基于改进 Faster RCNN 的毫米波图像实时目标检测

侯冰基[1,2,3]，杨明辉[1]，孙晓玮[1*]

[1]中国科学院上海微系统与信息技术研究所太赫兹固态技术重点实验室，上海 200050；

[2]中国科学院大学，北京 100049；

[3]上海科技大学信息科学与技术学院，上海 201210

摘要 采用反卷积与捷径连接，针对毫米波图像提出了一种高效、快速的卷积神经网络，在保留图像低阶细粒度特征的同时，检测速度由原框架的 9 frame/s 大幅提升至 27 frame/s，并取消了 Faster RCNN（Regions with Convolutional Neural Networks）中的 RCNN 部分。为了使网络更好地收敛，基于聚类思想设计了初始候选框的大小。使用在线困难样本挖掘（OHEM）优化了 Faster RCNN 的损失函数，解决了毫米波图像中正负样本失衡的问题，大幅提升了训练速度。所提算法在测试集上取得了 87.6%的准确率和 81.2%的检出率，F_1 分数相较于主流算法提升了 5%左右。

关键词 图像处理；图像识别；卷积神经网络；反卷积；毫米波图像；目标检测

中图分类号 TP751.2 **文献标识码** A **doi**：10.3788/LOP56.131009

Real-Time Object Detection for Millimeter-Wave Images Based on Improved Faster Regions with Convolutional Neural Networks

Hou Bingji[1,2,3], Yang Minghui[1], Sun Xiaowei[1*]

[1] *Key Laboratory of Terahertz Solid Technology, Shanghai Institute of Microsystem and Information Technology (SIMIT), Chinese Academy of Sciences, Shanghai* 200050, *China*;

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

[3] *School of Information Science and Technology, ShanghaiTech University, Shanghai* 201210, *China*

Abstract An efficient and fast convolution neural network for millimeter-wave images that uses deconvolution and a shortcut connection is proposed. The proposed network retains the low-order fine-grained features of the image and significantly improves the detection speed to 27 frame/s from 9 frame/s of original frame. The RCNN (Regions with Convolutional Neural Networks) part of the Faster RCNN is removed. To achieve better network convergence, the initial candidate box size is designed based on thought clustering. The online hard example mining is applied to optimize the loss function of the Faster RCNN such that the imbalance problem between positive and negative samples in millimeter wave images is solved and the training speed is improved significantly. By using the proposed algorithm, the accuracy of 87.6% and the detection rate of 81.2% are obtained on the test set. Compared with mainstream algorithms, the proposed algorithm improves the F_1 score by approximately 5%.

Key words imaging processing; image recognition; convolutional neural network; deconvolution; millimeter wave image; object detection

OCIS codes 100.3008; 100.4996;100.1830; 110.2970; 280.4750

1 引 言

传统的安检设备主要包括 X 射线成像仪、金属探测器、危险液体检测仪等。在各个机场、火车站等人流密集区域，已经形成了一套针对个人携带大件包裹、大件行李的完备检测机制。但对于人本身，除

收稿日期：2018-12-17；**修回日期**：2019-01-15；**录用日期**：2019-02-17

基金项目：国家自然科学基金(61731021,61671439)

* **E-mail**：xwsun@mail.sim.ac.cn

了通过安检门，依然需要通过安检人员手持金属探测仪进行人工检查。毫米波成像[1-3]基于合成孔径算法，采用无电离辐射技术代替人工检查，为旅客提供更加安全高效的安检服务。同时，实现安检的自动化与智能化，使用计算机来做危险物品的目标检测可以在很大程度上节约“人工判图”的时间与人力成本。在光学领域，针对图像分类和检测任务，借助大规模数据集的优势和神经网络强大的拟合能力，已经出现了一大批优秀的网络结构，如 VGG（Visual Geometry Group）[4]、ResNet（Residual Network）[5]、Inception 系列[6-9]等，以及检测框架，如 Faster RCNN[10]、SSD（Single Shot MultiBox Detector）[11]、YOLO（You Only Look Once）[12-14]等。Faster RCNN 等二阶分类器检测精度高但速度较慢，SSD、YOLO 等一阶分类器检测精度较低但检测速度快。

本文通过对毫米波的特点进行分析，总结了毫米波图像与光学图像的不同，并根据这些不同对 Faster RCNN 框架进行改进和优化：1）利用反卷积恢复被高度采样的特征图，保留图像的原始信息，降低了网络的采样倍数，提升了对小物体的检测能力，同时检测速度可达每张图 36 ms，快于 SSD 算法；2）考虑到毫米波目标检测是一个二分类问题，相对于精确的位置对检出率有更高的要求，取消了 Faster RCNN 框架的 RCNN（Regions with Convolutional Neural Networks）部分；3）使用 K-means++算法合理地设计初始候选框；4）将在线困难样本挖掘（Online Hard Example Mining，OHEM[15]）与 Faster RCNN 框架相结合，在训练过程中对负样本进行排序采样，使网络能够在训练中学习到当前的不足，提升收敛速度，也进一步提升了检出率。（本文中检测速度计算方式为：从读取图片后算起，中途将图像拷入图形处理器（Graphics Processing Unit，GPU），到检测出结果止，计算整个过程所消耗的时间。这种计算方式得到的时间为实际使用时的检测用时）。

2 毫米波图像分析

2.1 成像原理简介

毫米波收发机在扫描平面上作上下移动扫描，如图 1 所示[1-3]，扫描的频域范围从 28 GHz 到 33 GHz，涵盖之间的 64 个频点。记收发机在某一时刻的位置为(a,b,Z)，此时的频率为 ω，光速常量记为 c，波数为 $k=\omega/c$，目标物体在(x,y,z)位置处的反射系数记为 $f(x,y,z)$，对整个目标视场像素点积分得到电磁场数据为

$$s(a,b,\omega)=\iiint f(x,y,z)\exp\left[-\mathrm{j}2(\omega/c)\times\sqrt{(a-x)^2+(b-y)^2+(Z-z)^2}\right]\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z, \tag{1}$$

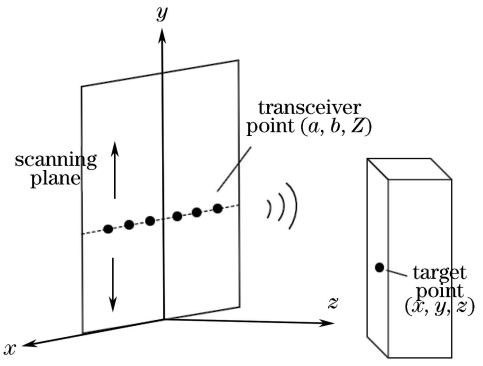

图 1 毫米波成像系统示意图

Fig. 1 Diagram of millimeter wave imaging system

推导出反射系数 $f(x,y,z)$，即可重构出物体的图像

$$I(x,y)=\max_z f(x,y,z)=\max_z\mathscr{F}_{3\mathrm{D}}^{-1}\{\mathscr{F}_{2\mathrm{D}}[s(a,b,\omega)]\cdot\exp\left[-\mathrm{j}\sqrt{4(\omega/c)^2-k_x^2-k_y^2}Z\right]\}, \tag{2}$$

式中：$\mathscr{F}_{2\mathrm{D}}$为二维傅里叶变换；$\mathscr{F}_{3\mathrm{D}}^{-1}$为三维傅里叶逆变换；$k_x$ 与 k_y 分别为波数 k 在 x 与 y 方向的分量。

2.2 图像分析

毫米波所在频谱介于红外与微波之间，这段频谱对于许多物体都具有一定的穿透性，使之能够透过衣服来检测被衣服遮挡的物体。与光学图像相比，毫米波图像具有如下特点：

1）从图像外观上看，如图 2 所示，毫米波照相机所采集的图像更相似于医学影像，图像为灰度图，没有光学图像丰富的颜色信息，甚至没有千姿百态的背景。网络在设计的时候需要考虑泛化性。主流的计算机视觉检测框架在毫米波图像上表现不够理想，常常出现过拟合现象；

2）图像没有远近造成的尺度变化，因此不需要像 SSD[11]、FPN[16]那样去获取不同尺度的特征图；

3）毫米波图像目标检测暂时为一个二分类问题，只需要检测出前景和背景，没有涉及多分类，且检出率比目标的精确位置更重要；

4）对于一些特别的物体，物体本身较模糊，难以根据物体所在像素来判断是否存在物体，还需要结合周围像素作判断；

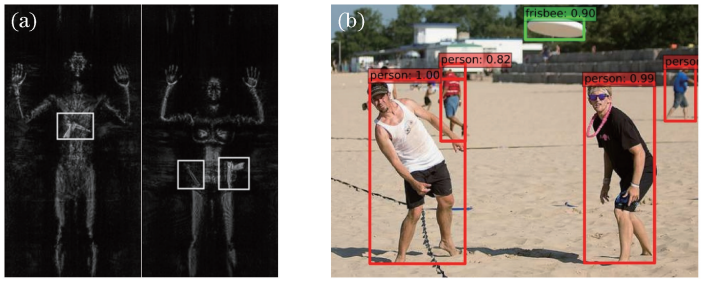

图 2 毫米波图像与光学图像对比。(a)毫米波图像,图中方框为真实标记;(b)光学图像,图中方框为检测结果,数字为目标置信度

Fig. 2 Comparison between millimeter wave images and optical images. (a) Millimeter wave images, boxes in figure are real markers; (b) optical images, boxes in figure are the test results, and the numbers are the target confidence

5) 毫米波图像中人所携带的物体普遍偏小,目标尺寸分布不同于光学图像,需要统计和设计初始化的尺寸;

6) 毫米波图像数据集中存在正负样本不平衡问题,表现为正样本偏少而负样本居多,使得模型在训练的过程中更倾向于判断为负样本,降低了检出率。

3 网络模型

3.1 反卷积与网络结构

在 CNN 的整个结构中,卷积层负责特征提取,池化层负责图像降采样与维持旋转不变性。随着网络结构越来越深,图像到达网络高层时感受野越来越大,特征图尺寸也越来越小,这对于小目标的检测来说也越来越困难。一般认为,网络的低层保留了图像的一些轮廓、纹理等低阶特征,而网络高层拥有更高阶的语义特征。在普通的网络结构中,网络经常经过降采样设计,对于图像中的小目标物体,一些比较原始的信息可能不会被保留。为恢复被高度采样的特征图,引入反卷积(deconvolution)。运算关系如图 3 所示,设 k_{en} 表示卷积核的大小,s_{tr} 表示卷积核的滑动步长,p_{ad} 表示卷积时候的补边。卷积过程中,输入图像尺寸 I_{in} 与输出图像尺寸 I_{out} 的变化关系可表示为

$$I_{out}=(I_{in}+2\times p_{ad}-k_{en})/s_{tr}+1, \quad (3)$$

反卷积过程中的输入输出关系为

$$I_{out}=(I_{in}-1)\times s_{tr}-2\times p_{ad}+k_{en}。 \quad (4)$$

在各个网络泛化能力的测试中,类 VGG 网络要优于类 ResNet 网络[17]。考虑到网络的泛化性要求,特征提取网络结构如表 1 所示,整体上采用了 VGG16+U-Net[18] 的思想。U-Net 是为了解决医疗影像的数据分割问题而被提出来的,可以兼顾感受野

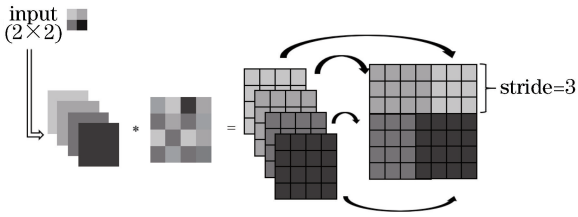

图 3 反卷积示意图

Fig. 3 Deconvolution diagram

和定位精度。在原检测框架中图像是经线性插值放大后再输入网络的,而线性插值不会额外带来更多的信息。将输入尺寸由原来的1000 pixel×500 pixel 调整为 512 pixel×256 pixel,间接扩大了整个网络的感受野。在网络 Convolution5 后添加反卷积层,再与 Convolution4 进行捷径连接,使得整个网络在最后输出时的特征图和原来调整之前的保持一致,而浮点计算量为原来的三分之一。

表 1 特征提取网络结构

Table 1 Feature extraction network structure

Type	Layers	Filters	Size/stride	Output
Convolution1	2	64	3×3/1	(512,256)
MaxPool			2×2/2	(256,128)
Convolution2	2	128	3×3/1	(256,128)
MaxPool			2×2/2	(128,64)
Convolution3	3	256	3×3/1	(128,64)
MaxPool			2×2/2	(64,32)
Convolution4	3	512	3×3/1	(64,32)
MaxPool			2×2/2	(32,16)
Convolution5	3	512	3×3/1	(32,16)
Deconvolution	1	512	2×2/2	(64,32)
Concate convolution5				(64,32)
Convolution6	2	512	3×3/1	(64,32)

在原 Faster RCNN 框架中,RCNN 负责解决多分类的问题,以及对 RPN(Region Proposal

Network)中预测的目标位置进行精修。但对于毫米波数据集,只需要检测出人是否携带有物品,且对物品的位置没有光学图像那么敏感。在 RCNN 之前,需要对特征图中的目标进行 ROI(Region of interest)池化,根据毫米波图像的第 4 个特征,这部分操作不会再带来性能上的提升,因此需在改进的 Faster RCNN 中将 RCNN 去除。

3.2 候选框聚类

图 4 为毫米波数据集中标记框的面积与光学图像 VOC(Visual Object Classes)数据集中标记框的面积直方图,纵坐标为不同面积的框占总框数的百分比。从直方图分布上看,毫米波图像中的物品尺度较小,变化也较小。为了引导网络更好地收敛,考虑使用 K-means++聚类的方法设计初始候选框的大小。

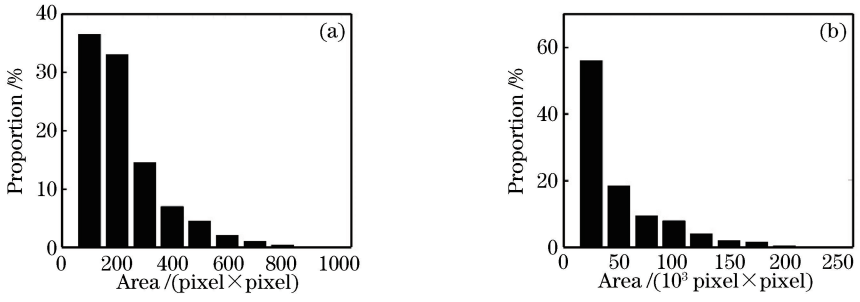

图 4 两个数据集中标记框的面积统计图。(a)毫米波图像数据集的统计图;(b) VOC 数据集的统计图

Fig. 4 Area statistics of label boxes in two data sets. (a) Statistical chart of millimeter wave image data sets; (b) statistical chart of VOC data sets

标记框集合表示为$\{(w^{(1)},h^{(1)}),\cdots,(w^{(i)},h^{(i)}),\cdots,(w^{(n)},h^{(n)})\}$,$w$ 和 h 分别表示框的宽和高,n 为总框数,i 为序号。原始的 K-means 算法从这 n 个框中随机选取 9 个作为初始聚类中心,在 K-means++算法中,首先从集合中随机选取一个成员记为$(\mu^{(1)},\eta^{(1)})$,在余下所有的成员中计算距离其最近聚类中心的欧式距离 $D^{(i)}(\mu^{(m)},\eta^{(m)})$,其中 m 为聚类族序号,在集合$\{D^{(i)},i\in[1,n]\}$中选取值最大的记为 $D^{(j)}$,此时将使 $D^{(j)}$ 最大的$(w^{(j)},h^{(j)})$作为下一个聚类中心,重复此过程直到 9 个聚类中心$\{(\mu^{(1)},\eta^{(1)}),\cdots,(\mu^{(9)},\eta^{(9)})\}$被找出。集合中所有的框根据与聚类中心的距离进行分类,更新 9 个聚类中心至这些中心不再变化为止。特征图上每个像素点产生预先设置好的 9 个候选框,去除超出边界的框后与真实标记作对比,交并比大于阈值的部分设为正样本,背景为负样本。

3.3 样本平衡与困难样本挖掘

毫米波数据集中每张图像由于目标较少,存在正负样本失衡的问题,部分图像正负样本比例可高达 1∶10000,故对原 Faster RCNN 的样本平衡机制进行修改。Faster RCNN 中目标的分类损失函数为

$$L_{\text{conf}}(p)=-\sum_{i}^{A}p_i^{*}\ln p_i-\sum_{i}^{A}(1-p_i^{*})\ln(1-p_i),\tag{5}$$

式中:A 为设置参与训练的候选框数量;p_i^{*} 为第 i 个候选框的标记;p_i 为网络预测第 i 个候选框为正样本的概率。该模型假设目标物体的存在概率服从一个概率为 p^{*} 的伯努利分布。原 RPN 中平衡正负样本的方式为

1) 正负样本总数不超过 A;

2) 正样本总数不超过 λA,λ 为一超参数,若超过以随机采样的方式保留 λA 个;

3) 负样本总数不超过 $A-K^{*}$,若超过,以随机采样的方式保留 $A-K^{*}$ 个,其中 K^{*} 为保留的正样本个数。

随机采样的方式会遗漏许多预测错误的候选框,降低网络性能,延长收敛速度。经 OHEM[15] 优化后分类损失函数设计为

$$P_{\text{os}}=\underset{i\in M}{\overset{\text{top}\,K^{*}}{\text{argsort}}}\,p_i^{*}(1-p_i),\tag{6}$$

$$N_{\text{eg}}=\underset{i\in M}{\overset{\text{top}\,A-K^{*}}{\text{argsort}}}\,(1-p_i^{*})p_i^{(0)},\tag{7}$$

$$L_{\text{conf}}(p,p^{*})=-\sum_{i\in P_{\text{os}}}^{M}\ln p_i-\sum_{i\in N_{\text{eg}}}^{M}\ln p_i^{(0)},\tag{8}$$

式中:P_{os}为筛选后的正样本集合;N_{eg}为筛选后的负样本集合;$L_{\text{conf}}(p,p^{*})$为分类损失函数;$p_i^{(0)}$ 为网络预测第 i 个候选框为背景的概率;M 为所有的不超出图像边界的候选框;$\underset{}{\overset{\text{top}\,K}{\text{argsort}}}$表示对函数值排序,取使其值最大的 K 个自变量的集合。RPN 中平衡样本的方式改进为

1) 正样本总数若超过 λA,就以 $p^{*}(1-p_i)$排序,选取最大的前 λA 个;

2) 负样本总数若超过 $A-K^*$，就以$(1-p^*)p_i^{(0)}$ 排序，选取最大的前 $A-K^*$ 个。

损失函数在进行反向传播时，每次自动选取当前状态下预测错误的值进行惩罚训练。这类较难预测的样本被称为困难样本。在实际训练过程中，相比于其他模型二十余次的循环训练过程，本文模型在一次循环后即可达到一个较理想的水平，训练速度有很大提升。

4 实 验

4.1 数据集与评价指标

数据集由中国科学院上海微系统与信息技术研究所采集，采集的设备为自主研制的 SimImage-MD 系列毫米波成像仪，参与采集的模特近千人，每张图均有人工标注，现有训练集 143334 张；测试集名称为 MMW-19，采集于另外的 19 位模特，包含 10 位女性、9 位男性，训练集中模特分别携带有 8 样最常见的物品，且物品均摆放在训练集中常出现的位置，共计 12198 张图像。每张图像的大小均为 380 pixel×190 pixel。

与一般的目标检测任务类似，使用准确率 p_{re}、检出率 r_{ec} 以及 F_β 作为模型性能评价指标，它们的表达式为

$$p_{\mathrm{re}}=\frac{c_{\mathrm{BoxNum}}}{t_{\mathrm{BoxNum}}},\tag{9}$$

$$r_{\mathrm{ec}}=\frac{c_{\mathrm{ObjectNum}}}{t_{\mathrm{ObjectNum}}},\tag{10}$$

$$F_\beta=(1+\beta^2)\frac{p_{\mathrm{re}}\times r_{\mathrm{ec}}}{\beta^2\times p_{\mathrm{re}}+r_{\mathrm{ec}}},\tag{11}$$

式中：c_{BoxNum} 为预测正确的框数；t_{BoxNum} 为预测出来的总的框数；$c_{\mathrm{ObjectNum}}$ 为预测正确的物品数；$t_{\mathrm{ObjectNum}}$ 为总物品数，这里参数 β 取 1。

4.2 训练设置

文中所使用的深度学习框架为 caffe[19]，目标检测框架基于 Faster RCNN，在此基础上作了改进。整个训练过程在 4 块 GTX TITAN XP 上完成，优化方法为 SGD(Stochastic Gradient Descent)，冲量设置为 0.9，正则化方法为 L2 范数正则化，参数为 0.0005。训练耗时 9.5 h，累计训练 7 epoch，在第 6 epoch时将学习率由 0.001 调整为原来的 0.1 倍。Batchsize 设置为 64。实际输入图像大小为 380 pixel×190 pixel，线性插值放大后的大小调整为 512 pixel ×256 pixel，再输入网络，选用翻转作数据增强，并随机化训练顺序。设置图像像素均值为 28。根据物品实际大小，修正了初始候选框的尺寸。若标签与初始候选框的最大交并比仍小于阈值，选取最大交并比的候选框作为正样本。初始化预训练模型使用 VGG16 在 VOC 数据集上训练得到。最终，选取置信度高于 0.5 的候选框作为输出，部分训练结果如图 5 所示。

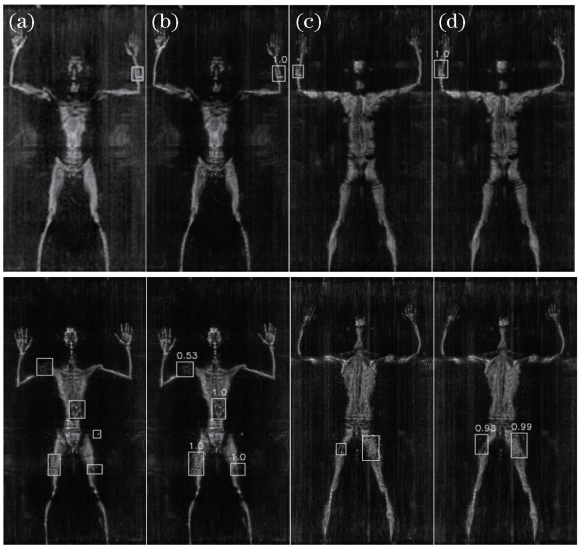

图 5 部分训练结果。(a1)(b1)(c1)(d1)真实标记；(a2)(b2)(c2)(d2)对应的检测结果

Fig. 5 Partial training results. (a1)(b1)(c1)(d1) real marks; (a2)(b2)(c2)(d2) test results of corresponding graph

4.3 与其他网络框架的对比

表 2 列举了目前主流所使用的模型与本文模型之间的对比。VGG-1、VGG-2、ZF-Net、ResNet、DenseNet[20]使用的是 Faster RCNN 框架，VGG-1 保留了 RCNN，随后的 4 个模型使用 OHEM 优化后的损失函数，移除了 RCNN。SSD 检测框架模型使用的是 VGG16，为避免最后的图像特征图过小，将第 4 个 MaxPool 层的核由 2×2/2 改为 3×3/1。YOLOv3 使用的模型为 DarkNet。由于 YOLOv3 的分类打分偏低，为方便评价指标的比较，阈值设置为 0.3。相同框架下的模型在训练时使用的优化方法相同。

从表 2 可以看出，本文模型的准确率和检出率在这几类模型中表现最优。VGG-2 的测试结果高于 VGG-1，表明 RCNN 相对于整个数据集而言作用并不明显。YOLOv3 相较于之前的 YOLO 系列提升了对小目标的检测能力，但整个框架缺乏正负样本的平衡机制，对于毫米波图像这类正负样本极为不平衡的情况不太适用。ZF-Net 由于卷积层数少的原因，检测速度略快于本文模型。尽管卷积层数为 19 层，但本文模型依然达到了每张图 36 ms，即 27 frame/s 的实时检测速度。

表 2 网络测试 GPU 环境为 GTX 1080 时的实验结果对比

Table 2 Comparison of experimental results when the GPU environment for network testing is GTX 1080

Model	Input size /(pixel×pixel)	Sampling	Layers	Without RCNN	p_{re}	r_{ec}	F_1 score	Time /ms
VGG-1	1000×500	16	16	×	0.859	0.744	0.797	107
VGG-2	1000×500	16	16	√	0.864	0.785	0.822	97
ZF-Net	1000×500	16	5	√	0.837	0.757	0.794	27
ResNet	1000×500	16	41	√	0.867	0.766	0.813	66
DenseNet	1000×500	16	90	√	0.850	0.540	0.660	94
SSD	300×300	8	16	√	0.854	0.763	0.805	44
YOLOv3	416×416	8	59	√	0.822	0.784	0.802	29
Ours	512×256	8	19	√	0.876	0.812	0.843	36

相比较之下，DenseNet 在这几类网络中表现最差，为此，统计了 5 类网络在相同框架、相同优化方法下训练后期的损失值。此时的损失值已经趋于稳定，起伏非常小，可以用作网络收敛性上的对比。

计算相同框架下最后 10000 次训练各模型的平均损失，5 类网络 ZF-Net、VGG-2、Ours、ResNet 和 DenseNet 的平均损失分别为 0.044、0.034、0.032、0.043和 0.028。就收敛性而言，DenseNet 是这几类网络中收敛性最好的网络，这也可以由 DenseNet 的网络结构得以解释。DenseNet 在进行前向传播时将本层的特征值保留，并直接传递给后面的每一层，得益于这种结构，网络本身就是许多不同深度网络的集合[21]，但这种强大的拟合能力在特征较少的情况下并不适用。神经网络的可解释性值得作进一步讨论。

5 结 论

利用反卷积提出了一个全新的网络结构，在低层的特征图上结合高层特征作预测，可较好地应用于毫米波小目标检测。通过优化 Faster RCNN 的损失函数，每次训练时自动选择预测错误的样本参与训练，进一步提升了性能，也提升了训练速度，利于工程化。尤其是在检测速度上，所提算法在安检这一环境背景下更具优势，可以应用于毫米波智能摄像机检测视频中的目标。

参 考 文 献

[1] Sheen D M, McMakin D L, Hall T E. Three-dimensional millimeter-wave imaging for concealed weapon detection [J]. IEEE Transactions on Microwave Theory and Techniques, 2001, 49(9): 1581-1592.

[2] Zhu Y K, Yang M H, Wu L, *et al*. Millimeter-wave holographic imaging algorithm with amplitude corrections [J]. Progress in Electromagnetics Research M, 2016, 49: 33-39.

[3] Zhu Y K, Yang M H, Wu L, *et al*. Practical millimeter-wave holographic imaging system with good robustness[J]. Chinese Optics Letters, 2016, 14(10): 101101.

[4] Albelwi S, Mahmood A. Automated optimal architecture of deep convolutional neural networks for image recognition[C]// 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), December 18-20, 2016, Anaheim, CA, USA. New York: IEEE, 2016: 53-60.

[5] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[6] Szegedy C, Liu W, Jia Y Q, *et al*. Going deeper with convolutions [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 7298594.

[7] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C] // Proceedings of the 32nd International Conference on International Conference on Machine Learning, July 6-11, 2015, Lille, France. Massachusetts: JMLR.org, 2015, 37: 448-456.

[8] Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 2818-2826.

[9] Szegedy C, Ioffe S, Vanhoucke V. Inception-v4, inception-ResNet and the impact of residual connections on learning [EB/OL]. (2016-08-23) [2018-12-01]. https://arxiv.org/abs/1602.07261.

[10] Ren S Q, He K M, Girshick R, *et al*. 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.

[11] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot multibox detector[M]// Leibe B, Matas J, Sebe N, *et al*. Lecture Notes in Computer Science. Cham: Springer, 2016, 9905: 21-37.

[12] 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), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 779-788.

[13] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6517-6525.

[14] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2018-12-01]. https://arxiv.org/abs/1804.02767.

[15] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 761-769.

[16] Lin T Y, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 936-944.

[17] Azulay A, Weiss Y. Why do deep convolutional networks generalize so poorly to small image transformations[EB/OL]. (2018-03-30)[2018-12-01]. https://arxiv.org/abs/1805.12177.

[18] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[M]// Navab N, Hornegger J, Wells W, *et al*. Lecture Notes in Computer Science. Cham: Springer, 2015, 9351: 234-241.

[19] Jia Y Q, Shelhamer E, Donahue J, *et al*. Caffe: convolutional architecture for fast feature embedding[C]// Proceedings of the 22nd ACM International Conference on Multimedia, November 3-7, 2014, Orlando, Florida, USA. New York: ACM, 2014: 675-678.

[20] Huang G, Liu Z, Maaten L V D, *et al*. Densely connected convolutional networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2261-2269.

[21] Veit A, Wilber M, Belongie S. Residual networks are exponential ensembles of relatively shallow networks[EB/OL]. (2016-10-27)[2018-12-01]. https://arxiv.org/abs/1605.06431v1.