计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(8):116−125 [doi: 10.15888/j.cnki.csa.009203]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

轻量化自监督单目深度估计①

刘　佳[1,2,3,4], 林　潇[1,2,3,4], 陈大鹏[1,2,3,4], 徐　闯[1,2,3,4], 石　豪[1,2,3,4]

[1](南京信息工程大学 自动化学院, 南京 210044)
[2](江苏省智能气象探测机器人工程研究中心, 南京 210044)
[3](江苏省大数据分析技术重点实验室, 南京 210044)
[4](江苏省大气环境与装备技术协同创新中心, 南京 210044)
通信作者: 陈大鹏, E-mail: dpchen@nuist.edu.cn

摘　要: 目前, 大多数的增强现实和自动驾驶应用不仅会使用到深度网络估计的深度信息, 还会使用到位姿网络估计的位姿信息. 将位姿网络和深度网络同时集成到嵌入式设备上, 会极大地消耗内存. 为解决这一问题, 提出一种深度网络和位姿网络共用特征提取器的方法, 使模型保持在一个轻量级的尺寸. 此外, 通过带有线性结构的深度可分离卷积轻量化深度网络, 使网络在不丢失过多细节信息前提下还可获得更少的参数量. 最后, 通过在 KITTI 数据集上的实验表明, 与同类算法相比, 该位姿网络和深度网络参数量只有的 35.33 MB. 同时, 恢复深度图的平均绝对误差也保持在 0.129.

关键词: 深度学习; 单目深度估计; 自监督学习; 轻量化; 计算机视觉

引用格式: 刘佳,林潇,陈大鹏,徐闯,石豪.轻量化自监督单目深度估计.计算机系统应用,2023,32(8):116–125. http://www.c-s-a.org.cn/1003-3254/9203.html

Lightweight Self-supervised Monocular Depth Estimation

LIU Jia[1,2,3,4], LIN Xiao[1,2,3,4], CHEN Da-Peng[1,2,3,4], XU Chuang[1,2,3,4], SHI Hao[1,2,3,4]

[1](School of Automation, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](Jiangsu Province Engineering Research Center of Intelligent Meteorological Exploration Robot, Nanjing 210044, China)
[3](Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing 210044, China)
[4](Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing 210044, China)

Abstract: Currently, most augmented reality and autonomous driving applications use not only the depth information estimated by the depth network but also the pose information estimated by the pose network. Integrating both the pose network and the depth network into an embedded device can be extremely memory-consuming. In view of this problem, a method of the depth and pose networks sharing feature extractors is proposed to keep the model at a lightweight size. In addition, the depth-separable convolutional lightweight depth network with linear structure allows the network to obtain fewer parameters without losing too much detailed information. Finally, experiments on the KITTI dataset show that compared with the algorithms of the same type, the size of the pose and deep network parameters is only 35.33 MB. At the same time, the average absolute error of the restored depth map is also maintained at 0.129.

Key words: deep learning; monocular depth estimation; self-supervised learning; lightweight; computer vision

深度图在许多计算机视觉任务中得到应用, 如解决增强现实中的虚实遮挡问题[1] 和解决无人驾驶的距离判断问题. 虽然结构光或激光雷达传感器等专业硬件设备也可以提供逐像素的深度图, 但这些主动获得

① 基金项目: 国家自然科学基金 (61773219, 62003169); 江苏产业前瞻与关键技术重点项目 (BE2020006-2); 江苏省自然科学基金青年基金 (BK20200823)
收稿时间: 2023-02-03; 修改时间: 2023-03-01, 2023-03-14; 采用时间: 2023-03-21; csa 在线出版时间: 2023-05-19
CNKI 网络首发时间: 2023-05-22

深度图的传感器不仅笨重、昂贵, 还会受到噪声和人为因素的影响. 此外, 还可以通过多视角立体视觉[2,3]和三维重建[4]等方法获得深度信息. 使用双目摄像头并利用多视角立体视觉算法和三维重建算法获得深度图, 计算量相对较大, 且对于低纹理场景的深度估计效果不好[5]. 由于使用深度学习训练好的网络模型进行深度估计能够使用较少的计算量来获得更好的效果, 因此使用深度学习的方法是上述方法的一个不错的代替方案.

近些年来, 随着深度学习在计算机视觉任务中的广泛应用, 许多工作都将深度学习与深度估计相结合. 江俊君等[6]将基于深度学习的深度估计分为监督学习和自监督学习两种途径. 其中监督学习将图像与传感器收集到的真实深度图信息作为输入[7,8], 并利用估计的深度与真实深度的不一致性作为损失函数来训练网络. 然而, 监督学习方法依赖于大型标记的 RGB-D 数据集, 获取这样的大型数据集成本较高. 为了避免使用大型标记数据集, 单目深度估计的自监督方法被提出. 文献 [9–11] 使用图像重建损失取代以地面真实深度构建的损失. 自监督学习的方法一般通过一个位姿网络 (PoseNet) 来辅助训练深度网络 (DepthNet), 但在预测深度时只使用 DepthNet. 然而, 在许多实际应用中不仅需要使用 DepthNet 还需要使用 PoseNet , 如自动驾驶、增强现实以及将整个网络集成到移动设备上进行实时训练. 同时将 DepthNet 和 PoseNet 集成到嵌入式设备上, 对嵌入式设备的计算能力的要求非常高, 无法有效地在存储和内存空间有限的嵌入式设备上运行. 为了提供密集的逐像素预测, 现在主流的自监督单目深度估计网络还继续延用 Dosovitskiy 等[12]提出的结构, 对所有特征映射进行上采样, 来为高分辨率任务提供局部精细的信息. 如果只是使用主流的轻量化编码器, 来轻量化网络, 并和 Dosovitskiy 等[12]提出的结构进行结合使用, 会导致特征图的细节过多的丢失导致不好的预测结果.

为了解决这个问题, 本文提出了一个有别于已有自监督单目深度估计模型的轻量化模型 (LightDepth). LightDepth 的重点在于轻量化 PoseNet 和 DepthNet 的编码器以及 DepthNet 的解码器, 同时减少细节的缺失. 与已有的将 DepthNet 和 PoseNet 分成两个独立网络的方法不同, 本文的 DepthNet 与 PoseNet 共用一个特征提取器 (见图 1). 此外, 本文通过使用带线性结构的深度可分离卷积构建 DepthNet 的编码器和解码器, 其不仅具有非常轻量级的参数, 还能在 KITTI数据集[13]上取得不错的预测效果 (见表 1).

本文的主要工作总结如下.

(1) 本文提出了一种新的轻量化的单目深度估计框架, 该框架中的 DepthNet 和 PoseNet 共用一个特征提取器.

(2) 本文设计了一个高效且轻量的 DepthNet, 其能保持在一个非常轻的尺寸, 同时用于从视频序列进行实时高性能无监督单目深度预测.

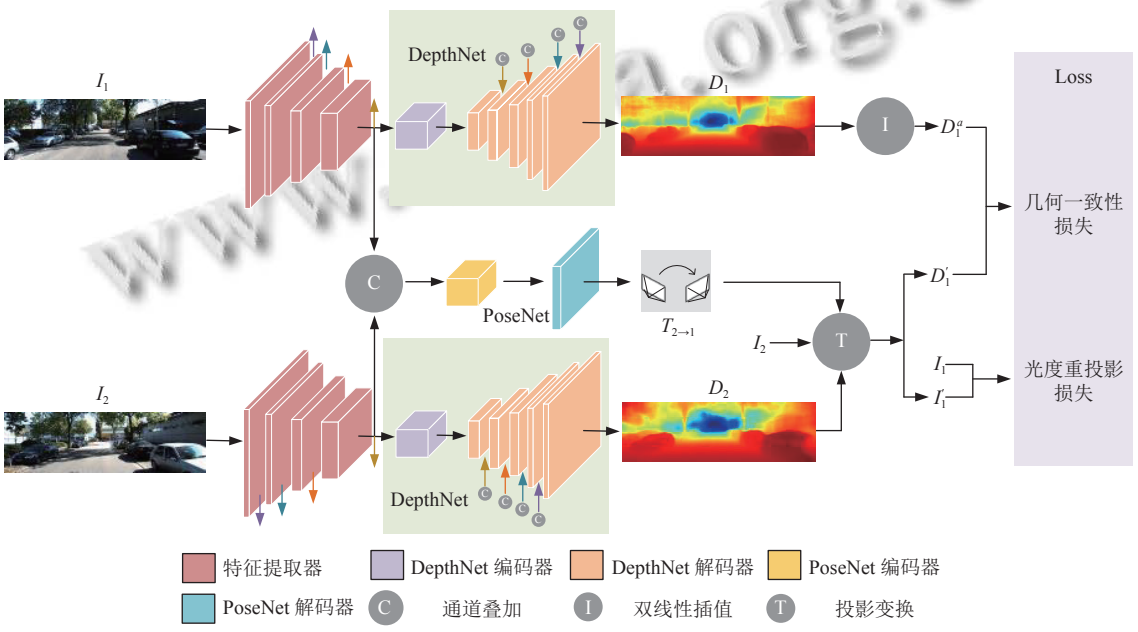


图 1 自监督单目深度估计网络结构框架

表 1 KITTI 数据集上的测试结果

方法	误差					准确率		
	Parameters (MB)	*AbsRel*	*SqRel*	*RMSE*	log*RMSE*	δ_1	δ_2	δ_3
Zhou等[10]	126	0.208	1.768	6.958	0.283	0.678	0.885	0.957
Mahjourian等[14]	144	0.163	1.240	6.220	0.250	0.762	0.916	0.968
EPC++[15]	146	0.141	1.029	5.250	0.216	0.816	0.944	0.974
Ranjan等[16]	527	0.140	1.070	5.326	0.217	0.826	0.941	0.975
Gordon等[17]	503	0.128	0.959	5.230	0.212	0.845	0.947	0.976
Monodepth2[18]	343	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Sc_depth[19]	181	0.114	0.813	4.706	0.191	0.873	0.960	0.982
Packnet-SFM[20]	495	0.106	0.838	4.545	0.186	0.895	0.964	0.982
Shu等[21]	630	0.104	0.729	4.481	0.179	0.893	0.965	0.984
ManyDepth[22]	191	0.093	0.715	4.245	0.172	0.909	0.966	0.983
LightDepth	35	0.129	0.880	5.209	0.205	0.836	0.949	0.981

1 相关工作

本节回顾了以 RGB 图像作为输入, 并在测试时估计每个像素的深度值作为输出的相关工作. 根据训练时是否使用真实深度, 可以将这些工作分为监督深度估计和自监督深度估计. 此外, 还对自监督单目深度估计任务的轻量化的工作进行了总结.

1.1 监督深度估计

单目深度估计的一种方法是监督学习. Eigen 等[7]最早使用卷积神经网络来解决深度估计问题. 对于输入的 RGB 图像中的每一个像素点, 模型都需要预测出深度图像对应像素的深度值, 预测的深度值可以作为监督信号来训练网络. Laina 等[23]首次将残差网络应用到深度估计领域. 该模型由一个编码器和一个解码器组成, 在编码端使用残差网络获得更强的特征提取能力. Liu 等[24]将连续条件随机场的知识与卷积神经网络相结合, 进行单幅图像的深度预测. Li 等[25]提出先用深度神经网络对超像素尺度的深度进行回归, 然后用条件随机场进行处理, 通过结合超像素和像素尺度的深度获得更好的预测结果. 由于深度值具有有序的特点, Fu 等[8]将深度分类问题归结为一个有序回归问题, 提出了使用间隔递增离散化 (SID) 策略来代替常规的均匀分段策略 (UD), 提高了估计精度. 受该工作的启发, Bhat 等[26]将分段策略进一步泛化, 使用网络训练出分段区间, 进一步提高了深度估计的精度. 监督单目深度估计能获得较高的估计结果, 但其需要地面真实深度作为监督信息来训练网络. 然而, 大量收集训练集所需的深度图是昂贵的.

1.2 自监督深度估计

单目深度估计的另一种是自监督学习. 其方法大致可以分为两类, 第 1 类使用立体图像对进行深度估计, 第 2 类使用单目相机获得的视频序列进行深度估计. Garg 等[9]最早提出使用无监督学习的方法来获得深度图, 其在训练阶段利用立体图像对. Godard 等[27]的工作将深度估计框架化为一个视图合成问题, 并通过最小化右视图重建误差和从左视图生成右视图之间的重建误差重建目标[28], 同时还引入了可微的插值函数来克服之前模型在训练阶段可能陷入局部最优的问题. Poggi 等[29]提出了使用 3 张图片之间的几何约束来训练模型, 这种方法有效地减少了遮挡物体对训练结果的影响.

相比使用立体图像对进行深度估计的网络, 在训练阶段使用单目视频序列的方法增加了一个姿态网络来估计帧间的相对位姿[10]. 尽管这种方法存在尺度模糊和动态物体对训练结果的影响[18]等限制, 但是单目自监督深度估计能够从原始视频中学习, 不需要真实的深度信息作为约束, 能够更容易的获得数据集, 更易于拓展. Godard 等[18]以像素为单位进行最小化重投影误差, 并提出以自动掩膜损失来减小动态物体对训练结果的影响. Shu 等[21]提出了特征度量损失来解决无纹理或低分辨率像素造成的局部最小问题. 马成齐等[30]提出一种自动屏蔽损失来损失函数来处理物体运动造成的边界伪影, 使预测效果细节更加饱满. Watson 等[22]认为在测试时可以使用视频帧形式的序列信息, 并提出了一种新的自监督深度估计模型, 其在测试时利用可用的多帧深度进行估计, 以达到比之前方法更好的效果. 这启发了相关的工作, Guizilini 等[20]提出一种新的结构用于成本体积生成, 并通过交叉注意力机制和自注意力机制来完善多视图特征匹配. 他们的网络获

得了较好的深度估计结果, 在很大程度上超过了其他自监督的方法, 甚至超过了监督的单帧架构. 但使用成本体积量的方式会极大地消耗计算量, 所有本文在网络中并未使用成本体积量的形式.

1.3 轻量化的自监督单目深度估计

虽然基于视频序列的自监督单目深度估计的数据容易获得, 也易于拓展, 但其需要更复杂、更深的网络架构. 由于深度估计在自动驾驶、增强现实等任务下对模型有推理速度、规模量级的限制要求, 需要使复杂的深度估计网络轻量化, 以减少参数量和计算量, 同时其估计的深度精度应保持在一个合理的范围内. 对于轻量化的自监督单目深度估计, Poggi 等[31] 利用立体图像对进行训练, 使用从单个输入图像中提取的特征金字塔快速地推断出精确的深度图, 其减少了参数量, 并在 CPU 上获得了实时性. Liu 等[32] 利用视频帧进行训练, 提出了使用循环神经网络 (RNN) 的一个单元循环提取特征的方法来减小模型参数量. 该网络只轻量化了 DepthNet, 并没有轻量化 PoseNet. 由于在增强现实等应用中, PoseNet 估计的两帧之间的位姿也可以被利用. 还有些移动端需要实时的进行训练, 也需要将 PoseNet 和 DpehtNet 一起集成到设备上. 本文从这个角度出发, 提出了一种轻量化网络 LightDepth, 其 DepthNet 和 PoseNet 总体参数量为 35.33 MB, 并且在轻量化网络的同时还能在 KITTI 数据集上取得不错的效果, 其恢复深度图的平均绝对误差也保持在 0.129.

2 本文方法

2.1 LightDepth 的整体流程

与之前的工作[21,30] 一样, LightDepth 也是以最小化光度重投影误差进行建模. 该模型主要是将目标帧 I_1与参考帧I_2分别送入到 DepthNet 网络中来估计它们的深度D_1和D_2. 然后使用 PoseNet 网络估计它们之间的 6 自由度的相对位姿$T_{2\to1}$. 根据预测的深度和相对位姿, 利用可微分双线性插值I_2[28] 来获得合成图像 I'_1. 最后利用目标帧I_1与合成图像I'_1之间的光度损失来训练网络. 为了更好地指导编码器提取带有平移和旋转信息的特征, 本文使用了几何一致性损失. 通过相对位姿$T_{2\to1}$和相机内参将参考帧的深度D_2投影到目标帧I_1的像素平面上获得D'_1. 最后利用D'_1和D_1插值得到的D_1^a之间的不一致性作为损失函数来训练网络.

本文的目的是在轻量化网络的同时, 取得不丢失过多细节信息的预测结果. 为了让 DepthNet 解码器获得精细的局部信息, 网络保留了 Dosovitskiy 等[12] 提出的结构, 其对所有特征映射进行上采样, 同时为了减少参数量, 将 PoseNet 和 DepthNet 共用一个特征提取器, 见图 1. ResNet-18[33] 的最后一层由 4 个通道数为 512 的 3×3 卷积和一个平均池化组成, 其参数量为 34 MB, 所以本文的特征提取器只使用了 ResNet-18 的前 3 层, 其输出的是通道数为 256 的张量, 参数量只有 10.4 MB. 这不仅保证了提取出的特征具有代表性, 还能保证网络的轻量性.

2.2 编码器

如图 2 所示, 本文的 DepthNet 编码器主要由一个 1×1 的卷积层和一个循环模块组成. 本文使用 1×1 的卷积层对特征编码器提取出来的特征进行升维, 然后通过 ReLU 激活, 将该层的输出通道数设置为 512. 本文提出的循环模块主要由 3 个带有线性结构的深度可分离卷积组成, 其 3 个深度可分离卷积交替放置. 每个模块循环使用两次. 为了实现复用性, 每个模块的输入通道和输出通道设计为相同都是 512. 图 2 中颜色相同的线性深度可分离卷积代表同一个卷积, 其使用同一组参数. 由于深度可分离卷积[34] 提取的图片特征不充分, 在本文任务中会导致部分物体轮廓丢失的情况. 从 Han 等[35] 中获得启发, 通过更廉价的线性操作来减少特征图的冗余, 获得更多信息的特征图. 本文的线性深度可分离卷积结构如图 2, 其主要由两个 1×1 的卷积层和一个 3×3 的分层卷积组成. 为了让 3×3 的可深度分离卷积和 1×1 的卷积结果的特征联系更加紧密, 本文通过将两个卷积出来的结果在通道数上叠加, 再通过一个 1×1 的卷积来学习跨通道信息.

如图 3 所示, PoseNet 的编码器由一个循环模块组成. 本文将从特征提取器获得的两张图片的特征在通道上进行叠加得到通道数为 512 的张量, 并将其作为 PoseNet 的输入. 为了将两张图片特征的信息更好的聚合起来, 循环模块的第 1 层使用的是一个标准的 1×1 卷积, 输入通道数为 512, 输出通道数为 256, stride 为 1. 从 RNN 循环网络中得到启发, 本文提出的循环模块由 1 个 3×3 的卷积组成. 该 3×3 的卷积的输入和输出通道数设计为相同为 256. 首先, 将输入张量送入到 1×1 的卷积层中, 获得大小为 512×16×52 的张量. 然后, 将获得的张量送入到 3×3 的循环卷积层中, 在反复

经过 3 次同一个 3×3 的循环卷积层后, 结束循环模块. 为了减少计算量, 在循环模块的最后两层卷积模块的后面使用最大池化层来降低空间维度. 最后, 循环模块输出的张量大小为 256×4×13.

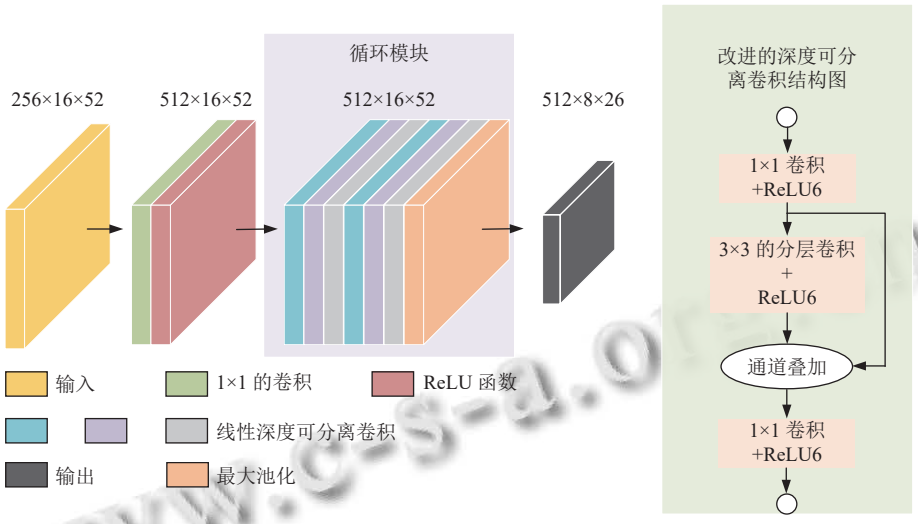


图 2 DepthNet 编码器

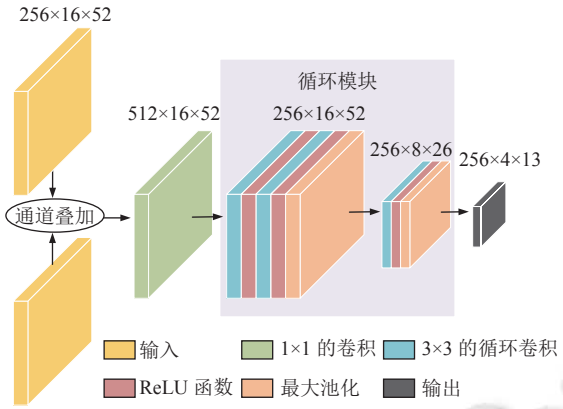


图 3 PoseNet 编码器

2.3 DepthNet 解码器

为了满足高精度和实时性的要求, 本文设计了一种新型高效的上采样模块对编码器输出的特征图进行上采样和聚合, 如图 4. 本文将从 DepthNet 编码器获得的特征图直接作为解码器的输入. 与解码器 disponet[10] 全使用 3×3 的卷积不同, 本文的解码器使用线性深度可分离卷积与标准的 3×3 卷积交替使用的方法. 交替使用线性深度可分离卷积与标准的 3×3 的卷积可以在轻量化网络的同时, 还能指导网络不丢失训练目标, 保证了不错的预测效果. 为了提高预测精度, 在经过改进的深度可分离卷积后, 采用双线性插值方法将多尺度的特征图插值到相同的空间分辨率, 再进行通道拼接. 本文解码器在输出层的激活是 Sigmoid, 其他地方的激活函数使用的是 ELU 函数. 本文将 Sigmoid 输出 x 转换为深度, $D=1/(ax+b)$, 其中选择a和b将D约束在 0.1 到 100 单位之间.

2.4 损失函数

在本节中, 将介绍训练 LightDepth 的损失函数. 本文的损失函数主要由 3 部分构成, 其分别为重投影光度损失L_p, 几何一致性损失L_g和平滑损失L_s:

$$L=\alpha ML_p+\beta L_g+\delta L_s \tag{1}$$

其中, α, β, δ是每一项的权重参数, 用于调整各个损失函数对训练的影响, 分别为 1.0, 0.5, 0.1. 与文献 [16,21] 类似, 光度损失 L_p 是由合成图像I_1'与目标帧 I_1 的差值的 L1 范数与 *SSIM* 函数[28] 组成, 定义为:

$$L_p=\sum_{p\in I}\left(\lambda\left\|I_1'(p)-I_1(p)\right\|_1+\gamma\frac{1-SSIM\left(I_1'(p),I_1(p)\right)}{2}\right) \tag{2}$$

其中, I表示I_1'与I_1成功匹配的点的个数. *SSIM* 函数代表两张图片的相似性, 用于更好地处理光照变化的场景. λ和γ为权重参数, 其分别为 0.15, 0.85.

为了减小动态物体对恢复深度图效果的影响, 本文引入了 Bian 等[19] 提出的动态掩膜M和几何一致性损失L_g, 其定义为:

$$L_g = \sum_{p\in I} \frac{\left|D_1^a(p) - D_1'(p)\right|}{D_1^a(p) + D_1'(p)} \tag{3}$$

$$M = 1 - \frac{\left|D_1^a(p) - D_1'(p)\right|}{D_1^a(p) + D_1'(p)} \tag{4}$$

其中, $D_1^a(p)$ 表示目标帧估计的深度图D_1通过双线性插值获得的合成深度图, $D_1'(p)$表示参考帧估计的深度图D_2投影到目标帧深度图D_1坐标上的投影深度图.

式 (1) 中的ML_p表示M与L_p之间对应的元素相乘. M中每个值对应图像中的每个像素点的重建误差权重, 通过减小掩膜的取值, 来减小图片中动态物体对训练结果的影响.

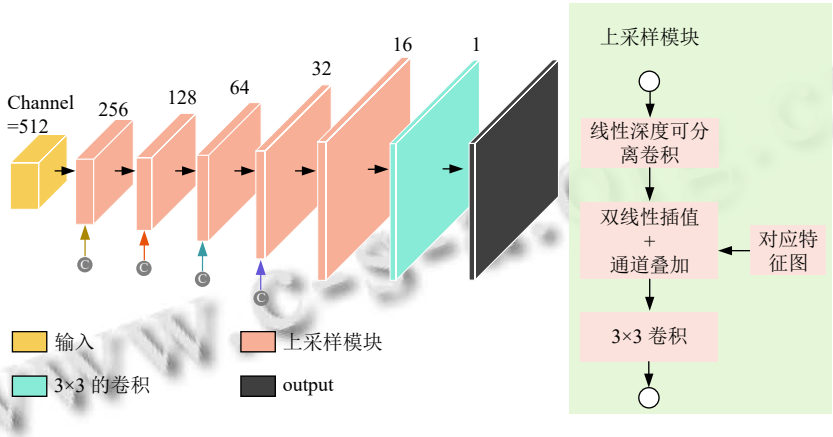


图 4 DepthNet 解码器

在低纹理场景下, 光度损失的效果不佳, 本文采用了平滑损失L_s[36], 其定义为:

$$L_s = \sum_p \left(\mathrm{e}^{-\nabla I1(p)}.\nabla D_1(p)\right)^2 \tag{5}$$

其中, p表示图片中的所有像素点, ∇表示一阶梯度.

3 实验分析

3.1 实验细节

本文在 KITTI 原始数据集[13] 上使用文献 [7] 的数据集分割方式进行模型的训练和测试. 其中 697 张图片作为测试集, 39 810 张图片作为训练集, 4 424 张图片作为验证集. 其中每幅图片的分辨率约为 1226×370 像素. 为了提高效率, 并保持原始图片的长宽比, 将 LightDepth 输入图片大小设置为 832×256. 为了增强数据的随机性, 在训练阶段本文通过随机裁剪缩放和水平翻转进行数据增强. LightDepth 使用的是 PyTorch Library[37] 框架, 在单张 NVIDIA RTX 3090 上进行训练, 系统为 Ubuntu 16.04.6, 训练时间为 16 h. 3 张连续帧被作为 LightDepth 的输入, 其中将第 2 帧作为参考帧, 计算参考帧与其他两帧的损失函数. 为了最大程度上利用数据, 将参考帧与其他两帧互换角色, 再次计算损失. 本文使用 Adam[38] 优化器, 学习率设置为 0.000 07, batch_size 设置为 8, 总计训练 200 轮. 在每一轮中随机抽取 1 000 张图片进行训练. 在对恢复深度图的效果进行评价时, 本文使用的是和 Zhou 等[10] 一样的指标. 使用平均绝对误差 (*AbsRel*)、均方根误差 (*RMSE*)、均方根对数误差 (log*RMSE*) 和在阈值 (δ) 下的精度, 其中除了在阈值下的精度要求越大越好, 其余的误差都是要求越小越好. 损失函数的表达式分别为:

$$\begin{cases} AbsRel = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\frac{\left|D_i - D_i^*\right|}{D_i^*} \\ SqRel = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\frac{\left|D_i - D_i^*\right|^2}{D_i^*} \\ RMSE = \sqrt{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left|D_i - D_i^*\right|^2} \\ \log RMSE = \sqrt{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left|\lg D_i - \lg D_i^*\right|^2} \\ \delta = \max\left(\dfrac{D_i}{D_i^*}, \dfrac{D_i^*}{D_i}\right) < T \end{cases}$$

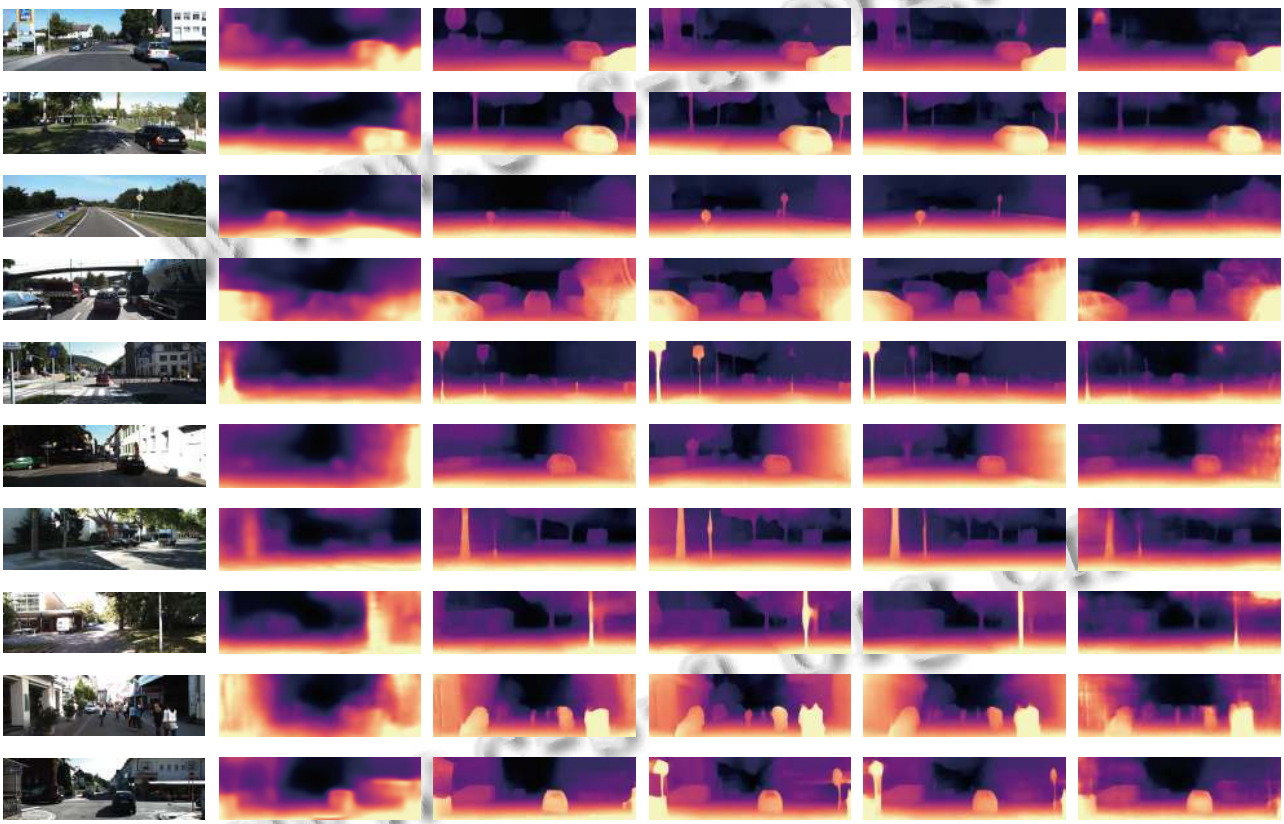
其中, N表示像素总数, D_i表示第 i 个像素的估计深度值, D_i^*表示第 i 个像素对应的真实深度值. δ_1表示小于

门槛$T = 1.25$的比例, δ_2表示小于门槛$T^2 = 1.25^2$的比例, δ_3表示小于门槛$T^3 = 1.25^3$的比例.

3.2 KITTI 数据集评测

表 1 显示了在 KITTI 数据集[13] 上的评估结果, 其第 1 列列出了每个网络总参数的大小. LightDepth 在总体参数量上获得了最小的参数 (35.33 MB). 在预测深度图精度方面, 本文的方法恢复深度图的 *AbsRel* 获得了 0.129. 值得注意的是 Guizilini 等[39] 使用了语义标签学习, Zhao 等[40] 使用了光流学习来提高精确度. 而该算法使用不附加信息的单目方法就获得了不错的预测精度, 并轻量化了网络, 这证明了共用一个编码器方法的有效性和本文使用的轻量化的 DepthNet 的合理性. 与 ManyDepth[22] 相比, 虽然 LightDepth 的 *AbsRel* 下降 0.036, 但是 ManyDepth 的参数量是 LightDepth 的 5.5 倍.

图 5 中展示了 10 张 RBG 图片和各个网络估计出来的视差图. 虽然 LightDepth 在勾勒物体细节方面要弱于最近的工作, 但其可以清晰地画出物体轮廓, 极少的存在物体丢失的现象. 这对改善在自动驾驶和无人机安全等应用中具有重要意义.



(a) RGB 图片 (b) Zhou等[10] (c) Sc_depth[19] (d) Monodepth2[24] (e) ManyDepth[12] (f) 本文算法

图 5 KITTI 数据集上的可视化结果

3.3 消融实验

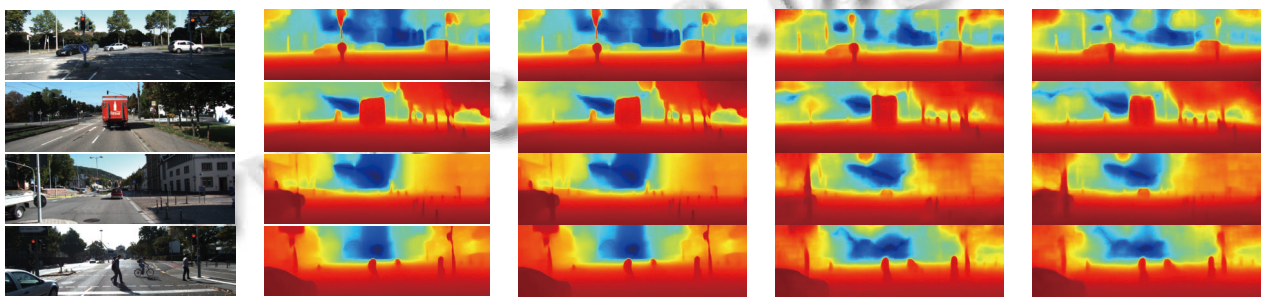
为了更好地展示共用特征提取器和轻量化的 DepthNet 在无监督单目深度估计中的整体性能, 本文进行了消融实验. 结果见表 2. “w/o R”方法表示 DepthNet 和 PoseNet 没有共用一个特征编码器. 其使用两个独立 ResNet-18 网络进行作为 PoseNet 和 DepthNet 网络的编码器, 再通过各自的解码器进行深度估计. “w/o E1”方法表示两个网络共用特征提取器, 并对 PoseNet 的编码器进行了轻量化的处理, 而 DepthNet 依然使用的是 ResNet-18 的最后一层作为编码器. “w/o E2”表示两个网络共用特征提取器, 并对两个网络的编码器进行了轻量化处理. “估计图片的时间”表示估计测试集中 697 张图片所需要的时间.

通过对表 2 的方法“w/o R”和方法“w/o E1”进行对比, 发现共用特征提取器的方法, 能够大量的减小网络的参数量, 但是其对预测结果的影响也比较大.

通过方法“w/o E1”和方法“w/o E2”的对比发现, 使用改进的深度可分离卷积作为编码器, 其能够轻量化编码器, 并且能实现深度网络的功能. 通过方法“w/o E2”和本文方法对比发现, 虽然本文的线性深度可分离卷积在通道数小的时候会增加参数量, 但是其可以将更多的性能放在小通道上, 用以获得更好的效果, 并且其在小通道数增加的参数量相比大通道减少的参数量可以忽略不计. 因此, 本文方法可以帮助嵌入式设备节省存储和内存空间. 图 6 给出消融实验估计的深度图.

表 2 消融实验结果

方法	特征提取器参数 (MB)	深度网络参数量 (MB)	位姿网络参数量 (MB)	估计图片的时间 (s)	*AbsRel*	δ_1
w/o R	20.8	43.8	39.3	13.068	0.117	0.867
w/o E1	10.4	46.2	7.28	12.337	0.125	0.849
w/o E2	10.4	21.95	7.28	10.391	0.131	0.829
本文算法	10.4	17.65	7.28	10.255	0.129	0.836



(a) RGB 图片 (b) w/o R (c) w/o E1 (d) w/o E2 (e) 本文算法

图 6 消融实验可视化结果

为了验证本文使用 3 个损失函数对本文网络的影响, 本文同样对损失函数进行消融实验. 表 3 为使用不同损失函数时网络训练结果, 其中 val-*AbsRel* 表示在验证集上的平均绝对误差, tes-*AbsRel* 表示在测试集上的平均绝对误差. 由于轻量化的原因, 当网络只使用损失函数L_p时网络无法拟合, 这时需要更多的约束来指导网络训练. 当同时使用L_p和L_g时, 网络获得了更多的约束条件, 这为网络提供训练目标, 使其获得不错的深度预测效果. 当同时使用L_p、L_g和L_s时网络的 val-*AbsRel* 和上一个并无太大差别, 但是网络的 tes-*AbsRel* 获得了较好的结果. 图 7 给出了使用不同损失函数时 val-*AbsRel* 的曲线图. 当只使用损失函数L_p时其误差一直在 0.49–0.5 之间波动, 未能找到训练目标, 所以在图 7 中是一条直线.

表 3 不同损失函数下的误差

方法	val-*AbsRel*	tes-*AbsRel*	δ_1
L_p	0.494	0.443	0.303
L_p+L_g	0.165	0.131	0.831
L_p+L_g+L_s	0.164	0.129	0.836

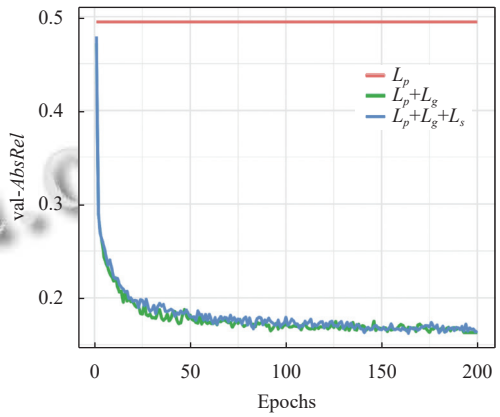


图 7 不同损失函数下的验证误差曲线

4 结论与展望

本文提出了一种轻量化的自监督单目深度估计网络 LightDepth, 其 DepthNet 和 PoseNet 共用一个特征提取器. 该轻量化结构大大降低了 PoseNet 的网络参数量. 此外, 还设计了一个高效的 DepthNet, 其使用线性深度可分离卷积大大的轻量化了网络参数量. 通过在 KITTI 数据集上进行了大量实验, 证明本文提出的

轻量化网络具有有效性和高效率. 该方法具有轻量化和估计精度较好的优点, 在增强现实和自动驾驶等应用上具有广泛的应用前景.

参考文献

1 刘万奎, 刘越. 用于增强现实的光照估计研究综述. 计算机辅助设计与图形学学报, 2016, 28(2): 197–207. [doi: 10.3969/j.issn.1003-9775.2016.02.001]

2 Hou YL, Peng JW, Hu ZH, *et al*. Planarity constrained multi-view depth map reconstruction for urban scenes. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 139: 133–145. [doi: 10.1016/j.isprsjprs.2018.03.003]

3 Mostegel C, Fraundorfer F, Bischof H. Prioritized multi-view stereo depth map generation using confidence prediction. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 143: 167–180. [doi: 10.1016/j.isprsjprs.2018.03.022]

4 Zeller N, Quint F, Stilla U. Depth estimation and camera calibration of a focused plenoptic camera for visual odometry. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 118: 83–100. [doi: 10.1016/j.isprsjprs.2016.04.010]

5 公冶佳楠, 李轲. 基于光场图像序列的自适应权值块匹配深度估计算法. 计算机系统应用, 2020, 29(4): 195–201. [doi: 10.15888/j.cnki.csa.007387]

6 江俊君, 李震宇, 刘贤明. 基于深度学习的单目深度估计方法综述. 计算机学报, 2022, 45(6): 1276–1307. [doi: 10.11897/SP.J.1016.2022.01276]

7 Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: ACM, 2014. 2366–2374.

8 Fu H, Gong MM, Wang CH, *et al*. Deep ordinal regression network for monocular depth estimation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2002–2011.

9 Garg R, B. G. VK, Carneiro G, *et al*. Unsupervised CNN for single view depth estimation: Geometry to the rescue. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 740–756.

10 Zhou TH, Brown M, Snavely N, *et al*. Unsupervised learning of depth and ego-motion from video. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6612–6619.

11 Zhan HY, Garg R, Weerasekera CS, *et al*. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 340–349.

12 Dosovitskiy A, Fischer P, Ilg E, *et al*. FlowNet: Learning optical flow with convolutional networks. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 2758–2766.

13 Geiger A, Lenz P, Stiller C, *et al*. Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research, 2013, 32(11): 1231–1237. [doi: 10.1177/0278364913491297]

14 Mahjourian R, Wicke M, Angelova A. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5667–5675.

15 Luo CX, Yang ZH, Wang P, *et al*. Every pixel counts++: Joint learning of geometry and motion with 3D holistic understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(10): 2624–2641. [doi: 10.1109/TPAMI.2019.2930258]

16 Ranjan A, Jampani V, Balles L, *et al*. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12232–12241.

17 Gordon A, Li HH, Jonschkowski R, *et al*. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 8976–8985.

18 Godard C, Mac Aodha O, Firman M, *et al*. Digging into self-supervised monocular depth estimation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 3827–3837.

19 Bian JW, Zhan HY, Wang NY, *et al*. Unsupervised scale-consistent depth learning from video. International Journal of Computer Vision, 2021, 129(9): 2548–2564. [doi: 10.1007/s11263-021-01484-6]

20 Guizilini V, Ambruş R, Chen D, *et al*. Multi-frame self-supervised depth with transformers. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 160–170.

21 Shu C, Yu K, Duan Z X, *et al*. Feature-metric loss for self-supervised learning of depth and egomotion. Proceedings of the 16th European Conference on Computer Vision.

Glasgow: Springer, 2020. 572–588.
22 Watson J, Mac Aodha O, Prisacariu V, *et al*. The temporal opportunist: Self-supervised multi-frame monocular depth. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 1164–1174.
23 Laina I, Rupprecht C, Belagiannis V, *et al*. Deeper depth prediction with fully convolutional residual networks. Proceedings of the 4th International Conference on 3D Vision (3DV). Stanford: IEEE, 2016. 239–248.
24 Liu FY, Shen CH, Lin GS. Deep convolutional neural fields for depth estimation from a single image. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 5162–5170.
25 Li B, Shen CH, Dai YC, *et al*. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1119–1127.
26 Bhat SF, Alhashim I, Wonka P. AdaBins: Depth estimation using adaptive bins. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 4008–4017.
27 Godard C, Mac Aodha O, Brostow GJ. Unsupervised monocular depth estimation with left-right consistency. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6602–6611.
28 Wang Z, Bovik AC, Sheikh HR, *et al*. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600–612. [doi: 10.1109/TIP.2003.819861]
29 Poggi M, Tosi F, Mattoccia S. Learning monocular depth estimation with unsupervised trinocular assumptions. Proceedings of the 2018 International Conference on 3D vision (3DV). Verona: IEEE, 2018. 324–333.
30 马成齐, 李学华, 张兰杰, 等. 抗遮挡的单目深度估计算法. 计算机工程与应用, 2021, 57(2): 217–222. [doi: 10.3778/j.issn.1002-8331.1911-0346]
31 Poggi M, Aleotti F, Tosi F, *et al*. Towards real-time unsupervised monocular depth estimation on CPU. Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2018. 5848–5854.
32 Liu J, Li Q, Cao R, *et al*. MiniNet: An extremely lightweight convolutional neural network for real-time unsupervised monocular depth estimation. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166: 255–267. [doi: 10.1016/j.isprsjprs.2020.06.004]
33 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: 2016. 770–778.
34 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017.
35 Han K, Wang YH, Tian Q, *et al*. Ghostnet: More features from cheap operations. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1577–1586.
36 Pilzer A, Lathuilière S, Sebe N, *et al*. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9760–9769.
37 Paszke A, Gross S, Chintala S, *et al*. Automatic differentiation in PyTorch. Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach: NIPS, 2017.
38 Kingma DP, Ba LJ. Adam: A method for stochastic optimization. Proceedings of the 2015 International Conference on Learning Representations. San Diego: ICLR, 2015.
39 Guizilini V, Hou R, Li J, *et al*. Semantically-guided representation learning for self-supervised monocular depth. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR, 2020.
40 Zhao W, Liu SH, Shu YZ, *et al*. Towards better generalization: Joint depth-pose learning without PoseNet. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9148–9158.

(校对责编: 孙君艳)