# 基于帧间信息提取的单幅红外图像深度估计

顾婷婷[1]，赵海涛[1]，孙韶媛[2]

[1]华东理工大学信息科学与工程学院，上海 200237；

[2]东华大学信息科学与技术学院，上海 201620

**摘要**　针对红外图像存在纹理信息不丰富和边缘信息较少导致深度估计精度难以提高的问题，本文设计一种深层神经网络估计红外图像的深度，该网络融合了一个二维(2D)残差神经网络和一个三维(3D)卷积神经网络。传统单幅红外图像的深度估计方法遗漏了帧间信息，容易出现物体轮廓模糊甚至丢失的情况。在2D和3D网络输入端分别加入稠密光流和前后帧图像。进一步将3D卷积网络提取的视频特征与2D残差网络的特征图做权值连接。不同于传统神经网络的全连接层，全卷积层突破了输入图片的尺寸限制。实验结果表明，本文提出的红外图像深度估计方法具有较高的精度，估计出的物体轮廓更清晰完整。

**关键词**　图像处理；红外图像；深度估计；光流信息；残差神经网络；卷积神经网络

**中图分类号**　TP391.9　　**文献标识码**　A　　**doi**：10.3788/LOP55.061010

# Depth Estimation of Single Infrared Image Based on Interframe Information Extraction

Gu Tingting[1], Zhao Haitao[1], Sun Shaoyuan[2]

[1] *School of Information Science and Engineering, East China University of Science and Technology, Shanghai* 200237, *China*;

[2] *School of Information Science and Technology, Donghua University, Shanghai* 201620, *China*

**Abstract**　In view of lacking of the texture information and the edge information in the infrared image, the accuracy of depth estimation is hard to be improved. We propose a deep neural network to estimate the depth of infrared images. The network combines a two-dimensional (2D) residual neural network and a three-dimensional (3D) convolution neural network. The traditional methods of estimating the depth of a single infrared image omits the interframe information and is prone to fuzzy or even missing object contour. The 2D and 3D network inputs are added dense optical flow and the frame before and after the image, respectively. Secondly, the feature map extracted from the 3D convolutional network is further connected to the feature maps of the 2D residual network. Unlike the fully connected layer of the traditional neural network, fully convolutional layer breaks through the size constraints of the input. The experimental results show that the accuracy of the proposed infrared image depth estimation method is improved, and the object contour estimated is clear and complete.

**Key words**　image processing; infrared images; depth estimation; optical flow information; residual neural networks; convolutional neural networks

**OCIS codes**　100.4996; 150.0155; 100.5010

**收稿日期**：2017-11-21；**收到修改稿日期**：2017-12-22

**基金项目**：国家自然科学基金(61375007)、上海市科委基础研究项目(15JC1400600)

**作者简介**：顾婷婷(1992—)，女，硕士研究生，主要从事模式识别、计算机视觉等方面的研究。E-mail：18751973506@163.com

**导师简介**：赵海涛(1974—)，男，博士，教授，主要从事模式识别、计算机视觉等方面的研究。E-mail：haitaozhao@ecust.edu.cn

# 1 引 言

从单目图像中估计深度一直都是计算视觉中比较棘手的问题，其在场景理解[1]、三维(3D)重建[2]和机器人[3]方面都有着广泛的应用。人眼利用视差从单幅图像中能轻易地推断出其3D结构，但是计算机视觉因缺少直接可用的信息，深度估计仍是一项具有挑战性的任务。

对于深度估计的研究大多数是依赖于视觉信息，例如双眼视差、运动信息、散焦等，然而很多情况下无法得知这些视觉信息，故图像中物体和观察者的绝对距离就无法得知。针对这些问题，Torralba等[4]提出一种基于整体场景结构来估计深度的方法，此种方法不需要事先获取已知物体的尺寸，通过识别出图像中的物体结构推断出其大小比例来估计深度。Saxena等[5]结合诸如纹理变化和梯度等单目视觉信息，提出了一种马尔科夫随机场(MRF)算法估计深度，并将其用于很多室内和室外环境的3D重建，获得了很好的效果。不同于前几种直接从外观特征映射到深度的方法，Liu等[6]首先对场景进行语义分割，并使用语义标签指导3D重建，通过了解像素或区域的语义类，可以很容易地执行深度和几何约束(例如“天空”很远，“地面”是水平的)。此外，通过测量与给定语义类的外观差异，可以更简便地预测深度。

这些方法大多依赖于单目视觉特征的挖掘提取，前期的视觉特征提取容易遗漏一些重要特征且耗费时间。近年来，随着卷积神经网络(CNNs)在图像处理如视觉感知[7]、目标跟踪[8]，以及医疗影像[9]等领域的杰出贡献，人们对CNNs的关注度越来越高，AlexNet和VGGNet等网络结构取得了很好的实验效果[10-11]。在可见光图像的深度估计领域，Eigen等[12]基于CNNs提出一种多尺度方法，首先对单幅图像进行粗略特征提取，然后对提取的特征图进行精细的特征提取估计深度。然而CNNs随着网络层数的加深，容易出现梯度消失的问题，残差神经网络[13]有效地解决了这一问题。Laina等[14]基于残差和全卷积神经网络，设计一种新型网络估计深度并取得了很好的实验效果[15]。Li等[16]基于残差神经网络提出了一种层间连接的网络结构，充分利用网络中的前后特征进行深度估计。在神经网络的无监督学习方法中，Zhou等[17]提出的方法与许多其他方法类似[18-20]，都是基于视点合成的端到端学习方法作为监控信号，使用单视图深度和多视角姿态网络计算深度。

激光雷达在雾霾等天气状况下的使用受到限制，但红外热成像不受烟雾等条件的影响，红外热成像广泛应用于夜视领域，进一步结合景物的深度估计可用于夜间车辆的辅助驾驶，因此本文研究基于红外图像的景物深度估计。在红外图像的深度估计领域，红外图像存在色彩缺失、纹理信息不丰富等特点，使得其深度估计比可见光图像的深度估计更具挑战性。红外图像深度估计的传统方法大多依赖前期的手工特征提取，Sun等[21]提出的方法基于前期的特征提取，例如劳斯掩膜和灰度特征，然后结合支持向量机(SVM)，简单的神经网络(MLP)、决策树(DT)，以及K近邻(KNN)进行深度估计。此类方法对单幅图像的深度估计未充分利用图像的帧间信息，容易出现目标丢失的情况，估计出的深度精度也较低。许路等[22]提出先提取劳斯掩膜特征再结合卷积神经网络，并以此估计红外图像的深度的方法。吴寿川等[23]采用双向递归的视频序列信息传递机制，使得提取到的每一帧图像的特征都包含了视频前后文的序列信息来估计红外图像深度。何建梅等[24]通过特征点密度与图像边缘检测聚焦测度来实现场景深度估计，弥补了SML方法的错误深度估计值。

神经网络可以自动学习特征，不受限于传统的手工特征提取。本文提出了一种新型的网络估计红外图像深度，其中包含二维(2D)残差神经网络和3D卷积神经网络。光流反映前后帧物体的运动信息，为了充分利用帧间物体的运动信息，本文先计算单幅红外图像前后帧的两张稠密光流[25]图像，再与当前的红外图像做像素加和处理，即在输入层面充分利用帧间运动信息，加大运动物体的像素占比。3D卷积神经网络[26]在处理视频时，不同于传统2D卷积神经网络缺少时间轴信息，其能够自动提取图像的时间和空间维度上的特征。本文将当前红外图像及相邻帧的图像作为3D网络的输入，充分提取视频的帧间信息后与2D网络的特征图进行权值连接。由于传统全连接层为全局连接，无法感知局部信息，本文将其修改为感受野更大的卷积层。

# 2 基本原理

## 2.1 红外图像的光流处理

对于第一幅图像中的一个像素 $I(x,y,t)$，在 $\mathrm{d}t$ 时间段后，在第二幅图像中移动了 $(\mathrm{d}x,\mathrm{d}y)$ 距离，因为这些像素的强度不变，所以有：

$$I(x,y,t)=I(x+\mathrm{d}x,y+\mathrm{d}y,t+\mathrm{d}t),\quad(1)$$

取(1)式等号右边的泰勒级数近似值,除去常数项和除以 d$t$ 得到光流公式:

$$\frac{\partial I}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial I}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}+I_t=0,\quad(2)$$

式中 $I_t$ 为时间梯度。根据以上方法,对数据集中的连续红外图像进行光流计算。

不同于稀疏光流只针对图像上若干个特征点,稠密光流计算图像上所有点的偏移量。根据稠密光流的计算方法,本文事先计算出数据集的帧间稠密光流图像。

### 2.2 红外图像的深度标签处理

图 1 是由雷达生成的红外图像深度标签的散点极坐标图,图 1 中显示了距离中心点 0~100 m 距离,360°方向的散点,进一步截取其中相应视角的数据与原始红外图像配准作为该红外图像的深度标签。

为了网络最后的分类操作,对于深度标签中的每个深度值 $d_i$ 有:

$$S=\frac{\max(\ln d_i)}{N},\quad(3)$$

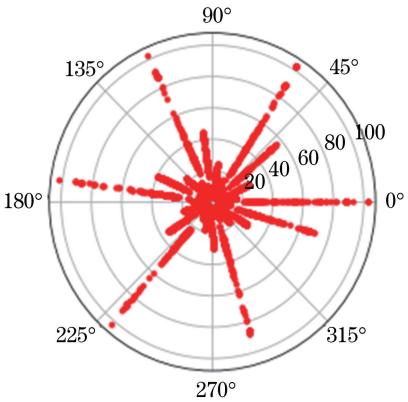


图 1 雷达散点图

Fig. 1 Radar scatter plot

在(3)式中,原始深度标签中的深度值被分为 $N$ 类,$S$ 为每一类对应的刻度值。对于同一个 $N$,根据对数特性,为了得出相同的刻度值,所有的深度值中离雷达近的点就会被密集分类,而远处的点就会被稀疏分类。在实际情况中,距离观察者越近的物体也希望被密集分类。对于不同的 $N$,如图 2 中取值,$N$ 越大,分类出的深度标签图更细致,也越接近原始深度图,本文选择 32 作为分类个数。

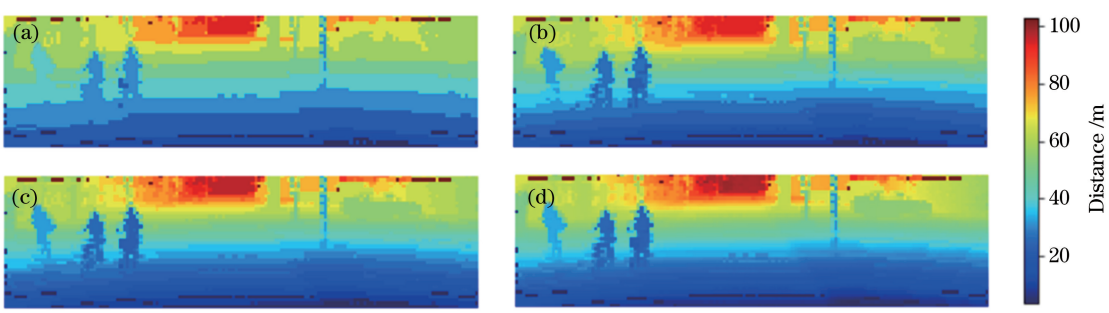


图 2 真实深度分层图。(a) 12 层;(b) 22 层;(c) 32 层;(d)原始深度图

Fig. 2 Hierarchies of the ground truth. (a) 12 hierarchies; (b) 22 hierarchies; (c) 32 hierarchies; (d) original ground truth

## 3 深层神经网络结构设计

红外图像对比可见光图像存在色彩和纹理方面的缺陷,但是红外图像在夜视领域更具优越性,能够显示出被遮挡物体,可见光在夜视下无法达到这种效果。图 3 为针对红外图像设计网络结构图,包括一个 2D 残差神经网络和一个 3D 卷积神经网络。对于当前单幅红外图像,取其前后一帧的图像,计算出其光流图像。光流图像中存在运动物体的信息,将其与当前单幅图像做像素级加和作为 2D 网络输入,而将连续三帧图像作为 3D 网络输入。红外图像分辨率为 576 pixel×160 pixel,深度标签为 144 pixel×40 pixel大小。图中注释如 16 @3×3 表示该操作的卷积核大小为 3×3,产生的特征图个数为 16。3 个 Residual block 中的卷积步长如表 1 所示,最终所得特征图的大小为原输入图的 1/4,即 144 pixel×40 pixel。3D 卷积神经网络中全连接最后一层的神经元个数设定为 5760,与 2D 输出特征图神经元个数一致。

### 3.1 2D 残差神经网络

对于神经网络而言,网络层数越深,特征提取越多。而传统深度神经网络随着网络层数的加深,在反向传播过程中会出现梯度消失的问题,导致训练效果变差。但残差神经网络[27]有效地解决了这一难点,对于一个残差的基本单元:

$$X_{l+1}=\mathrm{Loss}(X_l,W_l,b_l)+X_l,\quad(4)$$

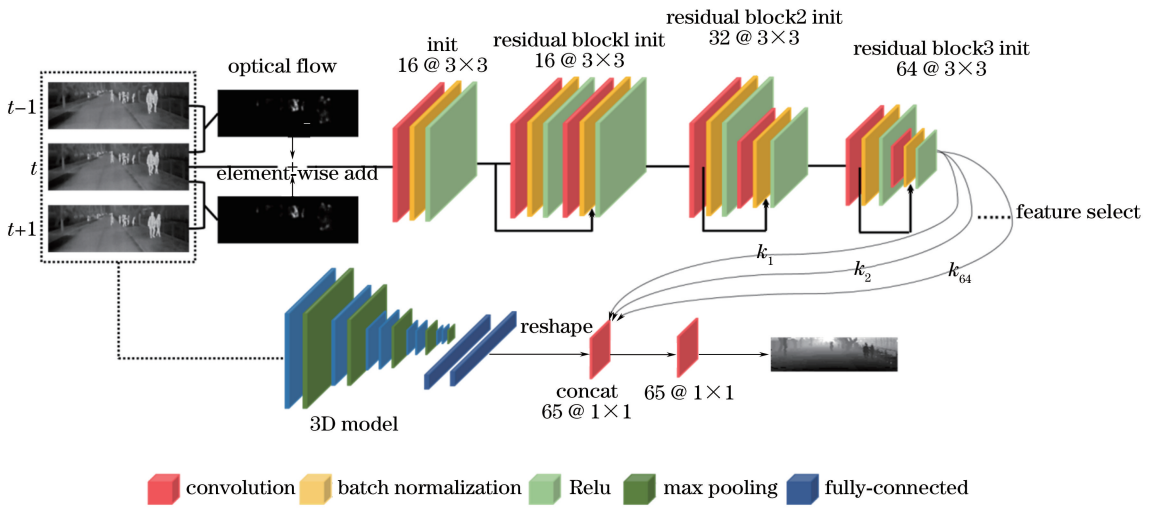


图 3 网络结构图

Fig. 3 Architecture of proposed network

式中 $X_l$ 为第 $l$ 层的输入，对应的损失函数为 Loss。$X_{l+1}$ 为第 $l$ 层的输出：

$$\frac{\partial X_L}{\partial X_l}=\frac{\partial\left[\mathrm{Loss}(X_l,W_l,b)+X_l\right]}{\partial X_l}=1+\frac{\partial \mathrm{Loss}(X_l,W_l,b)}{\partial X_l},\tag{5}$$

式中 $W_l$ 和 $b_l$ 分别为第 $l$ 层的权值和偏置。通过(5)式求偏导可以看出，反向传播时即使网络层数加深也不会出现梯度消失的问题。

表 1 2D 网络残差块卷积步长

Table 1 Convolutional kernel stride of 2D network residual blocks

| Layer | Conv1 kernel stride | Conv2 kernel stride |
|---|---|---|
| Init | 1×1 | 1×1 |
| Residual block1 init | 1×1 | 1×1 |
| Residual block2 init | 2×2 | 1×1 |
| Residual block2 init | 2×2 | 1×1 |

在 2D 卷积操作中，第 $l$ 层的第 $j$ 个特征图的 $(x,y)$ 位置单元的值 $X\_c_{lj}^{xy}$ 表示为

$$X\_c_{lj}^{xy}=F\_c2d\left(\sum_m\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}W_{ljm}^{pq}\times X\_c_{(l-1)m}^{(x+p)(y+q)}+b_{lj}\right),\tag{6}$$

式中 $Q_i$ 和 $P_i$ 分别为卷积核的高和宽，$W$ 和 $b$ 为需要学习的参数。例如初始残差模块的输入 1 个 $X\_C_{l-1}$ 特征图，输出 32 个 $X\_C_l$ 特征图，做卷积核大小为 3×3 的卷积操作，则权重 $\boldsymbol{W}$ 为[3,3,1,16]的矩阵，偏置 $b$ 为 32 维。卷积操作后使用 Relu 激活函数，缩短学习周期：

$$\mathrm{Relu}(X\_c_l)=\begin{cases}0, & X\_c_l<0\\ X\_c_l, & X\_c_l\geqslant 0\end{cases}。\tag{7}$$

残差模块中的批量归一化防止内部协变量转变，本文一个批次大小为 4：$B=\{X_{1,\cdots,4}\}$，当前层的输出$\{y_i=BN_{w,b}\}$，参数 $W$ 和 $b$ 更新如下：

$$\begin{cases}\mu_B\leftarrow\dfrac{1}{4}\sum_{i=1}^{4}X_i\\ \sigma_B^2\leftarrow\dfrac{1}{4}\sum_{i=1}^{4}(X_i-\mu_B)^2\\ \hat{X}_i\leftarrow\dfrac{X_i-\mu_B}{\sqrt{\sigma_B^2+\varepsilon}}\\ y_i\leftarrow W\hat{X}_i+b\equiv BN_{w,b}\end{cases},\tag{8}$$

式中 $\varepsilon=0.001$。

### 3.2 3D 卷积神经网络

2D 神经网络的输入综合了运动物体的光流特征来提取特征，卷积操作都是基于 2D 网络。为了综合运动信息，本文将引入 3D 网络，将连续三帧的红外图像作为该网络的输入，自动提取视频的时间和空间特征，如图 4 所示。

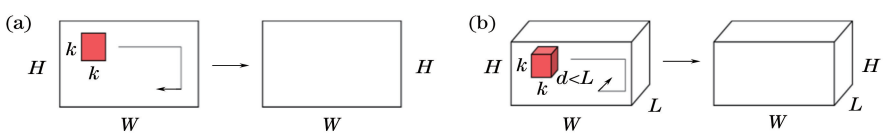


图 4 (a) 2D 和(b) 3D 卷积对比

Fig. 4 Comparison of (a) 2D convolution and (b) 3D convolution

与 2D 卷积操作不同的是图 4 中 3D 卷积操作能够提取时间和空间维度上的特征：

$$X\_c_{lj}^{xyz}=F\_c3d\left(\sum_m\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1}W_{ljm}^{pqr}X\_c_{(l-1)m}^{(x+p)(y+q)(z+r)}+b_{lj}\right), \tag{9}$$

相比与(6)式中的 2D 卷积,3D 卷积多了一个时间维度 $z$,$Q_i$、$P_i$、$R_i$ 为卷积核的三维大小,$W$ 和 $b$ 为需要学习的权值和偏置。

本文 3D 卷积神经网络定义如表 2 所示,与 2D 的 VGGNet 相比,该 3D 网络结构层数减少,卷积核变成三维大小。

表 2 3D 网络结构

Table 2 Architecture of 3D network

| Layer | Channel | Kernel | Stride | Layer | Channel | Kernel | Stride |
|---|---|---|---|---|---|---|---|
| Conv1 | 64 | 1×1×1 | 1×1×1 | Conv4a | 512 | 1×1×1 | 1×1×1 |
| Pool1 | 64 | 1×2×2 | 1×2×2 | Conv4b | 512 | 1×1×1 | 1×1×1 |
| Conv2 | 128 | 1×1×1 | 1×1×1 | Pool4 | 512 | 2×2×2 | 2×2×2 |
| Pool2 | 128 | 2×2×2 | 2×2×2 | Conv5a | 512 | 1×1×1 | 1×1×1 |
| Conv3a | 256 | 1×1×1 | 1×1×1 | Conv5b | 512 | 1×1×1 | 1×1×1 |
| Conv3b | 256 | 1×1×1 | 1×1×1 | Pool5 | 512 | 2×2×2 | 2×2×2 |
| Pool3 | 256 | 2×2×2 | 2×2×2 | Fc | - | - | |

为了综合 2D 网络和 3D 网络的特征,本文提出一种特征选择方法。在 2D 网络提取到的 64 张特征图中,为了让神经网络能自动学习其重要程度,对 64 张特征图各设定一个权重 $k_i$,由于 3D 网络最后的全连接层含有 5760 个神经元,可将其转换为 144×40的特征图,然后将与 2D 的 64 张特征图做权值连接,组成一个有 65 张特征图的特征提取组:

$$F_{\text{usion}}=\text{concat}\left[k_i\times(X_{2\text{d}}+\gamma),X_{3\text{d}}\right],\quad i=1,2,\cdots,64, \tag{10}$$

式中 concat 操作将 $X_{2\text{d}}$和 $X_{3\text{d}}$特征图纵向连接作为全卷积层的输入,为了防止出现学习到的参数全为零的情况,定义 $\gamma=1$。

特征融合后产生 144×40×65 个节点,传统的全连接层是全局连接,等同于 1×1 的卷积核对输入分辨率要求固定。但全卷积层是局部连接即稀疏的全连接,对输入大小没有限制,能够增大图像的感受野。根据预测深度 $m_i$ 和深度标签 $m_i^*$ 定义损失函数为

$$Loss(m,m^*;\alpha)=\lambda\{w\}-\frac{1}{N}\sum_i m_i^*\ln m_i, \tag{11}$$

式中 $w$ 为神经网络定义的参数集合,$\lambda$ 为惩罚项系数,可由神经网络自主学习,以防在更新参数集合 $\{w\}$时过拟合。

## 4 网络训练

实验数据集 NUSTMS 由南京理工大学车载红外和激光雷达采集,采集设备包括图 5 所示的 64 线激光雷达和分辨率为 768 pixel×576 pixel 的远红外热像仪。在红外成像的同时可以得到测量视场中物体轮廓和设备间的相对距离,图 6 分别展示了热红外和其对应时刻的雷达散点图,并将雷达测定的距离进行图 2 所示的深度分类作为训练标签。实验数据集中红外图像的大小为 576 pixel×160 pixel,深度标签尺寸为 144 pixel×40 pixel。

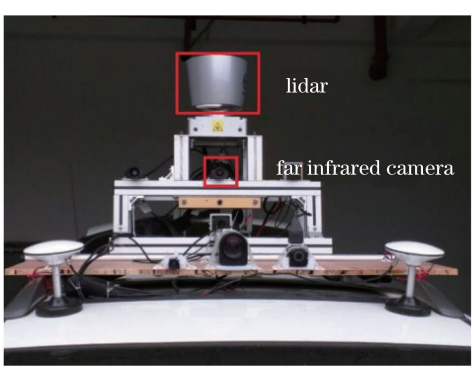


图 5 采集数据设备

Fig. 5 Data acquisition equipment

实验的数据集由 3000 张训练集和 1000 张测试集组成。设定最大步长为 60000,一个批次大小为 6,使用 AdamOptimizer 优化器最小化损失函数,从而计算反向传播。对 $l$ 层的输出 $X_l^j$ 卷积操作:

$$X_l^j=F\left(\sum_{i\in M_j}X_{l-1}^i*W_l^{ij}+b^j\right), \tag{12}$$

式中 $W_l^{ij}$ 为卷积核,令 $u_l$ 为(12)式括号内的表达式更新参数如下:

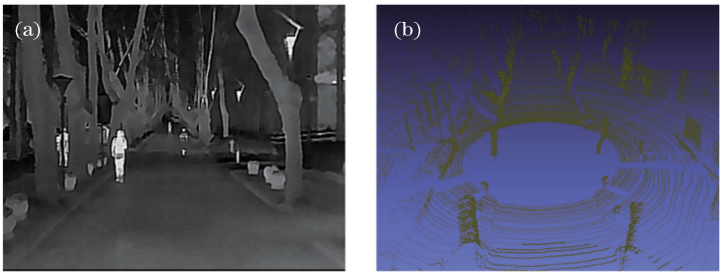


图 6 红外成像和对应时刻的雷达散点图。(a) 红外成像;(b) 雷达散点

Fig. 6 Infrared imaging and radar scatter plot at corresponding time. (a) Infrared imaging; (b) radar scatter points

$$\begin{cases}\delta_l=\dfrac{\partial Loss}{\partial X_l}\\ \delta_l^j=\delta_{l+1}^j W_{l+1}^j\cdot F'(u_l)\\ \dfrac{\partial L}{\partial w_l}=\sum_{u,v}(\delta_l^j)_{u,v}(p_{l-1}^i)_{u,v}\\ \dfrac{\partial L}{\partial b^j}=\sum_{u,v}(\delta_l^j)_{u,v}\end{cases},\tag{13}$$

式中·为点乘,$(p_{l-1}^i)_{u,v}$为 $X_l^j$ 与 $W_l^{ij}$ 做卷积操作时对应的区域块,$(u,v)$为区域中心。

## 5 实验结果与分析

对于红外图像的深度估计,传统方法大多是基于前期高维特征提取,并结合 MLP、SVM,以及 KNN 等机器学习方法估计出深度。图 7 给出了本文方法和 4 种传统方法的对比,由图 7 可知,传统方法和本文方法相比清晰度很低,物体轮廓也不完整。本文方法与深度标签图对比有细微差别,而 4 种特征提取方法和深度标签图相差较大,只能大致反映物体的整体轮廓。为了量化分析其区别,图 8 给出了本文和 4 种方法对应深度标签的拟合图,由于一张图片有 9 万多像素,本文随机选取 100 个像素。由图 7 可知,本文方法和深度标签拟合度较高,但是 4 种传统方法拟合效果很差,MLP 和 SVM 两种方法波动范围很小,只是反应了波动趋势,在预测的深度图像中表示为所有区域的明暗变化不明显;KNN 和 DT 拟合时波动范围太大,在预测深度图像中表示为明暗变化明显,但跟深度标签差距较大,这 4 种传统方法都不能很好地估计深度。

为了在测试集上评价本文方法的性能指标,深度估计领域通用的评价指标分准确率和误差[21-23]:

1) 平均相对误差(REL):$\dfrac{1}{N}\sum_{P^{(i)}\in N}\dfrac{|P_{\mathrm{gt}}^{(i)}-P^{(i)}|}{P_{\mathrm{gt}}^{(i)}}$。

2) 均方根误差(RMSE):$\sqrt{\dfrac{1}{N}\sum_{P^{(i)}\in N}(P_{\mathrm{gt}}^{(i)}-P^{(i)})^2}$。

3) 平均 log 10 误差:$\dfrac{1}{N}\sum_{P^{(i)}\in N}|P_{\mathrm{gt}}^{(i)}-P^{(i)}|$。

4) 阈值:$p$ 的百分比 s.t. $\max\left(\dfrac{P_{\mathrm{gt}}^{(i)}}{P^{(i)}},\dfrac{P^{(i)}}{P_{\mathrm{gt}}^{(i)}}\right)=\delta<t_{\mathrm{thr}}$。式中 $P$、$P_{\mathrm{gt}}^{(i)}$ 分别为深度标签和预测深度值。

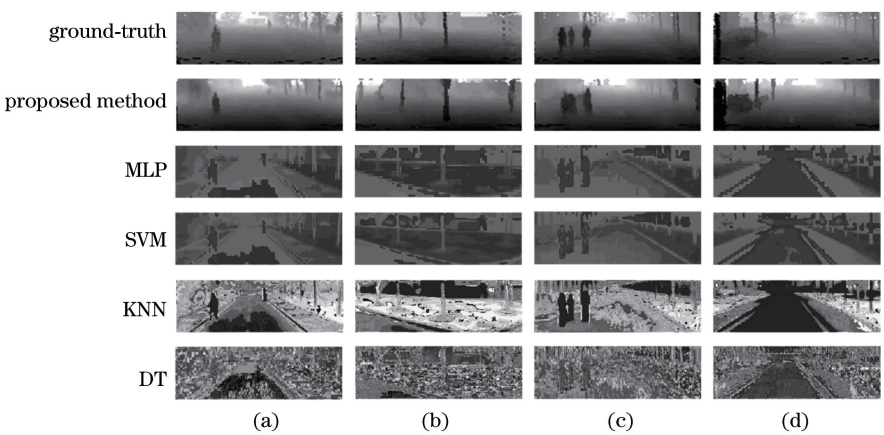


图 7 传统方法对比。(a)场景 1;(b)场景 2;(c)场景 3;(d)场景 4

Fig. 7 Comparison of traditional methods. (a) Scenario1; (b) scenario2; (c) scenario3; (d) scenario4

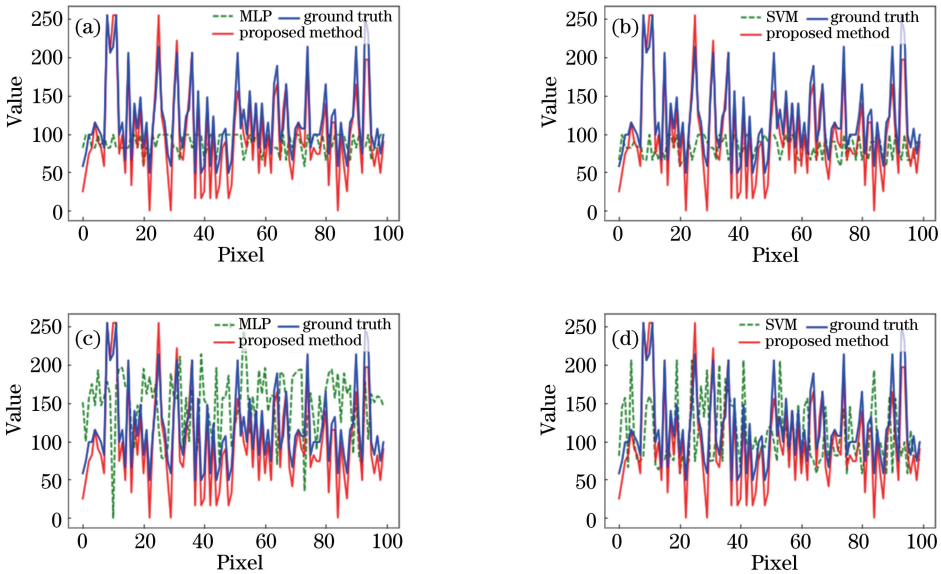


图 8 本文方法与传统方法对比

Fig. 8 Comparison of proposed method with traditional methods

性能评价指标结果如表 2 所示，本文给出 5 组对比实验，由评价指标可知，本文给出的方法在准确率和误差方面表现性能最优。这些对比实验都是针对单幅图像，没有综合利用帧间信息，从而会出现特征提取不充分的问题。本文方法首先结合了运动物体的光流信息，又利用 3D 卷积神经网络提取了视频时间和空间上的信息，最后综合二者的特征估计深度，实验效果较好。对比实验由神经网络方法和传统方法组成，由表 3 根据评价指标可得出结论：在深度估计方面，神经网络方法在准确率和误差上的表现好于传统方法；本文方法的表现优于其他 5 种神经网络方法。由图 7 和图 10 可知，传统方法的清晰度远低于神经网络方法，而且物体的轮廓模糊不清。

表 3 对比实验性能评价

Table 3 Performance evaluation of comparative experiments

| Method | $\delta<1.25$ | $\delta<1.25^2$ | $\delta<1.25^3$ | REL | log 10 | RMSE |
|---|---|---|---|---|---|---|
| Proposed method | **0.769** | **0.899** | **0.938** | **0.196** | **0.081** | **3.112** |
| method | 0.752 | 0.890 | 0.925 | 0.218 | 0.087 | 3.201 |
| ResNet-32 | 0.741 | 0.875 | 0.919 | 0.241 | 0.087 | 3.285 |
| FCN-Vgg | 0.737 | 0.879 | 0.913 | 0.251 | 0.080 | 3.290 |
| method | 0.625 | 0.826 | 0.892 | 0.297 | 0.115 | 4.098 |
| FCN-AlexNet | 0.575 | 0.806 | 0.890 | 0.320 | 0.121 | 4.101 |
| MLP | 0.175 | 0.377 | 0.601 | 6.698 | 0.294 | 9.883 |
| SVM | 0.182 | 0.379 | 0.622 | 6.808 | 0.298 | 9.615 |
| KNN | 0.182 | 0.656 | 0.505 | 5.420 | 0.385 | 10.250 |
| DT | 0.257 | 0.508 | 0.698 | 6.242 | 0.276 | 10.323 |

网络训练完成后，对测试集进行深度估计。图 9为随机从测试结果中选出的 4 张预测深度结果，做出其分别与深度标签的数据分布图。从图 9(a)～(d)可知，相应的深度标签和预测出的深度分布大体吻合，并且二者的相关性很高，尤其是在距离为 50～100 m 时，预测结果和深度标签相关性最高，表明在这个范围内预测结果很精确。由相关性系数可以看出，这 4 组图中，(b)组预测效果最差，(c)组效果最好。为了更直观地展现预测效果，图 10列出 4 组对比实验结果，场景(1)～(4)分别与图 7(a)～(d)对应。可以看出，神经网络方法的预测结果优于以上 4 种传统方法，然而这 4 种神经网络方法也存在差异，由上至下预测效果逐渐变差。对比每一个场景可以发现，本文方法预测出的深度清晰度跟原深度标签对比很接近，优于其他 5 种方法。以场景 3 和场景 4 为例，本文方法估计出的行人轮廓和路边车辆轮廓优于其他方法，说明本文网络能充分地提取帧间信息特征。

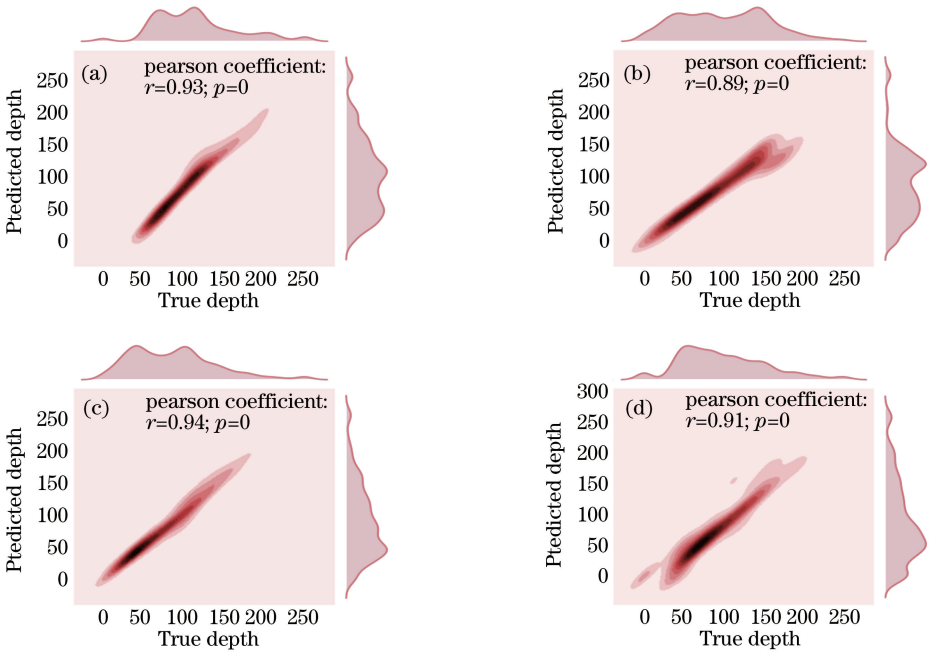


图 9 数据联合分布图。(a)场景 1;(b)场景 2;(c)场景 3;(d)场景 4

Fig. 9 Data joint distribution. (a) Scenario1; (b) scenario2; (c) scenario3; (d) scenario4

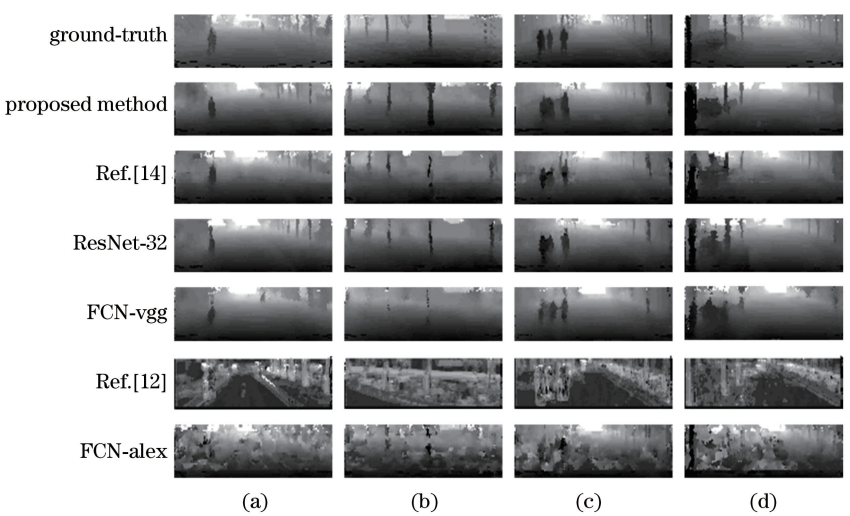


图 10 实验结果对比。(a)场景 1;(b)场景 2;(c)场景 3;(d)场景 4

Fig. 10 Comparison of experimental results. (a) Scenario1; (b) scenario2; (c) scenario3; (d) scenario4

## 6 结 论

激光雷达成本高,且在雾霾等天气状况下使用效果不佳,而红外热成像不受天气情况限制,故而本文研究基于红外图像的深度估计。传统红外图像估计深度的方法大多局限于当前图像本身,而忽略了图像的帧间信息。因此,提出了一种新颖的方法估计红外图像的深度。因光流能很好地反映运动物体的信息,将光流和当前红外图像做像素加和,以此作为本文 2D 残差神经网络的输入,增大运动物体的像素占比,后续的神经网络自动进行特征提取中,这些存在的运动目标就不易出现模糊甚至丢失的问题。此外还将当前图像和其前后一帧的图像作为 3D 卷积神经网络的输入,在时间和空间维度上提取视频特征。将两个网络的特征连接作为下一层输入时,提出一种特征选择方法,赋予每个特征图不同的权重,更新参数后网络可与很好地进行重要和次要特征选择。

在未来的研究工作中,期望能将研究工作进一步拓展,在深度图中能做到目标识别,在估计出深度

信息的同时，加上物体的类型信息，判断物体的类别，可以进一步为夜间车辆的辅助驾驶提供帮助。

## 参 考 文 献


[1] Lin D H, Fidler S, Urtasun R. Holistic scene understanding for 3D object detection with RGBD cameras[C]// Proceedings of IEEE International Conference on Computer Vision, 2013: 1417-1424.

[2] Saxena A, Chung S H, Ng A Y. 3-D depth reconstruction from a single still image [J]. International Journal of Computer Vision, 2008, 76 (1): 53-69.

[3] Biswas J, Veloso M. Depth camera based indoor mobile robot localization and navigation [C]// Proceedings of IEEE International Conference on Robotics and Automation, 2012: 1697-1702.

[4] Torralba A, Oliva A. Depth estimation from imagestructure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(9): 1226-1238.

[5] Saxena A, Schulte J, Ng A Y. Depth estimation using monocular and stereo cues[C]// Proceedings of the 20th International Joint Conference on Artifical Intelligence, 2007: 2197-2203.

[6] Liu B Y, Gould S, Koller D. Single image depth estimation from predicted semantic labels [C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010: 1253-1260.

[7] Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115 (3): 211-252.

[8] Guo J, Gould S. Deep CNN ensemble with data augmentation for object detection [J]. Computer Science, 2015. https://arxiv.org/pdf/1506.07224.pdf.

[9] Dou Q, Chen H, Yu L Q, *et al*. Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection[J]. IEEE Transactions on Biomedical Engineering, 2017, 64(7): 1558-1567.

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

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014. https://arxiv.org/pdf/1409.1556.pdf.

[12] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network [J]. Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014, 2: 2366-2374.

[13] He K M, Zhang Z Y, Ren S Q, *et al*. Deep residual learning for image recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[14] Laina I, Rupprecht C, Belagiannis V, *et al*. Deeper depth prediction with fully convolutional residual networks[C]// Proceedings of 4th International Conference on 3D Vision, 2016: 239-248.

[15] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015, 79(10): 3431-3440.

[16] Li B, Dai Y C, Chen H H, *et al*. Single image depth estimation by dilated deep residual convolutional neural network and soft-weight-sum inference[J]. Computer Science, 2017. https://arxiv.org/pdf/1705.00534.pdf.

[17] Zhou T H, Brown M, Snavely N, *et al*. Unsupervised learning of depth and ego-motion from video [C]// Proceedings of IEEE Conference on CComputer Vision and Pattern Recognition, 2017: 6612-6619.

[18] Garg R, Vijay K B G, Carneiro G, *et al*. Unsupervised CNN for single view depth estimation: geometry to the rescue[C]// Proceedings of European Conference on Computer Vision, 2016: 740-756.

[19] Godard C, mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6602-6611.

[20] Kuznietsov Y, Stückler J, Leibe B. Semi-supervised deep learning for monocular depth mappprediction[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6647-6655.

[21] Sun S Y, Li L N, Xi L. Depth estimation from monocular infrared images based on BP neural network model [C]// Proceedings of International Conference on Computer Vision in Remote Sensing, 2012: 237-241.

[22] Xu L, Zhao H T, Sun S Y. Monocular infrared

image depth estimation based on deep convolutional neural networks[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.

[23] 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.

[24] He J M, Qiu J, Liu C. Fusing feature point density and edge information for scene depth estimation[J]. Laser & Optoelectronics Progress, 2017, 54(7): 071101.
何建梅, 邱钧, 刘畅. 融合特征点密度与边缘信息的场景深度估计[J]. 激光与光电子学进展, 2017, 54(7): 071101.

[25] Farnebäck G. Two-frame motion estimation based on polynomial expansion [C] // Proceedings of Scandinavian Conference on Image Analysis, 2003: 363-370.

[26] Ji S W, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[27] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.