition and reconstruction[M]. Springer, 2010.

[5] Takala V, Pietikainen M. Multi-object tracking using color, texture and motion[C]. Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007: 1-7.

[6] Ma L, Cheng J, Lu H. Multi-cue collaborative kernel tracking with cross ratio invariant constraint[C]. Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. IEEE, 2008: 1-4.

[7] Kumar P, Brooks M J, Van Den Hengel A. An adaptive Bayesian technique for tracking multiple objects[M]. Pattern Recognition and Machine Intelligence. Springer Berlin Heidelberg, 2007: 657-665.

[8] Li Z, Gong H, Zhu S C, et al. Dynamic feature cascade for multiple object tracking with trackability analysis[C]. Energy Minimization Methods in Computer Vision and Pattern Recognition. Springer Berlin Heidelberg, 2007: 350-361.

[9] Pernkopf F. Tracking of multiple targets using online learning for reference model adaptation[J]. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2008, 38(6): 1465-1475.

[10] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 580-587.

[11] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1440-1448.

[12] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[13] Wenju Li, Jianguo Yao, Tianzhen Dong,等. Object Tracking Based on Fragment Template and Multi-Feature Adaptive Fusion[C]// 2015 8th International Symposium on Computational Intelligence and Design (ISCID). IEEE, 2015.

[14] Zhang J , Presti L L , Sclaroff S . Online Multi-person Tracking by Tracker Hierarchy[C]// 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance. IEEE Computer Society, 2012.

[15] Guan Y , Chen X , Yang D , et al. Multi-person tracking-by-detection with local particle filtering and global occlusion handling[C]// 2014 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2014.

[16] Minggang,Gan,Yulong,Cheng,Yanan,Wang,Jie,Chen. Hierarchical particle filter tracking algorithm based on multi-feature fusion[J]. 系统工程与电子技术：英文版(1期):51-62.

[17] Ju J , Kim D , Ku B , et al. Online Multi-object Tracking Based on Hierarchical Association Framework[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2016.

[18] 金哲, 刘传才. 加速的TLD算法及其在多目标跟踪中的应用 [J]. 计算机系统应用, 2016(6):196-201,共6页.

[19] Shuhe Chang, Yanqing Zhong, Zhenhua Quan,等. A real-time object tracking and image stabilization system for photographing in vibration environment using OpenTLD algorithm[C]// 2016 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO). IEEE, 2016.

[20] Leal-Taixe L , Canton-Ferrer C , Schindler K , et al. Learning by Tracking: Siamese CNN for Robust Target Association[C]// Computer Vision & Pattern Recognition Workshops. IEEE, 2016.

[21] Gan, Weihao, Wang, Shuo, Lei, Xuejing,等. Online CNN-based multiple object tracking with enhanced model updates and identity association[J]. Signal Processing Image Communication:S0923596518304156.

[22] Yu F , Li W , Li Q , et al. POI: Multiple Object Tracking with High Performance Detection and Appearance Feature[M]// Computer Vision – ECCV 2016 Workshops. Springer International Publishing, 2016.

[23] Y. Qi et al., "Hedged Deep Tracking," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 4303-4311.

[24] Milan A , Rezatofighi S H , Dick A , et al. Online Multi-Target Tracking Using Recurrent Neural Networks[J]. 2016.

[25] S. Yang and H. Li, "Application of EKF and UKF in Target Tracking Problem," 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, 2016, pp. 116-120.

[26] Kokul T , Ramanan A , Pinidiyaarachchi U A J . Online Multi-Person Tracking-by-Detection Method Using ACF and Particle Filter[C]// 2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS). IEEE, 2015.

[27] A. Milan, L. Leal-Taixé, K. Schindler and I. Reid, "Joint tracking and segmentation of multiple targets," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 5397-5406.

[28] Yu S , Meng D , Zuo W , et al. The Solution Path Algorithm for Identity-Aware Multi-object Tracking[C]// Computer Vision & Pattern Recognition. IEEE, 2016.

[29] Milan A , Roth S , Schindler K . Continuous Energy Minimization for Multitarget Tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(1):58-72.

[30] Milan A , Schindler K , Roth S . Multi-Target Tracking by Discrete-Continuous Energy Minimization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10):2054-2068.

[31] Rezatofighi S H , Milan A , Zhang Z , et al. Joint Probabilistic Data Association Revisited[C]// International Conference on Computer Vision. IEEE Computer Society, 2015.

[32] Shi X , Song Y Q , Yang Z , et al. Multiple target tracking under occlusions using modified Joint Probabilistic Data Association[C]// ICC 2015 - 2015 IEEE International Conference on Communications. IEEE, 2015.

[33] Milan A , Leal-Taixe L , Reid I , et al. MOT16: A Benchmark for Multi-Object Tracking[J]. 2016.

[34] Smith K , Gatica-Perez D , Odobez J , et al. Evaluating Multi-Object Tracking[C]// Computer Vision and Pattern Recognition, 2005 IEEE Computer Society Conference on. IEEE, 2005.