浙江科技学院学报，第 32 卷第 4 期，2020 年 8 月
Journal of Zhejiang University of Science and Technology
Vol. 32 No. 4, Aug. 2020

doi：10.3969/j.issn.1671-8798.2020.04.002

基于多层特征融合的单目深度估计模型

叶　绿[a]，段　婷[b]，朱家懿[b]，Nwobodo Samuel Chuwkuebuka[a]，Annor Arnold Antwi[a]
（浙江科技学院 a.信息与电子工程学院；b.机械与能源工程学院，杭州 310023）

摘　要：为了获取信息完整的深度图以提高预测深度图的质量，解决单目深度估计模型中特征融合的问题，提出一种融合多尺度和不同层特征的双流神经网络模型。该模型采用 ResNet-50 残差网络结构提取深度特征信息，利用金字塔结构融合不同层次的图像特征，实现低层、中层和高层的特征融合，保证不同层次特征的有效互补，改善多层间特征信息的传递，在一定程度上避免了信息的遗漏和缺失。在 KITTI(Karlsruhe Institute of Technology and Toyota Technological Institute)数据集上进行试验，结果表明，该模型的均方根误差为 2.370 4，对数均方根误差为 0.229，平均对数误差为 0.118，阈值精度分别为 0.686、0.951、0.977，实现了较好的评测结果。

关键词：特征融合；双流神经网络；金字塔结构
中图分类号：TP391.41　　**文献标志码**：A　　**文章编号**：1671-8798(2020)04-0257-07

Monocular depth estimation model based on multi-level feature fusion

YE Lü[a], DUAN Ting[b], ZHU Jiayi[b], Nwodobo Samuel Chuwkuebuka[a], Annor Arnold Antwi[a]
(a. School of Information and Electronic Engineering; b. School of Mechanical and Engineering, Zhejiang Universitity of Science and Technology, Hangzhou 310023, Zhejiang, China)

Abstract: In order to obtain the depth map with complete information, boost the quality of prediction depth map, and solve the problem of feature fusion in monocular depth estimation model, a dual-stream neural network model was proposed integrating multi-scale and multi-layer features. In this model, ResNet-50 residual network structure was used to extract depth feature information, and pyramid structure was used to fuse image features of different levels to realize feature fusion of low, middle and high levels, so as to ensure the effective complementarity of features at different levels, improve the transmission of feature information among multiple levels, and to a certain extent, avoid the omission and lack of information. The experimental results on the KITTI (Karlsruhe Institute of Technology and Toyota Technological Institute) dataset show that the root mean square error of the model is 2.370 4, the root mean squared log error is 0.229, the average log10 error is 0.118, and the threshold accuracy is 0.686, 0.951 and 0.977, respectively, which

收稿日期：2019-10-30

通信作者：叶　绿(1962—　)，女，浙江省杭州人，教授，博士，主要从事人工智能和计算机视觉研究。E-mail：yelue@zust.edu.cn。

achieves sound evaluation results.

Keywords: feature fusion; dual-stream neural network; pyramid structure

深度估计在计算机视觉中占重要地位,广泛应用于三维重建[1]、场景理解等上下文环境,同时对语义分割、显著性目标检测、边缘检测等具有一定的改良作用。早期的单目深度估计主要利用单目线索,如阴影、运动和视差等方法,但这些方法都具有一定的局限性,预测出的深度图较模糊,质量不高,且带有噪声,纹理不清楚。要解决这些问题,首先要获得更丰富的特征信息。随着信息技术的发展,网络模型对特征提取的要求越来越高,不仅限于单层特征提取,还包括多层次的特征提取。近几年,卷积神经网络(convolutional neural networks,CNN)在深度估计中能够提取不同层次的特征,包括低层特征的细节信息和高层的抽象语义特征,以及介于二者之间的中层特征,但卷积层初始参数的选取会对特征提取产生较大的影响,如果选取不恰当,会导致整个网络训练效果不佳,因此它仍存在较大的改进空间。在特征提取不断优化的情况下,神经网络又面临特征融合问题。Chen 等[2]使用基于注意力的聚集网络(attention-based context aggregation network,ACCN)来捕获连续的上下文信息,并集成图像级和像素级上下文信息;钟海军等[3]提出了一种有效的特征融合方法,通过消除高层次之间的语义差距来提高分割质量。这些方法虽然在一定程度上改善了特征融合,但仍然不能有效地融合多尺度特征。陈好等[4]提出了多模态融合模块(complementarity-aware fusion,CA-Fuse),采用跨模态残差函数和补充性感知监督,从深层到浅层逐层对 RGB(red,green,blue)图像和深度图进行特征融合;吴磊等[5]采用了最经典最简单的融合方式——拼接,将不同层的特征以拼接的方式在通道维度上实现特征融合;余春艳等[6]通过跳层将整体嵌套边缘检测(holistically-nested edge detection,HED)结构的侧面输出特征图从深层向浅层传递,进行不同层的多尺度融合。通过这些方法实现了多层特征融合,但这些融合方式主要从高层向低层或从低层向高层进行特征融合,融合方向具有单一性,特征信息在融合的过程中不全面,忽略了中层特征信息。基于此,我们提出了一种新的单目深度估计模型,采用了两个对称的金字塔式结构(以下简称"金字塔"),从中层特征出发向低层和高层实现双向的特征融合,最后再用中层特征和通过"金字塔"互补后的特征进行融合,从而实现了中层到高层,中层到低层,以及中层与融合后的高层、低层三个层面的融合。

1 总体模型设计

双流神经网络由编码和解码两个部分构成,编码部分主要进行下采样以提取特征,解码部分主要进行上采样和特征融合。编码部分是一个基于 ResNet-50[7]的双流网络,包括卷积层、池化层、残差块及全连接部分,前三个部分用于下采样,最后一个全连接部分(包括平均池化层和全连接)的输出通道数为1 000(主要用于分类),而在模型的搭建过程中,不需要这一部分,因此将其移除。本研究将 ResNet-50 分成了 5 个小块,即卷积模块(Conv_bn)、残差模块 1(Res1)、残差模块 2(Res2)、残差模块 3(Res3)和残差模块 4(Res4)。每经过一个小块,特征图的尺寸变为原来的 1/2,其中 Conv_bn 与 Res1 位于浅层网络,输出的特征图为低层特征图;Res2 位于中层网络,输出的特征图为中层特征图;Res3、Res4 输出的特征图为高层特征图。该双流网络有 2 个支流,右边支流输入的是一个 320×512 像素的 RGB 图,从 Conv_bn 输入,依次进行 5 次下采样,尺寸变为原来的 $1/2^5$,即 10×16 像素;左边支流输入的是一个 1/2 大小的 RGB 图,即 160×256 像素,从 Res1 输入,经过 4 次下采样,尺寸变成10×16像素,与右边支流输出的尺寸一样,通过两个不同尺寸图片的输入可以获得不同的深度特征,将其在 ResNet-50 最后一个输出进行特征叠加,实现了多尺度信息的融合。为了在编码过程中减少参数,我们在双流网络中相同的部分采用了参数共享。为了使输出的深度图大小等于输入的 RGB 图像,需要对模型进行解码。本研究解码部分采用了上采样和下采样两个上下文感知的"金字塔",下采样"金字塔"通过两个反卷积层(deconv),上采样"金字塔"通过卷积层(conv)和池化层,使得分辨率与 Res2 的侧边输出相同。为了保证通道数相同,Res2 的输出经过一个卷积核大小为 1×1 的卷积层,然后再与两个"金字塔"的输出进行叠加求和。通过

卷积层改变通道数后输入初始注意力(inception-attention，IA)模块进行特征选择。该解码部分通过“金字塔”对不同层的特征进行提取、融合来解码特征信息，从而获得上下文感知的多尺度特征，并在不同的特征层中实现信息传输。网络整体结构如图 1 所示。

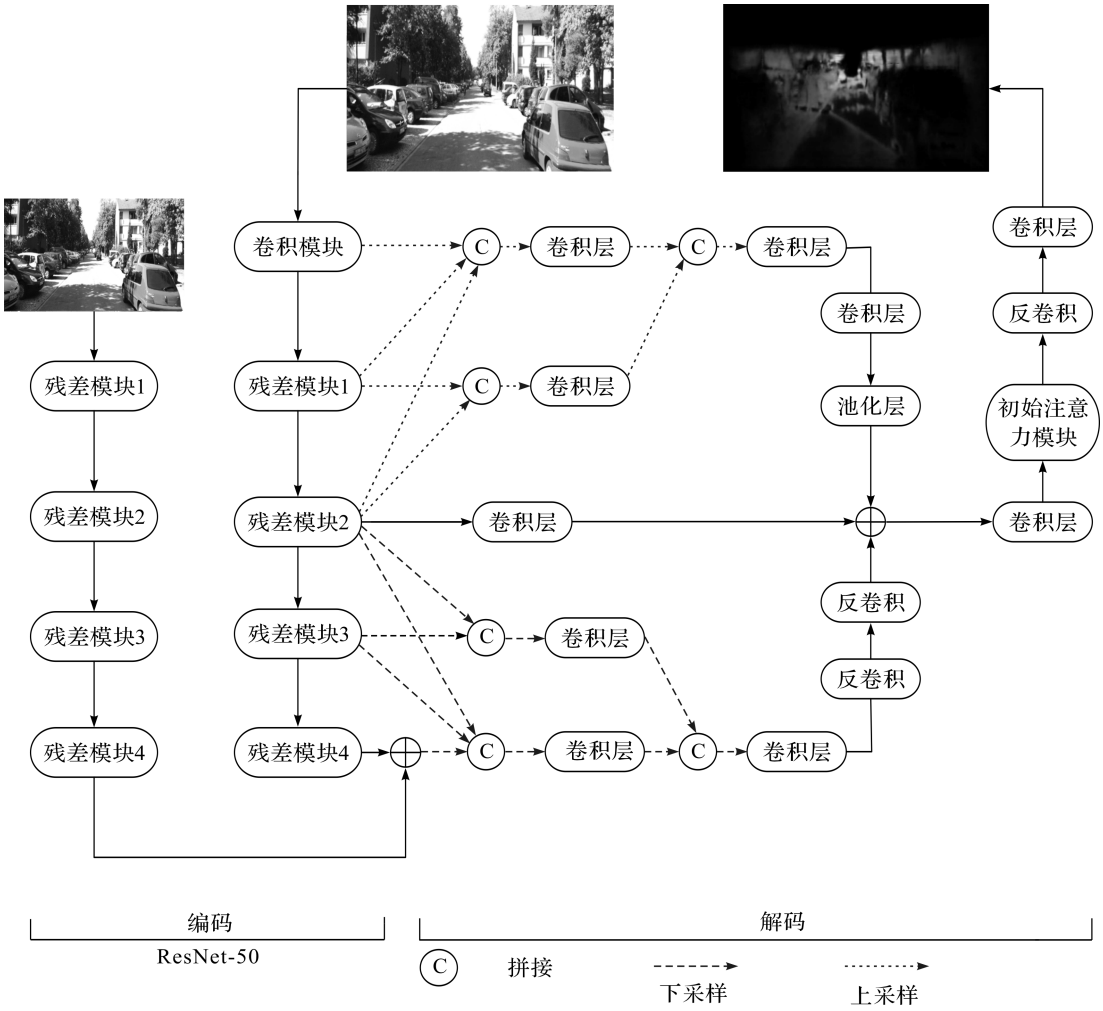


图 1　网络整体结构

Fig. 1　Overall architecture of network

2　模型方法

2.1　“金字塔”特征融合

在神经网络训练时，对局部和全局信息的提取非常重要。局部信息是关于特征图的一些细节信息，包括点、线、面的信息，这些局部信息的提取主要是在较低的层次上。全局信息是指语义信息，它是抽象的，一般从高层数据中获取，而在获取时通常会忽略中间层次的信息，且认为其不重要。实际上，中层特征介于高层特征和低层特征之间，既可以描述抽象的语义信息又包含了局部细节，深度估计不仅需要全局信息，而且还需要局部信息。因此，有效地实现各层信息融合是一个关键点，“金字塔”[8]融合图像可以起到一个很好的过渡作用。

图像经过上采样和下采样后信息会丢失，从而变得模糊。针对此问题，本研究采用了两种“金字塔”。上采样“金字塔”如图 2 所示，第一层由 3 个节点组成，每个节点包含了不同尺度的特征；接着将 Res1 和 Res2 的输出特征图分别通过上采样将尺寸扩大 2 倍和 4 倍，使其与 Conv_bn 的尺寸大小一致，并将它们拼接起来，作为第二层的第一个节点；然后将 Res2 的输出特征图通过上采样将尺寸扩大 2 倍后与 Res1 拼接，构建第二层上的第二个节点。同理，将第二层的第二个节点通过上采样将尺寸扩大 2 倍，与第二层

的第一个节点拼接作为第三层的第一个节点。下采样“金字塔”如图 3 所示，第一层也由 3 个节点组成，分别是 Res2、Res3 和 Res4，将 Res2 和 Res3 通过最大池化后分别经过下采样将尺寸缩小 1/4 和 1/2 后与 Res4 拼接，构成第二层中的第一个节点；同样地，将 Res2 经过下采样将尺寸缩小 1/2 后与 Res3 拼接，得到第二层的第二个节点；最后将第二层的第二个节点通过下采样缩小 1/2，与第二层的第一个节点拼接构成第三层的第一个节点。在每个拼接层后面都有一个卷积层用于通道数降维，且所有卷积层的卷积核都是 3×3。通过两个不同功能的“金字塔”，分辨率不断变化，有效地避免了融合过程中图像连通区域像素不连续的问题，实现了图像间的无缝连接。

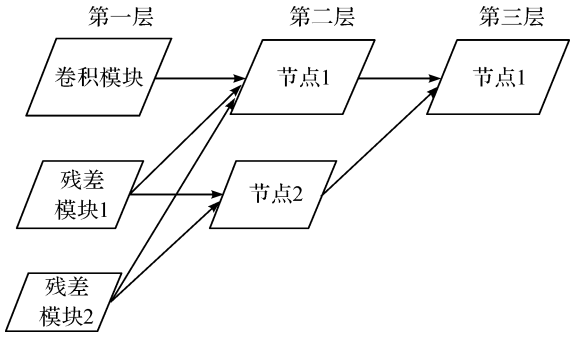


图 2 上采样“金字塔”

Fig. 2 Upsampling “pyramid”

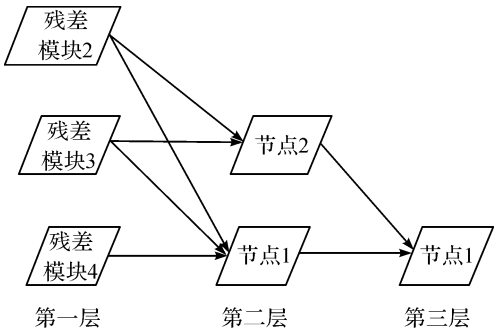


图 3 下采样“金字塔”

Fig. 3 Downsampling “pyramid”

2.2 IA 模块

网络深度的加深会导致参数的不断增加，增大对硬件的压力，导致模型的运行速度减慢。针对该问题，本研究在 Inception[9] 模块与注意力模块[10] 的基础上设计了 IA 模块，它结合 Inception 与注意力模块的优点，IA 模块如图 4 所示。IA 模块首先经过一个 1×1 的卷积层对输入的特征图进行通道数降维，再经过一个上采样层(upsample)将特征图尺寸放大 2 倍。上采样的方式选择使用双线性插值，使用双线性插值可使得上采样的特征图更平滑，因而可避免特征图中出现方块的现象，减少棋盘效应的出现；之后 IA 模块被分成 4 个分支，即 1×1 卷积层、3×3 卷积层、通道注意力(channel attention，CA)模块、空间注意力(spatial attention，SA)模块。

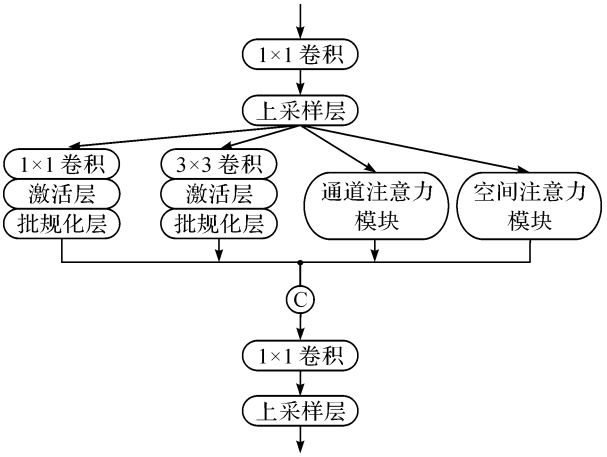


图 4 IA 模块

Fig. 4 Module of IA

通过不同大小的卷积核进行卷积，卷积层接收域大小要求不同，对于大的物体，需要更多的上下文信息来避免类间的不一致，这时需要更大的接收域来提取全局特征；同理，小的物体则需要更小的接收域来提取局部特征，才可以得到不同尺度的特征图。因此，本模块把两个卷积层分支输出的多尺度特征图与 CA 模块和 SA 模块的输出通过融合层进行拼接，进行不同尺度特征图的有效融合，拼接后的通道数变成了输入前的 4 倍，因此我们设计了 1×1 的卷积层对通道进行降维，再经过一个双线插值上采样层进行上采样。

2.3 注意力机制

在计算机视觉领域中，注意力机制是让计算机模仿人眼自觉地将注意力放在感兴趣的地方去学习图片。注意力模型在训练过程中学习图片时，给图片的每个部分赋予不同的权重，对感兴趣的地方给予较大的权重。因此，为了提取图片中的关键信息，让模型做出的判断更为准确，我们在 IA 模块中加入了注意力机制，在一定程度上减少了计算资源的消耗。

本文采用通道注意力机制和空间注意力机制这两种常用的注意力机制。CA 模块如图 5 所示，首先经过一个全局平均池化层(global average pooling，GAP)，将特征图整合成 1×1×C(C 为通道数)的形

式，高和宽都变成尺寸为 1，这时只对通道数进行改变，后面通过一个全连接的密集层（dense），用逻辑函数进行激活，将特征图标准化为 0 到 1 之间，将输入通过跳跃连接与输出进行点乘，将尺寸还原，特征图通过 CA 模块后，输出特征图的大小和通道数保持不变。SA 模块如图 6 所示，它主要作用在空间域中，将通道数变为 1 后，集中对特征图的高和宽进行处理。该模块由两个卷积层和一个逻辑函数及点乘组成，两个卷积层的卷积核宽和高是不等的，一个为 1×3，另一个为 3×1，对水平和垂直方向的特征进行捕捉，不同卷积核的大小产生不同的接收域，两个卷积层后面都有 Relu 非线性激活层与批规化层，经过卷积层后通过逻辑函数对特征图进行归一化，再通过一个跳层将 SA 模块的输入与归一化后的输出进行点乘，将尺寸和通道数恢复成原输入大小。

注意力模型的输出不改变特征图的尺寸和通道数量，但是经过注意力模型输出之后，模型会自动将注意力集中在图片的部分区域进行特征提取。将注意力模块加入 IA 模块中，替换了原 Inception 结构中的卷积层，为 Inception 结构增加了注意力机制，从而有效地改善了模型的性能。

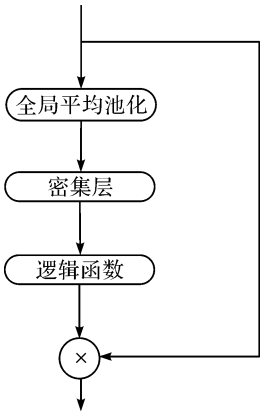


图 5　CA 模块

Fig. 5　CA module

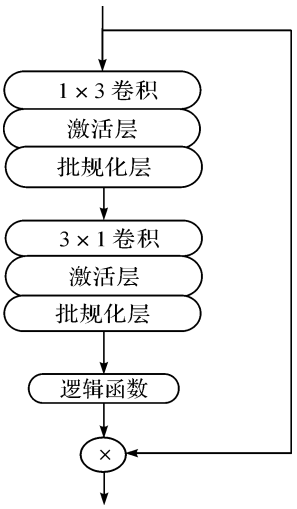


图 6　SA 模块

Fig. 6　SA module

3　试验分析

试验借助 Ubuntu 16.04 Keras 2.2.4 环境实现，编程语言采用 Python，使用 NVIDIA TITAN XP-12 GB 的显卡，数据集采用 KITTI（Karlsruhe Institute of Technology and Toyota Technological Institute）数据集[11]。KITTI 数据集包括了车、房子、树、行人，由于 KITTI 数据集的测试数据集标签还未在 KITTI 官网公布，因而对 KITTI 原始数据集中的训练集进行了随机拆分，将其分为 4 286 张图的训练集，343 张图的测试集，把数据集尺寸处理成 320×512 像素。训练时，将处理好的训练集输入搭建好的神经网络中，损失函数值最小时保存一个最优的模型权重，总共训练 20 个周期。测试时，把保存好的最优权重载入模型，将测试集输入神经网络中进行模型评估。损失函数选择均方误差（mean square error，MSE），用于估计预测深度图与对应标签的误差程度。用自适应矩估计优化器进行优化，以减少损失函数的值，学习率设置为 1e-4。

单目深度估计的评价指标有均方根误差（root mean squared error，RMS）、对数均方根误差（root mean squared log error，Log_RMS）、平均对数误差（average log10 error，Log10）、阈值精度 δ_i（$i=1,2,3$）6 个。前 3 个评价指标为误差指标，衡量预测值和标签之间的误差大小，值越小表示评测的结果越好，而阈值准确性表示精确度的指标，值越大表示评测的指标越好。

我们进行一系列对比试验来说明模型的有效性，不仅对比了一些经典模型，如 Eigen 等[12]、Laina 等[13]、Yin 等[14]的模型，同时也对比了一些新模型，如 Gur 等[15]、Heo 等[16]。这 5 个模型的复现符合原设定标准，试验环境配置同本文模型一致，训练集和测试集图片数量分别保持 4 286 张和 343 张。通过 6

个指标的对比来衡量预测深度图的优劣性，6 种模型的误差和精度比较结果见表 1(表中加粗的数值表示结果最优)。

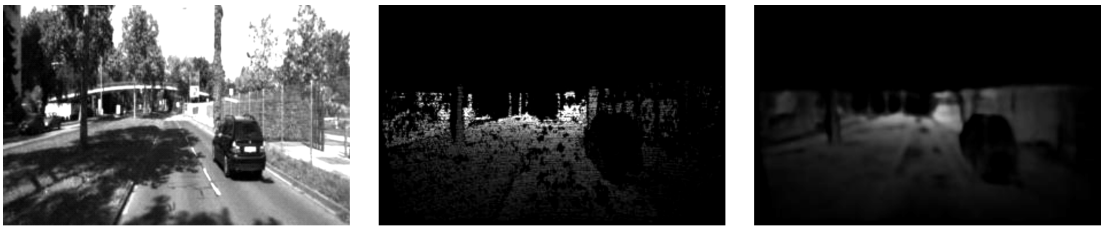
表 1 6 种模型的误差和精度比较结果

Table 1 Error and accuracy comparison results of six models

模型	误差			精度		
	RMS	Log_RMS	Log 10	δ_1	δ_2	δ_3
Eigen	2.644 0	0.272	0.167	0.488	0.948	0.972
Laina	2.461 8	0.243	0.126	0.674	0.943	0.972
Yin	2.519 3	0.247	0.134	0.640	0.947	**0.979**
Gur	2.399 6	0.234	0.120	0.684	0.949	0.975
Heo	2.408 2	0.235	0.121	0.685	0.946	0.974
本文模型	**2.370 4**	**0.229**	**0.118**	**0.686**	**0.951**	0.977

由表 1 可知，本文模型在评价指标上总体优于对比的其他 5 个模型。通过具体的数据分析可知，在参与对比的这 5 个模型中，近两年提出的模型测试出的指标要比复现经典模型的指标整体上有较大的提升。定量试验的目的是要将本研究模型测得的每个指标与之前最好的评测结果逐一对比，文献[15]的模型复现结果在 RMS、Log_RMS、Log 10 及 δ_2 上表现最优，依次为 2.399 6、0.234、0.120、0.949；文献[14]的模型复现结果在 δ_3 上最优，为 0.979；文献[16]的模型复现结果在 δ_1 上最优，为 0.685；与之前获得的模型性能相比，本文模型在 RMS 上相比最优的结果 2.399 6(文献[15])减少了 0.029 2。同理，在 Log_RMS 上相比 0.234 减少了 0.005，在 Log 10 上相比 0.120 的误差减少了 0.002。在 δ_1 和 δ_2 上相比 0.685 和 0.949 分别增加了 0.001 和 0.002，虽然 δ_3 比最优结果小 0.002，但是文献[14]中是降低了前 3 个误差指标的值来提升 δ_3 的值。本研究提出的模型上述前 5 个指标均是对比模型中最优的，因此本研究模型整体的评测结果指标优于对比的神经网络，这充分验证了本文神经网络的有效性。

本文模型的预测结果如图 7 所示。输入的彩色图片如图 7(a)所示；标签图片如图 7(b)所示，标签图是灰度图，通过深度相机拍摄而来，由一系列的像素点构成，每个像素点的值反映了该点的深度信息，因此通过深度图可以反映深度图中物体到摄像机的距离，不同的深度可以表示不同的距离；预测出的深度图片如图 7(c)所示，我们可以直观地看到，预测出的深度图比较接近标签图，该深度图可将整车轮廓及道路的场景输出。总体而言，生成的深度图信息比较完整，丢失的信息较少，得到的深度图更精确。因此，本文模型预测出了质量较高的深度图。



(a) 输入的彩色图片 (b) 标签图片 (c) 预测出的深度图片

图 7 本文模型的预测结果

Fig. 7 Predicted results by model of this paper

4 结 语

本研究提出了一种有监督单目深度估计模型，该模型通过双流网络对不同尺度的特征图进行融合，采用“金字塔”模块有效地提取多层特征，实现了不同层特征的传递和集成，避免了训练过程中的特征信息丢失，改进的 IA 模块有助于训练并提升了网络性能。试验结果表明，我们提出的模型优于对比的其他 5 种模型，可以预测出精确的深度图，从而验证了本模型的有效性。

参考文献：

[1] 陈加，张玉麒，宋鹏，等. 深度学习在基于单幅图像的物体三维重建中的应用[J]. 自动化学报，2019，45(4)：657.

[2] CHEN Y，ZHAO H T，HU Z W. Attention-based context aggregation network for monocular depth estimation[EB/OL]. (2019-01-29)[2019-10-04]. https://arxiv.org/pdf/1901.10137.pdf.

[3] 钟海军，胡步发. 基于高层特征融合的图像语义分割[J]. 机械制造与自动化，2019，48(3)：178.

[4] CHEN H，LI Y F. Progressively complementarity-aware fusion network for RGB-D salient object detection[C]//The IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE Computer Society，2018：3051.

[5] 吴磊，吕国强，薛治天，等. 基于多尺度递归网络的图像超分辨率重建[J]. 光学学报，2019，39(6)：90.

[6] 余春艳，徐小丹，钟诗俊. 融合去卷积与跳跃嵌套结构的显著性区域检测[J]. 计算机辅助设计与图形学学报，2018，30(11)：2150.

[7] 王若瑜. 基于 Resnet-50 的智能驾驶红绿灯分类研究[J]. 电子测试，2019(增刊 1)：145.

[8] 赵斐，张文凯，闫志远，等. 基于多特征图金字塔融合深度网络的遥感图像语义分割[J]. 电子与信息学报，2019，41(10)：2526.

[9] CAHALL D E，RASOOL G，BOUAYNAYA N C，et al. Inception modules enhance brain tumor segmentation[EB/OL]. (2019-07-12)[2019-11-04]. https://www.frontiersin.org/articles/10.3389/fncom.2019.00044/full.

[10] 杨康，宋慧慧，张开华. 基于双重注意力孪生网络的实时视觉跟踪[J]. 计算机应用，2019，39(6)：1654.

[11] GEIGER A，LENZ P，STILLER C，et al. Vision meets robotics：the KITTI dataset[J]. The International Journal of Robotics Research，2013，32(11)：1235.

[12] EIGEN D，PUHRSCH C，FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]//Advances in neural information processing systems. Montreal：Neural Information Processing Systems Foundation，2014：2368.

[13] LAINA I，RUPPRECHT C，BELAGIANNIS V，et al. Deeper depth prediction with fully convolutional residual networks[C]//International Conference on 3D Vision(3DV). Standod：IEEE，2016：241.

[14] YIN X，WANG X，DU X，et al. Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields[C]//2017 IEEE International Conference on Computer Vision. Venice：IEEE，2017：5873.

[15] GUR S，WOLF L. Single image depth estimation trained via depth from defocus cues[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Angeles：IEEE，2019：7687.

[16] HEO M，LEE J，KIM K R，et al. Monocular depth estimation using whole strip masking and reliability-based refinement[C]//Proceedings of the European Conference on Computer Vision. Munich：Spinger，2018：39.