计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202110956
60(2): 465−474, 2023

基于上下文增强和特征提纯的小目标检测网络

肖进胜[1] 赵 陶[1] 周 剑[2] 乐秋平[1] 杨力衡[1]
[1](武汉大学电子信息学院 武汉 430072)
[2](测绘遥感信息工程国家重点实验室(武汉大学) 武汉 430079)
(xiaojs@whu.edu.cn)

Small Target Detection Network Based on Context Augmentation and Feature Refinement

Xiao Jinsheng[1], Zhao Tao[1], Zhou Jian[2], Le Qiuping[1], and Yang Liheng[1]
[1] (*School of Electronic Information*, *Wuhan University*, *Wuhan* 430072)
[2] (*State Key Laboratory of Information Engineering in Surveying*, *Mapping and Remote Sensing* (*Wuhan University*), *Wuhan* 430079)

Abstract Small objects contain few and fuzzy features, which is a hard problem in the field of object detection. The poor performance of small object detection is mainly caused by the limitation of the network and the imbalance of the training dataset. A novel feature pyramid composite structure constructed by context augmentation module (CAM) and feature refinement module (FRM) is proposed. The feature fusion of multi-scale dilated convolution is applied to generate features on different receptive fields, and then the features are added to detection network to supplement context information. The channel and space feature refinement mechanism is introduced to suppress the conflict information generated by multi-scale feature fusion and prevent small objects from being submerged in the conflict information. Besides, a data augmentation method called copy-reduce-paste is proposed to increase the proportion of small targets, so that the contribution of small targets to the loss value during training is greater and the training is more balanced. Experimental results show that the Mean Average Precision(mAP) of object detection on the VOC dataset of the proposed network is 83.6% (IOU is 0.5). The AP value of small target detection is 16.9% (IOU changes from 0.5 to 0.95), which is 3.9%, 7.7% and 5.3% higher than that of YOLOV4, CenterNet and RefineDet, respectively. The AP value of small target detection on TinyPerson dataset is 55.1%, which is 0.8% and 3.5% higher than that of YOLOV5 and DSFD, respectively.

Key words small target detection; context augmentation; feature refinement; dilated convolution; data augmentation

摘 要 微小目标的纹理模糊、包含特征少,是目标检测领域的难点.针对小目标检测提出一种新的上下文增强模块(context augmentation module, CAM)和特征提纯模块(feature refinement module, FRM)相结合的特征金字塔复合结构.利用多尺度空洞卷积的特征融合,补充网络中的上下文信息;引入通道和空间的特征提纯机制来抑制多尺度特征融合后的冲突信息,防止小目标淹没在冲突信息中;同时,引入复制—缩小—粘贴(copy-reduce-paste)的数据增强方法提高小目标的占比,使训练时小目标对损失值的贡献更大,训练更加平衡.由实验结果可知,所提出的算法在VOC数据集上目标检测的平均精度均值(Mean Average Precision, mAP)达到了83.6%(交并比为0.5);对小目标检测的AP值达到了16.9%(交并比为0.5~0.95),比YOLOV4,CenterNet,RefineDet的分别提高3.9%, 7.7%和5.3%.在TinyPerson数据集上小目标检测的AP

收稿日期:2021−09−23;**修回日期**:2022−04−19
基金项目:国家自然科学基金青年科学基金项目(42101448);中国科学院光电信息处理重点实验室开放课题基金项目(OEIP-O-202009)
This work was supported by the National Natural Science Foundation of China for Young Scientists (42101448) and the Open Project Program Foundation of the CAS Key Laboratory of Opto-Electronics Information Processing (OEIP-O-202009).
通信作者:周剑(jianzhou @whu.edu.cn)

值为 55.1%，比 YOLOV5,DSFD 的分别提高 0.8% 和 3.5%.

关键词 小目标检测；上下文增强；特征提纯；空洞卷积；数据增强

中图法分类号 TP183

小目标检测作为目标检测中的难点技术，被广泛应用于自动驾驶、医学领域、无人机导航、卫星定位和工业检测等视觉任务中. 近些年基于深度学习的目标检测算法发展迅猛. 以 YOLO(You Only Look Once)[1] 和 SSD(Single Shot MultiBox Detector)[2] 为代表的一阶段算法直接预测出目标的位置和类别，具有较快的速度. 而二阶段算法[3-4] 在生成候选框的基础上再回归出目标区域，具有更高的精度.但是这些算法在检测只含有较少像素的小目标(小于 32×32 像素)时表现较差，检测率甚至不到较大目标的一半. 因此，小目标检测仍然具有很大的改进空间.

小目标检测效果差主要是由于网络本身的局限性以及训练数据不平衡所导致[5]. 为了获得较强的语义信息和较大的感受野，检测网络不断堆叠下采样层，使得小目标信息在前向传播的过程中逐渐丢失[6]，限制了小目标的检测性能. 特征金字塔网络(feature pyramid network, FPN)[7] 将低层特征图和高层特征横向融合，可以在一定程度上缓解信息丢失的问题[1-2]. 然而 FPN 直接融合不同层级的特征会造成语义冲突，限制多尺度特征的表达，使小目标容易淹没于冲突信息中. 同时，目前主流的公开数据集中，小目标的数量远远小于较大目标，使得小目标对损失的贡献小，网络收敛的方向不断向较大目标倾斜.

针对小目标检测效果差的问题，本文提出一种上下文增强和特征提纯相结合的复合 FPN 结构，该结构主要包括上下文增强模块(context augmentation module, CAM)和特征提纯模块(feature refinement module, FRM). 同时，提出一种复制—缩小—粘贴(copy-reduce-paste)的数据增强方法，具体有 3 点：

1)CAM 融合多尺度空洞卷积特征以获取丰富的上下文信息，补充检测所需信息；

2)FRM 引入通道和空间自适应融合的特征提纯机制以抑制特征中的冲突信息；

3)通过 copy-reduce-paste 数据增强来提高小目标在训练过程中对损失的贡献率.

1 相关工作

1.1 现代目标检测器

目标检测是一种基础的计算机视觉任务，经过多年的发展，基于卷积神经网络(convolutional neural network, CNN)的目标检测器逐渐成为主流. RCNN[3] 首先生成候选区域以匹配不同尺寸的目标，然后通过 CNN 筛选候选区域. FasterR-CNN[4] 将候选区域生成阶段和分类阶段结合在一起，以提高检测速度. EFPN[8] 提出超分辨率 FPN 结构以放大小目标的特征[9]. 一阶段网络 SSD 将锚盒密集的布置在图像上以回归出目标框，同时充分利用不同尺度的特征，以检测较小目标. YOLOV3[1] 利用特征金字塔的 3 层输出分别检测大、中、小目标，明显提高小目标检测性能. RefineDet[10] 引入一种新的损失函数以解决简单样本和复杂样本不平衡的问题. 同时也有研究者提出基于 anchor-free 架构的检测器[11].尽管目标检测算法发展迅速，但是小目标检测率却一直较低. 本文选用带有 FPN 的 YOLOV3 作为基础网络，并在此基础上做出改进.

1.2 多尺度特征融合

多尺度特征是一种提高小目标检测率的有效方法. SSD[2] 首次尝试在多尺度特征上预测目标位置和类别. FPN[7] 自上而下地将含有丰富语义信息的高层特征图和含有丰富几何信息的低层特征图横向融合. PANet[12] 在 FPN 的基础上添加了额外的自下而上的连接以更高效地传递浅层信息到高层. NAS-FPN[13] 利用神经架构搜索技术搜索出了一种新的连接方式. BiFPN[14] 改良了 PANet 的连接方式，使其更加高效，并在连接处引入了简单的注意力机制. 虽然文献 [12−14] 中的结构都能提升网络多尺度表达的能力，但是都忽略了不同尺度特征之间冲突信息的存在可能会阻碍性能的进一步提升，本文则充分考虑了冲突信息对检测精度的影响.

1.3 数据增强

深度学习是基于数据的方法，因而对训练数据的预处理是其关键的一环. 常见的数据预处理方法如旋转、变形、随机擦除、随机遮挡和光照畸变等. Stitcher[15] 将 4 张训练图像缩小为原图的 1/4，并且将它们拼接为 1 张图像来实现小目标的数据增强，同时将损失值作为反馈信号以指导数据增强的进行. YOLOV4[16] 将 4 张训练图像缩小为不同大小并且拼接为 1 张来实现小目标的数据增强. 文献 [15−16] 中的方式对于目标尺寸普遍很大的图像来说，会将大目标图像缩小为中等目标大小，最终提高中等目标图像的检测率. Kisantal 等人[5] 采用将图像的小目标

区域复制然后粘贴回原图的方式实现小目标数据增强. 但这种方式只能增加小目标个数而不能增加含有小目标的图像个数, 也会造成一定的不平衡. 本文提出的数据增强算法则基于较大目标广泛分布于训练的各个批次的事实, 保证训练平衡进行. 本文算法结构图如图 1 所示:

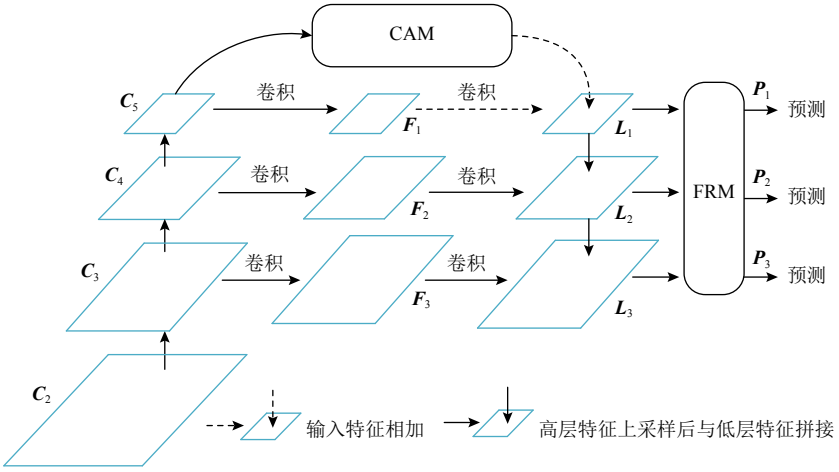


Fig. 1 Overall network structure of FPN

图 1 FPN 总体网络结构

2 本文算法

图 1 中{$\boldsymbol{C}_2$, $\boldsymbol{C}_3$, $\boldsymbol{C}_4$, $\boldsymbol{C}_5$}分别表示图像经过{4, 8, 16, 32}倍下采样后的特征图, {$\boldsymbol{C}_3$, $\boldsymbol{C}_4$, $\boldsymbol{C}_5$}经过 1 层卷积后分别生成{$\boldsymbol{F}_1$, $\boldsymbol{F}_2$, $\boldsymbol{F}_3$}, 其中 $\boldsymbol{C}_2$ 由于含有大量噪声而未被使用. {$\boldsymbol{L}_1$, $\boldsymbol{L}_2$, $\boldsymbol{L}_3$}分别是{$\boldsymbol{F}_1$, $\boldsymbol{F}_2$, $\boldsymbol{F}_3$}经过 FPN 后的结果, {$\boldsymbol{P}_1$, $\boldsymbol{P}_2$, $\boldsymbol{P}_3$}为{$\boldsymbol{L}_1$, $\boldsymbol{L}_2$, $\boldsymbol{L}_3$}经过 FRM 的输出.

CAM 启发于人类识别物体的模式. 如, 我们很难分辨很高天空中的小鸟, 但是考虑天空作为其背景, 我们就很容易分辨出, 因为从我们学习到的知识中可知, 在天空背景下的微小目标很有可能是小鸟, 而这种背景信息, 即是目标的上下文信息.因此如果目标检测网络也在图像中学习到这样的“知识”将会有助于检测小目标.

由于 FPN 不同层的特征密度不同, 因而含有大量的语义差异, 在实现信息共享的同时也引入了很多冲突信息. 因此, 本文提出了 FRM 用于过滤冲突信息, 减少语义差异. FRM 通过将不同层间的特征自适应融合, 以达到抑制层间冲突信息的目的.

针对小目标对损失贡献低的问题, 提出了一种 copy-reduce-paste 数据增强方法, 以提高小目标对损失的贡献.

2.1 上下文增强和特征提纯的特征金字塔网络

2.1.1 上下文增强模块（CAM）

目标检测需要定位信息也需要语义信息, 处于 FPN 最低层的 $\boldsymbol{L}_3$ 含有较多的定位信息而缺少语义信息. FPN 自上而下的信息共享结构在通道数减少之后才进行融合, 使得 $\boldsymbol{L}_3$ 未能获取充分的语义信息. 为此我们利用不同空洞卷积率的空洞卷积来获取上下文信息, 并将其注入到 FPN 中, 以补充上下文信息.

图 2(a)是 CAM 的结构图. 对于大小为 [*bs*, *C*, *H*, *W*] 的输入分别进行空洞卷积率为 1, 3, 5 的空洞卷积[17]. *bs*, *C*, *H*, *W* 分别为特征图的批次大小、通道数、高和宽.由于该模块输入的尺寸较小, 为了获取更多的细节特征, 不宜使用大卷积, 因此选用 3×3 的卷积. 同时为了避免引入较多的参数量, 选取卷积核的个数为 *C*/4, 即首先压缩通道数为输入的 1/4, 然后再通过 1×1 的卷积扩张通道数为 *C*, 得到 3 种大小相同而感受野不同的输出, 最后融合得到的特征. 特征融合可采用的方式如图 2(b)~(d)所示. 图 2(b), (c)分别为拼接融合和加权融合, 即分别在通道和空间维度上直接拼接和相加. 图 2(d)是自适应融合方式, 即通过卷积、拼接和归一化等操作将输入特征图压缩为通道为 3 的空间权重, 3 个通道分别与 3 个输入一一对应, 计算输入特征和空间权重的加权和可以将上下文信息聚合到输出中.

本文通过消融实验验证各个融合方式的有效性, 实验结果如表 1 所示.

由表 1 可知, 对于小目标来说, 拼接融合所取得的增益最大, AP_s 和 AR_s 分别提高了 1.8% 和 1.9%. 自适应融合对中目标的提升最为明显, AP_m 提升了 2.6%. 相加融合带来的提升则基本介于拼接融合和自适应融合两者之间, 因此本文选择拼接融合的方式.

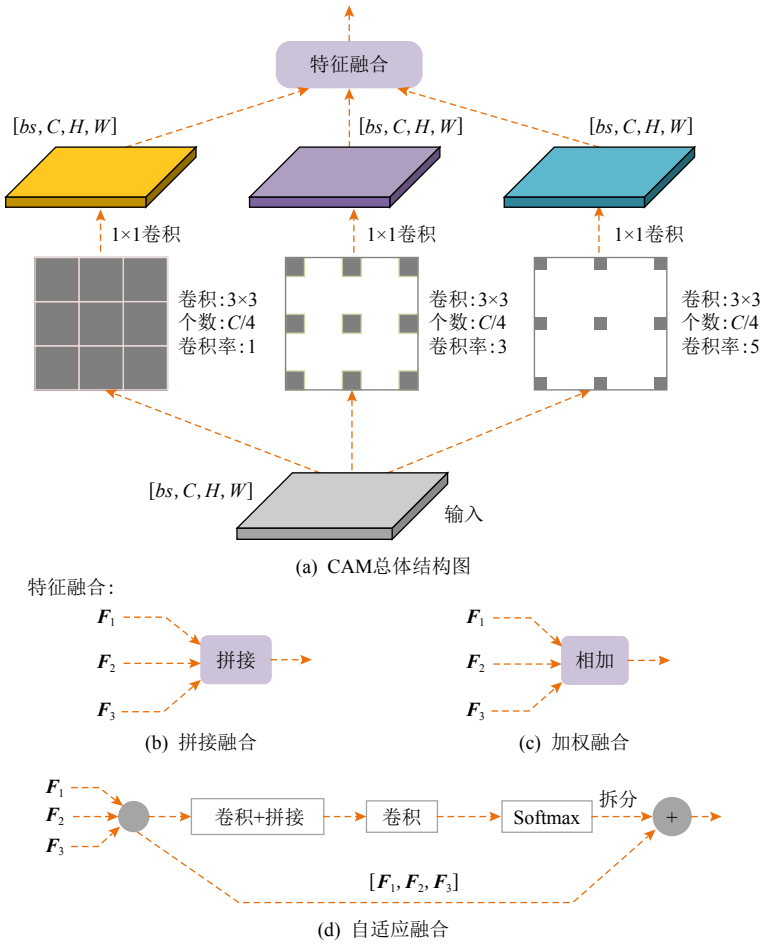


Fig. 2 The structure of CAM

图 2 CAM 结构图

Table 1 Ablation Experimental Results of CAM

表 1 CAM 消融实验结果 %

算法	AP(IOU=0.5)		AR(IOU=0.5)	
	AP_s	AP_m	AR_s	AR_m
基线模型	34.8	60.5	57.9	78.7
相加融合	35.6	63.0	60.5	81.8
自适应融合	36.0	63.1	58.9	81.0
拼接融合	36.6	61.0	59.8	79.5

注：基线模型为 YOLOV3，测试数据集为 VOC，IOU 为交并比. AP_s, AP_m 分别指小目标、中目标的平均精度；AR_s, AR_m 分别指小目标、中目标的平均召回率.

本文将部分特征图可视化以说明 CAM 的效果，可视化结果如图 3 所示.

图 3(b)为 CAM 输入特征图，从中可以发现在图像的目标处有微小响应，呈现为较小的“白点”. 图 3(c)为 CAM 输出特征图，可以明显看到目标处的响应明显增强，并且响应范围更大，这是因为 CAM 将目标周围的上下文信息也融入特征中，使得目标处的响应更强. 因此将 CAM 提取的上下文信息注入网络中将有助于小目标的检测.

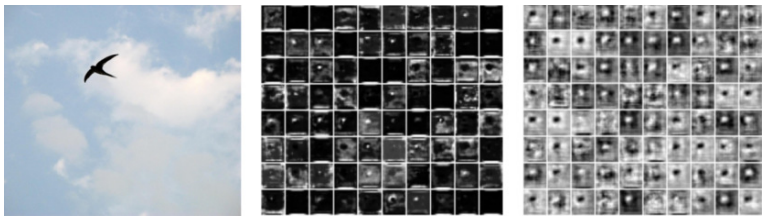
2.1.2 特征提纯模块（FRM）

FPN 用于融合不同尺度大小的特征，然而不同尺度的特征具有不可忽视的语义差异，将不同尺度的特征直接融合可能引入大量的冗余信息和冲突信息，降低多尺度表达的能力. 为了抑制冲突信息，本文提出 FRM，该模块结构如图 4 所示.

图 4(a)为接在 FPN 第 2 层后的 FRM 结构图. 从图(4)可看出，$\boldsymbol{X}^1$, $\boldsymbol{X}^2$, $\boldsymbol{X}^3$(FPN 的 3 层输出)为该模块的输入，首先将 $\boldsymbol{X}^1$, $\boldsymbol{X}^2$, $\boldsymbol{X}^3$ 3 个输入缩放到同一大小，分别为 $\boldsymbol{R}^1$, $\boldsymbol{R}^2$, $\boldsymbol{R}^3$，然后再利用拼接和卷积操作将所有输入特征的通道数压缩为 3，随后接上并联的通道提纯模块和空间提纯模块.

通道提纯模块的具体结构如图 4(b)所示，为了计算通道注意力，采用平均池化和最大池化相结合的方式来聚合图像的全局空间信息. 用 $\boldsymbol{X}^m$ 表示 FRM 的第 $m(m\in\{1, 2, 3\})$ 层输入特征图，其输出可表示为

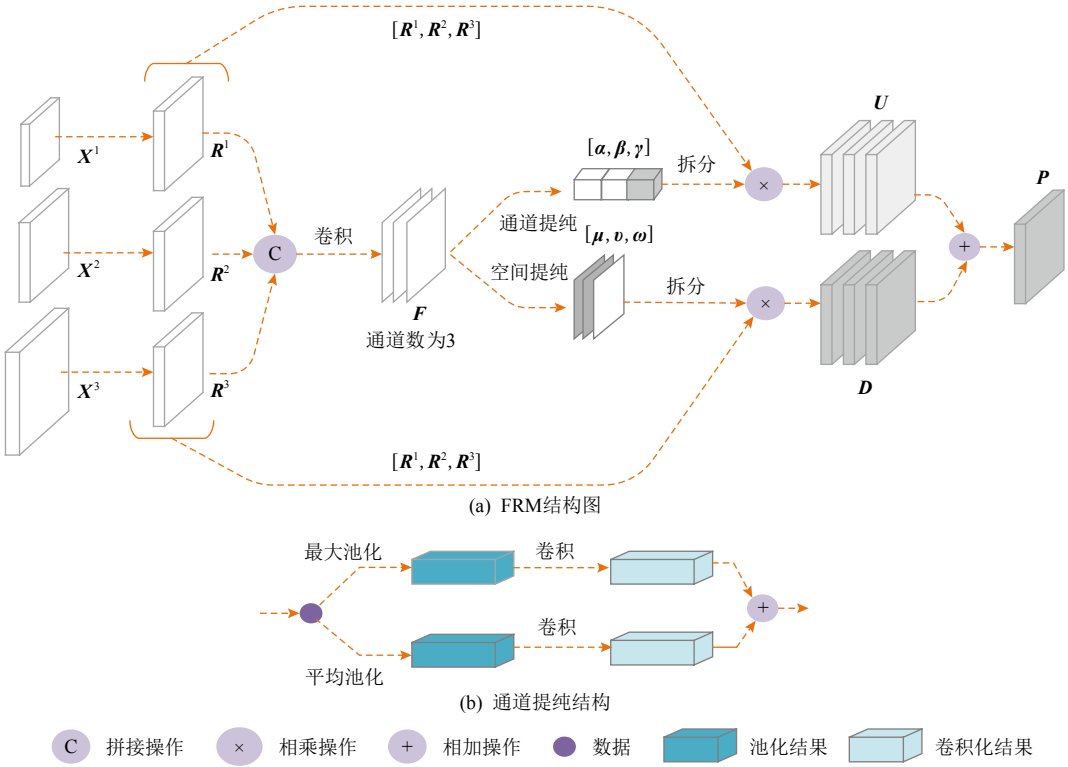
$$\boldsymbol{U}=\boldsymbol{\alpha}\times RS(\boldsymbol{X}^1)+\boldsymbol{\beta}\times\boldsymbol{X}^2+\boldsymbol{\gamma}\times RS(\boldsymbol{X}^3). \quad (1)$$



(a) 输入图像 (b) CAM输入特征图 (c) CAM输出特征图

Fig. 3 Context information augmentation effect diagrams

图 3 上下文信息增强效果图



(a) FRM结构图

(b) 通道提纯结构

Fig. 4 The structure of FRM

图 4 FRM 结构

其中 RS 表示 resize 函数，在式(1)中将 $\boldsymbol{X}^1$ 和 $\boldsymbol{X}^3$ 特征缩放到和 $\boldsymbol{X}^2$ 同一尺度. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ 为通道自适应权重，其尺度为 1×1×1. 经过归一化的 $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ 代表 3 个输入的相对权重，这 3 个值越大表示具有更大的响应，将它们与输入相乘，响应大的输入将被放大，响应小的输入将被抑制，以此将更加有用的信息增强而抑制不重要的噪声. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ 可表示为

$$[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}] = sigmoid\,[AvgPool(\boldsymbol{F}) + MaxPool(\boldsymbol{F})]. \quad (2)$$

其中 $\boldsymbol{F}$ 为图 4(a)中标识的特征图，$AvgPool$ 和 $MaxPool$ 分别为平均池化和最大池化操作.

空间提纯模块利用 softmax 函数将特征图在空间上归一化，得到特征图中某点关于其他所有位置的相对权重，然后将其与输入分别相乘. 其输出可表示为

$$\boldsymbol{D} = \boldsymbol{\mu}_{(x,y)} \times RS\left(\boldsymbol{X}^1_{(x,y)}\right) + \boldsymbol{\nu}_{(x,y)} \times \boldsymbol{X}^2_{(x,y)} + \boldsymbol{\omega}_{(x,y)} \times RS\left(\boldsymbol{X}^3_{(x,y)}\right). \quad (3)$$

(x, y) 表示特征图的空间坐标. $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, $\boldsymbol{\omega}$ 为空间自适应权重，目标区域的响应较大，将会获得更大的权重，反之背景区域获得的权重较小. $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, $\boldsymbol{\omega}$ 与输入具有相同的空间大小，因此将它们和输入直接相乘可以达到将目标特征放大和背景噪声抑制的目的. $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, $\boldsymbol{\omega}$ 可由式(4)表示.

$$[\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\omega}] = \mathrm{softmax}(\boldsymbol{F}). \quad (4)$$

softmax 函数用于归一化特征参数以提高模型的泛化能力.那么此模块的总输出为

$$\boldsymbol{P} = \boldsymbol{U} + \boldsymbol{D}. \quad (5)$$

FPN 所有层的特征都在自适应权重的引导下融合，融合的结果作为整个网络的输出.

为更加直观地说明特征提纯模块的作用,图 5 展示了部分可视化的特征图. 由于小目标的检测由 FPN 的最低层主导,因此我们仅可视化了最低层的特征. 图 5 中 $\boldsymbol{F}_3$, $\boldsymbol{L}_3$, $\boldsymbol{P}_3$ 分别对应图 1 中的标签 $\boldsymbol{F}_3$, $\boldsymbol{L}_3$, $\boldsymbol{P}_3$.

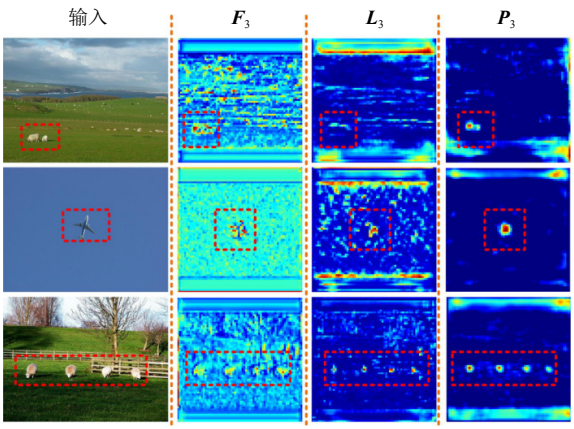


Fig. 5 Visualization results of FRM

图 5 FRM 可视化结果

由图 5 可知, $\boldsymbol{F}_3$ 特征可大致定位目标位置,但是包含较多背景噪声,具有较大误检的可能. $\boldsymbol{L}_3$ 相比于 $\boldsymbol{F}_3$,背景信息明显减少,这是 FPN 融合高层信息的结果. 高层信息更加关注于物体的抽象信息而不关注背景信息,因此背景信息会被中和. 但是由于特征的细腻度不同,引入了冲突信息,使得目标的响应被削弱.而 $\boldsymbol{P}_3$ 的目标特征被强化,并且目标和背景之间的边界更加明显. 由可视化分析可知,本文提出的 FRM 可减少干扰小目标的冲突信息,提高判别性,以此提高小目标的检测率.

2.2 copy-reduce-paste 数据增强

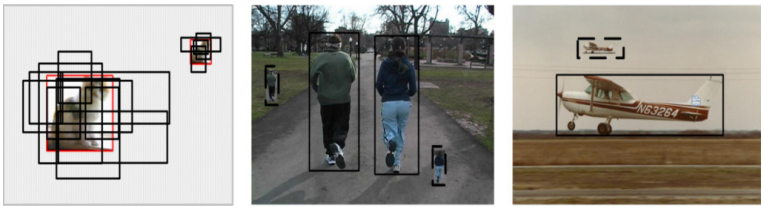
当前主流的公开数据集中,小目标的数量或包含小目标的图片数量远远小于较大目标的,如 VOC 数据集,统计情况如表 2 所示. 同时,如图 6(a)所示,小目标产生的正样本数量远远小于较大目标的,因而小目标对损失的贡献率远远小于较大目标的,使得网络收敛的方向不断向较大目标倾斜.

Table 2 Statistical Results of Target Size on VOC Datebase

表 2 VOC 数据集目标尺寸统计结果 %

统计值	小目标	中目标	大目标
目标数量占比	10.0	16.6	73.4
图片数量占比	8.2	16.2	75.6

为了缓解这个问题,我们在训练过程中复制、缩小、粘贴图像中的目标,以增加小目标产生的正样本数量以及对损失的贡献值,使得训练更为平衡. 数据增强效果如图 6(b)和图 6(c)所示.



(a) 目标的锚框 (b) 增强结果1 (c) 增强结果2

Fig. 6 Data augmentation examples

图 6 数据增强示例

图 6(b), 图 6(c)是粘贴 1 次的结果示例,实线框是原有的目标,虚线框为粘贴的目标. 首先复制大目标图像块,然后对图像块进行缩小,最后粘贴到原图的不同位置.我们提出的数据增强方法并没有直接复制小目标图像区域粘贴到不同位置,这是考虑到数据集中含有小目标的图像数量较少,如果仅仅复制粘贴小目标,在很多批次中小目标对损失的贡献仍然很低. 此外,我们研究了粘贴次数对小目标检测性能的影响,实验结果如表 3 所示.

从表 3 中可知,随着粘贴次数的增加,小目标的检测率逐渐减小,甚至会造成低于基线模型的情况.这可能是由于随着粘贴次数的增加,逐渐破坏了原始数据的分布,使得在测试集的表现较差. 在粘贴 1 次时, AP_s 提高了 2.5%, AR_s 提高了 1.9%,同时中目标的检测率也略有提升,结果表明粘贴 1 个目标是最佳的设定.

Table 3 Ablation Experimental Results of Data Augmentation

表 3 数据增强消融实验结果 %

粘贴次数	$AP(IOU=0.5)$		$AR(IOU=0.5)$	
	AP_s	AP_m	AR_s	AR_m
0(基线模型)	34.8	60.5	57.9	78.7
1	37.3	62.7	59.8	80.9
2	36.8	62.6	58.0	81.0
3	33.2	59.7	58.0	79.8

注:基线模型为 YOLOV$_3$, IOU 为交并比.AP_s, AP_m 分别指小目标、中目标的平均精度; AR_s, AR_m 分别指小目标、中目标的平均召回率.

3 实　　验

3.1 训练设置

本文实验在 VOC 和 TinyPerson 两种数据集[18]上进行. VOC 有 22 136 张训练图像和 4 952 张测试图像, 共 20 个类别. TinyPerson 数据集包含 2 个类别, 798 张训练图片和 816 张测试图片, 其场景多为远距离大背景下的图像, 所标注目标的平均大小为 18 像素, 是一个真正意义上的小目标数据集.

本文所使用的评估指标为:

精度(precision, P), 用来检测结果中相关类别占总结果的比重;

召回率(recall, R), 用来检测结果中相关类别占总类别的比重. 由 P-R 曲线可计算所有大、中、小目标平均检测精度的均值(mAP):

$$mAP = \frac{1}{k}\sum_{n=1}^{N} P(n)\times\Delta r(n). \tag{6}$$

其中 N 为测试集总数, $P(n)$ 表示 n 张图像的精确度, $\Delta r(n)$ 表示从 $n-1$ 增加到 n 时召回率的变化量, k 为类别数. 同时, 使用下标 s, m, l 分别表示在小尺度、中尺度和大尺度目标上的性能. 本文所有的实验在同样的软件和硬件条件下进行(pytorch[19]框架, Intel Core i7-5820k CPU@3.30 GHz 处理器, 16 GB 内存, GeForce GTX TITAN 显卡).

图 7 为训练时的损失变化曲线, 我们采用 SGD 优化器训练 50 轮次(前 2 个轮次预热), 批次设定为 8, 学习率初始值为 0.000 1, 训练的损失值平滑下降. 部分特征可视化结果如图 8 所示.

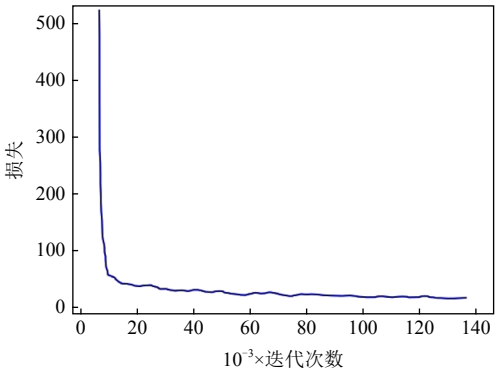
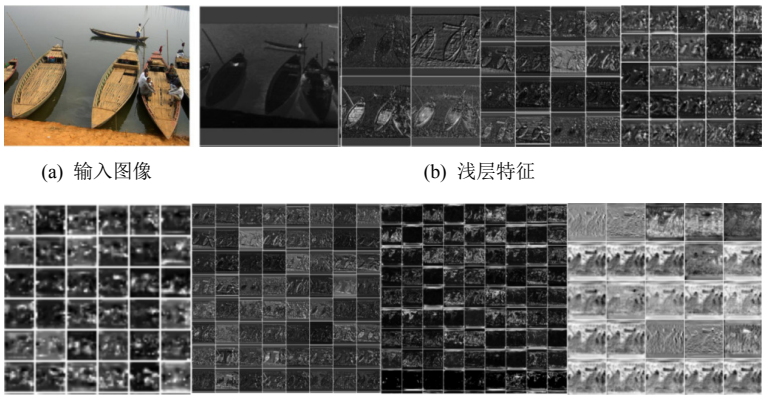


Fig. 7 The curve of loss

图 7 损失曲线



(a) 输入图像　(b) 浅层特征

(c) 深层特征

Fig. 8 Visualization results of feature maps in training

图 8 训练特征图可视化效果

如图 8 所示, 图 8(b)为浅层特征, 网络更关注物体的纹理信息. 图 8(c)为深层特征, 图像的信息逐渐抽象, 网络更关注物体的高层语义信息.

3.2 实验结果

为验证本文算法在小目标检测上的有效性, 本文在 TinyPerson 和 VOC 数据集上分别进行了实验.

本文复现了 4 种算法在 TinyPerson 数据集上的检测结果, 由于该数据集几乎全是小目标, 因此只进行 AP_s 指标的对比, 对比结果如表 4 所示.

由表 4 可知, 本文算法在该数据集上的 AP_s 达到

Table 4 Detection Results on TinyPerson Dataset

表 4 TinyPerson 数据集上的检测结果 %

算法	主干网络	AP_s
MaskR-CNN[20]	ResNet50	42.5
AL-MDN[21]	AGG16	34.1
DSFD[22]	ResNet152	51.6
YOLOV5[23]	CSPDarkNet	54.3
本文算法	Darknet53	55.1

注: AP_s 指小目标的平均精度.

55.1%. 相比 YOLOV5 和 DSFD 算法, 本文算法分别有 0.8% 和 3.5% 的提升, 而相比于 AL-MDN 和 MaskR-CNN 则分别高出 21% 和 12.6%.

本文复现了 3 种较为前沿的目标检测算法在 VOC 上的结果, 并且比较这些算法在小目标、中目标上的 *AP* 和 *AR*, 实验结果如表 5 所示:

Table 5 Results of Small Targets Detection on VOC Dataset

表 5 VOC 数据集上的小目标检测结果 %

算法	$AP(IOU \in [0.5, 0.95])$		$AR(IOU \in [0.5, 0.95])$	
	AP_s	AP_m	AR_s	AR_m
RefineDet[10]	11.6	34.9	20.2	39.9
CenterNet[24]	9.2	31.3	17.4	43.0
YOLOV4[16]	13.0	34.5	18.1	42.8
本文算法	16.9	33.4	29.4	45.8

注: *IOU* 为交并比.AP_s, AP_m 分别指小目标、中目标的平均精度; AR_s, AR_m 分别指小目标、中目标的平均召回率.

由表 5 可知, 本文算法相比于 YOLOV4, AP_s 高 3.9%, AR_s 高 11.3%; 相比于 RefineDet, AP_s 高 5.3%, AR_s 高 9.2%; 而相比于 CenterNet, 本文算法的 AP_s 和 AR_s 分别具有 7.7% 和 12.0% 的优势. 不难发现, 本文算法在小目标的召回率上具有较大优势, 说明本文算法具有较强的小目标查找能力.

将本文算法和近几年的一阶段算法和二阶段算法在 VOC 数据集上的 *mAP* 进行对比, 对比结果如表 6 所示.

由表 6 可知, 与一阶段算法相比, 本文算法比 PFPNet 的 *mAP* 高 1.3%, 具有最好的表现. 与二阶段算法相比, 本文算法优于大部分的二阶段算法, 但比 IPG-Net 的 *mAP* 低 1.2%, 这主要是由于本文算法的主干网络性能较差以及输入图像大小较小. 如果本文采用多尺度测试的方法, 则在 VOC 数据集上的检测率可达到 85.1%, 高于所有的对比算法.

Table 6 Experimental Results on VOC Dataset (*IOU*=0.5)

表 6 VOC 数据集上的实验结果 (*IOU*=0.5)

类型	算法	主干网络	输入尺寸	*mAP*/%
二阶段	Faster R-CNN[4]	ResNet101	1 000×600	76.4
	R-FCN[3]	ResNet101	1 000×600	80.5
	HyperNet[25]	VGG16	1 000×600	76.3
	CoupleNet[26]	ResNet101	1 000×600	82.7
	Reconfig[27]	ResNet101	1 000×600	82.4
	IPG-Net[28]	IPGNet101	1 000×600	84.8
一阶段	SSD[2]	VGG16	512×512	79.8
	RefineDet[10]	VGG16	512×512	81.8
	RFBNet[29]	VGG16	512×512	82.2
	ScratchDet[30]	RestNet34	320×320	80.4
	PFPNet[31]	VGG16	512×512	82.3
	本文算法	Darknet53	448×448	83.6
	本文算法+	Darknet53	448×448	85.1

注: "+"表示多尺度测试.

本文算法对小目标的检测具有较大优势, 不管是总体检测效果还是小目标的检测率、召回率都表现良好, 优于大多数检测算法.

3.3 消融实验

本文以消融实验验证每个模块的贡献.通过逐个添加数据增强方法、CAM 和 FRM 到基线模型 YOLOV3 中, 得出实验结果如表 7 所示:

Table 7 Ablation Experimental Results

表 7 消融实验结果

基线模型	增强	CAM	FRM	*AP*/% (*IOU* = 0.5)			*AR*/% (*IOU* = 0.5)		
				AP_s	AP_m	AP_l	AR_s	AR_m	AR_l
√				34.8	60.5	83.6	57.9	78.7	92.8
√	√			37.3	62.7	83.4	59.8	80.9	93.0
√		√		36.6	61.0	84.2	59.8	79.5	93.1
√			√	37.6	62.1	83.9	59.0	79.1	92.6
√	√	√	√	40.2	64.1	84.6	64.8	81.0	93.9

注: √表示包含该模块, *IOU*为交并比.AP_s, AP_m, AP_l分别指小目标、中目标和大目标的平均精度; AR_s, AR_m, AR_l分别指小目标、中目标和大目标的平均召回率.

总体来说, 本文提出的算法可显著提高目标检测率, 尤其是小目标和中等目标的检测率, 这也符合本文算法的初衷. 如表 7 所示, AP_s 提升 5.4%, AP_m 提升 3.6%, 而 AP_l 提升 1.0%. 同时对于不同尺度目标的召回率也有不同程度的提升. 具体来说, AR_s 提升 6.9%, AR_m 提升 1.3%, AR_l 提升 1.1%.

copy-reduce-paste 数据增强方法将 AP_s 和 AP_m 分别提高 2.5% 和 2.2%.而 AP_l 略有下降. 由此可知, 该方法可有效提高小目标检测率.

CAM 分别提高小目标的 AP_s 和 AR_s 1.8% 和 0.6%. 证实了补充上下文信息对于小目标检测的重要性.

FRM 将 AP_s 和 AP_m 分别提高 2.8% 和 1.6%, 而 AP_l 基本持平. 由此可见, FRM 可滤除特征的冲突信息, 提高较小目标特征的判别性.

4 总 结

小目标特征模糊, 能够提取的特征少, 是目标检测领域的难点. 为了解决小目标特征消散的问题, 本

文引入 CAM, 通过不同空洞卷积率的空洞卷积提取上下文信息, 以补充小目标的上下文信息. 由于小目标容易淹没在冲突信息中, 本文提出 FRM, 该模块结合通道和空间自适应融合来抑制冲突信息, 提高特征的判别性. 同时, 提出一种 copy-reduce-paste 的小目标增强方法来提高小目标对损失函数的贡献, 使得训练更加平衡.通过实验结果可知, 本文提出的小目标检测网络在 TinyPerson 和 VOC 数据集上均表现良好, 优于大多数的目标检测算法.

致谢 感谢武汉大学超级计算中心对本文的数值计算提供的支持.

作者贡献声明: 肖进胜和赵陶设计网络并实践; 肖进胜和周剑负责论文撰写; 乐秋平和杨力衡提供数据支持和文章的润色

参 考 文 献

[1] Joseph R, Ali F. YOLOV3: An incremental improvement[J]. arXiv preprint, arXiv: 1804.02767, 2018

[2] Liu Wei, Anguelov D, Erhan D, et al. SSD: Single shot multi-box detector[C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 21−37

[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C] //Proc of the 27th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 580−587

[4] Ren Shaoqing, He Kaiming, Girshick R. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137−1149

[5] Kisantal M, Wojna Z, Murawski J, et al. Augmentation for small object detection[J]. arXiv preprint, arXiv: 1902.07296, 2019

[6] Huang Jipeng, Shi Yinghuan, Gao Yang. Multi-scale faster-rcnn algorithm for small object detection[J]. Journal of Computer Research and Development, 2019, 56(2): 319−327 (in Chinese)
(黄继鹏, 史颖欢, 高阳. 面向小目标的多尺度Faster-RCNN检测算法[J]. 计算机研究与发展, 2019, 56(2): 319−327)

[7] Lin Tsungyi, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 936−944

[8] Deng Chunfang, Wang Mengmeng, Liu Liang, et al. Extended feature pyramid network for small object detection[J]. arXiv preprint, arXiv: 2003.07021, 2020

[9] Xiao Jinsheng, Rao Tianyu, Jia Qian, et al. Interpolation algorithm based on improved adaptive shock filter in image super-resolution[J]. Chinese Journal of Computers, 2015, 38(6): 1131−1139 (in Chinese)
(肖进胜, 饶天宇, 贾茜, 等. 改进的自适应冲击滤波图像超分辨率插值算法[J]. 计算机学报, 2015, 38(6): 1131−1139)

[10] Zhang Shifeng, Wen Longyin, Bian Xiao, et al. Single-shot refinement neural network for object detection[C] //Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4203−4212

[11] Jiang Hongyim Wang Yongjuan, Kang Jingyu. A survey of object detection models and its optimization method[J]. Acta Automatica Sinica, 2021, 47(6): 1232−1255 (in Chinese)
(蒋弘毅, 王永娟, 康锦煜. 目标检测模型及其优化方法综述[J]. 自动化学报, 2021, 47(6): 1232−1255)

[12] Liu Shu, Qi Lu, Qin Haifang, et al. Path aggregation network for instance segmentation[C] //Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8759−8768

[13] Ghiasi G, Lin Tsungyi, Le Q V. NAS-FPN: Learning scalable feature pyramid architecture for object detection[C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 7029−7038

[14] Tan Mingxing, Pang Ruoming, Le Q V. EfficientDet: Scalable and efficient object detection[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10778−10787

[15] Chen Yukang, Zhang Peizhen, Li Zeming, et al. Stitcher: Feedback-driven data provider for object detection[J]. arXiv preprint, arXiv: 2004.12432, 2020

[16] Bochkovskiy A, Wang Chienyao, Mark Liao H Y. YOLOV4: Optimal speed and accuracy of object detection[J]. arXiv preprint, arXiv: 2004.10934, 2020

[17] Yu F, Vladlen K. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint, arXiv: 1511.07122, 2015

[18] Yu Xuehui, Gong Yuqi, Jiang Nan, et al. Scale match for tiny person detection[C] //Proc of 2020 IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2020: 1257−1265

[19] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library[C] //Proc of the 33rd Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2019: 8026−8037

[20] He Kaiming, Gkioxari G, Dollárm P, et al. Mask R-CNN[C] //Proc of the 16th IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 2980−2988

[21] Choi J, Elezi I, Lee H J, et al. Active learning for deep object detection via probabilistic modeling[J]. arXiv preprint, arXiv: 2103.16130, 2021

[22] Li Jian, Wang Yabiao, Wang Changan, et al. DSFD: Dual shot face detector[C] //Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 5060−5069

[23] Jocher G. YOLOV5[CP/OL]. (2021-06-09) [2021-06-09]. https://github.com/ultralytics/yolov5

[24] Duan Kaiwen, Bai Song, Xie Lingxi, et al. CenterNet: Keypoint triplets for object detection[C] //Proc of the 17th IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019: 6568−6577

[25] Kong Tao, Yao Anbang, Chen Yurong, et al. HyperNet: Towards accurate region proposal generation and joint object detection[C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 845−853

[26] Zhu Yousong, Zhao Chaoyang, Wang Jinqiao, et al. CoupleNet:

Coupling global structure with local parts for object detection[C] //Proc of the 15th IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 4146−4154

[27] Kong Tao, Sun Fuchun, Huang Wenbing, et al. Deep feature pyramid reconfiguration for object detection[C] //Proc of the 15th European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 169−185

[28] Liu Ziming, Gao Guangyu, Sun Lin, et al. IPG-net: Image pyramid guidance network for small object detection[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway, NJ: IEEE, 2020: 4422−4430

[29] Liu Songtao, Huang Di, Wang Yunhong. Receptive field block net for accurate and fast object detection[C] //Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 404−419

[30] Zhu Rui, Zhang Shifeng, Wang Xiaobo. ScratchDet: Training single-shot object detectors from scratch[C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 2263−2272

[31] Kim S W, Kook H K, Sun J Y, et al. Parallel feature pyramid network for object detection[C] //Proc of the 15th European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 234−250

Xiao Jinsheng, born in 1975. PhD, associate professor. Member of CCF. His main research interests include image processing and computer vision.

肖进胜, 1975 年生.博士，副教授.CCF 会员.主要研究方向为图像处理和计算机视觉.

Zhao Tao, born in 1996. Master. His main research interests include image processing and computer vision.

赵　陶, 1996 年生.硕士.主要研究方向为图像处理和计算机视觉

Zhou Jian, born in 1989. PhD, postdoc. His main research interests include portable vision metrology and high definition map.

周　剑, 1989 年生.博士，博士后.主要研究方向为移动视觉测量和高精度地图.

Le Qiuping, born in 1997. Master. Her main research interests include image processing and computer vision.

乐秋平, 1997 年生.硕士.主要研究方向为图像处理和计算机视觉.

Yang Liheng, born in 1995. Master candidate. His main research interests include image processing and computer vision.

杨力衡, 1995 年生.硕士研究生.主要研究方向为图像处理和计算机视觉.