基于迭代式自主学习的三维目标检测

王康如[1,2*]，谭锦钢[1,2]，杜量[3]，陈利利[1]，李嘉茂[1]，张晓林[1]

[1]中国科学院上海微系统与信息技术研究所仿生视觉系统实验室，上海 200050；

[2]中国科学院大学，北京 100049；

[3]复旦大学类脑智能科学与技术研究院计算神经科学与类脑智能教育部重点实验室，上海 200433

摘要 为了提高基于双目视觉的三维目标检测的精度与鲁棒性，提出了一种基于迭代式自主学习的三维目标检测算法。首先，为了给三维目标检测任务提供更精准的目标点云信息，提出了一种基于迭代式自主学习的视差估计算法，通过迭代地增加目标区域的视差监督信号以及引入选择性优化策略，提高了视差估计在目标区域中的准确性。其次，在网络结构中，提出了一种自适应特征融合机制，将不同模态信息的特征进行自适应融合，进而得到准确且稳定的目标检测结果。结果表明，与近年来较流行的基于视觉系统的算法相比，所提出的三维目标检测算法在检测精度上有较大提升。

关键词 机器视觉；三维目标检测；立体视觉；卷积神经网络；自主学习

中图分类号 TP391.41 **文献标志码** A **doi**：10.3788/AOS202040.0915005

3D Object Detection Based on Iterative Self-Training

Wang Kangru[1,2*], Tan Jingang[1,2], Du Liang[3], Chen Lili[1], Li Jiamao[1], Zhang Xiaolin[1]

[1] *Bionic Vision System Laboratory, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai* 200050, *China*;

[2] *University of Chinese Academy of Sciences, Beijing*, 100049, *China*;

[3] *Key Laboratory of Computational Neuroscience and Brain Inspired Intelligence, Ministry of Education, Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai* 200433, *China*

Abstract To improve the precision and robustness of 3D object detection based on stereo vision, a novel 3D object detection algorithm based on iterative self-training is proposed. To acquire the precise object point clouds for 3D object detection task, a disparity estimation algorithm based on iterative self-training is first proposed, which is capable of improving the disparity accuracy of object region by increasing the supervised signal in object region iteratively and introducing a selective optimization strategy. Then a self-adaptive feature fusion mechanism is proposed in network architecture, which adaptively fuses the features from multimodal information to obtain the precise and robust object detection results. Compared with the recent and popular algorithms based on vision system, the proposed 3D object detection algorithm achieves a great improvement in precision.

Key words machine vision; 3D object detection; stereo vision; convolutional neural network; self-training

OCIS codes 150.0155; 110.2970; 100.4996

1 引 言

三维目标检测技术能够获取目标物体在真实世界中的三维信息，是实现机器与环境交互的关键技术，在机器人环境感知、自动驾驶等场景中有着巨大的应用前景，已逐渐成为研究热点。

在室外场景中，目前大多数的三维目标检测算法都是基于激光雷达传感器展开的[1-6]。虽然激光雷达能够采集精准的三维点云信息，但是激光雷达成本较高、无法测量高度反光的物体且只能提供稀疏的点云信息，无法获取稠密的场景图像信息。此外，从安全性上来看，即使再精准的传感器，单一使

收稿日期：2019-11-25；**修回日期**：2020-01-01；**录用日期**：2020-02-10

基金项目：国家自然科学基金（61806189）、上海市自然科学基金（17ZR1436000）、上海市市级科技重大专项（2018SHZDZX01）、上海市“科技创新行动计划”项目（17511105803）

* **E-mail**：wangkangru@mail.sim.ac.cn

用情况下还存在设备损坏、断电等固有安全风险。因此，研究基于其他传感器的三维目标检测算法是必然趋势。相机传感器比激光雷达便宜，且能以高帧率提供稠密的场景信息，因此基于视觉系统的检测算法引起了广泛关注。基于单目视觉的三维目标检测算法相继被提出[7-10]，但单目相机缺少绝对尺度，获取的深度信息并不精准且适用性较差，因此获得的三维目标检测结果准确性较低。双目相机利用采集到的双目图像计算视差，并利用相机内外参数获取整个视野内的绝对三维信息，无需进行尺度恢复。因此，基于双目视觉的三维目标检测算法正逐渐成为研究热点。Chen 等[11]提出了 3DOP 算法，首先利用立体匹配算法计算视差信息，并将其转换成场景点云，随后利用物体的三维先验构建能量函数以生成三维候选框，最后利用卷积神经网络回归三维目标框的坐标以及位姿，得到最终检测结果。Wang 等[12]提出的 Pseudo-LiDAR 算法将视差转换成场景点云，并利用已有的基于雷达的 AVOD 检测网络[1]进行三维物体检测。Königshof 等[13]利用语义信息、视差信息以及几何约束来共同实现三维目标检测。然而，目前基于双目视觉的三维目标检测算法仍有较大的研究空间：首先，准确的视差计算是完成目标检测任务的重要前提，然而大部分检测算法将视差计算与三维目标检测任务分开，并没有考虑两者的相关性；其次，3DOP[11]、Pseudo-LiDAR[12]等大部分三维目标检测算法在利用 RGB 信息与视差或点云信息进行目标检测时，一般采用简单的直接连接或取平均的模式融合不同模态特征，这种固定模式无法在变化的场景下自适应地实现高效合理的特征融合。

基于以上问题，本文提出了一个基于双目视觉的三维目标检测算法。首先，为了向三维目标检测任务提供精准的目标点云信息，提出了一种基于迭代式自主学习的视差估计算法，以提高视差估计在目标区域中的准确性。在视差估计网络训练中，通过迭代式自主学习，逐步增加目标区域中的监督信号，提高训练效果。此外，通过引入选择性优化策略来加强网络对目标区域的特征提取，进一步提升目标区域的视差精度。其次，在目标检测网络中，提出了一个自适应特征融合机制，实现了点云特征与 RGB 特征的有效融合，进而得到更加精准且稳定的三维目标检测结果。基于以上方法，所提三维目标检测算法在检测精度上得到了较大提升。

2　所提算法

所提出的三维目标检测算法流程图如图 1 所示。首先，利用基于迭代式自主学习的视差估计算法来计算场景中的视差信息。随后，利用相机内外参数将视差信息转换为场景点云。最后，在目标检测阶段，利用自适应特征融合模块(SAFFM)对点云信息与 RGB 信息进行自适应融合，实现精准的目标识别与定位。下面对算法细节进行详细介绍。

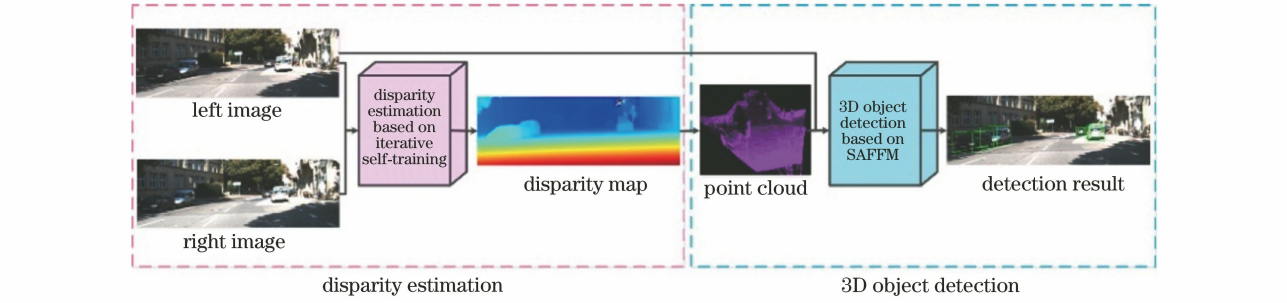

图 1　三维目标检测系统流程图

Fig. 1　Flow chart of 3D object detection system

2.1　基于迭代式自主学习的视差估计算法

人类在观察环境时，比起全局信息，大脑更加关注某些重要的局部区域，这是一种特殊的信号处理机制。当选定了需要关注的重要区域时，大脑会投入更多的资源来进一步获取区域细节信息，进而提高信息处理的准确度。这种信息处理机制是随着任务而变化的，比如在室外驾驶场景下，大脑会更关注车、人等动态目标。借鉴于人脑的这种信息处理机制，我们认为在三维目标检测任务中，目标区域的三维信息相对背景区域更加重要，因此应着重优化目标区域的视差信息。然而，在目前的三维目标检测系统中，大部分的算法在计算视差时对场景全图的关注是一致的，并没有突出关注目标区域。

因此，本文提出了一个基于迭代式自主学习的视差估计算法，该算法更专注于目标区域的视差估计，进而提高三维目标检测的准确性。首先，由于室

外三维目标检测数据集的视差真值大多是由雷达点云映射而成，因此视差真值是稀疏的。然而，在深度学习中，为了保证网络的精度和泛化能力，需要大量的样本点参与训练，因为足够多的训练样本才能表达数据的真实分布。因此，提出了一个基于迭代式自主学习的视差估计网络（IST-Net），其网络结构如图 2 所示，其中 CNN 为卷积神经网络。在网络训练过程中，利用迭代式自主学习这一半监督算法，不断增加目标区域内的视差监督信号，实现网络对目标区域视差的精准估计。此外，在视差估计的目标损失函数中引入了选择性优化策略，以引导网络增强对目标区域的特征提取，进一步提高网络在目标区域中的视差估计精度。下面对具体的算法细节进行详细介绍。

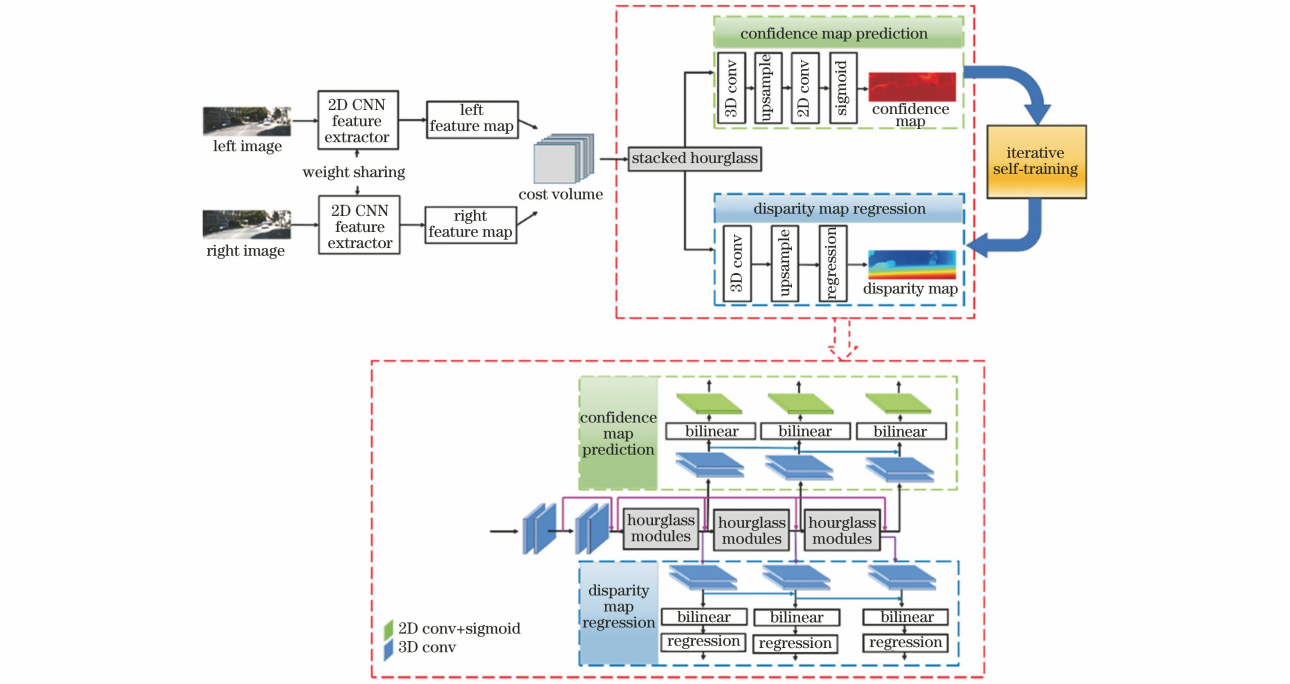

图 2　IST-Net 结构图

Fig. 2　Architectural diagram of IST-Net

2.1.1　IST-Net 网络结构

在 PSMNET[14] 网络框架的基础上，利用 IST-Net 网络进行视差估计，在网络中引入了一个置信图预测模块对没有视差真值的像素点进行预测，预测其估算视差值的置信度，置信度越高，估算的视差值越准确。随后对高置信度的样本点进行筛选，得到目标区域内的伪真值点，并将其加入到原始的视差真值中，作为下一轮网络训练的监督信号。随着不断地训练迭代，越来越多的目标区域样本点作为监督信号被加入到网络训练中。

IST-Net 网络结构如图 2 所示，网络的输入为左右目图像，首先利用 2D 卷积网络对左右目图像进行特征提取，得到代价体（cost volume），随后利用堆叠沙漏结构（stacked hourglass）进一步对代价体进行调整，最后利用置信度预测模块以及视差回归模块，分别预测视差置信图与视差图。其中，网络中的堆叠沙漏结构包含三个沙漏模型，实现由精到粗再到精的特征提取。利用视差图回归模块提取这三个沙漏模型所得到的特征，对应生成三个视差图，由网络深层得到的视差图融合了较浅层的特征，最终的视差图由最后一个输出得到。同样，置信图预测模块也利用这三个沙漏模型所提取的特征，进一步生成视差置信图。在置信图预测模块中，每个生成置信图的分支都包括两层 3D 卷积层、一层 2D 卷积层以及一层 sigmoid 激活层，第一层 3D 卷积层含有 32 个大小为 3×3×3 的卷积核，第二层 3D 卷积层含有 1 个大小为 3×3×3 的卷积核，随后 2D 卷积层里含有 1 个大小为 3×3 的卷积核，最后利用 sigmoid 激活层得到视差置信图。在测试时，置信图预测模块并不参与，其只在网络训练过程为本文提出的迭代式自主学习算法提供视差置信度。

2.1.2　迭代式自主学习算法

稀疏的视差真值与稠密的图像信息具有不对称性，这种不对称性启发我们采用自主学习这一半监督策略来辅助网络的训练过程。以往的自主学习算法一般先以初始真值作为监督信号训练出一个完整

模型，随后利用该模型生成新的标签信息，并将其添加到原有真值中以作为新的监督信号继续对模型进行训练。然而，由于视差估计网络的学习过程时间长且资源占用多，因此每次训练一个完整的模型后再更新监督信号的方法不适用于工程实践。鉴于以上原因，对自主学习策略进行了改进，提出了迭代式的自主学习策略以适应视差估计任务，其算法流程如图 3 所示。

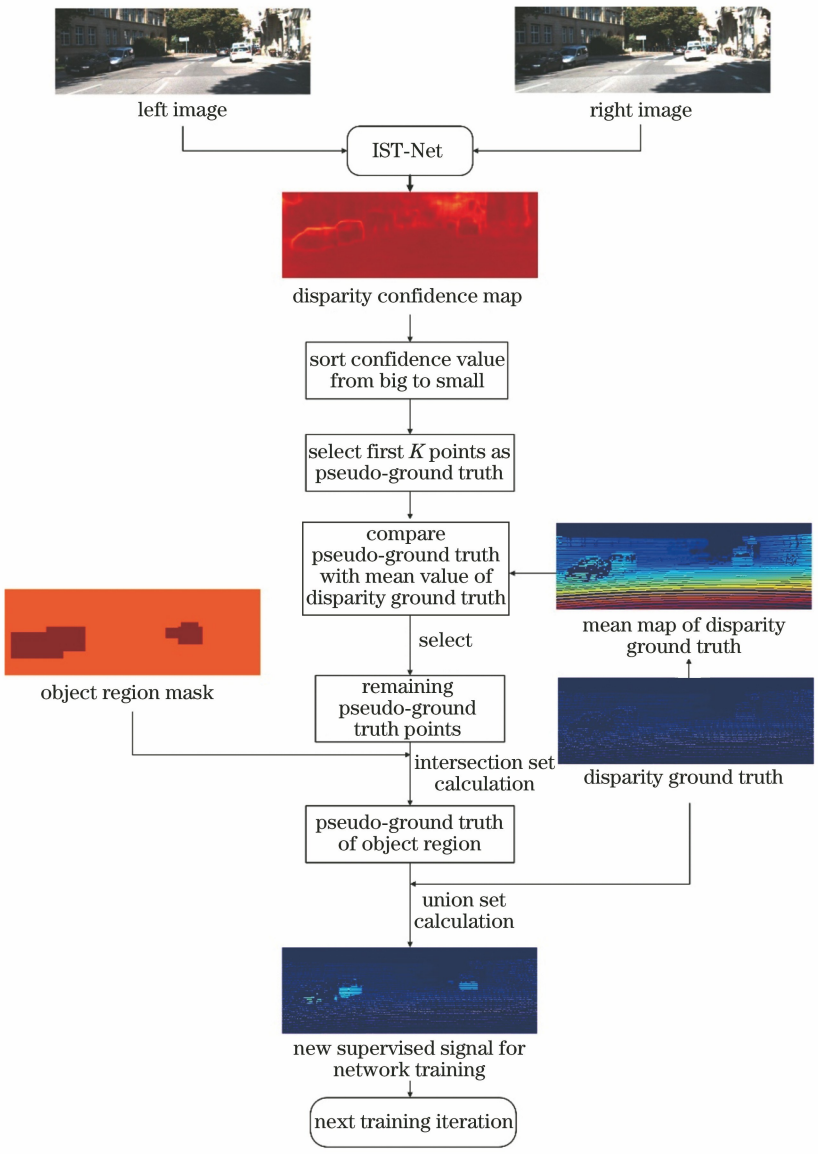

图 3　迭代式自主学习算法流程图

Fig. 3　Flow chart of iterative self-training

首先利用 IST-Net 网络中的置信图预测模块得到视差置信图，以衡量每个点的视差估计值和真实值之间的差值，置信度越高则差值越小。在得到视差置信图后，将没有视差真值的点的置信度从高到低排序，保留前 K 个点作为初始伪真值点。而对于 K 的选择，本文采用了步进的策略，由当前训练的迭代轮数（epoch）与总迭代轮数来决定。随着迭代轮数的增加，网络预测的视差置信图本身的可靠性也在增加，因此 K 值也逐渐增加，即

$$K = R \cdot \frac{e}{E}, \tag{1}$$

式中：e 为当前迭代轮数；E 为总的迭代轮数；R 为常数。

对于最终保留的伪真值点，仅仅以生成的视差置信图作为唯一的标准是不够的，对于一些复杂的场景区域，置信度的预测存在一定的难度。源于图像的连续性，图像中某点的视差值应该接近其周围区域的视差值。基于这一假设，在置信度的基础上提出了第二个伪真值点筛选标准。首先利用 3×3

的滤波窗口对视差真值图进行均值滤波，生成一张视差真值均值图。随后将前面根据置信度排名选出的 K 个初始伪真值点与视差真值均值进行比较，若差值超出三个像素或该点周围视差均值为空，则剔除该点。本文算法最终的任务是三维目标检测，因此应更加重视目标区域的视差估计。利用目标二维框来确定目标区域，利用保留下来的伪真值点与待检测目标区域的掩码求解交集，得到可靠的目标区域伪真值点。最后，将目标区域伪真值点与原真值点进行合并，并将其作为网络下一轮训练的监督信号。需要特别指出的是，在每一轮的训练过程中，只有由真实雷达信号产生的视差真值是始终保留的，而其余额外加入的伪真值点在每一轮训练中都有更新。另外，由于视差置信图的预测和视差图的估计在网络训练初期并不准确，因此在网络训练 100 轮后才加入迭代式自主学习算法。

将视差置信图的取值范围设置为 0～1，取值越接近 1 代表该点预测的视差越准确（即更接近于视差真值）。置信图的真值为

$$g_{\mathrm{gt,confidence}}=1-N\left(\sqrt{\left|g_{\mathrm{gt,disparity}}-P_{\mathrm{PRE,disparity}}\right|}\right),\tag{2}$$

式中：$g_{\mathrm{gt,confidence}}$ 为置信图真值；$g_{\mathrm{gt,disparity}}$ 为视差真值；$P_{\mathrm{PRE,disparity}}$ 为估计视差；$N(\cdot)$ 为归一化函数，计算公式为

$$N(x)=\frac{x-x_{\min}}{x_{\max}-x_{\min}},\tag{3}$$

式中：x 为输入向量；$x_{\max}$ 为输入向量中的最大值；$x_{\min}$ 为输入向量中的最小值。

在立体匹配网络中，置信图预测的目标损失函数设计为

$$L_{\mathrm{confidence}}=-\sum_{i=1}^{N}B(c_i,\hat{c}_i),\tag{4}$$

$$B(c_i,\hat{c}_i)=c_i\ln\hat{c}_i+(1-c_i)\ln(1-\hat{c}_i),\tag{5}$$

式中：$L_{\mathrm{confidence}}$ 为置信图预测的目标损失函数；N 为真值点的个数；c_i 为置信图真值中第 i 个点的值；$\hat{c}_i$ 为预测置信图中第 i 个点的值。

2.1.3　选择性优化损失函数

为了进一步生成更加精准的目标区域视差，在视差估计的目标损失函数中引入了选择性优化策略，引导网络重点优化目标区域的视差值。在目标区域和背景区域，分别赋予不同的权重对损失函数进行约束，使网络更侧重于目标区域的视差估计，这可以理解为应用于损失函数的由任务驱动的注意力机制。

与 PSMNET[14] 相同，使用视差回归的方式估算连续的视差图。对于回归问题，常用的损失函数一般包括均方误差（L_2）与平均绝对误差（L_1）。在 CSPN[15] 中已证明，与 L_2 损失函数相比，L_1 损失函数能取得更好效果，因此选取 L_1 损失函数，并在目标区域赋予其更大的权重，以引导网络加强对目标区域的特征提取，进而生成更加精准的目标区域视差。视差估计的目标损失函数计算公式为

$$L_{\mathrm{disparity}}=\eta_1\frac{1}{N_{\mathrm{F}}}\sum_{i=1}^{N_{\mathrm{F}}}\mathrm{smooth}_{L_1}(d_i-\hat{d}_i)+\eta_2\frac{1}{N_{\mathrm{B}}}\sum_{i=1}^{N_{\mathrm{B}}}\mathrm{smooth}_{L_1}(d_i-\hat{d}_i),\tag{6}$$

$$\mathrm{smooth}_{L_1}(D)=\begin{cases}0.5D^2, & |D|<1\\ |D|-0.5, & \text{otherwise}\end{cases},\tag{7}$$

式中：$L_{\mathrm{disparity}}$ 为视差估计的目标损失函数；d_i 为第 i 个视差真值；$\hat{d}_i$ 为对应的第 i 个视差预测值；N_{F} 为目标区域中的视差真值点的个数；N_{B} 为背景区域中的视差真值点的个数；η_1 与 η_2 为权重。

2.2　基于自适应特征融合机制的三维目标检测算法

由于点云能够提供场景的三维信息，RGB 图像则能提供丰富的色彩与纹理信息，因此在进行三维目标检测时，同时利用点云信息与 RGB 信息，可以实现较稳定的检测性能。然而，在检测过程中，面对场景中的不同区域，不同模态的信息对检测结果的重要性是不同的，如在目标存在遮挡的区域中，当目标物体之间的色彩、纹理相似时，若过分依赖 RGB 特征则可能引起检测性能的下降。相反，当非目标物体的三维特征与目标物体相似时，过分依赖点云信息则可能造成误检。此外，在远距离的小目标物体上，因点云分布较稀疏，若过分依赖点云信息也可能会造成漏检情况。因此，在目标检测过程中，不同模态信息的融合应根据场景中的不同情况实现自适应调整。

在 AVOD[1] 网络框架的基础上，设计了一个端到端的基于自适应特征融合机制的三维目标检测网络（SAFF-3DOD-Net），其结构如图 4 所示，网络包含多级自适应特征融合模块，该模块的设计受空间注意力机制的启发[16]，通过自动学习场景中不同区域的特征融合权重，实现多模态特征的有效融合，进而得到准确且稳定的目标检测结果。

2.2.1　SAFF-3DOD-Net 结构

SAFF-3DOD-Net 结构如图 4 所示，整个网络包含两个子网络——候选区域生成子网络（region proposal network）与检测子网络（detection network）。

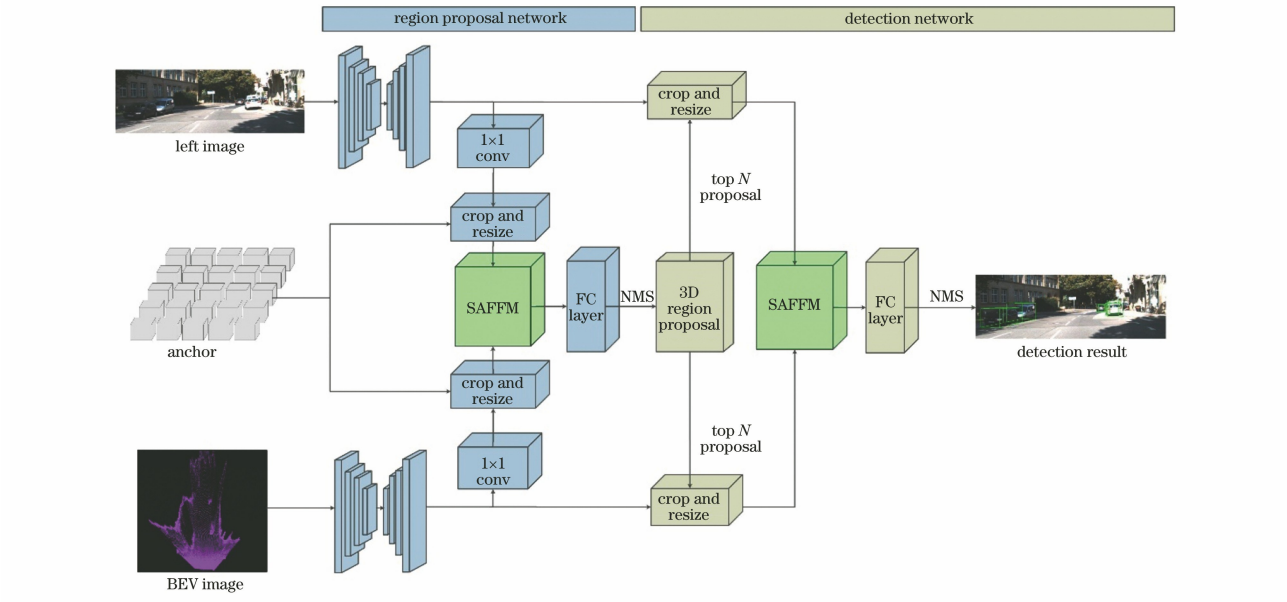

图 4　SAFF-3DOD Net 结构示意图

Fig. 4　Architectural diagram of SAFF-3DOD Net

首先，利用视差信息与相机内外参数计算场景点云，随后利用 Pseudo-LiDAR[12] 中的方法，计算点云鸟瞰图（BEV map）并将三维空间分割成若干 3D 锚（anchor）。最终，将计算得到的点云鸟瞰图与左目 RGB 图一同送入 SAFF-3DOD-Net 网络中进行目标检测。在第一阶段的候选区域生成子网络中，首先对 RGB 图与点云鸟瞰图分别提取特征，随后利用 SAFFM，将每个 3D anchor 对应的 RGB 图特征与点云鸟瞰图特征进行融合，并利用全卷积（FC）层生成无朝向的候选三维区域（3D region proposal）。在第二阶段的检测网络中，SAFFM 将每个候选三维区域对应的 RGB 图特征与点云鸟瞰图特征进行融合，利用全卷积层以及非极大值抑制（NMS）生成精确的有朝向的目标三维框，最终实现三维目标的识别与定位。

2.2.2　SAFFM

SAFFM 的具体结构如图 5 所示，对于输入的高为 H，宽为 W，通道数为 C 的 RGB 图特征（F_{RGB}）与点云鸟瞰图特征（F_{BEV}），首先进行重塑（reshape），分别得到 I_{RGB} 与 I_{BEV}，将其分别输入到

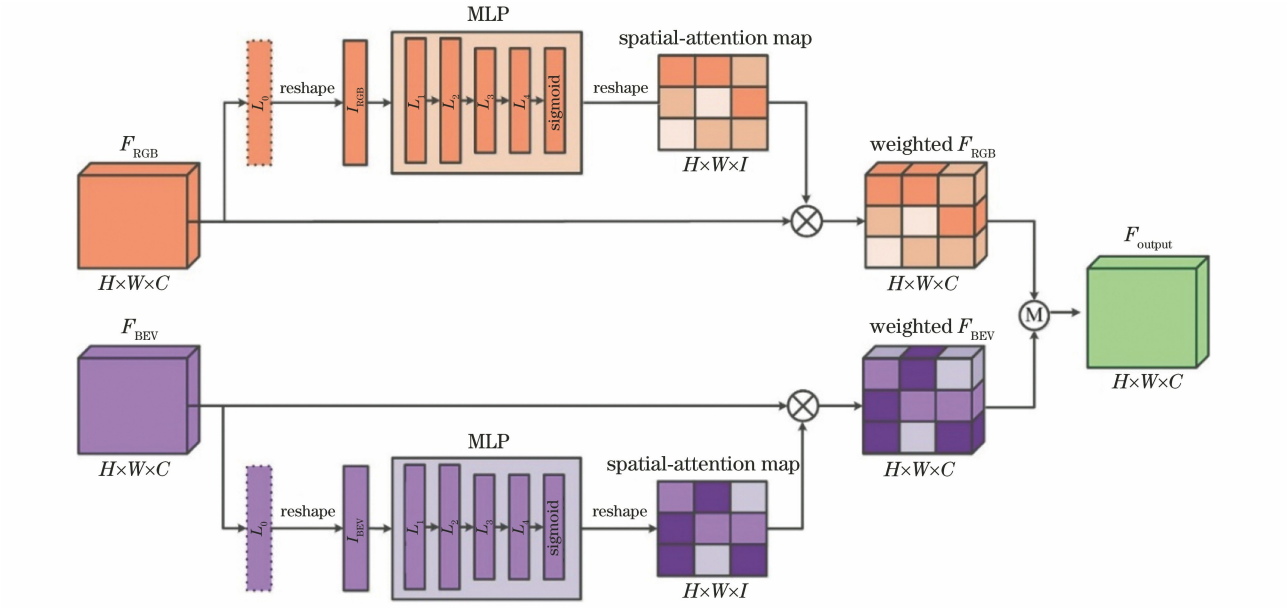

图 5　SAFFM 示意图

Fig. 5　Diagram of SAFFM

两个多层感知器(MLP)中,利用卷积层 L_1～L_4 进行特征的编码、解码,并利用 sigmoid 层对特征进行激活。随后,将 MLP 输出的特征再次重塑,得到两个空间注意力图(spatial-attention map)。利用空间注意力图,分别对 RGB 图特征与点云鸟瞰图特征进行加权,并将加权后的特征(weighted F_{RGB}、weighted F_{BEV})进行元素级的平均操作(element-wise mean operation),此操作在图 5 中用 M 表示,最终得到融合特征 F_{output}。此外,在检测子网络中的 SAFFM 里,增加了一层卷积核为 1×1 的 L_0 卷积层,对输入的特征进行降维,从而降低运算量。SAFFM 的具体参数设置如表 1 所示,其中卷积层设置中的参数表示为卷积核个数-卷积核尺寸。

表 1　SAFFM 参数

Table 1　Detailed configuration of SAFFM

Parameter	SAFFM in region proposal network		SAFFM in detection network	
	Layer setting	Output dimension	Layer setting	Output dimension
$F_{\mathrm{RGB}}/F_{\mathrm{BEV}}$		3×3×1		7×7×32
L_0			1-1×1	7×7×1
$I_{\mathrm{RGB}}/I_{\mathrm{BEV}}$		9×1×1		49×1×1
L_1	36-1×1	36×1×1	98-1×1	98×1×1
L_2	36-1×1	36×1×1	98-1×1	98×1×1
L_3	18-1×1	18×1×1	49-1×1	49×1×1
L_4	9-1×1	9×1×1	49-1×1	49×1×1
Sigmoid		9×1×1		49×1×1
Spatial-attention map		3×3×1		7×7×1
Weighted F_{RGB} and weighted F_{BEV}		3×3×1		7×7×32
F_{output}		3×3×1		7×7×32

3　实验结果与分析

3.1　数据集介绍

由于本文所提算法需要双目图像以及雷达监督信号,因此本文的实验评估基于以下两个数据集展开。

1) KITTI 立体匹配数据集(KITTI stereo 2012&2015)[17-18]。数据集包括真实场景下的 RGB 双目图像和视差真值,其中 KITTI stereo 2012 包含 194 对训练数据集与 195 对测试数据集,KITTI stereo 2015 包含 200 对训练数据集与 200 对测试数据集。数据集中每帧图像对应的视差真值图由雷达信息重建而成,视差真值点是相对稠密的,且在 KITTI stereo 2015 数据集中进一步利用 CAD 模型信息对运动物体的视差进行稠密重建。由于本文提出的基于迭代式自主学习的视差估计算法针对只有稀疏视差监督信号(由每帧图像对应的瞬时雷达信息计算而成)的情况,因此在网络训练时利用原有的视差真值并不适合。鉴于以上原因,根据 UnFlow[19] 中提供的列表,在 KITTI raw data[20] 里找到了 KITTI stereo 2012 和 KITTI stereo2015 训练数据集对应的瞬时雷达信息,并将其转换为稀疏视差真值用于视差估计网络的训练。此外,由于数据集并不提供目标物体的二维框,因此根据文献[21]与文献[22]中提供的语义以及实例标签,生成目标二维框用于实验。参考 PSMNET[14] 中的划分比例,将所有训练数据集的 80% 作为训练集,20% 的数据作为校验集。

2) KITTI 三维目标检测数据集(KITTI 3D object detection 2017)[17]。数据集包括真实室外场景下的 RGB 双目图像和雷达点云数据,其中有 7481 对训练图像对和 7518 对测试图像对,包含三种目标类别(汽车,行人和骑自行车者)。对于每个类别,数据集的检测难易程度可以划分为容易,中等和困难,在评估的时候也会分等级评估算法结果。由于测试集的真值并不公开,因此本文与文献[1,12,23]一样,将原有训练集划分为新的训练集和验证集,其中训练集样本 3712 对,用于模型的训练,验证集样本 3769 对,用于评估模型性能。

3.2　实验细节

IST-Net 网络的训练利用 PyTorch 平台实现，采用 Adam 优化器（一阶矩估计指数衰减率 $\beta_1=0.9$，二阶矩估计指数衰减率 $\beta_2=0.999$）。与 PSMNET 算法相同，先利用 Scene Flow 数据集[24]进行预训练，随后在对应数据集中对网络进行微调训练。在训练前，对图片进行归一化处理，并随机剪裁（$H=256$，$W=512$）。利用 Scene Flow 数据集进行模型训练时，学习率设为 0.001，训练轮数为 10。模型微调时，前 200 轮的学习率设为 0.001，后 100 轮的学习率设为 0.0001，视差的最大值设为 192，训练 batch size 设为 12，在 4 块 16G 显存的 Tesla-P100 GPU 上训练。在迭代自主学习的过程中，参数 R 取值 2%；在视差估计损失函数中，权重 η_1 与 η_2 分别为 1 与 0.8。同时，对于网络输出的三个视差图与置信图，基于总的网络损失函数，设置对应的损失函数权重分别为 0.5、0.7、1.0。利用 Tensorflow 平台训练 SAFF-3DOD-Net 网络，训练参数与 AVOD[1]算法保持一致，采用 Adam 优化器，初始学习率设置为 0.0001，每一个 30000 次迭代，衰减系数为 0.8，一共训练 12×10^4 次迭代。训练 batch size 设为 1，在一块 12G 显存的 Titan-X GPU 上进行训练。在本文实验中，目标物体为车辆。

3.3　评价指标

本文的评价指标包括用于评估视差计算结果的视差错误率（disparity error rate）和用于评估三维目标检测结果的平均准确率（AP）。下面对这两个评价指标逐一介绍。

1）视差错误率

本文选择了与 KITTI 评测网站一致的视差评价指标，即计算所得的视差图与真实视差图的错误率。视差错误率定义为

$$E_1(u,v)=|d(u,v)-d_{gt}(u,v)|, \tag{8}$$

$$E_2(u,v)=\frac{|d(u,v)-d_{gt}(u,v)|}{d_{gt}(u,v)}, \tag{9}$$

$$E_{Error\_rate}=\frac{\sum\limits_{(u,v)\in d\cap d_{gt}}(E_1(u,v)>\delta_1 \& E_2(u,v)>\delta_2)}{N_{d\cap d_{gt}}}, \tag{10}$$

式中：E_{Error_rate} 为视差错误率；d 为预测视差图；d_{gt} 为真值视差图；$N_{d\cap d_{gt}}$ 是在预测视差图和真值视差图中同时有值的像素的个数；$d(u,v)$ 为点 (u,v) 的预测视差；$d_{gt}(u,v)$ 为点 (u,v) 的真值视差；δ_1 和 δ_2 为容错阈值，分别取值 3 pixel 和 5%。通过上述定义可以得出视差错误率即为估计误差大于 3 pixel 且估计误差大于真值的 5%的像素所占的百分比。

2）平均精度

对于三维目标检测的评估，选取与 KITTI 测评网站一致的平均精度 AP 作为评价指标。

首先计算交并比（IoU），公式为

$$I_{IoU}=\frac{A(B_p\cap B_{gt})}{A(B_p\cup B_{gt})}, \tag{11}$$

式中：I_{IoU} 为交并比；B_p 与 B_{gt} 分别为预测的目标框和真值目标框；当目标框为二维框时，A 为面积，当目标框为三维框时，A 为体积。

基于得到的 IoU，按照不同类别的阈值要求，计算由当前模型得到的结果对应的精度（precision）和召回率（recall）。随后利用文献[25]中的算法计算平均精度，得到最终 AP 结果。值得注意的是，文献[25]分析得出，在计算 AP 时 40 个采样 recall 要比传统 11 个采样 recall 更加公正，因此 KITT 评测网站自 2019 年 10 月 8 日起更新了测评指标，在本文中所有 AP 的计算都基于 40 个采样 recall。本文具体的评估目标检测算法性能的评价指标包括 AP_{3D}、AP_{BEV} 两项，其区别在于计算 IoU 时使用的维度不同。计算 AP_{3D} 时使用的 IoU 为预测目标三维框与真值三维框的重叠区域；计算 AP_{BEV} 时使用的 IoU 则为预测目标三维框与真值目标三维框在鸟瞰图上的二维投影的重叠区域。

3.4　视差估计实验结果分析

1）定量实验结果分析

为了证明本文提出的基于迭代式自主学习的视差估计算法的有效性，首先在 KITTI 三维目标检测数据集中的校验集上对视差结果进行定量分析，并与基准算法 PSMNET[14]进行对比。对全图、目标区域以及背景区域上的视差结果分别进行了评估，实验结果如表 2 所示，只利用迭代式自主学习算法的检测结果用 Ours（IST）表示；只利用选择性优化损失函数的结果用 Ours（SOL）表示；同时采用以上两种改进策略的方法，即本文最终视差估计算法用 Ours（IST+SOL）表示。结果表明，无论是迭代式自主学习算法还是选择性优化损失函数，都在一定程度上提高了目标区域视差结果的准确性，当两种策略融合在一起时，通过相互促进进一步提高了目标区域视差的估计性能。同时，所提算法在优化目标区域视差的同时，并没有影响背景区域视差的准确性。

为了进一步验证本文视差估计算法的有效性，在KITTI立体匹配数据集中的校验集上，与具有代表性的PSMNET[14]、Stereonet[26]算法进行定量对比分析，结果如表3所示。可以看出，在较小的数据集中，所提算法仍能生成更加精准的视差信息。

表2　KITTI三维目标检测校验集上的视差估计的定量比较

Table 2　Quantitative comparison of disparity estimation on KITTI 3D object detection validation set

Method	Disparity error rate /%		
	Object region	Background region	Global image
PSMNET(base)	8.96	4.35	5.49
Ours(IST)	8.69	4.18	5.27
Ours(SOL)	8.72	4.20	5.30
Ours(IST+SOL)	8.60	4.17	5.25

表3　KITTI立体匹配校验集上的视差估计的定量比较

Table 3　Quantitative comparison of disparity estimation on KITTI stereo matching validation set

Method	Disparity error rate /%		
	Object region	Background region	Global image
Stereonet	11.14	5.23	6.99
PSMNET	7.23	3.33	4.44
Ours(IST+SOL)	6.83	3.20	4.27

2）定性实验结果分析

图6为视差结果的定性比较，其中图6(a)为RGB原图，图6(b)为PSMNET视差结果，图6(c)为本文算法生成的视差结果。可以看出，所提算法在目标区域能够得到较精准的视差。同时，相较于PSMNET[14]生成的视差结果，所提算法的视差结果能够保持较完整的目标轮廓。

将视差结果进一步转换成点云，在三维空间下进行观察，图7为点云结果的定性比较，其中图7(a)为RGB原图，图7(b)为基准算法PSMNET对应的点云结果，图7(c)为本文算法对应的点云结果。为了便于观察，在图7(a)、(b)中，三维框为车辆三维框真值，用于标注车辆在点云空间中的位置。如图7(b)所示，基准算法结果中第一行与第四行的车头部分的点云以及第二行与第三行图车侧区域的点云都超出了三维框真值；由所提算法得到的车辆点云则较好地保持在三维框真值内，保证了车辆三维信息的准确性，如图7(a)所示。因此，本文提出的视差估计算法得到的目标区域视差更加准确。

3.5　三维目标检测实验结果分析

1）定量实验结果分析

在3769张校验集上的三维目标检测实验结果的定量分析如表4所示，评价指标包括鸟瞰图上的平均精度(A_{BEV})与三维空间的平均精度(A_{3D})，分

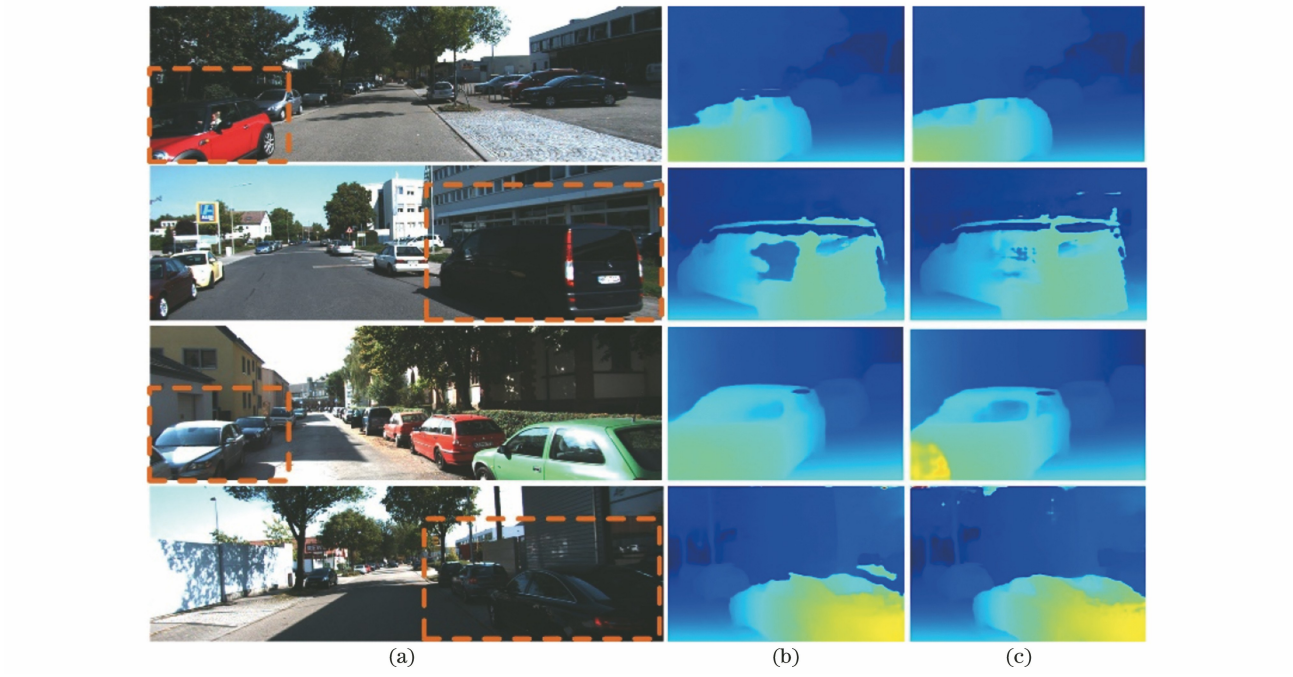

图6　视差定性结果比较。(a) RGB左目图；(b) PSMNET算法；(c)本文视差估计算法

Fig. 6　Qualitative comparison of baseline and our method on estimated disparity map. (a) RGB left image; (b) PSMNET method; (c) our disparity estimation method

容易(Easy)、中等(Moderate)和困难(Hard)三个难度等级，IoU 的取值分别为 0.5 与 0.7。在表 4 中，只利用自适应特征融合算法的结果用 Ours(SAFF)表示，同时采用迭代式自主学习算法、选择性优化损失函数、自适应特征融合算法策略(即本文最终算法)的实验结果用 Ours 表示。可以看出，当 IoU 取值 0.5 时，本文最终的检测算法相对基准算法 Pseudo-LiDAR[12]在 Hard 数据集上的检测精度提高显著，A_{3D}与 A_{BEV}均提高 6.9。当 IoU 取值 0.7 时，本文最终检测算法相对基准算法 Pseudo-LiDAR[12]，在 Easy 数据集上 A_{BEV}与 A_{3D}分别提高了 5.3%与 4.3%，在 Moderate 数据集上 A_{BEV}和 A_{3D}分别提高 5.1%与 2.8%，而在 Hard 数据集上 A_{BEV}提高了 4%，A_{3D}提高了 2.6%。实验结果证明，本文的检测算法能够得到更加精准的目标检测结果。

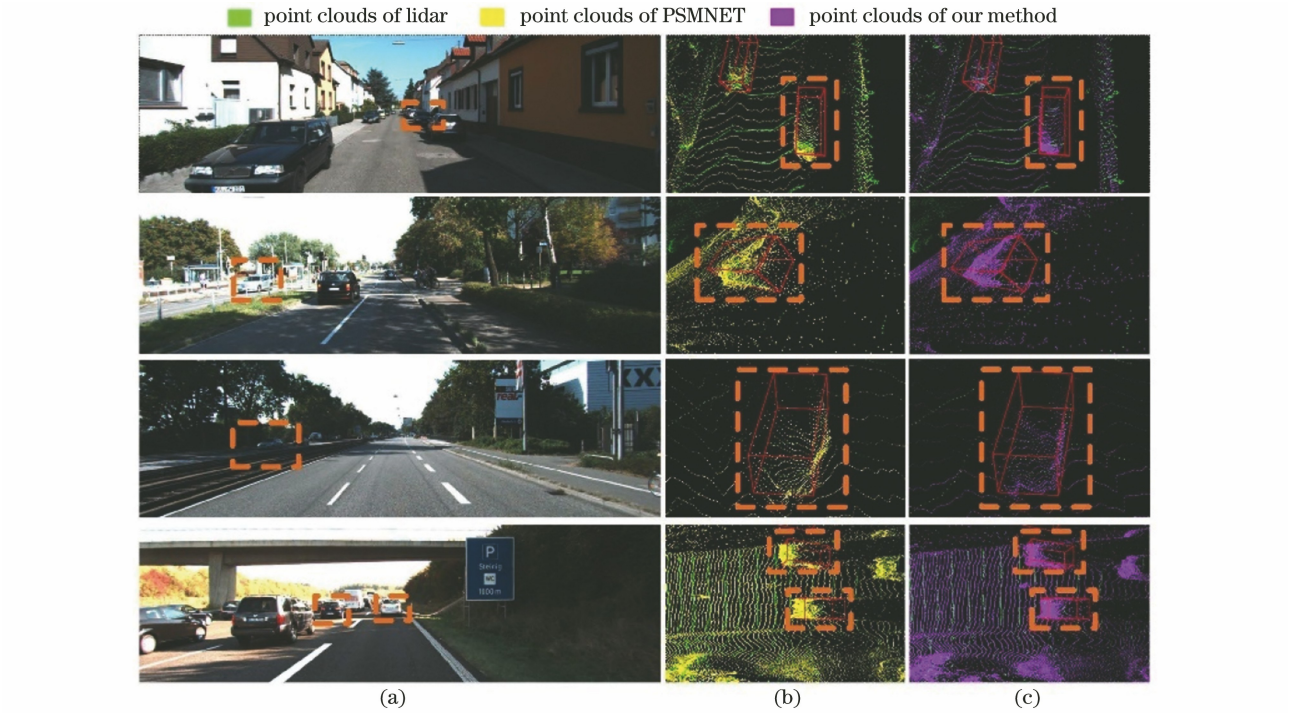

图 7　点云定性结果比较。(a) RGB 左目图；(b) PSMNET 算法；(c)本文视差估计算法

Fig. 7　Qualitative comparison of baseline and our method on estimated point cloud. (a) RGB left image; (b) PSMNET method; (c) our disparity estimation method

表 4　KITTI 三维目标检测校验集上的三维目标检测的定量比较(A_{BEV}和 A_{3D}单位均为%)

Table 4　Quantitative comparison of 3D object detection on KITTI 3D object detection validation set (units of A_{BEV} and A_{3D} are both %)

Method	IoU is 0.5			IoU is 0.7		
	Easy	Moderate	Hard	Easy	Moderate	Hard
Pseudo-LiDAR(base)	92.1/91.6	78.3/75.3	66.7/63.8	75.6/61.5	55.6/43.3	48.3/36.8
Ours(IST)	92.1/91.0	80.4/77.4	70.8/67.8	77.5/61.3	59.5/43.3	50.6/36.9
Ours(SOL)	92.3/91.5	80.6/75.9	69.1/66.2	78.8/63.1	58.2/43.7	50.1/37.4
Ours(IST+SOL)	92.1/91.4	81.0/78.0	69.2/66.3	78.4/63.5	59.6/45.0	50.8/38.6
Ours(SAFF)	92.0/91.5	78.3/75.4	68.5/65.5	77.7/63.0	57.1/43.3	48.6/37.0
Ours	94.5/92.5	81.6/78.6	73.6/70.7	80.9/65.8	60.7/46.1	52.3/39.4

2) 定性实验结果分析

图 8 为三维目标检测的定性结果比较，其中图 8(a)为基准算法 Pseudo-LiDAR[12]得到的车辆三维检测结果，图 8(b)为本文算法得到的结果。可以看出，本文算法的检测结果更加准确，如第三行、第六行、第七行检测结果，图 8(b)相对图 8(a)更接近真

值。除此之外，相对于基准算法，本文算法在漏检问题上也有改善，如第二行与第三行，车辆间遮挡以及视野截断在图 8(a)中造成的漏检，在图 8(b)中得到了改善。此外，对于远处小目标车辆的检测，本文算法也展现了较好的优越性，如图 8(b)第四、五行所示，本文算法能够检测出图 8(a)中漏检的远处车辆目标。同时，在解决目标误检问题上，本文算法也展现了较好的性能，图 8(a)第一行出现右下角的误检情况，在 8(b)中得到了改善。

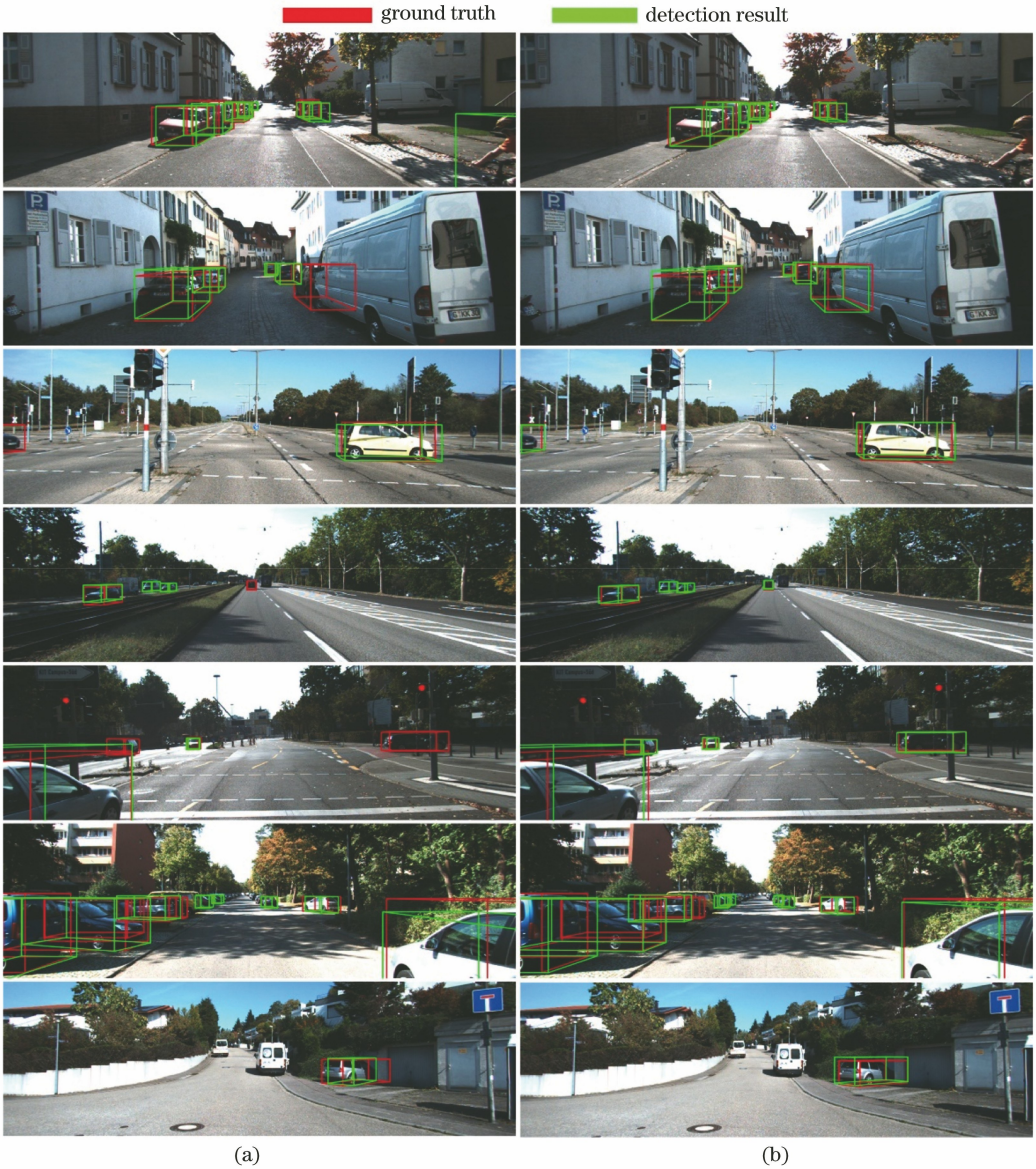

图 8　三维目标定性结果比较。(a) Pseudo- LiDAR 算法；(b)本文算法

Fig. 8　Qualitative comparison of 3D object detection results. (a) Pseudo- LiDAR; (b) our method

3.6　KITTI 网络平台测试结果

将本文所提算法结果上传到 KITTI 三维目标测试平台，在 7518 张测试集上进行测评，在汽车目标检测任务中，测试平台在计算 AP 时 IoU 取值为 0.7。实验结果如表 5 所示，相对基准算法 Pseudo-LiDAR[12]，本文所提算法在 Easy 数据集上的 A_{BEV} 和 A_{3D} 均提高了 4.17%；在 Moderate 数据集上的 A_{3D} 和 A_{BEV} 分别提高了 3.87% 与 4.61%；在 Hard 数据集上的 A_{3D} 提高了 3.74%，A_{BEV} 提高了 4.31%。此外，与其他近年较流行的检测算法进行对比，结果表明，本文算法具有更优的性能。

4　结　　论

提出一种基于双目视觉的三维目标检测算法。首先，为了获取更加精准的目标点云用于三维目标检测任务，提出一种基于迭代式自主学习的视差估

表 5　KITTI 平台上的三维目标检测结果(A_{BEV} 和 A_{3D} 单位均为%)

Table 5　3D object detection results on KITTI test benchmark(units of A_{BEV} and A_{3D} are both %)

Method	Input	Easy	Moderate	Hard
MonoPSR[7]	Monocular	18.33/10.76	12.58/7.25	9.91/5.85
Mono3D_PLiDAR[8]	Monocular	21.27/10.76	13.92/7.50	11.25/6.10
TopNet-HighRes[2]	Lidar	67.84/12.67	53.05/9.28	46.99/7.95
M3D-RPN[9]	Monocular	21.02/14.76	13.67/9.71	10.23/7.42
AM3D[10]	Monocular	25.03/16.50	17.32/10.47	14.91/9.52
RT3D[3]	Lidar	56.44/23.74	44.00/19.14	42.34/18.86
RT3DStereo[13]	Stereo	58.81/29.90	46.82/23.28	38.38/18.96
Stereo R-CNN[27]	Stereo	61.92/47.58	41.31/30.23	33.42/23.72
Pseudo-LiDAR[12]	Stereo	67.30/54.53	45.00/34.05	38.40/28.25
Ours	Stereo	71.47/58.70	49.61/37.92	42.71/31.99

计算法，通过增加视差监督信号以及引入选择性优化策略，提升了目标区域的视差估计精度。随后，提出一种自适应特征融合机制，通过 RGB 信息与点云信息的自适应有效融合，实现了目标物体的精准识别与定位。基于较精准的目标区域视差以及自适应特征融合策略，所提算法的三维目标检测精度与目前较流行的基于视觉系统的检测算法相比得到了较大提升。在今后的工作中，将在现实场景中采集更多数据，对本文算法进行进一步验证。

参　考　文　献

[1] Ku J, Mozifian M, Lee J, et al. Joint 3D proposal generation and object detection from view aggregation [C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 1-5, 2018, Madrid. New York: IEEE, 2018: 1-8.

[2] Wirges S, Fischer T, Stiller C, et al. Object detection and classification in occupancy grid maps using deep convolutional networks [C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC), November 4-7, 2018, Maui, HI. New York: IEEE, 2018: 3530-3535.

[3] Zeng Y M, Hu Y, Liu S C, et al. RT3D: real-time 3-D vehicle detection in LiDAR point cloud for autonomous driving [J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3434-3440.

[4] Yang Z T, Sun Y N, Liu S, et al. STD: sparse-to-dense 3D object detector for point cloud [EB/OL]. (2019-07-22) [2019-10-08]. https://arxiv.org/abs/1907.10471.

[5] Shi S S, Wang X G, Li H S. PointRCNN: 3D object proposal generation and detection from point cloud [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 770-779.

[6] Zhang Y, Ren G Q, Cheng Z Y, et al. Application research of there-dimensional LiDAR in unmanned vehicle environment perception [J]. Laser & Optoelectronics Progress, 2019, 56(13): 130001.
张银，任国全，程子阳，等. 三维激光雷达在无人车环境感知中的应用研究[J]. 激光与光电子学进展，2019，56(13)：130001.

[7] Ku J, Pon A D, Waslander S L. Monocular 3D object detection leveraging accurate proposals and shape reconstruction [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 11867-11876.

[8] Weng X S, Kitani K. Monocular 3D object detection with pseudo-LiDAR point cloud [EB/OL]. (2019-03-23) [2019-10-08]. https://arxiv.org/abs/1903.09847.

[9] Brazil G, Liu X M. M3D-RPN: monocular 3D region proposal network for object detection [EB/OL]. (2019-07-13) [2019-10-08]. https://arxiv.org/abs/1907.06038.

[10] Ma X Z, Wang Z H, Li H J, et al. Accurate monocular object detection via color-embedded 3D reconstruction for autonomous driving [EB/OL]. (2019-03-27) [2019-10-08]. https://arxiv.org/abs/1903.11444.

[11] Chen X, Kundu K, Zhu Y, et al. 3D object proposals for accurate object class detection [C]// Proceedings of the 28th International Conference on Neural Information Processing Systems, December 7-12,

2015, Montreal, Quebec, Canada. New York: Curran Associates, 2015: 424-432.

[12] Wang Y, Chao W L, Garg D, et al. Pseudo-LiDAR from visual depth estimation: bridging the gap in 3D object detection for autonomous driving[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 8445-8453.

[13] Konigshof H, Salscheider N O, Stiller C. Realtime 3D object detection for automated driving using stereo vision and semantic information[C]//2019 IEEE Intelligent Transportation Systems Conference (ITSC), October 27-30, 2019, Auckland, New Zealand. New York: IEEE, 2019: 19211370.

[14] Chang J R, Chen Y S. Pyramid stereo matching network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT. New York: IEEE, 2018: 5410-5418.

[15] Cheng X J, Wang P, Yang R G. Depth estimation via affinity learned with convolutional spatial propagation network[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 108-125.

[16] Park J, Woo S, Lee J, et al. BAM: bottleneck attention module[EB/OL]. (2018-07-18)[2019-10-03]. https://arxiv.org/abs/1807.06514.

[17] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE, 2012: 3354-3361.

[18] Menze M, Heipke C, Geiger A. Joint 3D estimation of vehicles and scene flow[J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015, II-3/W5: 427-434.

[19] Meister S, Hur J, Roth S. UnFlow: unsupervised learning of optical flow with a bidirectional census loss[EB/OL]. (2017-11-21)[2019-10-03]. https://arxiv.org/abs/1711.07837.

[20] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The KITTI dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[21] Guney F, Geiger A. Displets: Resolving stereo ambiguities using object knowledge[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 4165-4175.

[22] Abu Alhaija H, Mustikovela S K, Mescheder L, et al. Augmented reality meets computer vision: efficient data generation for urban driving scenes[J]. International Journal of Computer Vision, 2018, 126(9): 961-972.

[23] Qi C R, Liu W, Wu C X, et al. Frustum PointNets for 3D object detection from RGB-D data[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 918-927.

[24] Mayer N, Ilg E, Hausser P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4040-4048.

[25] Simonelli A, Bulò S R R, Porzi L, et al. Disentangling monocular 3D object detection[EB/OL]. (2019-05-29)[2019-10-03]. https://arxiv.org/abs/1905.12365.

[26] Khamis S, Fanello S, Rhemann C, et al. StereoNet: guided hierarchical refinement for real-time edge-aware depth prediction[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 596-613.

[27] Li P L, Chen X Z, Shen S J. Stereo R-CNN based 3D object detection for autonomous driving[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 7644-7652.