基于改进密集网络和小波分解的自监督单目深度估计

程德强[1],王伟臣[1],韩成功[1],吕 晨[1],寇旗旗[2]

(1.中国矿业大学 信息与控制工程学院,江苏 徐州 221116;2.中国矿业大学 计算机科学与技术学院,江苏 徐州 221116)

摘 要:针对传统自监督单目深度估计模型对浅层特征的提取和融合不充分,容易导致小物体漏检、物体边缘模糊等问题,本文提出了一种基于改进密集网络和小波分解的自监督单目深度估计模型。该模型沿用了U-Net结构,其中,编码器采用改进的密集网络,提高了编码器的特征提取和融合能力;跳跃连接中加入细节增强模块,对编码器输出的多尺度特征进一步细化整合;解码器引入小波分解,迫使解码器更加关注高频信息,实现对图像边缘的精细化处理。实验结果表明,本文提出的深度估计模型对小物体特征的捕捉能力更强,生成的深度图边缘更清晰准确。

关键词:信号与信息处理;深度估计;自监督;密集网络;小波分解;细节增强

中图分类号:TP391.41 **文献标志码**:A **文章编号**:1671-5497(2025)05-1682-10

DOI:10.13229/j.cnki.jdxbgxb.20230820

Self-supervised monocular depth estimation based on improved densenet and wavelet decomposition

CHENG De-qiang[1],WANG Wei-chen[1],HAN Cheng-gong[1],LYU Chen[1],KOU Qi-qi[2]

(1. *School of Information and Control Engineering*, *China University of Mining and Technology*, *Xuzhou* 221116, *China*; 2. *School of Computer Science and Technology*, *China University of Mining and Technology*, *Xuzhou*, 221116 *China*)

Abstract: The traditional self-supervised monocular depth estimation model has limitations in extracting and fusing shallow features, leading to issues such as omission detection of small objects and blurring of object edges. To address these problems, a self-supervised monocular depth estimation model based on improved dense network and wavelet decomposition is proposed in this paper. The whole framework of the model follows the structure of U-net, in which the encoder adopts the improved densenet to improve the ability of feature extraction and fusion. A detail enhancement module is introduced in the skipping connections to further refine and integrate the multi-scale features generated by the encoder. The decoder

收稿日期:2023-08-04.

基金项目:国家自然科学基金项目(52204177,52304182);中央高校基本科研业务费专项资金项目(2020QN49).

作者简介:程德强(1979-),男,教授,博士.研究方向:机器视觉与模式识别,图像智能检测与信息处理.
E-mail: chengdq@cumt.edu.cn

通信作者:寇旗旗(1988-),男,讲师,博士.研究方向:机器视觉与模式识别,图像智能检测与信息处理.
E-mail: kouqiqi@cumt.edu.cn

incorporates wavelet decomposition, enabling better focus on high-frequency information during decoding to achieve precise edge refinement. Experimental results demonstrate that our method exhibits stronger capability in capturing depth estimation for small objects, resulting in clearer and more accurate edges in the generated depth map.

Key words: signal and information processing; depth estimation; self-supervision; densenet; wavelet decomposition; detail enhancement

0　引　言

二维图像的像素点仅能记录场景的颜色信息，而丢失了场景第三维度的深度信息。单目深度估计是通过输入单张RGB图像预测出其像素点深度。在大多数应用场景中，机器通常只使用一个摄像机对三维场景进行图像采集。因此，单目深度估计技术在三维场景重建中至关重要。

单目深度估计技术在许多实时三维场景重建中有着广泛的应用，比如自动驾驶、虚拟现实、增强现实等[1-3]。早期的深度估计算法大多数是有监督的[4,5]，但训练这些算法需要获取成本很高的地面真值。随着计算机算力和深度学习算法信息挖掘能力的增强，单目深度估计的自监督[6-8]算法的实现成为了可能。相比有监督[9-12]算法，自监督方法不需要地面真值进行训练，只需使用普通的单目摄像头即可实现深度估计。

近年来，自监督方法[13-15]因其不需要真实深度数据进行训练，受到了广泛关注，且取得了显著的成果。值得注意的是，Monodepth2[8]的提出显著提高了以往方法的性能，其通过引入处理遮挡的每像素最小光度损失、遮蔽静态像素的自动掩码方法，以及缓解深度纹理复制问题的多尺度深度估计策略，成为近几年自监督单目深度估计领域的主流算法。Xiang等[16]通过引入注意力机制提升深度估计网络性能。Suri[17]通过引入时间一致性损失约束姿态网络从而提高深度和自我运动的预测性能。Houssem等[18]提出一种新的时空注意力方法，对时间帧序列进行约束，获得了更高准确度的深度预测结果。最近一些方法[19,20]尝试对深度估计网络进行轻量化。Michael等[21]也尝试在Monodepth2的基础上，利用小波分解的稀疏性对深度估计的解码器进行改进，以轻量化深度估计网络。该方法关注了边缘特征，但由于编码器特征提取能力不足，仍存在预测深度信息不够精确、物体边缘模糊等问题。

目前单目深度估计方法仍存在对特征的提取和细节捕捉不够充分的问题，导致输出深度图存在小物体缺失和边缘模糊的情况。针对上述问题，本文提出了一种基于改进密集网络和小波分解的自监督深度估计模型。本文贡献如下：①编码器引入并重新设计了密集网络的卷积层，使之与深度估计网络更匹配，以提高对图像浅层特征的提取能力；②提出了细节增强模块（Detail enhancement module, DEM），通过跳跃连接强化编码器输出的多尺度浅层特征信息，实现编-解码器的多层级特征信息融合；③解码器引入小波分解与上采样特征图相结合的方法输出深度图，以细化深度边缘信息。实验表明，本文提出的深度估计模型在KITTI和CityScapes数据集上都取得了优异的深度估计性能。

1　本文方法

基于改进密集网络和小波分解的自监督深度估计网络遵循U-Net[22]结构，可以充分提取浅层特征信息并与深层特征更好地融合，有效地缓解了网络输出深度图中小物体缺失和边缘模糊的问题，其总体架构如图1所示。本文提出的深度估计网络架构由编码器、跳跃连接、解码器3个部分组成。编码器引入了密集网络（DenseNet）[23]，并提出一种改进的DenseNet-D特征提取网络替换原来的残差网络（ResNet）[24]作为编码器，加强了浅层与深层之间的特征融合，促进了浅层特征复用，从而提升了对图像特征的提取能力。为了细化编码器输出的浅层多尺度特征空间信息，在跳跃连接处设计了DEM。解码器采用双线性插值法上采样特征图，并在不同尺度特征图上预测小波高频系数图，同时采用逆小波分解的方法将低频深度图和高频系数图融合，对深度图进行迭代升级和细化，最终输出与输入尺度一致的深度图。实验表明，本文所提网络与原有基础网络相比，提高了深度估计的精度，优化了深度图的细节。

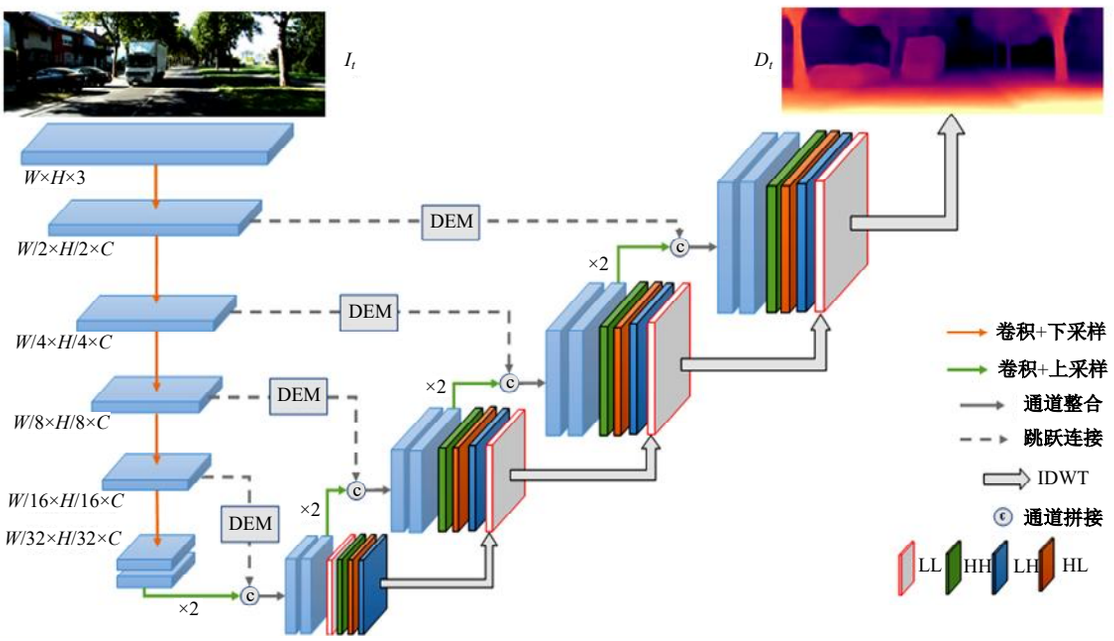


图1 本文深度估计网络架构

Fig. 1 The deep estimation network architecture in this paper

1.1 改进的DenseNet编码器

现有大多数自监督单目深度估计方法的网络编码器都基于ResNet实现。其通过让神经网络的卷积层跨层相连,从而弱化每层之间的强联系,使梯度可以直接通过恒等函数从深层传向浅层,有效缓解深层网络中存在的退化、梯度消失以及梯度爆炸等问题。但是ResNet的跨层连接只是对当前层输入特征与输出特征进行简单求和后输入下一层卷积,没有充分利用浅层和深层的特征信息,且这种特征求和过程是累加的,会阻碍网络的梯度和信息流传输。

为了确保网络中各层之间的信息流最大化,本文使用DenseNet作为深度估计网络的编码器。与ResNet不同的是:①DenseNet直接将所有层相互连接;②DenseNet不是通过特征求和组合特征,而是通过拼接通道实现特征组合,其非线性映射关系如下:

$$\begin{cases}\text{ResNet}: x_t = H(x_{t-1}) + x_{t-1} \\ \text{DenseNet}: x_t = H([x_0, x_1, x_2, \ldots, x_{t-1}])\end{cases} \tag{1}$$

式中:$H(\ )$为卷积处理;x_t为第t层特征图。

本文使用的基于DenseNet的编码器主要由卷积层、池化层、Dense Block、Transition Layer组成。与大多数深度估计网络类似,本文保留了原U-Net编码器输出的多尺度特征数。为使DenseNet网络更好地与深度估计网络架构适配,在原网络基础上移除了第3层的最大池化层,以及用于图像分类的全局池化层和全连接层。作为代替,在第4个Dense Block后加入一层下采样模块,用于整合特征并缩小特征图的尺度,最后将下采样模块的输出直接接入解码器。这种设计解决了原DenseNet网络结构在用于深度估计架构时存在的最大池化输出层导致的特征提取能力不足问题,从而在卷积层数增加的情况下充分发挥DenseNet强大的特征提取能力。本文分别使用在ImageNet数据集上预训练的DenseNet121和DenseNet169作为深度估计编码器。为加以区分,将基于DenseNet网络设计的适用于深度估计网络的编码器命名为DenseNet-D。以DenseNet121-D和DenseNet169-D为例,具体网络结构参数见表1。编码器网络特征图分别以尺度$S/2$、$S/4$、$S/8$、$S/16$、$S/32$在Convolution、Transition layer(1)、Transition layer(2)、Transition layer(3)、Downsample layer层与解码器进行连接,其中S为输入图片的尺度。

1.2 细节增强模块

传统U-Net网络架构通过直接将编码器输出的多尺度特征图与解码器进行通道拼接实现浅层与深层信息的融合。但是简单地利用跳跃连接将特征叠加缺乏对局部细节的有效处理,会导致预测深度图出现模糊伪影。为有效利用编码器通过跳跃连接输出的丰富空间特征信息,在原有深度估计网络编码器和解码器之间的跳跃连接处加

表 1　编码器网络结构参数

Table 1　Encoder network structure parameters

层	尺度	DenseNet121-D	DenseNet169-D
卷积	S/2	7×7 conv，64，stride 2	
Dense block (1)	S/2	$\begin{bmatrix}1\times1\text{conv}\\3\times3\text{conv}\end{bmatrix}\times6$，stirde 1	$\begin{bmatrix}1\times1\text{conv}\\3\times3\text{conv}\end{bmatrix}\times6$，stirde 1
Transition layer (1)	S/4	1×1 conv，128，stride 1 2×2 平均池化，stride 2	
Dense block (2)	S/4	$\begin{bmatrix}1\times1\text{conv}\\3\times3\text{conv}\end{bmatrix}\times$ 12，stirde 1	$\begin{bmatrix}1\times1\text{conv}\\3\times3\text{conv}\end{bmatrix}\times$ 12，stirde 1
Transition layer (2)	S/8	1×1 conv，256，stride 1 2×2 平均池化，stride 2	
Dense block (3)	S/8	$\begin{bmatrix}1\times1\text{conv}\\3\times3\text{conv}\end{bmatrix}\times$ 24，stride 1	$\begin{bmatrix}1\times1\text{conv}\\3\times3\text{conv}\end{bmatrix}\times$ 32，stride 1
Transition layer (3)	S/16	1×1 conv，512，stride 1 2×2 平均池化，stride 2	1×1 conv，640，stride 1 2×2 平均池化，stride 2
Dense block (4)	S/16	$\begin{bmatrix}1\times1\text{conv}\\3\times3\text{conv}\end{bmatrix}\times$ 16，stirde 1，1 024	$\begin{bmatrix}1\times1\text{conv}\\3\times3\text{conv}\end{bmatrix}\times$ 32，stirde 1，1 664
Downsample layer	S/32	Batch norm，ReLU， 2×2 平均池化，stride 2	

入了自研的 DEM。该模块不会改变输入与输出通道，仅对编码器每个阶段的输出特征图进行细化以提取边缘特征。具体模块结构如图 2 所示。

具体来说，首先将编码器输出的特征图通过一个 3×3 卷积层探取特征 $\boldsymbol{X}$。随后，通过全局池化层将特征 $\boldsymbol{X}$ 压缩成一个向量，获取上下文信息，并通过 2 个 1×1 卷积层和一个 Sigmoid 激活函数得出权重向量 $\boldsymbol{Y}$，以重新校准不同通道的重要性。接着，对 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 进行逐像素点积运算以重新生成加权特征，通过这个操作，包含关键信息的通道特征将获得更大的权重，从而增强多尺度特征图的边缘细节。最后，将重新生成的加权特征与特征 $\boldsymbol{X}$ 进行整合，其数学表达式为：

$$\boldsymbol{Z}=\boldsymbol{X}\odot\boldsymbol{Y}+\boldsymbol{X} \tag{2}$$

式中：$\odot$ 为逐元素卷积；$\boldsymbol{Z}$ 为最终的输出特征。

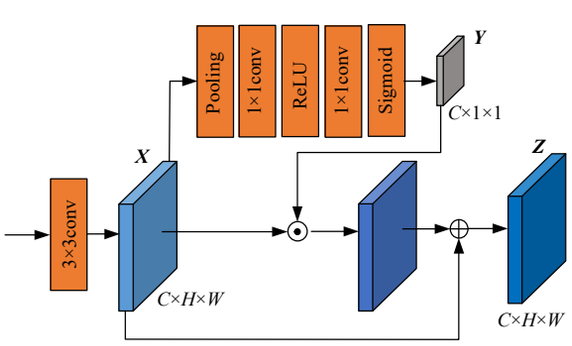


图 2　DEM 结构

Fig. 2　DEM structure

1.3　基于小波分解的解码器

大多数深度估计 U-Net 网络架构解码器只是将通道特征进行简单融合，导致特征图的高低频信息无法准确整合。在深度估计中，清晰的边缘信息不仅能够直观呈现物体轮廓，而且能够区别不同的深度区域。而这些边缘信息通常也是高频信息。文献[4,25]表明，大部分信号以编码器最低分辨率在深度图的低频估计中被捕获，即深度图的粗略估计足以捕获场景的全局几何结构。深度图中的边缘细节可以通过小波高频系数添加到深度图中，以获得更清晰的结果。

本文基于文献[21]，使用 Haar 小波的逆离散小波变换（Inverse discrete wavelet transform，IDWT），以系数图的 2 倍分辨率将 4 个系数图转换为 2D 图像，即 1 个低频分量（LL）和 3 个高频分量（LH、HL、HH）。4 个系数图分辨率为输出 LL 图像的 1/2。通过将 IDWT 递归应用于低频系数图 LL 以重建全分辨率图像，如图 3 所示，其中不同颜色框的频率分量对应图 1 中与之颜色相同的部分。

如图 1 所示，解码器建立在 IDWT 对预测系数图的递归使用上。网络首先在 U 型架构的最深层尺度上进行粗略深度估计，然后通过预测高频系数图迭代上采样和细化深度估计，从而在原始输入尺度上重建深度图。

具体来说，网络在解码器特征图尺度为 S/16 处进行粗略深度估计获得低频深度图 LL3，然后预测稀疏小波系数{LH3，HL3，HH3}，通过 IDWT 融合得到分辨率 S/8 的新深度图 LL2，由此不断迭代上采样和细化该深度图，最终生成 5 个

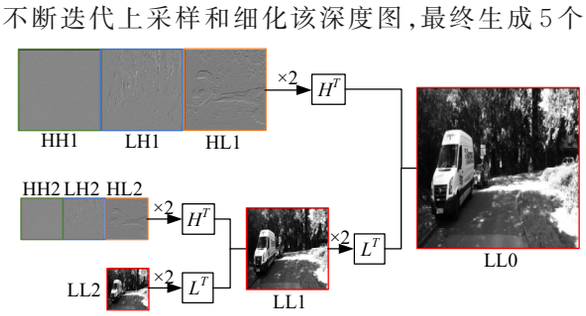


图 3　逆离散小波变换

Fig. 3　Inverse discrete wavelet transform

不同尺度的深度图LL的集合,其尺度分别为输入图片的1/16、1/8、1/4、1/2、1。值得注意的是,高频系数图只需在$S/16$、$S/8$、$S/4$、$S/2$处预测,无需在全分辨率处计算,如图1所示。与大多数解码器不同的是,基于小波分解的解码器在每个尺度上用3个通道输出层替代原本1个单通道输出视差的层,以预测{LH,HL,HH}。相比于传统解码器以1/16、1/8、1/4、1/2、1的比例监督解码器特征图,由于IDWT以2倍分辨率输出视差,基于小波分解的解码器方法只需要监督4个尺度(1/16、1/8、1/4、1/2),即可最终输出尺度为1的深度图。

2 损失函数

本文沿用文献[14]的自监督损失函数框架。给定目标图像I_t和另一视图的源图像$I_{t'}$,通过联合训练网络同时预测目标图像的密集深度图D_t和目标到原图像的相对相机姿态$T_{t\to t'}$,其损失函数构建流程如图4所示。

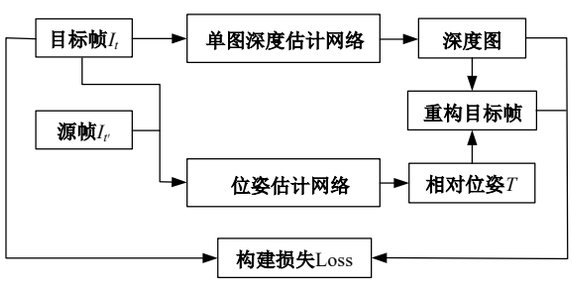


图4 自监督深度估计框架

Fig. 4 Self-supervised depth estimation framework

构造光度重投影损失函数,表达式为:

$$L_p=\sum_{t'}\rho(I_t,I_{t'\to t}) \tag{3}$$

式中:ρ为光度重建误差。该参数由光度损失L_1和结构相似度损失(SSIM)的加权组合构成,这种设计是因为在真实环境中光照条件不是固定不变的,引入SSIM能够更好地处理复杂光照的变化,具体定义为:

$$\rho(I_t,I_{t'\to t})=\frac{a}{2}(1-\mathrm{SSIM}(I_t,I_{t'\to t}))+(1-a)\left\|I_t,I_{t'\to t}\right\|_1 \tag{4}$$

式中:a计算过程中一般取0.85。SSIM可以定量比较2张图片的相似性,表达式为:

$$\mathrm{SSIM}(I_a,I_b)=[l(I_a,I_b)]^{\alpha}[c(I_a,I_b)]^{\beta}[s(I_a,I_b)]^{\gamma} \tag{5}$$

式中:α、β、γ为常数,一般取1。$I_{t'\to t}$是根据目标图像的深度(视差的倒数)和相机位姿扭曲到目标坐标系的源图像,可表示为:

$$I_{t'\to t}=I_{t'}\langle proj(D_t,T_{t\to t'},K)\rangle \tag{6}$$

式中:$proj()$为转换函数,可将目标图像的像素p_t

$$p_{t'}\sim KT_{t\to t'}D_t(p_t)K^{-1}p_t \tag{7}$$

映射到源图像$p_{t'}$上;$\langle\cdot\rangle$为局部亚可微的双线性采样算子;K为相机内部参数,这里假设它固定不变。此外,当图像出现纹理较弱的区域时,边缘平滑损失能在上述光度重投影损失失效的情况下对深度预测值进行约束,具体损失函数为:

$$L_s=\left|\partial_x d_t^*\right|\mathrm{e}^{-\left|\partial_x I_t\right|}+\left|\partial_y d_t^*\right|\mathrm{e}^{-\left|\partial_y I_t\right|} \tag{8}$$

式中:$d_t^*=d/\overline{d_t}$为平均归一化反深度,用于阻止估计深度收缩。

为进一步保证深度预测的一致性,引入尺度一致损失,即:

$$L_c=\frac{\left|D_t-\tilde{D}_{t'\to t}\right|}{D_t+\tilde{D}_{t'\to t}} \tag{9}$$

式中:$\tilde{D}_{t'\to t}$为将源图像深度图$D_{t'}$根据相机姿态$T_{t\to t'}$向目标深度图扭曲投影后,再将像素网格通过式(7)对齐至目标图像的深度图。

综上,本文总损失函数的计算公式为:

$$L=\mu L_p+\lambda L_s+\tau L_c \tag{10}$$

$$\mu=[\min_{t'}\rho(I_t,I_{t'\to t})<\min_{t'}\rho(I_t,I_{t'})] \tag{11}$$

μ用于静态掩模,判断重投影的光度误差是否小于原光度误差:若成立,则$\mu=1$;反之,$\mu=0$。λ和τ分别为边缘光滑损失和尺度一致性损失的权重,训练时分别设为0.1和0.5。

3 实验结果

3.1 实验设置

本文使用深度估计经典的自动驾驶室外场景数据集KITTI[26]对网络进行训练、测试并与其他方法进行对比分析。为与其他方法进行公平对比,参照Eigen等[4]对数据集的划分方法,选取29个场景中的697张图像作为测试集,其他32个场景的22 600张图片和888张图片分别作为训练集和验证集。

本文模型使用Pytorch框架实现,Python版本为3.7,Pytorch版本为1.8,CUDA版本为11.3。实验设备显存为24 GB的NVIDIA RTX 3090Ti显卡和Ubuntu操作系统。模型使用Adam优化器进行训练,学习率为10^{-4},训练迭代轮数

（epoch）为 20，批处理（batch size）大小为 12。在训练时，使用文献[27]所提的 efficient memory 方法以减少训练所占显存。将网络模型的训练过程以伪代码的形式呈现，见算法 1。

算法 1：自监督单目深度估计网络模型训练

输入：从数据集中选取目标图像 I_t 和源图像 $I_{t'}$

输出：训练后的自监督单目深度估计网络模型

1. 初始化深度估计网络模型
2. 定义深度估计网络的损失函数 L
3. 定义 Adam 优化器
4. 设置训练迭代次数 20、批处理大小 12 和学习率 10^{-4}
5. **while** 训练迭代次数≤20 **do**：
6. 　　从数据集中选取目标图像 I_t 以及其源图像 $I_{t'}$
7. 　　目标图像 I_t 通过深度估计网络得到深度图 D_t
8. 　　位姿估计网络预测 I_t 与 $I_{t'}$ 之间位姿 $T_{t\to t'}$
9. 　　根据 D_t 与位姿 $T_{t\to t'}$ 将 $I_{t'}$ 向 I_t 扭曲得到 $I_{t'\to t}$
10. 　　根据 L 计算 $I_{t'\to t}$ 与 I_t 图像之间的重构误差
11. 　　根据损失函数梯度反向传播更新模型参数优化
12. **end while**

3.2　评估指标

采用 Eigen 等[10]提出的标准度量作为深度估计模型性能的评估指标，分别为绝对相对误差（AbsRel）、平方相对误差（SqRel）、均方根误差（RMSE）、均方根对数误差（RMSElog）以及不同阈值下的准确度（δ），分别具体表示为：

$$\mathrm{AbsRel}=\frac{1}{|N|}\sum_{i\in N}\frac{\|d_i-d_i^*\|}{d_i^*} \tag{12}$$

$$\mathrm{SqRel}=\frac{1}{|N|}\sum_{i\in N}\frac{\|d_i-d_i^*\|^2}{d_i^*} \tag{13}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{|N|}\sum_{i\in N}\|d_i-d_i^*\|^2} \tag{14}$$

$$\mathrm{RMSElog}=\sqrt{\frac{1}{|N|}\sum_{i\in N}\|\log(d_i)-\log(d_i^*)\|^2} \tag{15}$$

$$\delta=\max\left(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i}\right)<T,\ T=1.25,1.25^2,1.25^3 \tag{16}$$

式中：d_i 为像素点 i 的预测深度值；d_i^* 为像素点 i 的真实深度值；N 为可获取真实深度的像素点个数；δ 为比值；T 为阈值。

3.3　结果对比

本文选取了近几年的主流先进算法与本文方法在 KITTI 和 CityScapes 数据集上进行了测试比较，如表 2 和表 3 所示。结果表明，本文方法具有一定的先进性和有效性，在多项指标上优于现有单目深度估计方法。

自监督深度估计的训练方式包括前后帧、立体图像对以及前后帧加立体图像对，分别对应表 2 中的 M、S 和 MS。本文的深度估计网络都使用 DenseNet121-D 编码器与其他方法进行对比。此外，还将本文方法与其他方法在高分辨率（1 024×320）输入图像的情况下进行比较，对应表 2 中的 HD。除此之外，所有输入图像大小都为 640×192。

从表 2 中可以看出，本文方法在不同监督方式下的性能指标几乎都优于基线方法[21]。与基线方法相比，本文方法在 S 监督方式训练后测试的绝对相对误差、平方相对误差、均方根误差以及均方根对数误差分别降低了 6.3%、8.2%、3.8%、2.4%，在阈值 $\delta<1.25$、$\delta<1.25^2$ 时的准确度分别提升了 1.3%、0.6%。此外，对高分辨率输入图像的网络以及 MS 监督方式训练的网络进行了训练测试，并与基线算法进行对比，结果也表明本文方法的测试指标有显著提高。与基线方法相比，本文方法在高分辨率输入图像训练后测试的绝对相对误差、平方相对误差、均方根误差以及均方根对数误差分别降低了 7.6%、6.4%、4.2%、3.0%，在阈值 $\delta<1.25$、$\delta<1.25^2$、$\delta<1.25^3$ 时的准确度分别提升了 1.0%、0.2%、0.1%。与先进方法相比，本文方法在 MS 监督方式训练后测试的绝对相对误差、平方相对误差、均方根误差以及均方根对数误差分别降低了 5.7%、10.2%、4.5%、3.1%，在阈值 $\delta<1.25$、$\delta<1.25^2$ 时的准确度分别提升了 0.8%、0.2%。为了进一步证明本文方法的有效性和泛化性能，对在 KITTI 数据集训练后的网络在 CityScapes 数据集上进行迁移测试，表 3 中的测试结果表明本文方法相比主流算法也具有更好的性能。

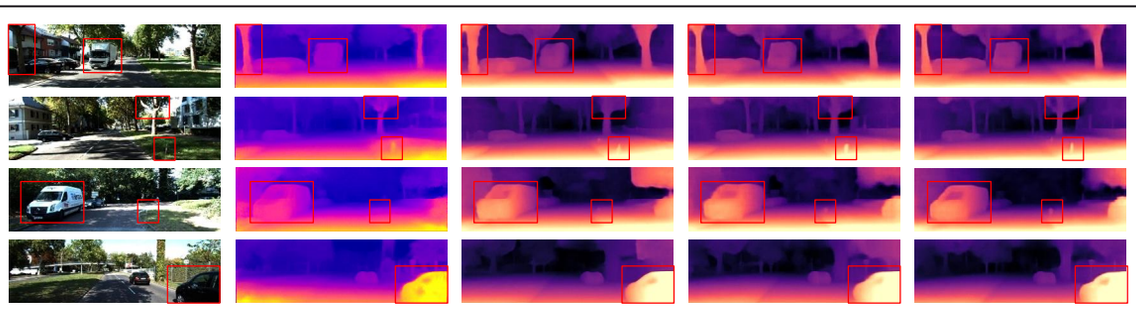
为更直观地展示本文方法的有效性，比较了本文方法与 3 个先进方法在 KITTI 和 CityScapes 数据集上的可视化效果图，分别如图 5 和图 6 所示。从图中可以看出，相比于现有方法，本文方法预测的深度图中物体边缘更加完整且更加贴合物体本身形状，并且不同深度的区域之间边缘更加清晰锐利，深度一致的区域上纹理也更加平滑。从图 5 中第 2 行和第 3 行深度图中红框圈出的小柱子轮廓可以看出，本文方法对小物体的深度预测能力明显优于其他方法，尤其是第 3 行图中

Monodepth和Monodepth2方法甚至没有检测到目标;此外,本文方法在对图5中第1行、第3行和第4行深度图中汽车车窗的轮廓以及第2行深度图中树杈边缘的深度估计也明显比其他方法更加

表2 KITTI数据集上的测试结果

Table 2 Test results on the KITTI dataset

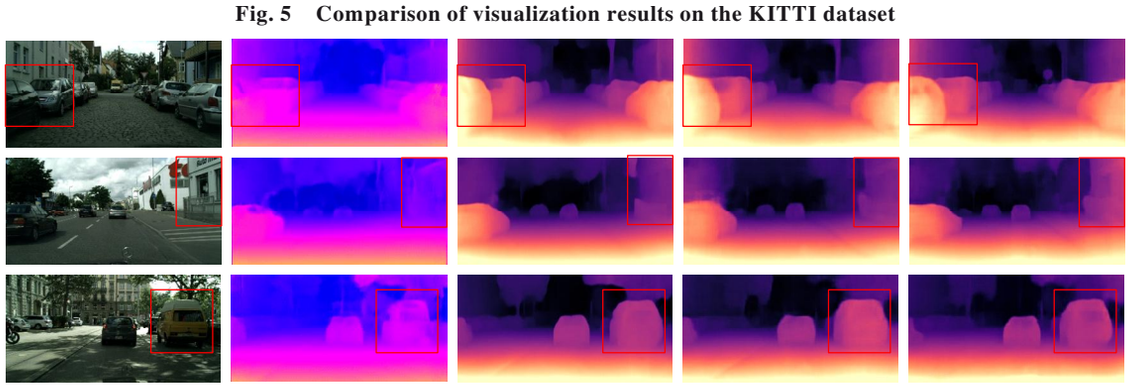
方法	监督方式	误差				准确度		
		AbsRel	SqRel	RMSE	RMSElog	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Garg[6]	S	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Monodepth R50[13]	S	0.133	1.142	5.533	0.230	0.830	0.936	0.970
StrAT[28]	S	0.128	1.019	5.403	0.227	0.827	0.935	0.971
3Net R50[29]	S	0.129	0.996	5.281	0.223	0.831	0.939	0.974
3Net VGG[29]	S	0.119	1.201	5.888	0.208	0.844	0.941	0.978
SuperDepth[30]	S	0.112	0.875	4.958	0.207	0.852	0.947	0.977
VA-Depth[16]	M	0.112	0.864	4.804	0.190	0.877	0.959	0.982
Zeeshan[17]	M	0.113	0.903	4.863	0.193	0.877	0.959	0.981
STDepthFormer[18]	M	0.110	0.805	4.678	0.187	0.878	0.961	0.983
Monodepth2[8]	S	0.109	0.873	4.960	0.209	0.864	0.948	0.975
WaveletMonodepth[21](基线)	S	0.110	0.876	4.916	0.206	0.864	0.950	0.976
本文	S	0.103	0.801	4.727	0.201	0.877	0.953	0.976
Monodepth2(HD)[8]	S	0.107	0.849	4.764	0.201	0.874	0.953	0.977
WaveletMonodepth(HD)[21]	S	0.105	0.797	4.732	0.203	0.869	0.952	0.977
本文(HD)	S	0.097	0.726	4.531	0.195	0.884	0.955	0.978
Monodepth2[8]	MS	0.106	0.818	4.750	0.196	0.874	0.957	0.979
WaveletMonodepth[21]	MS	0.109	0.814	4.808	0.198	0.868	0.955	0.980
本文	MS	0.100	0.731	4.536	0.190	0.882	0.959	0.980



(a)输入图像 (b)Monodepth (c)Monodepth2 (d)WaveletMonodepth (e)本文方法

图5 KITTI数据集的可视化结果对比

Fig. 5 Comparison of visualization results on the KITTI dataset



(a)输入图像 (b)Monodepth (c)Monodepth2 (d)WaveletMonodepth (e)本文方法

图6 CityScapes数据集的可视化结果对比

Fig. 6 Comparison of visualization results on the CityScapes dataset

清晰。在CitySpaces数据集上的性能对比中（图6）同样可以看出本文方法对物体边缘和轮廓细节的深度估计准确性明显优于其他方法。特别是在图6第2行深度图中，红框圈出的建筑物的轮廓以及其窗户与墙面之间的深度差都被准确呈现；且在第3行深度图中，面包车的轮廓更加符合实际，后窗玻璃也在本文方法中更加明显。

表3　CityScapes数据集上的迁移测试结果

Table 3　Migration test results on the CityScapes dataset

方法	AbsRel	SqRel	RMSE	RMSElog
Monodepth R50[13]	0.210	2.230	9.430	0.311
Monodepth2[8]	0.182	1.880	8.870	0.253
WaveletMonodepth[21]	0.185	1.950	8.659	0.248
本文	0.175	1.831	8.590	0.242

3.4　消融实验

通过消融实验进一步验证各模块的有效性，结果如表4所示，其中未使用的模块用×表示，使用的模块用√表示。将文献[8]的方法作为基线方法（表4中实验1），将只引入小波分解的解码器方法记为实验2。结果表明，设计的DenseNet-D编码网络和DEM均能有效提高单目深度估计网络的性能。比较实验2和实验3指标可以看出，DEM能够在一定程度上有效提升深度估计网络的精度。而比较实验2和实验4指标可以看出，各指标有了显著改善，这说明本文设计的基于DenseNet-D网络的编码器能够对图像进行更有效的特征提取，提升深度估计网络的深度线索捕捉能力，从而提高深度估计精度并降低误差。此外，相比于使用ResNet作为编码器，DenseNet-D的参数量仅为6.95 M，远小于ResNet18的11.2 M。

为了证明本文对DenseNet编码网络层的改动是有效的，进行消融实验，结果如表5所示。对比表5实验1和实验2可以看出，直接将DenseNet网络用于深度估计网络的编码器是有效的。但是，在编码器网络层数增加的情况下，整个网络的深度估计误差没有降低反而增加。这是因为本文保持了原有深度估计网络编码器输出的多尺度特征数量，直接使用原有的DenseNet网络结构替换ResNet网络架构作为编码器并不能很好地适配深度估计网络的编码器。在使用改进的DenseNet-D编码器后，误差和精度指标均有所改善。重要的是，在编码器网络层数增加的情况下，评价指标继续提高，见实验4。

表4 本文模块消融实验

Table 4　Ablation experiment of modules

实验	Densenet-D	DEM	Wavelet	AbsRel	SqRel	RMSE	RMSElog	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
1	×	×	×	0.108	0.842	4.891	0.207	0.865	0.949	0.976
2	×	×	√	0.110	0.876	4.916	0.206	0.864	0.950	0.976
3	×	√	√	0.107	0.865	4.891	0.204	0.867	0.950	0.976
4	√	×	√	0.105	0.810	4.739	0.202	0.876	0.952	0.976
5	√	√	√	0.103	0.801	4.727	0.201	0.877	0.953	0.976

表5 本文编码器设计消融实验

Table 5　Ablation experiment of encoder design

实验	编码器	AbsRel	SqRel	RMSE	RMSElog	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
1	DenseNet121	0.106	0.862	4.882	0.203	0.872	0.951	0.976
2	DenseNet169	0.108	0.892	4.972	0.205	0.871	0.950	0.976
3	DenseNet121-D	0.103	0.801	4.727	0.201	0.877	0.953	0.976
4	DenseNet169-D	0.102	0.797	4.745	0.200	0.879	0.953	0.976

4　结束语

为提高单目深度估计精度，本文提出了一种改进密集网络和小波分解的自监督深度估计网络。该网络编码器由改进后的DenseNet-D网络组成，采用密集连接的方式促进浅层特征复用和特征的前向传播；在跳跃连接处设计了DEM，用于多尺度边缘特征进一步细化整合；解码器基于IDWT的递归使用，通过预测小波系数图迭代还原深度图。实验结果表明，相较于现有方法，本方法在KITTI和CityScapes数据集上实现了更小的误差和更高的精度；在KITTI和CityScapes数据集上的可视化结果显示，获得了同深度平面区域更加平滑、不同深度区域之间边缘更加清晰的

深度图。因为是自监督方法,无需深度真值图的监督,所以本文提出的网络也有良好的泛化能力。本文提出的单目深度估计网络仅在精度上有明显提高,在以后的工作中,将继续完善该网络以获得更加轻量化的单目深度估计网络。

参考文献:

[1] 王新竹, 李骏, 李红建, 等. 基于三维激光雷达和深度图像的自动驾驶汽车障碍物检测方法[J]. 吉林大学学报(工学版), 2016, 46(2): 360-365.
Wang Xin-zhu, Li Jun, Li Hong-jian, et al. Obstacle detection based on 3D laser scanner and range image for intelligent vehicle[J]. Journal of Jilin University (Engineering and Technology Edition), 2016, 46(2): 360-365.

[2] 张宇翔, 任爽. 定位技术在虚拟现实中的应用综述[J]. 计算机科学, 2021, 48(1): 308-318.
Zhang Yu-xiang, Ren Shuang. Overview of the application of location technology in virtual reality[J]. Computer Science, 2021, 48(1): 308-318.

[3] 史晓刚, 薛正辉, 李会会, 等. 增强现实显示技术综述[J]. 中国光学, 2021, 14(5): 1146-1161.
Shi Xiao-gang, Xue Zheng-hui, Li Hui-hui, et al. Overview of augmented reality display technology [J]. China Optics, 2021, 14(5): 1146-1161.

[4] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]//2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015: 2650-2658.

[5] Fu H, Gong M, Wang C, et al. Deep ordinal regression network for monocular depth estimation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 2002-2011.

[6] Garg R, Vijay K B G, Carneiro G, et al. Unsupervised CNN for single view depth estimation: geomery to the rescue[C]//European Conference Computer Vision, Amsterdam, Netherlands, 2016: 740-756.

[7] Zhou T H, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, 2017: 1851-1858.

[8] Clément G, Oisin M A, Michael F, et al. Digging into self-supervised monocular depth estimation[C]// 2015 IEEE International Conference on Computer Vision (ICCV), Seoul, South Korea, 2019: 3828-3838.

[9] Ashutosh S, Sun M, Andrew Y N. Make3D: learning 3D scene structure from a single still image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 31(5): 824-840.

[10] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]//Advances in Neural Information Processing Systems, Montreal, Canada, 2014: 2366-2374.

[11] Zachary T, Jia D. Deepv2D: video to depth with differentiable structure from motion[C]//International Conference on Learning Representations (ICLR) 2020, Addis Ababa, Ethiopian, 2020: 181204605.

[12] Benjamin U, Zhou H Z, Jonas U, et al. Demon: depth and motion network for learning monocular stereo[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, 2017: 5038-5047.

[13] Clément G, Oisin M A, Gabriel J B. Unsupervised monocular depth estimation with left-right consistency [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, 2017: 270-279.

[14] Bian J W, Li Z C, Wang N, et al. Unsupervised scale-consistent depth and ego-motion learning from monocular video[C]//33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2019: 1-12.

[15] Han C, Cheng D, Kou Q, et al. Self-supervised monocular depth estimation with multi-scale structure similarity loss[J]. Multimedia Tools and Applications, 2022, 31: 3251-3266.

[16] Xiang J, Wang Y, An L, et al. Visual attention-based self-supervised absolute depth estimation using geometric priors in autonomous driving[J/OL]. (2022-10-06) [2023-06-13]. http://arxiv.org/abs/2205.08780v1.

[17] Suri Z K. Pose constraints for consistent self-supervised monocular depth and ego-motion[J/OL]. (2023-04-18) [2023-06-13]. http://arxiv.org/abs/2304.08916.

[18] Houssem B, Adrian V, Andrew C. STDepthFormer: predicting spatio-temporal depth from video with a self-supervised transformer model[C] //Detroit, USA, 2023: No. 230301196.

[19] Matteo P, Filippo A, Fabio T, et al. Towards real-time unsupervised monocular depth estimation on CPU[C]//2018 IEEE/RSJ international Conference

Intelligent Robots and Systems (IROS), Madrid, Spain, 2018: 5848-5854.

[20] Diana W, Ma F C, Yang T J, et al. FastDepth: fast monocular depth estimation on embedded systems [C]//2019 International Conference on Robotics and Automation (ICRA), Montreal, Canada, 2019: 6101-6108.

[21] Michael R, Michael F, Jamie W, et al. Single image depth prediction with wavelet decomposition[C]//Conference on Computer Vision and Pattern Recognition (CVPR), Online, 2021: 11089-11098.

[22] Olaf R, Philipp F, Thomas B. U-Net: convolutional networks for biomedical image segmentation[C]//International Conference On Medical Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 2015: 234-241.

[23] Huang G, Liu Z, Maaten L V D, et al. Densely connected convolutional networks[C]//2017 Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, 2017: 2261-2269.

[24] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 770-778.

[25] Chen X T, Chen X J, Zha Z J. Structure-aware residual pyramid network for monocular depth estimation[C]//28th International Joint Conference on Artificial Intelligence, Macau, China, 2019: 694-700.

[26] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: the kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[27] Pleiss G, Chen D, Huang G, et al. Memory-efficient implementation of densenets[J/OL]. (2017-07-21)[2023-06-13]. https://arxiv.org/abs/1707.06990v1.

[28] Mehta I, Sakurikar P, Narayanan P J. Structured adversarial training for unsupervised monocular depth estimation[C]//2018 International Conference on 3D Vision, Verona, Italy, 2018: 314-323.

[29] Matteo P, Fabio T, Stefano M. Learning monocular depth estimation with unsupervised trinocular assumptions[C] //International Conference on 3D Vision (3DV), Verona, Italy, 2018: 324-333.

[30] Sudeep P, Rares A, Ambrus G, et al. Superdepth: self-supervised, super-resolved monocular depth estimation[C]//2019 International Conference on Robotics and Automation (ICRA), Montreal, Canada, 2019: 9250-9256.