基于密集特征融合的无监督单目深度估计

陈 莹* 王一良

(江南大学轻工过程先进控制教育部重点实验室 无锡 214122)

摘 要：针对无监督单目深度估计生成深度图质量低、边界模糊、伪影过多等问题，该文提出基于密集特征融合的深度网络编解码结构。设计密集特征融合层(DFFL)并将其以密集连接的形式填充U型编解码器，同时精简编码器部分，实现编、解码器的性能均衡。在训练过程中，将校正后的双目图像输入给网络，以重构视图的相似性约束网络生成视差图。测试时，根据已知的相机基线距离与焦距将生成的视差图转换为深度图。在KITTI数据集上的实验结果表明，该方法在预测精度和误差值上优于现有的算法。

关键词：深度估计；无监督；密集特征融合层；编解码器

中图分类号：TN911.73; TP391 **文献标识码**：A **文章编号**：1009-5896(2021)10-2976-09

DOI: 10.11999/JEIT200590

Unsupervised Monocular Depth Estimation Based on Dense Feature Fusion

CHEN Ying WANG Yiliang

(*Key Laboratory of Advanced Process Control for Light Industry* (*Ministry of Education*), *Jiangnan University*, *Wuxi* 214122, *China*)

Abstract: In view of the problems of low quality, blurred borders and excessive artifacts generated by unsupervised monocular depth estimation, a deep network encoder-decoder structure based on dense feature fusion is proposed. A Dense Feature Fusion Layer(DFFL) is designed and it is filled with U-shaped encoder-decoder in the form of dense connection, while simplifying the encoder part to achieve a balanced performance of the encoder and decoder. During the training process, the calibrated stereo pair is input to the network to constrain the network to generate disparity maps by the similarity of reconstructed views. During the test process, the generated disparity map is converted into a depth map based on the known camera baseline distance and focal length. The experimental results on the KITTI data set show that this method is superior to the existing algorithms in terms of prediction accuracy and error value.

Key words: Depth estimation; Unsupervised learning; Dense Feature Fusion Layer(DFFL); Encoder-decoder

1 引言

从单张2维图片中恢复深度信息是计算机视觉领域的重要课题。利用深度信息可以有效地重建场景的3维结构，在自动驾驶、虚拟现实、视觉SLAM等领域有着广泛的应用前景。在过去的研究中，对深度的预测依赖运动推断结构(Structure From Motion, SFM)[1]、双目或多视角几何(binocular or multi-view stereo)[2]等shape-from-X算法。这些传统算法通常都需要一定的限制条件，比如需要使用多个视角或连续的图片帧序列，不同的光照条件，亦或是已知的纹理特性。此外，传统算法往往依赖图像间的特征匹配，而这些特征算子是人工设计的，因此其应用场景是受限的，没有很好的鲁棒性。基于深度学习的单目深度估计直接为单张图片的每一个像素点预测其对应的深度值，解决了传统算法的约束条件，同时也带来了新的问题。从单张RGB图片恢复对应的3维结构是一个不适定的问题，可以有很多解符合要求。但值得注意的是，人类从日常生活中的不断训练中获得了从单目视觉中推理深度线索的能力，例如，物体的相对大小、纹理信息、物体之间的遮挡、视觉的透视效果等等。

基于卷积神经网络的单目深度估计，采取了和人类获取深度线索相似的训练过程。网络通过对数

收稿日期：2020-07-17；改回日期：2020-12-29；网络出版：2021-02-03
*通信作者： 陈莹 chenying@jiangnan.edu.cn
基金项目：国家自然科学基金(61573168)
Foundation Item: The National Natural Science Foundation of China (61573168)

据的不断学习，利用多层的卷积和非线性激活单元，提取出非常抽象的特征，这些抽象特征帮助网络推理当前场景的深度信息，抽象特征的提取和人类获取深度线索的过程是相似的。Eigen等人[3]首先提出了利用全局粗尺度和局部细尺度，两种尺度的网络估计逐像素的深度值。Liu等人[4]引入了条件随机场(Conditional Random Fields, CRFs)来提高预测精度。Laina等人[5]受到ResNet[6]的启发提出了基于残差的全卷积网络来预测深度，得益于ResNet的优异性能，预测精度得到很大的提高。周武杰等人[7]加入了金字塔池化模块增强网络的特征融合能力。Zhao等人[8]利用合成的虚拟深度数据集结合真实的深度数据集，利用生成对抗网络(Generative Adversarial Network, GAN)做真实数据与合成数据之间的风格迁移，增高精度的同时减少了网络对于真实数据集的需求。上述的方法均是有监督的，依赖大规模、高精度、逐像素对齐的彩色图和深度图。

近年来，一些全新的无监督算法的提出，尝试去处理数据对网络的限制。无监督的算法总共分为两种思路：(1)基于连续时间的图像序列。在这种结构中，网络需要同时预测深度和相机姿态。Zhou等人[9]在训练深度估计网络的同时，独立地训练了相机姿态估计网络。基于时间序列的无监督算法只能在刚性场景下成立，当运动物体与相机速度保持一致或者相机静止时，会使网络预测出无穷大深度值的“空洞”。为了避免对刚性运动假设的破坏，Zhou等人又附加了可解释性的掩膜来处理有问题的区域。(2)基于双目图像对。这种设计需要用到校正的双目图像输入给网络预测出对应的视差图，再用得到的视差图对双目图像进行重建，将深度估计问题转换成图像重建问题。在已知两相机的基线距离与焦距的情况下，就可以通过视差推导出深度。Garg等人[10]首次利用这样的思路设计了无监督的深度估计网络。 Godard等人[11]加入了左右一致性项来约束网络的输出，获得了更高的精度。但是随着网络提取出的特征越来越复杂，特征图的分辨率也在不断下降，使得网络难以恢复清晰的深度边界。

受到Zhou等人[12]的启发，本文在文献[11]的编解码器网络基础上设计并引入了全新的密集特征融合层DFFL。在提高网络预测精度的同时减少了网络的参数量。首先，通过将DFFL以密集连接的形式放置在一般的编解码器结构中，实现了各级解码器之间的信息互通，提高了不同层次特征的复用率，恢复出更精细的图像细节；其次，考虑到无监督深度估计的精度不仅仅取决于编码器提取抽象特征的能力，也取决于如何合理利用所得到的不同层次的特征，论文设计编码器的修剪策略，使得编码器、解码器的性能更加匹配，合理的裁剪加快了网络的预测速度并且提高了预测的精度。实验证明，本文设计出的网络在KITTI驾驶数据集[13]上的表现优于现有的算法。

2　基于图像重建的无监督深度估计及问题分析

2.1　基于图像重建的无监督深度估计

无监督的核心思想是不使用RGB图像与其对应的真实深度图作为训练的监督信号。为了使网络具有估计深度的能力，就必须找到一种与深度有关并且可以获得的替代监督信号。在双目视觉中，视差与深度成反比，经过校准的双目相机，其左右视图的视差是可以通过匹配对应点来获得的。对于传统方法，精确地匹配对应点是非常困难的，但是这项工作非常适合卷积神经网络。网络以重建前后的左右视图外观相似性与左右视差图的一致性为约束条件，促使网络生成正确的左右视差图。算法流程如图1所示。

首先，从已经校准的双目相机获取同一时刻的左右视图$\boldsymbol{I}^{\mathrm{l}}$, $\boldsymbol{I}^{\mathrm{r}}$[14]，通过网络预测出对应的左右视差图$\boldsymbol{d}^{\mathrm{l}}$, $\boldsymbol{d}^{\mathrm{r}}$。以左视差图$\boldsymbol{d}^{\mathrm{l}}$为例，根据每一个视差值在右视图$\boldsymbol{I}^{\mathrm{r}}$进行检索，将检索到的RGB信息返回并填充获得重构的左视图$\tilde{\boldsymbol{I}}^{\mathrm{l}}$，右视图$\tilde{\boldsymbol{I}}^{r}$的重构方法是完全相同的。在测试过程中，只需要单目视图作为输入，在已知双目相机的基线距离b和相机焦距f的条件下，根据网络的预测视差d，可以通过公式$\hat{d}=bf/d$获得预测的深度$\hat{d}$。

2.2　卷积神经网络设计中的问题分析

无监督的单目深度估计以单张视图作为卷积神经网络的输入，预测出左右两张视差图，属于图片到图片(image to image)转换的问题。该问题通常使用编解码器结构来解决，其中以U-Net[15]为代表的U型编解码器结构最为常用。如图2(a)所示，U-Net的跳转连接在一定程度上补充了编码过程中丢失的图像细节，但是仅仅在同一层编、解码器之间使用跳转连接对于特征的使用是不充分的。为解决这个问题，U-Net++[12]将U型结构中间的空缺填满，将融合上下文信息后的特征补充给解码器，其网络拓扑如图2(b)。

针对无监督单目深度估计生成的深度图比较模糊、边界不清晰等问题，本文对U-Net++的融合策略进行改造。如图2(c)所示，本文在特征融合时，使用反卷积代替双线性插值，具有学习性的上采样

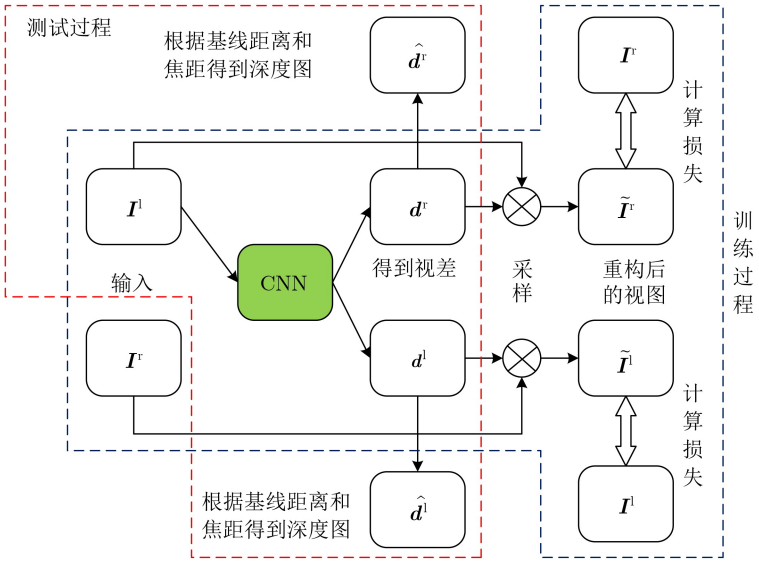
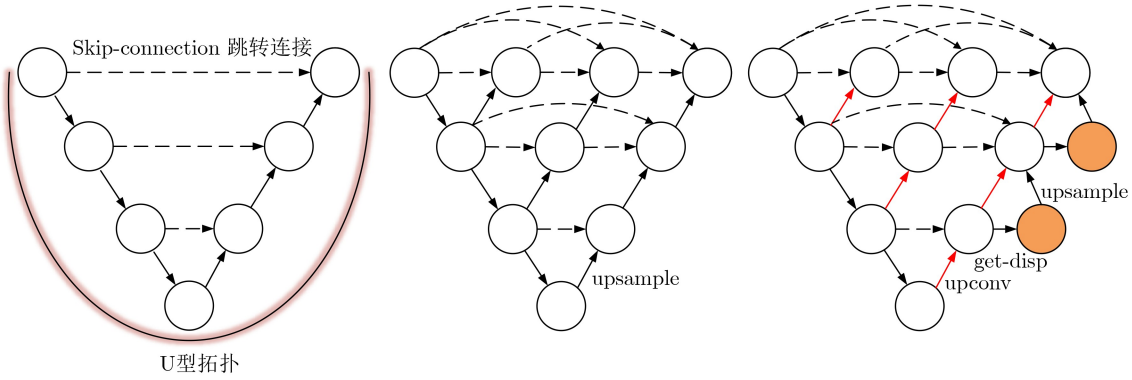


图 1 无监督深度估计算法框图



(a) U-Net网络拓扑 (b) U-Net++网络拓扑 (c) 本文网络拓扑

图 2 U-Net, U-Net++和本文的网络拓扑图

操作更适合深度估计问题，反卷积操作在图2(c)中用红色的箭头表示。并且将预测出的低分辨率视差图也当作特征进行融合，引导网络逐步生成更高分辨率、边界更清晰的视差图。本文将这种全新的特征融合策略命名为密集特征融合层DFFL。

3 本文方法

本节主要介绍应用DFFL的无监督单目深度估计网络，该网络基于编解码器结构实现了从单张RGB图像到对应深度图的端到端预测。本节对传统的编解码器进行改造，降低了编码器的复杂程度，将提出的DFFL密集地部署在解码器上，提高了网络从抽象特征图中恢复深度信息的能力。通过权衡编解码器的性能差异，不仅提高了网络的预测精度，相较于之前的工作，参数量也得到了降低。

3.1 密集特征融合层DFFL及其密集连接

为了消除传统编解码器仅仅在同一层级的编、解码器之间使用跳转连接导致特征利用率低，各级特征之间融合程度不足的问题。本文提出了DFFL，每一个独立的DFFL均是一个解码器节点。DFFL的输入是自适应的，根据其所在位置的不同可能有3种输入：(1)上采样的下一层特征；(2)第1种输入加上同一层通过密集连接引入的特征；(3)第2种输入加上下一层预测的视差图。DFFL将所有输入按通道堆叠在一起，接一个卷积将拼接后的特征进行融合。图3以编解码器的第1层为例，展示在3种不同的输入下DFFL如何对密集特征进行融合。

图3最上方一行是本文提出的编解码器网络的第1层，下面3行展示了DFFL的内部结构，其中最左侧为连续3层编码器的输出特征图。第1种输入情况首先将相邻的两个特征图按通道叠加后，接卷积融合。第2种输入情况再将融合生成的两个特征图与第1层编码器的输出按通道叠加融合。第3种输入情况接收本层的编码器的输出，所有与之相连接的DFFL的输出以及下一层生成的两张低分辨率左右视差图进行融合。基于这样的融合策略，使得图3

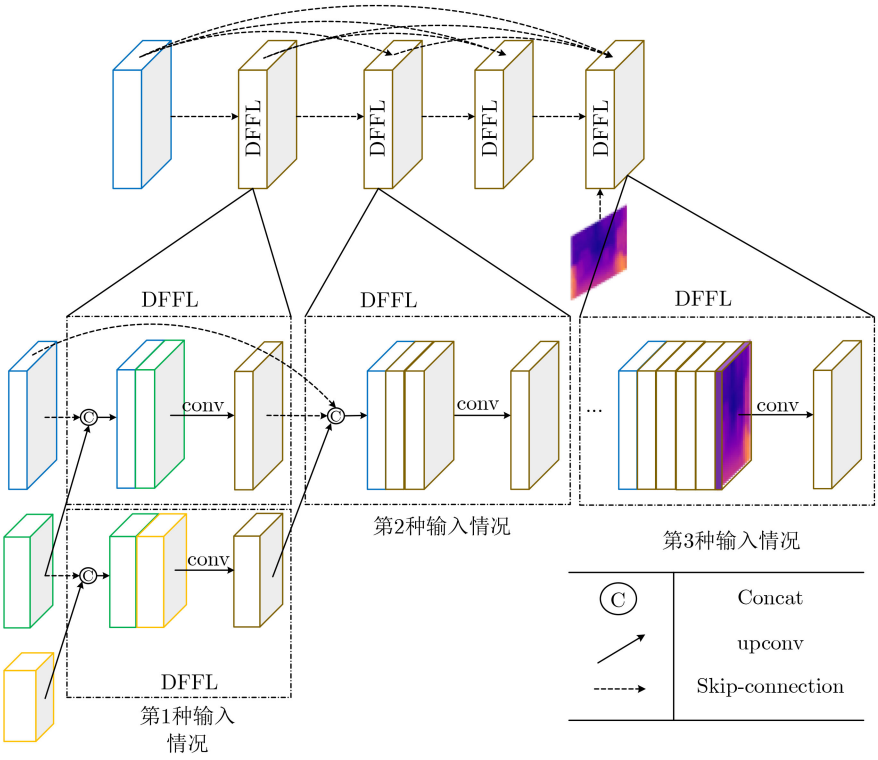


图 3 密集特征融合层及其密集连接

中第1层的第2个DFFL虽然处在第1层编解码器之间但是获得了连续3层的特征信息。同时，将低分辨率的视差图当作DFFL的输入，利用融合得到的上下文信息对低分辨率视差图不断精细化。逐步指导网络生成更高分辨率、细节更清晰的视差图。

在该结构中，同一层特征之间放弃了U-Net的长连接结构，采用密集连接的形式，大大提高了密集特征融合层的特征复用率。密集连接的思想来自DenseNet[16]，这种结构的另一个优势是训练时梯度更容易传播，不容易出现梯度消失的问题。通过引入DFFL，充分地融合了各级特征，使得最终用于估计视差的特征图既包含全局的语义信息，也包括图像的细节信息。

3.2 网络结构

本文基于上述的DFFL，设计出改进后的编解码器网络。整个编解码器网络以左视图作为输入，输出4个空间分辨率下的左右视差图。网络的结构图如图4所示。

在编码器部分，使用修剪后的ResNet-50作为特征提取器，ResNet通过对恒等映射的学习，允许网络规模进一步加深提取出更抽象更丰富的特征信息，但是考虑到编码器一般都是一些精心设计的，已经被图像处理的各个领域广泛使用的基础网络，比如VGG, ResNet, DenseNet等，解码器部分相对来说要简单得多，成为整个网络的短板，使得编码器即使提取出了非常好的特征表示，解码器也未必能将其很好地还原。直观地体现在网络最终输出的深度图边界不清晰，有很多伪影。U型编解码器结构具有很强的对称性，为此将ResNet的第1个7×7卷积与max pool替换为相同作用的Resblock，使编码器的每一层都是Resblock，并且减半了每一级Resblock的通道数来控制编码器的能力，详细的修改见表1。

其中，R50代表ResNet-50, PR50代表修剪后的ResNet-50(Pruned ResNet-50)。

在解码器部分，本文通过密集放置所提出的DFFL，组成互相交织的多路解码器网络，每一个DFFL都是解码器的一个节点。同时，对U-Net的跳转连接进行改造，原始的U-Net每一层的跳转连接一定程度上补充了网络因连续的下采样而丢失的图像细节。但是，特征的提取过程以及使用提取出的特征重建图片的过程都是抽象的，每一层解码器所需要的补充信息并不一定来自对应层的编码器。基于这样的出发点，重新设计的解码器由多个不同规模的解码器组合形成，相邻的解码器之间相互连接。先前大多数的工作更关注优秀的特征提取，即如何使网络变“深”，忽视了怎样去充分利用提取出的优秀特征，即如何使网络变“宽”。将U-Net“填满”，丰富横向的拓扑结构的思想与Inception[17]类似，不同的是，本文希望网络在特征融

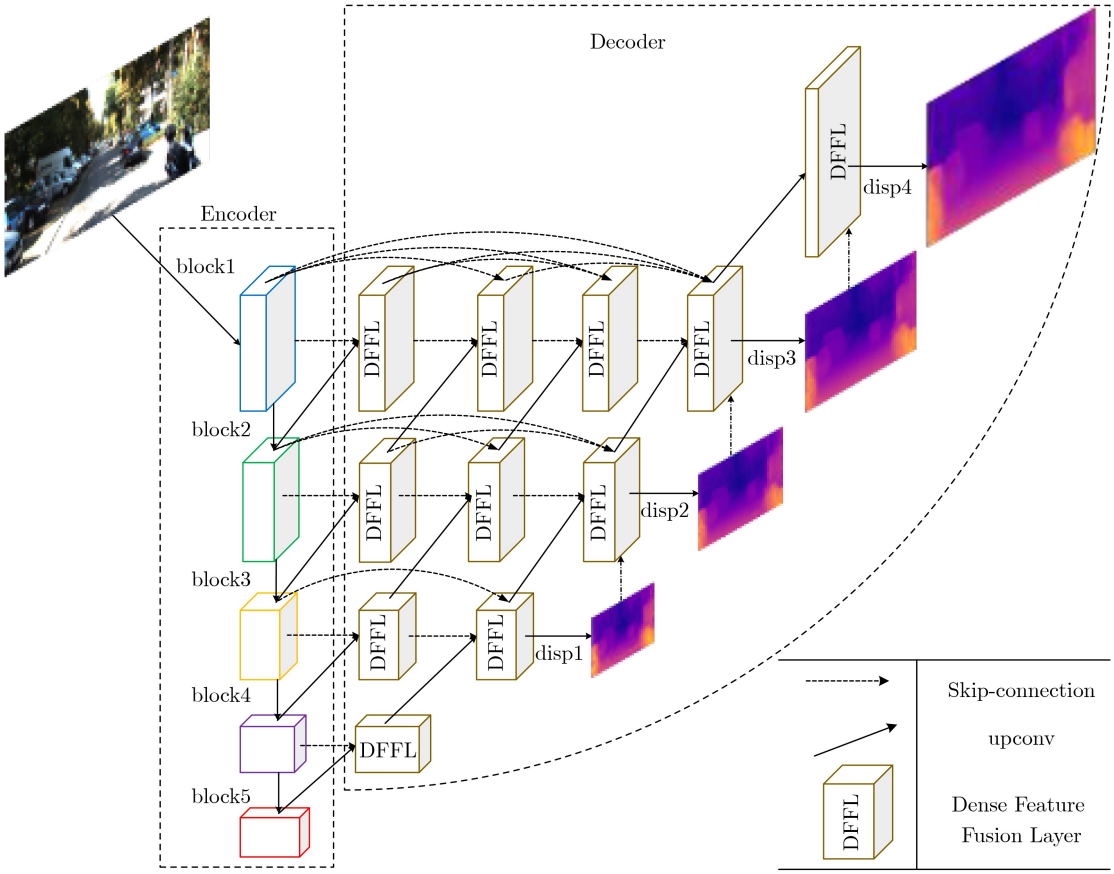


图 4 网络框架

表 1 修改前后的编码器参数

层	R50	PR50
block1	7×7,64,stride2	$\begin{bmatrix}1\times1,8\\3\times3,8\\1\times1,32\end{bmatrix}\times2$
block2_x	3×3 max pool,stride2, $\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,32\\3\times3,32\\1\times1,128\end{bmatrix}\times3$
block3_x	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times4$
block4_x	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times6$
block5_x	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times3$

合部分变得更“宽”。实验表明，对于无监督单目深度估计，提高精度的瓶颈不在于编码器使用多么复杂的特征提取网络，解码器如何充分利用提取出的抽象特征，如何调度各层特征之间的融合才是瓶颈所在。

3.3 损失函数

本文沿用了文献[11]中使用的损失函数，用s表示不同的输出尺度。总的损失L是4个输出尺度损失之和：$L=\sum_{s=1}^{4}L_{\mathrm{s}}$。每一个尺度下的损失$L_{\mathrm{s}}$由3部分组成，分别为重构匹配损失$L_{\mathrm{m}}$、平滑损失$L_{\mathrm{ds}}$以及左右视差一致损失$L_{\mathrm{lr}}$。$L_{\mathrm{s}}$可以表示为

$$L_{\mathrm{s}}=\alpha_{\mathrm{m}}\left(L_{\mathrm{m}}{}^{\mathrm{l}}+L_{\mathrm{m}}{}^{\mathrm{r}}\right)+\alpha_{\mathrm{ds}}\left(L_{\mathrm{ds}}{}^{\mathrm{l}}+L_{\mathrm{ds}}{}^{\mathrm{r}}\right)+\alpha_{\mathrm{lr}}\left(L_{\mathrm{lr}}{}^{\mathrm{l}}+L_{\mathrm{lr}}{}^{\mathrm{r}}\right) \tag{1}$$

其中，α_{m}, α_{ds}, α_{lr}为3个损失的权重。网络以左视图为输入，同时输出左右视差图，因此每一个损失同时拥有左右两个版本。下面以左视图版本为例介绍3种损失各自的作用：

(1)重构匹配损失L_{m}：网络根据预测出的视差图，在对应视图上进行采样。为了验证采样后的重构视图与原视图是否相似，这里除了使用常用的$L1$范数，还引入了结构相似性指标SSIM[18]，具体公式为

$$L_{\mathrm{m}}^{\mathrm{l}}=\frac{1}{N}\sum_{i,j}\alpha\frac{1-\mathrm{SSIM}(I_{ij}^{\mathrm{l}},\tilde{I}_{ij}^{\mathrm{l}})}{2}+(1-\alpha)\left\|I_{ij}^{\mathrm{l}}-\tilde{I}_{ij}^{\mathrm{l}}\right\| \tag{2}$$

其中，I_{ij}^{l}为原视图某一像素位置RGB值的平均，$\tilde{I}_{ij}^{\mathrm{l}}$为重构后的视图某一像素位置RGB值的平均，$N$为所有图片的总像素数，$\alpha$为权重系数。

(2)平滑损失L_{ds}：该损失使用原图在x,y方向上的梯度信息约束视差图的梯度。原图较为平滑的区域视差图也应该较为平滑，减少了人为伪影的出现。而原图的梯度变化较大的边界区域也指引视差图获得更清晰的边界。具体公式为

$$L_{\mathrm{ds}}^{\mathrm{l}}=\frac{1}{N}\sum_{i,j}\left|\partial_x d_{ij}^{\mathrm{l}}\right|\mathrm{e}^{-\left\|\partial_x I_{ij}^{\mathrm{l}}\right\|}+\sum_{i,j}\left|\partial_y d_{ij}^{\mathrm{l}}\right|\mathrm{e}^{-\left\|\partial_y I_{ij}^{\mathrm{l}}\right\|} \tag{3}$$

(3)左右视差一致损失L_{lr}：为了使输出正确的左右视差图，应使其具有一致性。一致性的含义是：根据左视差图中的视差信息为索引在右视差图采样，使得重构出的左视差图与原始的左视差图尽可能相似。具体公式为

$$L_{\mathrm{lr}}^{\mathrm{l}}=\frac{1}{N}\sum_{i,j}\left|d_{ij}^{\mathrm{l}}-d_{ij+d_{ij}^{\mathrm{l}}}^{\mathrm{r}}\right| \tag{4}$$

4　实验结果与分析

本章使用应用最为广泛的KITTI数据集与其他深度估计算法进行了比较。其中包括：有监督的算法[3,4,19]，基于单目视频序列的无监督算法[9,20]，基于双目图像对的无监督算法[10,11,21–23]。同时，通过消融实验验证了本文各项改进的作用。

4.1　实施细节

本网络具体实验环境如下：网络使用PyTorch编程实现，硬件方面为单张RTX2080Ti显卡，12 GB运行内存，操作系统为Ubuntu18.04。输入图片被缩放到512×256大小。优化器选择Adam优化器，优化器参数为β_1= 0.9, β_2= 0.999, ε= 10^{-8}。网络总共训练50个epochs，初始学习率为10^{-4}，在第30个epoch学习率减半，在第40个epoch再减半。Upconv操作由一个放大率为2的双线性插值后跟一个3×3卷积实现。

为了避免过拟合，采用的数据增强操作为：以0.5的概率分别对图片进行水平翻转，在[0.8,1.2]范围内改变gamma值，在[0.5,2.0]范围内改变亮度，在[0.8,1.2]范围内改变图片的彩色3通道。

4.2　数据集

KITTI数据集总共包含了来自61个场景的42382张校正的双目图像对。绝大多数图片分辨率为1242×375。为了与其他工作进行对比，本文使用了Eigen 等人[3]拆分出的训练集与测试集。Eigen使用29个场景中的697张图进行测试，剩下的32个场景包含了22600张训练图片与888张验证图片。为了与其他工作保持一致，所有的测试结果都使用Garg等人[10]的裁剪方式进行裁剪。

后处理：因为双目遮挡的缘故，生成的左视差图的左边界往往比较模糊。文献[11,23]为了解决这个问题引入了后处理操作，将图像$\boldsymbol{I}$及其水平镜像$h(\boldsymbol{I})$输入给网络，分别得到两个视差图$\boldsymbol{d}$, $\boldsymbol{d}_{\mathrm{h}}$。再次对$\boldsymbol{d}_{\mathrm{h}}$进行水平镜像得到与$\boldsymbol{d}$对齐的$\boldsymbol{d}_{\mathrm{h}}'$。综合$\boldsymbol{d}$的前5%，$\boldsymbol{d}_{\mathrm{h}}'$的后5%，中间部分为$\boldsymbol{d}$与$\boldsymbol{d}_{\mathrm{h}}'$的平均，得到最终的视差图。使用后处理的方法在表1中以黑体**pp**标明。

4.3　评价指标

在评估的过程中，本文使用了与之前工作相同的评价指标。分别为阈值精度，平均相对误差(Absolute Relative error, Abs Rel)，平方相对误差(Square Relative error, Sq Rel)，均方根误差(Root Mean Square Error, RMSE)，对数均方根误差(Root Mean Square logarithmic Error, RMSE ln)。公式为

$$\text{阈值精度}:\%\text{ of }d_i\text{ s.t. }\max\left(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i}\right)=\delta<\mathrm{thr} \tag{5}$$

$$\text{Abs Rel}:\frac{1}{|T|}\sum_{d\in T}|d-d^*|/d^* \tag{6}$$

$$\text{Sq Rel}:\frac{1}{|T|}\sum_{d\in T}\|d-d^*\|^2/d^* \tag{7}$$

$$\text{RMSE}:\sqrt{\frac{1}{|T|}\sum_{d\in T}\|d-d^*\|^2} \tag{8}$$

$$\text{RMSE ln}:\sqrt{\frac{1}{|T|}\sum_{d\in T}\|\ln d-\ln d^*\|^2} \tag{9}$$

其中，d为某一像素的预测深度值，d^*为某一像素的真实深度值，T为真实深度图中可获取的像素总数。

4.4 结果对比及分析

为证明本文方法的有效性和先进性，在KITTI数据集上将本文方法与近年相关方法进行对比，结果见表2。

监督方式一栏中，D代表有监督的方法，M代表基于单目视频序列的无监督方法，S代表基于双目图像对的无监督方法。黑体**pp**表示加入了后处理操作。每一项指标的最优结果用黑体标注。

从表2中可以看出，几乎所有评价指标，本文的结果均优于先前的方法。值得注意的是，本文在提高模型精度的同时，并没有扩大网络的参数量，因此推理深度的速度很快。以简化的ResNet-50作为编码器的网络可以做到21 fps的推理速度，使用简化的ResNet-18可以达到33 fps的推理速度。

图5给出了一些可视化的结果，可以看到本文所提算法估计出的深度图像边界更加清晰，并且在深度不变的区域也更为平滑，很少有伪影的出现。与同样比较精确的Monodepth[11]相比，本文在细节处理上更为优秀，图5的最后两列给出了两者的细节对比。

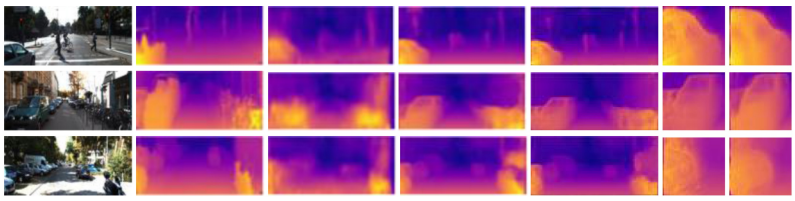
4.5 消融实验

为了验证本文所提DFFL和对编码器修剪的有效性，通过消融实验进行对比，结果如表3所示。

其中，R50代表ResNet-50, PR50代表修剪后的ResNet-50, R18, PR18同理。DFFL代表本文提出的密集特征融合层。

表 2 KITTI数据集使用Eigen拆分集的验证结果

方法	监督方式	越小越好				越大越好		
		Abs Rel	Sq Rel	RMSE	RMSE ln	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Eigen[3]	D	0.203	1.548	6.307	0.282	0.702	0.890	0.890
Liu[4]	D	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Klodt[19]	D+M	0.166	1.490	5.998	–	0.778	0.919	0.966
Zhou[9]	M	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Struct2depth[20]	M	0.141	1.026	5.291	0.215	0.816	0.945	**0.979**
Garg[10]	S	0.152	1.226	5.849	0.246	0.784	0.921	0.967
StrAT[21]	S	0.128	1.019	5.403	0.227	0.827	0.935	0.971
Monodepth2[22]	S	0.130	1.144	5.485	0.232	0.831	0.932	0.968
Monodepth+**pp**[11]	S	0.128	1.038	5.355	0.223	0.833	0.939	0.972
3Net+**pp**[23]	S	0.126	0.961	5.205	0.220	0.835	0.941	0.974
本文	S	0.131	1.110	5.426	0.224	0.839	0.941	0.972
本文+pp	S	**0.122**	**0.939**	**5.063**	**0.212**	**0.850**	**0.947**	0.976



(a) Input (b) GT (c) Garg[10] (d) Godard[11] (e) 本文 (f) 本文细节 (g) 文献[11]细节

图 5 KITTI数据集上可视化结果对比

表 3 KITTI数据集消融实验的结果

方法	编码器网络	参数量($\times10^6$)	Abs Rel	$\delta<1.25$	预测速度(fps)
Baseline	R50	58.5	0.143	0.812	21
Baseline+DFFL	R50	158.0	0.135	0.833	11
网络修剪+DFFL	PR50	39.4	0.131	0.839	21
Baseline	R18	20.2	0.149	0.794	46
Baseline+DFFL	R18	20.4	0.139	0.820	22
网络修剪+DFFL	PR18	19.1	0.129	0.835	33

从表3中可以观察到，在baseline上添加DFFL后精度和误差值都有了一定的优化。以R50作为编码器网络最终输出通道数为2048的特征图，因此进行密集特征融合将引入较大的参数量，但是如果编码器比较精简，例如在R18的基础上加入DFFL就只会增加0.2M的参数。此外，本文通过对编码器进行修剪，同时利用DFFL对各级特征进行融合，使得编码器和解码器的能力做到了很好的权衡，更大程度上发掘了网络的潜力。因此，无论是PR50还是PR18的版本，本文方法相较于所参考的baseline，不仅精度变得更高，参数量也得到了缩减。baseline的R18版本拥有最快的推理速度，但是其精度太低，本文的PR18版本拥有最少的参数量，较低的计算量以及与PR50版本相差无几的精度，甚至更低的平均相对误差Abs Rel，既保证了深度估计的准确性，又维持了预测的速度。

为了证明DFFL 3种不同的输入对于网络精度的影响，进行消融实验，结果如表4所示。从表4中可以观察到，第1种输入融合了上采样的下一层特征，提高了特征的融合程度，相较于baseline精度提高0.7%。第2种输入在第1种输入的基础上通过密集连接引入了同级特征，提高了特征的复用率，相较于第1种输入，网络的精度进一步提高0.8%，该实验也说明了密集连接在本模型中所起到的作用。网络预测出的低分辨率视差图作为指导信号结合DFFL得到的密集特征逐步恢复更精细化的高分辨率视差图是一个从简到难的过程。第3种输入通过融合低分辨率的视差图作为指引，降低了网络的预测难度，在第2种输入的基础上将精度提高了0.3%。

表 4　3种输入下消融实验的结果

方法	编码器网络	Abs Rel	$\delta < 1.25$
Baseline	PR50	0.137	0.821
Baseline +DFFL (第1种输入)	PR50	0.161	0.828
Baseline +DFFL (第2种输入)	PR50	0.132	0.836
Baseline +DFFL (第3种输入)	PR50	0.131	0.839

5　结束语

本文针对无监督的单目深度估计提出了一种全新的网络框架。该框架的核心思想是权衡编解码器的能力，即在合理控制编码器能力的同时，通过在解码的过程中密集放置本文所提出的DFFL，提高特征的融合程度和复用率，并且将多层解码器密集连接起来，提高了解码器的能力，做到编、解码器间的均衡。得益于这种丰富的融合策略，网络最终用于估计视差图的特征图中包含了全局、局部以及各个尺度下的特征信息。在KITTI数据集的实验结果表明，本文相较于之前的算法估计出更平滑、边界更清晰、伪影更少的深度图像，本文的精度高于一些有监督的方法，也预示着无监督深度估计的潜力。通过无监督的训练，避免了网络对于真实深度图的依赖，使得网络可以适用于更多的实际场景中。本文在提高预测精度的同时拥有着较快的预测速度，满足实时场景的深度估计。

参 考 文 献

[1] SNAVELY N, SEITZ S M, and SZELISKI R. Skeletal graphs for efficient structure from motion[C]. 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA, 2008: 1–8.

[2] 狄红卫, 柴颖, 李逵. 一种快速双目视觉立体匹配算法[J]. 光学学报, 2009, 29(8): 2180–2184. doi: 10.3788/AOS20092908.2180.
DI Hongwei, CHAI Ying, and LI Kui. A fast binocular vision stereo matching algorithm[J]. *Acta Optica Sinica*, 2009, 29(8): 2180–2184. doi: 10.3788/AOS20092908.2180.

[3] EIGEN D, PUHRSCH C, and FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]. The 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 2366–2374.

[4] LIU Fayao, SHEN Chunhua, LIN Guosheng, *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. doi: 10.1109/TPAMI.2015.2505283.

[5] LAINA I, RUPPRECHT C, BELAGIANNIS V, *et al*. Deeper depth prediction with fully convolutional residual networks[C]. The 2016 4th International Conference on 3D Vision, Stanford, USA, 2016: 239–248.

[6] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778.

[7] 周武杰, 潘婷, 顾鹏笠, 等. 基于金字塔池化网络的道路场景深度估计方法[J]. 电子与信息学报, 2019, 41(10): 2509–2515.

doi: 10.11999/JEIT180957.

ZHOU Wujie, PAN Ting, GU Pengli, *et al*. Depth estimation of monocular road images based on pyramid scene analysis network[J]. *Journal of Electronics & Information Technology*, 2019, 41(10): 2509–2515. doi: 10.11999/JEIT180957.

[8] ZHAO Shanshan, FU Huan, GONG Mingming, *et al*. Geometry-aware symmetric domain adaptation for monocular depth estimation[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 9780–9790.

[9] ZHOU Tinghui, 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, Honolulu, USA, 2017: 6612–6619.

[10] GARG R, B G V K, CARNEIRO G, *et al*. Unsupervised CNN for single view depth estimation: Geometry to the rescue[C]. The 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 740–756.

[11] GODARD C, MAC AODHA O, and BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6602–6611.

[12] ZHOU Zongwei, SIDDIQUEE M R, TAJBAKHSH N, *et al*. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation[J]. *IEEE Transactions on Medical Imaging*, 2020, 39(6): 1856–1867. doi: 10.1109/TMI.2019.2959609.

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

[14] HARTLEY R and ZISSERMAN A. Multiple View Geometry in Computer Vision[M]. 2nd ed. New York: Cambridge University Press, 2003: 262–263.

[15] RONNEBERGER O, FISCHER P, and BROX T. U-net: Convolutional networks for biomedical image segmentation[C]. The 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 2015: 234–241.

[16] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, *et al*. Densely connected convolutional networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 2261–2269.

[17] SZEGEDY C, VANHOUCKE V, IOFFE S, *et al*. Rethinking the inception architecture for computer vision[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2818–2826.

[18] WANG Zhou, 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. doi: 10.1109/TIP.2003.819861.

[19] KLODT M and VEDALDI A. Supervising the new with the old: Learning SFM from SFM[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 713–728.

[20] CASSER V, PIRK S, MAHJOURIAN R, *et al*. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos[C]. The 33rd AAAI Conference on Artificial Intelligence, Honolulu, USA, 2019: 8001–8008.

[21] MEHTA I, SAKURIKAR P, and NARAYANAN P J. Structured adversarial training for unsupervised monocular depth estimation[C]. 2018 International Conference on 3D Vision, Verona, Italy, 2018: 314–323.

[22] GODARD C, MAC AODHA O, FIRMAN M, *et al*. Digging into self-supervised monocular depth estimation[C]. 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 3827–3837.

[23] POGGI M, TOSI F, and MATTOCCIA S. Learning monocular depth estimation with unsupervised trinocular assumptions[C]. 2018 International Conference on 3D Vision, Verona, Italy, 2018: 324–333.

陈 莹：女，1976年生，教授，博士，研究方向为信息融合、模式识别.

王一良：男，1997年生，硕士生，研究方向为计算机视觉与模式识别.

责任编辑：马秀强