(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114663509 B
(45)授权公告日 2022.09.27

(21)申请号 202210290488.0
(22)申请日 2022.03.23
(65)同一申请的已公布的文献号
申请公布号 CN 114663509 A
(43)申请公布日 2022.06.24
(73)专利权人 北京科技大学
地址 100083 北京市海淀区学院路30号
专利权人 北京科技大学顺德研究生院
(72)发明人 曾慧 修海鑫 刘红敏 樊彬 张利欣
(74)专利代理机构 北京市广友专利事务所有限责任公司 11237
专利代理师 张仲波
(51)Int.Cl.
G06T 7/73(2017.01)
G06T 3/40(2006.01)
G06T 7/55(2017.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)
审查员 王慧琴

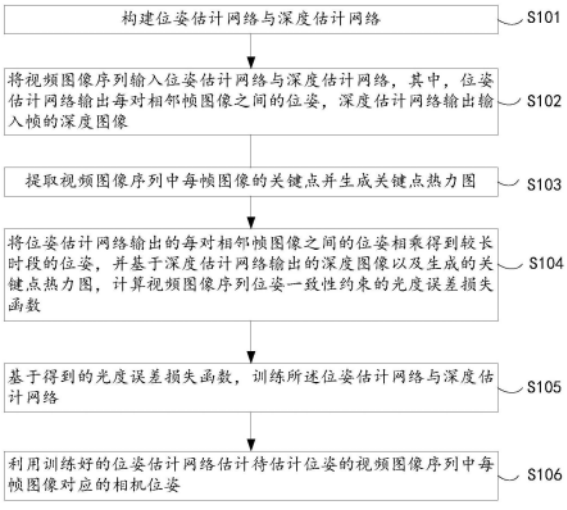
权利要求书3页 说明书11页 附图2页

(54)发明名称

一种关键点热力图引导的自监督单目视觉里程计方法

(57)摘要

本发明提供一种关键点热力图引导的自监督单目视觉里程计方法，属于计算机视觉领域。所述方法包括：构建位姿估计网络与深度估计网络；将视频图像序列输入位姿估计网络与深度估计网络；提取视频图像序列中每帧图像的关键点并生成关键点热力图；将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿，并基于深度估计网络输出的深度图像以及生成的关键点热力图，计算视频图像序列位姿一致性约束的光度误差损失函数；基于得到的光度误差损失函数，训练所述位姿估计网络与深度估计网络；利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿。采用本发明，能够提高相机位姿估计的精度。



CN 114663509 B

1.一种关键点热力图引导的自监督单目视觉里程计方法,其特征在于,包括:

构建位姿估计网络与深度估计网络;

将视频图像序列输入位姿估计网络与深度估计网络,其中,位姿估计网络输出每对相邻帧图像之间的位姿,深度估计网络输出输入帧的深度图像;

提取视频图像序列中每帧图像的关键点并生成关键点热力图;

将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿,并基于深度估计网络输出的深度图像以及生成的关键点热力图,计算视频图像序列位姿一致性约束的光度误差损失函数;

基于得到的光度误差损失函数,训练所述位姿估计网络与深度估计网络;

利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿;

其中,所述将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿,并基于深度估计网络输出的深度图像以及生成的关键点热力图,计算视频图像序列位姿一致性约束的光度误差损失函数包括:

将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿,并基于深度估计网络输出的深度图像以及生成的关键点热力图,计算图像之间的关键点热力图加权的光度误差;

根据计算得到的光度误差,计算视频图像序列位姿一致性约束的光度误差损失函数;

其中,所述将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿,并基于深度估计网络输出的深度图像以及生成的关键点热力图,计算图像之间的关键点热力图加权的光度误差包括:

对于长度为N的一段视频图像序列,其对应的时刻为$t_0,t_1,\ldots,t_{N-1}$,将位姿估计网络输出的每对相邻帧图像之间的位姿进行累积相乘,得到较长时段的位姿:

$$\begin{bmatrix} R_{t_j\to t_i} & t_{t_j\to t_i} \\ 0 & 1 \end{bmatrix}=\begin{bmatrix} R_{t_{i+1}\to t_i} & t_{t_{i+1}\to t_i} \\ 0 & 1 \end{bmatrix}\begin{bmatrix} R_{t_{i+2}\to t_{i+1}} & t_{t_{i+2}\to t_{i+1}} \\ 0 & 1 \end{bmatrix}\cdots\begin{bmatrix} R_{t_j\to t_{j-1}} & t_{t_j\to t_{j-1}} \\ 0 & 1 \end{bmatrix}$$

其中,$R_{t_j\to t_i}$为由时刻t_j到时刻t_i的旋转变换矩阵;$t_{t_j\to t_i}$为由时刻t_j到时刻t_i的平移变换向量;$R_{t_j\to t_i}$和$t_{t_j\to t_i}$构成图像I_{t_j}和I_{t_i}之间的位姿;N为输入位姿估计网络与深度估计网络的每个批次的视频图像序列的长度;

基于得到的较长时段的位姿、深度估计网络输出的图像的深度图像以及生成的关键点热力图,计算t_i和t_j时刻的图像I_{t_i}和I_{t_j}之间的关键点热力图加权的光度误差$L'_p[I_{t_j},I'_{t_i\to t_j}]$;

其中,所述光度误差损失函数L_p表示为:

$$L_p=\sum_{i=0}^{N-1}\sum_{j=i+1}^{N}L'_p[I_{t_j},I'_{t_i\to t_j}]$$

其中,所述基于得到的较长时段的位姿、深度估计网络输出的图像的深度图像以及生成的关键点热力图,计算t_i和t_j时刻的图像I_{t_i}和I_{t_j}之间的关键点热力图加权的光度误差$L'_p[I_{t_j},I'_{t_i\to t_j}]$包括:

设P_{t_j}为t_j时刻时图像I_{t_j}上的像素齐次坐标，则点P_{t_j}在t_i时刻图像I_{t_i}上对应的像素点的齐次坐标P_{t_i}表示为：

$$P_{t_i} = K(R_{t_j \to t_i} d_{t_j}[P_{t_j}] K^{-1} P_{t_j} + t_{t_j \to t_i})$$

其中，K为摄相机内参数；$R_{t_j \to t_i}$为由时刻t_j到时刻t_i的旋转变换矩阵；$t_{t_j \to t_i}$为由时刻t_j到时刻t_i的平移变换向量；d_{t_j}为图像I_{t_j}的深度图像；$R_{t_j \to t_i}$和$t_{t_j \to t_i}$构成图像I_{t_i}和I_{t_j}之间的位姿；

设$I'_{t_i \to t_j}$为利用t_i时刻的图像I_{t_i}重构得到的t_j时刻的重构图像，则$I'_{t_i \to t_j}$表示为：

$$I'_{t_i \to t_j}[P_{t_j}] = I_{t_i}[P_{t_i}]$$

其中，对于P_{t_i}坐标不为整数的情况，采用双线性插值的方法进行采样；

基于得到的重构图像$I'_{t_i \to t_j}$，确定t_j和t_i时刻的图像I_{t_j}和I_{t_i}之间的关键点热力图加权的光度误差$L'_p[I_{t_j}, I'_{t_i \to t_j}]$：

$$L'_p[I_{t_j}, I'_{t_i \to t_j}] = \frac{\alpha_0}{2}(1 - SSIM(I_{t_j}, I'_{t_i \to t_j}) * H) + \alpha_1 \left\| (I_{t_j} - I'_{t_i \to t_j}) * H \right\|_1 + \alpha_2 \left\| (I_{t_j} - I'_{t_i \to t_j}) * H \right\|_2$$

其中，$SSIM(I_{t_j}, I'_{t_i \to t_j})$表示源图像$I_{t_j}$与重构图像$I'_{t_i \to t_j}$的结构相似性，$||\cdot||_1$、$||\cdot||_2$分别为L1范数及L2范数，$\alpha_0$、$\alpha_1$、$\alpha_2$为超参数，*表示逐像素相乘，H表示关键点热力图。

2.根据权利要求1所述的关键点热力图引导的自监督单目视觉里程计方法，其特征在于，所述提取视频图像序列中每帧图像的关键点并生成关键点热力图包括：

对视频图像序列中图像I提取关键点，使用高斯核函数生成一幅仅关注关键点周围局部区域的关键点热力图，其中，图像I为视频图像序列中的任一图像；

生成的关键点热力图H[p]表示为：

$$H[p] = \sum_{f \in F, |p-f| < \delta} \frac{1}{\sqrt{2\pi \frac{\delta}{3}}} e^{-\frac{\|p-f\|_2}{2(\frac{\delta}{3})^2}}$$

其中，p为关键点热力图中的像素点坐标，f∈F为关键点的坐标，F表示特征点集，δ为关键点的影响半径。

3.根据权利要求1所述的关键点热力图引导的自监督单目视觉里程计方法，其特征在于，所述基于得到的光度误差损失函数，训练所述位姿估计网络与深度估计网络包括：

对于深度估计网络的输出，确定深度平滑损失函数L_s：

$$L_s = \left|\partial_x d_t^*\right| e^{-|\partial_x I_t|} + \left|\partial_y d_t^*\right| e^{-|\partial_y I_t|}$$

其中，$d_t^* = 1/d_t$为视差，即深度d_t的倒数，∂_x、∂_y分别表示x方向与y方向上的偏导数，I_t为t时刻的图像；

根据确定的深度平滑损失函数L_s以及所述光度误差损失函数L_p，得到最终的损失函数

L：

$L=L_p+\lambda L_s$

其中，λ为控制深度平滑损失函数比例的超参数；

利用最终的损失函数训练所述位姿估计网络与深度估计网络。

4.根据权利要求1所述的关键点热力图引导的自监督单目视觉里程计方法，其特征在于，所述利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿包括：

将待估计位姿的图像序列中每对相邻帧图像输入到训练好的位姿估计网络中，得到每对相邻帧图像之间的位姿；

将位姿估计网络输出的每对相邻帧图像之间的位姿累积相乘，得到每帧图像对应的相机位姿。

一种关键点热力图引导的自监督单目视觉里程计方法

技术领域

[0001] 本发明涉及计算机视觉领域，特别是指一种关键点热力图引导的自监督单目视觉里程计方法。

背景技术

[0002] 视觉里程计是指根据输入视频图像帧估计相机当前的位置与姿态的方法，可被广泛应用在机器人导航、自动驾驶、增强现实、可穿戴计算等领域。根据采用传感器的种类和数目不同，视觉里程计可分为单目视觉里程计、双目视觉里程计以及融合惯性信息的视觉里程计等。其中，单目视觉里程计具有着仅需要一个相机，对硬件要求较低、无需矫正等优点。

[0003] 传统的视觉里程计方法首先进行图像特征提取与匹配，然后根据几何关系估计相邻两帧之间的相对位姿。这种方法在实际应用中取得了不错的结果，是当前视觉里程计的主流方法，但其存在计算性能与鲁棒性难以平衡的问题。

[0004] 基于深度学习的单目视觉里程计可分为有监督的方法和自监督的方法。自监督的方法仅仅需要输入视频图像帧，不需要采集真实的位姿，没有对额外设备的依赖，适用性比有监督的方法更为广泛。

[0005] 现有的自监督方法在训练过程中使用了过多的冗余像素，使得深度神经网络在学习过程中没有重点，导致网络估计的位姿会产生累积误差。此外，这些方法仅考虑了相邻帧间的位姿一致性，没有考虑视频图像序列的位姿一致性。

发明内容

[0006] 本发明实施例提供了一种关键点热力图引导的自监督单目视觉里程计方法，能够提高相机位姿估计的精度。所述技术方案如下：

[0007] 本发明实施例提供了一种关键点热力图引导的自监督单目视觉里程计方法，包括：

[0008] 构建位姿估计网络与深度估计网络；

[0009] 将视频图像序列输入位姿估计网络与深度估计网络，其中，位姿估计网络输出每对相邻帧图像之间的位姿，深度估计网络输出输入帧的深度图像；

[0010] 提取视频图像序列中每帧图像的关键点并生成关键点热力图；

[0011] 将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿，并基于深度估计网络输出的深度图像以及生成的关键点热力图，计算视频图像序列位姿一致性约束的光度误差损失函数；

[0012] 基于得到的光度误差损失函数，训练所述位姿估计网络与深度估计网络；

[0013] 利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿。

[0014] 进一步地，所述提取视频图像序列中每帧图像的关键点并生成关键点热力图包

括：

[0015] 对视频图像序列中图像I提取关键点，使用高斯核函数生成一幅仅关注关键点周围局部区域的关键点热力图，其中，图像I为视频图像序列中的任一图像；

[0016] 生成的关键点热力图H[p]表示为：

[0017] $$H[p]=\sum_{f\in F,|p-f|<\delta}\frac{1}{\sqrt{2\pi\frac{\delta}{3}}}e^{-\frac{\|p-f\|_2}{2(\frac{\delta}{3})^2}}$$

[0018] 其中，p为关键点热力图中的像素点坐标，f∈F为关键点的坐标，F表示特征点集，δ为关键点的影响半径。

[0019] 进一步地，所述将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿，并基于深度估计网络输出的深度图像以及生成的关键点热力图，计算视频图像序列位姿一致性约束的光度误差损失函数包括：

[0020] 将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿，并基于深度估计网络输出的深度图像以及生成的关键点热力图，计算图像之间的关键点热力图加权的光度误差；

[0021] 根据计算得到的光度误差，计算视频图像序列位姿一致性约束的光度误差损失函数。

[0022] 进一步地，所述将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿，并基于深度估计网络输出的深度图像以及生成的关键点热力图，计算图像之间的关键点热力图加权的光度误差包括：

[0023] 对于长度为N的一段视频图像序列，其对应的时刻为$t_0,t_1,\ldots,t_{N-1}$，将位姿估计网络输出的每对相邻帧图像之间的位姿进行累积相乘，得到较长时段的位姿：

[0024] $$\begin{bmatrix}R_{t_j\to t_i} & t_{t_j\to t_i}\\ 0 & 1\end{bmatrix}=\begin{bmatrix}R_{t_{i+1}\to t_i} & t_{t_{i+1}\to t_i}\\ 0 & 1\end{bmatrix}\begin{bmatrix}R_{t_{i+2}\to t_{i+1}} & t_{t_{i+2}\to t_{i+1}}\\ 0 & 1\end{bmatrix}\cdots\begin{bmatrix}R_{t_j\to t_{j-1}} & t_{t_j\to t_{j-1}}\\ 0 & 1\end{bmatrix}$$

[0025] 其中，$R_{t_j\to t_i}$为由时刻t_j到时刻t_i的旋转变换矩阵；$t_{t_j\to t_i}$为由时刻t_j到时刻t_i的平移变换向量；$R_{t_j\to t_i}$和$t_{t_j\to t_i}$构成图像I_{t_j}和I_{t_i}之间的位姿；N为输入位姿估计网络与深度估计网络的每个批次的视频图像序列的长度；

[0026] 基于得到的较长时段的位姿、深度估计网络输出的图像的深度图像以及生成的关键点热力图，计算t_i和t_j时刻的图像I_{t_i}和I_{t_j}之间的关键点热力图加权的光度误差$L'_p[I_{t_j},I'_{t_i\to t_j}]$。

[0027] 进一步地，所述光度误差损失函数L_p表示为：

[0028] $$L_p=\sum_{i=0}^{N-1}\sum_{j=i+1}^{N}L'_p[I_{t_j},I'_{t_i\to t_j}]。$$

[0029] 进一步地，所述基于得到的较长时段的位姿、深度估计网络输出的图像的深度图像以及生成的关键点热力图，计算t_i和t_j时刻的图像I_{t_i}和I_{t_j}之间的关键点热力图加权的光

度误差$L'_p[I_{t_j}, I'_{t_i \to t_j}]$包括：

[0030] 设P_{t_j}为t_j时刻时图像I_{t_j}上的像素齐次坐标，则点P_{t_j}在t_i时刻图像I_{t_i}上对应的像素点的齐次坐标P_{t_i}表示为：

[0031] $$P_{t_i} = K(R_{t_j \to t_i} d_{t_j}[P_{t_j}] K^{-1} P_{t_j} + t_{t_j \to t_i})$$

[0032] 其中，K为摄相机内参数；$R_{t_j \to t_i}$为由时刻t_j到时刻t_i的旋转变换矩阵；$t_{t_j \to t_i}$为由时刻t_j到时刻t_i的平移变换向量；d_{t_j}为图像I_{t_j}的深度图像；$R_{t_j \to t_i}$和$t_{t_j \to t_i}$构成图像I_{t_i}和I_{t_j}之间的位姿；

[0033] 设$I'_{t_i \to t_j}$为利用t_i时刻的图像I_{t_i}重构得到的t_j时刻的重构图像，则$I'_{t_i \to t_j}$表示为：

[0034] $$I'_{t_i \to t_j}[P_{t_j}] = I_{t_i}[P_{t_i}]$$

[0035] 其中，对于P_{t_i}坐标不为整数的情况，采用双线性插值的方法进行采样；

[0036] 基于得到的重构图像$I'_{t_i \to t_j}$，确定t_j和t_i时刻的图像I_{t_j}和I_{t_i}之间的关键点热力图加权的光度误差$L'_p[I_{t_j}, I'_{t_i \to t_j}]$：

[0037] $$L'_p[I_{t_j}, I'_{t_i \to t_j}] = \frac{\alpha_0}{2}(1 - SSIM(I_{t_j}, I'_{t_i \to t_j}) * H) + \alpha_1 \left\| (I_{t_j} - I'_{t_i \to t_j}) * H \right\|_1 + \alpha_2 \left\| (I_{t_j} - I'_{t_i \to t_j}) * H \right\|_2$$

[0038] 其中，$SSIM(I_{t_j}, I'_{t_i \to t_j})$表示源图像$I_{t_j}$与重构图像$I'_{t_i \to t_j}$的结构相似性，$||\cdot||_1$、$||\cdot||_2$分别为L1范数及L2范数，$\alpha_0$、$\alpha_1$、$\alpha_2$为超参数，*表示逐像素相乘，H表示关键点热力图。

[0039] 进一步地，所述基于得到的光度误差损失函数，训练所述位姿估计网络与深度估计网络包括：

[0040] 对于深度估计网络的输出，确定深度平滑损失函数L_s：

[0041] $$L_s = \left|\partial_x d_t^*\right| e^{-|\partial_x I_t|} + \left|\partial_y d_t^*\right| e^{-|\partial_y I_t|}$$

[0042] 其中，$d_t^* = 1/d_t$为视差，即深度d_t的倒数，∂_x、∂_y分别表示x方向与y方向上的偏导数，I_t为t时刻的图像；

[0043] 根据确定的深度平滑损失函数L_s以及所述光度误差损失函数L_p，得到最终的损失函数L：

[0044] $L = L_p + \lambda L_s$

[0045] 其中，λ为控制深度平滑损失函数比例的超参数；

[0046] 利用最终的损失函数训练所述位姿估计网络与深度估计网络。

[0047] 进一步地，所述利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿包括：

[0048] 将待估计位姿的图像序列中每对相邻帧图像输入到训练好的位姿估计网络中，得到每对相邻帧图像之间的位姿；

[0049] 将位姿估计网络输出的每对相邻帧图像之间的位姿累积相乘，得到每帧图像对应的相机位姿。

[0050] 本发明实施例提供的技术方案带来的有益效果至少包括：

[0051] (1)针对视频图像中包含冗余像素使得深度神经网络缺乏学习重点的问题，本发明计算关键点热力图，进而计算出关键点热力图加权的光度误差。这样，可以为网络学习指出关注的重点，以减少图像中冗余像素点对网络学习的影响，从而解决现有技术所存在的在训练过程中使用了过多的冗余像素，使得深度神经网络在学习过程中没有重点的问题。

[0052] (2)针对视觉里程计对于较长时间的序列会存在累积误差的问题，本发明将连续视频图像帧之间的位姿相乘得到较长时段的位姿，并在此基础上计算图像序列位姿一致性约束的光度误差损失函数，进而训练位姿估计网络与深度估计网络，并利用训练的位姿估计网络估计图像序列中每帧图像对应的相机位姿。这样，可以在训练过程中在更长的输入序列上约束位姿估计网络的输出结果，以减小累积误差，提高相机位姿估计的精度，从而解决现有技术所存在的仅考虑了相邻帧间的位姿一致性，没有考虑视频图像序列的位姿一致性的问题。

附图说明

[0053] 为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0054] 图1为本发明实施例提供的关键点热力图引导的自监督单目视觉里程计方法的流程示意图；

[0055] 图2为本发明实施例提供的关键点热力图引导的自监督单目视觉里程计方法的结构示意图；

[0056] 图3为本发明实施例提供的SIFT关键点热力图生成示意图；

[0057] 图4为本发明实施例提供的用于训练和测试的KITTI数据集的样图；

[0058] 图5为本发明实施例提供的方法在KITTI里程计数据集中序列09、10上估计的轨迹图。

具体实施方式

[0059] 为使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明实施方式作进一步地详细描述。

[0060] 如图1和图2所示，本发明实施例提供了一种关键点热力图引导的自监督单目视觉里程计方法，包括：

[0061] S101，构建位姿估计网络(PoseNet)与深度估计网络(DepthNet)；

[0062] 本实施例中，为了控制内存占用，将位姿估计网络与深度估计网络的输入图像(指RGB图像)缩放为了416×128的大小。

[0063] 本实施例中，位姿估计网络包括：编码器和解码器，其中，可以选择ResNet50作为编码器，编码器输出2048通道的编码后输入位姿估计网络解码器。位姿估计网络的解码器输入为编码器输出的2048通道的编码，经过一层核为1的卷积层和ReLU激活函数调整通道数后，再依次经过两层核为3、激活函数同样为ReLU的卷积层，再经过核为1的卷积层，得到6

通道的张量，再经过全局平均池化层，得到6维的向量。本实施例中，位姿估计网络用来估计相邻两帧图像之间的位姿变换，输入为相邻两帧图像，输出为相应的6自由度位姿变换向量，即位姿变换(简称:位姿)，包括:3自由度旋转矩阵和3自由度平移向量。

[0064] 本实施例中，深度估计网络同样选择ResNet50结构作为编码器，以类似于DispNet解码器的多层反卷积结构作为解码器，并通过跳跃链接结构与编码器连接，输出层激活函数为Sigmoid。

[0065] 本实施例中，深度估计网络用来估计一帧图像的深度图像，输入为单帧图像，输出为相应的深度图像，具体为:归一化的视差d^*。要获得深度，需要对获得的视差取倒数$d=1/(ad^*+b)$，其中，a和b为限制输出取值范围的参数，使输出深度为0.1到100之间。表1、表2、表3分别给出了本实施例中所使用的神经网络结构，表1为位姿估计网络与深度估计网络公共的编码器结构。表2为位姿估计网络的解码器结构，表3为深度估计网络的解码器结构。

[0066] 表1 编码器结构

[0067]

层	输入通道数	卷积核	步幅	输出通道数
conv+bn+relu	6	7		64
maxpool	-	3	2	-
res block	64	-	-	256
res block	256	-	-	512
res block	512	-	-	1024
res block	1024	-	-	2048

[0068] 表2 位姿估计网络的解码器结构

[0069]

层	输入通道数	卷积核	输出通道数
conv+relu	2048	1	256
conv+relu	256	3	256
conv+relu	256	3	256
conv	256	1	6

[0070] 表3 深度估计网络的解码器结构

[0071]

层	输入通道数	卷积核	步幅	输出通道数	备注
upconv+relu	2048	3	2	1024	
conv+relu	1024+1024	3	-	1024	
upconv+relu	1024	3	2	512	
conv+relu	512+512	3	-	512	
upconv+relu	512	3	2	256	
conv+relu	256+256	3	-	256	
conv+sigmoid	256	1	-	1	output depth
upconv+relu	256	3	2	64	
conv+relu	64+64	3	-	64	
conv+sigmoid	64	1	-	1	output depth
upconv+relu	64	3	2	16	
conv+relu	16+16	3	-	16	
conv+ sigmoid	16	1	-	1	output depth

[0072] S102,将视频图像序列输入位姿估计网络与深度估计网络,其中,位姿估计网络输出每对相邻帧图像之间的位姿,深度估计网络输出输入帧的深度图像;

[0073] 在本实施例中,设每对相邻图像帧为当前时刻t的图像I_t与上一时刻t-1的图像I_{t-1}。将图像I_t和I_{t-1}输入S101中构建的位姿估计网络与深度估计网络中,得到相邻帧图像I_t和I_{t-1}之间的位姿以及图像I_t和I_{t-1}的深度图像。

[0074] S103,提取视频图像序列中每帧图像的关键点并生成关键点热力图;

[0075] 对于已有的自监督单目视觉里程计方法来说,深度估计网络和位姿估计网络在训练时损失函数的定义往往考虑了原始图像与重构图像中所有的像素点,在整个参数空间中搜索合适的网络参数。这种训练方法将不同的像素同等的对待,缺乏重点的搜索使得训练过程中使用到了大量特征信息较少,不宜匹配的像素。为了解决上述问题,本实施例中设计了关键点热力图引导的加权网络训练方法,具体的:

[0076] 首先,选择一种特征点提取算法对输入图像I提取特征点,得到特征点集F;图像I为视频图像序列中的任一图像;

[0077] 接着,可以使用SIFT关键点对输入图像I提取SIFT关键点。使用高斯核函数生成一幅仅关注关键点周围局部区域的SIFT关键点热力图,其中,生成的关键点热力图H[p]表示为:

[0078] $$H[p]=\sum_{f\in F,|p-f|<\delta}\frac{1}{\sqrt{2\pi\frac{\delta}{3}}}e^{-\frac{\|p-f\|_2}{2(\frac{\delta}{3})^2}}$$

[0079] 其中,p为关键点热力图中的像素点坐标,f∈F为关键点的坐标,F表示特征点集,δ为关键点的影响半径。

[0080] 如图3所示,图3(a)为原图像的示意图,图3(b)为提取的SIFT关键点的示意图,图3(c)为生成的SIFT关键点热力图的示意图。

[0081] 需要说明的是:

[0082] SIFT关键点热力图仅用于在训练阶段计算损失函数。在测试阶段,对于测试图像不需要计算其对应的SIFT关键点热力图。因此,SIFT关键点热力图虽然计算较为耗时,但是不会增加实际应用中位姿估计的计算负担。

[0083] S104,将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿,并基于深度估计网络输出的深度图像以及生成的关键点热力图,利用多帧图像计算视频图像序列位姿一致性约束的光度误差损失函数;具体可以包括以下步骤:

[0084] A1,将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿,并基于深度估计网络输出的深度图像以及生成的关键点热力图,计算图像之间的关键点热力图加权的光度误差;

[0085] 本实施例中,位姿估计网络的估计结果是相邻的两帧图像之间的位姿变换。当输入一段连续图像序列后,得到的是一系列相邻两帧图像之间的位姿变换。由于视觉里程计是一个长期的、连续的过程,仅使用相邻两帧之间的位姿变换计算损失函数,会使得网络仅关注两帧之间的变换,没有考虑较长时间内整体的位姿变换的一致性。为了使网络能够适应较长时间上的位姿变换,本实施例设计了基于视频图像序列位姿一致性约束的光度误差

损失函数。

[0086] 本实施例中，设输入S101中构建的位姿估计网络与深度估计网络的每个批次的视频图像序列的长度为N，则每个批次中的每对相邻图像帧皆为S102中所述相邻图像帧，在S102中输入到位姿估计网络与深度图及网络中，得到每个批次中每对相邻图像帧之间的位姿和每一帧的深度图像。

[0087] 本实施例中，对于长度为N的一段视频图像序列，其对应的时刻为$t_0, t_1, \ldots, t_{N-1}$，将位姿估计网络输出的每对相邻帧图像之间的位姿进行累积相乘，得到较长时段的位姿：

[0088] $$\begin{bmatrix} R_{t_j\to t_i} & t_{t_j\to t_i} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} R_{t_{i+1}\to t_i} & t_{t_{i+1}\to t_i} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R_{t_{i+2}\to t_{i+1}} & t_{t_{i+2}\to t_{i+1}} \\ 0 & 1 \end{bmatrix} \cdots \begin{bmatrix} R_{t_j\to t_{j-1}} & t_{t_j\to t_{j-1}} \\ 0 & 1 \end{bmatrix}$$

[0089] 其中，$R_{t_j\to t_i}$为由时刻t_j到时刻t_i的旋转变换矩阵；$t_{t_j\to t_i}$为由时刻t_j到时刻t_i的平移变换向量；$R_{t_j\to t_i}$和$t_{t_j\to t_i}$构成图像I_{t_j}和I_{t_i}之间的位姿；

[0090] 基于得到的较长时段的位姿、深度估计网络输出的图像的深度图像以及生成的关键点热力图，计算t_i和t_j时刻的图像I_{t_i}和I_{t_j}之间的关键点热力图加权的光度误差$L'_p[I_{t_j}, I'_{t_i\to t_j}]$，具体可以包括以下步骤：

[0091] 设P_{t_j}为t_j时刻时图像I_{t_j}上的像素齐次坐标，则点P_{t_j}在t_i时刻图像I_{t_i}上对应的像素点的齐次坐标P_{t_i}表示为：

[0092] $$P_{t_i} = K(R_{t_j\to t_i} d_{t_j}[P_{t_j}] K^{-1} P_{t_j} + t_{t_j\to t_i})$$

[0093] 其中，K为摄相机内参数；$R_{t_j\to t_i}$为由时刻t_j到时刻t_i的旋转变换矩阵；$t_{t_j\to t_i}$为由时刻t_j到时刻t_i的平移变换向量；d_{t_j}为图像I_{t_j}的深度图像；$R_{t_j\to t_i}$和$t_{t_j\to t_i}$构成图像I_{t_i}和I_{t_j}之间的位姿；

[0094] 设$I'_{t_i\to t_j}$为利用t_i时刻的图像I_{t_i}重构得到的t_j时刻的重构图像，则$I'_{t_i\to t_j}$表示为：

[0095] $$I'_{t_i\to t_j}[P_{t_j}] = I_{t_i}[P_{t_i}]$$

[0096] 其中，对于P_{t_i}坐标不为整数的情况，采用双线性插值的方法进行采样；

[0097] 基于得到的重构图像$I'_{t_i\to t_j}$，确定t_j和t_i时刻的图像I_{t_j}和I_{t_i}之间的关键点热力图加权的光度误差$L'_p[I_{t_j}, I'_{t_i\to t_j}]$：

[0098] $$L'_p[I_{t_j}, I'_{t_i\to t_j}] = \frac{\alpha_0}{2}(1 - SSIM(I_{t_j}, I'_{t_i\to t_j}) * H) + \alpha_1 \left\| (I_{t_j} - I'_{t_i\to t_j}) * H \right\|_1 + \alpha_2 \left\| (I_{t_j} - I'_{t_i\to t_j}) * H \right\|_2$$

[0099] 其中，$SSIM(I_{t_j}, I'_{t_i\to t_j})$表示源图像$I_{t_j}$与重构图像$I'_{t_i\to t_j}$的结构相似性，$||_1$、$||\cdot||_2$分别为L1范数及L2范数，$\alpha_0$、$\alpha_1$、$\alpha_2$为超参数，*表示逐像素相乘，H表示关键点热力图。

[0100] 本实施例中，以t-1时刻和t时刻的两帧图像为例，说明光度误差的计算方法：根据步骤S102可知，将t-1时刻和t时刻的两帧图像输入位姿估计网络可得到两帧之间的位姿变换。将t时刻的视频图像送入深度估计网络可得到其对应的深度图像。在获得t-1时刻和t时

刻两帧视频图像之间的位姿变换以及t时刻视频图像的深度图像之后，可利用它们对t-1时刻的视频图像进行重采样，得到t时刻的重构图像，并利用重构图像计算光度误差，以指导神经网络训练。

[0101] 本实施例中，在计算光度误差的过程中使用生成的关键点热力图对图像的不同区域采用不同的关注度，即对图像的不同区域采用不同的权重来计算光度误差。

[0102] A2，根据计算得到的光度误差，计算视频图像序列位姿一致性约束的光度误差损失函数。

[0103] 本实施例中，根据计算得到的t_i和t_j时刻的图像I_{t_i}和I_{t_j}之间的关键点热力图加权的光度误差$L'_p[I_{t_j}, I'_{t_i \to t_j}]$，计算视频图像序列位姿一致性约束的光度误差损失函数L_p：

[0104] $$L_p = \sum_{i=0}^{N-1} \sum_{j=i+1}^{N} L'_p[I_{t_j}, I'_{t_i \to t_j}];$$

[0105] 本实施例中，根据上述公式可知，需要对上述长度为N的视频图像序列的每2，3，...，N个子序列都进行了累积相乘，从而得到每个子序列的首、尾两帧的位姿，以便进一步利用各子序列的首、尾两帧的位姿计算视频图像序列位姿一致性约束的光度误差损失函数L_p。

[0106] 本实施例中，考虑到随着时间的推移，过长的时间跨度下，场景中的物体也会出现较大变化，以至于失去相关性，因此N的取值不宜过大。

[0107] S105，基于得到的光度误差损失函数，训练所述位姿估计网络与深度估计网络；

[0108] 考虑到一帧图像对应的深度图像中，原图像纹理平滑的区域，在深度图像中对应的区域同样是平滑的。因此，在本实施例中，对于深度估计网络的输出，按如下公式计算深度平滑损失函数：

[0109] $$L_s = \left|\partial_x d_t^*\right| e^{-\left|\partial_x I_t\right|} + \left|\partial_y d_t^*\right| e^{-\left|\partial_y I_t\right|}$$

[0110] 其中，$d_t^* = 1/d_t$为视差，即深度d_t的倒数，∂_x、∂_y分别表示x方向与y方向上的偏导数。

[0111] 在本实施例中，上述深度平滑损失函数对每个批次中的每帧图像都进行了计算；

[0112] 根据确定的深度平滑损失函数L_s以及所述光度误差损失函数L_p，则最终的损失函数L可表示为：

[0113] $L=L_p+\lambda L_s$

[0114] 其中，λ为控制深度平滑损失函数比例的超参数。

[0115] 利用最终的损失函数$L=L_p+\lambda L_s$，训练所述位姿估计网络与深度估计网络。

[0116] S106，利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿。

[0117] 本实施例中，将待估计位姿的图像序列中每对相邻帧图像输入到训练好的位姿估计网络中，得到每对相邻帧图像之间的位姿；将位姿估计网络输出的每对相邻帧图像之间的位姿累积相乘，得到每帧图像对应的相机位姿。

[0118] 本发明实施例针对现有的基于深度学习的单目视觉里程计方法包含了大量对冗余像素的计算、在进行位姿估计时只考虑相邻两帧图像之间的位姿变换导致误差不断累积

的问题，提供一种关键点热力图引导的自监督单目视觉里程计方法，该方法能够较为有效地根据输入的待估计位姿的图像序列估计每一帧图像对应的相机位姿，适用于用于自监督单目视觉里程计。

[0119] 本实施例所述的关键点热力图引导的自监督单目视觉里程计方法至少具有以下优点：

[0120] (1)针对视频图像中包含冗余像素使得深度神经网络缺乏学习重点的问题，本发明计算关键点热力图，进而计算出关键点热力图加权的光度误差。这样，可以为网络学习指出关注的重点，以减少图像中冗余像素点对网络学习的影响，从而解决现有技术所存在的在训练过程中使用了过多的冗余像素，使得深度神经网络在学习过程中没有重点的问题。

[0121] (2)针对视觉里程计对于较长时间的序列会存在累积误差的问题，本发明将连续视频图像帧之间的位姿相乘得到较长时段的位姿，并在此基础上计算图像序列位姿一致性约束的光度误差损失函数，进而训练位姿估计网络与深度估计网络，并利用训练的位姿估计网络估计图像序列中每帧图像对应的相机位姿。这样，可以在训练过程中在更长的输入序列上约束位姿估计网络的输出结果，以减小累积误差，提高相机位姿估计的精度，从而解决现有技术所存在的仅考虑了相邻帧间的位姿一致性，没有考虑视频图像序列的位姿一致性的问题。

[0122] 为了验证本发明实施例所述的关键点热力图引导的自监督单目视觉里程计方法的效果，本实施例使用KITTI里程计数据集中提供的评估指标测试其性能：

[0123] (1)相对位移均方误差(Rel.trans.)：一个序列中全部长度为100、200、……、800米的子序列的平均位移RMSE(Root Mean Square Error)，以％度量，即每100米偏差的米数，数值越小越好。

[0124] (2)相对旋转均方误差(Rel.rot.)：一个序列中全部长度为100、200、……、800米的子序列的平均旋转RMSE，以deg/m度量，数值越小越好。

[0125] 本实施例中，应用了KITTI里程计数据集中00-07这八个视频图像序列作为训练集与验证集训练位姿估计网络与深度估计网络，并用09-10这两个视频图像序列来测试所述的关键点热力图引导的自监督单目视觉里程计方法的性能。

[0126] 如图4所示，图4为KITTI里程计数据集中的样图。KITTI里程计数据集是车载相机等设备采集的城市中公路环境的双目图像，雷达点以及实际轨迹。

[0127] 在实施过程中，首先构建位姿估计网络与深度估计网络；将视频图像序列输入位姿估计网络与深度估计网络，其中，位姿估计网络输出每对相邻帧图像之间的位姿，深度估计网络输出输入帧的深度图像；提取视频图像序列中每帧图像的关键点并生成关键点热力图；将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿，并基于深度估计网络输出的深度图像以及生成的关键点热力图，计算视频图像序列位姿一致性约束的光度误差损失函数；基于得到的光度误差损失函数，训练所述位姿估计网络与深度估计网络；利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿。

[0128] 在本实施例中，光度误差损失函数的超参数$\alpha_0=0.85$，$\alpha_1=0.1$，$\alpha_2=0.05$，深度平滑损失函数的参数$\lambda=10^{-3}$。关键点热力图参数δ由多次实验确定为$\delta=16$。图像序列位姿一致性约束参数N的确定，考虑了服务器显存，并通过实验确定为N＝5。网络的训练过程中，初

始学习率为10^{-4}，并随着训练的进行逐渐减小，每经过一轮迭代，学习率变为上一轮的0.97倍，采用Adam优化器进行30次迭代，每轮迭代的批量大小为4。在训练时还对输入进行了增广，即对输入进行亮度、对比度、饱和度以及色相的随机变换，以增加网络对不同色调、亮度、饱和度等的情况的适应能力，增强网络的泛化能力。

[0129] 为了验证本发明实施例提供的关键点热力图引导的自监督单目视觉里程计方法的性能，本实施例中，选择了近几年基于深度学习的自监督的单目视觉里程计方法进行了对比，对比结果如表4所示。本实施例生成的轨迹如图5所示，图5(a)为本发明实施例提供的方法在KITTI里程计数据集中序列09上估计的轨迹图，图5(b)为本发明实施例提供的方法在KITTI里程计数据集中序列10上估计的轨迹图，其中，小方块表示起点，红色虚线轨迹为真实的轨迹，蓝色实现轨迹为本实施例中估计出的轨迹。

[0130] 表4 本实施例所述的方法与其他方法对比

[0131]

方法	09 Rel. trans. (%)	09 Rel. rot. (deg/m)	10 Rel. trans. (%)	10 Rel. rot. (deg/m)
SfMLearner	8.28	0.031	12.2	0.03
GeoNet	28.72	0.098	23.9	0.09
DepthVOFeat	11.93	0.039	12.45	0.035
Vid2depth	-	-	21.54	0.125
UnDeepVO	7.01	0.036	10.63	0.046
Wang et al.	9.88	0.034	12.24	0.052
Ranjan et al.	**6.92**	0.018	**7.97**	**0.031**
DeepMatchVO	9.91	0.038	12.18	0.059
PoseGraph	8.1	**0.028**	12.9	0.032
MonoDepth2	11.47	0.032	**7.73**	0.034
SC-SfMLearner	11.2	0.034	10.1	0.05
Jau et al.	8.64	0.66	11.72	0.95
本发明实施例	**5.79**	**0.0296**	8.72	**0.0291**

[0132] 由表4可以看出，相比于其他基于FlowNet等多层卷积网络的方法，如Wang et al.，本发明实施例提供的关键点热力图引导的自监督单目视觉里程计方法取得了更好的性能。相比于SC-SfMLearner、GeoNet等基于ResNet结构的方法，本发明所述的图像序列位姿一致性约束和关键点热力图引导的方法也使得性能有所提升。

[0133] 为了验证本发明实施例提供的关键点热力图引导的自监督单目视觉里程计方法各部分的意义，本实施例中还进行了消融实验。实验结果如表5所示。表5中“basic”为没有融入关键点热力图引导和图像序列位姿一致性约束的方法，“kphm r12”、“kphm r16”、“kphm r32”分别表示融入了关键点的影响半径δ为12、16、32的关键点热力图引导的方法，“acc”表示融入了图像序列位姿一致性约束的方法，“res50”表示将编码器结构从多层卷积结构改为ResNet50结构的方法。

[0134] 表5 消融实验结果

[0135]

方法	Rel. rot. (deg/m)	Rel. trans. (%)
basic	0.0443	11.88
kphm r12	0.0477	16.36
kphm r16	0.0362	9.80
kphm r32	0.0421	11.65
kphm r16 acc	0.0331	8.95
kphm r16 acc res50	0.0296	5.79

[0136] 本实施例中，深度估计网络和位姿估计网络最初采用的分别是ResNet18和FlowNet作为编码器，得到的结果如表5中前5行所示。实验时测试了不同关键点的影响半径δ下的实验结果，如表5中第二至第四行所示，发现在半径δ为16的时候结果最好。因此在之后的实验中，关键点热力图引导的半径δ将设为16。第五行为在融入了半径δ为16的关键点热力图引导的基础上融入图像序列位姿一致性约束的实验结果。可以看到，关键点热力图引导的方法使深度网络的训练更容易关注重点，使得网络的性能显著增强；而图像序列位姿一致性约束使得网络的学习更容易关注到更长跨度的帧之间的联系，使得方法的性能进一步提升。第六行为将深度估计网络和位姿估计网络的编码器部分改为ResNet50的结果。可以看到网络容量的增加使得性能有了进一步的提升。本发明实施例所述的关键点热力图引导的自监督单目视觉里程计方法的性能随着各个部分的增加而逐渐上升，证明了本发明实施例所述的关键点热力图引导的自监督单目视觉里程计方法中各个部分的意义。

[0137] 以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

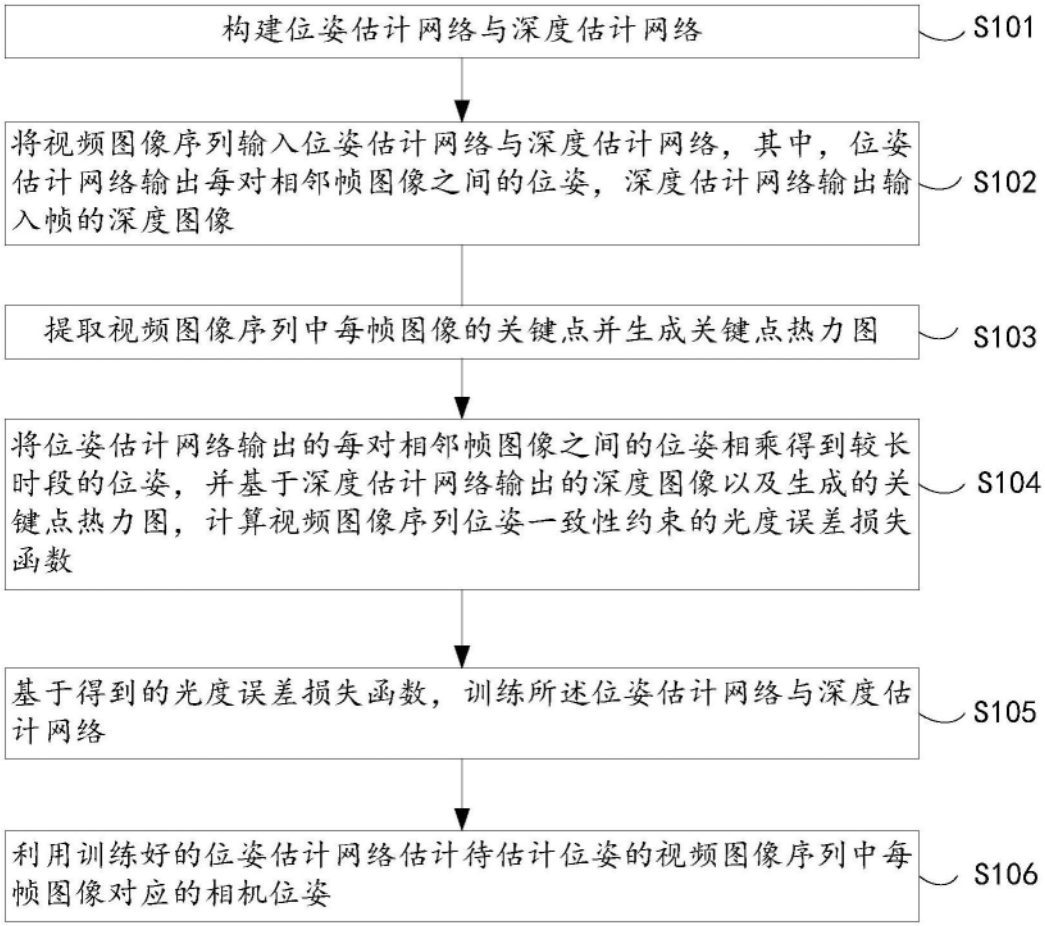

构建位姿估计网络与深度估计网络
S101
将视频图像序列输入位姿估计网络与深度估计网络，其中，位姿估计网络输出每对相邻帧图像之间的位姿，深度估计网络输出输入帧的深度图像
S102
提取视频图像序列中每帧图像的关键点并生成关键点热力图
S103
将位姿估计网络输出的每对相邻帧图像之间的位姿相乘得到较长时段的位姿，并基于深度估计网络输出的深度图像以及生成的关键点热力图，计算视频图像序列位姿一致性约束的光度误差损失函数
S104
基于得到的光度误差损失函数，训练所述位姿估计网络与深度估计网络
S105
利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿
S106

图1

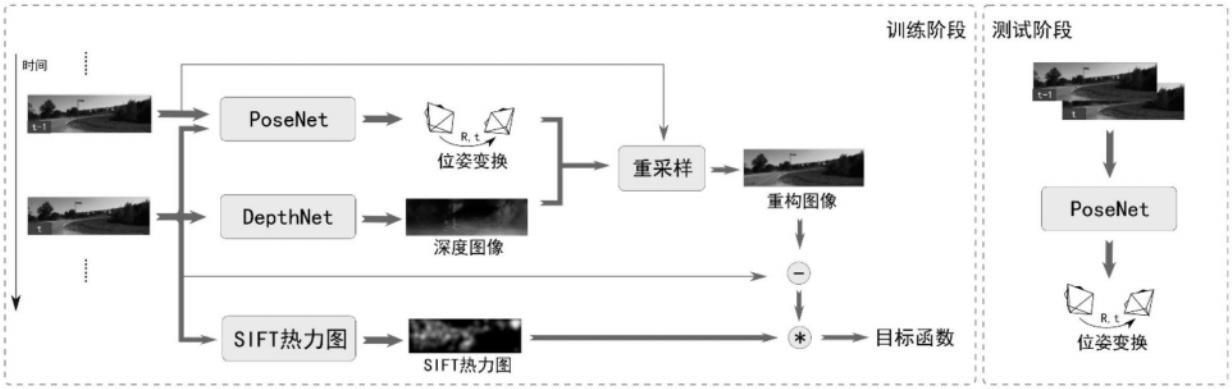
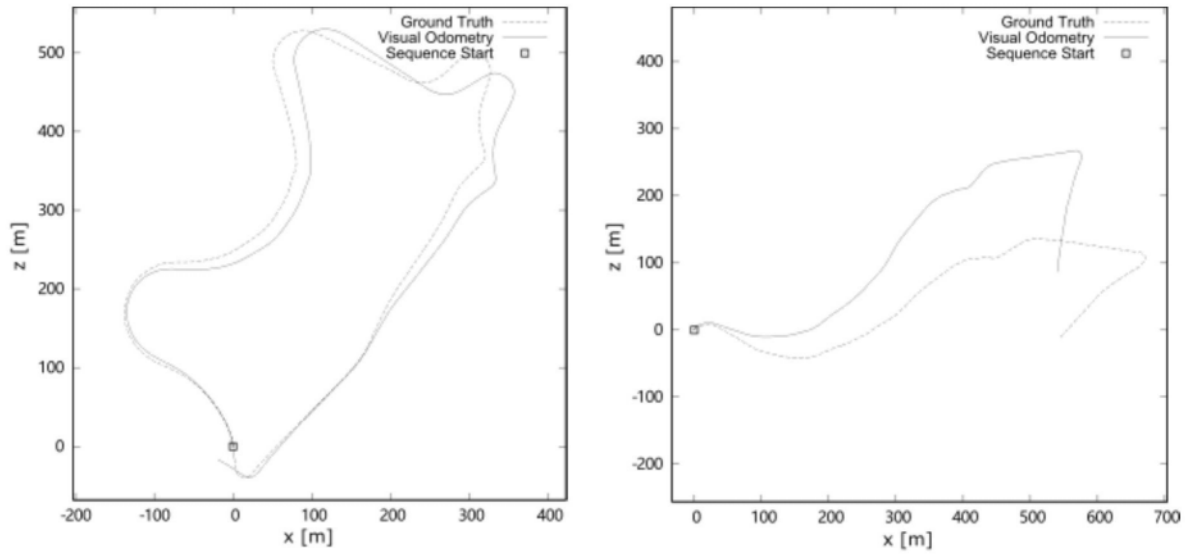

训练阶段
测试阶段
时间
PoseNet
位姿变换
DepthNet
深度图像
重采样
重构图像
SIFT热力图
SIFT热力图
目标函数
PoseNet
位姿变换

图2

(a) (b) (c)

图3

图4



(a) (b)

图5