

基于视觉注意的自监督单目深度估计

程德强¹⁾, 徐帅¹⁾, 韩成功¹⁾, 吕晨¹⁾, 寇旗旗²⁾, 张剑英^{1)*}

¹⁾ (中国矿业大学信息与控制工程学院 徐州 221116)

²⁾ (中国矿业大学计算机科学与技术学院 徐州 221116)

(zjysice@cumt.edu.cn)

摘 要: 为解决单目深度估计方法在面对复杂纹理区域时预测效果差的问题, 提出了一种基于视觉注意的自监督单目深度估计方法. 首先构建多尺度源图像输入编码器, 从而更好地融合多路特征; 其次, 通过并行的中间注意力模块跨区域交互, 分别在空间维度和通道维度建模语义依赖关系, 得到丰富的上下文信息; 此外, 采用连续的外部注意力特征聚合模块构成解码器部分, 有效利用上下文信息解决复杂区域的不适应问题. 在 KITTI 和 Cityscapes 数据集上的实验结果表明, 该方法优于目前主流的方法, 在复杂纹理区域拥有更好的深度预测性能. KITTI 数据集的 RMS 和 RMSln 分别达到了 4.486 和 0.181.

关键词: 图像处理; 深度估计; 自监督学习; 视觉注意

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.20106

Visual Attention-Based Self-Supervised Monocular Depth Estimation

Cheng Deqiang¹⁾, Xu Shuai¹⁾, Han Chenggong¹⁾, Lyu Chen¹⁾, Kou Qiqi²⁾, and Zhang Jianying^{1)*}

¹⁾ (School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116)

²⁾ (School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116)

Abstract: To solve the problem of poor prediction of complex texture areas with existing monocular depth estimation methods. We propose a visual attention-based self-supervised monocular depth estimation method. Firstly, this method provides better fusion of multiple features by fusing multi-scale source images as input to the encoder. Then through parallel intermediate attention modules interacted across regions, we modelled semantic dependencies in the spatial dimension and channel dimension respectively, to obtain rich contextual information. In addition, the continuous external attentional feature aggregation module was used to form the decoder part, which effectively used contextual information to solve the maladjustment problem in complex regions. Experimental results on KITTI and Cityscapes datasets show that our method is better than the current mainstream methods, with better depth prediction performance in complex texture regions. In the KITTI dataset, RMS and RMSln reached 4.486 and 0.181, respectively.

Key words: image processing; depth estimation; self-supervised learning; visual attention

深度估计在机器人导航、增强现实、三维重建等领域有着广阔的应用前景^[1]. 传统方法利用传感

器测量成像设备中所有或部分像素到目标物体的距离, 但是使用这些设备需要高昂的成本. 单目深

收稿日期: 2022-12-29; 修回日期: 2023-07-01. 基金项目: 国家自然科学基金(52204177); 中央高校基本科研业务费专项资金(2020QN49). 程德强(1979—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为图像处理、机器视觉; 徐帅(1999—), 男, 硕士研究生, 主要研究方向为三维重建; 韩成功(1998—), 男, 博士研究生, 主要研究方向为三维重建; 吕晨(1994—), 男, 博士研究生, 主要研究方向为三维重建; 寇旗旗(1988—), 男, 博士, 讲师, 主要研究方向为模式识别; 张剑英(1963—), 女, 博士, 教授, 硕士生导师, 论文通信作者, 主要研究方向为图像处理.

度估计可以在不使用额外测量的情况下估计给定二维图像中像素的深度, 因此, 成为了目前计算机视觉领域的研究热点. 得益于深度学习的快速发展, 卷积神经网络(convolutional neural network, CNN)以其自适应的特征提取和强大的非线性映射能力, 在单目深度估计中发挥了重要作用. 大多数基于 CNN 的方法使用有监督深度标签训练网络, 尽管这些方法在深度估计上取得了不错的效果, 但真实深度标签的信息采集比较困难、代价较高. 因此, 不需要标准真实深度标签的自监督单目深度估计得到广泛使用.

Zhou 等^[2]首次提出自监督单目深度估计方法, 利用视频序列中连续帧的集合联合训练深度和位姿网络, 但是其使用运动恢复结构构造监督信息, 会受到低纹理区域、遮挡和运动物体的影响; 此外, 位姿网络的不确定性会破坏训练的稳定性, 导致产生错误的深度值. 为了解决这些内在的问题, Yin 等^[3]提出联合深度和光流的自监督学习框架, 可以通过光流将自我运动和深度信息关联起来. Godard 等^[4]使用最小重投影误差和自动掩模损失处理遮挡场景以及运动物体的问题. Jung 等^[5]将语义特征作为辅助信息优化深度估计, 从而减小对光度一致性损失的依赖.

上述方法极大地促进了自监督单目深度估计的发展, 但最后生成的深度图仍存在着很多问题, 它们都采用基于 ResNet(residual network)^[6]的编码器-解码器架构. 该架构使用跳跃连接恢复语义特征和空间信息, 尽管对提取局部信息和抽象层次特征的效果显著, 但不足以区分重要细节和干扰的信息, 无法充分地利用高层次特征和低层次特征, 因此很难在一些复杂纹理区域以及远景区域预测准确的深度估计结果. 针对上述问题, 本文提出了基于视觉注意的多尺度特征融合(visual attention-based multi-scale feature fusion, VAMF)单目深度估计网络. 首先, 采用多尺度数据增强的方法, 将源图像多次下采样后的不同尺度图像融合输入到编码器网络, 同时基于设计的视觉注意模块构建解码器. 其次, 提出一种有效的注意力模块用于连接编码器和解码器, 该模块采用位置注意模块和通道注意模块并联的方式, 并将二者的输出融合以进一步增强特征表示. 这种结合的框架可以充分地挖掘不同层次特征, 从而提高深度估计的准确性. 基于充分的实验比较和结果分析, 在 KITTI 数据集^[7]和 Cityscapes 数据集^[8]上验证了 VAMF 网络

结构和注意力模块的有效性.

1 相关工作

1.1 监督单目深度估计

单目深度估计最早采用监督算法, 旨在不进行任何额外测量的情况下预测单幅二维图像中的像素深度. Eigen 等^[9]和 Liu 等^[10]提出多尺度网络结构的单目深度估计网络框架, 通过最小化预测深度与真实深度之间的误差, 逐像素地估计输入图像的深度图, Xu 等^[11-12]将连续条件随机场模块和 CNN 结合以融合多尺度特征. 此外, 一些方法利用先验知识^[13-14]、时空一致性^[15]和多任务约束^[16]提高深度估计网络的性能. 然而, 这些连续回归方法通常会遇到收敛速度慢和局部解不理想的问题. Fu 等^[17]使用间隔离散化策略将连续深度离散化为多个间隔, 并将深度网络学习转换为序列回归问题. Bhat 等^[18]提出了基于 Transformer 架构的单目深度估计方法, 它不使用传统的卷积, 而是将深度范围划分为不同的区间, 同时依赖注意力机制使其预测更加高效, 进一步提升了单目深度估计的效果. 但是基于监督学习的单目深度估计方法需要大量标准的真实深度标签, 因此该方法的使用受到限制.

1.2 自监督单目深度估计

基于自监督学习的方法将单目深度估计任务转换为图像重建问题, 并使用立体图像对或单目序列中的光度损失作为监督信号训练网络. Garg 等^[19]提出一种基于立体视觉的自监督单目深度估计网络框架, 利用立体图像对构建重建损失, 在预测阶段使用单目左视差图作为模型的输入. Godard 等^[20]基于文献[19]的方法进行改进, 可以在仅输入左视差图的情况下对左右图像对应的视差图进行预测. 由于单目视频序列比立体图像对更容易获取, 因此, Zhou 等^[2]利用单目视频序列训练深度估计模型, 用 2 个网络分别估计单帧图像的深度和视频序列中相机的姿态变化. Godard 等^[4]提出了每像素最小重投影损失和多尺度损失来解决遮挡问题和局部梯度问题. Wang 等^[21]在文献[4]的基础上, 利用特征度量损失、时间几何一致性进一步提高了单目深度估计精度.

此外, 很多研究者还将其他信息作为辅助特征引入到单目深度估计的任务中, Yin 等^[3]和 Zou 等^[22]使用光流信息处理移动对象; Chen 等^[23]通过度量深度图和语义图之间的一致性, 提出了将语

义信息作为监督信号,引导网络学习语义特征,进而提高深度估计的精度. 这些方法仍然依赖跳跃连接实现不同层次特征融合,导致得到的特征细节丢失且表示能力弱,因此很难实现对复杂纹理区域的深度估计.

1.3 注意力机制

注意力机制是将注意力转移到重要特征的动态过程. Vaswani 等^[24]首次使用自注意力机制提取全局依赖关系,并将其应用于机器翻译. 之后,通道注意力和空间注意力等更多的注意力机制被提出,并应用于计算机视觉任务中. 对于自监督单目深度估计, Jung 等^[5]使用多头注意力模块将深度特征和语义特征融合; Yan 等^[25]将通道注意力应用于结构感知和跳跃连接,捕获更多场景的上下文信息,增强特征表示; Ma 等^[26]结合局部注意力和跨任务注意力同时增强深度估计任务和语义分割任务,再利用语义分割网络得到的语义信息细化深度特征. 然而,这些方法只是在原有网络的基础上增加了一些注意力模块,深度网络的主体仍然是传统的局部化卷积网络. 与之前的工作不同,本文在整个 VAMF 网络中均采用视觉注意融合机制,本文方法能够自适应地提取重要特征,并在各个阶段去除噪声干扰,从而获得更准确的深度预测结果.

2 本文方法

2.1 自监督单目深度估计方法

在自监督单目深度估计模型训练过程中,通过最小化目标图像与源图像和预测目标深度图合成图像之间的差异优化模型. 深度估计框架采用联合深度估计网络和位姿网络的方式进行自监督训练,其中位姿网络仅用于预测相邻帧间的位姿变化,深度信息的准确性最大限度地受到深度估计网络的影响. 文献[5]将语义特征作为辅助信息细化深度特征,用于提高深度估计网络的性能. 但深度估计网络仍采用传统编解码器架构,该架构在编解码的过程中会丢失大量的细节信息,无法充分地利用高层次特征和低层次特征. 本文在文献[5]的基础上替换了深度估计网络的整体架构,提出了基于视觉注意的多尺度特征融合单目深度估计网络框架 VAMF. VAMF 包括使用多尺度图像特征聚合(multi-scale image feature integration, MFI)的编码器模块、中间注意力模块(intermediate attention module, IAM)以及由外部注意力特征融合模块(external attention feature fusion, EAFF)组成的解码器模块,如图 1 所示,整体网络框架需要联合深度估计网络、姿态估计网络和语义分割网络进行训练.

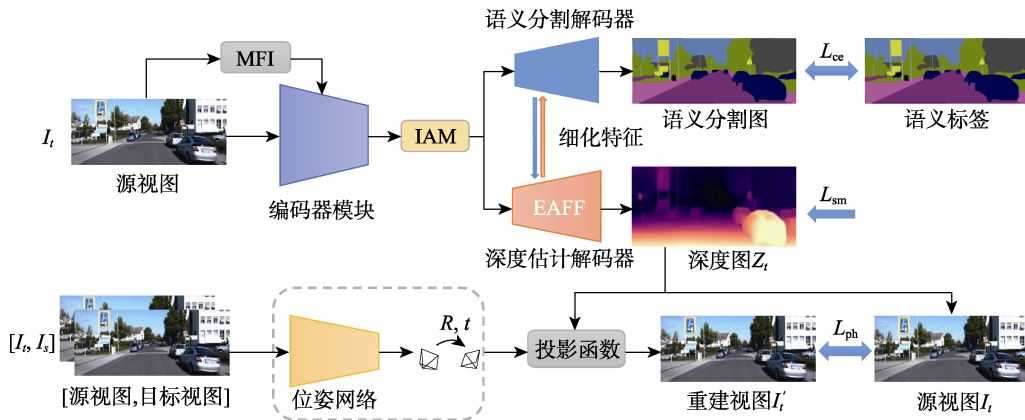


图 1 整体网络框架

在使用视频序列训练的过程中,需要从相同的单目图像序列中选取源帧的目标帧,一般选取源帧 I_t 的相邻帧 $I_s (s \in \{t-1, t+1\})$ 作为其目标帧. 由于不清楚源帧和目标帧之间的位置关系,首先将 I_t 和 I_s 通过位姿估计网络计算出六自由度相对位姿信息 $T_{t \rightarrow s}$. 在已知相机内参的情况下,可以得到投影后的像素坐标,并通过深度估计网络得到的深度图 Z_t 构建出原视图 I_t 的重建视图 $I'_t = W_t(I_s, p')$. 其中,

$p' = K T_{t \rightarrow s} Z_t K^{-1} p$, K 是相机内参矩阵, $K \in \mathbb{R}^{3 \times 3}$; p 是像素的齐次坐标; p' 是 p 经过 $T_{t \rightarrow s}$ 变换后的坐标; $W(\cdot)$ 是一个可微双线性采样器^[27],用于获取 I_s 中 p' 处的像素,并在 I_t 中的 p 处制定线性插值像素.

理想情况下,完全收敛时网络的重建视图应与原视图是一致的,利用结构相似性指标(structural similarity, SSIM)^[28],结合 L_1 损失作为光度一致性损

失^[20] $L_{ph} = \alpha \frac{1 - \text{SSIM}(I_t, I'_t)}{2} + (1 - \alpha) |I_t - I'_t|$ 约束重建视图。

此外, 为了处理移动物体等问题, 本文应用自动掩码方法^[4]过滤掉在视频帧间保持运动的物体, 使用最小重投影^[4]处理遮挡问题, 同时采用边缘平滑损失^[20] $L_{sm} = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|}$ 对低梯度区域的深度进行正则化, 解决视差的不连续性问题。其中, ∂_x 和 ∂_y 分别为水平方向和垂直方向的梯度; $d_t^* = d_t / \bar{d}$ 表示平均归一化深度, d_t 表示绝对深度误差, $\bar{d}$ 表示真实深度。

在训练语义分割网络时, 本文使用已有的分割模型^[29]生成的伪标签进行训练, 并使用交叉熵损失^[29]计算代价。因此, 总损失函数为

$$L = L_{ph} + \beta L_{sm} + \gamma L_{ce};$$

其中, β 和 γ 分别为边缘平滑损失 L_{sm} 和交叉熵损失 L_{ce} 的权重系数。

2.2 VAMF 网络结构

现有的许多深度估计网络都是建立在 ResNet 作为主干网络的基础上, ResNet 可以将输入的图像编码成特征图。其中, 低层特征分辨率更高, 包含更多位置、细节信息, 但是由于经过的卷积更少, 其语义性更低, 噪声更多; 高层特征具有更强的语义信息, 但是分辨率较低, 对细节的感知能力较差。因此, 实现自适应融合高低层特征就显得极为重要。本文提出了融合新型深度网络 VAMF, 如图 2 所示, 采用改进的编码器-解码器结构, 将单幅 RGB 图像作为输入, 输出多尺度深度图。不同于以往在网络内部融合特征的方式, 本文分别对输入编码器和输出解码器的特征进行外部融合, 一方面, 外部融合避免了加深网络而导致的网络退化和梯度爆炸、消失等问题; 另一方面, 在减少参

数量的同时, 保留了大量的特征信息, 提高了网络的预测精度。此外, VAMF 使用 IAM 连接编码器和解码器, IAM 由基于查询无关的简化位置注意力模块(simplified position attention module, SPAM)和通道注意力模块(channel attention module, CAM)并行组成。

2.2.1 基于多尺度特征融合的编解码器模块

(1) 编码器模块, 本文采用 ResNet18 作为编码器的主干网络。从输入图像 $I_t \in \mathbb{R}^{3 \times H \times W}$ 中提取多尺度特征; 去除了 ResNet18 的平均池化层和全连接层, 得到通道数分别为 $[64, 64, 128, 256, 512]$ 、分辨率为 $\left[\frac{H}{2} \times \frac{W}{2}, \frac{H}{4} \times \frac{W}{4}, \frac{H}{8} \times \frac{W}{8}, \frac{H}{16} \times \frac{W}{16}, \frac{H}{32} \times \frac{W}{32}\right]$ 的中间特征 $[C_1, C_2, C_3, C_4, C_5]$ 。考虑丰富的空间信息对深度估计的重要性, 为了更好地利用多尺度图像信息, 本文在编码器的输入阶段对源图像预处理。首先, 将源图像 I_t 多次下采样得到子图像

I'_t , 其分辨率为源图像 I_t 分辨率的 $\frac{1}{2^r}$ ($r=1, 2, 3, 4$);

其中 I'_1 和 I'_2 保持着相同的通道数, 其他子图像随着 r 的增加, 通道数是前一子图像的 2 倍。然后将编码器上一层的输出和子图像 I'_t 通过逐像素求和输入到下一层, 编码器将输出 5 个不同尺度的特征图 C'_i ($i=1, 2, 3, 4, 5$)。最后将这些特征图通过跳跃连接送入到解码器中, 并将分辨率最低的特征图作为 IAM 的输入, IAM 采用并行注意力模块自适应地选择判别性特征作为解码器的初始输入。

(2) 解码器模块。解码器接收从编码器提取的特征图 C'_i , 并输出深度特征和视差图作为场景深度表示。He 等^[30]在解码器模块采用 HRDecoder, 通过继承编码器的多尺度特征, 再逐步融合低尺度特征实现高效的特征融合, 但这种方法会加深网

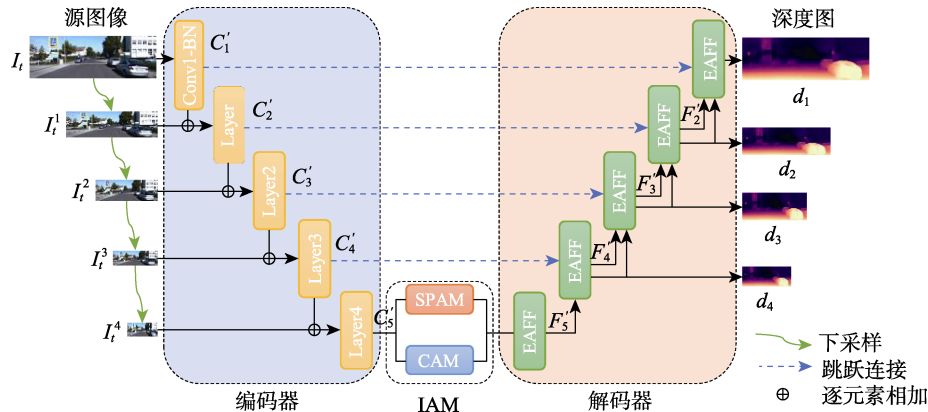


图 2 VAMF 网络结构

络的层数, 计算量也会相应增大. 因此, 本文提出一种有效的轻量级外部注意力特征聚合模块 EAFF, 在不加深网络的同时, 实现对解码器输出的深度特征和视差图的有效融合. 解码器由 5 个连续的 EAFF 模块组成, 第 i 个 EAFF 模块将输出深度特征 F_i^d 和视差图 d_i , 同时需要来自编码器的内部特征 C_i' 和上一个 EAFF 模块的特征 F_{i+1}^d , d_{i+1} 作为输入. 需要注意的是, 最底层的 EAFF 模块仅需来自经过 IAM 的编码器内部特征图 C_5' 作为输入, 且仅输出深度特征 F_5^d . 如图 3 所示, F_{i+1}^d 首先连续通过 2 个带有线性指数单元(exponential linear unit, ELU)的卷积层直接输出 F_i^d , 同时 d_{i+1} 相应的通过一个卷积层和一个采样因子为 2 的上采样层恢复分辨率. 然后使用注意力模块融合编码器内部特征 C_i' 、直接输出的低级特征 F_i^d 和来自上采样操作的高级特征. 对于注意力模块, 通过全局平均池化层、带有修正线性单元(rectified linear unit, ReLU)的全连接层和使用 Sigmoid 函数的全连接层处理串联特征, 以生成通道注意力图, 注意力模块的最终输出 d_i 由注意力图和原始特征元素相乘得到.

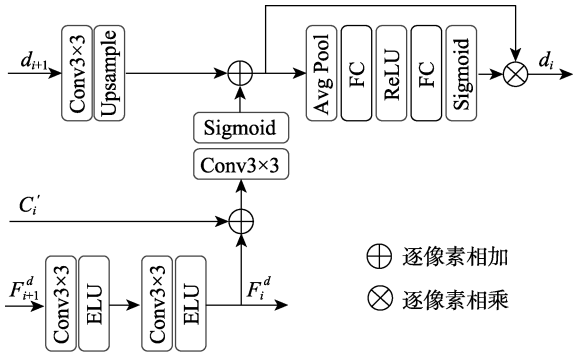


图 3 EAFF 模块

2.2.2 IAM

给定一幅用于深度估计的图像, 每个物体之间的尺寸、光照和视图都是不同的, 由于卷积操作会导致局部感受野, 因此网络往往使用的都是局部特征. 虽然主流的编码器-解码器结构可以融合中高级语义特征, 但它不能利用全局视图中的物体或物体之间的关系, 这会影响网络在复杂纹理区域的深度估计效果. 为了解决上述问题, Fu 等^[31]引入一种自注意力机制分别捕获空间维度和通道维度中的特征依赖关系, 通过在残差网络生成的局部特征中捕获全局上下文信息. 但这种注意力机制会导致模型计算量成倍数增大, 推理速度极

慢. 本文提出的 IAM 去除了无关紧要的查询点计算, 在减少参数量的同时自适应地整合局部特征和全局依赖, 为像素级预测获得更好的特征表示. IAM 由基于查询无关的 SPAM 和 CAM 并行组成, 如图 4 所示.

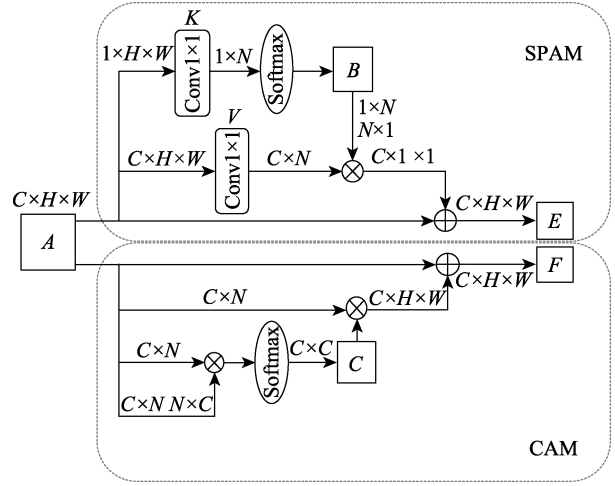


图 4 IAM

(1) SPAM. 位置注意力模块是通过计算每一个查询点的位置与图中其他像素之间的关系来提取该点的特征, 然而, Cao 等^[32]的研究表明, 在深层次网络中不同查询点对应的注意力图几乎一样, 因此, 不需要单独计算每一个查询点的特征图. 通过计算一个查询点的全局注意力图, 并为所有查询位置共享这个全局注意力图简化整个网络. 给定编码器输出的特征图 $A = C_5' \in \mathbb{R}^{C \times H \times W}$, 首先将其通过卷积层生成 2 个新的特征图 $K \in \mathbb{R}^{1 \times H \times W}$ 和 $V \in \mathbb{R}^{C \times H \times W}$. 将 K 重塑为 $\mathbb{R}^{1 \times N}$ ($N = H \times W$) 后应用 Softmax 层计算空间注意力映射

$$B_{ij} = \exp(K_i A_j) / \sum_{j=1}^N \exp(K_i A_j).$$

其中, B_{ij} 衡量第 j 个位置对第 i 个位置的影响. 2 个位置更相似的特征表示有助于加强它们之间的相关性. 同时, 将特征图 V 重塑为 $\mathbb{R}^{C \times N}$, 然后与 B 的转置进行矩阵乘法. 最后, 将其乘以尺度参数 λ , 并与特征 A 执行逐元素求和操作, 获得最终输出

$$E_i = A_i + \lambda V \sum_{j=1}^N B_{ij} A_j.$$

其中, 初始化 $\lambda = 0$, 并逐渐学习分配更多的权重. 每个位置得到的特征 E 都是所有位置特征和原始特征的加权和, 因此可以从全局的角度自适应地

集合其他位置上的相似特征。

(2) CAM. 在深度估计中, 高层特征图的每个通道映射都可以看作是一个区域特定的响应, 并且不同的区域响应相互关联, 通过利用通道图之间的相互依赖, 可以捕获不同区域的相对深度信息, 显著地提高特定场景的特征表示。因此, 本文使用了一个 CAM 建模通道之间的相关性。首先将输入特征 $A_1 \in \mathbb{R}^{C \times H \times W}$ 重塑为 $A_2 \in \mathbb{R}^{C \times N}$, 在 A_1 和 A_2 的转置之间进行矩阵乘法, 即

$$C_{ij} = \exp(A_i A_j) / \sum_{j=1}^N \exp(A_i A_j);$$

然后应用 Softmax 层得到通道注意力图 $C \in \mathbb{R}^{C \times C}$ 。其中, C_{ij} 用于度量第 j 个通道对第 i 个通道的影响。此外, 在 C 和 A 的转置之间执行矩阵乘法, 并将其结果重塑为 $\mathbb{R}^{C \times H \times W}$ 。最后, 在 A 和结果之间执行逐元素求和操作, 以获得最终输出

$$F_i = A_i + \eta \sum_{j=1}^N D_{ij} A_j,$$

其中, 初始化 $\eta = 0$, 并逐渐学习分配更多的权重。每个通道的最终特征是所有通道的特征与原始特征的加权和。通过捕获特征图之间的长程依赖关系, 使得各个通道之间产生全局的关联, 获得更好的场景特征表示。

为了进一步获得全局依赖关系的特征, 将 2 个模块的输出通过卷积层变换后进行逐元素求和实现注意力特征融合, 从而获得丰富上下文信息的聚合特征。

3 实验

3.1 数据集

KITTI 数据集^[7]是评测多种自动驾驶任务常用的数据集。KITTI stereo 是 KITTI 数据集中的立体匹配数据集, 使用 KITTI stereo 数据集的 Eigen split^[9]评估本文方法。在训练前对图像进行预处理以移除静态帧^[2], 因此, 将有 39 810 幅像用于训练, 4 424 幅图像用于验证, 697 幅图像用于测试。这些图像的分辨率约为 $375 \times 1\,242$ 像素, 在训练和评估时, 所有图像大小都调整到 192×640 像素; 在测试时, 采用原始 RGB 图像的全分辨率进行评估, 但深度网络仅输出分辨率为 192×640 像素的深度图。

Cityscapes 数据集^[8]面向自动驾驶场景, 包含

50 个不同城市街道场景的立体图像序列, 图像分辨率为 $1\,024 \times 2\,048$ 像素, 共 19 个类别。主要用于研究场景语义与实例分割任务, 本文选用该数据集的图像用于测试可视化结果, 以评估本文方法的泛化性。

3.2 实施细节

本节实现了本文的深度估计网络, 并与已有方法^[4-5,20,25,33-43]进行对比实验。本文采用 Pytorch 作为深度学习框架, 并在 NVIDIA GeForce RTX 2080 Ti 上实现本文方法。在所有的实验中, 使用 VAMF 作为深度估计网络的训练架构, 输入/输出分辨率为 192×640 像素; 对于位姿估计, 使用 ResNet18 预测相邻帧的六自由度相对位姿; 位姿网络采用与深度网络和语义分割网络相同的输入分辨率。此外, 通过在 ImageNet^[44]上预训练的权重进行初始化深度估计网络、语义分割网络和位姿网络的编码器。在训练过程中, 设置训练学习批次 (batch size) 为 12; epoch 为 20; 使用 Adam 优化器进行优化; 设置初始学习率为 0.000 15, 在训练 10 和 15 个 epoch 后, 学习率分别下降 10% 和 50%。根据之前的工作^[4], 设定损失参数为 $\alpha = 0.85$, $\beta = 0.001$, $\gamma = 0.3$ 。

3.3 评价指标

对于单目深度估计问题, 为了评价不同方法的单目深度估计效果, 通常采用以下 4 项误差指标以及 3 项准确率指标作为评测指标, 包括绝对值相对误差 (absolute relative error, AbsRel), 平方相对误差 (square relative error, SqRel), 均方误差 (root mean squared error, RMS), 对数均方误差 (root mean squared error in logarithmic space, RMSln) 以及 3 项不同阈值的准确率。各指标分别定义为

$$\begin{aligned} \text{Abs Rel} &= \frac{1}{|N|} \sum_{d \in N} \frac{|d - d^*|}{d^*}, \\ \text{SqRel} &= \frac{1}{|N|} \sum_{d \in N} \frac{\|d - d^*\|^2}{d^*}, \\ \text{RMS} &= \sqrt{\frac{1}{|N|} \sum_{d \in N} (d - d^*)^2}, \\ \text{RMSln} &= \sqrt{\frac{1}{|N|} \sum_{d \in N} \|\ln(d) - \ln(d^*)\|^2}, \\ \delta &= \max \left(\frac{d}{d^*}, \frac{d^*}{d} \right) < t, \quad t = 1.25, 1.25^2, 1.25^3. \end{aligned}$$

其中, d 和 d^* 分别表示地面真值和预测的深度; N 表示单幅图像中所有预测深度值的集合; $|\cdot|$ 返回每个输入集合中元素的个数。AbsRel, SqRel,

RMSE, RMSEln 的值越小, 表示深度估计效果越好; δ 是衡量方法准确率的指标, 值越大, 表示深度估计效果越好.

3.4 实验结果

3.4.1 定量分析

为评估深度网络的性能, 本文与目前表现良好的单目深度估计方法进行了对比, 所有的方法都以与训练时相同的分辨率进行测试, 结果如表 1 所示. 其中, M 表示在单目视频中训练, Se 表示使用语义标签训练. 通过比较所有的指标, 本文方法比其他方法有更准确的深度预测结果. 对于单目深度估计中最重要的精度指标, $\delta < 1.25$ 达到了 0.891, $\delta < 1.25^2$ 达到了 0.965. 对比次优效果, RMS 指标降低了 0.061. 文献[5]在 SqRel 指标上取得了最好的效果, 主要原因在于该方法采用三元组损失优化跨语义边界像素的距离, 使其在纹理单一的物体边界区域产生更具辨别力的特征, 当检测的物体背景比较复杂时, 该方法会造成较差的效果. 此外, 本文方法也使用了 ResNet50 作为主干

网络与其他方法进行比较, 以验证提出网络的通用性和鲁棒性, 其中最重要的精度指标 $\delta < 1.25$ 达到了 0.896, $\delta < 1.25^2$ 达到了 0.967; 与次优方法相比, RMSln 指标降低了 0.002, AbsRel 和 RMS 也取得了与之前方法相当甚至更优的性能.

3.4.2 消融实验

为了评估本文所提模块的有效性, 在 KITTI 数据集^[7]上进行了消融实验, 所有的消融实验编码器主干网络均采用 ResNet18, 文献[5]作为基准方法. 对 MFI, EAff, IAM 这 3 种不同方法进行了消融实验, 结果如表 2 所示, 其中 EAff 将精度从 0.886 提升到 0.889, 添加 IAM 后精度得到进一步的提升, 因为 IAM 使得局部信息和全局信息得以关联, 值得注意的是, MFI 的加入在没有降低深度估计精度的同时使得 4 项误差指标达到最优.

在表 3 中, 本文方法在使用 MFI 以及 EAff 模块的情况下, 对 IAM 模块的 SPAM 和 CAM 进行消融实验. 结果表明, 仅使用 SPAM 的情况下 RMS 和 RMSln 指标达到最好, 但最重要的 $\delta < 1.25$ 指标

表 1 不同方法在 KITTI 数据集上的定量结果

骨干网络	方法	训练方法	AbsRel ↓	SqRel ↓	RMS ↓	RMSln ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
ResNet18	Godard 等 ^[20]	M	0.114	0.864	4.817	0.192	0.877	0.959	0.981
	Zou 等 ^[33]	M	0.115	0.871	4.778	0.191	0.874	0.963	0.984
	Lu 等 ^[34]	M	0.120	0.896	4.869	0.198	0.868	0.957	0.981
	Klingner 等 ^[35]	M+Se	0.113	0.835	4.693	0.191	0.879	0.961	0.981
	Lee 等 ^[36]	M+Se	0.112	0.777	4.772	0.191	0.872	0.959	0.982
	Poggi 等 ^[37]	M	0.111	0.863	4.756	0.188	0.881	0.961	0.982
	Patil 等 ^[38]	M	0.111	0.821	4.650	0.187	0.883	0.961	0.982
	Choi 等 ^[39]	M+Se	0.112	0.788	4.582	0.187	0.878	0.963	<u>0.983</u>
	赵霖等 ^[40]	M	0.108	0.756	4.677	0.186	0.882	0.961	<u>0.983</u>
	Wang 等 ^[41]	M	<u>0.109</u>	0.792	4.632	0.185	0.884	0.962	<u>0.983</u>
	Yan 等 ^[25]	M	0.110	0.812	4.686	0.187	0.882	0.962	<u>0.983</u>
	Jung 等 ^[5]	M+Se	0.105	0.722	<u>4.547</u>	<u>0.182</u>	<u>0.886</u>	<u>0.964</u>	0.984
ResNet50	本文	M+Se	0.105	<u>0.750</u>	4.486	0.181	0.891	0.965	0.984
	Klingner 等 ^[35]	M+Se	0.112	0.833	4.688	0.190	0.884	0.961	0.981
	Godard 等 ^[4]	M	0.110	0.831	4.642	0.187	0.883	0.962	0.982
	Lu 等 ^[34]	M	0.117	0.900	4.826	0.196	0.873	0.967	0.980
	Shu 等 ^[42]	M	<u>0.104</u>	0.792	4.481	0.179	<u>0.893</u>	0.965	<u>0.984</u>
	Yan 等 ^[25]	M	0.105	0.769	4.535	0.181	0.892	0.964	0.983
	Jung 等 ^[5]	M+Se	0.102	0.675	<u>4.393</u>	<u>0.178</u>	<u>0.893</u>	0.966	<u>0.984</u>
PackNet	本文	M+Se	0.102	<u>0.705</u>	4.390	0.176	0.896	0.967	0.985
	Guizilini 等 ^[43]	M	0.111	0.785	4.601	0.189	0.878	0.960	0.984

注. 粗体表示最优值, 下划线表示次优值. AbsRel, SqRel, RMS, RMSln 值越小效果越好, δ 值越大效果越好.

表 2 KITTI 数据集上的消融实验结果

方法	MFI	EAFf	IAM	AbsRel ↓	SqRel ↓	RMS ↓	RMSln ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Baseline				0.105	0.722	4.547	0.182	0.886	0.964	0.984
VAMF	√	√		0.107	0.762	4.501	0.181	0.889	0.964	0.984
	√		√	0.106	0.738	4.496	0.181	0.887	0.964	0.984
	√	√	√	0.105	0.750	4.486	0.181	0.891	0.965	0.984
		√	√	0.106	0.754	4.498	0.181	0.891	0.965	0.984

注. 粗体表示最优值. AbsRel, SqRel, RMS, RMSln 值越小效果越好, δ 值越大效果越好.

表 3 KITTI 数据集上不同注意力模块的对比结果

SPAM	CAM	AbsRel ↓	SqRel ↓	RMS ↓	RMSln ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
		0.107	0.762	4.501	0.181	0.889	0.965	0.984
√		0.107	0.759	4.486	0.180	0.888	0.964	0.984
	√	0.106	0.764	4.496	0.182	0.890	0.965	0.984
√	√	0.105	0.750	4.486	0.181	0.891	0.965	0.984

注. 粗体表示最优值. AbsRel, SqRel, RMS, RMSln 值越小效果越好, δ 值越大效果越好.

精度较低, 而采用 2 个注意力模块并联的方式, 以 RMSln 指标较小的损失代价使得其他所有指标取得最好的效果.

3.4.3 定性分析

图 5 中显示了在 KITTI 数据集^[7]中的定性分析结果, 将本文方法与 Jung 等^[5], Godard 等^[4]和 Guizilini 等^[43]的结果进行对比. 可以发现, 本文方法对复杂纹理区域的物体取得了更好的预测效果,

例如, 背景是树木和房子的路标以及电线杆; 并且可以更详细地描绘物体的轮廓(第 3 行和第 8 行的路标). 在图 6 中, 更加直观地展现了深度估计细节, 文献[5]的结果受背景的影响较大, 将背景中的房屋错误地估计成路标的一部分, 文献[4]的结果直接丢失了远景区域的电线杆和路标, 而本文方法得到了很好的预测效果.

图 7 显示了在 Cityscapes 数据集^[8]中的定性分

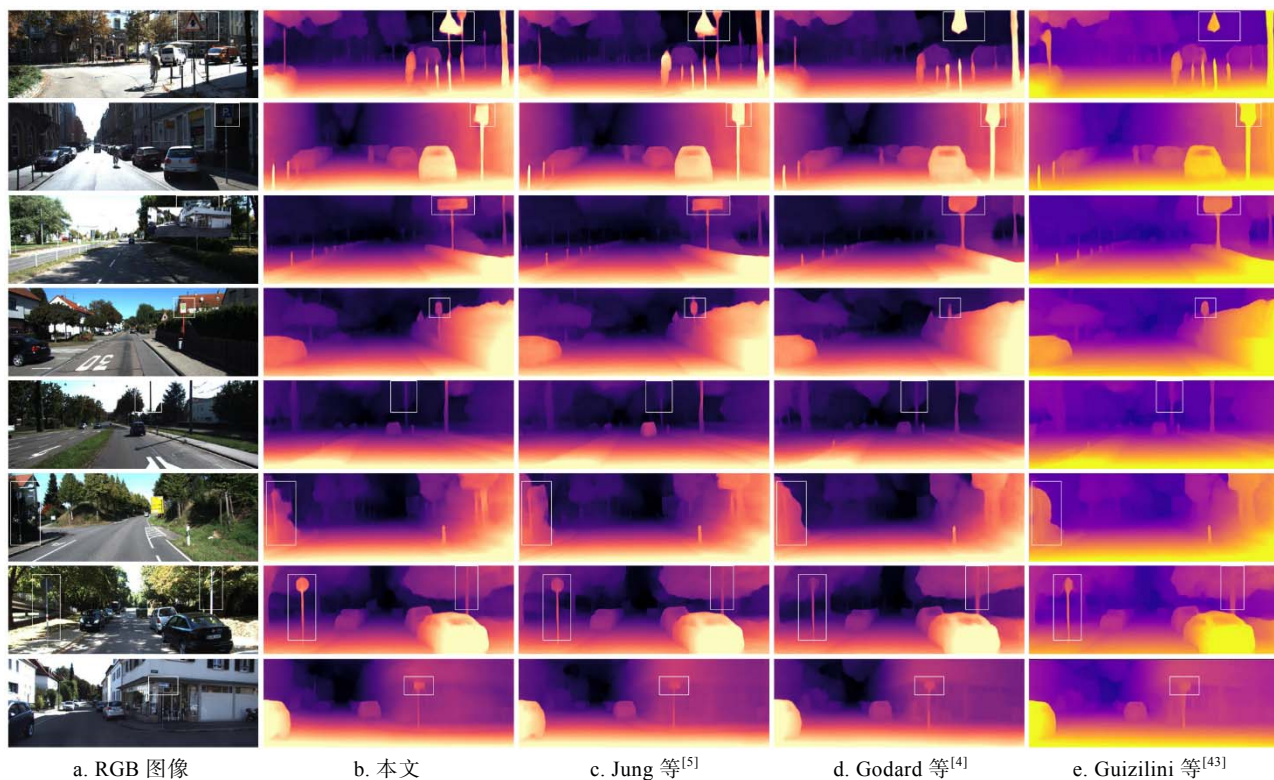


图 5 不同方法在 KITTI 数据集上可视化结果

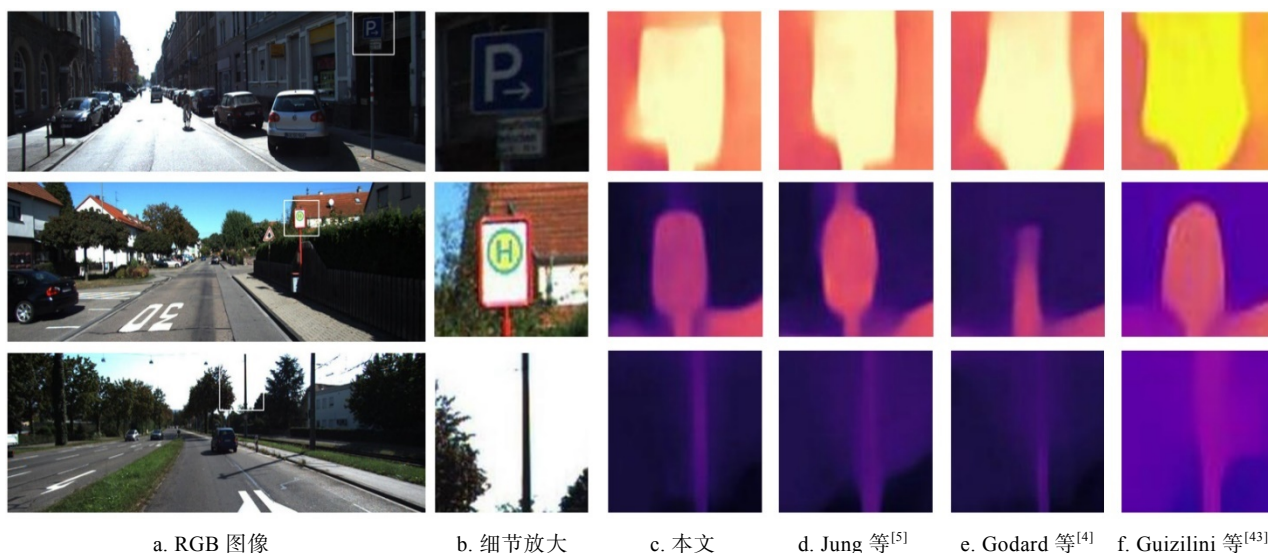


图 6 不同方法在 KITTI 数据集上深度估计细节对比

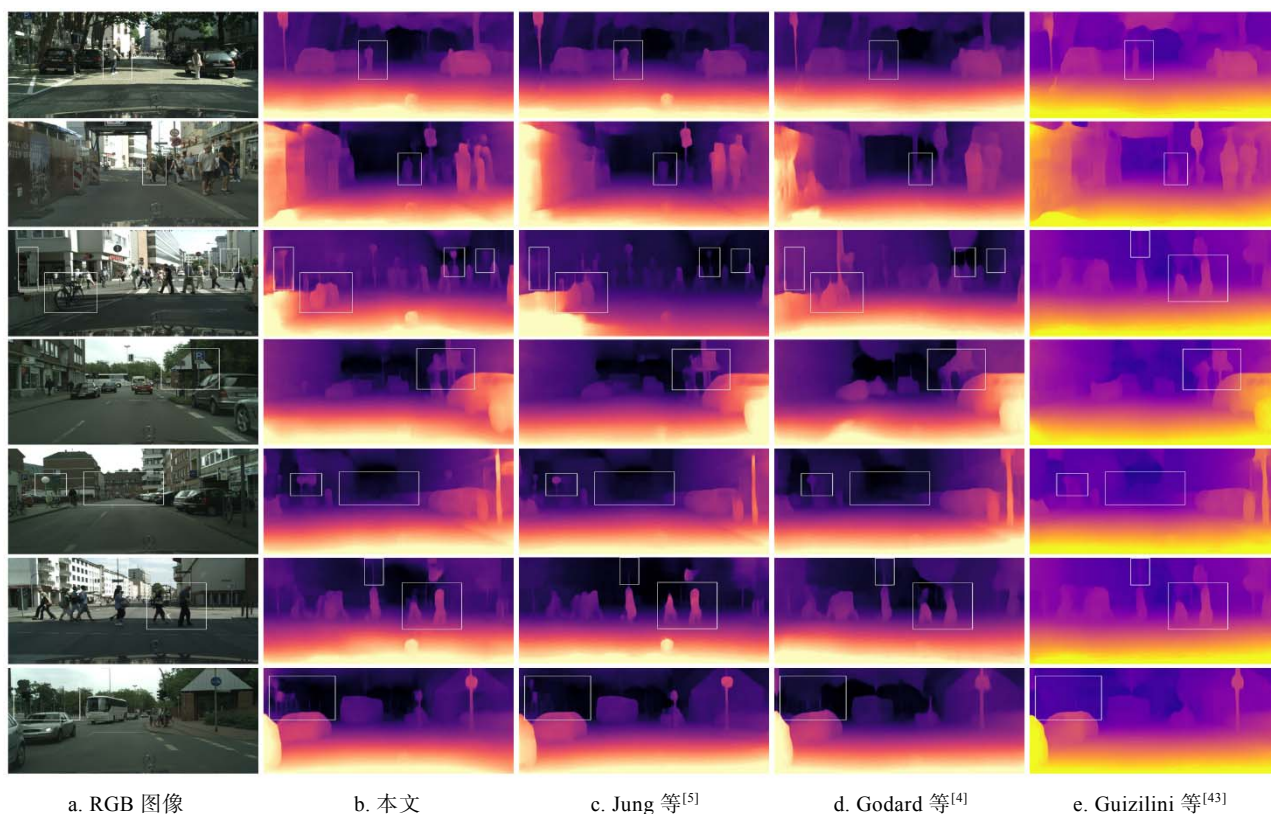


图 7 不同方法在 Cityscapes 数据集上可视化结果

析结果. 为了评估本文方法的通用性, 直接使用 Cityscapes 数据集进行测试, 将本文方法与其他结果进行对比. 本文方法无论是在处理近景区域(第1行、第2行和第6行的行人), 还是远景区域(第5行的房屋、第7行的电线杆), 效果都要优于其他方法. 具体的深度预测细节如图8所示, 其中文献

[4-5,43]的方法在预测远景区域的房屋和路灯时, 会丢失大量的细节信息, 不能比较完整地预测物体的深度. 在面对复杂纹理区域时, 这些方法受到背景信息的干扰, 难以将自行车的轮廓与背景分割, 产生了错误的深度估计. 与之对比, 本文方法更好地保持了物体的完整性和深度的一致性.

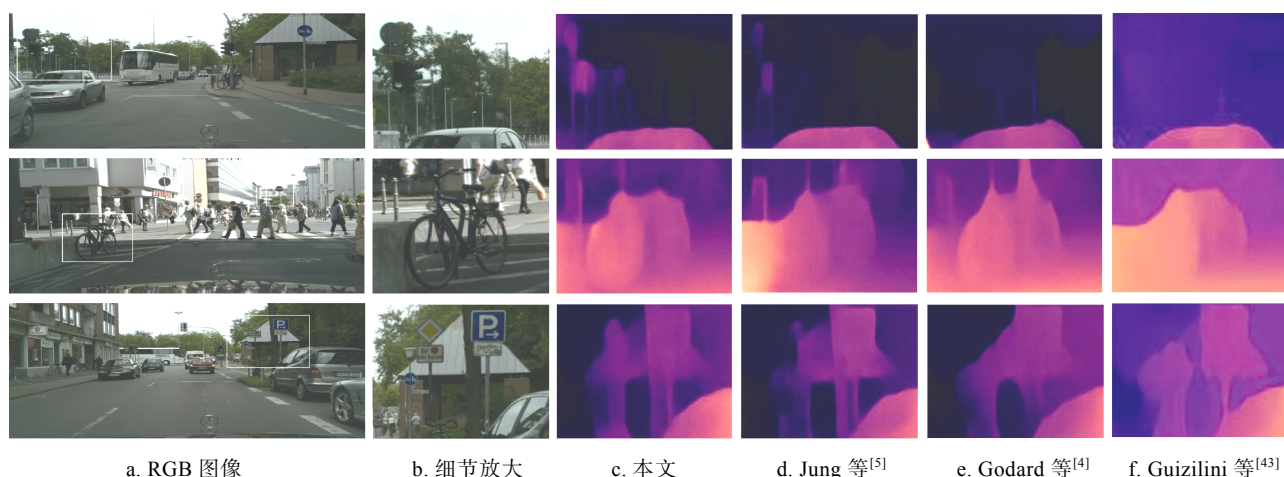


图 8 不同方法在 Cityscapes 数据集上深度估计细节对比

4 结 语

本文提出了一种基于视觉注意的自监督单目深度估计方法, 通过在网络的各个阶段利用视觉注意机制, 本文方法被赋予提取重要上下文信息和有效融合多层次特征的强大能力, 以解决深度估计在复杂纹理区域和远景区域的不适应问题. 在 KITTI 数据集上进行了实验并且取得了很好的效果. 为进一步验证本文方法的鲁棒性, 在 Cityscapes 数据集上进行了测试和比较, 结果表明, 本文方法在深度估计方向取得了不错的性能, 在面对复杂纹理区域和远景区域时能得到更好的预测结果, 并且在未知变化的环境中具有良好的泛化性. 未来, 计划进一步提高深度估计网络的泛化能力和解释能力.

参考文献(References):

- [1] Bi Tianteng, Liu Yue, Weng Dongdong, *et al.* Survey on supervised learning based depth estimation from a single image[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2018, 30(8): 1383-1393(in Chinese)
(毕天腾, 刘越, 翁冬冬, 等. 基于监督学习的单幅图像深度估计综述[J]. *计算机辅助设计与图形学学报*, 2018, 30(8): 1383-1393)
- [2] Zhou T H, Brown M, Snavely N, *et al.* Unsupervised learning of depth and ego-motion from video[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6612-6619
- [3] Yin Z C, Shi J P. GeoNet: unsupervised learning of dense depth, optical flow and camera pose[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 1983-1992
- [4] Godard C, Mac Aodha O, Firman M, *et al.* Digging into self-supervised monocular depth estimation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 3827-3837
- [5] Jung H, Park E, Yoo S. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 12622-12632
- [6] He K M, Zhang X Y, Ren S Q, *et al.* Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778
- [7] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 3354-3361
- [8] Cordts M, Omran M, Ramos S, *et al.* The Cityscapes dataset for semantic urban scene understanding[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 3213-3223
- [9] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C] //Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2366-2374
- [10] Liu F Y, Shen C H, Lin G S. Deep convolutional neural fields for depth estimation from a single image[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 5162-5170
- [11] Xu D, Ricci E, Ouyang W L, *et al.* Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 161-169
- [12] Xu D, Wang W, Tang H, *et al.* Structured attention guided convolutional neural fields for monocular depth estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 3917-3925

- [13] Zheng C X, Cham T J, Cai J F. T²Net: synthetic-to-realistic translation for solving single-image depth estimation tasks[C] // Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 798-814
- [14] Zhao S S, Fu H, Gong M M, *et al.* Geometry-aware symmetric domain adaptation for monocular depth estimation[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 9780-9790
- [15] Kumar A C S, Bhandarkar S M, Prasad M. DepthNet: a recurrent neural network architecture for monocular depth prediction[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 396-404
- [16] Dharmasiri T, Spek A, Drummond T. Joint prediction of depths, normals and surface curvature from RGB Images using CNNs[C] // Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2017: 1505-1512
- [17] Fu H, Gong M M, Wang C H, *et al.* Deep ordinal regression network for monocular depth estimation[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2002-2011
- [18] Bhat S F, Alhashim I, Wonka P. AdaBins: depth estimation using adaptive bins[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 4008-4017
- [19] Garg R, VijayKumar B G, Carneiro G, *et al.* Unsupervised CNN for single view depth estimation: geometry to the rescue[C] // Proceedings of the 14th European Conference on Computer Vision. Heidelberg: Springer, 2016: 740-756
- [20] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6602-6611
- [21] Wang L J, Wang Y F, Wang L Z, *et al.* Can scale-consistent monocular depth be learned in a self-supervised scale-invariant manner?[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 12707-12716
- [22] Zou Y L, Luo Z L, Huang J B. DF-Net: unsupervised joint learning of depth and flow using cross-task consistency[C] // Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 38-55
- [23] Chen P Y, Liu A H, Liu Y C, *et al.* Towards scene understanding: unsupervised monocular depth estimation with semantic-aware representation[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 2619-2627
- [24] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need[C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 6000-6010
- [25] Yan J X, Zhao H, Bu P H, *et al.* Channel-wise attention-based network for self-supervised monocular depth estimation[C] // Proceedings of the International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2021: 464-473
- [26] Ma J Y, Lei X Y, Liu N, *et al.* Towards comprehensive representation enhancement in semantics-guided self-supervised monocular depth estimation[C] // Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 304-321
- [27] Jaderberg M, Simonyan K, Zisserman A, *et al.* Spatial Transformer networks[C] // Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 2017-2025
- [28] Wang Z, Bovik A C, Sheikh H R, *et al.* Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612
- [29] Zhu Y, Sapra K, Reda F A, *et al.* Improving semantic segmentation via video propagation and label relaxation[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 8848-8857
- [30] He M, Hui L, Bian Y K, *et al.* RA-Depth: resolution adaptive self-supervised monocular depth estimation[C] // Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 565-581
- [31] Fu J, Liu J, Tian H J, *et al.* Dual attention network for scene segmentation[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3141-3149
- [32] Cao Y, Xu J R, Lin S, *et al.* GCNet: non-local networks meet squeeze-excitation networks and beyond[C] // Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 1971-1980
- [33] Zou Y, Ji P, Tran Q H, *et al.* Learning monocular visual odometry via self-supervised long-term modeling[C] // Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 710-727
- [34] Lu X, Sun H R, Wang X L, *et al.* Semantically guided self-supervised monocular depth estimation[J]. IET Image Process, 2022, 16(5): 1293-1304.
- [35] Klingner M, Termöhlen J A, Mikołajczyk J, *et al.* Self-supervised monocular depth estimation: solving the dynamic object problem by semantic guidance[C] // Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 582-600
- [36] Lee S, Im S, Lin S, *et al.* Learning monocular depth in dynamic scenes via instance-aware projection consistency[C] // Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 1863-1872
- [37] Poggi M, Aleotti F, Tosi F, *et al.* On the uncertainty of self-supervised monocular depth estimation[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 3224-3234
- [38] Patil V, Van Gansbeke W, Dai D X, *et al.* Don't forget the past: recurrent depth estimation from monocular video[J]. IEEE Robotics and Automation Letters, 2020, 5(4): 6813-6820
- [39] Choi J, Jung D, Lee D, *et al.* SafeNet: self-supervised monocular depth estimation with semantic-aware feature extrac-

- tion[OL]. [2022-12-29]. <https://arxiv.org/abs/2010.02893>
- [40] Zhao Lin, Zhao Yan, Jin Jie. A self-supervised monocular depth estimation algorithm based on local attention and iterative pose refinement[J]. *Journal of Signal Processing*, 2022, 38(5): 1088-1097(in Chinese)
(赵霖, 赵艳, 靳捷. 基于局部注意力和位姿迭代优化的自监督单目深度估计算法[J]. *信号处理*, 2022, 38(5): 1088-1097)
- [41] Wang J D, Sun K, Cheng T H, *et al.* Deep high-resolution representation learning for visual recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(10): 3349-3364
- [42] Shu C, Yu K, Duan Z X, *et al.* Feature-metric loss for self-supervised learning of depth and egomotion[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 572-588
- [43] Guizilini V, Ambruş R, Pillai S, *et al.* 3D packing for self-supervised monocular depth estimation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 2482-2491
- [44] Russakovsky O, Deng J, Su H, *et al.* ImageNet large scale visual recognition challenge[J]. *International Journal of Computer Vision*, 2015, 115(3): 211-252