Opto-Electronic Engineering
光 电 工 程

Article

2021 年，第 48 卷，第 12 期

DOI: 10.12086/oee.2021.210340

基于点云与图像交叉融合的道路分割方法

张　莹，黄影平*，郭志阳，张　冲

上海理工大学光电信息与计算机工程学院，上海 200093

摘要：道路检测是车辆实现自动驾驶的前提。近年来，基于深度学习的多源数据融合成为当前自动驾驶研究的一个热点。本文采用卷积神经网络对激光雷达点云和图像数据加以融合，实现对交通场景中道路的分割。本文提出了像素级、特征级和决策级多种融合方案，尤其是在特征级融合中设计了四种交叉融合方案，对各种方案进行对比研究，给出最佳融合方案。在网络构架上，采用编码解码结构的语义分割卷积神经网络作为基础网络，将点云法线特征与 RGB 图像特征在不同的层级进行交叉融合。融合后的数据进入解码器还原，最后使用激活函数得到检测结果。实验使用 KITTI 数据集进行评估，验证了各种融合方案的性能，实验结果表明，本文提出的融合方案 E 具有最好的分割性能。与其他道路检测方法的比较实验表明，本文方法可以获得较好的整体性能。

关键词：自动驾驶；道路检测；语义分割；数据融合

中图分类号：TP391.41　　**文献标志码：**A

张莹，黄影平，郭志阳，等. 基于点云与图像交叉融合的道路分割方法[J]. 光电工程，2021，**48**(12): 210340
Zhang Y, Huang Y P, Guo Z Y, *et al*. Point cloud-image data fusion for road segmentation[J]. *Opto-Electron Eng*, 2021, **48**(12): 210340

Point cloud-image data fusion for road segmentation

Zhang Ying, Huang Yingping*, Guo Zhiyang, Zhang Chong

School of Optical-Electronic and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract: Road detection is the premise of vehicle automatic driving. In recent years, multi-modal data fusion based on deep learning has become a hot spot in the research of automatic driving. In this paper, convolutional neural network is used to fuse LiDAR point cloud and image data to realize road segmentation in traffic scenes. In this paper, a variety of fusion schemes at pixel level, feature level and decision level are proposed. Especially, four cross-fusion schemes are designed in feature level fusion. Various schemes are compared, and the best fusion scheme is given. In the network architecture, the semantic segmentation convolutional neural network with encoding and decoding structure is used as the basic network to cross-fuse the point cloud normal features and RGB image features at different levels. The fused data is restored by the decoder, and finally the detection results are obtained by using the activation function. The substantial experiments have been conducted on public KITTI data set to evaluate the

收稿日期：2021-10-30；　收到修改稿日期：2021-12-13
基金项目：上海市自然科学基金资助项目(20ZR1439007)；国家自然科学基金资助项目(61374197)
作者简介：张莹(1996-)，女，硕士研究生，主要从事计算机视觉、机器学习的研究。E-mail：192420365@st.usst.edu.cn
通信作者：黄影平(1966-)，男，教授，主要从事汽车电子、计算机视觉的研究。E-mail：huangyingping@usst.edu.cn

performance of various fusion schemes. The results show that the fusion scheme E proposed in this paper has the best segmentation performance. Compared with other road-detection methods, our method gives better overall performance.

Keywords: autonomous driving; road detection; semantic segmentation; data fusion

1 引 言

道路检测是自动驾驶中环境辨识的重要内容，是车辆实现自动驾驶的前提。目前，自动驾驶车辆大多采用多传感器数据融合的方式实现对道路的检测。其中最为常见的是将激光雷达数据与 RGB 图像数据进行融合，现有的研究表明将这两种传感器的数据进行融合，可以提高道路检测精度。最新的融合方法是采用卷积神经网络(convolutional neural network，CNN)作为融合工具对两种模态的数据进行融合，采用语义分割的方式实现对道路的检测。然而，如何将两种传感器数据更好地融合仍是本研究领域亟待解决的问题。针对上述问题，本文提出了像素级、特征级和决策级多种融合方案，尤其是在特征级融合中设计了四种交叉融合方案，对各种方案进行对比研究，得到最佳的融合方案。在网络构架上，采用编码解码结构的语义分割卷积神经网络作为基础网络，将点云深度图以法线图的方式来表示，法线图特征与 RGB 图像特征在不同的层级进行交叉融合。此方法可以更好地学习到激光雷达点云信息与相机图像信息的关联性，交叉补充点云和图像信息以及减少特征信息的丢失。

本文主要贡献如下：1) 提出了基于 CNN 的点云与图像数据融合的像素级、特征级和决策级多种融合方案，实现对交通场景中道路的检测。尤其是在特征级融合中设计了四种交叉融合方案，对各种方案进行对比研究，得到最佳的融合方案。2) 使用 KITTI 数据集进行实验评估，并对多种融合方式的实验结果进行对比分析。实验结果表明，本文提出的最佳融合方法(交叉融合方案 E)可以显著提高道路的分割效果。

2 相关工作

传统的路面检测方法是依据场景中的几何性质将路面与直立目标加以区分，以实现路面检测的目的。近年来，CNN 强大的特征提取能力与表征能力使其成为路面分割的主流方式。深度学习 CNN 的道路分割方法又分为基于图像的语义分割方法和基于激光雷达—图像融合的语义分割方法。

1) 基于图像的语义分割方法

基于图像的语义分割是将道路检测看做一个语义分割任务。语义分割网络多采用编码器—解码器结构。编码器提取有效特征，解码器对特征进行复原，再通过全连接层综合所有特征及优化函数实现对道路的分割(分类)。U-Net[1]是编码器—解码器结构中常见的一种分割模型，现如今已经有许多基于 U-Net[1]结构而设计的新型卷积神经网络。U-Net++[2]针对 U-Net 中解码器的连接方式作出改进，增加了类似于 DenseNet[3]中的密集连接机制，对精度的提升有所贡献。理论上增加网络深度，可以进行更加复杂的特征提取，分割性能也会变得更好。但是网络的加深往往会带来退化问题，并且会出现过拟合现象。Res-UNet[4]受到 ResNet[5]原理的启发，通过短路机制加入残差单元，极大地消除了深层神经网络所带来的退化过拟合问题。Chen 等人[6]使用 DeepLabv3 作为编码器模块和一个简单有效的解码器模块细化分割结果，并将深度可分卷积应用于 ASPP 模块和解码器模块中，得到一个更快、更强的编解码器网络进行语义分割。SegNet[7]在解码器中使用编码器中进行最大池化的像素索引来进行反池化，从而省去学习上采样的需要，节省了计算时间，并用 Softmax 分类对每个像素输出一个类别的概率。

OFA Net[8]使用一种"1-N 替代"的策略进行训练，探讨了检测任务和语义分割之间的相互增强效果，极大地解决了数据集过少带来的一系列问题。MultiNet[9]提出了一种将分类、检测和语义分割联合起来的方法，三个任务的编码器阶段是共享的，利用深层的 CNN 产生能够在所有任务中使用的丰富共享特征。这些特征再被三个以任务为导向的解码器使用，解码器实时产生结果，共享计算降低了执行所有任务所耗时长，性能方面还有待提高。RBNet[10]同时进行道路检测和道路边界检测，研究道路之间的语境关系结构及其边界排列，然后通过贝叶斯模型同时估计图像上像素的概率属于道路和道路的边界，消除了边界以外的潜在误判。Multi-task CNN[11]提出了紧凑的多任务 CNN 架构，在嵌入式系统的计算资源约束下，有效检测和估计物体以及基本汽车环境模型的可干燥地形，并引入

了基于检测解码器和分析几何的简单扩展的 3D 边界框估计方案。

2) 基于激光雷达与图像融合方法

多传感器融合是对多源的信息数据利用一定的方法、准则进行处理，以实现所需要的估计决策。在自动驾驶领域，大多采用激光雷达传感器、相机等数据信息进行融合，以感知周围环境。Schlosser 等人[12]将激光雷达的 3D 点云数据预处理成了 HHA(水平视差、地面高度、角度)数据，与 RGB 图像一同输入，在 CNN 网络的不同特定层采用像素相加的融合方式，证明了在网络的中间层融合会得到最强的效果。LidCamNet[13]采用了特征融合的方式，采用可训练的线性叠加，将实验结果与前期、后期融合的结果进行对比。可训练的参数在数据融合时有一定的灵活性，较为良好的分割结果进一步验证了该思路在语义分割领域的可行性。Chen 等人[14]采用了渐进式激光雷达自适应级联融合结构，用激光雷达数据去辅助图像数据进行道路分割，使用可训练参数的同时将激光雷达特征与 RGB 特征进行自适应处理，在强光或者强阴影条件下达到更好的融合效果。Neven 等人[15]提出了以 RGB 图像为指导，利用其目标信息去纠正点云信息的预测的融合方式，降低了点云的误判概率。Wang 等人[16]利用激光雷达传感器和立体双目相机，用两种增强技术的立体匹配网络来估计深度，而不是直接融合，一定程度上提高了检测精度。

Zhang 等人[17]采用了基于深度学习的 RGB-D 深度图补全的方法，输入 RGB-D 图去预测 RGB 图中所有平面的表面法线和物体边缘遮挡，用深度图作为正则化，求解全局线性优化问题，最终得到补全的深度图，为自动驾驶环境感知提供了更好的数据信息。文献[18]为了能够同时提取 RGB 图像和深度图特征，将两者融合，并将融合后的图像变成 HHG 图像。文献[19]提出了一种基于双传感器信息融合的三维物体姿态估计—视锥体 PointNet 目标定位算法，进一步证明了多数据融合的可行性。SNE-RoadSeg[20]采用编码器—解码器结构，在编码器部分对双传感器数据输入进行特征融合，实现精准的自由空间检测。并提出了将点云深度图转换为法线特征图的方法，将表面法线估计问题转化为最小二乘平面拟合估计问题，对三维曲面上的每个点估计法线，难点在于道路和人行道上三维点具有非常相似的表面法线。

3 本文方法

网络基础结构如图 1 所示，由采用残差网络(ResNet[5])的编码器、采用密集连接的跳跃连接的解码器(如图 2 所示)、表面法线估计器(surface normal estimator，SNE，如图 3 所示)组成。输入图像为 RGB-D 图，激光雷达深度图经过表面法线估计器处理为法线图；两路输入信号经两路编码器提取特征，解码器还原特征，最后使用 sigmoid 激活函数生成道路分割结果。

法线的作用在于丰富特征信息并矫正光源产生的阴影和其他视觉效果，深度图只有单层的少量深度特征信息，处理得到的法线图根据每个点所处平面不同、表面法线方向也不同的原理，更好地区分路面与非路面。RGB 编码器和表面法线编码器的主干为 ResNet[5]，它们的结构彼此相同。如图 1 所示，输入数据先经过一个初始块(由卷积核 7×7、步长 2 的卷积层，批量规范化层(BN)和 ReLU 激活层组成)，然后依次使用一个最大池化层和四个 Res-layer 来逐渐降低分辨率并增加

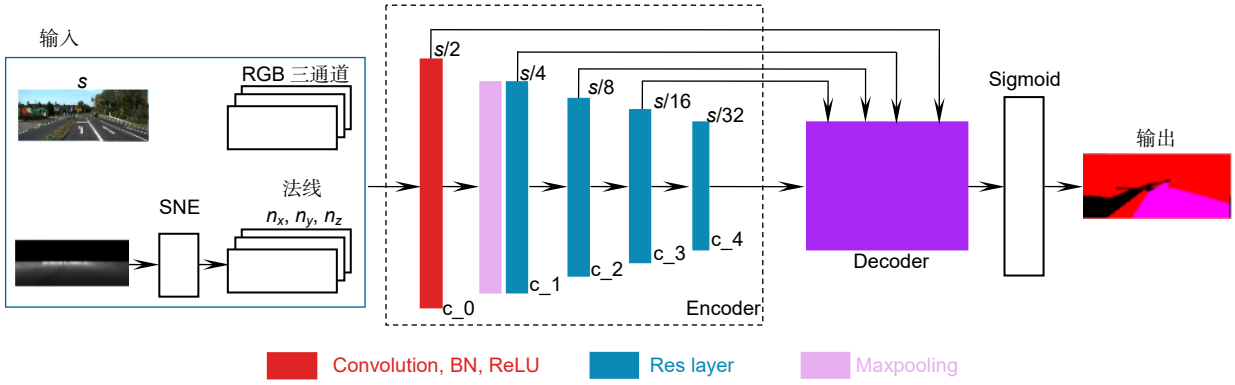

图 1 网络基础结构图

Fig. 1 Network infrastructure diagram

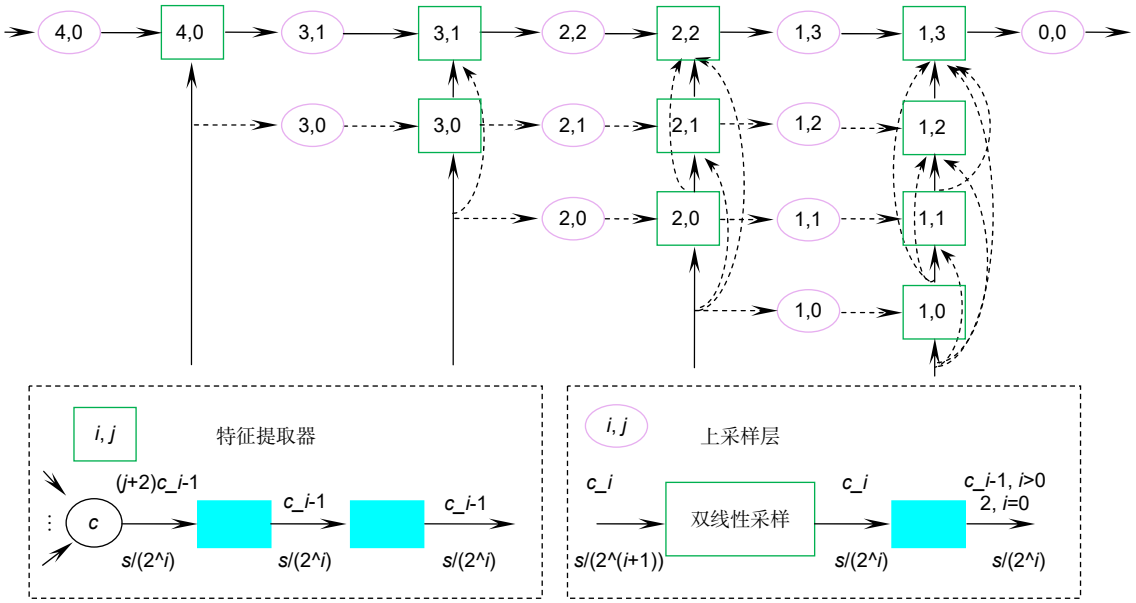

图 2 解码器结构图

Fig. 2 Decoder structure diagram

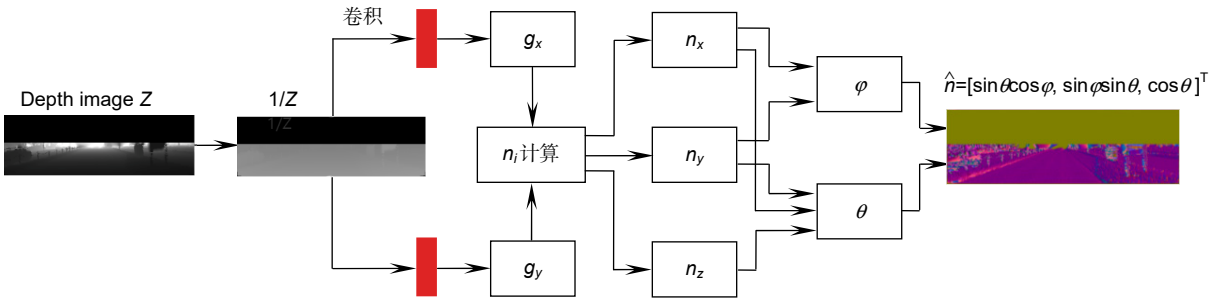

图 3 表面法线估计器

Fig. 3 Surface normal estimator

特征图通道的数量，四个 Res-layer 分别由 n 个 bottleneck block 构成，bottleneck block 由卷积核分别为 1×1、3×3、1×1 的三个卷积层组成。ResNet[5]有多种体系结构，本文采用 ResNet-152，特征映射通道的数量 c_0~c_4 分别为 64、256、512、1024、2048，四个 Res-layer 的 bottleneck block 数目 n 分别为 3、8、36、3，s 代表图像输入的分辨率。

解码器(图 1 中 decoder 方形块)如图 2 所示，由两种不同类型的模块组成——特征提取器和上采样层，对编码后的特征图进行解码，以恢复特征映射的分辨率。在解码器的每一层分别引入相应的编码阶段产生的特征层，它们紧密连接以实现灵活的特征融合。曲面箭头表示跳跃连接，自下而上的直箭头表示引入编码阶段产生的特征图。利用特征提取器提取特征，并确保特征图分辨率不变，利用上采样层提高分辨率和减少特征图的通道数。特征提取器和上采样层共有的矩形框由卷积核为 3×3、步长 1、padding1 的卷积层、BN 层和 ReLU 层组成。

3.1 表面法线估计器

曲面法线是几何表面的重要属性，是指经过曲面上一点并与该点的切平面垂直的直线(即向量)。曲面法线在三维建模中应用较为广泛，可以矫正光源产生的阴影和其他视觉效果。将深度图处理成法线图，可以更好地区分不同平面不同高度的物体。

曲面法线的计算，可以通过对逆深度图像或视差图像执行三个滤波操作，即两个图像梯度滤波器(分别

在水平和垂直方向)和一个平均/中值滤波器。表面法线估计器(surface normal estimator，SNE)如图 3 所示，由 3F2N[21]的方法发展而来，文献[20]中多次实验证明采用这种深度数据处理方式可以得到更好的分割效果。而对表面法线的估计，可以转化为最小二乘平面拟合估计问题，对三维曲面上的每个点估计在该位置与表面相切的平面的法线。

首先将深度图上的每个点 $p=[u,v]^{\mathrm{T}}$ 通过坐标转换式(1)与空间上的点 $\boldsymbol{P}=[X,Y,Z]^{\mathrm{T}}$ 进行连接，进而拟合出一个局部平面(如式(2))。通过将反深度图像 $1/Z$ 分别与水平图像梯度过滤器、垂直图像梯度过滤器卷积，得 n_x，n_y，并代入平面公式，得式(3)：

$$\boldsymbol{K}\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}f_x&0&u_0\\0&f_y&v_0\\0&0&1\end{bmatrix}\begin{bmatrix}X\\Y\\Z\end{bmatrix}, \tag{1}$$

$$n_xX+n_yY+n_zZ+d=0, \tag{2}$$

$$n_x=-\mathrm{d}f_xg_x,\quad n_y=-\mathrm{d}f_yg_y,$$

$$n_z=\mathrm{d}\frac{f_x\Delta X_ig_x+f_y\Delta Y_ig_y}{\Delta Z_i}\quad(i=1,2,3,.....,8), \tag{3}$$

其中：$\boldsymbol{K}$ 为相机内参矩阵，$\boldsymbol{p}_0=[u_0,v_0]^{\mathrm{T}}$ 为图像中心，f_x、f_y 是以像素为单位的相机焦距。$\boldsymbol{n}=[n_x,n_y,n_z]^{\mathrm{T}}$ 为曲面法线，d 为常量，g_x、g_y 是对反深度图像 $1/Z$ 做 x、y 方向的微分。

估计点 $\boldsymbol{P}$ 处的表面法线需要周围点的信息(也称为 k 邻域)，$\boldsymbol{N}_P=[\boldsymbol{Q}_1,...,\boldsymbol{Q}_k]^{\mathrm{T}}$ 为 $\boldsymbol{P}$ 的 k 个近邻点，给一个任意的 $\boldsymbol{Q}_i\in\boldsymbol{N}_P$，$\boldsymbol{Q}_i-\boldsymbol{P}=[\Delta X_i,\Delta Y_i,\Delta Z_i]^{\mathrm{T}}$。$\boldsymbol{N}_P$ 关于 $\boldsymbol{P}$ 的 k-连通域可以生成 k(1，2，3，……，8)个标准化曲面法线 $\bar{n}_1,...,\bar{n}_k$，其中：

$$\bar{\boldsymbol{n}}_i=\frac{\boldsymbol{n}_i}{\|\boldsymbol{n}_i\|_2}=[\bar{n}_{xi},\bar{n}_{yi},\bar{n}_{zi}]^{\mathrm{T}}。$$

由于任何标准化曲面法线都可以投影在圆心(0，0，0)、半径为 1 的球体上，因此我们认为 $\boldsymbol{P}$ 的最佳曲面法线 $\hat{\boldsymbol{n}}$ 也投影在同一球体的某个位置，所以将同一个点的 k 个标准化曲面法线归一化，并将 $\hat{\boldsymbol{n}}$ 用球坐标系表示(式(4))，得到最优曲面法线。其中 $\theta\in[0,\pi]$ 表示倾角，$\varphi\in[0,2\pi)$ 表示方位角。

$$\hat{\boldsymbol{n}}=[\sin\theta\cos\varphi,\ \sin\theta\sin\varphi,\ \cos\theta]^{\mathrm{T}}, \tag{4}$$

$$\varphi=\arctan\left(\frac{f_yg_y}{f_xg_x}\right), \tag{5}$$

$$\theta=\arctan\left(\frac{\sum_{i=1}^{k}\bar{n}_{xi}\cos\varphi+\sum_{i=1}^{k}\bar{n}_{yi}\sin\varphi}{\sum_{i=1}^{k}\bar{n}_{zi}}\right)。 \tag{6}$$

假设任意一对归一化曲面法线之间的角度小于 $\pi/2$，因此 $\hat{\boldsymbol{n}}$ 可以通过最小化 $E=-\sum_{i=1}^{k}\hat{\boldsymbol{n}}\cdot\bar{\boldsymbol{n}}_i$，$\frac{\partial E}{\partial\theta}=0$ 得出 θ 值。

3.2 融合方式

在多传感器信息融合中，按其在融合系统中信息处理的抽象程度可分为三个层次：像素级融合、特征级融合和决策级融合。针对采用怎样的方式以及在什么阶段融合能得到更优效果的问题，本文设计并实验了多种融合策略(如图 4 所示)。

像素级融合属于底层数据融合方法(如融合 A)，将两路传感器的原始观测信息在数据预处理结束后直接进行通道融合，以六通道观测数据进入编码器—解码器结构，提取特征并进行判断识别。

特征级融合属于中间层次级融合(如融合 B、C、D、E)，先从两路传感器的原始观测信息中提取代表性特征，选择合适的特征进行交叉融合：

融合 B：将原始数据分别进入编码器结构中提取特征，然后将编码后的两路特征数据进行融合，再将融合后的数据送进解码器部分得出分割结果；

融合 C：将原始数据分别进入编解码网络结构，在编码器五个阶段采用交叉方法 1(图 4 中的菱形框)，如图 5 中的(a)所示，对 RGB 特征图进行信息补充；

融合 D：将原始数据分别进入编解码网络结构，在编码器五个阶段采用交叉方法 2(图 4 中的椭圆框)，如图 5 中的 5(b)所示，对 RGB 特征图进行信息补充。

融合 E：将原始数据分别进入编解码网络结构，在编码器五个阶段采用交叉方法 3(图 4 中的圆角矩形框)，如图 5 中的 5(c)所示。该融合方法是方案 C，D 的综合，单从一路数据讲就是将法线特征与 RGB 特征通道拼接，通过训练学习到 α、β 两个参数，根据这两个参数得到转换后的法线数据特征图，与 RGB 特征图叠加得到转换后的 RGB 特征图，同理得到转换后的法线特征图。然后将转换后的法线特征图与可训练参数 b_i 再次相乘，最后与转换后的 RGB 特征图再次叠加得到新 RGB 特征图。另一路同理可得融合后的新法线特征图。然后将两路融合数据均送入解码器结构还原，最后在 Sigmoid 层再次进行融合。

决策级融合属于高层次级融合(如融合 F)，输出是一个联合决策结果，理论上这种联合决策比基于单传感器的决策要更优。将两路传感器数据信息分别进入编解码网络，在解码后拼接，然后在 sigmoid 层进行融合，得出分割结果。

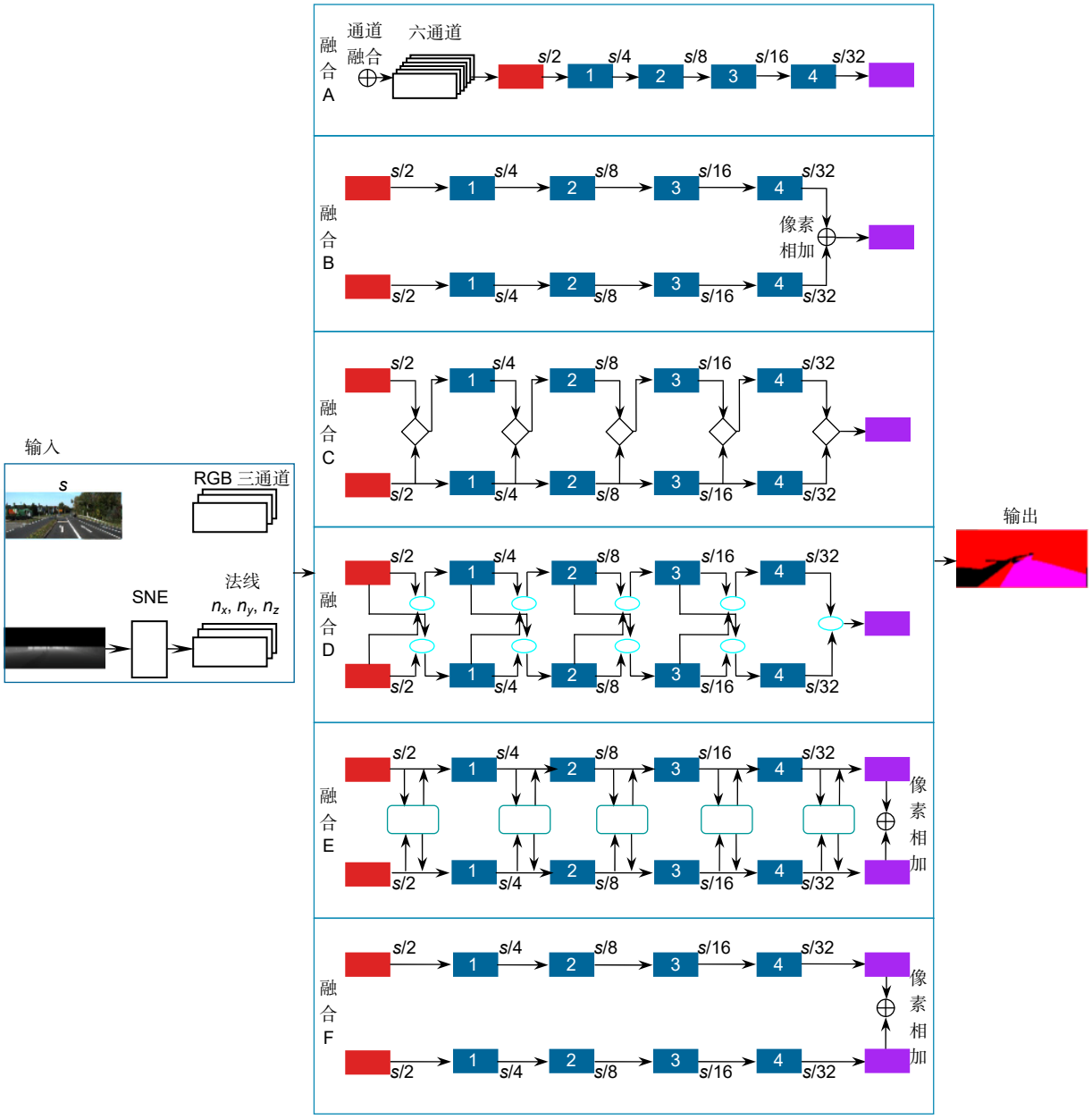

图 4 采用不同融合策略的网络结构

Fig. 4 Network structure with different fusion strategies

4 实验结果

实验数据来自 KITTI 的道路数据集，包含三个子集：训练集(289 张图像)，验证集(32 张图像)，测试集(290 张图像)。

验证集是训练集中留出的用于模型验证的图像集，KITTI 提供真值，用于调整模型的超参数和评估模型的能力。

测试集仅用于评估最终模型的性能，KITTI 不提供真值，需要研究者提供检测结果，由 KITTI 将检测结果与真值进行比较，这样可以保证不同方法比较的公正性。KITTI 图像序列包含三种场景：UU(城市无标记)、UM(城市标记)、UMM(城市多条标记车道)。实验结果采用 KITTI 的评价方法，性能评估有五个常用的指标：准确率(Accuracy, A_{cc})，精确度(Precision, P)，召回率(Recall, R)，F1 值(F1-score, F_1)，PR 曲线(AP)：

$$A_{cc}=\frac{n_{tp}+n_{tn}}{n_{tp}+n_{tn}+n_{fp}+n_{fn}}\ , \tag{7}$$

$$P=\frac{n_{tp}}{n_{tp}+n_{tn}}\ , \tag{8}$$

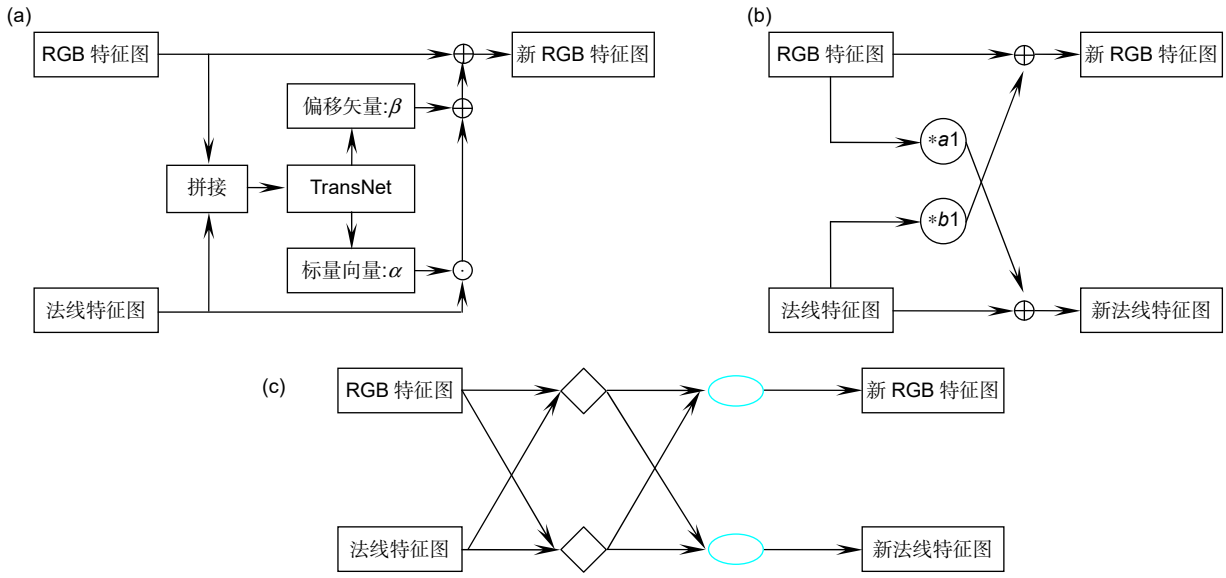

图 5 交叉融合方法。(a) 交叉方法 1; (b) 交叉方法 2; (c) 交叉方法 3

Fig. 5 Cross fusion method. (a) Cross method 1; (b) Cross method 2; (3) Cross method 3

$$R=\frac{n_{\rm tp}}{n_{\rm tp}+n_{\rm fn}}\quad , \tag{9}$$

$$F_1=2\cdot\frac{P\cdot R}{P+R}=\frac{2n_{\rm tp}^2}{2n_{\rm tp}^2+n_{\rm tp}(n_{\rm fp}+n_{\rm fn})}\quad , \tag{10}$$

$$IoU=\frac{n_{\rm tp}}{n_{\rm tp}+n_{\rm fp}+n_{\rm fn}}\quad , \tag{11}$$

$$AP=\int_0^1 P(R){\rm d}R\quad , \tag{12}$$

其中：$n_{\rm tp}$、$n_{\rm tn}$、$n_{\rm fp}$、$n_{\rm fn}$为所有图像中真正像素数(true positive pixel numbers)、真负像素数(true negative pixel numbers)、假正像素数(false positive pixel numbers)、假负像素数(false negative pixel numbers)。AP 即平均精度，是 PR 曲线(以 Recall(R)为横轴，Precision(P)为纵轴)下的面积，$P(R)$为不同召回率所对应的准确度。

此外，采用随机梯度下降动量(stochastic gradient descent with momentum，SGDM)优化器最小化损失函数，初始学习率设置为 0.1。在验证子集上采用了早期停止机制，以避免过度拟合，然后使用测试子集对性能进行量化。

实验主要分两个部分：第一，在同一基础网络结构上比较不同的融合方式的分割结果，确定最佳融合方法。第二，比较本文方法与其他道路分割方法的分割效果，验证本文提出的方法对道路分割性能的提升。

4.1 各种融合方案的实验结果比较

各种融合方案的比较是在验证集图像上进行的，由我们自己与真值进行比较得到各项指标。网络的输入数据均为相机采集的 RGB 图像和激光雷达得到的深度图像，在数据预处理中实现了对深度数据的表面法线估计，采用不同的融合方式，对特征信息进行补充，利用编码器—解码器结构提取特征并进行道路分割。

表 1 给出了采用不同融合方式在验证集上得到的实验结果性能指标和 loss 值。对比像素级融合(融合

表 1 不同融合方式之间的性能比较

Table 1 Performance comparison between different fusion methods

		Loss	Accuracy	Precision	Recall	F1-score	IoU
像素级融合	融合 A	0.049	0.984	**0.959**	0.948	0.954	0.911
	融合 B	0.065	0.982	0.941	0.951	0.946	0.897
特征级融合	融合 C	0.050	0.985	0.943	**0.969**	**0.956**	**0.915**
	融合 D	0.047	0.983	**0.951**	0.946	0.948	0.902
	融合 E(ours)	**0.022**	**0.994**	**0.979**	**0.988**	**0.984**	**0.968**
决策级融合	融合 F	0.058	0.982	0.934	0.962	0.948	0.901

A)和决策级融合(融合 F)，融合 A 的 accuracy、precision、F1-score 及 IoU 分别比融合 F 高 0.2%、2.5%、0.6%、1%，仅 recall 低了 1.4%。在所有的特征级融合方法中，融合 E 各方面性能指标均有非常不错的表现，Loss 只有 0.022，Accuracy 提升至 99.4%，Precision 提升至 97.9%，Recall 提升了 1.9%，F1-score 提升了 2.8%，IoU 值提升至 96.8%。

现有的 2D 道路分割方法多采用激光雷达的数据信息去补充 RGB 图像信息，交叉方法 3 可以对两路特征信息都进行补充，将两路传感器数据置于同等重要的地位。原始特征的组合形式特征增加了特征维数，提高目标分割的准确率，解决了像素级融合易受环境噪声干扰的不稳定以及算法实现的费时。决策级融合有很好的纠错性，可以消除单个传感器造成的误差，同时具有很好的分割速度。将两者结合，提高分割的准确率，同时有很好的纠错性。

图 6 为同一张道路图不同融合方式的分割结果示例。通过多组图片对比，可以看出本文提出的融合 E 分割结果与真值图最为接近，道路轮廓分割较为完整，并没有过多的误检区域。对于处于同一水平面的人行

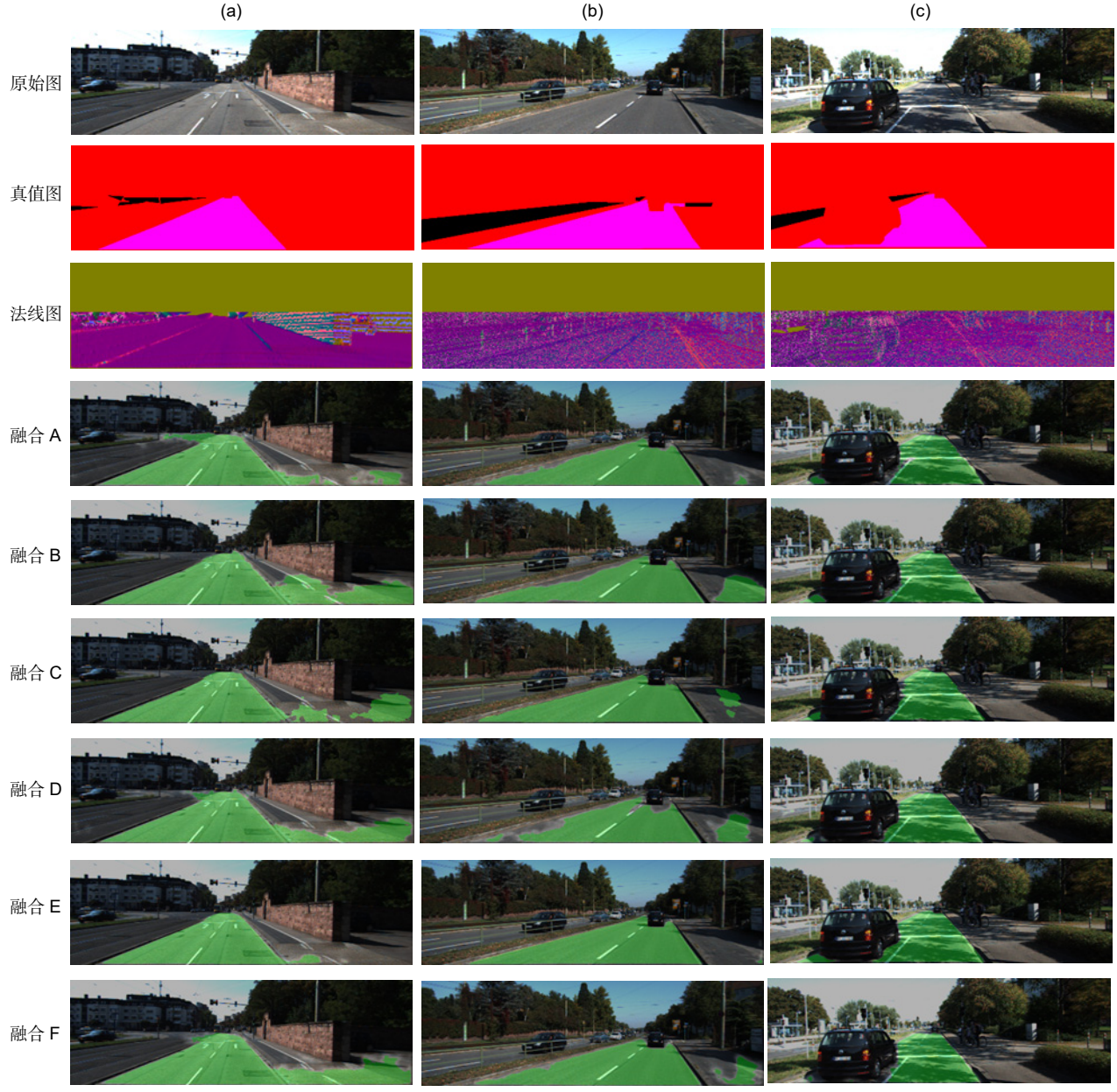

图 6 不同融合方式实验结果示例

Fig. 6 Examples of experimental results of different fusion methods

道、远处的路口区域以及车辆周围的区域，融合 E 对于非道路区域的剔除最干净。

4.2 与其他方法的比较

4.2.1 定性分割结果比较

图 7 给出了对于 KITTI 数据集中几个典型场景的测试结果，将本文提出的最佳融合方法(融合 E)与 OFA Net[8]、MultiNet[9]、RBNet[10]、multi-task CNN[11]、SNE-RoadSeg[20]进行比较。其中第一列为 OFA Net[8]的分割结果图，第二列为 RBNet[10]的分割结果图，第三列为 multi-task CNN[11]的分割结果，第四列为 SNE-RoadSeg[20]的分割结果图，第五列为 LidCamNet[13]的分割结果，第六列为融合方案 E 的方法的分割结果图。图 7(a)、7(b)为 UM 场景，图 7(c)、7(d)为 UMM 场景，图 7(e)、7(f)为 UU 场景，绿色区域为正确的驾驶区域(真阳性)，蓝色区域对应于缺失驾驶区域(假阳性，即错检区域)，红色区域表示假驾驶区域(假阴性，即误检区域)。

对比 UM 场景，对于图 7(a)，OFA Net[8]检测出绿色区域更为完整，红色误检区域很少，但是道路边缘有一圈蓝色错检区域；SNE-RoadSeg[20]蓝色错检区域最少，有少量红色误检区域；融合 E 在阴影处有少量蓝色错检区域，在接近车辆位置有少量红色误检区域，绿色区域较为完整。对于图 7(b)，虽然融合 E 对于车辆下方的人行区域产生了误判，但是绿色区域是最为完整，与右边车辆交界处处理得也很好，其他方法都有少量红色或者蓝色区域。对比 UMM 场景，对于图 7(c)，各方法检测结果都较为理想，误检与错检区域都非常少。而对于图 7(d)，融合 E、OFA Net[8]、RBNet[10]、LidCamNet[13]检测结果最好，对于铁轨区域基本完全剔除。

对比 UU 场景，对于图 7(e)，可以看出融合 E 对于车辆与道路交界位置处理非常好，绿色道路区域绕着车辆的边缘，基本没有红色误检区域；其他方法或多或少存在一些误检区域或者错检区域。对于图 7(f)是同样的，右边部分检测较为完整，虽然左边有少量人行区域的错检。而 multi-task CNN[11]的每次检测结果虽然也比较完整，但是蓝色错检区域太多。综合考虑，融合 E 对于道路与车辆交界处处理非常好。

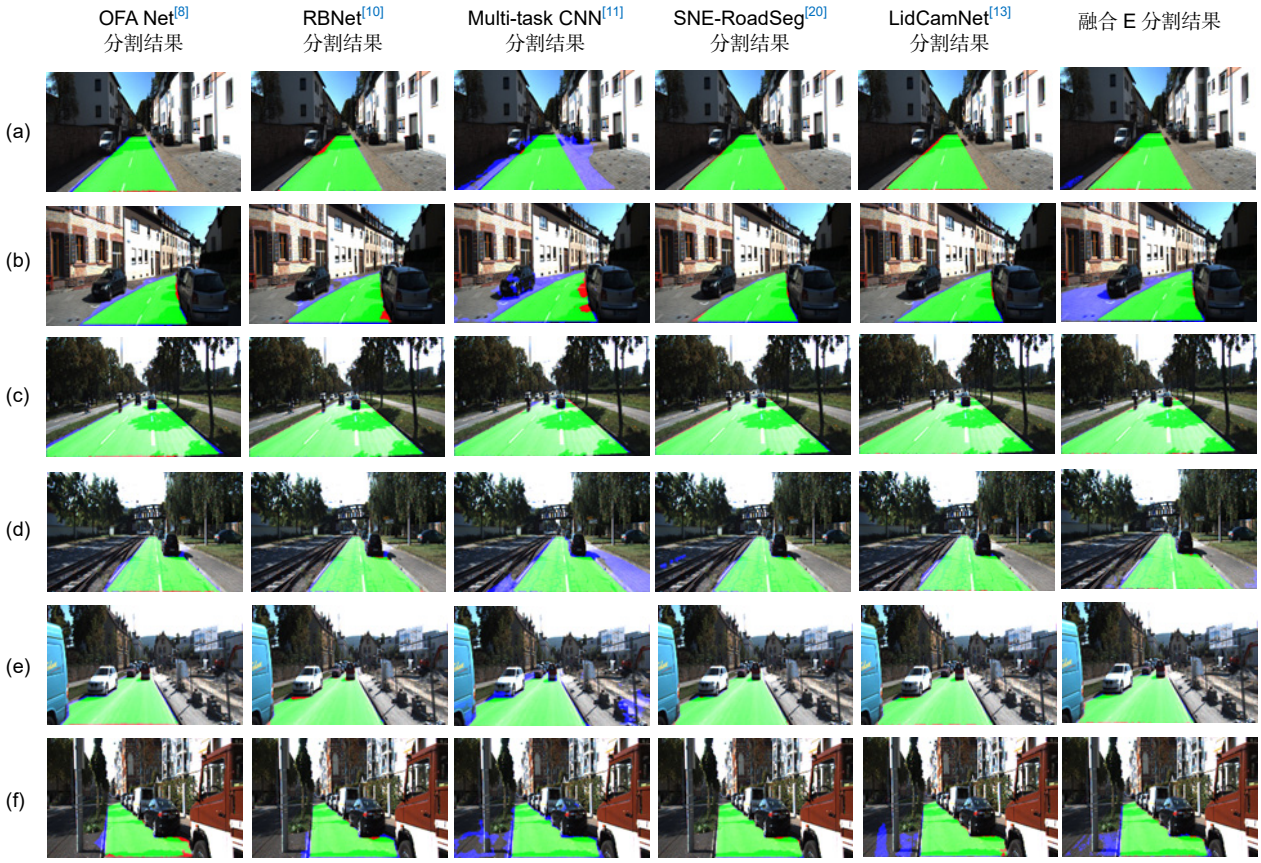

图 7 KITTI 数据集实验结果示例

Fig. 7 Example of KITTI dataset experimental results

融合方案 E 采用可训练参数交叉融合，对图像和法线数据进行特征级融合，综合利用图像数据密集纹理信息和法线数据的方向信息，对两路传感器分割信息进行融合，有效降低了道路分割的误检率。

4.2.2 定量比较

与其他方法的定量比较是在测试集图像上进行的，将我们对测试集图像的分割结果提交 KITTI，由 KITTI 与真值进行比较得到各项指标。将本文提出的最佳融合方法(融合 E)与 KITTI road 基准上发布的 OFA Net[8]、MultiNet[9]、RBNet[10]、multi-task CNN[11]、SNE-RoadSeg[20]、LidCamNet[13]在不同的场景下进行比较，输入数据均为深度数据、RGB 图像数据、融合数据。表 2 中给出了几种方法在测试集上的定量比较。其中，OFA Net[8]、MultiNet[9]、RBNet[10]、multi-task CNN[11] 属于单纯基于图像的分割方法，SNE-RoadSeg[20]、LidCamNet[13]和我们的融合 E 方法属于点云与图像融合的方法。

Precision 表示模型检测出的目标有多大比例是真正的目标物体，Recall 代表所有真实的目标有多大比例被模型检测出。由表中数据可看出，基于图像分割的 OFA Net[8]、multi-task CNN[11]在 recall 方面很高，UMM 场景下可达百分之九十八点几，而 precision 方面却不尽人意，说明基于图像的分割方法检测正确的道路像素数很多，但出现了很多误判情况；而基于点云—图像融合的分割方法在 MaxF(max F1-score)、AP(average precision 平均精度)、Precision 等方面均有不错的表现，Recall 方面略有逊色，说明多数据融合模型检测出的道路是真实道路的比例更高，存在少量漏检情况。结果对比，证明了多数据融合对于道路的误判有显著降低。

在基于点云—图像融合的分割方法中，对比使用特征融合的 LidCamNet[13]，我们的融合 E(交叉方法 3)UM 和 UU 场景下各方面性能均有所提升，而在 UMM 场景下 AP 提升了 0.28%，Recall 提升了 0.22%，Precision 降低了 0.95%，MaxF 降低了 0.37%；对比 SNE-RoadSeg[20]，我们的融合 E 方法在各场景的 AP 值

表 2 KITTI 道路基准测试结果

Table 2 The KITTI road benchmark results

	方法类型		MaxF/%	AP/%	Precision/%	Recall/%
UM	基于图像的语义分割方法	OFA Net[8]	92.08	83.73	87.87	**96.72**
		MultiNet[9]	93.99	93.24	94.51	93.48
		RBNet[10]	94.77	91.42	95.16	94.37
		Multi-task CNN[11]	85.95	81.28	77.40	96.64
	基于点云—图像融合的语义分割方法	SNE-RoadSeg[20]	**96.42**	93.67	**96.59**	**96.26**
		LidCamNet[13]	95.62	93.54	95.77	95.48
		融合 E(ours)	**95.72**	**95.12**	**95.87**	**95.59**
UMM	基于图像的语义分割方法	OFA Net[8]	95.43	89.10	92.78	**98.24**
		MultiNet[9]	96.15	95.36	95.79	96.51
		RBNet[10]	96.06	93.49	95.80	96.31
		Multi-task CNN[11]	91.15	87.45	85.08	98.15
	基于点云—图像融合的语义分割方法	SNE-RoadSeg[20]	**97.47**	95.63	**97.32**	**97.61**
		LidCamNet[13]	97.08	95.51	97.28	96.88
		融合 E(ours)	**96.71**	**95.79**	**96.33**	**97.10**
UU	基于图像的语义分割方法	OFA Net[8]	92.62	83.12	88.97	**96.58**
		MultiNet[9]	93.69	92.55	94.24	93.14
		RBNet[10]	93.21	89.18	92.81	93.60
		Multi-task CNN[11]	80.45	75.87	68.63	97.19
	基于点云—图像融合的语义分割方法	SNE-RoadSeg[20]	**96.03**	93.03	**96.22**	95.83
		LidCamNet[13]	94.54	92.74	94.64	94.45
		融合 E(ours)	**95.38**	**93.23**	**94.95**	**95.83**

均为最高，在 UU 场景下 recall 方面不相上下，其他方面均有不足。Precision 反映了被模型判定为道路的正例中真实道路的比重，体现了检测的准确度。融合 E 的 precision 低于 SNE-RoadSeg[20]，说明被判断为道路的像素中有不少误判的情况。Recall 反映了被正确判断为道路的正例占总的真实道路的比重，体现了检测的完整性。两个方法均为 95.83%，说明被正确判断为道路的像素数基本一致。对于道路检测任务而言，Precision 和 recall 往往是此消彼长的，AP 是两者的结合，AP 越高代表检测失误越少。Precision 的降低，说明我们的融合 E 方法出现了道路误检的情况。从图 7(a)、7(b)可以看出，在 UM(城市标记)场景下，高度与道路一致的非道路区域出现车辆的情况，检测结果出现了严重偏差。从图 7(c)、7(d)可以看出，在 UMM(城市多条标记道路)场景下，路面情况较为复杂时，检测结果较为良好。而且融合 E 在 AP 方面有所提高，说明交叉方法 3 对于模型性能有所改善，但对于个别道路与人行道高度一致且有混淆因素(车辆)的情况仍有不足。

5 总 结

本文研究基于点云与图像数据融合的道路分割方法，设计了像素级、特征级和决策级多种融合方案，尤其是在特征级融合中设计了四种交叉融合方案。采用 KITTI 数据集进行多种融合方式的实验验证，融合方案 E 能够更好地获取图像和法线的特征信息，具有最佳的道路分割效果。对比其他多种道路检测方法，本文提出的最佳融合方法表现出平均检测精度上的优势，具有较好的整体性能。

参考文献

[1] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C]//*Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[2] Zhou Z W, Siddiquee M R, Tajbakhsh N, *et al*. UNet++: a nested U-Net architecture for medical image segmentation[C]//*Proceedings of the 4th International Workshop on Deep Learning in Medical Image Analysis*, 2018.

[3] Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[4] Xiao X, Lian S, Luo Z M, *et al*. Weighted Res-UNet for high-quality retina vessel segmentation[C]//*Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education*, 2018.

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

[6] Chen L C, Zhu Y K, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018.

[7] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. *IEEE Trans Pattern Anal Mach Intell*, 2017, **39**(12): 2481–2495.

[8] Zhang S C, Zhang Z, Sun L B, *et al*. One for all: a mutual enhancement method for object detection and semantic segmentation[J]. *Appl Sci*, 2020, **10**(1): 13.

[9] Teichmann M, Weber M, Zöllner M, *et al*. MultiNet: real-time joint semantic reasoning for autonomous driving[C]//*Proceedings of 2018 IEEE Intelligent Vehicles Symposium*, 2018.

[10] Chen Z, Chen Z J. RBNet: a deep neural network for unified road and road boundary detection[C]//*Proceedings of the 24th International Conference on Neural Information Processing*, 2017.

[11] Oeljeklaus M, Hoffmann F, Bertram T. A fast multi-task CNN for spatial understanding of traffic scenes[C]//*Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems*, 2018.

[12] Schlosser J, Chow C K, Kira Z. Fusing LIDAR and images for pedestrian detection using convolutional neural networks[C]//*Proceedings of 2016 IEEE International Conference on Robotics and Automation*, 2016.

[13] Caltagirone L, Bellone M, Svensson L, *et al*. LIDAR–camera fusion for road detection using fully convolutional neural networks[J]. *Rob Auton Syst*, 2019, **111**: 125–131.

[14] Chen Z, Zhang J, Tao D C. Progressive LiDAR adaptation for road detection[J]. *IEEE/CAA J Automat Sin*, 2019, **6**(3): 693–702.

[15] van Gansbeke W, Neven D, de Brabandere B, *et al*. Sparse and noisy LiDAR completion with RGB guidance and uncertainty[C]//*Proceedings of the 2019 16th International Conference on Machine Vision Applications*, 2019.

[16] Wang T H, Hu H N, Lin C H, *et al*. 3D LiDAR and stereo fusion using stereo matching network with conditional cost volume normalization[C]//*Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019.

[17] Zhang Y D, Funkhouser T. Deep depth completion of a single RGB-D image[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[18] Deng G H. Object detection and semantic segmentation for RGB-D images with convolutional neural networks[D]. Beijing: Beijing University of Technology, 2017.
邓广晖. 基于卷积神经网络的 RGB-D 图像物体检测和语义分割[D]. 北京: 北京工业大学, 2017.

[19] Cao P. Dual sensor information fusion for target detection and attitude estimation in autonomous driving[D]. Harbin: Harbin Institute of Technology, 2019.
曹培. 面向自动驾驶的双传感器信息融合目标检测及姿态估计[D]. 哈尔滨: 哈尔滨工业大学, 2019.

[20] Fan R, Wang H L, Cai P D, *et al*. SNE-RoadSeg: incorporating surface normal information into semantic segmentation for accurate freespace detection[C]//*Proceedings of the 16th European Conference on Computer Vision*, 2020.

[21] Fan R, Wang H L, Xue B H, *et al*. Three-filters-to-normal: an accurate and ultrafast surface normal estimator[J]. *IEEE Rob Automat Lett*, 2021, **6**(3): 5405–5412.

Point cloud-image data fusion for road segmentation

Zhang Ying, Huang Yingping*, Guo Zhiyang, Zhang Chong

School of Optical-Electronic and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Fusion scheme E detection results(right side)

Overview: Road detection is an important content of environmental identification in the field of automatic driving, and it is an important prerequisite for vehicles to realize automatic driving. Multi-source data fusion based on deep learning has become a hot topic in the field of automatic driving. RGB data can provide dense texture and color information, LiDAR data can provide accurate spatial information, and multi-sensor data fusion can improve the robustness and accuracy of detection. The latest fusion method uses convolutional neural network (CNN) as a fusion tool to fuse the LiDAR data and RGB image data, and semantic segmentation to realize road detection and segmentation. In this paper, different fusion methods of LiDAR point cloud and image data are adopted by encoder-decoder structure to realize road segmentation in traffic scenes. Aiming at the fusion methods of point cloud and image data, this paper proposes a variety of fusion schemes at pixel level, feature level, and decision level. In particular, four kinds of cross-fusion schemes are designed in feature level fusion. Various schemes are compared and studied to give the best fusion scheme. As for the network architecture, we use the encoder with residual network and the decoder with dense connection and jump connection as the basic network. The input image is RGB-D, and the LiDAR depth map is processed into a normal map by a surface normal estimator. The normal map features and RGB image features are fused at different levels of the network. The features are extracted through two input signals generated by two encoders, restored by a decoder, and finally road detection results are obtained by using sigmoid activation function. KITTI data set is used to verify the performances of various fusion schemes. The contrast experiments show that the proposed fusion scheme E can better learn the LiDAR point cloud information, the camera image information, the correlation of cross added point cloud, and image information. Also, it can reduce the loss of characteristic information, and thus has the best road segmentation effect. Through quantitative analysis of the average accuracy (AP) of different road detection methods, the optimal fusion method proposed in this paper shows the advantages of average detection accuracy, and has good overall performance. Through qualitative analysis of the performance of different detection methods in different scenarios, the results show that the fusion scheme E proposed in this paper has good detection results for the boundary area between vehicles and roads, and could effectively reduce the false detection rate of road detection.

Zhang Y, Huang Y P, Guo Z Y, *et al.* Point cloud-image data fusion for road segmentation[J]. *Opto-Electron Eng*, 2021, **48**(12): 210340; DOI: 10.12086/oee.2021.210340

Foundation item: the Shanghai Natural Science Foundation of Shanghai Science and Technology Commission, China (20ZR14379007), and National Natural Science Foundation of China (61374197)

* E-mail: huangyingping@usst.edu.cn