文章编号　1004-924X(2023)20-2993-17

基于层级特征融合的室内自监督单目深度估计

程德强[1]，张华强[1]，寇旗旗[2]，吕　晨[1]，钱建生[1*]
(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 计算机与科学技术学院，江苏 徐州 221116)

摘要：针对目前自监督单目深度估计网络在充斥着大量低纹理、低光照区域的室内复杂场景中存在预测深度信息不精确、物体边缘模糊以及细节丢失严重等问题，本文提出一种基于层级特征融合的室内自监督单目深度估计网络模型。首先，通过映射一致性图像增强模块来处理室内图像，提升低光照区域可见性并且保持亮度一致性，丰富纹理细节，一定程度上解决了训练网络时出现模糊假平面恶化模型的问题。然后，设计结合基于注意力机制的跨层级特征调整模块的深度估计网络，充分融合编码器以及编-解码器多层级特征信息，提升网络的特征利用能力，缩小预测深度与真实深度的语义差距。最后，设计基于图像风格特征的格拉姆矩阵相似性损失函数作为额外的自监督信号约束网络模型，提升网络预测深度的能力，进一步提高了预测深度的精度。在NYU Depth V2 和ScanNet室内数据集上进行训练与测试，正确预测深度像素的比例能够分别达到81.9%和76.0%。实验结果表明，相比现有主要的室内自监督单目深度估计网络，本文网络模型很好地保持了物体边缘和细节信息，有效地提高了预测深度的精度。

关　键　词：自监督；单目深度估计；图像增强；层级特征融合；格拉姆矩阵
中图分类号：TP391.4　　**文献标识码**：A　　**doi**：10.37188/OPE.20233120.2993

Indoor self-supervised monocular depth estimation based on level feature fusion

CHENG Deqiang[1], ZHANG Huaqiang[1], KOU Qiqi[2], LÜ Chen[1], QIAN Jiansheng[1*]

(1. *School of Information and Control Engineering, University of Mining and Technology, Xuzhou* 221116, *China*;
2. *School of Computer Science and Technology, University of Mining and Technology, Xuzhou* 221116, *China*)
* *Corresponding author, E-mail: qianjsh@cumt.edu.cn*

Abstract: Due to a high number of areas with low texture and lighting in complex indoor scenes, current self-supervised monocular depth estimation network models suffer from certain issues. These problems include imprecise depth predictions, noticeable blurriness around object edges in the predictions, and significant loss of details. This paper introduces an indoor self-supervised monocular depth estimation network model based on level feature fusion. First, to enhance the visibility of poorly lit areas and address the issue of pseudo planes deteriorating the model, the Mapping-Consistent Image Enhancement module was applied to process indoor images. This module simultaneously maintained brightness consistency. Subse-

收稿日期：2023-03-01；**修订日期**：2023-04-13.
基金项目：国家自然科学基金资助项目(No. 52204177)；中央高校基本科研业务费专项资金资助项目(No. 2020QN49)

quently, a novel self-supervised monocular depth estimation network model that incorporates the Cross-Level Feature Adjustment module was proposed, utilizing an attention mechanism. This module effectively fused multilevel feature information from the encoder to the decoder, enhancing the network's ability to utilize feature information and reducing the semantic gap between predicted depth and true depth. Finally, the Gram Matrix Similarity Loss function was introduced based on image style features, as an additional self-supervised signal to further constrain the network model. This addition enhanced the network's depth prediction capabilities, leading to improved accuracy. Through training and testing on NYU Depth V2 and ScanNet indoor datasets, this paper achieves a pixel accuracy rate of 81.9% and 76.0%, respectively. The experimental results also include a comparative analysis with existing main indoor self-supervised monocular depth estimation network models. The network model proposed in this paper excels in preserving object edges and details, effectively enhancing the accuracy of predicted depth.

Key words: self-supervision; monocular depth estimation; image enhancement; feature fusion; gram matrix

1 引　言

深度估计作为计算机视觉领域中一个重要的研究方向,在各种3D感知任务中发挥着重要的作用。它在增强现实[1]、虚拟现实、机器人[2]和自动驾驶[3]等领域有着广泛的应用,这些应用目前大多数依靠双目立体视觉[4]、连续多帧图像匹配[5]和激光雷达[6]等方式进行估计,存在计算复杂、相机校准困难以及估计成本较高等问题。与这些方式相比,单目深度估计计算需求小,结构形式简单且开销小,优势明显,备受研究者们的青睐,已经成为目前研究的热点。

从数学的角度来看,单目深度估计就是一个病态问题[7],因为图像实质为物理世界的二维投影,三维场景投影到成像平面时会丢失大量的空间信息,导致单幅图像对应多个可能的实际场景,但是又缺少稳定的物理线索来约束这些不确定性。因此,运用传统的图像处理方法,仅从单幅图像来准确估计出稠密的三维图像十分具有挑战性。直到神经网络的出现,我们才有了很好的解决办法,利用深度卷积神经网络(Convolutional Neural Network,CNN),通过编-解码器的训练网络结构和深度真值标签,目前已经可以得到较高质量的预测深度。但大规模获取深度真值标签难度较大、费用昂贵,因此,不需要深度真值标签作为约束的自监督深度估计已成为一种强有力的替代方法。早期的研究中,GODARD等[8]提出利用立体对当作自监督深度估计的训练范式,引入左右深度一致性损失进行自监督学习。ZHOU等人[9]提出利用位姿网络来估计相邻两帧之间的相对位姿,很好地利用单目图像序列进行训练,减少了立体相机的限制。Sfm-Learner通过视频序列中的连续帧集合进行深度估计网络与位姿网络的联合训练,取得了较好的性能,然而这需要一些理想状态比如视图场景中没有运动物体且不存在遮挡情况,这在现实中通常难以达到。直到MonoDepth2[10]中最小化重投影损失的提出,才很好地解决了物体遮挡问题,同时,提出的应对违反相机运动的自动遮挡损失用于消除运动物体对预测深度的破坏,极大地促进了研究。后来,PackNet-Sfm[11]通过加入额外的语义分割模型来辅助训练深度估计网络。P^2Net[12]采用区别于单个像素、更具辨别力的像素块来计算光度损失,进一步提高了自监督单目深度估计性能。

目前大多数自监督单目深度估计采用端到端的U-Net网络框架,通过压缩和扩张特征信息图以及跳跃连接方式融合编码器和解码器之间的特征信息图来实现图像的特征提取和深度重建工作。尽管现有的自监督模型在室外数据集(如KITTI[13]和Make3D[14])上取得了不错的性能,但是在一些室内数据集(如NYU Depth V2和ScanNet)上表现不佳。究其原因是室内场景复杂多变,充斥着大量低纹理、低光照区域,如光

滑墙壁、天花板和物体背光阴影等，使得自监督信号—光度一致性损失约束力明显下降，造成预测深度图存在深度不精确、物体边缘模糊以及细节丢失等一系列问题。针对上述问题，本文提出了一种基于层级特征融合的室内自监督单目深度估计网络模型，贡献如下：(1)提出了一个能够提升室内低光照区域可见性并且保持亮度一致性的映射一致性图像增强模块，改善了该区域造成估计深度细节缺失，出现假平面恶化深度模型的问题；(2)提出了一个基于注意力机制的跨层级特征调整模块。通过跳跃连接方式，实现了编码器端低层级与高层级以及编-解码器端多层级间的特征信息融合，提高了深度估计网络的特征利用能力，缩小了预测深度与真实深度的语义差距。(3)提出了一个用于室内自监督单目深度估计的格拉姆矩阵相似性损失函数。该损失函数从深度图像的风格特征角度出发，通过计算本文预测深度图和预测平面深度图之间的格拉姆矩阵相似性作为额外的自监督信号约束网络，提升网络预测深度的能力，进一步提高了预测深度图的精度。经过实验验证，本文在 NYU Depth V2 和 ScanNet 室内数据集上获得了优于现有模型的预测深度精度。

2　相关工作

2.1　单目深度估计

利用单幅图像进行深度估计是一项难以完成的工作。传统的单目深度估计只是通过人工手动提取图像特征以及可视化概率模型，其表达能力有限，鲁棒性很差，无力应对室内复杂场景。自从先驱工作[7,15]使用 CNN 直接回归深度以来，已经提出了许多基于 CNN 的单目深度估计方法[16-19]，在基准数据集中产生了令人印象深刻的预测结果，其中大多数是依靠深度真值标签进行训练的监督方法。由于大规模地获取深度真值标签具有挑战性，因此，不需要真值标签的自监督深度学习已成为一种有希望的替代方法。在文献[8]中首次引入图像外观来代替真值标签作为训练网络的监督信号，立体对中的一个图像被预测的深度扭曲到另一个视图，然后计算合成图像与真实图像之间的差异或光度误差用于监督。这种方法进一步扩展到单目深度估计，通过设计网络架构[10]、损失函数[20]以及在线细化[21]等方法，使得自监督单目深度估计在基准数据集上获得了令人满意的效果。然而，现有自监督模型在室内数据集上表现比在室外数据集上逊色很多，原因是室内复杂场景充斥着大量低纹理、低光照区域，导致光度一致性损失变得太弱，无法用来监督深度学习。Zhou 等人[22]通过光流网络的流场监督和稀疏 SURF[23]初始化，提出了一种基于光流的训练范式，取得了很好的效果。最近的 StructDepth[24]充分利用曼哈顿世界模型中的室内结构规律在自监督单目深度学习中进一步取得了优异的效果。尽管它们一定程度地提升了预测精度，但仍存在预测深度信息不够精确、物体边缘模糊和细节丢失严重等问题。此外，StructDepth[24]忽略了室内存在大量低光照、低纹理区域的缺点，导致训练网络时出现模糊假平面恶化深度模型。

2.2　低光照图像增强

图像增强处理是提高图像亮度和对比度的有效方法[25]，对图像的特征提取和降噪有着很大的帮助。目前 Retinex 模型[26]是将图像分解成反射和光照两部分来增强处理。基于直方图均衡化的模型是通过重新调整图像中像素的亮度水平对图像增强处理。最近的研究方法[27]将 Retinex 模型与 CNN 相结合取得了令人印象深刻的结果。对于基于学习的方法[28-29]需要配对数据，但是在图像增强领域获取美学质量较高的配对数据集是比较困难的。为了解决这个难题，研究者们已经努力探索使用非配对的输入[30]或零参考[31]的方法来增强低光照图像，尽管这些方法已经被证明是有效的，但它们仍然没有考虑到视频序列中帧间的亮度对应关系，而这对于自监督单目深度估计训练是必不可少的。直到文献[32]在增强夜间视频序列图像的同时保持了亮度一致性，才很好地解决了帧间亮度对应关系的问题。

2.3　深度估计与位姿网络

目前大多数单目深度估计网络是根据运动重构模型原理在 CNN 网络中同时训练深度估计网络和位姿网络来完成深度估计任务。整个训

练网络的输入为视频序列的连续多帧RGB图像。深度估计网络输入目标图像，经过CNN网络处理输出预测深度图。位姿网络输入目标图像和其上一帧图像，输出相机运动姿态的变化矩阵。根据两个网络的输出结果共同构建重投影图像，然后将重投影误差引入至损失函数，进而结合其他自监督损失函数来反向传播更新模型，迭代优化训练网络输出。目前流行的深度估计网络采用U-Net网络框架，以ResNet[33]、DenseNet[34]和VGG[35]等作为编码器，通过下采样处理来提取图像的特征信息。编码器浅层能够提取图像中颜色、边界以及纹理等较为直接、简单的特征信息。通过加深网络层数和堆叠卷积操作，使感受野变大，其深层就能够提取图像中内在的重要特征信息。解码器将编码器输出的特征信息，通过上采样、反卷积以及跳跃连接等手段[36]，逐步恢复至原图像尺寸。最后使用Sigmoid激活函数映射处理得到预测深度图。位姿网络整体流程与深度估计网络类似，在位姿编码器中使用预训练权重模型，可直接解决相似问题。将预训练模型中第一个卷积核的通道数进行扩展，使网络可以接收六通道信息作为输入，将扩展后的卷积核中的权重缩小一倍，保证卷积操作结束后与单张图像进入网络的数值范围相同，最终输出图像特征。位姿解码器整合提取的图像特征，通过降维、按行并排特征、卷积和图像缩放等手段，最终输出轴角矩阵和平移矩阵来预测出相机位置变化的平移运动和旋转运动。

3　算法与网络模型

本文基于StructDepth[24]的网络模型框架，以ResNet18作为编码器。如图1所示的本文深度估计网络模型，Encoder部分表示编码器网络，Decoder部分表示解码器网络，Skip表示跳跃连接。图中立方体F_1-F_5，V_1-V_5分别表示编码器、解码器各层的特征信息图，上方数字表示该特征的通道数。网络利用映射一致性图像增强(Mapping-Consistent Image Enhancement，MC-IE)模块提升三通道的RGB图像中低光照区域可见性和丰富纹理细节后，送入添加跨层级特征调整(Cross-Level Feature Adjustment，CLFA)模块来实现层级特征融合的编-解码器网络中预测深度(图中CLFA下标表示不同添加位置)，通过在StructDepth[24]原有损失约束基础上利用预测深度图与预测平面深度图之间的格拉姆矩阵相似性损失作为额外的自监督信号约束本文构建的室内自监督单目深度估计网络，最后输出与原尺寸大小一致的单通道预测深度图。本文模型一定程度上解决了室内低光照区域的低能见度对模型带来的不利影响，很好地改善了现有室内自监督模型预测深度不精确，物体边缘模糊和细节丢失等问题。

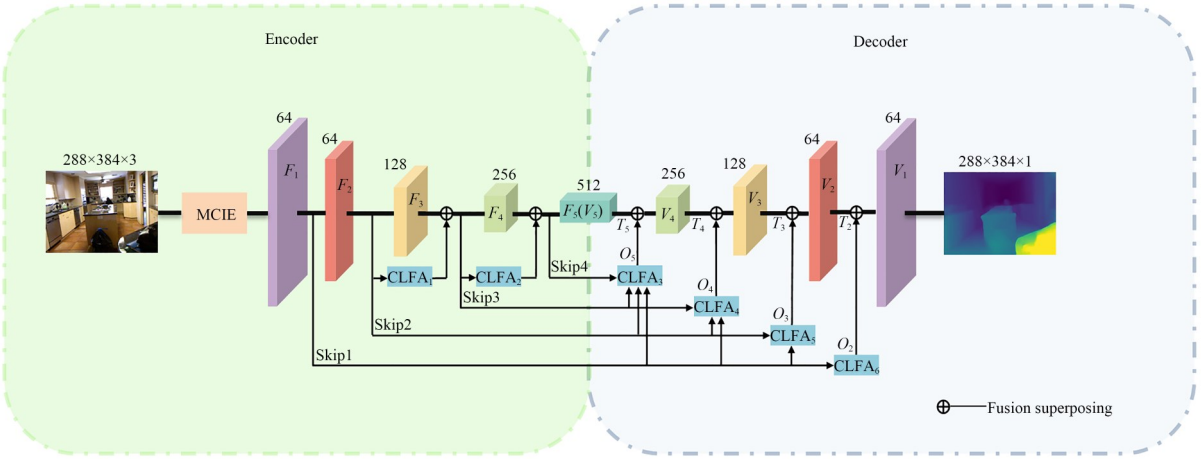

图1　本文深度估计网络模型

Fig. 1　Depth estimation network model in this paper

3.1　映射一致性图像增强

映射一致性图像增强根据对比度有限的直方图均衡化（Contrast Limited Histogram Equalization，CLHE）算法改进而来[32]。其目的在于丰富纹理细节且满足保持亮度一致性的需要，这可以很大程度地提升自监督深度估计算法在低光照区域的预测精度。MCIE模块通过使用亮度映射函数 $b'=\nu(b)$ 并将其应用于增强图像 $\boldsymbol{I}_z$ 和原图像 $\boldsymbol{I}_s$ 来增强图像，计算公式为：

$$\boldsymbol{I}_z'=\nu(\boldsymbol{I}_z),\boldsymbol{I}_s'=\nu(\boldsymbol{I}_s), \tag{1}$$

其中：ν 是亮度单值映射函数，它能够将输入的亮度映射为单个特定输出，这种方式很好地保持了增强图像和原图像之间的亮度一致性。如图2所示的计算 ν 的主要步骤，假设输入图像的频率分布为 $f_b=h(b)$，其中 f_b 是亮度水平 b 的频率。首先，裁剪大于预设参数 μ 的频率（本文预设参数 $\mu=0.12$），以避免噪声信号的放大。其次，如子图(b)所示，被削波的频率被均匀填充到每个亮度级别。最后，通过累积分布函数获得 ν，计算公式为：

$$\nu(b)=\frac{\mathrm{cdf}(b)-\mathrm{cdf}_{\min}}{\mathrm{cdf}_{\max}-\mathrm{cdf}_{\min}}\times(L-1), \tag{2}$$

其中：$\mathrm{cdf}_{\max}$，$\mathrm{cdf}_{\min}$ 分别代表累积分布的最大值和最小值，L 代表亮度级别的数量（RGB图像中通常为256）。如图3所示，MCIE模块为NYU Depth V2，ScanNet室内数据集中低光照区域带来了更高的可见性和更多细节，可以见到在室内低光照区域亮度和对比度显著改善，特别是在红色框内。此外，本文还将MCIE模块应用在计算光度一致性损失时增强图像。

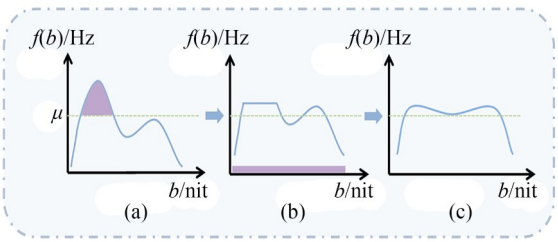

图2　计算亮度单值映射函数 ν 步骤

Fig. 2　Main steps to compute the brightness mapping function ν

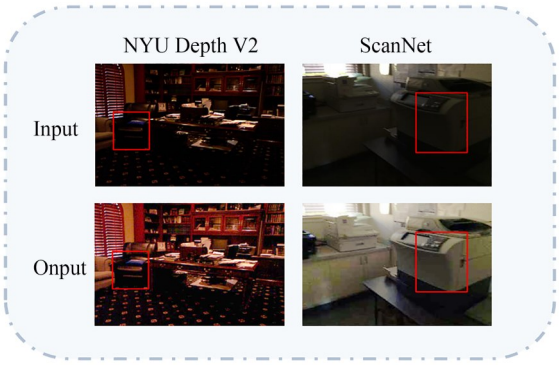

图3　MCIE模块图像增强效果对比

Fig. 3　Comparison of image enhancement effect of MCIE module

3.2　层级特征信息融合

本节共分为3.2.1编码器特征信息融合、3.2.2编-解码器特征信息融合两小节。在3.2.1节中通过跳跃连接方式添加CLFA模块，在下采样提取特征过程中，实现编码器低层级与高层级间的特征信息融合。不仅增强了语义信息表达能力而且增强了编码器输出特征对几何空间位置的内在特征关联性的保持能力。在3.2.2节中同样通过跳跃连接方式添加CLFA模块，在解码上采样过程中实现编码器与解码器层级间的特征信息融合，充分利用了编码器层级中重要的特征信息，缩小了预测深度与真实深度的语义差距。通过编码器、编-解码器层级间特征信息融合，增强了深度估计网络对特征信息的利用能力，改善了特征细节丢失严重等问题，有效地提升了网络预测深度的精度。

3.2.1　编码器特征信息融合

下采样提取特征过程中，特征信息图的分辨率由高到低，通道数由低到高变化。同时，特征信息也由具体逐渐变为抽象，导致室内场景内在几何空间关联性逐渐变弱。为了使编码器在提取特征过程中能够更好地保持原始室内图像中的场景空间关系，提升网络对几何空间位置的内在特征关联性的保持能力，本文在编码器网络中添加如图1中典型双层级形式的特征调整结构。由4.5.1节中表4可知，这种典型双层级形式的特征调整结构实验效果优于图4～图5中所示的典型单层级、三层级以及其他等形式。

通过添加基于注意力机制[37]的CLFA模块，使层级间特征融合，极大地增强了编码器特征信

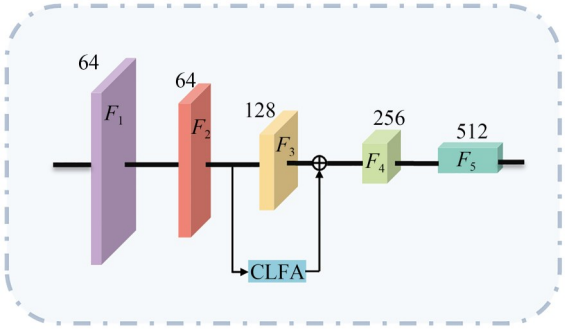

图 4　单层级形式的特征调整结构

Fig. 4　Structure of feature adjustment in single level

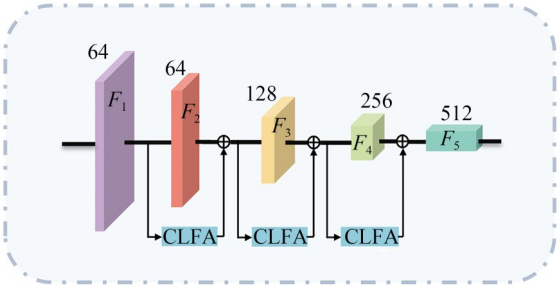

图 5　三层级形式的特征调整结构

Fig. 5　Structure of feature adjustment in three levels

息图中原始室内场景的几何空间位置关联性。CLFA 模块由 Parallel adjustment 和 Attention mechanism 两部分构成，如图 6 所示。Parallel adjustment 部分由两个分支并联组成，以减少直接叠加融合时可能会因为不同层级间的语义差距较大带来负面影响。上分支由一个 1×1 卷积层串联最大池化层组成，1×1 卷积层用来提升输入特征信息的通道数，最大池化层用来缩小此特征信息图尺寸。下分支为一个 3×3 卷积层，两个分支的输出像素叠加生成特征信息图 F。为了使 F 能够在重要的特征通道以及在通道方向上信息聚合多的位置分配到更大的权重，抑制不必要的特征信息，将其送入图 6 中 Attention mechanism 部分。该部分由通道注意力机制（Channel Attention Mechanism，CAM）[38]和空间注意力机制（Spatial Attention Mechanism，SAM）[39]串联组成。将 F 送入 CAM，首先在空间维度上使用最大池化和平均池化操作对 F 处理得到两个新特征块，接着利用共享多层感知机（Multilayer Perceptron，MLP）分别学习特征块中通道特征以及每个通道的重要性程度后将 MLP 的输出直接叠加，最后使用 Sigmoid 激活函数对叠加结果进行映射处理得到通道注意力权重，该权重与 F 相乘结果即为 CAM 的输出 F'，过程计算公式为：

$$F'=\sigma(MLP(AvgPool(F))+MLP(MaxPool(F)))\otimes F, \tag{3}$$

其中：σ 为 Sigmoid 激活函数映射处理。进一步地，将 F' 送入 SAM，首先在通道维度上使用最大池化和平均池化操作对 F' 处理后，拼接得到的两个新特征块。然后使用一个 7×7 卷积层和 Sigmoid 激活函数对拼接结果进行卷积、映射处理得到空间注意力权重，该权重与 F' 相乘结果即为 SAM 的输出 F''，过程计算公式为：

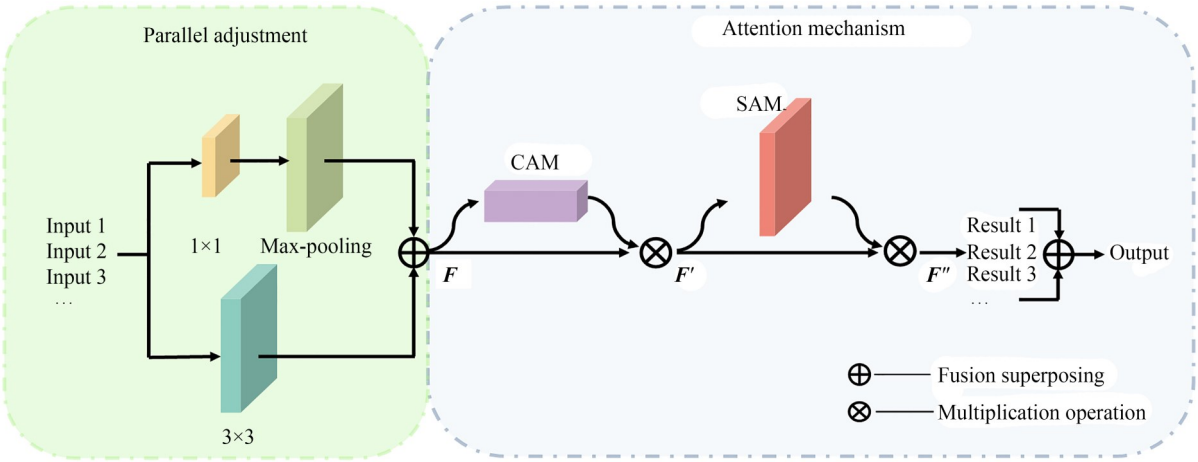

图 6　跨层级特征调整模块

Fig. 6　Cross-Level Feature Adjustment module

$$F''=\sigma(f^{7\times7}([AvgPool(F');MaxPool(F')]))\otimes F', \tag{4}$$

其中：$f^{7\times7}$为上述7×7卷积操作。本文在图1编码器的第二层、第三层添加CLFA模块，拓宽了编码器网络结构。将图像依次通过7×7卷积层、批标准化层、线性整流函数以及最大池化层，使之变成64通道的高分辨率特征信息图F_1，然后送入ResNet18中4个BasicBlock块[33]进行下采样处理，依次得到特征信息图F_2，F_3，F_4和F_5，通道数分别为图1中的64，128，256，512，实现主干特征提取任务。将F_2送入CLFA模块缩小尺寸、提升通道数以及分配特征信息权重后，与F_3像素叠加融合的结果作为下一层输入。同时将此结果送入CLFA模块得到输出后，与F_4像素叠加融合作为下一层输入，以此完成本文深度估计的特征提取任务。

3.2.2　编-解码器特征信息融合

解码器对编码器端输出的多通道、低分辨率的特征信息（本文编码器端输出为9×12×512）通过上采样、反卷积以及跳跃连接方式逐层提高特征信息图的分辨率和降低通道数，最后使用Sigmoid激活函数映射处理获得与原图像尺寸一致的单通道预测深度图。StructDepth[24]通过跳跃连接方式仅将相同层级编-解码器的特征融合，虽一定程度上解决了架构退化和梯度消失的问题，但没有充分利用编码器端的不同层级特征，造成特征信息的浪费，导致解码输出的预测深度图与深度真值仍存在较大语义差距、深度不精确以及细节缺失严重等问题。为了使预测深度更精确，将编码器每一层特征信息采用具有CLFA模块的跳跃连接方式，将这些特征分配不同权重后，按通道方向拼接到解码器层级中融合信息。

如图1所示，解码器与编码器第5层特征信息图V_5，F_5实质为同一特征信息，T_2-T_5，O_2-O_5分别表示解码器F_2-F_5层上采样结果以及CLFA模块输出。解码过程是对输入的特征信息图V_i先使用反卷积操作将通道数缩减至原来一半，再使用双线性插值方法，上采样使其尺寸扩大一倍变成特征信息图T_i。同时，如表1所示关系，将编码器各层级的特征信息图$F_1,F_2,\cdots,F_{i-1}(2\leqslant i\leqslant5)$送入各CLFA模块作为输入（Input），根据图6中特征调整方法以及各输入调整后结果（Result）的通道数总和与T_i通道数保持一致原则，对各输入按比例调整其通道数，然后在通道方向上拼接调整后结果获得模块输出（Output）$O_2,O_3,\cdots,O_i(2\leqslant i\leqslant5)$，最后将$T_i$与通道数相同的$O_i$直接叠加融合结果作为解码器第$i-1$层输出$V_{i-1}$。重复上述步骤，依次进行4次解码处理得到解码器输出，经过Sigmoid函数映射处理获得预测深度图D_t。通过编-解码器多层级特征融合，有效健壮了深度估计网络利用特征的能力，使得预测深度精度更高、内容更丰富以及细节更加清晰。

表1　解码器网络中CLFA模块相关参数关系

Tab. 1　Relationship between CLFA module related parameters in Decoder network

Module	Input	Channels (Input)	Channels (Result)	Output	Channels (Output)	T_i	Channels (T_i)
$CLFA_3$	F_1	64	32	O_5	256	T_5	256
	F_2	64	32				
	F_3	128	64				
	F_4	256	128				
$CLFA_4$	F_1	64	32	O_4	128	T_4	128
	F_2	64	32				
	F_3	128	64				
$CLFA_5$	F_1	64	32	O_3	64	T_3	64
	F_2	64	32				
$CLFA_6$	F_1	64	64	O_2	64	T_2	64

3.3　格拉姆矩阵相似性损失

格拉姆矩阵相似性损失(Gramm Matrix Similarity Loss,GMSL)是受到图像风格迁移[40]启发而来。我们认为图像矩阵中的每一个值都是来自于特定滤波器在特定位置的卷积结果,即每个值代表这个特征信息的强度。Gram Matrix的数学定义解释为特征值之间的偏心协方差矩阵。也就是说,该矩阵实质上表示的是图像中特征信息之间的相关性,同时其对角线能够体现每个特征在图像中出现的频率。因此,Gram Matrix能够用来很好地表达图像的整体风格即图像风格特征。本文将解码输出的预测深度图$\boldsymbol{D}_t$与利用共平面约束预测的平面深度图$\boldsymbol{D}_t^{\text{plane}}$[24]看作是具有不同风格的图像,通过计算两者的格拉姆矩阵相似性损失值,来衡量$\boldsymbol{D}_t$与$\boldsymbol{D}_t^{\text{plane}}$之间的相似性程度,损失值越小,代表两者相似性越高,进一步说明预测的$\boldsymbol{D}_t$更接近于深度真值。GMSL作为一个额外的自监督信号与StructDepth[24]中的自监督信号共同约束本文网络,网络通过前向传播计算损失误差值,然后通过反向传播更新网络中所有权值,直到计算的损失获得最小值且收敛,较好地提升网络预测深度的能力,使预测深度更加精确。

如图7所示的格拉姆矩阵相似性计算步骤:(1)将预测的不同风格深度图$\boldsymbol{D}_t$和$\boldsymbol{D}_t^{\text{plane}}$分别通过同一编码器提取特征(编码器与深度估计网络共享),编码器端输出代表两个深度图的风格特征;(2)根据提取出来的风格特征分别计算两者的Gram Matrix—$\boldsymbol{G}_1$,$\boldsymbol{G}_2$;(3)将$\boldsymbol{G}_1$,$\boldsymbol{G}_2$由多维拉直转换成一维特征向量$f(\boldsymbol{G}_1)$,$f(\boldsymbol{G}_2)$[41];(4)计算两者的余弦相似度。因此,格拉姆矩阵相似性损失约束计算公式为:

$$L_{\text{gram}}=1-\cos\left(f(\boldsymbol{G}(\boldsymbol{D}_t)),f(\boldsymbol{G}(\boldsymbol{D}_t^{\text{plane}}))\right), \tag{5}$$

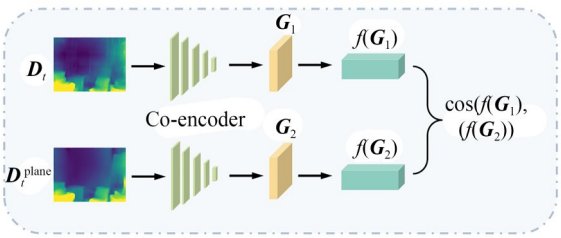

图7　格拉姆矩阵相似性计算步骤

Fig. 7　Steps to calculate Gram matrix similarity

其中:$\boldsymbol{G}$为提取深度图的Gram Matrix处理,f为将多维特征向量转换为一维特征向量的处理,$\cos(\cdot,\cdot)$为余弦相似度计算,定义可以表示为:

$$cos(\alpha,\beta)=(\alpha\cdot\beta)/(\|\alpha\|\cdot\|\beta\|). \tag{6}$$

3.4　损失函数

使用损失函数对模型进行优化在单目深度估计网络中是一项非常重要的环节,越好的损失函数越能大幅提升深度估计模型的性能,当前单目深度估计模型训练采用单目视频序列的训练方式。在单目深度估计中,学习问题被认为是一个视图合成过程。通过使用预测深度图$\boldsymbol{D}_t$和相对姿态$T_{t\to s}$执行重投影处理,能够从每个原图像$\boldsymbol{I}_s$的视点重建目标图像$\boldsymbol{I}_t$。预测深度图$\boldsymbol{D}_t$和相对姿态$T_{t\to s}$分别由两个神经网络$\boldsymbol{D}_t=\phi_d(\boldsymbol{I}_t)$(本文深度估计网络)和$T_{t\to s}=\phi_p(\boldsymbol{I}_t,\boldsymbol{I}_s)$(本文位姿网络)预测得到。此时,可以求得目标图像$\boldsymbol{I}_t$的任意像素点$p_t$和原图像$\boldsymbol{I}_s$对应的点$p_s$,计算公式为:

$$p_s\sim\boldsymbol{K}T_{t\to s}\boldsymbol{D}_t(p_t)\boldsymbol{K}^{-1}p_t, \tag{7}$$

其中:$\boldsymbol{K}$为相机内参矩阵。通过可微的双线性采样[42]操作$s(\cdot,\cdot)$能够从原图像$\boldsymbol{I}_s$中重建出目标图像$\boldsymbol{I}_t$,计算公式为:

$$\hat{\boldsymbol{I}}_t=s(\boldsymbol{I}_s,p_s). \tag{8}$$

模型学习基于上述重建过程,即从原图像重建目标图像,目标是通过优化ϕ_d和ϕ_p以产生更精确的输出来减少重建误差。根据L1损失以及SSIM图像结构相似性[43]来构建光度一致性损失函数L_{pe},计算公式为:

$$L_{\text{pe}}(\boldsymbol{I}_t,\hat{\boldsymbol{I}}_t)=$$

$$\frac{\alpha}{2}(1-SSIM(\boldsymbol{I}_t,\hat{\boldsymbol{I}}_t))+(1-\alpha)\|\boldsymbol{I}_t-\hat{\boldsymbol{I}}_t\|. \tag{9}$$

本文将α设为0.85。此外,本文3.1节中在计算光度一致性损失时使用MCIE模块增强图像,网络的扭曲过程重新定义为:

$$\hat{\boldsymbol{I}}_t'=s(\nu(\boldsymbol{I}_s),p_s). \tag{10}$$

因此,本文光度一致性损失函数变为:

$$L_{\text{pe}}'(\boldsymbol{I}_t',\hat{\boldsymbol{I}}_t')=$$

$$\frac{\alpha}{2}(1-SSIM(\boldsymbol{I}_t',\hat{\boldsymbol{I}}_t'))+(1-\alpha)\|\boldsymbol{I}_t'-\hat{\boldsymbol{I}}_t'\|_1. \tag{11}$$

由于预测时存在着可能的不正确深度,这导致在给定相对姿态$T_{t\to s}$的情况下不能完全正确地重建目标图像,为了更好地解决这种深度模糊问题,本文采用与MonoDepth2[10]一致的逐像素

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。计算公式为：

$$L_{\text{smooth}}=|\partial_x D_t|e^{-|\partial_x I_t|}+|\partial_y D_t|e^{-|\partial_y I_t|}. \quad (12)$$

同时，本文采用了StructDepth[24]中的法向量损失函数L_{nrom}和共平面损失约束L_{plane}，因此总损失函数计算公式为：

$$L=L_{\text{pe}}{}'+\lambda_1 L_{\text{smooth}}+\lambda_2 L_{\text{nrom}}+\lambda_3 L_{\text{plane}}+\lambda_4 L_{\text{gram}}, \quad (13)$$

其中：根据实验效果，$\lambda_1,\lambda_2,\lambda_3,\lambda_4$分别取值为0.001，0.05，0.1，0.17，本文损失函数训练迭代优化曲线如图8所示。

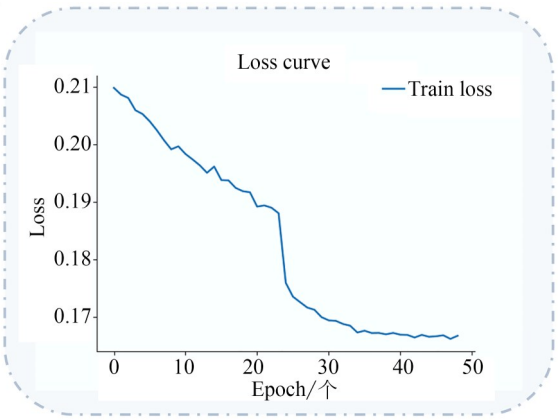

图8　损失函数迭代优化曲线

Fig. 8　Iterative optimization curve of loss function

4　实验结果与分析

4.1　实验数据集

本文使用室内数据集NYU Depth V2[44]，ScanNet[45]训练以及测试。NYU Depth V2数据集是目前最大的RGB-D数据集之一，包含了464个通过使用kinect-V1传感器捕获的室内场景。本文使用了官方所提供的方法对训练集以及测试集进行分割，即使用了249个室内场景作为训练集，余下215个室内场景用作测试集。ScanNet数据集包含1 513个通过使用一个连接在iPad上的深度相机拍摄捕捉的室内场景，包含大约2.5 M的RGB-D视频，本文使用StructDepth[24]提出的分割方法进行实验。

4.2　实验细节与参数设置

本文网络模型在目前主流的深度学习框架PyTorch上搭建的，并在一张内存为24 GB的NVIDIA GeForce GTX 3090 GPU上进行训练与测试，根据网络模型以及GPU性能，将批尺寸(Batch size)设置为18，初始学习率设置为1×10^{-3}。在训练过程中，网络模型采用Adam算法优化，权重衰减参数分别设置为$\beta_1=0.9$，$\beta_2=0.999$，共训练了50个Epoch。

4.3　性能评价指标

为了准确分析模型的精度，本文使用了Eigen中提出的评价指标[7]，这也是当前主流算法采取的深度估计精度评价方法，其方法有：绝对相对误差(Absolute Relative Error，$Abs\,\mathrm{Re}\,l$)、平方相对误差(Square Relative Error，$sq\,\mathrm{Re}\,l$)、均方根误差(Root Mean Square Error，$RMSE$)以及对数空间下的均方根误差(log-root-mean-square error，$RMSE\log$)，计算公式为：

$$RMSE\log=\sqrt{\frac{1}{N}\sum_{i\in N}\|\log(d_i)-\log(d_i^*)\|^2}, \quad (14)$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i\in N}\|d_i-d_i^*\|^2}, \quad (15)$$

$$Abs\,\mathrm{Re}\,l=\frac{1}{N}\sum_{i\in N}\frac{|d_i-d_i^*|}{d_i^*}, \quad (16)$$

$$Sq\,\mathrm{Re}\,l=\frac{1}{N}\sum_{i\in N}\frac{\|d_i-d_i^*\|^2}{d_i^*}, \quad (17)$$

$$Accuracies=\max\left(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i}\right)=\sigma<thr, \quad (18)$$

其中：d_i是像素i的预测深度值，d_i^*表示真实的深度值，N为具有真实值的像素总数，thr为阈值。精度计算是对多个图像中所有的像素点分别计算预测深度与真实深度之比并取最大值，最后将结果赋值给σ。统计σ小于阈值thr的像素点占总像素点的比例即为预测正确率($Accuracies$)，其值越接近于1，预测效果越好。thr一般取值为1.25，1.25^2，1.25^3。

4.4　结果分析

4.4.1　NYU Depth V2数据集结果

为了验证本文模型的有效性，按照4.1节所述方法分割训练集与测试集、4.2节实验参数设置进行实验。在NYU Depth V2数据集上实验结果如表2所示，本文模型与现有主要的几种室内自监督单目深度估计方法进行比较，表中指标数值加粗代表实验最优的结果。

从表2可以看出，相比于其他室内自监督单

表2　本文模型与现有主要方法在NYU Depth V2数据集上的实验结果对比

Tab. 2　Comparison of experimental results between proposed model and existing main methods on NYU Depth V2 dataset

Method	Method of Supervision	Indicator of Error (Lower is better)			Accuracy of prediction(Higher is better)		
		RMSE	Abs Rel	RMS Log10	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
DORN[18]	Supervised	0.509	0.115	0.051	0.828	0.965	0.992
Hu et al.[46]		0.530	0.115	0.050	0.866	0.975	0.993
Yin et al.[47]		0.416	0.108	0.108	0.875	0.976	0.994
AdaBins[48]		0.364	0.103	0.044	0.903	0.984	0.997
Niklaus et al.[49]		0.300	0.080	0.030	0.940	0.990	1.000
MovingIndoor[22]	Self-supervised	0.712	0.208	0.086	0.674	0.900	0.968
TrainFlow[50]		0.686	0.208	0.086	0.701	0.912	0.978
Monodepth2[10]		0.600	0.161	0.068	0.771	0.948	0.987
SC-Depth[51]		0.608	0.159	0.068	0.772	0.939	0.982
P^2Net[12]		0.561	0.150	0.064	0.796	0.948	0.986
P^2Net(5 frames PP)[12]		0.553	0.147	0.062	0.801	0.951	0.987
Bian et al.[52]		0.536	0.147	0.062	0.804	0.950	0.986
PLNet (5 frames)[53]		0.540	0.144	0.061	0.807	0.957	**0.990**
Zhan et al.[54]		0.538	0.143	0.060	0.812	0.951	0.986
StructDepth[24]		0.540	0.142	0.060	0.813	0.954	0.988
Our		**0.530**	**0.138**	**0.059**	**0.819**	**0.959**	**0.990**

目深度估计方法，本文模型在4.3节所述的6个单目深度估计的评价指标中，都达到了目前最优的效果。其中，在衡量单目深度估计效果最重要的指标$\delta<1.25$上，相比StructDepth[24]模型提高了0.6%，RMSE指标下降了1.0%，Abs Rel指标下降了0.4%，RMS Log10指标下降0.1%，说明本文提出的单目深度估计网络模型在现有主要的室内自监督单目深度估计方法中效果最佳、预测的深度更准确。将预测深度可视化，预测深度效果如图9所示，与目前两种主要的室内自监督单目深度估计方法P^2Net[12]和StructDepth[24]比较，本文模型很好地改善了物体边缘模糊、不清晰以及细节缺失的问题，预测深度值更接近于深度真值(Ground truth)，更好地解决了深度模糊的问题。尤其是在图9中红色框内(彩图见期刊电子版)，预测的柜面、墙面拐角、台灯以及花瓶外形轮廓深度更加准确。

4.4.2　ScanNet数据集结果

本文使用在NYU Depth V2数据集上训练的模型来评估推广到其他室内数据集(本文使用ScanNet数据集)的方法，按照4.1节所述方法分割训练集与测试集、4.2节实验参数设置进行实验。在ScanNet数据集上实验结果如表3所示，本文模型与现有主要的几种室内自监督单目深度估计方法进行比较，表中指标数值加粗为最优的结果。

在表3中，本文模型在6个评价指标中，与现有主要的几种室内自监督算法相比，也同样达到了目前最佳预测效果，在最重要的$\delta<1.25$指标上，相比StructDepth[24]模型提高了0.6%，RMSE指标下降了0.9%，Abs Rel指标下降了0.3%，RMS Log10指标下降0.1%。将本文模型和P^2Net[12]、StructDepth[24]模型的深度预测效果可视化如图10所示(彩图见期刊电子版)，本文模型进一步改善了预测深度时物体轮廓模糊、细节丢失等问题。尤其是在图中红色框内，本文预测乒乓球台、椅子等物体轮廓的效果，细节更丰富、更接近真实深度。

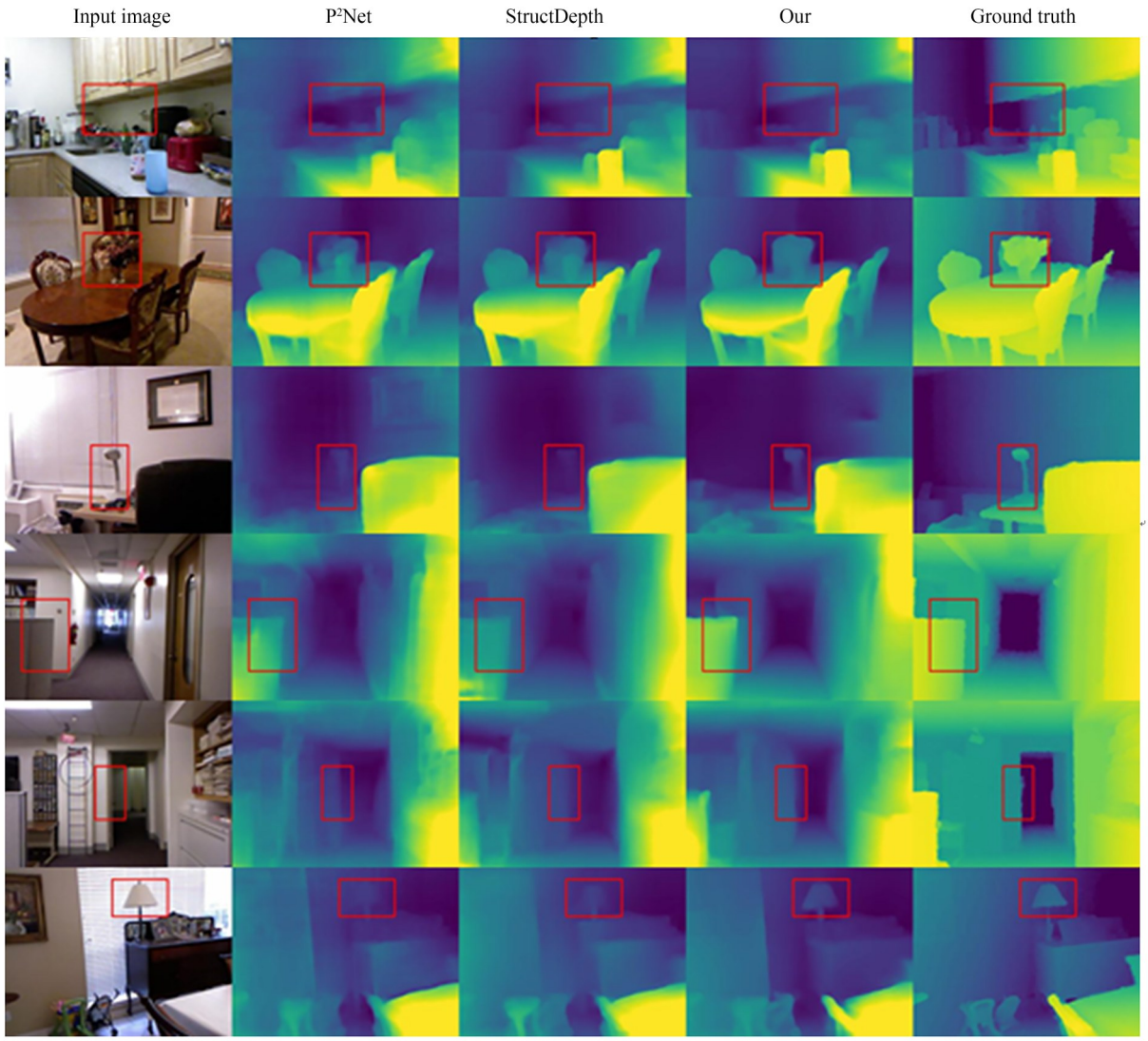

图 9　本文模型与现有主要方法在 NYU Depth V2 数据集上的预测深度图对比

Fig. 9　Comparison of predicted depth maps between proposed model and existing main methods on NYU Depth V2 dataset

表 3　本文模型与现有主要方法在 ScanNet 数据集上的实验结果对比

Tab. 3　Comparison of experimental results between the model in this paper and existing main methods on ScanNet dataset

Method	Method of Supervision	Indicator of Error (Lower is better)			Accuracy of prediction (Higher is better)		
		RMSE	Abs Rel	RMS Log10	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
MovingIndoor[22]	Self-supervised	0.483	0.212	0.088	0.650	0.905	0.976
Monodepth2[10]		0.451	0.1911	0.080	0.693	0.926	0.983
P²Net[12]		0.420	0.175	0.074	0.740	0.932	0.982
P²Net-finetune[24]		0.412	0.172	0.073	0.743	0.935	0.984
StructDepth[24]		0.400	0.165	0.070	0.754	0.939	0.985
Our		**0.391**	**0.162**	**0.069**	**0.760**	**0.946**	**0.987**

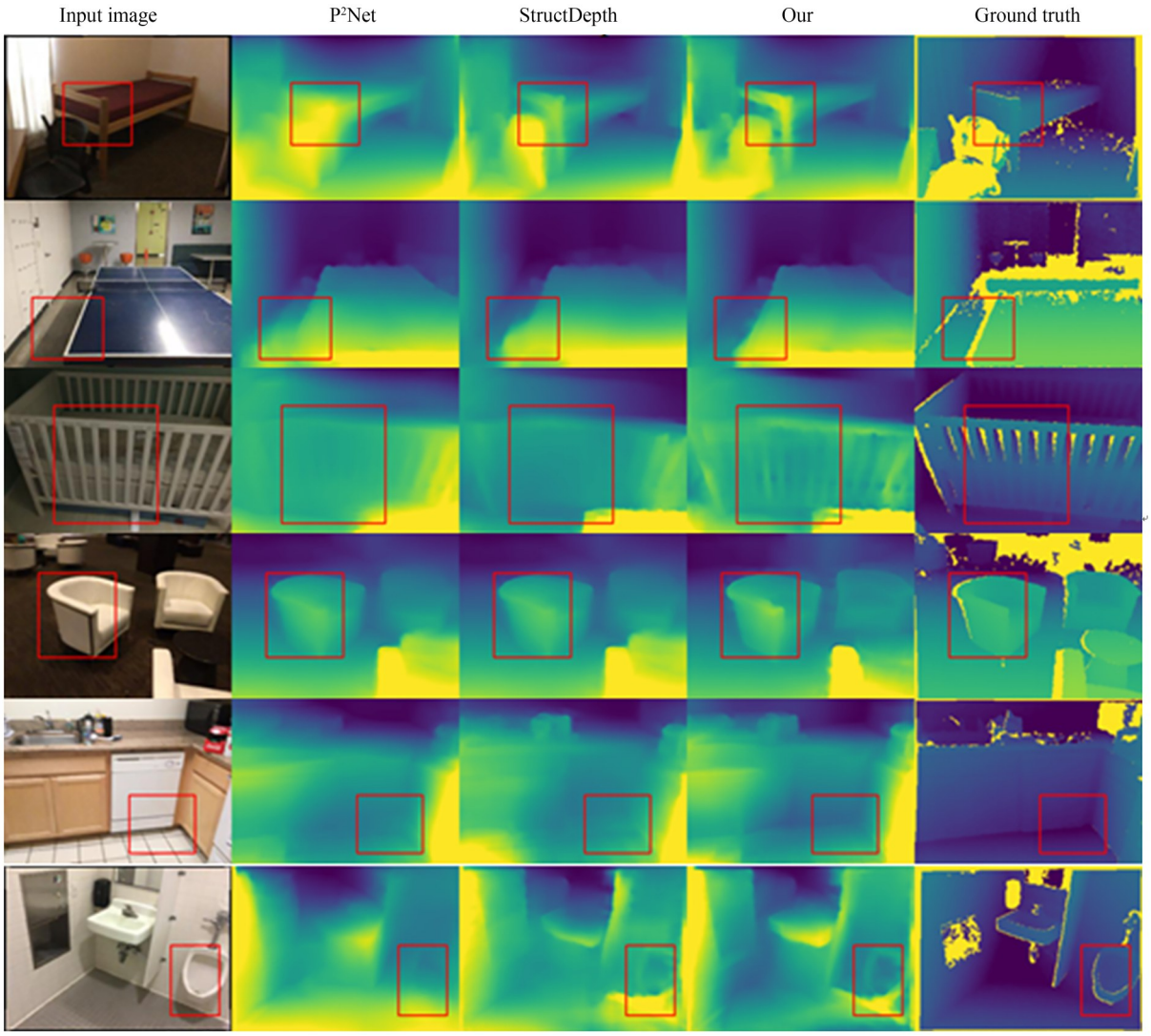

图 10　本文模型与现有主要方法 ScanNet 数据集上的预测深度图对比

Fig. 10　Comparison of predicted depth maps between proposed model and existing main methods on ScanNet dataset

4.5　消融实验

4.5.1　跨层级形式消融实验

为了验证模块“CLFA”的不同添加形式对本文室内自监督单目深度估计网络的性能提升效果，本文在编-解码器层级融合结构形式不变的情况下，进一步地对编码器层级不同融合结构形式进行了消融实验。如表 4 所示，方法 Single Level，Double Levels 以及 Three Levels 分别表示在编码器中同时添加仅跨单层级、跨双层级、跨三层级的结构形式。位置 F_1，F_2，F_3，F_4 分别表示跨编码器的 F_1，F_2，F_3，F_4 层级添加 CLFA 模块。图 1 表示同时跨 F_3，F_4 双层级添加 CLFA 模块，图 4 表示仅跨 F_3 单层级形式，图 5 表示同时跨 F_2，F_3，F_4 三层级形式，表中指标数值加粗为实验最优结果。从表 4 可以看出，并不是任意形式添加 CLFA 模块就能提升网络性能，添加仅跨 F_1，F_2，F_3 单层级融合形式在最重要的 $\delta<1.25$ 指标上，相比表中 Baseline 方法 StructDepth[24] 模型实验指标反而下降了 0.4%，0.2%，0.1%。相对于添加跨单层级、三层级形式，跨双层级形式的实验效果更好，其中同时跨 F_3，F_4 双层级融合形式实验效果最佳，在 $\delta<1.25$ 指标上能够提高 0.4%。不同添加形式的实验效果不一，出现表中实验结果，其可能原因是一方面跨过多的层级会干扰高维特征图的输出，另一方面过低的层级不能保证有效信息的导入，而实验效果更好的双层级融合形式可以在低维干扰和信息导入之间达到平衡状态。根据实验结果，添加仅

表 4　NYU Depth V2 数据集上多种跨层级形式的特征融合结构消融实验

Tab. 4　Ablation experiment of Several Cross-Level feature fusion structure on NYU Depth V2 dataset

Method	Position	Indicator of Error (Lower is better)			Accuracy of prediction (Higher is better)		
		RMSE	Abs Rel	RMS Log10	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Baseline	/	0.540	0.142	**0.060**	0.813	0.954	0.988
Single Level	F_1	0.543	0.143	0.061	0.810	0.953	0.988
	F_2	0.543	0.142	0.061	0.811	0.954	0.988
	F_3	0.540	0.142	0.061	0.812	0.954	0.988
	F_4	0.540	0.141	**0.060**	0.814	0.955	**0.989**
Double Levels	F_2,F_3	0.540	0.140	0.061	0.815	**0.956**	0.988
	F_2,F_4	0.539	0.140	**0.060**	0.816	0.955	**0.989**
	F_3,F_4	**0.537**	**0.139**	**0.060**	**0.817**	0.955	**0.989**
Three Levels	F_2,F_3,F_4	0.540	0.140	**0.060**	0.815	0.955	0.988

跨单层级形式时，仅跨 F_1 层级的实验效果最差；添加跨双层级形式时，其实验效果与分别仅跨单层级形式时效果的叠加呈正相关关系；添加跨三层级形式时，实验效果不如跨双层级形式。综上所述，添加含有 F_1 这种最低层级的跨双层级、三层级形式时最容易干扰正确特征输出，其实验效果不如本文最终添加的跨 F_3，F_4 双层级形式表现优秀，因此本文只对仅跨 F_1 层级进行了实验。

4.5.2　创新模块消融实验

为了验证本文提出的各创新模块对室内自监督单目深度估计的有效性，本文在 NYU Depth V2 数据集上进行了消融实验。如表 5 所示，模块“CLFA”，“GMSL”和“MCIE”分别对应添加跨层级特征调整模块、格拉姆矩阵相似性损失函数和映射一致性图像增强模块。表中√表示实验中添加此创新模块，×表示没有添加此创新模块，表中指标数值加粗表示实验最优的结果。实验结果表明，在最重要的 $\delta<1.25$ 指标上，模块“CLFA”相比表中 Baseline 方法—Struct-Depth[24] 模型实验效果提高了 0.4%，模块“GMSL”提高了 0.3%，模块“MCIE”提高了 0.2%，本文所提出的每个创新模块均能够不同程度地提升预测深度的正确率。并且，不论是其创新模块两两组合添加还是三者组合添加都能优于单独添加时的实验效果，其中三者组合添加实验效果最佳，在最重要的 $\delta<1.25$ 指标上提高了 0.6%。

表 5　NYU Depth V2 数据集上不同创新模块的消融实验

Tab. 5　Ablation experiment of different innovative modules on NYU Depth V2 dataset

Method	CLFA	GMSL	MCIE	Indicator of Error (Lower is better)			Accuracy of prediction (Higher is better)		
				RMSE	Abs Rel	RMS Log10	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Baseline	×	×	×	0.540	0.142	0.060	0.813	0.954	0.988
1	√	×	×	0.537	0.139	0.060	0.817	0.955	0.989
2	×	√	×	0.536	0.140	0.060	0.816	0.956	0.989
3	×	×	√	0.539	0.141	0.061	0.815	0.955	0.988
4	√	√	×	0.533	**0.138**	**0.059**	0.818	0.957	**0.990**
5	√	×	√	0.536	0.139	**0.059**	0.817	0.956	0.989
6	×	√	√	0.537	0.139	0.060	0.817	0.955	0.988
7	√	√	√	**0.530**	**0.138**	**0.059**	**0.819**	**0.959**	**0.990**

5 结　论

针对目前室内自监督单目深度估计网络预测的深度图像存在深度不准确、物体边缘模糊以及细节丢失严重等问题，本文提出了基于层级特征融合的室内自监督单目深度估计的网络模型。该网络模型通过引入映射一致性图像增强模块，提高了室内低光照区域可见性，改善了预测深度时出现细节丢失和模糊假平面恶化模型的问题。同时，在编码器和编-解码器中仔细设计含有基于注意力机制的跨层级特征调整模块的跳跃连接形式，实现了多层级特征信息融合，提高了深度估计网络的特征利用能力，缩小了预测深度与真实深度的语义差距。在网络自监督约束方面，提出图像格拉姆矩阵相似性损失作为额外的自监督信号进行损失约束，进一步增加了预测深度的准确性。本文在NYU Depth V2和ScanNet室内数据集上训练与测试，正确预测深度像素的比例分别能够达到81.9%和76.0%。实验结果表明，在相同的评价指标下，本文模型在所对比现有的室内自监督单目深度估计方法中性能最佳，更好地保留了物体细节信息，改善了预测物体边缘模糊问题，提高了预测深度的精度，从而有效增强了单目相机对室内复杂场景的三维理解能力。虽然本文模型在对比现有方法中指标得到了提升，但是仍存在边缘模糊、细节丢失等问题，与真实深度仍有差距。因此，下一步工作将利用结构相似度参数与图像梯度相结合作为评价指标来评判预测深度的优劣，以进一步改善预测深度物体边缘模糊、语义差距较大的现状。

参考文献：

[1] ZHANG S G, CHEN Y F, ZHANG L J, *et al*. Study on robot grasping system of SSVEP-BCI based on augmented reality stimulus[J]. *Tsinghua Science and Technology*, 2022, 28(2): 322-329.

[2] LIN D H, FIDLER S, URTASUN R. Holistic Scene Understanding for 3D Object Detection with RGBD Cameras[C]. 2013 *IEEE International Conference on Computer Vision*. 1-8, 2013, *Sydney*, *NSW*, *Australia*. IEEE, 2014: 1417-1424.

[3] RASOULI A, TSOTSOS J K. Autonomous vehicles that interact with pedestrians: a survey of theory and practice[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2019, 21(3): 900-918.

[4] KHAMIS S, FANELLO S, RHEMANN C, *et al*. StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction [EB/OL]. 2018: *arXiv*: 1807.08865. https://arxiv.org/abs/1807.08865.pdf

[5] RANFTL R, VINEET V, CHEN Q F, *et al*. Dense Monocular Depth Estimation in Complex Dynamic Scenes[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 27-30, 2016, *Las Vegas*, *NV*, *USA*. IEEE, 2016: 4058-4066.

[6] 伍锡如，薛其威. 基于激光雷达的无人驾驶系统三维车辆检测[J]. 光学 精密工程，2022，30(4)：489-497.
WU X R, XUE Q W. 3D vehicle detection for unmanned driving systerm based on lidar[J]. *Opt. Precision Eng.*, 2022, 30(4): 489-497. (in Chinese)

[7] 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 - Volume 2. December 8 - 13, 2014, Montreal, Canada*. New York: ACM, 2014: 2366-2374.

[8] GODARD C, AODHA OMAC, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 6602-6611.

[9] 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*). 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 6612-6619.

[10] GODARD C, AODHA OMAC, FIRMAN M, *et al*. Digging into self-supervised monocular depth estimation [C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October 27-November 2*, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 3827-3837.

[11] GUIZILINI V, AMBRUS R, PILLAI S, *et al*. 3D Packing for self-supervised monocular depth estimation [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 2482-2491.

[12] YU Z H, JIN L, GAO S H. *P2Net: Patch-Match and Plane-Regularization for Unsupervised Indoor Depth Estimation* [M]. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 206-222.

[13] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]. 2012 *IEEE Conference on Computer Vision and Pattern Recognition*. 16-21, 2012, *Providence*, *RI*, *USA*. IEEE, 2012: 3354-3361.

[14] SAXENA A, SUN M, NG A Y. Make3D: learning 3D scene structure from a single still image [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 31(5): 824-840.

[15] 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*). 7-13, 2015, *Santiago*, *Chile*. IEEE, 2016: 2650-2658.

[16] LIU F Y, SHEN C H, LIN G S, *et al*. Learning depth from single monocular images using deep convolutional neural fields [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(10): 2024-2039.

[17] KIM S, PARK K, SOHN K, *et al*. *Unified Depth Prediction and Intrinsic Image Decomposition From A Single Image Via Joint Convolutional Neural Fields* [M]. Computer Vision - ECCV 2016. Cham: Springer International Publishing, 2016: 143-159.

[18] FU H, GONG M M, WANG C H, *et al*. Deep ordinal regression network for monocular depth estimation [C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 2002-2011.

[19] WOFK D, 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*). 20-24, 2019, *Montreal*, *QC*, *Canada*. IEEE, 2019: 6101-6108.

[20] SHU C, YU K, DUAN Z X, *et al*. *Feature-Metric Loss for Self-Supervised Learning of Depth and Egomotion* [M]. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 572-588.

[21] CHEN Y H, SCHMID C, SMINCHISESCU C. Self-Supervised learning with geometric constraints in monocular video: connecting flow, depth, and camera [C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27-*November* 2, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 7062-7071.

[22] ZHOU J S, WANG Y W, QIN K H, *et al*. Moving indoor: unsupervised video depth learning in challenging environments [C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27-*November* 2, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 8617-8626.

[23] BAY H, TUYTELAARS T, VAN GOOL L. *SURF: Speeded Up Robust Features* [M]. Computer Vision-ECCV 2006. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 404-417.

[24] LI B Y, HUANG Y, LIU Z Y, *et al*. Structdepth: leveraging the structural regularities for self-supervised indoor depth estimation [C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 12643-12653.

[25] CHENG D Q, CHEN L L, LV C, *et al*. Light-guided and cross-fusion U-net for anti-illumination image super-resolution [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, 32(12): 8436-8449.

[26] 黄慧，董林鹭，刘小芳，等. 改进Retinex的低光照图像增强 [J]. 光学 精密工程，2020，28(8)：1835-1849.
HUANG H, DONG L L, LIU X F, *et al*. Improved retinex low light image enhancement method [J]. *Opt. Precision Eng.*, 2020, 28(8): 1835-1849. (in Chinese)

[27] ZHANG Y H, ZHANG J W, GUO X J. Kindling the darkness: a practical low-light image enhancer [C]. *Proceedings of the 27th ACM International Conference on Multimedia*. *October* 21-25, 2019,

Nice, *France*. New York: ACM, 2019: 1632-1640.

[28] CAI J R, GU S H, ZHANG L. Learning a deep single image contrast enhancer from multi-exposure images[J]. *IEEE Transactions on Image Processing*, 2018, 27(4): 2049-2062.

[29] CHEN C, CHEN Q F, XU J, *et al*. Learning to see in the dark[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 3291-3300.

[30] JIANG Y F, GONG X Y, LIU D, *et al*. EnlightenGAN: deep light enhancement without paired supervision[J]. *IEEE Transactions on Image Processing*, 2021, 30: 2340-2349.

[31] GUO C L, LI C Y, GUO J C, *et al*. Zero-reference deep curve estimation for low-light image enhancement[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 1777-1786.

[32] WANG K, ZHANG Z Y, YAN Z Q, *et al*. Regularizing nighttime weirdness: efficient self-supervised monocular depth estimation in the dark[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 16035-16044.

[33] HE K M, ZHANG X Y, REN S Q, *et al*. Deep residual learning for image recognition[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 27-30, 2016, *Las Vegas*, *NV*, *USA*. IEEE, 2016: 770-778.

[34] HUANG G, LIU Z, VAN DER MAATEN L, *et al*. Densely connected convolutional networks[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 2261-2269.

[35] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[EB/OL]. 2014: *arXiv*: 1409.1556. https://arxiv.org/abs/1409.1556.pdf

[36] 程德强, 赵佳敏, 寇旗旗, 等. 多尺度密集特征融合的图像超分辨率重建[J]. 光学 精密工程, 2022, 30(20): 2489-2500.
CHENG D Q, ZHAO J M, KOU Q Q, *et al*. Multi-scale dense feature fusion network for image super-resolution[J]. *Opt. Precision Eng.*, 2022, 30(20): 2489-2500. (in Chinese)

[37] 程德强, 陈杰, 寇旗旗, 等. 融合层次特征和注意力机制的轻量化矿井图像超分辨率重建方法[J]. 仪器仪表学报, 2022, 43(8): 73-84.
CHENG D Q, CHEN J, KOU Q Q, *et al*. Lightweight super-resolution reconstruction method based on hierarchical features fusion and attention mechanism for mine image[J]. *Chinese Journal of Scientific Instrument*, 2022, 43(8): 73-84. (in Chinese)

[38] 蔡体健, 彭潇雨, 石亚鹏, 等. 通道注意力与残差级联的图像超分辨率重建[J]. 光学 精密工程, 2021, 29(1): 142-151.
CAI T J, PENG X Y, SHI Y P, *et al*. Channel attention and residual concatenation network for image super-resolution[J]. *Opt. Precision Eng.*, 2021, 29(1): 142-151. (in Chinese)

[39] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 7132-7141.

[40] GHIASI G, LEE H, KUDLUR M, *et al*. Exploring The Structure of a Real-Time, Arbitrary Neural Artistic Stylization Network[EB/OL]. 2017: *arXiv*: 1705.06830. https://arxiv.org/abs/1705.06830.pdf

[41] LIU L N, SONG X B, WANG M M, *et al*. Self-supervised monocular depth estimation for all day images using domain separation[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 12717-12726.

[42] JADERBERG M, SIMONYAN K, ZISSERMAN A, *et al*. Spatial Transformer Networks[EB/OL]. 2015: *arXiv*: 1506.02025. https://arxiv.org/abs/1506.02025.pdf

[43] 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: a Publication of the IEEE Signal Processing Society*, 2004, 13(4): 600-612.

[44] SILBERMAN N, HOIEM D, KOHLI P, *et al*. *Indoor Segmentation and Support Inference from RGBD Images*[M]. Computer Vision - ECCV 2012. Berlin, Heidelberg: Springer Berlin Heidel-

berg, 2012: 746-760.

[45] DAI A, CHANG A X, SAVVA M, *et al*. ScanNet: richly-annotated 3d reconstructions of indoor scenes[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 2432-2443.

[46] HU J J, OZAY M, ZHANG Y, *et al*. Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries[C]. 2019 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*). 7-11, 2019, *Waikoloa*, *HI*, *USA*. IEEE, 2019: 1043-1051.

[47] YIN W, LIU Y F, SHEN C H, *et al*. Enforcing geometric constraints of virtual normal for depth prediction [C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27-*November* 2, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 5683-5692.

[48] FAROOQ BHAT S, ALHASHIM I, WONKA P. AdaBins: depth estimation using adaptive bins [C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 4008-4017.

[49] NIKLAUS S, MAI L, YANG J M, *et al*. 3D Ken Burns effect from a single image[J]. *ACM Transactions on Graphics*, 38(6): 1-15.

[50] ZHAO W, LIU S H, SHU Y Z, *et al*. Towards better generalization: joint depth-pose learning without PoseNet[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 9148-9158.

[51] BIAN J W, ZHAN H Y, WANG N Y, *et al*. Unsupervised scale-consistent depth learning from video[J]. *International Journal of Computer Vision*, 2021, 129(9): 2548-2564.

[52] BIAN JW, ZHAN H, WANG N, *et al*. Unsupervised depth learning in challenging indoor video: Weak rectification to rescue [J]. *arXiv preprint arXiv*:2006.02708, 2020.

[53] JIANG H L, DING L Y, HU J J, *et al*. PLNet: plane and line priors for unsupervised indoor depth estimation[C]. 2021 *International Conference on 3D Vision* (*3DV*). 1-3, 2021, *London*, *United Kingdom*. IEEE, 2022: 741-750.

[54] BIAN J W, ZHAN H Y, WANG N Y, *et al*. Auto-rectify network for unsupervised indoor depth estimation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(12): 9802-9813.

作者简介：

程德强(1979－)，男，河南洛阳人，博士，教授，博士生导师。主要研究方向为图像智能检测与模式识别，图像处理与视频编码。

E-mail: chengdq@cumt.edu.cn

通讯作者：

钱建生(1964－)，男，浙江桐乡人，教授，博士生导师。主要从事：智能化信息处理、人工智能的技术研究。

E-mail: qianjsh@cumt.edu.cn