基于深度卷积神经网络的道路场景深度估计

袁建中[1]，周武杰[1,2*]，潘婷[1]，顾鹏笠[1]
[1]浙江科技学院信息与电子工程学院，浙江 杭州 310023；
[2]浙江大学信息与电子工程学院，浙江 杭州 310027

摘要 提出了一种基于深度卷积神经网络的单目视觉深度估计方法，该方法采用端到端学习框架来构建模型。采用残差网络（ResNet）作为神经网络模型框架的编码部分来提取深度信息特征。采用密集连接卷积网络（DenseNet）对编码后的信息进行译码。通过 Skip-Connections 实现编码和解码的信息流的集成，避免了层间信息传输的丢失。实验结果表明，与其他单目视觉深度估计方法相比，使用深度卷积神经网络可以更有效准确地估计视觉深度。

关键词 机器视觉；深度卷积神经网路；深度估计；单目图像；深度学习
中图分类号 TP391 **文献标识码** A **doi**：10.3788/LOP56.081501

Road Scene Depth Estimation Based on Deep Convolutional Neural Networks

Yuan Jianzhong[1], Zhou Wujie[1,2*], Pan Ting[1], Gu Pengli[1]
[1] *School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, Zhejiang* 310023, *China*;
[2] *College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, Zhejiang* 310027, *China*

Abstract A monocular visual depth estimation method is proposed based on deep convolutional neural networks, in which an end-to-end learning framework is used to construct a model. A residual network (ResNet) is used as the coding part of the neural network model framework to extract the depth information features. The encoded information is decoded by a densely concatenated convolution network (DenseNet). The integration of the encoded and decoded information streams is realized by Skip-Connections, which avoids the loss of inter-layer information under transmission. The experimental results show that the depth convolution neural network can be used to estimate visual depth more effectively and accurately than other monocular visual depth estimation methods.

Key words machine vision; deep convolutional neural network; depth estimation; monocular image; deep learning
OCIS codes 150.1135; 100.2960; 350.4600

1 引 言

深度估计旨在预测单个或多个图像的像素深度，在单眼深度估计中描绘多种尺度的复杂组合是一项具有挑战性的任务。通过深度信息可以理解周围的三维空间，这对于道路安全起着很重要的作用。道路场景的深度估计也是无人驾驶领域必不可少的环节，成熟精确的道路场景深度估计技术可以有效地保障行驶过程中汽车的安全[1]。近年来由于深度学习方法的应用和端到端学习框架的提出[2]，深度卷积神经网络（DCNN）的模型被应用到绝大多数计算机视觉领域中去。在单目视觉深度估计领域，与传统方法相比，借助深度卷积神经网络和深度学习的方法来进行单目深度估计，其结果得到了质的提升。

早期的工作主要集中在立体图像的深度估计

收稿日期：2018-11-02；**修回日期**：2018-11-14；**录用日期**：2018-11-22
基金项目：国家自然科学基金（61502429）、浙江省自然科学基金（LY18F020012）、浙江科技学院研究生科研创新基金（2017YJSKC004）

* **E-mail**：wujiezhou@163.com

上，通过开发基于几何的算法[3-4]，基于图像依靠三角测量之间的点对应来估计深度。在一项开创性的研究中，Saxena 等[5]通过监督学习从二维图像单眼线索中了解了深度。在这种方法之后，又有很多种利用手工表示的单眼线索[6-16]方法被提出。由于手工制作的特征本身只能捕获局部信息，因此常常基于这些特征构建诸如马尔可夫随机场（MRF）的概率图形模型，目的是使其包含长距离和全局信息[6,17-18]。使用全局信息的另一种成功方法是 DepthTransfer 方法[19]，该方法在包含 RGB-D（Red，Green，Blue-Depth）图像的数据库中，使用 GIST 全局场景特征[20]来搜索与输入图像“相似”的候选图像。

鉴于深度卷积神经网络在图像识别方面的成功应用，近年来越来越多的深度估计网络被提出。对于具有足够深度的网络［如 VGG[21]（Visual Geometry Group）和 ResNet[22]（Residual Network）］，其多级上下文和结构信息的深度估计已被提升到新的准确度。许路等[23]使用深度卷积神经网络与手工特征结合的方法来提取图像的深度信息。Eigen 等[24-25]提出了一种用于预测深度、表面法线和语义标签的多尺度架构。Xie 等[26]采用 Skip-Connections 策略将较深层中的低空间分辨率深度图与较低层中的高空间分辨率深度图融合。李素梅等[27]基于卷积神经网络（CNN），模拟人类的视觉系统，对原始深度图进行层次化的抽象处理，提出一种自主提取特征的深度图超分辨率重建算法。吴寿川等[28]提出一种用于单目红外视频深度估计的双向递归卷积神经网络。徐冉等[29]针对获取的深度图分辨率低、边缘信息丢失等缺点又提出基于双通道卷积神经网络的深度图超分辨率重建模型。Liu 等[30]提出了一种深度完全卷积神经网络模型和一种新颖的超像素池方法，进行深度估计，得到了统一深度网络中连续条件随机场（CRF）的一元和二元电位。同样，Li 等[31-32]还将 CNN 与 CRF 相结合，他们在双层分级 CRF 中制定了深度估计，以加强全局和局部预测之间的协同作用。Li 等[33]将单目深度估计作为一种多类别密集标记任务，以分层方式融合前端扩张的卷积神经网络的不同侧面输出，以利用多尺度深度线索进行深度估计。Li 等[34]提出了一种快速训练的双流 CNN，可以预测深度和深度梯度，然后将它们融合在一起形成精确而详细的深度图。Lee 等[35]提出了一种基于傅里叶频域分析的单图像深度估计的深度学习算法。Ummenhofer 等[36]从连续的、无约束的图像对中训练一个端到端的卷积神经网络来计算图像像素的深度值。为了提高效率，Fu 等[37]提出了深度顺序回归网络（DORN），首先将真值深度按照区间递增的方法预分为许多深度子区间，然后设计了一个像素到像素的有序回归损失函数来模拟这些深度子区间的有序关系。这些方法的提出极大地促进了深度估计技术的发展，但是其中有些方法的神经网络没有足够的深度，有些方法虽然保证了神经网络的足够深度却丢失了浅层的特征信息，两者均会影响其整体深度估计结果的精度。

残差神经网络（ResNet）采用标准的卷积神经网络的搭建模式，同时借助 Skip-Connections 的方式绕过一些卷积层来构造残差块。针对单目图像深度预测，ResNet 具有足够深的网络深度，可以提取到具有更高准确性和更多深度信息的特征；并且残差特性能够降低在增加神经网络层数时出现梯度消失而导致网络性能退化的可能性。密集连接卷积网络（DenseNet）[38]通过对特征的极致利用可以达到更好的效果和更少的参数，并在保证网络中层与层之间最大程度的信息传输的前提下，直接将网络中的所有层连接起来。DenseNet 应用于深度预测能够充分有效地利用特征并减轻梯度消失的问题。

考虑到 ResNet 和 DenseNet 的优越性能，本文将两者结合在一起，构建出一个单目视觉深度估计模型，不仅保证了神经网络足够的深度，而且由于其残差特性，浅层信息不但没有丢失，还得到了充分利用。该模型主要包括 3 个过程：1）使用 ResNet 作为编码器，从原始图像中提取特征信息；2）将 DenseNet 作为译码器，提供一种上采样功能，将特征以原始图像尺寸的预测深度图的形式输出；3）使用 Skip-Connections 级联 ResNet 和 DenseNet 的层间信息，可以更加有效地利用特征信息。本文采用网络深度更深的卷积神经网络来进行深度估计，实验结果表明，相比较其他方法，所提网络模型可以更有效准确地估计单目图像的深度。本文的创新点在于充分利用了神经网络 ResNet 和 DenseNet 的优点，以网络块为基础，结合多处的 Skip-Connections 来构建出简洁但有足够网络深度的神经网络模型，充分发挥残差属性的优势，最大化 Skip-Connections 对层间信息的利用，来提取丰富而准确的特征信息，并生成高精度的深度预测图。

2 模型方法提出

2.1 网络体系框架

针对当前现有方法处理单目图像深度估计的挑战，本文使用端到端的单眼深度估计学习框架，该框架学习从彩色图像到相应深度图的直接映射。CNN 架构的深度对于网络的性能非常重要，近年来的许多研究表明，VGG 网络胜过较浅的 AlexNet[39]。然而，简单地将更多图层堆叠到现有的 CNN 架构并不一定会提高性能，因为梯度消失的问题会阻碍从训练开始的融合。本文则在 ResNet 和 DenseNet 之间通过添加 Skip-Connections 来解决这个问题。

本文应用残差神经网络和密集连接卷积网络来构建单目视觉深度估计的网络模型，如图 1 所示，该网络以 ResNet 中的 Conv_block 和 Identity_block 的形式组成编码部分，译码的部分主要借助于 DenseNet 中的 Dense_block 和 TransitionUp 来完成。

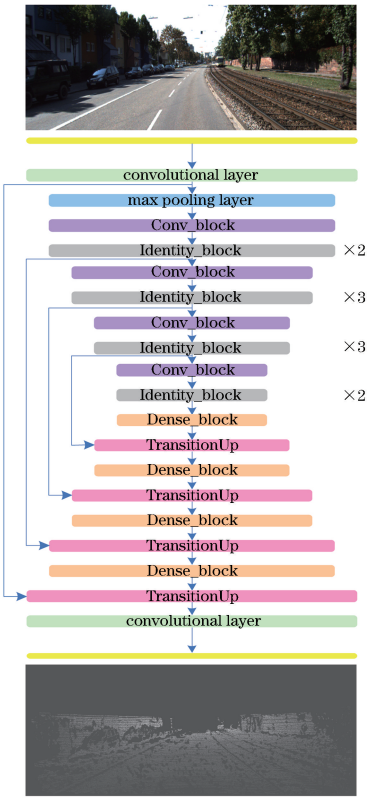

图 1 神经网络结构图

Fig. 1 Structural diagram of neural network

本文的网络体系结构主要分为编码部分和译码部分，以及 Skip-Connection。编码部分由卷积层、最大池化层和 4 个 Conv_block 和 Identity_block 组合构成。同样地，由 4 个 Dense_block 和 4 个 TransitionUp 以及卷积层构成了译码部分。

神经网络由编码部分开始，输入图像先经过一个卷积层和一个最大池化层的操作后，再将图片送入 Conv_block 和 Identity_block 中，首先从 Conv_block 开始，随后是不同数量的 Identity_block。在本实验中，上述 4 个卷积块中 Identity_block 的数量分别是 2，3，3，2。

网络的译码阶段从 Dense_block 开始，编码阶段的输出经 4 个 Dense_block 和 4 个 TransitionUp 后会被上采样到原始图像的尺寸大小。最后再经过一个通道数为 1 的卷积层后，就能得到神经网络最终产生的预测深度图。Skip-Connections 的主要作用是将译码部分 TransitionUp 中反卷积后的输出与编码部分对应尺寸相同的输出特征使用 concatenate 层融合起来，以加强特征信息的有效使用。

2.2 深度残差神经网络

加深网络的深度有助于提高神经网络的性能，但是在不断增加神经网络深度的同时网络的训练难度也会不断变大。不适当的网络深度的叠加甚至会导致神经网络在训练时出现性能退化问题。

深度残差神经网络的使用不会因为网络深度的增加而出现网络性能的退化，故本文方法可以具有更强的稳健性。考虑到残差神经网络的这些优势，在本文的模型中，采用了两种具有这种特性优势的网络块：Conv_block 和 Identity_block，构建模块如图 2 所示。

图 2(a)中所示的 Identity_block 定义为

$$\boldsymbol{y}=F(\boldsymbol{x},\boldsymbol{W}_i)+\boldsymbol{x}, \tag{1}$$

式中：$\boldsymbol{x}$ 和 $\boldsymbol{y}$ 分别是堆叠层的输入和输出矩阵；函数 $F(\boldsymbol{x},\boldsymbol{W}_i)$ 是需要学习的残差映射，$\boldsymbol{W}_i$ 代表学习到的权重矩阵，i 代表矩阵中的元素。由于 Identity_block 中的 Skip-Connections 使用的融合方式是按元素相加的，因此 $\boldsymbol{x}$ 和 $F(\boldsymbol{x},\boldsymbol{W}_i)$ 的尺寸必须相同。

图 2(b)中所示的 Conv_block 定义为

$$\boldsymbol{y}=F(\boldsymbol{x},\boldsymbol{W}_i)+\boldsymbol{W}_s\boldsymbol{x}, \tag{2}$$

式中：$\boldsymbol{W}_s$ 代表线性投影使用的权重矩阵，s 代表矩阵中的元素。如果 $\boldsymbol{x}$ 和 $F(\boldsymbol{x},\boldsymbol{W}_i)$ 尺寸不相等，可以通过 Skip-Connections 执行线性投影 $\boldsymbol{W}_s$ 来匹配 $\boldsymbol{x}$ 和 $F(\boldsymbol{x},\boldsymbol{W}_i)$ 的尺寸。

在本实验中，Conv_block 的右侧第一个卷积层和左侧卷积层的步长都是 2，其余卷积层(包含本文神经网络使用的其他网络块中的卷积层)步长都为 1。在编码部分，由于最大池化层已经将图像的尺

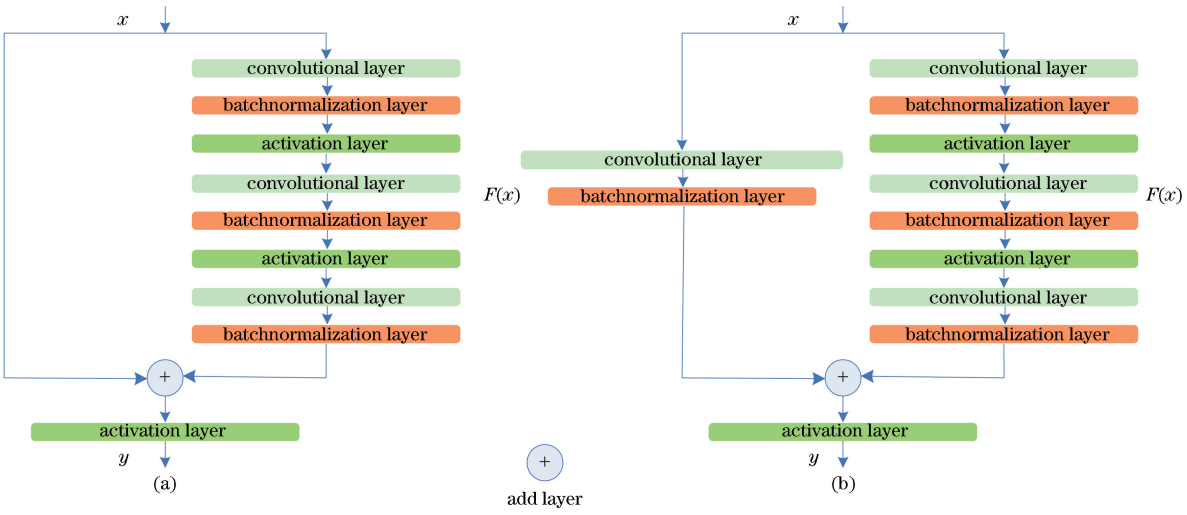

图 2 两种类型的网络块。(a) Identity_block;(b) Conv_block

Fig. 2 Two types of network blocks. (a) Identity_block; (b) Conv_block

寸缩减为原来的一半,因此将本文使用的第一个 Conv_block 的卷积步长设为 1,其余的 Conv_block 不变。使用这两种网络块来搭建本文神经网络的编码部分,合理搭配块的数量不仅可以加深神经网络的深度,提取到更为准确的特征信息,而且加深网络深度后网络性能不会出现退化现象。

2.3 密集连接卷积网络

原始图像经过编码部分提取到具有有效准确的特征信息的图像,同时该图像的尺寸与原始图像相比已经变小了很多,但是神经网络最终输出的预测深度图尺寸要和原始图像一致。因此,需要在尽量不损失这些特征信息的前提下进行译码操作,将编码器的输出上采样到原始图像的尺寸大小。采用密集连接卷积网络 DenseNet 中的 Dense_block 和 TransitionUp 来实现具有这样功能要求的译码操作。

调用 Dense_block 来连接给定分辨率下创建的新特征图。图 3 显示了 Dense_block 构造的示例。从具有 m 个特征映射的输入 $\boldsymbol{x}_0$ 开始,进过 Dense_block 的第一层生成维度为 k 的输出 $\boldsymbol{x}_1$。然后通过串联($[\boldsymbol{x}_1,\boldsymbol{x}_0]$)将 k 个特征映射堆叠到先前的 m 个特征映射,并用作第二层的输入。相同的操作重复 n 次,具有 $n\times k$ 个特征映射的新的特征图。Dense_block 中的 Layer 层由规范化(Batch Normalization,BN)层、ReLU 激活层和卷积核大小为 3×3 的卷积层组成。

为了恢复到原始图像的分辨率,用 DenseNet 中的 TransitionUp 的上采样功能来实现。TransitionUp 的构造图如图 4 所示。TransitionUp 由转置卷积组成,转置卷积可以对图像进行向上采样。使用 Skip-Connections 将转置卷积层上采样得到的特征映射和分辨率相同的 Identity_block 输出连接在一起作为 Dense_block 新的输入。TransitionUp 主要由步长为 2 的 3×3 转置卷积层组成。

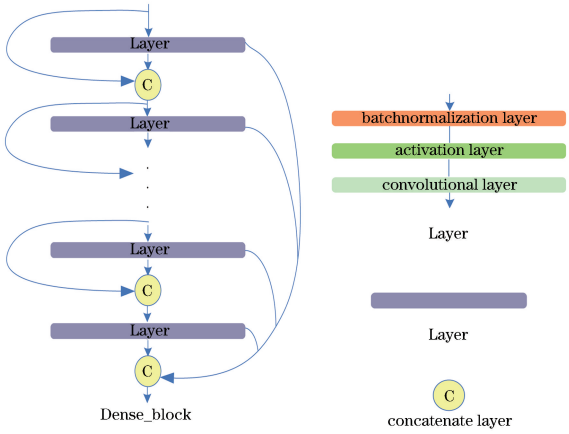

图 3 N 层 Dense_block 图

Fig. 3 Schematic of N-layer Dense_block

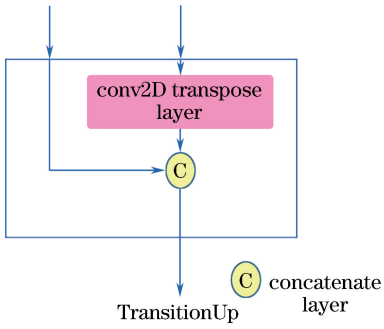

图 4 TransitionUp 图

Fig. 4 Schematic of TransitionUp

使用 Dense_block 和 TransitionUp 进行译码操作，其中的 Skip-Connections 有助于重用特征映射，以获得很多的特征信息。最后一层是卷积核大小为 1×1、通道数为 1 的卷积层，用来输出模型最终的深度预测图。

3 实验与结果分析

3.1 使用数据集

KITTI 数据集是现有公开的最大室外深度估计数据集，KITTI 包含城市街道、乡村道路和高速公路等场景采集的真实图像数据，每张图像的内容多样，包含多辆车和许多行人，并且还有各种程度的遮挡与截断。鉴于大量的训练数据，该数据集允许训练复杂的深度学习模型，用于单个图像深度预测。由于 KITTI 含有丰富的户外街景场景，在诸多计算机视觉任务中被当作数据集来训练，并且其基准成为测试网络模型的公认标准。

本文的深度估计网络也是在 KITTI 数据集上作单目视觉的深度估计，数据集分成两个部分，训练数据集部分由 4286 张 RGB 图像和其对应的深度标签构成，再用 343 张 RGB 图像及相应的深度标签组成测试数据集。训练数据集用来训练本文神经网络模型，本文将整个图像输入到网络以获取更多的背景信息。然后再使用测试数据集来测试本文模型的预测结果，并与其他方法进行对比。

3.2 训练方法

本文的神经网络模型训练与测试的实验都是在具有 11 GB 显存的 GTX 1080 Ti GPU 的计算机上使用基于 TensorFlow 后端的 Keras 学习框架上完成的。训练实验阶段对于模型的优化采用的是学习率为 0.001 的 Adam[40] 算法。

网络采用的是端到端的训练方式，根据 KITTI 提供的数据集，其中图像的大小并不完全一致。所以本文在训练神经网络模型之前需将图像的大小作统一处理，把处理后的图像输入模型，再经过深度神经网络作编码和译码操作，输出与输入图像相同尺寸的深度估计图。在模型训练的过程中使用均方误差函数(mean squared error，MSE)作为 loss 函数来获得训练阶段的最优权重 $\boldsymbol{W}^{\text{best}}$。MSE 的计算公式为

$$E_{\text{MS}}=\frac{1}{|T|}\sum_{n=1}^{|T|}|\boldsymbol{Y}_n-\overline{\boldsymbol{Y}_n}|^2, \tag{3}$$

式中：$\boldsymbol{Y}_n$ 为第 n 幅原始图像对应的深度标签；$\overline{\boldsymbol{Y}_n}$ 为第 n 幅原始图像对应的预测深度；T 为训练数据集。

3.3 测试结果评价分析

测试数据集中的图像按照训练集中图像的处理方式处理，把测试集中的每一张原始图像输入到所提模型框架中去，再载入训练阶段得到的最优权重 $\boldsymbol{W}^{\text{best}}$，最后经过模型的提取特征获得预测深度图。将预测深度图和对应原图中的深度标签作评价，计算获得评价指标。本文采用单目视觉深度预测评价方法中的 6 个常用客观参量作为评价指标，包括方均根误差(R_{rms})、对数方均根误差($R_{\text{log\_rms}}$)、平均对数误差(R_{lg})、阈值准确性(R^{e}_{thr}，$e=1,2,3$，$R_{\text{thr}}=1.25$)，其表达式分别为

$$R_{\text{rms}}=\sqrt{E_{\text{MS}}}=\sqrt{\frac{1}{|T|}\sum_{n=1}^{|T|}|\boldsymbol{Y}_n-\overline{\boldsymbol{Y}_n}|^2}, \tag{4}$$

$$R_{\text{log\_rms}}=\sqrt{\frac{1}{|T|}\sum_{n=1}^{|T|}|\log|\boldsymbol{Y}_n|-\log|\overline{\boldsymbol{Y}_n}||^2}, \tag{5}$$

$$E_{\text{lg}}=\frac{1}{|T|}\sum_{n=1}^{|T|}|\lg|\boldsymbol{Y}_n|-\lg|\overline{\boldsymbol{Y}_n}||, \tag{6}$$

$$\delta=\max\left(\frac{|\boldsymbol{Y}_n|}{|\overline{\boldsymbol{Y}_n}|},\frac{|\overline{\boldsymbol{Y}_n}|}{|\boldsymbol{Y}_n|}\right)<R^{e}_{\text{thr}}。 \tag{7}$$

提出基于深度学习方法的深度卷积神经网络，促进了单目视觉深度估计的高速发展。随着这种方法的不断改进，对深度估计精度的要求越来越高。但是，许多基于这种方法的神经网络模型都很难有效地使用全局特征和局部特征，所以诸多神经网络出现了瓶颈，深度估计的精度难以进一步提升。

由于所提深度神经网络模型具有足够的网络深度，且能够充分利用网络的残差特性，本文神经网络在图像特征提取过程中可以提取到足够的局部特征信息和全局特征信息。同时，在本文神经网络中存在着大量的 Skip-Connections，这些 Skip-Connections 可以将提取到的全局特征信息和局部特征信息进行有效融合。另外，使用的 KITTI 包含有丰富的户外街景场景，所以基于本文网络模型的预测结果更加精确。

表 1 对比了相同数据集(KITTI)下本文网络模型与其他一些经典模型的估计结果，加粗表示本文结果优于其他文章的结果。从表 1 可以看出，$\delta<1.25^3$ 时，本文结果未能超越 Yin 等[42] 的结果，其原因在于 Yin 等在他们的方法中加入了 CRF，但经过大量的实验发现，CRF 在提高 Accuracy 中的 3 个性能指标的同时也会导致 3 个 Error 性能指标提高，因此，调用 CRF 并不能使模型在各个性能指标上都实现提升。所提方法虽然有一个性能指标不如其他方法中的一

个，但是在其他性能指标上都取得了进步，整体的网络模型性能还是优于其他网络模型的。

相比于其他的神经网络模型，本文的模型基于网络块的构造，不仅简洁工整，而且包含足够的深度。在不繁琐的网络基础上能够提取丰富的特征信息并保证深度估计结果的准确性。所提出的神经网络模型不依赖任何预处理步骤即可实现准确的视觉深度估计。由此不仅突显了本文的神经网络模型的优越性，也突出了深度卷积神经网络结合残差方式的强大性能。

表 1　KITTI 数据集的深度估计结果

Table 1　Depth estimation results on KITTI dataset

Method	Accuracy (higher is better)			Error (lower is better)		
	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$	R_{rms}	$R_{\mathrm{log\_rms}}$	E_{lg}
Ref. [24]	0.488	0.947	0.972	2.6440	0.272	0.167
Ref. [41]	0.674	0.943	0.972	2.4618	0.243	0.126
Ref. [42]	0.640	0.947	0.979	2.5193	0.247	0.134
Ref. [43]	0.634	0.916	0.945	2.8246	0.305	0.127
Ref. [44]	0.566	0.945	0.970	2.6507	0.264	0.145
Proposed	**0.717**	**0.947**	0.974	**2.4225**	**0.234**	**0.111**

最后，采用切片实验的方法来完成对本文使用的 Conv_block 和 Identity_block、Dense_block 和 TransitionUp 优势的验证。

使用 4 个常规的卷积层加 4 个最大池化层的编码方式，和本文方法中的 Dense_block、TransitionUp 构成神经网络（将该方法称之为 Proposed1）来验证 Conv_block 和 Identity_block 的优势。同时，使用 Conv_block 和 Identity_block，以及 4 个转置卷积层作为译码方式构成的神经网络（将该方法称之为 Proposed2）来验证 Dense_block 和 TransitionUp 的优势。表 2 给出切片实验得到的结果数据。

表 2　切片实验深度估计结果

Table 2　Depth estimation results of slice experiment

Method	Accuracy (higher is better)			Error (lower is better)		
	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$	R_{rms}	$R_{\mathrm{log\_rms}}$	E_{lg}
Proposed	0.717	0.947	0.974	2.4225	0.234	0.111
Proposed1	0.696	0.941	0.971	2.3877	0.242	0.124
Proposed2	0.678	0.946	0.974	2.412	0.237	0.122

将表 2 中 Proposed 的实验结果与 Proposed1 的实验结果相比可知，在使用相同译码器的情况下，使用 Conv_block 和 Identity_block 作为编码中主要组成部分来搭建的神经网络 Proposed 要优于卷积层加最大池化层组成编码来搭建的神经网络 Proposed1。相同地，使用一样的编码器，将反卷积作为译码方式搭建的神经网络 Proposed2 的实验结果也不如使用 Dense_block 和 TransitionUp 作为译码方式搭建的神经网络 Proposed。

对于 Proposed 方法中性能指标 R_{rms} 不如 Proposed1 和 Proposed2，经过实验发现，这是由于 Conv_block 和 Identity_block、Dense_block 和 TransitionUp 在内部使用了大量的 Skip-Connections，Skip-Connections 的使用能够融合所有特征信息，故导致一些稀少的具有负面作用的特征信息也得到了加强，进而提升了性能指标 R_{rms} 的数据，但提升幅度很小，对于神经网络整体性能影响也非常小。

本文的实验结果图如图 5 所示，图 5(a)为原始 RGB 图像，图 5(b)给出地面实况，图 5(c)给出实验结果深度预测图。

4　结　　论

汽车环境的单目深度估计对于汽车的行驶和安全有着重要的影响，精确的深度估计能够给汽车带来很大的裨益。针对汽车前行方向上的环境，即车前图像的深度预测，本文提出了一种基于 ResNet 和 DenseNet 相结合的以网络块的形式搭建的深度神经网络模型的单目视觉深度估计方法。结合使用 4 种不同网络块，搭建出具有足够深度的神经网络，以提取到更多的特征信息；其中嵌入的 Skip-Connections 增加了层间信息的传递与特征融合。实验在 KITTI 数据集上进行，以验证方法的有效性。实验结果表明，该算法可以提高单目视觉深度估计任务的精度。

图 5 实验结果。(a) RGB 图像;(b)地面实况;(c)深度预测图

Fig. 5 Experimental results. (a) RGB image; (b) ground truth; (c) depth prediction map

在所提神经网络模型的基础上,考虑到图像特征的维度,下一步工作将引用带孔卷积的方法,在不增加训练参数量的基础上扩大卷积神经元的感受野,从而获取更多的特征信息,以加强神经网络的稳健性。

参 考 文 献

[1] Wang F, Chen C, Huang J X. A review of research on driverless vehicles[J]. China Water Transport, 2016, 16(12): 126-128.
王芳, 陈超, 黄见曦. 无人驾驶汽车研究综述[J]. 中国水运, 2016, 16(12): 126-128.

[2] Silver D, van Hasselt H, Hessel M, *et al*. The predictron: End-to-end learning and planning[EB/OL]. (2017-07-20)[2018-09-30]. https://arxiv.org/abs/1612.08810.

[3] Scharstein D, Szeliski R, Zabih R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms[C] // Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001), December 9-10, 2001, Kauai, HI, USA. New York: IEEE, 2001: 131-140.

[4] Flynn J, Neulander I, Philbin J, *et al*. Deep stereo: Learning to predict new views from the world′s imagery[C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 5515-5524.

[5] Saxena A, Chung S H, Ng A Y. Learning depth from single monocular images[C] // Conference and Workshop on Neural Information Processing Systems. [S.l.: s.n.]. 2005, 18: 1161-1168.

[6] 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.

[7] Hoiem D, Efros A A, Hebert M. Recovering surface layout from an image[J]. International Journal of Computer Vision, 2007, 75(1): 151-172.

[8] Ladický L, Shi J B, Pollefeys M. Pulling things out of perspective[C] // 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 89-96.

[9] Choi S, Min D B, Ham B, *et al*. Depth analogy: data-driven approach for single image depth estimation using gradient samples[J]. IEEE Transactions on Image Processing, 2015, 24(12): 5953-5966.

[10] Konrad J, Wang M, Ishwar P, *et al*. Learning-based, automatic 2D-to-3D image and video conversion[J]. IEEE Transactions on Image

Processing, 2013, 22(9): 3485-3496.

[11] Baig M H, Torresani L. Coupled depth learning[C]// 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), March 7-10, 2016, Lake Placid, NY, USA. New York: IEEE, 2016: 1-10.

[12] Shi J P, Tao X, Xu L, *et al*. Break Ames room illusion[J]. ACM Transactions on Graphics, 2015, 34(6): 1-11.

[13] 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), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4058-4066.

[14] Furukawa R, Sagawa R, Kawasaki H. Depth estimation using structured light flow: Analysis of projected pattern flow on an Object′s surface[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 4650-4658.

[15] Häne C, Ladický L, Pollefeys M. Direction matters: Depth estimation with a surface normal classifier[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 381-389.

[16] You X G, Li Q, Tao D C, *et al*. Local metric learning for exemplar-based object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(8): 1265-1276.

[17] Zhuo W, Salzmann M, He X M, *et al*. Indoor scene structure analysis for single image depth estimation[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 614-622.

[18] Liu M M, Salzmann M, He X M. Discrete-continuous depth estimation from a single image[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 716-723.

[19] Karsch K, Liu C, Kang S B. Depth transfer: depth extraction from video using non-parametric sampling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(11): 2144-2158.

[20] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope[J]. International Journal of Computer Vision, 2001, 42(3): 145-175.

[21] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2018-09-30]. https://arxiv.org/abs/1409.1556.

[22] 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), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[23] Xu L, Zhao H T, Sun S Y. The Predictron: End-to-end learning and planning[J]. Acta Optica Sinica, 2016, 36(7): 0715002.
许路, 赵海涛, 孙韶媛. 基于深层卷积神经网络的单目红外图像深度估计[J]. 光学学报, 2016, 36(7): 0715002.

[24] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[EB/OL]. (2014-06-09)[2018-09-30]. https://arxiv.org/pdf/1406.2283v1.pdf.

[25] 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), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 2650-2658.

[26] Xie J Y, Girshick R, Farhadi A. Deep3D: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks[M]// Xie J Y, Girshick R, Farhadi A. eds. Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 842-857.

[27] Li S M, Lei G Q, Fan R. Depth map super-resolution reconstruction based on convolutional neural networks[J]. Acta Optica Sinica, 2017, 37(12): 1210002.
李素梅, 雷国庆, 范如. 基于卷积神经网络的深度图超分辨率重建[J]. 光学学报, 2017, 37(12): 1210002.

[28] Wu S C, Zhao H T, Sun S Y. Depth estimation from monocular infrared video based on Bi-recursive convolutional neural network[J]. Acta Optica Sinica, 2017, 37(12): 1215003.
吴寿川, 赵海涛, 孙韶媛. 基于双向递归卷积神经网络的单目红外视频深度估计[J]. 光学学报, 2017, 37(12): 1215003.

[29] Xu R, Zhang J G, Huang K Q. Image super-resolution using two-channel convolutional neural

networks[J]. Journal of Image and Graphics, 2016, 21(5): 556-564.
徐冉, 张俊格, 黄凯奇. 利用双通道卷积神经网络的图像超分辨率算法[J]. 中国图象图形学报, 2016, 21(5): 556-564.

[30] 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.

[31] Li B, Shen C H, Dai Y C, *et al*. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs[C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 1119-1127.

[32] Wang P, Shen X H, Lin Z, *et al*. Towards unified depth and semantic prediction from a single image[C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 2800-2809.

[33] Li B, Dai Y C, He M Y. Monocular depth estimation with hierarchical fusion of dilated CNNs and soft-weighted-sum inference [J]. Pattern Recognition, 2018, 83: 328-339.

[34] Li J, Klein R, Yao A. A two-streamed network for estimating fine-scaled depth maps from single RGB images[C] // 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 3392-3400.

[35] Lee J H, Heo M, Kim K R, *et al*. Single-image depth estimation based on Fourier domain analysis [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 330-339.

[36] Ummenhofer B, Zhou H Z, Uhrig J, *et al*. DeMoN: depth and motion network for learning monocular stereo [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 5622-5631.

[37] 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, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 2002-2011.

[38] Jégou S, Drozdzal M, Vazquez D, *et al*. The one hundred layers tiramisu: fully convolutional DenseNets for semantic segmentation [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1175-1183.

[39] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[40] Kingma D P, Ba J. Adam: A method for stochastic optimization[EB/OL]. (2017-01-30)[2018-09-30]. https://arxiv.org/abs/1412.6980.

[41] Laina I, Rupprecht C, Belagiannis V, *et al*. Deeper depth prediction with fully convolutional residual networks[C] // 2016 Fourth International Conference on 3D Vision (3DV), October 25-28, 2016, Stanford, CA, USA. New York: IEEE, 2016: 239-248.

[42] Yin X C, Wang X W, Du X G, *et al*. Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields[C] // 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 5871-5879.

[43] Dimitrievski M, Goossens B, Veelaert P, *et al*. High resolution depth reconstruction from monocular images and sparse point clouds using deep convolutional neural network [J]. Proceedings of SPIE, 2017, 10410: 104100H.

[44] Mancini M, Costante G, Valigi P, *et al*. Toward domain independence for learning-based monocular depth estimation[J]. IEEE Robotics and Automation Letters, 2017, 2(3): 1778-1785.