智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

多视角数据融合的特征平衡YOLOv3行人检测研究

陈丽， 马楠， 逄桂林， 高跃， 李佳洪， 张国平， 吴祉璇， 姚永强

引用本文:
陈丽, 马楠, 逄桂林, 等. 多视角数据融合的特征平衡YOLOv3行人检测研究[J]. 智能系统学报, 2021, 16(1): 57–65.
CHEN Li, MA Nan, PANG Guilin, et al. Research on multi–view data fusion and balanced YOLOv3 for pedestrian detection[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(1): 57–65.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202010003

DOI: 10.11992/tis.202010003

多视角数据融合的特征平衡 YOLOv3 行人检测研究

陈丽[1], 马楠[1,2], 逄桂林[3], 高跃[4], 李佳洪[1,2], 张国平[1], 吴祉璇[1], 姚永强[1]

（1. 北京联合大学 北京市信息服务工程重点实验室, 北京 100101; 2. 北京联合大学 机器人学院, 北京 100101; 3. 北京交通大学 计算机与信息技术学院, 北京 100044; 4. 清华大学 软件学院, 北京 100085）

摘　要: 针对复杂场景下行人发生遮挡检测困难以及远距离行人检测精确度低的问题, 本文提出一种多视角数据融合的特征平衡 YOLOv3 行人检测模型 (MVBYOLO), 包括 2 部分: 自监督学习的多视角特征点融合模型 (Self-MVFM) 和特征平衡 YOLOv3 网络 (BYOLO)。Self-MVFM 对输入的 2 个及以上的视角数据进行自监督学习特征, 通过特征点的匹配实现多视角信息融合, 在融合时使用加权平滑算法解决产生的色差问题; BYOLO 使用相同分辨率融合高层语义特征和低层细节特征, 得到平衡的语义增强多层级特征, 提高复杂场景下车辆前方行人检测的精确度。为了验证所提出方法的有效性, 在 VOC 数据集上进行对比实验, 最终 AP 值达到 80.14%。与原 YOLOv3 网络相比, 本文提出的 MVBYOLO 模型精度提高了 2.89%。

关键词: 多视数据; 自监督学习; 特征点匹配; 特征融合; YOLOv3 网络; 平衡特征; 复杂场景; 行人检测

中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2021)01−0057−09

中文引用格式: 陈丽, 马楠, 逄桂林, 等. 多视角数据融合的特征平衡 YOLOv3 行人检测研究 [J]. 智能系统学报, 2021, 16(1): 57–65.

英文引用格式: CHEN Li, MA Nan, PANG Guilin, et al. Research on multi-view data fusion and balanced YOLOv3 for pedestrian detection[J]. CAAI transactions on intelligent systems, 2021, 16(1): 57–65.

Research on multi-view data fusion and balanced YOLOv3 for pedestrian detection

CHEN Li[1], MA Nan[1,2], PANG Guilin[3], GAO Yue[4], LI Jiahong[1,2], ZHANG Guoping[1], WU Zhixuan[1], YAO Yongqiang[1]

(1. Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China; 2. College of Robotics, Beijing Union University, Beijing 100101, China; 3. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044; 4. School of Software, Tsinghua University, Beijing 100085)

Abstract: Because of the occlusion and low accuracy of long-distance detection, pedestrian detection in complex scenes is difficult. Therefore, a pedestrian detection method based on multi-view data fusion and balanced YOLOv3 (MVBYOLO) is proposed, including the self-supervised network for multi-view fusion model (Self-MVFM) and balanced YOLOv3 network (BYOLO). Self-MVFM fuses two or more input perspective data through a self-supervised network and incorporates a weighted smoothing algorithm to solve the color difference problem during the fusion; BYOLO uses the same resolution to fuse high- and low-level semantic features to obtain balanced semantic information, thereby enhancing multi-level features and improving the accuracy of pedestrian detection in front of vehicles in complex scenes. A comparative experiment is conducted on the VOC dataset to verify the effectiveness of the proposed method. The final AP value reaches 80.14%. The experimental results indicate that compared with the original YOLOv3 network, the accuracy of the MVBYOLO is increased by 2.89%.

Keywords: multi-view data; self- supervised learning; feature point matching; feature fusion; YOLOv3 network; balanced feature; complex scene; pedestrian detection

安全性是无人驾驶技术研究成果落地应用的重要需求。无人驾驶技术需要与周围环境形成良好的交互[1]。无人驾驶需要具备认知能力, 才能更好地学习。对周围环境的感知、主动学习是无人驾驶技术必须攻克的一个难点[2]。其中, 行人

收稿日期: 2020−10−07.

基金项目: 国家自然科学基金项目 (61871038, 61931012, 6183034); 军委装备发展部共性预研计划项目 (41412040302); 北京联合大学“人才强校优选计划”领军计划 (BPHR2020AZ02); 北京联合大学研究生科研创新资助项目 (YZ2020K001).

通信作者: 马楠. E-mail: xxtmanan@buu.edu.cn.

检测就是无人驾驶进行环境认知的一个必备环节。行人检测工作主要是判别在输入的视频、图像中是否含有行人并返回其位置。在无人驾驶场景下，一旦未能及时、准确地检测出行人，就会造成伤亡，后果不堪设想，所以无人驾驶条件下对行人检测的准确性有极高的要求。因为行人存在不同的运动姿态、不同的穿衣风格，行人被别的障碍物遮挡以及行人之间互相遮挡[3]，复杂交通场景下光线不统一等问题，行人检测一直是无人驾驶领域重点研究的问题[4]。

复杂交通场景下的行人检测要求在发生部分遮挡时，仍能检测出行人，并且要求能快速有效地检测出车辆前方远距离的小目标行人 (小目标指在整张图片中目标的像素点小于 32×32，或者目标尺寸低于原图像尺寸的 10%[5])。但是，在实际实验中，依靠单一视角的数据，行人发生遮挡时很难被检测到。

为了解决发生遮挡以及远距离行人检测困难的问题，本文提出一种基于多视角数据融合的特征平衡 YOLOv3 行人检测模型 (multi-view data and balanced YOLOv3, MVBYOLO)。首先输入不同视角的图像，使用自监督学习的多视角特征点融合网络模型 (self-supervised network for multi-view fusion model, Self-MVFM) 对其进行特征点提取与匹配，实现多视角图像融合。但是在实际问题中不同角度的摄像机采集的图像融后会产生色差。本文在多视角图像融合时引入改进的加权平滑算法，有效解决不同视角图像融合时产生色差的问题。此外，为了提高复杂交通场景下车辆前方远距离行人的检测精度，本文提出了一个特征平衡的 YOLOv3 网络 (balanced YOLOv3, BYOLO)，在接收到经过 Self-MVFM 网络融合的多视角图像后，用 Darknet-53 网络对图像进行特征提取，可以获得分辨率不同的特征。分辨率高的低层特征包括行人的轮廓、衣着颜色、纹理等信息；分辨率低的高层特征包括肢体、人脸等语义信息。对获得的低层特征与高层特征进行采样，映射到中间层级的分辨率进行特征融合、修正，再通过相反的采样方式适配到原分辨率的特征图，与 Darknet-53 提取的原始特征进行融合，再利用融合后的特征预测行人。在公共数据集 VOC 上的实验结果表明，本文提出的 MVBYOLO 行人检测模型可以有效提高复杂场景下的行人检测精度。

1 基于多视角数据的行人检测研究

1.1 多视角数据融合算法

针对多角度、多尺度的特征如何进行融合的问题，一直受到研究者们的关注。Farenzena 等[6] 提出了一种对称驱动的局部特征累积方法，该方法从结构元素成分分析模型[7] 提取的行人轮廓中找到垂直对称轴，然后根据像素的权重提取颜色和纹理特征。Wen 等 [8] 提出从几张已知相机位置的多视角彩色图片生成三角网格模型的网络结构，使用图卷积神经网络从多视角图片的交叉信息学习进一步提升形状质量。相比于直接建立从图像到最终 3D 形状的映射，本文预测一系列形变，逐渐将由多视角图片生成的粗略形状精细化。Chen 等[9] 通过输入多张不同角度的图片，提取不同的点云特征，再进行融合，从而生成最终的点云。与基于代价体的同类网络相比，这种基于点云的网络结构具有更高的准确性，更高的计算效率和更大的灵活性。Yi 等[10] 引入 2 种新颖的自适应视图融合 (逐像素视图融合和体素视图融合)，考虑在不同视角图像间多重匹配的不同重要性，优化了代价体的计算方法并且引入了新的深度图聚合结构，提高了 3D 点云重建的鲁棒性和完整性。旷世科技公司提出的双向网络[11]，利用深度学习模型，对提取的空间信息特征和全局语义特征进行融合，兼顾了语义分割任务的速度与语义信息。Su 等[12] 提出多视角卷积网络 (multi-view convolutional neural networks,MVCNN)，利用二维的 CNN 网络对多个视角的图像进行融合，实验结果显示比直接用 3D 检测方法更好。Feng 等[13] 提出的组视图卷积网络框架，在 MVCNN 基础上增加分组模型，将不同视角的信息根据相关性进行分组后，再进行特征融合。Dong 等[14] 在 CVPR2019 上提出一种利用外观特征和几何约束相似性矩阵共同寻找各个视角中满足回路一致性的二维姿态匹配关系，实现了较好的多视角下多人的三维姿态估计结果。澳洲国立大学郑良老师实验室提出多视角检测模型[15] 联合考虑多个相机，利用特征图的投影变换进行多相机信息融合，提高虚拟场景下行人发生遮挡时的检测效率。

1.2 行人检测方法

行人检测是目标检测领域的一个重要分支，其主要任务是找出输入的图像或视频帧中存在的行人，并用矩形框输出行人位置和大小。然而行人的着装风格、姿势、形状不同，并且面临被物体遮挡以及行人互相遮挡、拍摄光照不同、拍摄角度不同等因素的影响，使得行人检测任务一直受到视觉研究者们的关注。从研究历史来看，行人检测方法可以分为 2 个主要方向：基于传统算法的行人检测和基于深度学习的行人检测。

1.2.1 基于传统算法的行人检测

传统算法的典型代表是利用方向梯度直方图(histogram of oriented gradient, HOG)进行行人特征提取, 并利用支持向量机(support vector machine, SVM)算法进行分类[16]。HOG是一种重要有效的图像局部纹理特征描述子。在深度学习特征提取方法未普及之前, 被研究者们广泛使用。Girshick[17] 等提出形变部件模型(deformable parts model, DPM)算法, 使用 HOG 提取特征, 并独立地对行人的不同部位进行建模, 从而在一定程度上解决了行人遮挡难以检测的问题。DPM 中包含 2 个部分: 根部模型和部位模型。根部模型主要是定位对象的潜在区域, 找出可能存在物体对象的区域, 再与部位模型进行确认, 最终采用 SVM 和 AdaBoost 进行分类。另外, 也有部分学者从运动特征角度进行研究。假设捕捉行人运动的摄像机是固定不动的, 则使用背景建模算法提取出运动的前景目标, 再对前景目标进行分类。背景建模算法的思路是: 通过学习前一帧获得背景模型, 把当前帧与背景帧数据进行对比, 得到运动的目标, 代表性方法是高斯混合模型[18]、视频前景提取算法算法[19]、样本一致性建模算法[20]、基于像素的参数自适应算法[21]。

1.2.2 基于深度学习的行人检测

基于传统算法的行人检测在一定条件下可以达到较好的检测效率或准确性, 但仍不能满足实际的应用需求。2012 年 Krizhevsky 等[22] 将深度学习技术应用到图像分类并取得良好效果, 研究者们发现通过神经网络提取的特征具有很强的表达能力和鲁棒性, 使计算机视觉的发展迈上了一个新台阶。因此, 对于行人检测任务, 基于深度学习的方法受到越来越多研究者的青睐。

基于深度学习的行人检测又可分为双阶段检测与单阶段检测。双阶段检测方法首先生成一组稀疏的目标候选框, 然后对候选框进行分类和回归。Girshick[23] 等提出区域卷积神经网络(regions with CNN features, R-CNN), 首次将 CNN 用于目标检测, 极大提高了目标检测的性能。后来 Girshick 在 R-CNN 基础上进行改进, 提出了快速区域卷积神经网络模型[24], 将感兴趣区域提取与特征分类合并在同一个网络结构, 提高了模型训练的速度和检测的准确率。Ren 等[25] 在 Fast R-CNN 上增加区域卷积网络来生成候选区域, 构成一种更快的区域卷积神经网络模型, 端到端的训练方式大大提高了运算速度。

单阶段的方法通过直接对图像中的不同位置, 尺度和长宽比进行规则和密集采样, 以此来预测图像中的目标。以 YOLO[26] 为代表的单阶段检测方法将目标检测任务转换为回归问题, 是一种快速的行人检测方法。除了 YOLO 系列算法, 单阶段检测的模型还包括单次检测模型[27]。Zhang 等[28] 提出基于单次精化神经网络的目标检测方法, 结合单阶段检测速度快及双阶段检测准确率高的优点。方法包括锚窗精化模块和目标检测模块, 2 个模块互相连接, 兼顾了检测的准确率与速度。

2 多视角数据融合的特征平衡

本文提出的多视角数据融合的特征平衡 YOLOv3 行人检测网络 (MVBYOLO) 包括 2 部分: 1) 自监督学习的多视角特征点融合网络模型 (Self-MVFM); 2) 特征平衡 YOLOv3 网络 (BYOLO)。首先对输入的多视角图像做特征匹配, 融合成一个完备的图像, 之后再利用目标检测网络对融合后的图像做训练, 提高遮挡及远距离小尺寸行人检测的精度。网络总体框架如图 1 所示。

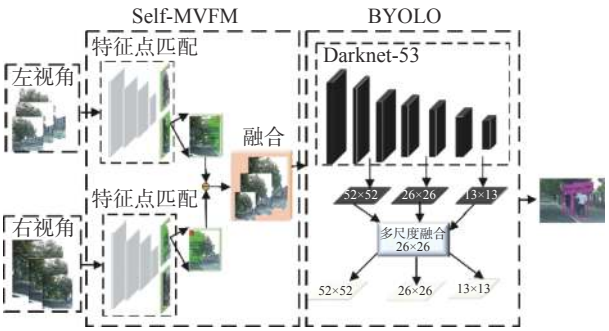

图 1　MVBYOLO 行人检测网络

Fig. 1　Multi-view data fusion and balanced YOLOv3 for pedestrian detection

2.1 自监督学习的多视角特征点融合网络模型

自监督学习的多视角数据融合模型工作流程如下: 图像获取、自监督特征点与描述子提取、特征匹配, 最后进行多视角图像融合。本文提出自监督学习的多视角特征点融合网络模型, 网络结构如图 2 所示。

2.1.1 数据集自标注与模型训练

多视角数据融合过程中的数据集特征点提取任务很难利用人工标注。对于传统的检测、分割任务的标注, 给定一个图像, 通过标注矩形框或者标注物体的轮廓, 可以得到确定的语义真值。但是对于特征点检测任务, 人工很难判断哪一个像素点可以作为特征点, 因此本文利用仅包含简单几何形状的基本数据集和自行采集数据集进行数据集的自标注[29], 具体流程为

1) 利用简单几何形状数据集进行模型的预

训练

简单几何形状数据集是由一些线段、多边形、立方体等特征点较为容易确定的图像构成的。利用尺度不变特征变换等进行基本数据集的特征点提取，可以得到数据集和特征点真值。因为线段、三角形等基础几何形状图像的特征点是真实图像特征点的子集。利用标注好的简单几何形状数据集对特征点检测网络进行训练，得到了一个初级特征点检测网络。与尺度不变特征变换等传统算法相比，在简单几何形状数据集训练得到的初级特征点检测网络在精度方面具有一定的优势，但是在对真实图像数据集进行提取特征点时会出现一些特征点的遗漏问题，检测精确度较低。因此本文利用单应性适应变换和初级特征点检测网络训练得到新的模型，提升真实图像特征点提取的精度。

2) 自行采集图像自标注

利用多次复合几何变换对输入的图像进行处理，本文设置超参数 N_h=80；即 N_1 是未经过复合几何变换的原始图像，剩余的 79 帧图像是原始图像经过随机生成的复合简单几何变换形成的图像。利用步骤 1) 中生成的初级特征点检测网络对真实图像数据集伪特征点进行提取，将与源图像对应的 79 帧图像映射回原图像的特征点累加起来形成新的源图像特征点。至此本文完成了真实图像数据集的特征点标注。

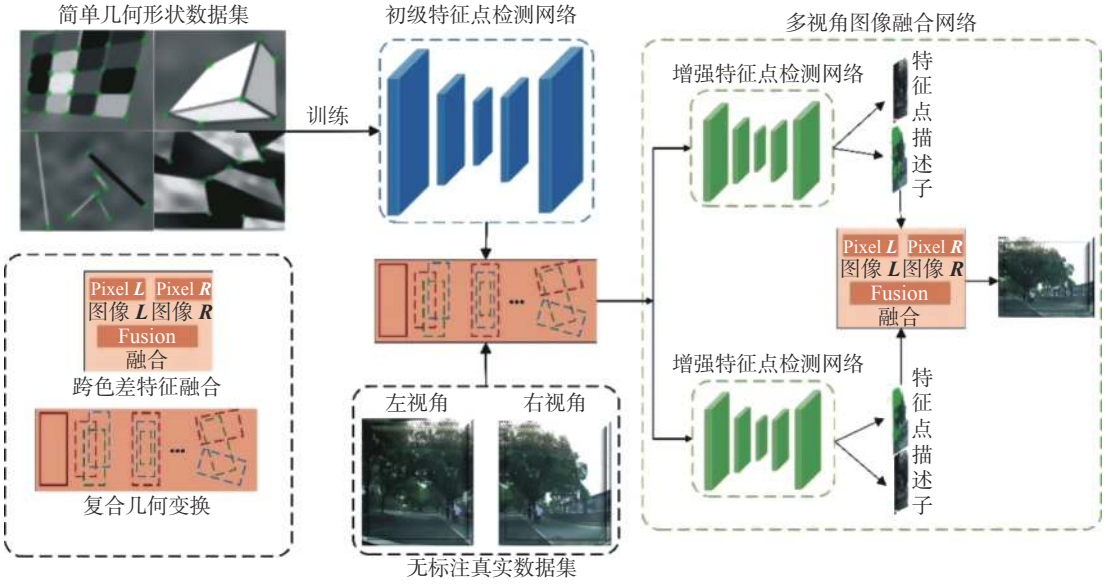

图 2　Self-MVFM 网络模型

Fig. 2　Self-supervised multi-view feature fusion model

在复合简单几何变换中，本文获取了 79 帧经过已知变换矩阵形成的源图像变换图像，因此获得了源图像和其对应的 79 帧图像的 79 组已知位姿变换的图像对。这样就得到了原始图像与变换图像之间映射关系的真值。最终的自行采集数据集包含特征点和特征点描述子真值，用于特征点检测网络中特征点检测和描述子检测 2 个网络分支的联合训练。

为了实现特征点检测子网络和描述子检测子网络在初级特征点检测网络中的联合训练，将 2 个检测子网络的损失函数值加权相加，得到统一的损失函数。

2.1.2　复合几何变换

为了将不同视角的信息进行融合，需要先找到不同视角的对应关系。利用自适应单应性变换求解不同视角的对应关系矩阵 $\boldsymbol{H}$。单应性变换为

$$\begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = \boldsymbol{H} \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} \tag{1}$$

式中：(x_1,y_1) 代表来自第一个视角的图片中的某一点；(x_2,y_2) 代表来自另一个视角图片中与 (x_1,y_1) 对应的某点。需要通过 2 张不同视角的照片计算出复合几何变换矩阵 $\boldsymbol{H}$。

通过自监督学习到的复合简单几何变换矩阵并非都是有用的，需要进行选择。为了选取表现较好的复合简单几何变换矩阵，使用截断正态分布在预定范围内进行平移、缩放、平面内旋转和对称透视变换采样。

2.1.3　增强特征点检测网络

在获得数据集的原始图像与真实图像之间映射关系的真值之后，就完成了真实数据集的自标注，实现了难以人工进行标注的真实图像数据集自标注。增强特征点检测网络[11]用于训练前面

获得的自标注图像数据集, 以提高特征点提取的准确性。增强特征点检测网络如图 3 所示。

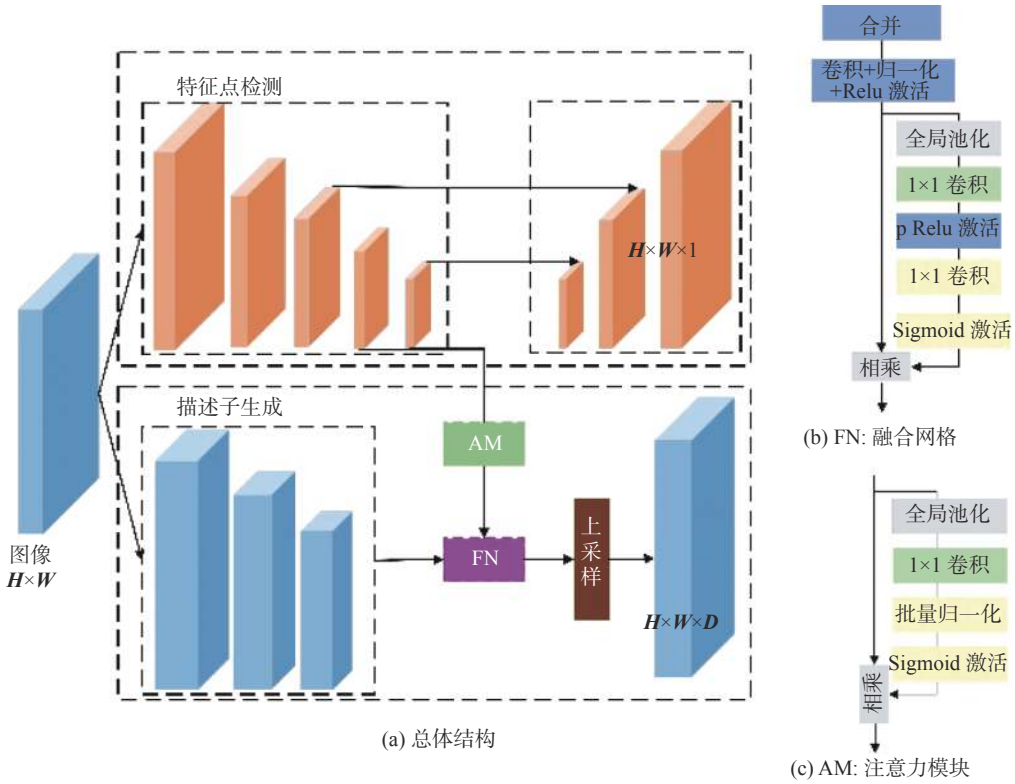

图 3　增强特征点检测网络结构

Fig. 3　Enhanced feature point detection network

多层次编码器: 为了兼顾实时性与精确性, 增强特征点检测网络被设计成 2 个分支, 分别用来处理不同的任务。上边的分支通过非对称的编码解码网络实现对原始图像进行深层特征点的提取。对原始单视图像进行特征描述子的生成, 通过多通道、低层次的编码器网络 (图 3 的下方分支), 提取原始图像表层的特征描述。

特征点检测: 在特征点检测网络部分时, 经过深层、少通道、非对称的编码解码网络得到图像的特征点。

融合网络 (fusion network, FN): 由于网络的特征图并不具有相同的通道和尺寸, 描述子生成网络提取到的特征是浅层的, 包含大量的位置信息, 而特征点检测网络经过多层编码器之后得到的是深层的特征点, 包含胳膊、人脸等信息。为了融合不同层级的特征, 融合网络先通过 Concatenate 操作实现不同层次特征图的简单融合。为了平衡不同尺寸的特征, 在 Concatenate 之后使用了 BatchNorm 操作。把相连接的特征经过全局池化、1×1 卷积得到一个新的权重。这样做的目的是对连接后的特征进行一个新的特征选择和结合。至此, 本文得到了 $W\times H\times D$ 的描述子检测结果, 其中 W 是原始图像的宽、H 是原始图像的长、D 是原始图像的通道。

注意力模块 (attention model, AM): 经过全局池化之后, 可以简单得到深层全局语义信息, 并通过 1×1 卷积操作平衡多层次编码器得到的不同特征图通道之间的差异。

2.1.4　加权平滑算法

在实际应用中, 自行采集的数据集由于相机的架设位置和光照条件变化原因, 存在 2 个视角点信息因光场变化产生的色差问题, 影响后续融合效果。因此, 在融合时本文采用加权平滑算法来解决存在的色差问题。加权平滑算法主要思想: 用 $\boldsymbol{f}(x,y)$ 表示重叠区域融合后的图像, 由 2 幅待融合图像 $\boldsymbol{f}_L$ 和 $\boldsymbol{f}_R$ 加权平均得到, 即: $\boldsymbol{f}(x,y)=\alpha\times\boldsymbol{f}_L(x,y)+(1-\alpha)\boldsymbol{f}_R(x,y)$, 其中 α 是可调因子。

一般情况下 $0<\alpha<1$, 即在图像交叉区域中, 沿视角 1 图像向视角 2 图像的方向, α 由 1 渐变为 0, 从而实现交叉区域的平滑融合。为了给 2 幅图像建立更大的相关性, 使用式 (2) 进行融合处理:

$$\boldsymbol{f}(x,y)=\begin{cases}\boldsymbol{f}_L(x,y), & (x,y)\in\boldsymbol{f}_L\\ \alpha\times\boldsymbol{f}_L(x,y)+(1-\alpha)\times\boldsymbol{f}_R(x,y), & (x,y)\in\boldsymbol{f}_L\cap\boldsymbol{f}_R\\ \boldsymbol{f}_R(x,y), & (x,y)\in\boldsymbol{f}_R\end{cases}\tag{2}$$

令 $\alpha=\dfrac{d_1^2}{d_1^2+d_2^2}$, 则 $1-\alpha=\dfrac{d_2^2}{d_1^2+d_2^2}$, 其中 d_1、d_2 分别表示交叉区域中的点到 2 个不同视角图像交叉区

域的左边界和右边界的距离。

2.2　特征平衡的 YOLOv3 网络

YOLOv3 网络是一种单阶段目标检测方法，与 RCNN 系列的目标检测框架不同，YOLOv3 网络不生成候选框，直接在输出层返回边界框的位置及其所属类别。YOLOv3 借鉴残差网络（residual network, ResNet）[30]、特征金字塔网络[31]网络的思想，添加跨层跳跃连接，融合粗细粒度的特征，能更好地实现检测任务。添加多尺度预测，即在 3 个不同尺寸的特征图层进行预测，每种尺度预测 3 个锚框。锚框的设计方式使用聚类，得到 9 个聚类中心，将其按照大小均分给 3 个特征图层。尺寸分别为 13×13、26×26、52×52。本文将对 3 个不同尺寸的特征进行融合。

YOLOv3 的特征提取网络为 Darknet-53，其网络结构如图 4 所示。Darknet-53 网络中的 Convolutional 代表一个激活函数 (darknetconv2d_BN_leaky, DBL) 操作流程，包含卷积层、批量归一化层（batch normalization, BN）和 Leaky_Relu 激活函数。对于 YOLOv3来说，BN 层和 Leaky_Relu 是和卷积层不可分离的部分，共同构成了最小组件。此外，还包括 Resn残差模块，图 4 中最左面的数字 1、2、8、8、4 表示残差单元的个数。

Darknet-53 加深了网络结构，处理速度为 78 张/s，比 Darknet-19 慢，但是与相同精度的 ResNet-152 相比，处理速度快了 1 倍，所以 Darknet-53 是兼顾速度与精度的特征提取网络架构。

原 YOLOv3 网络中通过 3 种不同尺度的特征图直接做预测，不同尺度分别包括 13×13、26×26、52×52。为了更好地使用深层特征与浅层特征进行小尺寸行人检测，本文提出一种特征平衡的 YOLOv3 网络结构，如图 5 所示。

	类型	卷积核数量	卷积核大小	步长	输出特征图像素大小
	卷积	32	3×3	1	256×256
	卷积	64	3×3	2	128×128
1×	卷积	32	1×1	1	
	卷积	64	3×3	1	
	残差链接				128×128
	卷积	128	3×3	2	64×64
2×	卷积	64	1×1	1	
	卷积	128	3×3	1	
	残差链接				64×64
	卷积	256	3×3	2	32×32
8×	卷积	128	1×1	1	
	卷积	256	3×3	1	
	残差链接				32×32
	卷积	512	3×3	2	16×16
8×	卷积	256	1×1	1	
	卷积	512	3×3	1	
	残差链接				16×16
	卷积	1 024	3×3	2	8×8
4×	卷积	512	1×1	1	
	卷积	1 024	3×3	1	
	残差链接				8×8
	平均池化		全局		
	全连接层		1 000		
	Softmax 分类				

图 4　Darknet-53 网络结构

Fig. 4　Darknet-53 Network

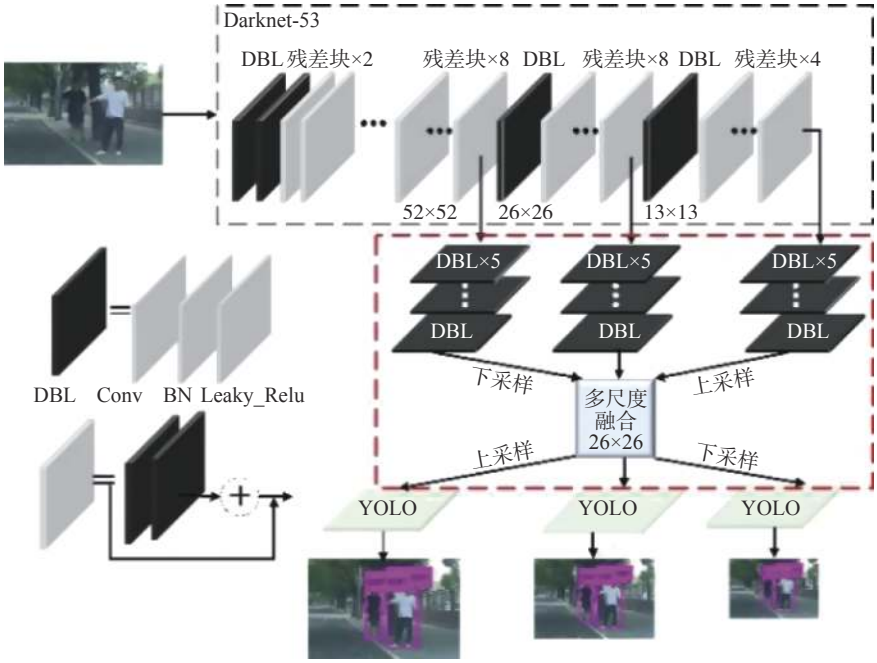

图 5　特征平衡 YOLOv3 网络结构

Fig. 5　Architecture of balance YOLOv3 network

特征融合是将不同类型、不同尺度的特征进行整合，去除冗余信息，从而得到更好的特征表达。在神经网络中直观的融合方式一般分为 Add 和 Concatenate 2 种。Add 方式[32]是特征图相加，从而增加描述图像特征的信息量，即图像本身的维度没有增加，只是每一维下的信息量增加了，这

样的融合方式有利于图像分类任务。Concatenate 方式 [33] 则是通道数的合并，也就是说描述图像本身的特征增加了，而每一特征下的信息并没有增加。深度网络中多层信息的直接拼接并不能更好地利用特征之间的互补性，所以本文考虑将特征采样到相同分辨率大小进行加权融合。

神经网络提取的低层特征分辨率高，可以学习到一幅图像中的细节特征，高层特征分辨率低，可以学习到更好的语义特征。为了更好地结合细节信息和语义信息的优势，本文采用对数据相加取平均的方式来进行特征融合。假设 C_l 代表不同层级的特征数据，则 C_1 代表 52×52 的特征，C_2 代表 26×26 特征数据，C_3 代表 13×13 的特征。本文将 3 个分辨率的特征进行不同的采样方式统一到 26×26 的大小，再利用式 (3) 进行相加取平均，得到一个融合后的特征：

$$C=\frac{1}{3}\sum_{l=1}^{3}C_l \tag{3}$$

在进行尺度缩放的具体操作中，针对 13×13 大小的特征图，对其进行 2 倍的上采样，对于 52×52 大小的特征图，对其进行 2 倍的下采样，这样将原来不同尺度的特征图全部变成了 26×26 的特征，可以直接进行加权求和。得到融合后的特征 C 后，再通过与之前相反的采样操作，即对特征分别进行下采样与上采样的操作，还原成 13×13、52×52 的尺寸，再与原来 Darknet-53 网络提取的第 36、61 与 74 层的原始特征进行融合，利用最终得到的具有细节信息和语义信息的特征去做预测。

3　实验结果与分析

3.1　实验环境

本实验平台为云服务器，操作系统为 Ubuntu 16.04，显卡型号为 GeForce GTX 2080Ti，显存 11 GB，内存 16 GB，Cuda 版本：10.0.130，OpenCV 版本：3.2.0。

3.2　实验数据集

本实验的训练与测试所使用的数据集全部来自 PASCAL VOC 数据集。训练使用 VOC2007 train、valid 与 VOC2012 train、valid 数据集，为了验证算法的有效性，在 VOC2007 test 数据集上做验证。总训练数据共 22 136 张图片，其中包含行人的图片为 6 496 张；总验证数据共 4 952 张图片，其中包含行人的图片为 2 097 张。

3.3　实验参数设置

本文只对行人这一类别做训练，输入的图片大小默认为 416×416，输入通道数为 3，本文设定的迭代次数是 50 200，batchsize 为 64，学习率为 0.001，在迭代到 40 000 次的时候学习率更新为 0.01。将处理好的数据集在同一性能服务器下用 YOLOv3 原模型进行训练。在相同实验环境以及实验参数下，对 MVBYOLO 网络进行训练。将得到的检测结果与 YOLOv3 原模型进行对比，观察改进后的检测模型针对有遮挡远距离行人检测中存在的问题优化效果及性能。

3.4　实验评价指标及结果分析

本文应用准确率 (precision, P)、漏检率 (recall, R) 来衡量检测算法的性能。因为本算法只检测行人，可以看做是一个二分类问题。为了计算准确率和漏检率，引入以下定义：

1) True_Pedetrian(TP)：真实目标是行人且被训练模型检测出来是行人；

2) True_N-Pedestrian(TN)：真实目标不是行人且没有被训练模型错误检测为行人；

3) False_Pedestrian(FP)：表示为真实目标不是行人但被模型错误检测为行人 (误检)；

4) False_N-Pedestrian(FN)：表示真实目标是行人 但是模型没有将其检测为行人 (漏检)。

则评价的标准为

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$
$$\mathrm{AP}=\int_0^1 P(r)\mathrm{d}r \tag{4}$$

将本文提出的 MVBYOLO 模型与原来 YOLOv2 和 YOLOv3 模型作对比，比较损失值下降趋势，PR 曲线以及 AP 值。

从图 6 可以看出，本文 MVBYOLO 网络的训练损失值下降趋势基本与原 YOLOv3 网络保持一致，下降速度快于 YOLOv2，经过相同的训练批次，MVBYOLO 网络的损失值明显低于 YOLOv2，可以更快地收敛。

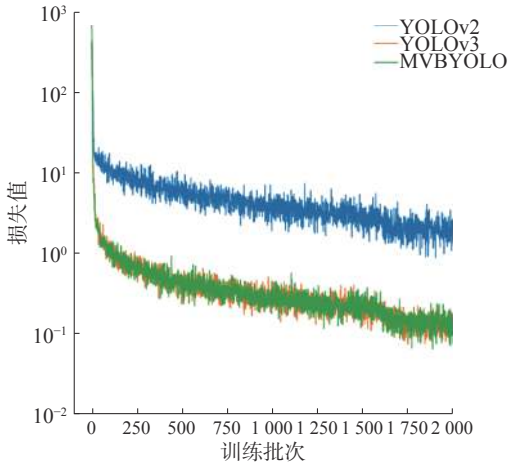

图 6　不同网络的训练损失值

Fig. 6　Train loss of different network

PR 曲线与横纵坐标轴形成了一个平面，面积越大，AP 值越高。图 7 为不同模型的 PR 曲线，显示了本文的模型具有更高的检测精度。

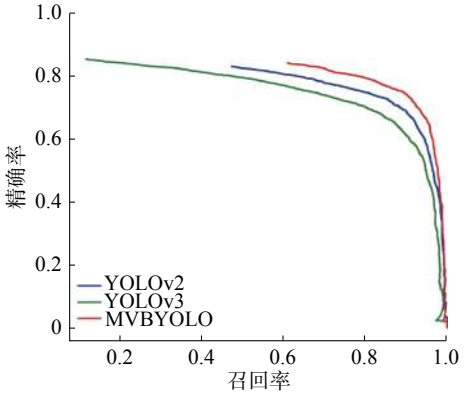

图 7　不同网络的 PR 曲线

Fig. 7　Precision of different network

从表 1 可以看出，本文提出的 MVBYOLO 的 2 个模块 Self-MVFM 与 BYOLO 对行人检测的精度均有贡献。结合提出的 2 个模块，行人检测的精度得到更好的提升，与 YOLOv2 相比，AP 值提高了 3.34%，与 YOLOv3 相比，AP 值提高了 2.89%。

表 1　不同网络在 VOC 数据集上的 AP 值

Table 1　AP of different network in the VOC dataset

网络	AP值
Faster R-CNN	76.70
MR-CNN[34]	76.40
YOLOv2	76.80
YOLOv3	77.25
Self-MVFM+YOLOv3	79.03
BYOLO	78.96
MVBYOLO	80.14

本文网络在真实场景下采集的 240 张图像上进行了检测，从中挑选了在 2 种不同真实场景下拍摄的图像进行展示，图 8 为检测结果。

从图 8 可以看出，在相同场景下，与 YOLOv3 网络相比，本文能检测出更多的行人。

(a) YOLOv3　　(b) MVBYOLO

图 8　不同网络的行人检测结果

Fig. 8　Pedestrian detection results for different network

4　结束语

本文提出的 MVBYOLO 行人检测模型，通过 Self-MVFM 进行自监督多视角信息融合，之后利用平衡的 YOLOv3 网络，准确地进行复杂场景下车辆前方小尺寸行人检测，提高行人检测的效率。本文在 VOC2007test 做测试，AP 值达到 80.14，与原 YOLOv3 网络相比，检测精度提高了 2.89%，取得较好的实验结果。但检测性能还有待优化。下一步研究工作主要针对 2 点：1) 优化损失函数，使模型更快收敛；2) 将多视角行人检测模型作为动作识别的数据预处理模型，将预测的行人检测框直接输入骨架提取网络，降低后续骨架提取任务的难度。

参考文献：

[1] 马楠, 高跃, 李佳洪, 等. 自驾驶中的交互认知 [J]. 中国科学: 信息科学, 2018, 48(8): 1083–1096.
MA Nan, GAO Yue, LI Jiahong, et al. Interactive cognition in self-driving[J]. Scientia sinca informationis, 2018, 48(8): 1083–1096.

[2] LI Deyi, MA Nan, GAO Yue. Future vehicles: learnable wheeled robots[J]. Science China information sciences, 2020, 63(9): 193201.

[3] 贲晛烨, 徐森, 王科俊. 行人步态的特征表达及识别综述 [J]. 模式识别与人工智能, 2012, 25(1): 71–81.
BENXianye, XU Sen, WANG Kejun. Review on pedestrian gait feature expression and recognition[J]. PR and AI, 2012, 25(1): 71–81.

[4] CHEN Li, MA Nan, WANG P, et al. Survey of pedestrian action recognition techniques for autonomous driving[J]. Tsinghua science and technology, 2020, 25(4): 458–470.

[5] 赵永强, 饶元, 董世鹏, 等. 深度学习目标检测方法综述 [J]. 中国图象图形学报, 2020, 25(4): 629–654.
ZHAO Yongqiang, RAO Yuan, DONG Shipeng, et al. Survey on deep learning object detection[J]. Journal of image and graphics, 2020, 25(4): 629–654.

[6] FARENZENA M, BAZZANI L, PERINA A, et al. Person re-identification by symmetry-driven accumulation of local features[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 2360–2367.

[7] ANDRILUKA M, ROTH S, SCHIELE B. Pictorial structures revisited: people detection and articulated pose estimation[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1014–1021.

[8] WEN Chao, ZHANG Yinda, LI Zhuwen, et al. Pixel2Mesh++: multi-view 3D mesh generation via deformation[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South), 2019: 1042–1051.

[9] CHEN Rui, HAN Songfang, XU Jing, et al. Point-based multi-view stereo network[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South), 2019: 1538–1547.

[10] YI Hongwei, WEI Zizhuang, DING Mingyu, et al. Pyramid multi-view stereo net with self-adaptive view aggregation[C]//16th European Conference on Computer Vision. Glasgow, UK, 2020: 766–782.

[11] YU Changqian, WANG Jingbo, PENG Chao, et al. Bisenet: bilateral segmentation network for real-time semantic segmentation[C]//15th European Conference on Computer Vision. Munich, Germany, 2018: 334–349.

[12] SU Hang, MAJI S, KALOGERAKIS E, et al. Multi-view convolutional neural networks for 3D shape recognition[C]//

2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 945–953.
[13] FENG Yifan, ZHANG Zizhao, ZHAO Xibin, et al. GVCNN: group-view convolutional neural networks for 3D shape recognition[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 264–272.
[14] DONG Junting, JIANG Wen, HUANG Qixing, et al. Fast and robust multi-person 3D pose estimation from multiple views[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 7784–7793.
[15] HOU Yunzhong, ZHENG Liang, GOULD S. Multiview detection with feature perspective transformation[C]//16th European Conference on Computer Vision. Glasgow, UK, 2020: 1–18.
[16] LIU Hong, XU Tao, WANG Xiangdong, et al. Related HOG features for human detection using cascaded adaboost and SVM classifiers[C]//19th International Conference on Advances in Multimedia Modeling. Huangshan, China, 2013: 345–355.
[17] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 32(9): 1627–1645.
[18] KAEWTRAKULPONG P, BOWDEN R. An improved adaptive background mixture model for real-time tracking with shadow detection[M]//REMAGNINO P, JONES G A, PARAGIOS N, et al. Video-Based Surveillance Systems. Boston, MA: Springer, 2002: 135–144.
[19] BARNICH O, VAN DROOGENBROECK M. ViBe: a universal background subtraction algorithm for video sequences[J]. IEEE transactions on image processing, 2010, 20(6): 1709–1724.
[20] WANG Hanzi, SUTER D. A consensus-based method for tracking: modelling background scenario and foreground appearance[J]. Pattern recognition, 2007, 40(3): 1091–1105.
[21] HOFMANN M, TIEFENBACHER P, RIGOLL G. Background segmentation with feedback: the pixel-based adaptive segmenter[C]//2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence, USA, 2012: 38–43.
[22] KRIZHEVSKY A, SUTSKEVER I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097–1105.
[23] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 580–587.
[24] GIRSHICK R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 1440–1448.
[25] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of the 2015 Conference and Workshop on Neural Information Processing Systems. Montreal, Canada, 2015: 91–99.
[26] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 779–788.
[27] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//14th European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 21–37.
[28] ZHANG Shifeng, WEN Longyin, BIAN Xiao, et al. Single-shot refinement neural network for object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4203–4212.
[29] DETONE D, MALISIEWICZ T, RABINOVICH A. SuperPoint: self-supervised interest point detection and description[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Salt Lake City, USA, 2018: 224–236.
[30] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 770–778.
[31] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 936–944.
[32] CAO Guimei, XIE Xuemei, YANG Wenzhe, et al. Feature-fused SSD: fast detection for small objects[C]//Proceedings Volume 10615, Ninth International Conference on Graphic and Image Processing (ICGIP 2017). Qingdao, China, 2018: 106151E.
[33] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 2261–2269.
[34] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 2980–2988.

作者简介：

陈丽，硕士研究生，主要研究方向为多视角数据融合、行人动作识别。

马楠，教授，博士，主要研究方向为交互认知、知识发现与智能系统，带领团队分别在 2018、2019、2020WIC 世界无人驾驶挑战赛虚拟场景赛项获得冠军（领军奖）。授权发明专利 7 项、软件著作权 13 项。发表学术论文 50 余篇，主编专著和教材 3 部。

逄桂林，硕士研究生，主要研究方向为计算机视觉、车道线检测。