DOI:10.16136/j.joel.2022.09.0824

基于层级特征融合的单目深度估计算法

郑秋梅[1],于 涛[1],王风华[1*],林 超[2]

(中国石油大学(华东)计算机科学与技术学院,山东 青岛 266580;中国石油大学(华东)信息化建设处,山东 青岛 266580)

摘要:MonoDepth2的提出使自监督单目深度估计取得了重大的进展,但该网络在大的无语义区域和边界处预测效果并不理想,主要原因是基础的U-Net框架没有充分利用多尺度特征信息,导致来自于大梯度区域的深度估计较差。针对此问题,本文提出了一个改进的DepthNet,层级特征融合网络(hierarchical integration net,HINet)。优化了U-Net网络结构,使编码器端在每一层都能产生不同尺度的特征信息,从而让解码器端在每一层都能够充分融合多尺度特征。由于不同尺度的特征信息对于特定的解码器层都有不同程度的贡献,本文提出的层级特征融合算法还增加了通道注意力模块,提升重要特征尺度的权重。当采用立体图像对进行训练时,本文对数据进行了预处理,并增加了立体对的深度暗示损失函数。在KITTI数据集上的实验结果表明,所有指标均获得了不同程度的提升,其中绝对相对误差减少了0.09,平方相对误差减少了0.093。

关键词:单目深度;特征提取;U-Net;自监督;层级融合

中图分类号:TP391.41 **文献标识码**:A **文章编号**:1005-0086(2022)09-0925-07

Monocular depth estimation algorithm based on hierarchical integration

ZHENG Qiumei[1], YU Tao[1], WANG Fenghua[1*], LIN Chao[2]

(1. College of Computer Science and Technology, China University of Petroleum Huadong, Qingdao, Shandong 266580, China; 2. Information Construction Department, China University of Petroleum Huadong, Qingdao, Shandong 266580, China)

Abstract: The proposal of MonoDepth2 has made significant progress in self-supervised monocular depth estimation, but the prediction effect of the network in large non semantic regions and boundaries is not ideal. The main reason is that the basic U-Net framework does not make full use of multi-scale feature information, resulting in poor depth estimation from large gradient regions. To address this problem, this paper proposed an improved DepthNet, a hierarchical integration net (HINet). The U-Net network structure is optimized so that the encoder side can generate feature information of different scales at each layer, thus allowing the decoder side to fully fuse multi-scale features at each layer. Since the feature information of different scales contributes to a specific decoder layer to different degrees, the hierarchical integration (HINet) algorithm proposed in this paper also adds a channel attention module to enhance the weight of important feature scales. When stereo pairs are used for training, this paper preprocesses the data and adds a depth-implying loss function for stereo pairs. The experimental results on the KITTI dataset show that all indicators are improved to varying degrees, in which the absolute relative error is reduced by 0.09 and the squared relative error is reduced by 0.093.

Key words: monocular depth; feature extraction; U-Net; self-supervisied; hierarchical integration (HINet)

* **E-mail**: fenghuawang@upc.edu.cn

收稿日期:2022-01-27 **修订日期**:2022-02-20

基金项目:国家自然科学基金(52074341,51874340)和中央高校基本科研业务费专项资金资助(19CX02030A)资助项目

1 引 言

深度估计在计算机视觉领域一直是一个被广泛研究的课题。随着深度学习的迅速发展，卷积神经网络在单目深度估计领域几乎全面取代了传统的方式。然而，目前很多先进的方法依赖于有监督[1-3]的训练方式。虽然有监督的网络拥有很高的精度，但由于其训练数据集需要地面真实标签的标注，因此收集起来是一个麻烦而艰巨的任务。而自监督的训练方式无需标注，大大降低了前期任务的难度。仅仅通过立体图像、单目视频或者两者的结合进行重建，就可以获得与有监督方式相媲美的实验结果。目前大多数的自监督单目深度估计都采用了在医学分割领域有出色性能的 U-Net 作为基础网络架构，因为其既结合了低分辨率信息(提供物体类别识别依据)又结合了高分辨率信息(提供精准定位依据)。

在早期工作中，GODARD 等[4]提出了使用立体对进行自监督深度估计的训练方法，通过引入左右深度一致性损失产生了优于当代监督方法的结果。作者采用编解码器网络，其中解码器从编码器的激活块做残差连接，这样可以让网络能够分解更高的特征图细节。为了减少立体相机的限制，ZHOU 等[5]首先提出利用 PoseCNN 估计相邻帧间的相对位姿。这项工作使网络完全依赖于单目图像序列进行训练。SfMLearner[5]进一步利用视频序列中连续帧的集合进行深度和姿态网络的联合训练，显示出了与现存的监督方法相当的性能；然而，在这项工作中单目视图需要以下假设：视图场景完全是静止的，不存在动态物体；目标视图和源视图之间没有遮挡的物体，在实际情况中并不能满足上述条件。MonoDepth2[6]通过最小化重投影损失，在很大程度上解决了遮挡问题。同时又提出了应对违反相机运动的自动遮挡损失，用于消除动态物体对于精度的破坏，该方法[6]将自监督单目工作推到了一个新的高度。在此后的工作中，PackNet-SfM[7]通过加入额外的语义分割模型来辅助训练深度估计网络。GUIZILINI 等[8]引入了预先训练的语义分割网络和像素自适应卷积，引导深度网络进一步利用语义信息。但这些工作有一个缺点就是增加了前期的标注成本和后期的参数量。而 Hr-depth[9]在不增加额外约束的前提下，通过修改网络本身的架构，从而得到了更高的精确度。

综上分析，基础的 U-Net 的网络架构没有充分利用每一个尺度的特征信息，并且编解码器端的特征存在语义差距，仍存在改进的空间。本文在参数量较少并且拥有较高性能的 MonoDepth2[6]基础上重新设计 U-Net 的连接方式。本文的创新点共有 3 个方面：1) 改进编解码器端的结构，使其在编码器端的每一层都能产生不同尺寸的特征信息，并且解码器端可以有效地融合多尺度的特征信息；2) 增加了通道注意力模块，用于提高重要特征的权重；3) 当使用立体对进行训练时，增加了深度提示优化了训练方式。实验结果表明，本文可以在不增加训练时间的前提下，明显提高网络的精度。

2 层级融合算法

先前的大多数网络都是基于编解码器通过加入额外复杂的模型体系结构和外部模式的额外约束从而取得了精度的提升，这在一定程度上违背了自监督的概念。

本文通过分析此前的模型结构，发现所采用的基础 U-Net 框架仅通过编码器端进行多尺度的特征提取，再通过解码器端逐步的上采样恢复到原图像的尺寸，让整个网络显得模态化。简单来说，对于特征提取阶段，浅层结构可以抓取图像的一些简单的特征，比如边界、颜色，而深层结构因为感受野大了，经过的卷积操作多了，能抓取到图像的一些内在的重要特征。总之，浅有浅的侧重，深有深的优势。

因此本文重新设计网络的连接方式，使其能充分利用不同尺度的特征信息。首先通过将编码器端的所有特征都分别连接到解码器端的不同层上，让网络学习哪一层的特征是最有用的，最终通过剪枝得到了如图 1 所示的网络结构。同时，还观察到基础的 U-Net 仅使用一个卷积层实现残差连接，不能很好地利用空间信息和语义信息，因此本文重新设计了残差连接。此设计不仅能提高特征融合的效果，还能进一步减少参数量。同时优化了解码器端的结构，通过模拟多尺度并行卷积，使网络能充分利用接收到的多尺度特征，提升了网络的精度。

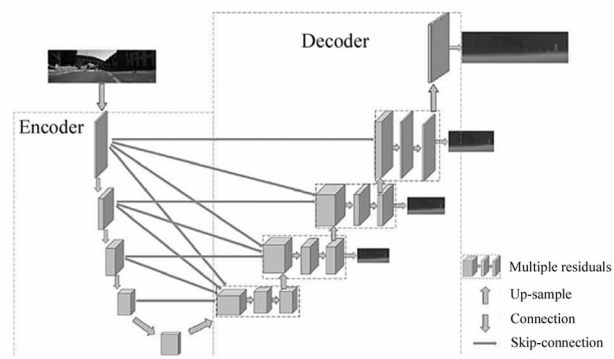

图 1 网络结构

Fig. 1 Network structure

2.1　多重残差连接的编码器模块

目前,大多数自监督单目深度估计的网络框架都采用基础的 U-Net 网络,仅通过每一层上的残差连接进行编解码器的特征图融合,虽然在一定程度上解决了梯度消失和架构退化问题,但无法充分利用不同层次的特征,导致在每一层的特征信息较少,从而造成了特征信息的浪费。

基于上述分析,为了得到预测更精准的深度图,本文尝试从语义信息和空间信息两方面对边界进行增强。因为此前的研究证明了(1)语义信息可以产生不同类别之间的边界,减少误分类造成的深度估计误差,(2)空间信息可以帮助网络了解边界的位置,从而更好地估计边界。

由于残差连接是 U-Net 网络的核心组件,其目的便是恢复在下采样过程中所丢失的信息。因此通过将编码器端每一层都分别与解码器端相连,使其有足够的特征信息进行学习。同时考虑到解码器端的特征图是经过大量的卷积运算所得到的,因此解码器端的特征信息与编码器端的特征信息可能存在较大的语义差距,直接相连可能会对实验结果产生不利影响。为了减轻编码器-解码器之间的差异,受到 IBTEHAZ[10] 的启发,本文引入了改进的残差连接。不是简单地将特征图从编码器连接到解码器,而是先将它们穿过带有残差连接的卷积层,然后再与解码器特征连接。同时由于层次越深的编解码器特征图语义差距越小,所以在不同层所采用的残差连接个数也不同。通过在残差连接中的多个卷积层,进一步缩小了编解码器特征图的语义差距,结构如图 2 所示。

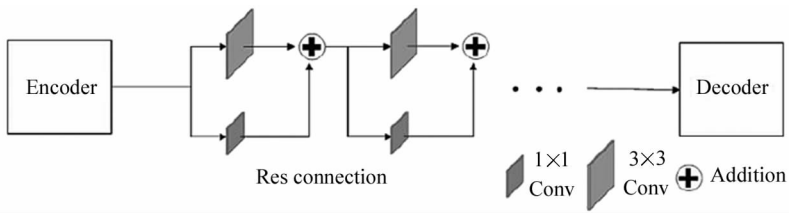

图 2　残差连接结构

Fig. 2　Residual connection structure

2.2　多尺度卷积的解码器模块

解码器的主要任务是接收编码器端的特征图信息,通过激活函数得到与之相对应的深度图。但由于解码器端接收到了不同尺寸的特征,因此网络应该能健壮地处理这些特征。

在基础的 U-Net 架构中,通过在解码器端依次使用两个 3×3 卷积层来处理接收到的特征图。正如 SZEGEDY 等[11] 所解释的,这两个 3×3 卷积运算系列实际上类似于 5×5 卷积运算。革命性的 Inception[12] 架构引入了 Inception 块,它利用不同内核大小的卷积层并行地处理来自不同尺度的图像中的关键信息点。因此按照这个思路处理多分辨率特征的最好方法就是将 3×3 卷积运算、7×7 卷积运算与 5×5 卷积运算并行结合。

但由于每一层解码器都需要处理大量不同尺寸的特征,因此在并行中引入额外的卷积层极大地增加了内存需求。本文进一步采用了参数更少的连接方式,如图 3 所示。第 2 个和第 3 个 3×3 卷积块

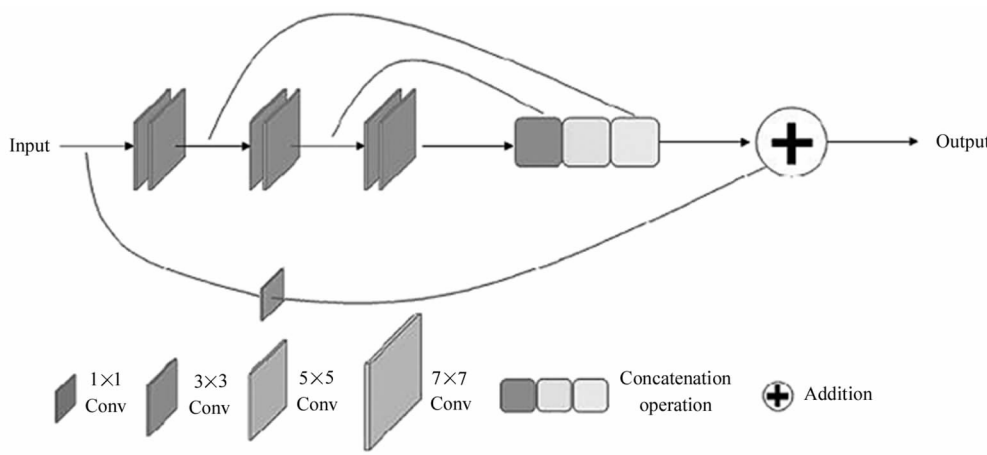

图 3　特征融合结构

Fig. 3　Feature fusion structure

的输出分别有效地逼近了 5×5 和 7×7 卷积运算。因此，将 3 个卷积块的输出拼接在一起，用于融合不同尺度的空间特征。

2.3　辅助增强策略

由于本文所采用的多层级特征图的连接会在一定程度上增加参数量，从而造成网络训练速度降低。针对此问题，本文采用了通道注意力模块。该模块不仅能减少网络的参数量，而且能增加重要通道的权重，让网络能更充分地利用重要特征，其结构如图 4 所示。

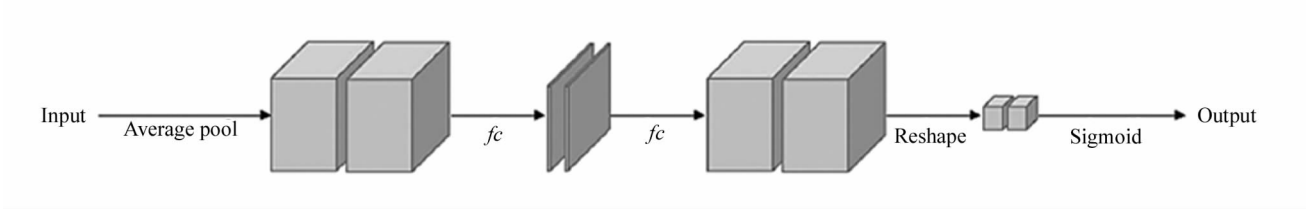

图 4　通道注意力模块

Fig. 4　Channel attention module

来自编码器的特征图 $H\times W\times C$，通过全局平均池化层和全连接层，拉伸成 $1\times 1\times C$，然后再与原图像相乘，将每个通道赋予权重，在增加了重要通道作用的同时削弱了不重要特征的权重。最后跟一个 Sigmoid 函数来测量每个特征的重要性并同时加权，再还原成与原图像大小形同的特征图。在基础的 U-Net 网络中，解码器融合编码器和上采样的特征，参数量为：

$$C_{\text{in}}\times C_{\text{out}}\times K^2+C_{\text{out}}, \tag{1}$$

式中，C_{in} 是输入通道的数量，C_{out} 是输出通道的数量，K 是卷积核的大小。而经过此通道注意力模块后，参数数量变为：

$$\frac{2}{r}\times C_{\text{in}}^2+(C_{\text{in}}+1)\times C_{\text{out}}, \tag{2}$$

式中，r 是缩减比例，在本文中设置为 16，对比式(1)和式(2)可以看出参数量明显下降。单目深度估计面临一个棘手的问题就是缺乏数据集的数量和多样性，从而导致许多深度模型会产生过拟合，导致泛化能力较差。因此有一些研究只是使用了图像翻转等简单数据扰动，自监督单目深度模型训练时需要严格的像素对应(极线约束)来确保匹配误差仅来自估计的视差。显然这种数据扰动会扰乱像素之间的对应关系，从而损害模型性能。

当采用立体图像对进行训练时，这种限制就被放宽了。因为这两个视图是用平行摄像机拍摄并校正的，它们之间的匹配将只发生在水平方向，所以可以通过在垂直方向上添加扰动，从而增强数据。受到 PENG 等[13]的启发，本任务采用了数据拼接的数据增强方式。通过将两个具有不同语义的图像嫁接在一起可以有效地缓解过度匹配风险，并鼓励模型在不破坏极线约束的情况下更好地利用输入的上下文。同时，将 WATSON 等[14]证明对薄结构有效的暗示损失纳入利用立体对图像训练时的模型。深度提示是由半全局匹配(semi-global block matching，SGM)算法[15]生成的。

2.4　损失函数

通过损失函数对模型进行优化是一个极为重要的环节，好的损失函数可以大幅提高模型的性能。当采用单目视频序列和其与立体图像对结合训练时，训练方式为采用单目视频帧，本文采用与 MonoDepth2 相同的逐像素平滑损失和遮蔽光度损失，并对每个像素、比例和批次进行平均。

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

$$I_{t'\to t}=I_{t'}<\text{proj}(D_t,\boldsymbol{T}_{t\to t'},P)>, \tag{4}$$

式中，<>是采样符，I_t 为目标图像，$I_{t'-t}$ 为源图像对目标图像的投影，P 为相机内参，$\boldsymbol{T}_{t\to t'}$ 为旋转矩阵，D_t 为 I_t 处的深度，其中 pe 为光度误差函数，即：

$$\text{pe}(I_a,I_b)=\frac{\alpha}{2}(1-SSIM(I_a,I_b))+(1-\alpha)\parallel I_a-I_b\parallel。 \tag{5}$$

本任务将 α 设为 0.85。边缘平滑损失为：

$$L_S=|\partial_x d_t^*|\mathrm{e}^{-|\partial_x I_t|}+|\partial_y d_t^*|\mathrm{e}^{-|\partial_y I_t|}。 \tag{6}$$

当训练方式为采用立体图像对时，采用光度损失和深度提示损失在每个尺度上的平均值进行模型优化。

$$L=\frac{1}{|D|}\sum_{d\in D}(L_p(d)+L_h(d)), \tag{7}$$

其中：

$$L_h(d_i)\begin{cases}\log(|h_i-d_i|+1),\text{if } l_p(I,\tilde{I}_h)_i<l_p(I,\tilde{I})_i\\ 0,\text{otherwise}\end{cases}, \tag{8}$$

式中，$\tilde{I}_h$ 表示使用深度提示重建的视图。

3　实验与结果

3.1　数据集

在 KITTI 立体数据集上评估了本文的模型。KITTI 基准最广泛用于深度评估。本文采用了 Eigen 的数据分割，并采用了 GODARD 去除静态帧的做法。最终，使用了 39 810 张图像进行训练，4 424 张用于验证，697 张用于评估。此外，对所有图像使用相同的内在设置，将相机的主点设置为图像中心，将焦距设置为 KITTI 中所有焦距的平均值。对于立体训练，将两个立体帧之间的变换设置为固定长度的纯水平平移，并采用相同的数据增强和深度提示。

3.2　实验环节及参数设置

最终在 PyTorch 上实现本文的模型，并在一台 Telsa V100s GPU 上训练它们。并使用 Adam Optimizer 进行优化，$\beta_1=0.9$，$\beta_2=0.999$。DepthNet 和 PoseNet 被训练了 20 个纪元，每批 12 个。在使用单目和组合训练时两个网络的初始学习率都是 1×10^{-3}，在使用立体图像对进行训练时，初始学习率被设置为 1×10^{-4}。并且都在 15 个纪元以后衰减了 10 倍。训练序列由 3 幅连续图像组成。设置 *SSIM* 权重为 $\alpha=0.85$，平滑损失权重为 $\lambda=1\times10^{-3}$。使用 ResNet-18 作为编码器，4 个尺度的视差图均用于训练时的损耗计算。为了评估，本文只使用最大输出尺度，然后使用双线性插值调整为地面真实深度分辨率。

3.3　评估指标

为了能够定量分析模型的精度，同时能够与主流算法做精度对比，本文使用 Eigen 中提出的评价指标，这也是绝大部分主流算法采取的精度评价标准，包含：均方根误差（root-mean-square error，*RMSE*）、对数空间下的均方根误差（log-root-mean-square error，*RMSE log*）、绝对相对误差（absolute relative error，*Abs Rel*）、平方相对误差（square relative error，*Sq Rel*）。具体如下：

$$RMSE=\sqrt{\frac{1}{|N|}\sum_{i\in N}\|d_i-d_i^*\|^2},\tag{9}$$

$$RMSElog=\sqrt{\frac{1}{|N|}\sum_{i\in N}\|\log(d_i)-\log(d_i^*)\|^{2^2}},\tag{10}$$

$$Abs\ Rel=\sqrt{\frac{1}{|N|}\sum_{i\in N}\frac{|d_i-d_i^*|}{d_i^*}},\tag{11}$$

$$Sq\ Rel=\frac{1}{|N|}\sum_{i\in N}\frac{\|d_i-d_i^*\|^2}{d_i^*},\tag{12}$$

$$Accuracies=\%\text{of}\ d_i\ \text{s.t.}\ \max\left(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i}\right)=\delta<thr,\tag{13}$$

式中，d_i 是像素 i 的预测深度值，d_i^* 表示深度的真实值，N 为具有真实值的像素总数，thr 为阈值。式（13）的阈值精度计算是对多图像中所有的像素点分别计算其预测深度与真实深度之比并取最大值，最后将结果赋值给 δ。统计 δ 小于阈值 thr 的像素点占总像素点的比例即为正确率（*Accuracies*），结果越接近于1，效果越好。thr 一般取为1.250、1.225、1.235。

3.4　实验结果分析

首先在 KITTI 数据集上验证模型的性能，并对每个组件进行了全面消融研究。如表 1 所示，将 HINet 与表 1 中的自监督和半监督单目深度估计方法进行了比较（后 3 个指标为正确率，数值越大越好）。结果表明，HINet 在所有指标上优于此前的自监督方法。与基线模型[6]相比，仅在立体声对上训练的方法在 $\delta<1.25$ 上改善了 0.013。并且在 HINet 中并没有像基线模型一样微调训练完的模型权重。

表 1　实验结果

Tab. 1　Experimental result

Method	Supervision	$H\times W$	*Abs Rel*	*Sq Rel*	*RMSE*	*RMSE log*	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
EPC++[9]	M	640×192	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2Depth[16]	M	640×192	0.141	1.026	5.291	0.215	0.816	0.945	0.979
MonoDepth2[6]	M	640×192	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Ours	M	640×192	**0.113**	**0.887**	**4.799**	**0.189**	**0.878**	**0.960**	**0.982**
MonoDepthR50[17]	S	640×192	0.133	1.142	5.533	0.230	0.830	0.936	0.970
MonoDepth2[6]	S	640×192	0.109	0.873	4.960	0.209	0.864	0.948	0.975
Ours	S	640×192	**0.101**	**0.780**	**4.544**	**0.185**	**0.884**	**0.961**	**0.982**

3.5　消融实验

为了进一步探索网络组件提供的性能改进，将引入的不同架构组件进行了消融分析。结果如表 2 所示。所有的实验结果都未进行微调。

残差连接模块(Res)的贡献：为了减轻编码器和解码器之间的差异，本文沿着编解码器之间的连接加入了卷积层和额外的残差连接。通过增加这些额外的非线性变换可以减小编码器和解码器特征之间的语义差距，并且加入的残差连接可以使网络学习变得更加容易。在实验结果中，当去除残差连接模块时，*Sq Rel* 下降了0.014，*RSME* 下降了0.049。

通道注意力模块(SE)的贡献：通过将经过残差连接的特征输入到通道注意力模块，可以增加重要特征的权重，同时提高计算的效率，大大减少密集跳接引入的参数，进一步提高网络性能。在实验结果中，当去除通道注意力模块时，*Abs Rel* 下降最为明显，主要原因为大量不重要特征混合在一起，导致网络学习出现偏差。

表 2　消融实验

Tab. 2　Ablation experiment

Method	Res	SE	MR	DC	Supervision	*Abs Rel*	*Sq Rel*	*RMSE*	*RMSE log*	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Baseline					S	0.109	0.873	4.960	0.209	0.864	0.948	0.975
HINet (woDC)	√	√	√		S	0.102	0.807	4.573	0.184	0.886	0.962	0.982
HINet (woMR)	√	√	√		S	0.102	0.788	4.605	0.186	0.880	0.961	0.982
HINet (woSE)	√		√	√	S	0.116	0.829	4.614	0.186	0.884	0.963	0.983
HINet (woRes)		√	√	√	S	0.101	0.794	4.593	0.185	0.885	0.961	0.982
HINet (Full)	√	√	√	√	S	0.101	0.780	4.544	0.185	0.884	0.961	0.982

多重残差模块(MR)的贡献：通过多重残差模块来模拟3×3，5×5 和7×7 卷积的并行运算，在减少参数量的基础上，可以有效地融合多尺度特征。在实验结果中，当去除多重残差模块时，*RMSE* 下降了0.012。

密集连接模块(DC)的贡献：将编码器端的每一层特征都连接到解码器端，可以使解码器的每一个层级都能充分利用多尺度特征，避免特征的不充分利用，进一步提高深度预测的精准度。在实验结果中，当去除密集连接模块时，*SqRel* 下降了0.027，下降程度最为明显。

4　结　论

本文在充分解析基线模型的基础上，重新设计了网络结构，并额外增加了辅助模块，即 HINet。通过不同层的特征连接，并通过重新设计的捷径连接和通道注意力模块，充分利用不同层级的特征信息。并通过多重残差连接模块进一步融合所接收到的特征，用于进行自监督单目深度估计。结果表明：本方法与一些经典的方法相比获得了较优的实验结果，并且可以和一些有监督的单目深度估计方法进行比较。

参考文献：

[1] FU H，GONG M M，WANG C H，et al. Deep ordinal regression network for monocular depth estimation[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，June 19-21，2018，Salt Lake City，Utah，USA. NewYork：IEEE，2018：2002-2011.

[2] RANFTL R，LASINGER K，HAFNER D，et al. Towards robust monocular depth estimation：mixing datasets for zero-shot cross-dataset transfer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2022，44(3)：1623-1637.

[3] BHAT S F，ALHASHIM I，WONKA P. Adabins：depth estimation using adaptive bins[C]//2021 IEEE Conference on Computer Visionand Pattern Recognition(CVPR)，June 19-25，2021，Nashville，TN，USA. New York：IEEE，2021：4009-4018.

[4] GODARD C，AODHA O M，BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency [C]//IEEE Computer Vision & Pattern Recognition，July 21-26，2017，Honolulu，HI，USA. New York：IEEE 2017：6602-6611.

[5] ZHOU T，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，July 21-26，2017，Honolulu，HI，USA. New York：IEEE，2017：6612-6619.

[6] GODARD C，AODHA O M，FIRMAN M，et al. Digging into self-supervised monocular depth estimation[C]//IEEE/CVF International Conference on Computer Vision，Octo-

ber 27-November 2, 2019, Seoul, South Korea. New York:IEEE,2019:3828-3838.

[7] GUIZILINI V, AMBRUS R, PILLAI S, et al. PackNet-SfM: 3D Packing for self-supervised monocular depth estimation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. New York:IEEE, 2020:2485-2494.

[8] GUIZILINI V, HOU R, LI J, et al. Semantically-guided representation learning for self-supervised monocular depth [EB/OL]. (2020-02-27)[2021-01-27]. https://arxiv.org/abs/2002.12319.

[9] LUO C X, YANG Z H, WANG P, et al. Every pixel counts++: Joint learning of geometry and motion with 3D holistic understanding[J]. IEEE Transactionson Pattern Analysisand Machine Intelligence, 2018, 45(2):2624-2641.

[10] IBTEHAZ N, RAHMAN M S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation[J]. Neural Networks, 2020, 121(2):74-87.

[11] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016:2818-2826.

[12] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York:IEEE, 2015:1-9.

[13] PENG R, WANG R G, LAI Y L, et al. Excavating the potential capacity of self-supervised monocular depth estimation[C]//IEEE International Conferenceon Computer Vision (ICCV), October 11-17, 2021, Montreal, Canada. Berlin:Springer, 2021:15560-15569.

[14] WATSON J, FIRMAN M, BROSTOW G J, et al. Self-supervised monocular depth hints[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 02, 2019, Seoul, Korea (South). New York: IEEE, 2019:2162-2171.

[15] HIRSCHMOÜER H. Stereo processing by semiglobal matching and mutual information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(2): 328-341.

[16] CASSER V, PIRK S, MAHJOURIAN R, et al. Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos[C]//AAAI Conference on Artificial Intelligence, January 27-February 01, 2019, Honolulu, Hawaii, USA. Menlo Park: AAAI Press, 2019:8001-8008.

[17] GODARD C, AOOHA O M, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York:IEEE, 2017:17355313.

作者简介:

王风华　(1979—),男,博士,讲师,主要从事图像处理与模式识别、嵌入式系统设计方面的研究.