(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116485867 A
(43)申请公布日 2023.07.25

(21)申请号 202310591809.5

(22)申请日 2023.05.24

(71)申请人 电子科技大学
地址 611731 四川省成都市高新区(西区)西源大道2006号

(72)发明人 陈浩然 李曙光 郑珂 刘斌

(74)专利代理机构 电子科技大学专利中心 51203
专利代理师 余涛

(51)Int.Cl.
G06T 7/55(2017.01)
G06V 10/80(2022.01)
G06V 10/764(2022.01)
G06V 10/26(2022.01)
G06V 20/70(2022.01)
G06V 10/40(2022.01)
G06V 10/774(2022.01)
G06V 10/82(2022.01)

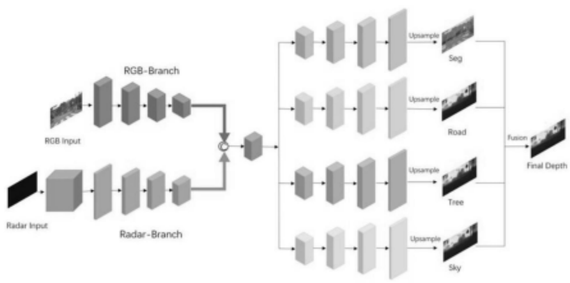
权利要求书2页 说明书6页 附图3页

(54)发明名称

一种面向自动驾驶的结构化场景深度估计方法

(57)摘要

本发明属于自动驾驶技术领域,具体为一种面向自动驾驶的结构化场景深度估计方法,通过双编码器对输入的RGB图像和毫米波雷达数据,采用稀疏前置映射模块提取稀疏的毫米波雷达特征并与图像特征进行融合,得到第一融合特征图。通过四个解码器中的其中1个解码器对第一融合特征图进行解码得到语义分割图,利用语义分割图将场景分类为三个特征类别;通过另外3个深度解码器分别对第一特征融合图解码,各得到1张初始预测图;3张初始预测图与三个特征类别一一对应融合,由此实现场景中的语义信息引入。结合本发明设计的基于L1loss的改进损失函数,该函数是在L1loss的基础上,对场景中不同类别目标赋予不同权重以提高网络性能。



1.一种面向自动驾驶的结构化场景深度估计方法，其特征在于，包括以下步骤：

步骤1、设计双编码-四解码网络

双编码-四解码网络由双编码网络和四解码网络组成；双编码网络以RGB图像和毫米波雷达数据作为输入，分别提取特征后融合，得到第一融合特征图；

四解码网络由四个解码器组成，四个解码器分别为第一解码器、第二解码器、第三解码器和第四解码器，第一解码器为分割解码器，第二解码器、第三解码器和第四解码器均为深度解码器：首先，将第一融合特征图分别输入四个解码器中，第一解码器根据第一融合特征图解码生成语义分割图，并根据语义分割图将场景划分三个特征类别，三个特征类别为道路及交通参与者特征、树木及建筑特征和天空特征三个特征类别；三个深度解码器分别对接收的第一融合特征图解码，各得到一张初始预测深度图；三张初始预测图与三个特征类别一一对应融合，获得不同场景类别下的深度图；然后再对不同场景类别下的深度图进行融合，得到预测深度图；

步骤2、设计双编码-四解码网络的损失函数

双编码-四解码网络的损失函数由四部分组成，分别是深度损失L_{depth}、平滑损失L_{smooth}，对稀疏前置模块生成的特征图的监督损失L_{map}、对语义分割结果的监督损失L_{seg}；其中，深度损失L_{depth}是以L1 loss为基础，将场景中的道路及交通参与者、树木及建筑、天空分别赋予不同权重后的改进函数；

深度损失L_{depth}如式(2)所示：

$$L_{depth}=\frac{1}{m}\left(\omega\sum_{(p,q)\in S_1}\left|d(p,q)-\hat{d}(p,q)\right|+(2-\omega)\sum_{(p,q)\in S_2}\left|d(p,q)-\hat{d}(p,q)\right|\right)\quad(2)$$

式(2)中，d和$\hat{d}$分别表示真实深度图和预测深度图，S_1表示d中属于道路及交通参与者的集合，S_1表示d中不属于道路及交通参与者的集合，m为有效深度的数量，ω为需要调节的超参数；当ω取值1.4时，自动驾驶场景中各类别特征点之间平衡达到最优；

平滑损失L_{smooth}如式(3)所示：

$$L_{smooth}=e^{-|\partial_x(I)|}\left|\partial_x(\hat{d})\right|+e^{-|\partial_y(I)|}\left|\partial_y(\hat{d})\right|\quad(3)$$

式(3)中，∂_x、∂_y分别表示沿x和y方向的梯度，I表示输入图像；

完整的双编码-四解码网络的损失函数如式(4)所示：

$$L_{total}=\lambda_1(L_{depth}+\lambda_2 L_{smooth}+\lambda_3 L_{map})+L_{seg}\quad(4)$$

式(4)中，λ_1，λ_2，λ_3均为加权因子，根据经验设定；

步骤3、以深度标签与分割标签为真值对网络进行监督，使用步骤2得到的损失函数进行反馈来训练双编码-四解码网络；

步骤4、将待估RGB图像和毫米波雷达数据输入训练好的双编码-四解码网络，对场景进行深度估计，得到最终的预测深度图。

2.根据权利要求1所述的一种面向自动驾驶的结构化场景深度估计方法，其特征在于：在构建并训练双编码-四解码网络时，均采用了nuScenes数据集。

3.根据权利要求1所述的一种面向自动驾驶的结构化场景深度估计方法，其特征在于：所述双编码网络包括图像编码器和深度编码器；其中所述图像编码器为预先训练过并去除

了全连接层的ResNet-34网络；所述深度编码器，包括稀疏前置映射模块和残差模块，通过稀疏前置映射模块提取毫米波雷达数据的初步特征，再采用残差模块进一步提取特征。

4.根据权利要求1所述的一种面向自动驾驶的结构化场景深度估计方法，其特征在于：所述深度解码器由4个依次连接的上采样模块组成，根据输入的第一融合特征图，首先生成一个分辨率为输入图像一半的16通道的特征映射，然后通过3×3卷积将生成的特征映射到单通道，最后使用双线性上采样到原分辨率后，直接作为初始预测图输出。

5.根据权利要求4所述的一种面向自动驾驶的结构化场景深度估计方法，其特征在于：所述分割解码器与深度解码器结构类似，其区别在于通过3×3卷积将生成的特征映射到不同分割类别的十九个通道，再采用softmax函数对其进行分类，得到三个特征类别输出。

一种面向自动驾驶的结构化场景深度估计方法

技术领域

[0001] 本发明涉及自动驾驶技术领域，具体为一种面向自动驾驶的结构化场景深度估计方法。

背景技术

[0002] 单目深度估计是计算机视觉领域中长期存在的一个不适定问题，它利用单张RGB图像估计场景中每个点到相机的距离，在机器人、自动驾驶、三维重建等多个领域中都有着广泛应用。

[0003] 传统的单目深度估计方法主要利用手工设计的特征，代表方法有运动恢复结构(SFM)和基于传统机器学习方法。运动恢复结构(SFM)是将摄像机运动作为线索进行深度估计，基于传统机器学习方法，通过使用马尔科夫随机场(MRF)或条件随机场(CRF)在图像与深度之间建立模型，学习输入特征与输出深度之间的映射关系，以获得深度估计信息。

[0004] 近年来，深度神经网络快速发展，已经在图像分类、图像检测、图像分割等图像处理任务中表现出了极为优秀的性能，因此研究者们将其引入到了单目深度估计中。2014年，Eigen等人首次使用深度卷积神经网络进行单目深度估计，它以RGB图像作为输入，经由两阶段网络分别粗略预测图像全局信息和细调图像局部信息。自从深度学习被应用到单目深度估计领域后，相关方法不断改进，如搭建多尺度网络改进性能，利用编码解码结构进行深度估计，或者按照深度分层，将深度估计从回归任务转化为分类任务。上述方法的训练均依赖于场景的真实深度标签，由于逐像素标注成本高昂，因此无监督学习方法也受到广泛的关注。其通常使用成对的立体图片或图片序列进行训练，通过图像重建的损失监督网络的训练，避免了标注过程中大量人力资源的投入。

[0005] 深度补全任务引入深度传感器，如激光雷达和毫米波雷达，将从深度传感器获得的粗糙深度图恢复成稠密的深度图。尽管纯视觉的深度估计方法已经可以取得较为满意的结果，利用传感器获取的额外深度信息与RGB图像信息相融合依然大幅度提高了深度估计的精度。深度补全任务关键点在于输入深度图十分稀疏且包含较大噪音，以及如何将图像与深度两个维度的信息充分融合以获得更好的结果。目前的深度补全方法利用多分支网络，使用编码器分别从稀疏深度图及其对应的RGB图像中提取特征，然后在不同层级上将特征融合，经解码器得到稠密深度图。随着深度补全技术的推进，表面法线、亲和矩阵等也被研究者们引入到网络模型之中，它们都促进了深度补全的发展。

[0006] 在自动驾驶场景中，深度估计任务发挥着重要的作用。结构化场景下的深度估计具有相对标准的场景特点，但是过去的方法并未考虑到利用场景信息对深度估计预测结果进行提升，也并未充分利用场景中的语义信息，因此，有必要对现有的结构化场景深度估计方法进行改进研究，以提高深度估计的精度。

发明内容

[0007] 本发明的目的在于:针对上述现有深度估计方法存在的不足，提出一种面向自动

驾驶的结构化场景深度估计方法。该方法以RGB图像和稀疏深度图作为输入，构建基于场景中的语义信息的双编码-四解码网络结构，以实现深度估计的精度提升。在构建双编码-四解码网络结构过程中，设计基于L1 loss改进的损失函数，对场景中不同类别目标，赋予不同权重以提高网络性能。

[0008] 为实现上述目的，本发明采用如下技术方案：

[0009] 一种面向自动驾驶的结构化场景深度估计方法，包括以下步骤：

[0010] 步骤1、设计双编码-四解码网络

[0011] 双编码-四解码网络由双编码网络和四解码网络组成；双编码网络以RGB图像和毫米波雷达数据作为输入，分别提取特征后融合，得到第一融合特征图；

[0012] 四解码网络由四个解码器组成，四个解码器分别为第一解码器、第二解码器、第三解码器和第四解码器，第一解码器为分割解码器，第二解码器、第三解码器和第四解码器均为深度解码器:首先，将第一融合特征图分别输入四个解码器中，第一解码器根据第一融合特征图解码生成语义分割图，并根据语义分割图将场景划分三个特征类别，三个特征类别为道路及交通参与者特征、树木及建筑特征和天空特征三个特征类别；三个深度解码器分别对接收的第一融合特征图解码，各得到一张初始预测深度图；三张初始预测图与三个特征类别一一对应融合，获得不同场景类别下的深度图；然后再对不同场景类别下的深度图进行融合，得到预测深度图；

[0013] 步骤2、设计双编码-四解码网络的损失函数

[0014] 双编码-四解码网络的损失函数由四部分组成，分别是深度损失L_{depth}、平滑损失L_{smooth}，对稀疏前置模块生成的特征图的监督损失L_{map}、对语义分割结果的监督损失L_{seg}；其中，深度损失L_{depth}是以L1 loss为基础，将场景中的道路及交通参与者、树木及建筑、天空分别赋予不同权重后的改进函数；

[0015] 深度损失L_{depth}如式(2)所示：

[0016] $$L_{depth}=\frac{1}{m}\left(\omega\sum_{(p,q)\in S_1}\left|d(p,q)-\hat{d}(p,q)\right|+(2-\omega)\sum_{(p,q)\in S_2}\left|d(p,q)-\hat{d}(p,q)\right|\right) \quad (2)$$

[0017] 式(2)中，d和$\hat{d}$分别表示真实深度图和预测深度图。S_1表示d中属于道路及交通参与者的集合，S_1表示d中不属于道路及交通参与者的集合，m为有效深度的数量，ω为需要调节的超参数；当ω取值1.4时，自动驾驶场景中各类别特征点之间平衡达到最优；

[0018] 平滑损失L_{smooth}如式(3)所示：

[0019] $$L_{smooth}=e^{-|\partial_x(I)|}\left|\partial_x(\hat{d})\right|+e^{-|\partial_y(I)|}\left|\partial_y(\hat{d})\right| \quad (3)$$

[0020] 式(3)中，∂_x、∂_y分别表示沿x和y方向的梯度，I表示输入图像。

[0021] 完整的双编码-四解码网络的损失函数如式(4)所示：

[0022] $L_{total}=\lambda_1(L_{depth}+\lambda_2 L_{smooth}+\lambda_3 L_{map})+L_{seg}$ (4)

[0023] 式(4)中，λ_1，λ_2，λ_3均为加权因子，根据经验设定；

[0024] 步骤3、以深度标签与分割标签为真值对网络进行监督，使用步骤2得到的损失函数进行反馈来训练双编码-四解码网络；

[0025] 步骤4、将待估RGB图像和毫米波雷达数据输入训练好的双编码-四解码网络，对场

景进行深度估计，得到最终的预测深度图。

[0026]　进一步的，所述构建并训练双编码-四解码网络时，均采用了nuScenes数据集。

[0027]　进一步的，所述双编码网络包括图像编码器和深度编码器；其中所述图像编码器为预先训练过并去除了全连接层的ResNet-34网络；所述深度编码器，包括稀疏前置映射模块和残差模块，通过稀疏前置映射模块提取毫米波雷达数据的初步特征，再采用残差模块进一步提取特征。

[0028]　进一步的，所述深度解码器由4个依次连接的上采样模块组成，根据输入的第一融合特征图，首先生成一个分辨率为输入图像一半的16通道的特征映射，然后通过3×3卷积将生成的特征映射到单通道，最后使用双线性上采样到原分辨率后，直接作为初始预测图输出。

[0029]　更进一步的，所述分割解码器与深度解码器结构类似，其区别在于通过3×3卷积将生成的特征映射到不同分割类别的十九个通道，再采用softmax函数对其进行分类，得到三个特征类别输出。

[0030]　本发明提供的一种面向自动驾驶的结构化场景深度估计方法，是以RGB图像和稀疏深度图作为输入，构建基于场景中的语义信息的双编码-四解码网络结构。该网络结构通过双编码器对输入的RGB图像和毫米波雷达数据，采用稀疏前置映射模块提取稀疏的毫米波雷达特征并与图像特征进行融合，得到第一融合特征图。通过四个解码器对第一融合特征图解码；解码过程中，利用其中1个解码器对第一融合特征图进行解码得到语义分割图，利用语义分割图将场景分类为三个特征类别；通过另外3个深度解码器来分别预测场景中三类目标的深度图，即三个解码器分别对第一特征融合图解码，各得到1张初始预测图；3张初始预测图与三个特征类别一一对应融合，由此实现场景中的语义信息引入。结合本发明设计的基于L1 loss的改进损失函数，该函数是在L1 loss的基础上，对场景中不同类别目标赋予不同权重以提高网络性能。

[0031]　与现有技术相比，本发明其深度估计的精度更高。

附图说明

[0032]　图1为实施例双编码-四解码网络架构示意图；

[0033]　图2为实施例的稀疏前置映射模块示意图；

[0034]　图3为实施例不同场景类别下的深度图融合过程；

[0035]　图4为本实施例双编码-四解码网络训练和推导示意图；

[0036]　图5为实施例得到的深度估计结果展示图。

具体实施方式

[0037]　下面结合附图和实施例对本发明作详细说明。

[0038]　本实施例提供的一种面向自动驾驶的结构化场景深度估计方法，包括以下步骤：

[0039]　步骤1、设计双编码-四解码网络

[0040]　如图1所示，双编码-四解码网络由双编码网络和四解码网络组成。

[0041]　所述双编码网络包括图像编码器和深度编码器。其中所述图像编码器为在ImageNet上预先训练过，并去除了全连接层的ResNet-34网络。包括4个依次连接的卷积模

块,4个卷积模块按连接顺序依次生成原图尺寸1/4,1/8,1/16,1/32的特征图,4个卷积模块的通道按连接顺序数依次为64,128,256,512。

[0042] 所述深度编码器,包括稀疏前置映射模块和残差模块,通过稀疏前置映射模块提取毫米波雷达数据的初步特征,再采用残差模块进一步提取特征。如图2所示,稀疏前置映射模块通过5个堆叠的稀疏不变卷积来获得更稠密的特征图,并在其输出处双线性上采样到原分辨率后,对此处的输出施加监督。其中稀疏不变卷积采用逐渐减少的卷积核依次为7,5,3,3,1,前4个卷积的输出通道数为16,最后一个卷积的输出通道数为1,第1个卷积的步幅为2,其余卷积步幅都为1,用于得到更稠密的输出以便施加监督。最后,将第4个卷积的输出作为残差模块的输入,进一步采用残差模块提取更高级的特征。本实施例中,稀疏前置映射模块采用的计算公式为:

[0043] $$f_{u,v}(x,o)=\frac{\sum_{i,j=-k}^{k} o_{u+i,v+j}x_{u+i,v+j}w_{i,j}}{\sum_{i,j=-k}^{k} o_{u+i,v+j}+\varepsilon}+b \qquad (1)$$

[0044] 式(1)中,x为输入;o代表对应于输入x的二值1或0,1表示有观测值)或者0表示没有观测值);W表示;为权重参数;b表示偏置;u、v为像素点坐标;ε为防止除数为0的一个极小的正数;

[0045] 残差模块采用层数更少的4个卷积模块,沿输出方向4个卷积模块分别得到的特征图,其尺寸依次为原图尺寸1/4,1/8,1/16,1/32的特征图,通道数依次分别为16,32,64,128。

[0046] 四解码网络由四个解码器组成,四个解码器分别为第一解码器、第二解码器、第三解码器和第四解码器,第一解码器为分割解码器,第二解码器、第三解码器和第四解码器均为深度解码器。

[0047] 首先,将第一融合特征图分别输入四个解码器中,第一解码器用于生成语义分割图,其包含4个依次连接的上采样模块,输入的第一特征融合图经4个上采样模块后,分别得到原图尺寸1/16,1/8,1/4,1/2的特征图,4个上采样模块的通道数分别为128,64,32,16。最后一个上采样模块的输出经过双线性上采样至19个通道后,再由softmax函数分类得到最终的分割结果,即得到道路及交通参与者特征、树木及建筑特征和天空特征三个特征类别输出。三个深度解码器结构与分割解码器类似,同样包含4个依次连接的上采样模块。只是深度解码器的最后一个上采样模块的输出经过双线性采样到原分辨率后,直接作为初始预测图输出。如图3所示,三个深度解码器各自生成的三张初始预测图与三个特征类别一一对应融合,获得不同场景类别下的深度图;然后再对不同场景类别下的深度图进行融合,得到预测深度图。

[0048] 步骤2、设计双编码-四解码网络的损失函数,损失函数由四部分组成,分别是深度损失L_{depth}、平滑损失L_{smooth},对稀疏前置模块生成的特征图的监督损失L_{map}、对语义分割结果的监督损失L_{seg}。包括以下子步骤:

[0049] 2.1、改进L1 loss

[0050] 在自动驾驶场景中各像素点存在一定关系,对网络参数进行优化时,需要考虑到各类别点的平衡关系,设计合适的损失函数。基于此,本实施将场景中的道路及交通参与者、树木及建筑、天空分别赋予不同权重,以L1 loss为基础,设计深度损失函数如下所示:

[0051] $$L_{depth}=\frac{1}{m}\left(\omega\sum_{(p,q)\in S_1}\left|d(p,q)-\hat{d}(p,q)\right|+(2-\omega)\sum_{(p,q)\in S_2}\left|d(p,q)-\hat{d}(p,q)\right|\right) \quad (2)$$

[0052] 式(2)中，d和$\hat{d}$分别表示ground truth depth map和预测深度图。S_1表示d中属于道路及交通参与者的集合，S_1表示d中不属于道路及交通参与者的集合，m为有效深度的数量，ω为需要调节的超参数。

[0053] 通过大量实验表明，合适的参数ω，能够使场景中各类别像素点达到平衡，在训练时使优化效果进一步提升。对ω取不同值，从0开始，以0.2为步长，得到其对各类别的误差如

[0054] 表1所示：

[0055]

ω值	道路上信息		树木建筑		天空	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
0.0	2.351	4.978	4.889	8.262	6.931	8.749
0.2	1.053	3.062	4.515	7.914	6.717	8.597
0.4	1.021	3.023	4.559	7.974	6.691	8.609
0.6	0.984	2.990	4.655	8.056	6.899	8.719
0.8	0.968	2.962	4.524	7.959	6.331	8.149
1.0	0.932	2.922	4.480	7.895	6.956	9.074
1.2	0.931	2.902	4.637	8.057	7.170	9.243
1.4	**0.905**	**2.853**	**4.457**	**7.845**	**6.575**	**8.449**
1.6	0.905	2.856	4.746	8.228	7.158	9.032
1.8	1.015	2.943	5.396	8.843	9.019	10.924
2.0	0.953	2.921	9.950	15.089	18.233	20.668

[0056] 从表中不难看出，当ω为1.4时，自动驾驶场景中各类别特征像素点之间平衡达到最优；获得的预测效果最好。

[0057] 2.2、对平滑损失定义

[0058] 由于深度不连续通常发生在交界处，因此使用图像梯度进行加权，平滑损失$L_{sm}oo_{th}$定义为：

[0059] $$L_{smooth}=e^{-|\partial_x(I)|}\left|\partial_x(\hat{d})\right|+e^{-|\partial_y(I)|}\left|\partial_y(\hat{d})\right| \quad (3)$$

[0060] 其中∂_x，∂_y分别表示沿x和y方向的梯度。I表示输入图像。

[0061] 2.3、引入监督损失，监督损失包含两部分：一是对稀疏前置映射模块生成的深度图map监督损失，记为L_{map}；二是对引入分割解码器语义分割结果的监督损失，记为L_{seg}

[0062] 因此，双编码-四解码网络的损失函数为：

[0063] $L_{total}=\lambda_1(L_{depth}+\lambda_2L_{smooth}+\lambda_3L_{map})+L_{seg}$ (4)

[0064] 其中λ_1，λ_2，λ_3是根据经验设定的超参数。

[0065] 步骤3、以深度标签与分割标签为真值对网络进行监督，使用步骤2得到的损失函数进行反馈来训练双编码-四解码网络。如图4所示，本实施训练时仅以图像和毫米波雷达

作为输入生成深度图。

[0066] 步骤4、将待估RGB图像和毫米波雷达数据输入训练好的双编码-四解码网络,对场景进行深度估计。结果如图5所示。其中预测结果中的颜色从蓝到红渐变,表示深度值增大,估计的最大深度值为120米。

[0067] 本实施例对双编码-四解码网络训练和测试均采用nuScenes数据集,nuScenes数据集中不仅包含相机和激光雷达数据,也记录了毫米波雷达数据,是为数不多包含毫米波雷达数据的大型数据集。该数据集每个场景时长20秒,其中有40个关键帧,每帧图像的分辨率为1600×900。并且nuScenes中包含各种情况下的驾驶场景,如雨天、夜晚等,这也增加了在该数据集上进行深度估计的难度。本发明使用了850个场景,并将它们分为810个场景用于训练,40个场景用于评估。(训练集共计32564张图片,测试集共计1585张图片)。最终估计得到的深度图在所有像素点,即144万个像素点上都估计出最终深度,相比于初始毫米波雷达每帧探测的有效点数仅为40-50点,提升约两万倍的稠密度。

[0068] 本实施例使用Pytorch部署网络,并且在NVIDIA GeForce GTX TITAN X上训练。批大小设置为4,使用Adam优化器其学习率为0.0005,并且每5个轮次学习率下降一半,参数设置为$\lambda_1=0.5$,$\lambda_2=0.001$,$\lambda_3=0.3$。所得结果在所有像素位置处计算误差,结果如表2所示,可以看出本发明的各项指标都优于现有的先进方案,证明了对不同类别分别进行深度估计和采用本发明提出的损失函数有效提升了网络的性能。令d和$\hat{d}$分别表示预测的深度图和标签,n表示每幅图像存在激光雷达深度值的观测点个数,Y表示测量范围。所采用的评价指标如下所示:

[0069] 均方根误差(RMSE):$\sqrt{\frac{1}{n}\sum_{p}^{n}(d_p-\widehat{d_p})^2}$;

[0070] 平均绝对误差(MAE):$\frac{1}{n}\sum_{p}^{n}(d_p-\widehat{d_p})$。

[0071] 表2深度估计结果

[0072]

方法	误差(↓)	
	MAE	RMSE
本发明	**2.424**	**5.516**

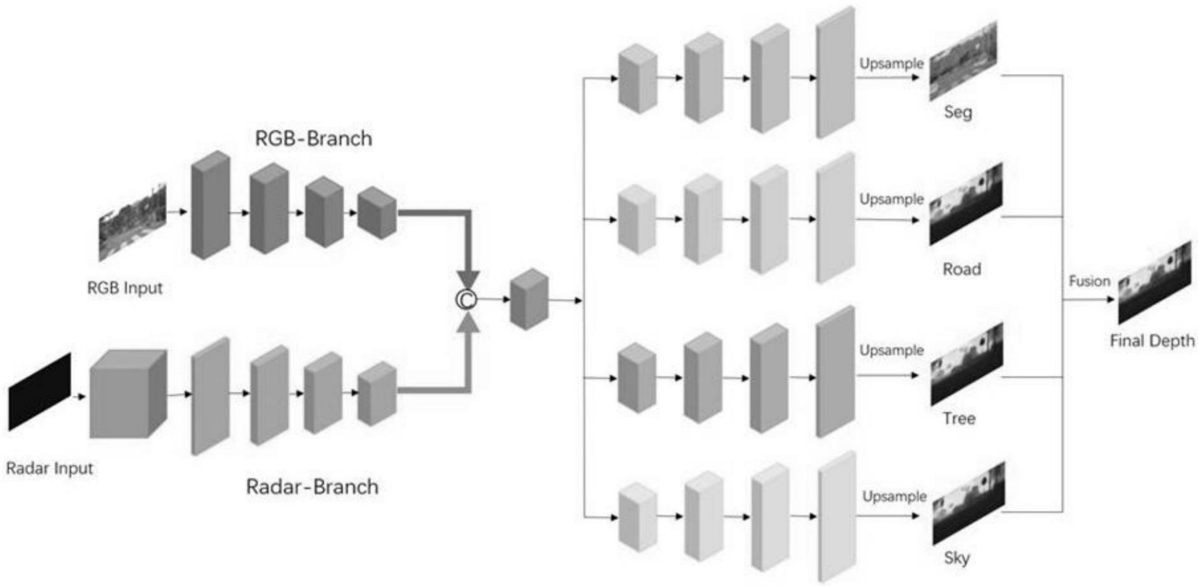

RGB-Branch
RGB Input
Radar Input
Radar-Branch
Upsample
Seg
Road
Tree
Sky
Fusion
Final Depth

图1

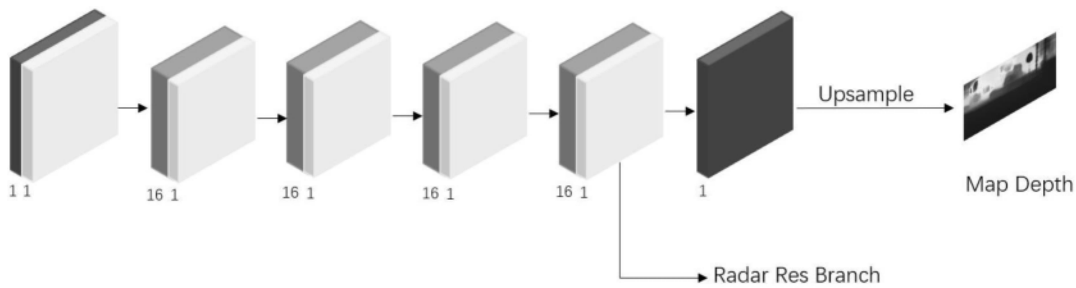

1 1
16 1
16 1
16 1
16 1
1
Upsample
Map Depth
Radar Res Branch

图2

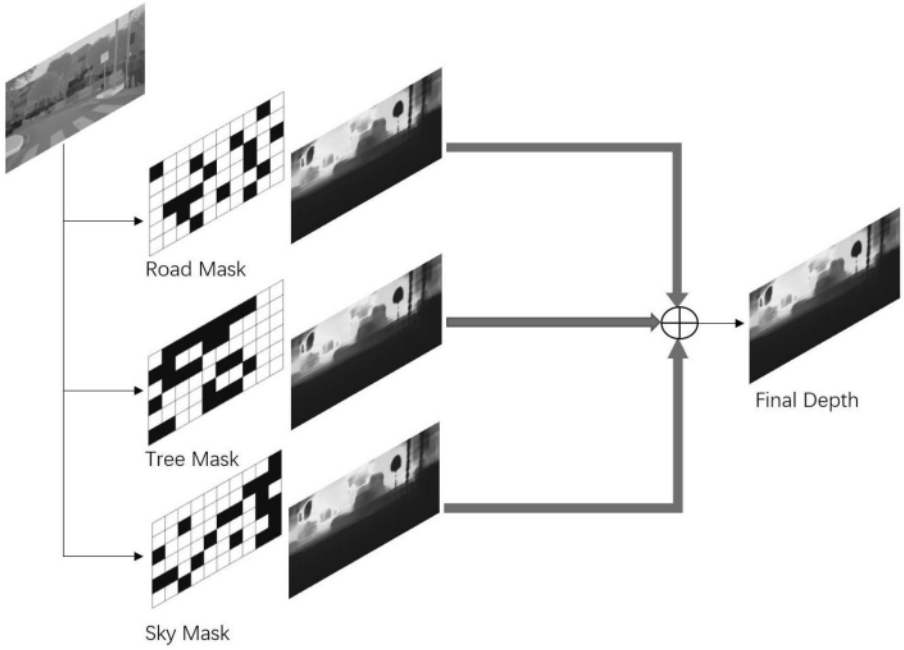

Road Mask
Tree Mask
Sky Mask
Final Depth

图3

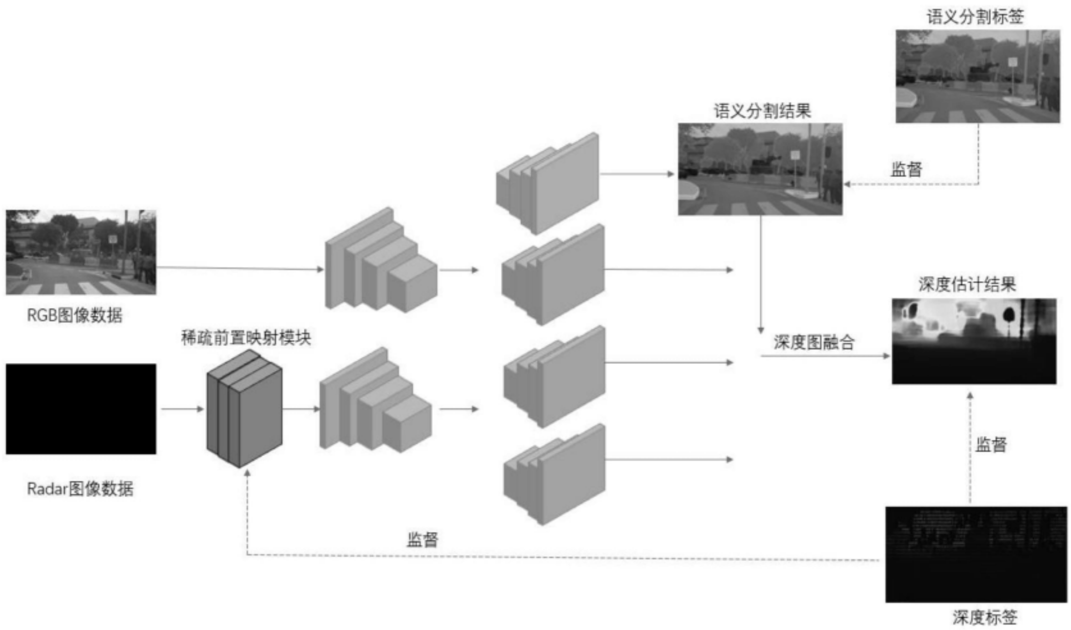

语义分割标签
语义分割结果
监督
RGB图像数据
稀疏前置映射模块
深度估计结果
深度图融合
Radar图像数据
监督
监督
深度标签

图4

图5