文章编号:1001-9014(2025)01-0118-11 DOI:10. 11972/j. issn. 1001-9014. 2025. 01. 016

基于红外双目视觉的同步目标检测与匹配网络

曾长紊[1,2,3], 杨支羽[1,2,3], 代作晓[1], 顾明剑[1,3*]
(1. 中国科学院上海技术物理研究所,上海 200083;
2. 中国科学院大学,北京 100049;
3. 上海市空间光电感知融合创新中心,上海 200083)

摘要:特殊环境下道路目标的三维感知对汽车的全天时、全气候自动驾驶具有重要意义,红外双目视觉模仿人眼实现微光/无光等特殊环境下目标的立体感知,目标检测与匹配是双目视觉立体感知的关键技术。针对当前分步实现目标检测与目标匹配的过程冗杂问题,提出了一个可以同步检测与匹配红外目标的深度学习网络——SODMNet(Synchronous Object Detection and Matching Network)。SODMNet 融合了目标检测网络和目标匹配模块,以目标检测网络为主要架构,取其分类与回归分支深层特征为目标匹配模块的输入,与特征图相对位置编码拼接后通过卷积网络输出左右图像特征描述子,根据特征描述子之间的欧氏距离得到目标匹配结果,实现双目视觉目标检测与匹配。与此同时,采集并制作了一个包含人、车辆等标注目标的夜间红外双目数据集。实验结果表明,SODMNet 在该红外双目数据集上的目标检测精度 mAP(Mean Average Precision)提升 84. 9% 以上,同时目标匹配精度 AP(Average Precision)达到 0. 5777。结果证明,SODMNet 能够高精度地同步实现红外双目目标检测与匹配。

关 键 词:红外双目视觉;目标检测;目标匹配;卷积网络
中图分类号:TP722.5 **文献标识码**: A

Synchronous object detection and matching network based on infrared binocular vision

ZENG Chang-Wen[1,2,3], YANG Zhi-Yu[1,2,3], DAI Zuo-Xiao[1], GU Ming-Jian[1,3*]
(1. Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Shanghai Integrated Innovation Center for Space Optoelectronic Perception, Shanghai 200083, China)

Abstract: The three-dimensional perception of road objects in challenging environments is crucial for the development of autonomous vehicles operating in all conditions, at all hours. Infrared binocular vision mimics the human binocular system, facilitating stereoscopic perception of objects in challenging conditions such as dim or zero-light environments. The core technology for stereoscopic perception in binocular vision systems is object detection and matching. To streamline the complex sequence of object detection and matching procedures, a synchronous object detection and matching network (SODMNet) is proposed, which can perform synchronous detection and matching of infrared objects. SODMNet innovatively combines an object detection network with an object matching module, leveraging the deep features from the classification and regression branches as inputs for the object matching module. By concatenating these features with relative position encoding from the feature maps and processing the concatenated features through a convolutional network, the network generates feature descriptors for the left and right images. Object matching is then achieved by calculating the Euclidean distances between these descriptors, thus facilitating synchronous object detection and matching in binocular vision. In addition, a novel nighttime infrared binocular dataset, annotated with targets such as pedestrians and vehicles, is created to support the development and evaluation of the proposed network. Experimental results indicate that SODMNet achieves a significant improvement of more than 84. 9% in object detection mean average precision (mAP) on this dataset, with an object matching average precision (AP) of 0. 5777. These results demonstrate that SODMNet is capable of high-precision, synchronized object detection and matching in infrared binocular vision,

收稿日期:2024- 04- 17,**修回日期**:2024- 09- 27 **Received date**:2024- 04- 17,**Revised date**:2024-09-27
基金项目:国家重点研发计划(2023YFB3905400)
Foundation item:Supported by the National Key Research and Development Program of China (2023YFB3905400)
作者简介(Biography):曾长紊(1993—),男,江西吉安人,博士研究生,主要研究领域为自动驾驶感知. E-mail: johnny_zeng@126. com
*通讯作者(Corresponding author): E-mail: gumingj@sina. com

marking a significant advancement in the field.

Key words: infrared binocular vision, object detection, object matching, convolutional network

引言

可见光相机因其出色的成像能力,成为目标检测[1]领域最常见、最受欢迎的传感器,也是自动驾驶汽车[2]最主要的传感器之一。作为当前的热门研究方向,单目相机三维目标检测[3]能够给自动驾驶汽车或移动机器人[4]提供重要的三维目标信息。但是在一些特殊环境条件下,可见光相机的探测能力被严重限制,例如夜间微光/无光等环境。有不少研究者对低光照条件下的可见光图像增强进行了研究[5],包括利用传统方法和基于深度学习的方法。这些方法可以有效改善图像质量,但是对无光等极端环境下的可见光图像却也无能为力。红外相机因其独特的热成像优势,不受环境光照影响,在无光等特殊环境下依旧能清晰成像。人和车辆是道路环境中最主要的动态目标,也是自动驾驶汽车最关心的目标。特别地,人和移动的车辆都是比较显著的热源,在红外相机中成像明显,这有助于对特殊道路环境下人和车辆的准确检测。因此,红外相机既可以在白天作为辅助感知,也可以在夜间作为可见光相机等传感器的互补,是自动驾驶汽车实现全天候目标检测的重要传感器之一。而红外双目相机可以在目标检测的基础上,通过三角测距原理实现目标的立体感知[6],为自动驾驶汽车提供更全面的道路环境信息。

目标检测与目标匹配[7]是实现红外双目视觉立体感知的关键技术,也是计算机视觉领域重要的研究方向。目标检测的目的是在图像中识别和定位特定目标,目标匹配的目的是在不同图像或图像序列之间识别和对应相同的目标或特征。早期的方法依赖于手工设计的特征提取器。目标检测方法主要有 Haar[8]、FAST(Features from accelerated segment test)[9]、HOG(Histogram of oriented gradients)[10]、DPM(Deformable Part Model)[11]等,这些方法往往需要复杂的机器学习分类器,如支持向量机(Support Vector Machine,SVM)[12]或 AdaBoost[13]。目标匹配通常基于特征匹配实现,传统特征匹配方法包括 SIFT(Scale-invariant feature transform)[14]、SURF(Speeded Up Robust Features)[15]、ORB(Oriented FAST and Rotated BRIEF)[16]等。

基于 CNN(Convolutional Neural Networks)的深度学习因其对空间特征的强大提取能力成为目标检测与匹配的主流方法之一,内部稀疏连接和参数共享使其对比基于 Transformer[17]的网络模型需要更少的计算成本。代表性的深度学习目标检测模型包括 R-CNN(Region CNN)系列[18]、YOLO(You Only Look Once)系列[19]、SSD(Single Shot MultiBox Detector)[20]、FCOS(Fully Convolutional One-Stage Object Detection)[21]等,这些深度学习方法在红外图像目标检测领域也取得不错的成就。Krišto 等[22]比较了几种标准的最先进的目标检测器(如 Faster R-CNN[23]、SSD、Cascade R-CNN[24]、FCOS 和 YOLOv3[25])在热图像数据集上的性能,在性能相当的前提下,FCOS 和 YOLOv3 是其中最快的。Yao 等[26]在标准的 FCOS 网络基础上,提出了一个轻量级网络模型,结合传统滤波方法,增强了对小红外目标的响应,同时抑制了背景响应。此外,部分目标检测网络[27, 28]针对小目标红外成像特点进行改进,实现红外小目标的有效识别。可见光图像特征匹配领域涌现了一大批诸如 Superglue[29]、LoFTR(Local Feature Transformer)[30]、CREStereo(Cascaded Recurrent Stereo matching network)[31]等优秀方法,但是这些匹配方法都是针对像素级特征提出的,对于分辨率较低的红外图像难以达到好的效果。

目标检测与目标匹配是两个独立的步骤,在目标匹配过程中缺乏信息的相互交流,深度学习网络特征提取过程中往往伴随冗余操作和冗余信息的生成。本文针对红外双目视觉提出一个同步目标检测与匹配网络(SODMNet),在目标检测的同时对目标进行实例级的匹配。本文的主要贡献包括:1)设计了一个同步目标检测与匹配的深度学习网络;2)采集并制作了一个包含 1 593 组红外双目图片的数据集,该数据集为搭载在汽车上的红外双目相机在夜间采集的城市道路数据,并对数据集中的人和车辆目标进行了标注;3)实现了红外双目图像目标高精度匹配,并大大提升了红外图像行人车辆检测精度。

本文的结构如下:第 1 节详细描述网络架构,包括各网络部分和模块,以及损失函数设计;第 2 节详细介绍了红外双目数据集及其制作过程,并在该数据集上进行不同对比实验及结果分析;第 3 节对本文工作进行总结及展望。

1 网络架构

SODMNet由目标检测网络主架构和目标匹配模块组成，如图1所示。对输入的图像进行预处理得到标准化数据，骨干网络对数据的基本特征进行提取，通过颈部网络的进一步处理和优化，头部网络根据优化特征图实现目标的识别分类以及定位，目标匹配模块（如图1紫色虚线框所示）实现左右相机图像相同目标的匹配。接下来对各部分进行详细介绍。

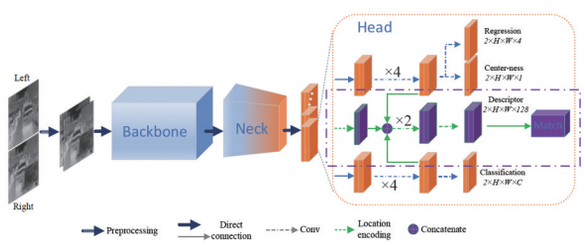

图1 SODMNet架构

Fig. 1 Architecture of the SODMNet

1.1 图像预处理

图像预处理包括数据增强[32]、归一化[33]及数据组合。数据增强可以增加网络训练的数据量，提高模型泛化能力，考虑到道路真实数据的特征，本文采用的数据增强方法主要包括图像缩放和图像水平翻转。归一化操作能够消除不同特征之间的尺度差异，使得模型能够更好地学习到数据的内在规律。最后，将处理好的左右图像进行叠加组合。

1.2 目标检测网络架构

SODMNet的目标检测主架构采用FCOS网络，它是一个高准确度且高效率的端到端全卷积网络，主要包括骨干网络、颈部网络和头部网络三个部分。

1.2.1 骨干网络

骨干网络从输入数据中提取高层次、语义丰富的特征，它决定了网络模型的基本性能。ResNet50（Residual Neural Network）[34]是本文采用的主要骨干网络，如图2所示，它分为5个阶段，第1个阶段包含卷积层、BN（Batch Normalization）层、ReLU（Rectified Linear Unit）激活函数和最大池化层，剩余4个阶段均由多个Bottleneck结构（如图3）组成，残差连接结构的使用可以很好地解决深层网络训练时梯度消失的问题。

此外，其他主流骨干网络也被采用并进行对比验证。MobileNetv2[35]利用深度可分离卷积（Depthwise separable convolution）结构替代传统卷积结构，大大减少模型参数和运算量，其中深度可分离卷积由深度卷积（Depthwise convolution）和逐点卷积（Pointwise convolution）组成，如图4。ShuffleNetv2[36]利用逐点群卷积（Pointwise group convolution）以降低卷积的计算复杂度（如图5），并通过通道混洗实现特征通道之间的信息流动。EfficientNetv2[37]通过自动化搜索网络结构的方法来提高模型的效率，结合模型复合缩放方法，综合考虑网络的深度、宽度和输入分辨率进行扩展，它的主要结构是MBConv（Mobile inverted bottleneck convolution），如图6。

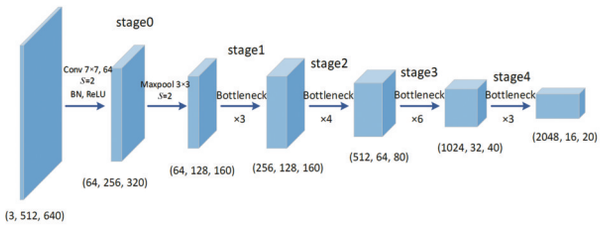

图2 ResNet50骨干网络

Fig. 2 ResNet50 backbone network

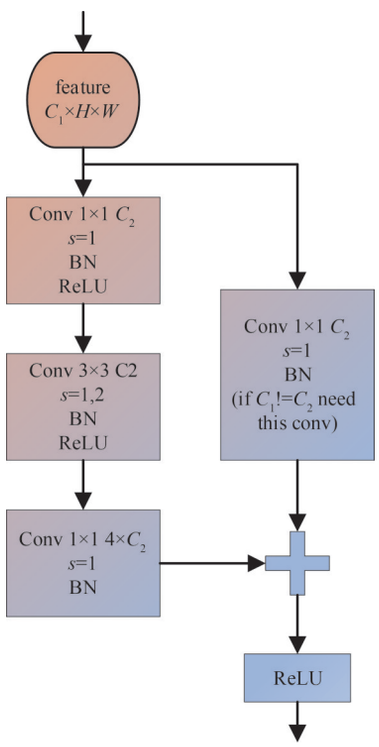

图3 Bottleneck结构

Fig. 3 Bottleneck structure

1.2.2 颈部网络

颈部网络在骨干网络提取的特征上进一步进行不同尺度特征的融合，通过上下文增强帮助网络

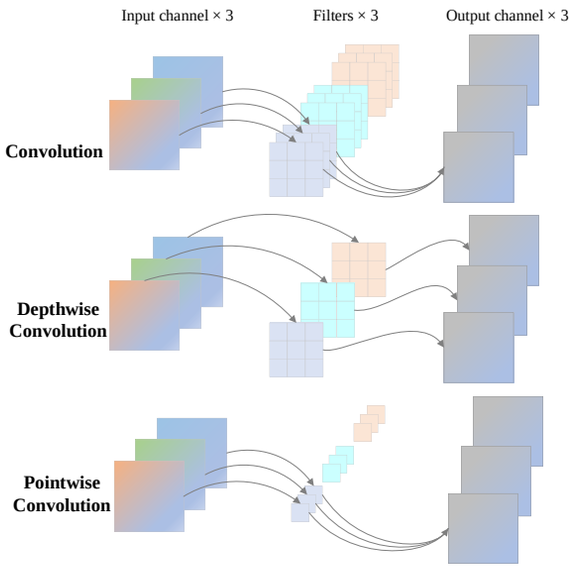

图4　深度卷积和逐点卷积与传统卷积对比

Fig. 4　Comparison of depthwise convolution and pointwise convolution with traditional convolution

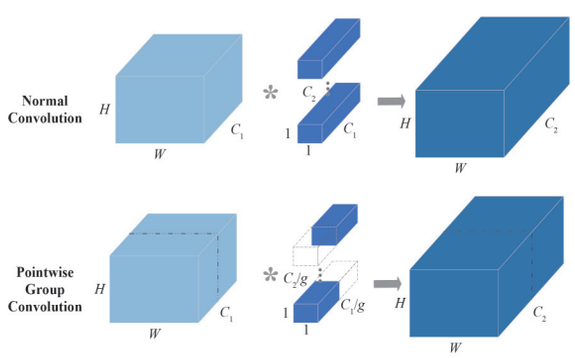

图5　逐点群卷积与传统卷积对比

Fig. 5　Comparison of pointwise group convolution with traditional convolution

感知不同大小的目标。FPN(Feature Pyramid Network)[38]是最常见的一种目标检测颈部网络,如图7,它将生成的金字塔特征图与对应的上采样特征图进行加和,融合特征层浅层的细节信息和深层的语义信息以区分不同特征的目标,融合特征层浅层低感受野信息和深层高感受野信息以区分不同尺寸的目标。

1.2.3　头部网络

头部网络是负责任务预测的输出部分。FCOS目标检测网络包括用于预测目标类别的分类头部和用于预测目标框位置的回归头部,同时在回归分支增加了一个中心度约束,将离标注框中心很远的低质量预测框筛除。分类头部和回归头部分别对颈部网络输出的特征图进行4组卷积,最后通过卷积输出分类标签和回归位置。SODMNet在此基础上增加了一个目标匹配模块,将在1.3节详细介绍。

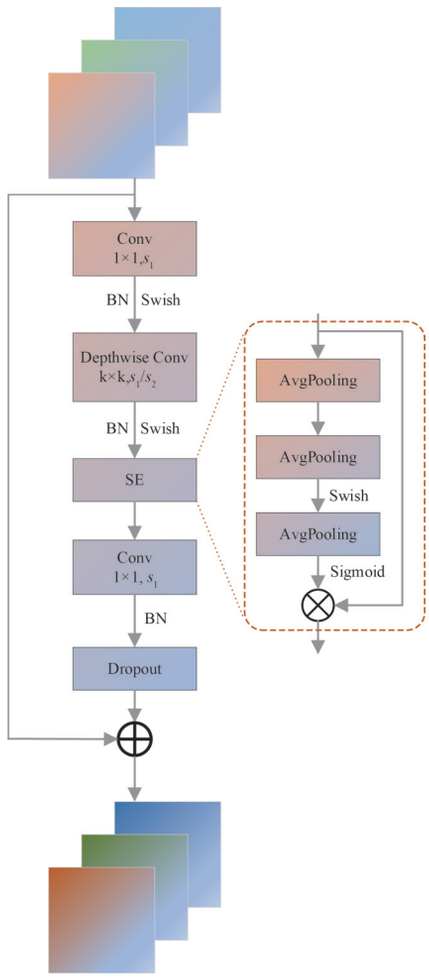

图6　MBConv结构

Fig. 6　MBConv structure

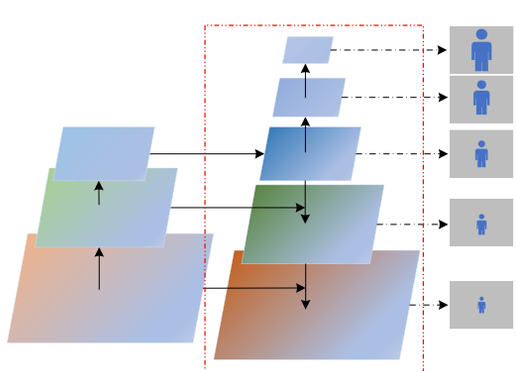

图7　FPN结构

Fig. 7　Structure of FPN

1.3　目标匹配模块

在红外双目相机的左右图像内,同一个目标通常存在相似特征。图8为红外双目相机同步采集的

道路环境图片，视场内未遮挡目标car1、car4和部分遮挡目标car3的相似特征较多，视场内遮挡目标car2、car5和视场边缘目标car6、car7存在部分相似特征。目标检测网络深层特征包含更丰富的信息，因此分别取分类和回归分支输出前一层特征图作为目标匹配模块的输入。同时，左右图像内的目标存在某种相对位置关系，因此，对特征图进行位置编码并输入模块。

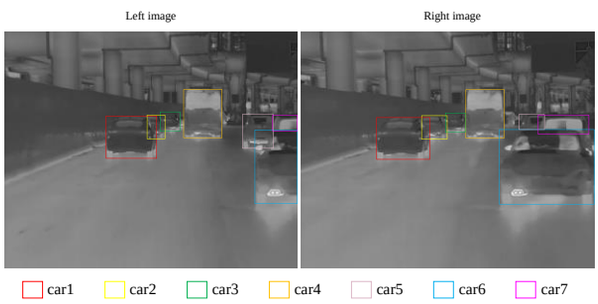

图8 红外双目标注图片

Fig. 8 Annotated infrared stereo images

目标匹配模块主要包括目标描述子生成和匹配子模块。输入经过两组卷积，每组卷积包含3×3卷积层、BN层和ReLU激活函数，最后输出特征图对应的128维描述子。匹配子模块根据左右图像输出描述子之间的欧氏距离，得到目标匹配对。平行式双目相机目标成像俯视图如图9，无论左相机与右相机基线距离远或近，同一目标在左相机成像位置比在右相机成像位置靠右。图10描述了左右特征图对应行描述子距离计算，本文红外双目相机采用平行式结构，故只需要考虑右上角黄色区域欧氏距离值。由于目标框内的特征都可以作为正样本预测目标，如图10中红色框内部分，而左右特征图中的最优预测可能不在同一行特征，甚至可能因为遮挡导致预测目标框有差异（如图8中的car5）。为了建立不同行之间的联系，增加右特征图与左特征图对应的前后行特征描述子欧氏距离进行匹配，如图11。对于右特征图首行特征，只考虑与左特征图对应行及下一行特征描述子之间的欧氏距离；对于右特征图末行特征，只考虑与左特征图对应行及上一行特征描述子之间的欧氏距离。

1.4 损失函数

损失函数是用来衡量模型对输入图像的预测值与真实值之间的接近程度，预测值与真实值越接近则损失越小，反之损失越大。然后通过反向传播

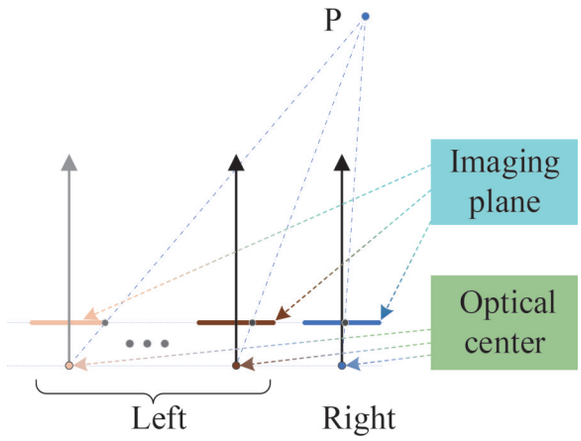

图9 平行式双目相机目标成像俯视图

Fig. 9 Top-down view of object imaging with a parallel stereo camera setup

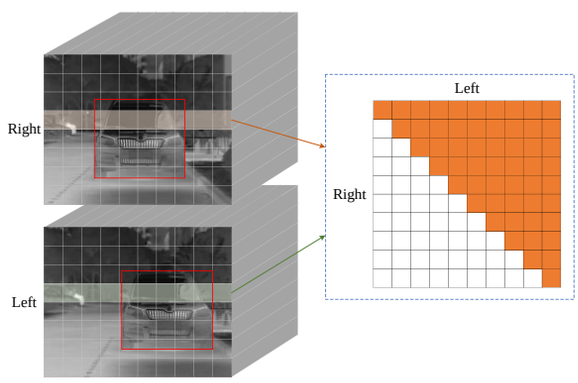

图10 特征图描述子欧氏距离计算

Fig. 10 Euclidean distance calculation for feature map descriptors

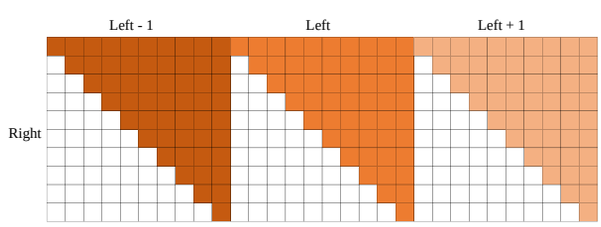

图11 右特征图与左特征图对应行及前后行描述子欧氏距离计算

Fig. 11 Euclidean distance calculation between right and left feature map descriptors for corresponding and adjacent rows

和梯度下降更新参数并重复训练，使得模型输出与样本真值之间的差距越来越小，即模型的预测能力越来越强。SODMNet的损失函数主要包括分类损失、回归损失、中心度损失和匹配损失。

1.4.1 分类损失

分类损失函数采用的是F_{loss}[39]，它能够解决端到端目标检测网络中正负样本数量极不平衡的问

题,公式如下:

$$F_{\text{loss}}=\begin{cases}-a(1-\hat{p})^{\gamma}\log p & \text{if } y=1\\ -(1-a)\hat{p}^{\gamma}\log(1-p) & \text{if } y=0\end{cases}, \quad (1)$$

其中,$\hat{p}$ 表示预测概率,y=0表示预测与真实标签不一致,y=1表示预测与真实标签一致,γ 和 α 为可调节因子,γ 可以控制难易区分样本数量失衡,α 可以抑制正负样本数量失衡。本文取 F_{loss} 参数经验值,即 γ=2,α=0.25。

1.4.2　回归损失

IoU(Intersection over Union)损失[40]作为回归损失将预测框的所有位置信息作为一个整体进行回归,实现高效、准确的定位,具有很好的尺度不变性。IoU公式如下:

$$L_{\text{IoU}}=1-\frac{S_{\text{Intersection}}}{S_{\text{Union}}}, \quad (2)$$

其中,$S_{\text{Intersection}}$ 表示目标真实框和目标预测框交集的面积,S_{Union} 表示目标真实框和目标预测框并集的面积,如图12。

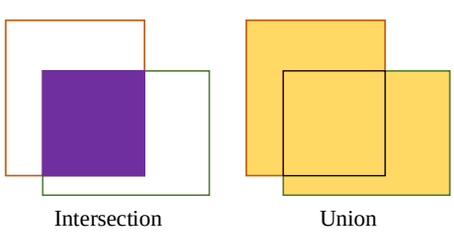

图12　目标框的交集与并集

Fig. 12　Intersection and union of the object bounding boxes

1.4.3　中心度损失

中心度损失通过二元交叉熵损失函数(Binary Cross Entropy Loss,BCE Loss)[41]计算,公式如下:

$$L_{\text{BCE}}=\frac{1}{N}\sum_{i=1}^{N}y_i\cdot\log[p(y_i)]-(1-y_i)\cdot\log[1-p(y_i)], \quad (3)$$

其中,$p(y_i)$表示预测值,y_i为真值,l_i,r_i,t_i,b_i表示预测位置中心与目标框之间的距离,如图13。

$$y_i=\sqrt{\frac{\min(l_i,r_i)}{\max(l_i,r_i)}\times\frac{\min(t_i,b_i)}{\max(t_i,b_i)}}. \quad (4)$$

1.4.4　匹配损失

目标匹配损失是评价左右图像预测目标之间的相似程度,借鉴siamese网络[42]中评价两幅图像相似程度的对比损失,改进公式如下:

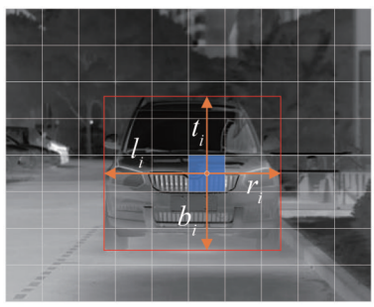

图13　预测位置中心与目标框之间的距离

Fig. 13　Distance from predicted center to object bounding box

$$L_{\text{Matching}}=\frac{1}{\sum Z_{ij}}\sum(Z_{ij}\cdot D_{ij}^2)+\frac{1}{\sum[(1-Z_{ij})\cdot M_{ij}]}\sum[(1-Z_{ij})\max(T-D_{ij},0)^2], \quad (5)$$

其中 X_i、Y_j 分别表示左右特征图不同位置描述子,Z_{ij} 表示对应位置真实匹配值,取值为1(匹配)和0(不匹配),T 为给定阈值,D_{ij} 和 M_{ij} 如下:

$$D_{ij}=\left\|X_i-Y_j\right\|_2, M_{ij}=\begin{cases}0 & \text{if } T-D_{ij}<0\\ 1 & \text{if } T-D_{ij}>0\end{cases}.$$

2　实验结果与分析

在本节中,首先介绍实验平台、数据准备和评估指标。然后,将不同骨干网络的实验结果进行比较,以证明该网络的有效性。最后展示了对网络匹配模块的输入和结构的消融研究,以验证网络的设计。

2.1　实验平台

本文的网络模型基于Pytorch框架实现,实验平台CPU为Intel i9-10900k,显卡为NVIDIA TITAN RTX,操作系统为Ubuntu 20.04。

2.2　数据准备

实验采用了公开数据集FLIR(FLIR Thermal Starter Dataset),由于目前没有公开的红外双目标注数据集,因此,基于搭建的红外双目系统采集并制作数据集。

2.2.1　FLIR数据集

FLIR数据集是通过安装在车辆上的可见光相机和红外相机采集的城市/高速道路环境数据,它一共包含人、自行车、汽车和狗4个标注类别,一共有14 452张标注的红外图片,训练集包含8 862张标注图片,验证集包含1 366张标注图片,训练集和验证集数据均从不同的短视频中采样得到。

2.2.2 红外双目数据集

红外双目数据集是通过安装在汽车车顶的红外双目相机采集的夜间市区道路环境数据，从多段视频中采样得到1 593组图片，每组包含时间同步的左右相机图像。采样数据标注类别标签包括人和车辆，其中车辆标签包括各种三轮及以上机动车辆，人标签包括行人及骑行的人。同时对每组标注图片内的目标进行匹配编号，将标号标签加在类别标签后面，如图8。

首先从红外双目相机采集的12个短视频中以每秒一帧的频率抽取视频帧组成了一个1 315组图片的数据集，图片分辨率为1 280×1 024个像素，并对每组图片中的人和车辆目标进行手动匹配标注，对于遮挡目标不进行脑补，标注框为该目标未遮挡部分的最大外接矩形框，如图8中的car2、car3、car4和car5。其中人标注框最小边长不小于12像素，车辆标注框最小边长不小于30像素。对所有图片目标数量的统计结果如图14(a)，由于选择的市区道路不够合理，且车辆行驶在道路上时红外双目相机的视场相对集中在机动车道上，采集的数据包含大量车辆目标，非机动车或行人目标相对偏少。为了使数据平衡，在一条机动车辆相对较少的道路采集了一段视频，并从中抽取278组图片，标注后加入数据集，其目标统计结果如图14(b)，最终获得红外双目数据集。

2.3 评价指标

混淆矩阵是后续评价指标的基础，它是对预测结果的一个粗略评价。如图15所示：N_{TP}表示实际为positive、模型预测为positive的样本数；N_{TN}表示实际为negtive、模型预测为negtive的样本数；N_{FP}表示实际为negtive、模型预测为positive的样本数；N_{FN}表示实际为positive、模型预测为negtive的样本数。

2.3.1 目标检测评价指标

精准率(P)和召回率(R)是目标检测领域常用的评价指标，其计算公式如式(6)所示。精准率能够反映模型对负样本的区分能力，召回率能够反映模型对正样本的识别能力。F1 score(F_1)综合了精准率和召回率，公式如式(7)，越稳健的模型F1 score越高。mAP(Mean Average Precision，P_{mAP})用于综合考虑模型在不同类别上的精准率和召回率，公式如式(8)，其中AP(Average Precision，P_{AP})是对一个类别所有预测框精准率和召回率曲线下面积的积分。

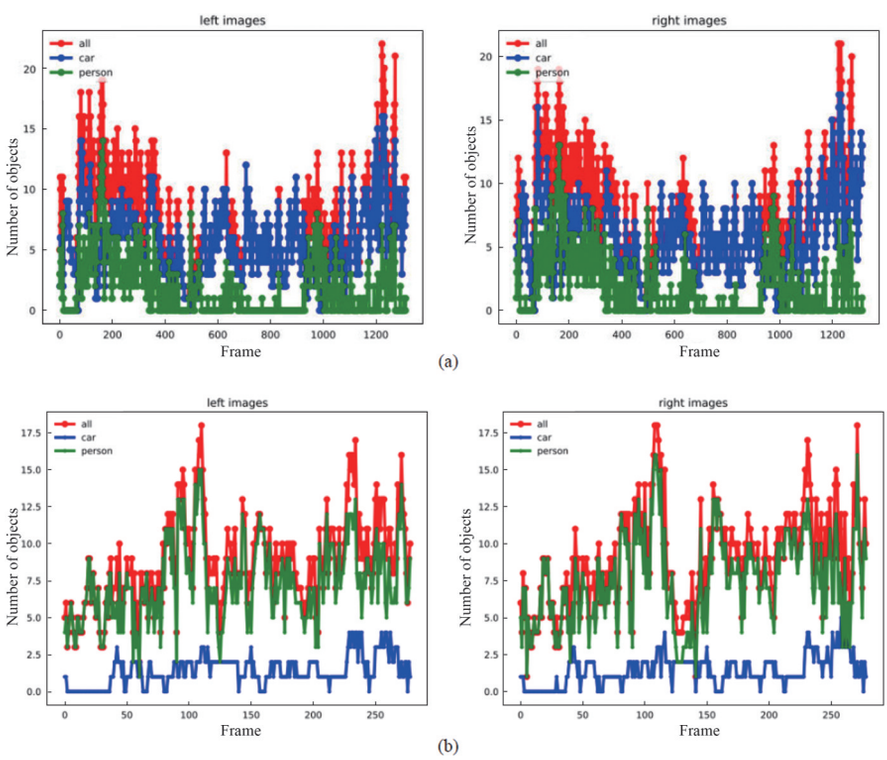

图14 数据集统计结果：(a)首次采集数据；(b)补充数据

Fig. 14 Statistical results for dataset objects: (a) initially acquired dataset; (b) supplement dataset

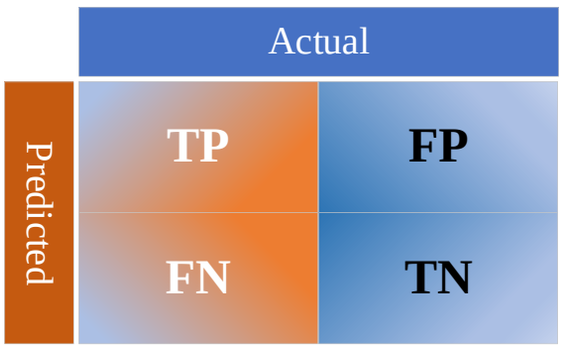

图 15　混淆矩阵

Fig. 15　Confusion matrix

$$P = N_{TP}/(N_{TP} + N_{FP}) \tag{6}$$

$$R = N_{TP}/(N_{TP} + N_{FN}) \tag{6}$$

$$F_1=2PR/(P+R) \tag{7}$$

$$P_{mAP} = \frac{\sum_{i=1}^{C} P_{AP,i}}{C} \tag{8}$$

2.3.2　目标匹配评价指标

模型的目标匹配能力评价指标也可以用精准率(M_precision)、召回率(M_recall)、M_F1 score 和 M_AP,其计算方法与目标检测评价指标精准率、召回率、F1 score、AP 相同,此时 N_{TP} 表示实际匹配预测为匹配的样本数,N_{FP} 表示实际不匹配预测为匹配的样本数,N_{FN} 表示实际匹配预测为不匹配的样本数,N_{TN} 表示实际不匹配预测为不匹配的样本数。

2.4　实验结果

为了证明 SODMNet 的有效性,本文在不同骨干网络条件下进行了对比实验,并与一些主流目标检测方法进行了对比。同时本文设计了消融实验以对比匹配模块不同输入及不同结构对模型的影响。最后分析了实验结果的可能原因。

2.4.1　不同骨干网络实验结果

为了提升 SODMNet 在小数据集训练中能得到较好的泛化能力,FCOS 目标检测网络首先针对 FLIR 数据集的人和车目标进行训练,将训练好的模型参数作为 SODMNet 中目标检测网络初始参数。随机提取红外双目数据集中 393 组图像作为测试集,剩下 1 200 组图像在每次训练时按 8∶2 随机分配为训练集和验证集。FCOS 和 SODMNet 在红外双目数据集中进行训练,改变骨干网络分得到模型稳定参数。

测试时设置 IoU 阈值为 0.5,即预测框与真实目标框的交并比大于等于 0.5 时,认为预测框是对真实目标框的正确预测,最终得到不同骨干网络下的测试集目标检测结果对比,如表 1。SODMNet 的目标匹配结果如表 2 所示。在 ResNet、MobileNet、ShuffleNet 和 EfficientNet 4 个主流骨干网络下,SODMNet 的 mAP 较 FCOS 分别提升了 106.3%、84.9%、89.0% 和 87.3%,大大提升了目标检测性能。同时 SODMNet 的 M_precision 均大于 0.84,M_AP 最高为 0.577 7。

图 16 为 SODMNet 在 ResNet 骨干网络下的预测实验结果,实验设置置信度阈值为 0.55,即置信度

表 1　不同骨干网络下 SODMNet 预测结果

Table 1　Prediction results of SODMNet under different backbone networks

Backbone		class	recall	precision	F1 score	AP	mAP
ResNet	FCOS	person	0.772 7	0.034 1	0.065 3	0.349 2	0.363 0
		car	0.688 7	0.047 8	0.089 4	0.376 7	
	SODMNet	person	0.668 3	0.916 4	0.772 9	0.656 3	0.748 9
		car	0.849 5	0.934 0	0.889 8	0.841 5	
MobileNet	FCOS	person	0.818 2	0.030 9	0.059 5	0.371 6	0.389 0
		car	0.722 6	0.051 3	0.095 8	0.406 4	
	SODMNet	person	0.632 1	0.914 4	0.747 5	0.621 3	0.719 4
		car	0.826 1	0.932 2	0.875 9	0.817 5	
ShuffleNet	FCOS	person	0.812 8	0.041 2	0.078 5	0.320 3	0.358 3
		car	0.704 0	0.055 2	0.102 3	0.396 3	
	SODMNet	person	0.593 2	0.909 3	0.718 0	0.577 1	0.677 3
		car	0.787 8	0.923 5	0.850 3	0.777 6	
EfficientNet	FCOS	person	0.778 1	0.053 2	0.099 5	0.383 3	0.404 0
		car	0.708 2	0.038 3	0.072 7	0.424 7	
	SODMNet	person	0.671 8	0.909 7	0.772 9	0.660 7	0.756 8
		car	0.859 6	0.939 4	0.897 7	0.852 9	

表 2 SODMNet目标匹配结果

Table 2 Object matching results of SODMNet

Backbone	M_recall	M_precision	M_F1 score	M_AP
ResNet	0.666 0	0.860 6	0.750 9	0.577 7
MobileNet	0.626 2	0.875 1	0.730 0	0.553 8
ShuffleNet	0.582 7	0.846 0	0.696 0	0.506 3
EfficientNet	0.645 5	0.863 8	0.738 8	0.561 6

小于阈值的目标被忽略,反之则被保留。目标框上面的标签表示目标类别、标签序号以及目标框置信度。场景①中的car1为几乎占满整个视场的近距离大目标,SODMNet也有能力检测出来。场景①中的car0和场景②中的car2匹配结果显示,对于左右相机成像存在明显差异的视场边缘目标,SODMNet依然能够根据部分相似区域将其很好地匹配。因视角差异,场景③中person0被树干部分遮挡不同区域,场景④中car7被更近距离的车辆遮挡不同部位,SODMNet能准确地进行识别和匹配。

2.4.2 与主流方法结果对比

YOLOv5和Swin Transformer分别是基于CNN和基于Transformer的主流目标检测网络,表3对比了YOLOv5、Swin Transformer与SODMNet(骨干网络为ResNet)在红外双目数据集上的目标检测性能。结果显示,虽然YOLOv5和Swin Transformer在该红外双目数据集上的目标检测性能都比较出色,mAP分别达到了0.710 2和0.721 1,但总体上还是略逊于SODMNet的0.748 9。此外,YOLOv5和Swin Transformer无法实现双目图像间的目标匹配。

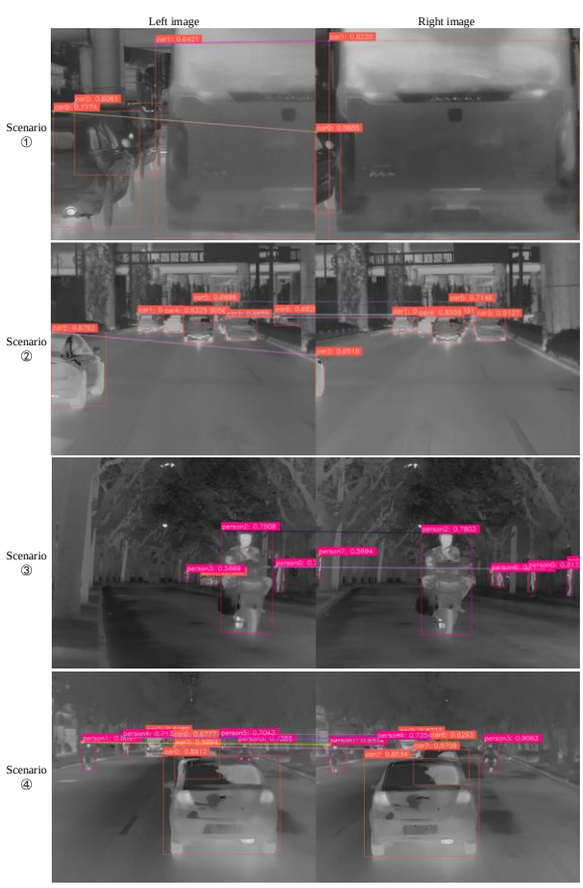

图16 不同场景SODMNet预测结果

Fig. 16 Predition results of SODMNet in different scenarios

表3 SODMNet与其他方法目标检测性能对比

Table 3 Performance comparison of object detection between SODMNet and other methods

Network	class	AP	mAP
YOLOv5	person	0.623 9	0.710 2
	car	0.796 5	
Swin Transformer	person	0.633 4	0.721 1
	car	0.808 7	
SODMNet	person	0.656 3	0.748 9
	car	0.841 5	

2.4.3 消融实验

分类信息和回归信息是影响目标匹配能力的最关键信息,在此基础上,本文设计消融实验验证特征图相对位置编码输入、浅层特征输入以及不同卷积层结构对SODMNet目标检测与匹配能力的影响。实验结果如表4所示。结果对比显示,缺少相对位置编码输入会使模型预测能力小幅下降,增加浅层特征会使模型预测能力进一步下降,增加或减少卷积层结构都无法提升模型预测能力。结果证明SODMNet的匹配模块对目标检测和匹配是最有效的。

表 4 消融实验结果

Table 4 Results of ablation experiment

Input			Layers			person_AP	car_AP	M_AP
cls&reg	location	feats	1 conv	2 conv	4 conv			
■				■		0.651 7	0.834 1	0.569 5
■	■			■		0.656 3	0.841 5	0.577 7
■		■		■		0.644 6	0.837 0	0.539 5
■	■	■		■		0.638 3	0.828 4	0.535 5
■	■		■			0.643 9	0.836 6	0.561 6
■	■				■	0.636 2	0.831 8	0.569 7

2.4.4 实验结果分析

SODMNet增加的目标匹配模块从双目相机左右图像中学习到额外的信息,损失函数增加了匹配

损失,训练阶段经过损失函数的反向传播对目标检测网络参数持续优化,这可能是其在目标检测精度上有较大提升的主要原因。同时,一个可靠的目标匹配标注的红外双目数据集对SODMNet的优越性能也至关重要。

双目图像的目标匹配主要是通过目标的特征信息进行,目标与周围目标的相对位置关系成为一个辅助判断依据,在目标特征信息缺乏时发挥重要作用。因此,相对位置编码的有无会使模型预测能力小幅变化。由于SODMNet颈部网络已经融合了浅层特征与深层特征,增加浅层特征输入可能导致匹配模块浅层特征权重占比过大,而增加或减少卷积结构可能使模型容易出现过拟合或欠拟合,这些原因都会影响模型性能。

3　结论

针对分步式实现目标检测与目标匹配效率低的问题,本文在目标检测网络的基础上增加目标匹配模块,增强模块间信息交流,提出SODMNet实现红外双目视觉同步目标检测与匹配。目标检测网络分类分支输出目标类别信息,回归分支输出目标位置信息。目标匹配模块由描述子生成子模块和匹配子模块组成,输入包括目标检测网络两个分支包含丰富信息的深层特征,同时增加特征图相对位置编码作为输入,经过卷积网络生成特征描述子,根据左右图像特征描述子之间的欧氏距离得到目标匹配结果。此外,本文采集并制作了一个夜间道路双目红外图像数据集,该数据集包含1 593组标注图片,标签类别包含人和车两类。SODMNet的目标检测网络初始参数经过FLIR公开数据集训练得到,然后对SODMNet在红外双目数据集上进行训练和优化。不同骨干网络的对比实验结果显示,SODMNet在实现准确目标匹配(AP达到0.577 7)的同时有效提升了目标检测的精度(mAP提升84.9%以上)。结果表明SODMNet可以为红外双目视觉目标立体感知提供高精度的目标检测与匹配,对实现全天候、全气候的自动驾驶提供重要基础。根据深度学习网络的通用性,SODMNet可以应用于任何双目视觉目标的同步检测与匹配。对于任务相似的目标跟踪领域视频前后帧图像之间的目标匹配也有一定适用性。由于本文红外双目数据集标注类别相对简单,训练模型无法对目标实现细分类别的检测,因此后续将深入研究SODMNet对于细分多类型目标检测与匹配的准确率,进一步分析细分目标类型对准确率的影响。

References

[1] Zou Z, Chen K, Shi Z, *et al*. Object detection in 20 years: A survey [J]. *Proceedings of the IEEE*, 2023, **111**(3): 257-276.

[2] Badue C, Guidolini R, Carneiro R V, *et al*. Self-driving cars: A survey [J]. *Expert Systems with Applications*, 2021, **165**: 113816.

[3] Wu X, Ma D, Qu X, *et al*. Depth dynamic center difference convolutions for monocular 3D object detection [J]. *Neurocomputing*, 2023, **520**: 73-81.

[4] Loganathan A, Ahmad N S. A systematic review on recent advances in autonomous mobile robot navigation [J]. *Engineering Science and Technology*, 2023, **40**: 101343.

[5] Wang W, Wu X, Yuan X, *et al*. An experiment-based review of low-light image enhancement methods [J]. *IEEE Access*, 2020, **8**: 87884-87917.

[6] Blake R, Wilson H. Binocular vision [J]. *Vision Research*, 2011, **51**(7): 754-770.

[7] Verma N K, Goyal A, Vardhan A H, et al. Object matching using speeded up robust features [C]//Intelligent and Evolutionary Systems: The 19th Asia Pacific Symposium, IES 2015, Bangkok, Thailand, November 2015, Proceedings. Springer International Publishing, 2016: 415-427.

[8] Pavani S K, Delgado D, Frangi A F. Haar-like features with optimally weighted rectangles for rapid object detection [J]. *Pattern Recognition*, 2010, **43**(1): 160-172.

[9] Li Y, Zheng W, Liu X, *et al*. Research and improvement of feature detection algorithm based on fast [J]. *Rendiconti Lincei Scienze Fisiche e Naturali*, 2021, **32**(4): 775-789.

[10] Chen P Y, Huang C C, Lien C Y, *et al*. An efficient hardware implementation of hog feature extraction for human detection [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2013, **15**(2): 656-662.

[11] Yebes J J, Bergasa L M, Arroyo R, et al. Supervised learning and evaluation of KITTI's cars detector with DPM [C]//2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE, 2014: 768-773.

[12] Wang H, Hu D. Comparison of SVM and LS-SVM for regression [C]//2005 International Conference on Neural Networks and Brain. IEEE, 2005, 1: 279-283.

[13] Hastie T, Rosset S, Zhu J, *et al*. Multi-class adaboost [J]. *Statistics and its Interface*, 2009, **2**(3): 349-360.

[14] Ng P C, Henikoff S. Sift: Predicting amino acid changes that affect protein function [J]. *Nucleic Acids Research*, 2003, **31**(13): 3812-3814.

[15] Bay H, Ess A, Tuytelaars T, *et al*. Speeded-up robust features (surf) [J]. *Computer Vision and Image Understanding*, 2008, **110**(3): 346-359.

[16] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF [C]//2011 International Conference on Computer Vision. IEEE, 2011: 2564-2571.

[17] Han K, Xiao A, Wu E, *et al*. Transformer in transformer [J]. *Advances in Neural Information Processing Systems*, 2021, **34**: 15908-15919.

[18] Bharati P, Pramanik A. Deep learning techniques—R-CNN to mask R-CNN: A survey [J]. *Computational Intel-*

ligence in Pattern Recognition: *Proceedings of CIPR* 2019, **2020**: 657-668.
[19] Cong X, Li S, Chen F, et al. A review of YOLO object detection algorithms based on deep learning[J]. *Frontiers in Computing and Intelligent Systems*, 2023, **4**(2): 17-20.
[20] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision - ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.
[21] Tian Z, Shen C, Chen H, *et al*. FCOS: A simple and strong anchor-free object detector [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **44**(4): 1922-1933.
[22] Krišto M, Ivasic-Kos M, Pobar M. Thermal object detection in difficult weather conditions using yolo [J]. *IEEE Access*, 2020, **8**: 125459-125476.
[23] Ren S, He K, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017,39(16):1139-1149.
[24] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.
[25] Redmon J, Farhadi A. Yolov3: An incremental improvement [J]. *arXiv preprint* arXiv:180402767, 2018.
[26] Yao S, Zhu Q, Zhang T, *et al*. Infrared image small-target detection based on improved FCOS and spatio-temporal features [J]. *Electronics*, 2022, **11**(6): 933.
[27] Lin F, Bao K, Li Y, *et al*. Learning contrast-enhanced shape-biased representations for infrared small target detection [J]. *IEEE Transactions on Image Processing*, 2024:33.
[28] Lin F, Ge S, Bao K, *et al*. Learning shape-biased representations for infrared small target detection [J]. *IEEE Transactions on Multimedia*, 2023,26:4681-4692.
[29] Sarlin P E, DeTone D, Malisiewicz T, et al. Superglue: Learning feature matching with graph neural networks [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 4938-4947..
[30] Sun J, Shen Z, Wang Y, et al. LoFTR: Detector-free local feature matching with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8922-8931.
[31] Li J, Wang P, Xiong P, *et al*. Practical stereo matching via cascaded recurrent network with adaptive correlation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16263-16272.
[32] Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning [J]. *Journal of Big Data*, 2019, **6**(1): 1-48.
[33] Patro S, Sahu K K. Normalization: A preprocessing stage [J]. *arXiv preprint* arXiv:150306462, 2015.
[34] Koonce B, Koonce B. Resnet 50 [J]. *Convolutional Neural Networks with Swift for Tensorflow*: *Image Recognition and Dataset Categorization*, 2021: 63-72.
[35] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4510-4520.
[36] Ma N, Zhang X, Zheng H T, *et al*. Shufflenet v2: Practical guidelines for efficient cnn architecture design [C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 116-131.
[37] Tan M, Le Q. Efficientnetv2: Smaller models and faster training; proceedings of the International conference on machine learning, F, 2021 [C]. PMLR.
[38] Lin T Y, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.
[39] Ross T Y, Dollár G. Focal loss for dense object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2980-2988.
[40] Zheng Z, Wang P, Liu W, *et al*. Distance-IoU loss: Faster and better learning for bounding box regression [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12993-13000.
[41] Su J, Liu Z, Zhang J, *et al*. DV-Net: Accurate liver vessel segmentation via dense connection model with D-BCE loss function [J]. *Knowledge-Based Systems*, 2021, **232**: 107471.
[42] Chicco D. Siamese neural networks: An overview [J]. *Artificial Neural Networks*, 2021,**2190**: 73-94.