Journal of Computer Applications
计算机应用

yyyy-mm-dd
http://www.joca.cn

文章编号:1001-9081(****)**-0000-00

doi:10.11772/j.issn.1001-9081.yyyymmnnnn

您的稿件编号：2019061075
本文章有任何疑问时请联系
作者：沈文祥
手机号： 18435130062
邮箱：SFLYswx@outlook.com

基于多级特征和混合注意力机制的室内人群检测网络

沈文祥，秦品乐，曾建朝*

(中北大学 大数据学院，太原 030051)

(*通信作者电子邮箱 zjc@nuc.edu.cn)

摘 要: 针对室内人群目标尺度和姿态多样性、人头目标易与周围物体特征混淆的问题，提出了一种基于多级特征和混合注意力机制的室内人群检测网络。该网络结构包括三部分，即特征融合模块、多尺度空洞卷积金字塔特征分解模块以及混合注意力模块。首先，通过将浅层特征和中间层特征信息融合，形成包含上下文信息的融合特征，用于解决浅层特征图中小目标语义信息不丰富、分类能力弱的问题；然后，利用空洞卷积增大感受野、不增加参数的特性，对融合特征进行多尺度分解，形成新的小目标检测分支，满足网络对多尺度目标的定位和检测；最后，局部混合注意力模块通过融合全局像素关联空间注意力和通道注意力，增强对关键信息贡献大的特征，来增强网络对目标和背景的区分能力。实验结果表明，所提方法在室内监控场景数据集 SCUT-HEAD 上达到了 0.94 的准确率、0.91 的召回率和 0.92 的 F1 分数，在召回率、准确率和 F1 指标上均明显优于当前用于室内人群检测的其他算法。

关键词: 室内人群检测；特征融合；注意力机制；空洞卷积 ；特征金字塔

中图分类号:TP389.1 人工神经网络计算机；TP391.41 图像识别及其装置 **文献标志码：A**

Indoor crowd detection network based on multi-level features and fusion attention mechanism

SHEN Wenxiang, QIN Pinle, ZENG Jianchao*

(College of Big Data, The North University of China, Taiyuan 030051, China)

Abstract: In order to solve the problem of indoor crowd target scale and attitude diversity and confusion of head targets with surrounding objects, a ~~method~~network based on Multi-level Features and hybrid Attention mechanism (MFANet) for indoor crowd detection was proposed. It included three parts, namely a feature fusion module, a multi-scale dilated convolution pyramid feature decomposition module, and a fusion attention module. Firstly, by combining the shallow features and the intermediate layer feature information, a fusion feature containing context information was formed to solve the problem that the small target semantic information is not rich and the classification ability is weak in the shallow feature map. Then, the dilated convolution was used to increase the receptive field. Without increasing the characteristics of the parameters, the multi-scale decomposition of the fusion features was performed to form a new small target detection branch, which satisfies the positioning and detection of the multi-scale target by the network. Finally, the local fusion attention module integrated the global pixels space attention and channel attention to enhance the contribution of key information and the ability of distinguishing target and background. Experiments show that the proposed method achieves an accuracy of 0.94, a recall rate of 0.91, and an F1 score of 0.92 on the indoor monitoring scene data set SCUT-HEAD. All are significantly better than other algorithms currently used for indoor crowd detection.

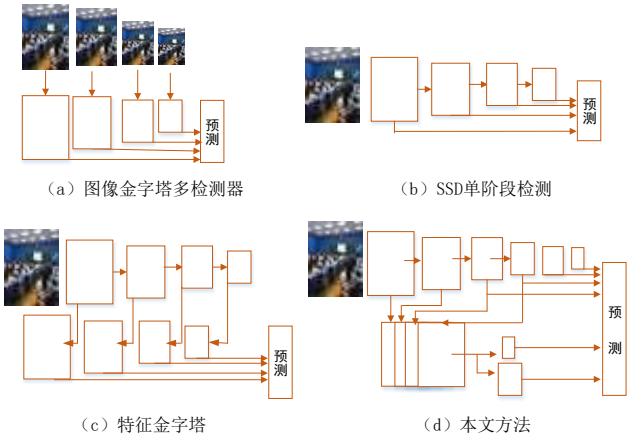
Keywords: indoor heads detection; single stage object detection; attention mechanism; dilate convolution; feature pyramid;

0 引言

计算机视觉一直是计算机科学领域研究热点之一。作为计算机视觉领域的一个典型应用，公共室内场所人数统计在人流量商业数据统计分析、公共安全等许多方面有着重要的应用价值。目前室内场景人群计数主要有两种思路：一种是直接通过回归的方式得到人群数量，另一种是采用检测的方式进行人群检测。基于回归的方法只能预测人群密度，得到一个粗略的结果；基于检测的方法可以得出精确的定位信息和人数统计。目前针对人的检测方法主要有两种，一类是人脸识别的算法[1-4]，一类是行人识别的算法[5-6]。但是，这两种方法在室内人群检测中性能均不好。人脸识别只能检测人

收稿日期: 2019-06-24; **修回日期**:2019-09-19; **录用日期**: yyyy-mm-dd。
基金项目: 山西省重点研发项目（201803D31212-1）
作者简介: 沈文祥(1995—)，男，安徽淮南人，硕士研究生，主要研究方向：深度学习、计算机视觉；秦品乐(1978—)，男，山西长治人，副教授，博士，CCF 会员，主要研究方向：机器视觉、大数据、医学影像；曾建朝（1963—）男，陕西大荔县人，教授，博士，CCF 会员，主要研究方向：演化计算，机器学习

脸，这意味着相机无法检测人的背面。由于室内场景人群的复杂性，很多身体部位被相互遮挡，因此，行人识别同样也无法很好地解决该问题。然而，人头检测却没有这些限制，可以很好的适用于室内人群定位和计数。当然，室内场景人头检测同样存在很多挑战。



（a）图像金字塔多检测器　（b）SSD单阶段检测

（c）特征金字塔　（d）本文方法

图 1 四种多尺度检测结构

Fig.1 Four multi-scale detection structure

头部姿态和尺度的多样性是人头检测的第一大关键难题。目前主要采用测试阶段输入多尺度图像和训练过程中对中间特征层进行多尺度变换两种主要的思路改善这个问题。第一种图像金字塔结构的思路，如图 1(a)所示。多任务卷积神经网络(Multi-Task Convolutional Neural Network, MTCNN)[4]直接通过下采样得到不同尺度的输入图像送入训练好的检测网络中进行预测，最终通过非极大值抑制(Non Maximum Suppression, NMS)[7]输出目标位置和种类。~~SNIP[8]被~~ Singh 等[8]提出在训练和测试中，建立大小不同的图像金字塔，在每张图上都运行一个检测网络，同时只保留那些大小在指定范围之内的输出结果，最终通过非极大抑制操作输出目标位置和种类。由于这类基于图像金字塔结构的算法计算复杂度高，内存消耗严重，耗时长，因此在检测任务上的效率非常低。第二种特征金字塔结构的思路是目前目标检测算法中出现最多的，Single Shot MultiBox Detector (SSD)[9]采用多层特征图独立检测输出，构成多尺度特征检测结构，如图 1(b)所示。Lin 等[10]提出了一种将高层特征和浅层特征图融合的至上而下的结构 Feature Pyramid Networks (FPN)，最后在融合后的不同层进行独立预测，如图 1(c)所示。Zhou 等[11]同样也提出了一种对称的结构进行多尺度融合。相比于第一种思路，第二种思路利用更少的内存和耗时，并且还可以作为组件嵌入到不同的检测网络中。因此本文也采用这种思路。通过分析，发现浅层特征对于小尺度目标有很好的定位能力，但是语义表征信息弱，由于连续下采样，小目标区域在中间层特征图中的表征区域已经降为 1×1 像素大小，因此，本文只利用上采样融合浅层和中间层特征，然后采用多尺度空洞卷积金字塔结构生成新的浅层和中间层检测分支，高层检测分支仍然采用原来的特征层，形成一个两级检测的混合结构，如图 1(d)所示。通过本文设计的特征融合结构和多尺度空洞卷积金字塔结构很好地改善了头部姿态和尺度多样性的问题。

图像质量不高容易使得人头区域与周围物体特征混淆，因此如何只关注目标特征，忽略背景特征干扰是室内场景人头检测另一个关键难题。目前很多算法引入注意力机制，引导神经网络关注目标区域，排除背景特征的干扰。Jaderberg 等[12]发现神经网络中池化和下采样操作直接将信息合并会导致关键信息无法识别出来，提出了一种新的空间转换模块结构，用于指导网络显式的学习目标的空间特性，例如旋转、平移等，这相当于空间域的注意力机制。Hu 等[13]发现不同的特征图对关键信息的贡献不同，因此通过学习的方式来自动获取到每个特征通道的重要程度，然后依照这个重要程度去提升有用的特征并抑制对当前任务用处不大的特征，这相当于通道域的注意力机制。Zhang 等[14]引入一种关注特征相似性，从而扩大图像感受野的注意力机制用于图像超分辨率。因此，注意力机制已经被很好的证明适用于关键特征的提取。本文提出了一种混合通道域和空间域的注意力模块，嵌入到不同的检测分支中，增强了不同分支对目标特征和背景特征的区分能力。

通过本文设计的特征融合模块和混合注意力模块很好的解决了上述两个难题，本文算法在标准数据集上达到了 0.91 的召回率(recall)，大大优于现有的算法[9,15-17]。

本文的主要工作有以下几点：

1)设计了一种新颖的特征融合结构。首先通过上采样操作将中间层特征图和浅层特征图尺度归一化，然后利用 concate 操作融合特征图，构成包含丰富小目标定位信息和语义信息的融合层，改善了网络浅层对小目标表征不足的问题。

2)设计了一种新颖的多尺度空洞卷积金字塔特征分解结构。利用多尺度空洞卷积金字塔结构对融合特征图进行多尺度分解，构成对小目标和中等目标检测的新分支，利用原网络针对大目标的检测分支和新生成的检测分支构成多特征层检测结构，有效地利用了网络不同层对目标检测的贡献，有效地改善了单阶段网络对多尺度和多姿态人头的检测性能不足的问题。

3)设计了一种混合空间域和通道域的注意力结构嵌入到不同的检测分支中，增强对关键信息贡献大的特征图，大大增强了网络对目标区域和背景区域的分辨能力。

4)以 Visual Geometry Group (VGG16)轻量级特征提取网络为基本网络结构，结合本文提出的特征融合分解结构和注意力机制，构成了单阶段两级检测的端到端网络，在训练和检测阶段实现了实时的人群检测网络。

1 相关工作

目前基于深度学习的目标检测算法主要分为两类。一类是两阶段检测算法，例如 Fast-RCNN[18]，Faster-RCNN[15]和 R-FCN[16]。这类方法都是通过先生成目标的候选区域并进行

粗筛选，然后对筛选后的候选区域进行目标分类和边界框回归。第二类是单阶段检测算法，主要有：OverFeat[19]，SSD[9]，YOLO[17]系列。

单阶段检测网络增强浅层特征对小目标表征能力的方法主要分为两类。第一类是是直接将输入图像放大提升小目标尺度，诸如：Multi-task convolutional neural network（MTCNN）、Scale Normalization for Image Pyramids (SNIP)，这一类算法都是将输入图像多尺度放大后用于训练或测试阶段。第二类是对特征图进行多尺度变换再利用，诸如：多尺度深度卷积神经网络(Multi-Scale deep Convolutional Neural Network, MS-CNN)[20]，反卷积单目标检测器(Deconvolutional Single Shot Detector, DSSD) [21]。

深度学习中的注意力是一种模拟人大脑处理视觉任务的机制，人类视觉只关注感兴趣区域，忽略其他背景干扰。注意力机制(Attention mechanism)[22]可以被解释为将可用的计算资源的分配偏向于包含最有用信息的特征部分，首先用于自然语言处理中关注对下文词语贡献高的词语，之后在很多图像处理任务中也已经证明了注意力机制的实用性，包括目标检测[13]，图像超分辨率[14]等。在这些任务中，注意力机制作为一种模块嵌入网络层中，表示用于模态之间的自适应高级抽象。

2 室内人群检测网络

本文提出的室内人群检测模型 MFANet 整体结构如图 2 所示，其和 SSD 一样是端到端的单阶段检测网络。主干网络采用轻量级网络 VGG16 的卷积层用于提取特征，替换用于分类的全连接层，并且额外增加了卷积层，形成特征提取主干网络，通过上采样操作将浅层和中间层特征尺度归一化，再通过 concate 操作构建融合层融合浅层和中间层特征图，对原有浅层和中间层特征进行融合形成新的融合特征图，再利用多尺度空洞卷积分解结构对融合层进行多尺度分解形成新的小目标检测分支，结合原有特征层形成的大目标检测分支，形成一个两级检测分支用于产生密集的预测框和分类置信度，最后通过软化非极大值抑制(soft-NMS)输出最终的检测结果。MFANet 主要包含 3 部分结构：特征融合模块(Feature Fusion Module, FFM)，多尺度空洞卷积金字塔特征分解结构(Multi-Scale Dilate convolution Feature Pyramid decomposition Module, MSDFPM)，混合注意力模块(Fusion Attention Module, FAM)。特征融合模块通过将浅层特征和中间层特征信息融合，形成包含上下文信息的融合特征，用于解决浅层特征图中小目标语义信息不丰富，分类能力弱的问题；多尺度空洞卷积金字塔结构主要利用空洞卷积感受野增大，参数不增加的特性，对融合特征进行多尺度分解，形成新的小目标检测分支，满足网络对多尺度目标的定位和检测；局部混合注意力模块通过融合全局像素关联空间注意力和通道注意力，增强对关键信息贡献大的特征，大大增强网络对目标和背景的区分能力。下面将详细描述三个模块的设计思路。

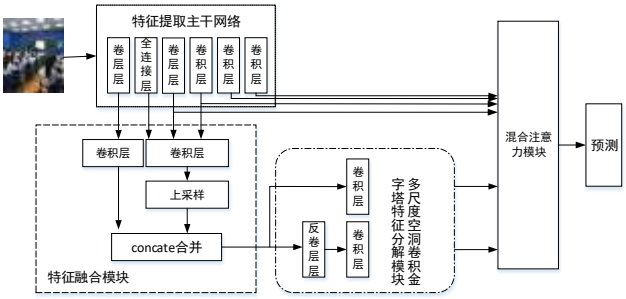


图 2 网络总体结构图

Fig. 2 Network structure diagram

2.1 特征融合模块

本文将绝对尺寸在图像中占据的区域小于 32×32 像素的目标定义为小目标。单次检测网络均存在小目标检测能力弱的问题，究其原因是由于用于特征提取的主干网络浅层特征图虽然包含丰富的细节定位信息，但是包含的小目标语义信息少，对小目标的分类能力弱。随着网络层加深，深层特征图包含丰富的语义信息，但是丢失了小目标的细节定位信息。因此，最直接的想法是将包含丰富细节定位信息的浅层特征图和包含丰富语义信息的深层特征图通过一定的融合规则融合形成既包含丰富细节定位信息又包含丰富语义信息的特征图。由于小目标特征在经过多层下采样之后，原有细节和语义信息已经丢失，在深层网络层中已经不再包含小目标的语义信息。如图 3 所示，当原图像中一个头部区域为 30×30 时，浅层特征图中的丰富细节特征随着网络层的加深，图像不断被下采样，最终在 conv7_2 特征图中目标区域已经被抽象成一个点特征，在 conv7_2 之后的特征图中已经丢失了该目标的特征信息。因此，直接使用最深层特征图对浅层进行语义增强的效果并不明显。

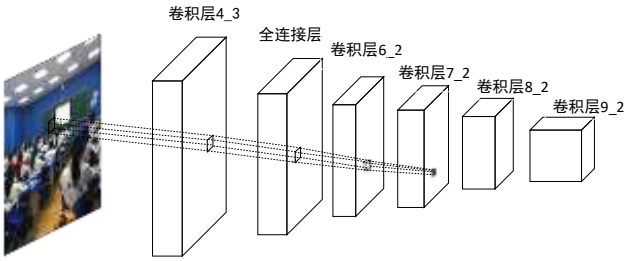


图 3 小目标在网络各层中的感受野

Fig. 3 Receptive field of small targets in all layers of the network

基于此分析，采用将中间层特征和浅层特征进行融合的思路，在 SSD 模型中原有的主干特征提取网络中嵌入了新颖的特征融合模块形成包含全局上下文信息的融合特征。如图 4 所示，首先利用 1×1 的卷积构建瓶颈层压缩浅层和中间层特征图的通道，然后分别将中间层特征图通过上采样放大到和浅层特征图相同的尺寸，最后这里没有采用将所有特征图

相加形成新的特征图，而是利用 concate 将所有相同尺寸的特征图连接起来形成第二层特征图，主要是由于像素级相加操作要求两个特征图有相同的长宽和通道，那么就需要在融合前确保两个特征图尺度完全一致，这么做的缺点是新增了额外的归整化操作，并限制了被融合 feature map 的灵活性，并且 concate 连接操作可以很好地保证不同特征图检测的同一个目标所包含的特征区域被相同激活。相比于主干网络提取的特征，新的融合特征图既包含特征提取主干网络中浅层特征图中小目标丰富的细节特征，同时又利用中间层特征图中小目标丰富的语义信息，这大大提升了检测小目标的准确率。

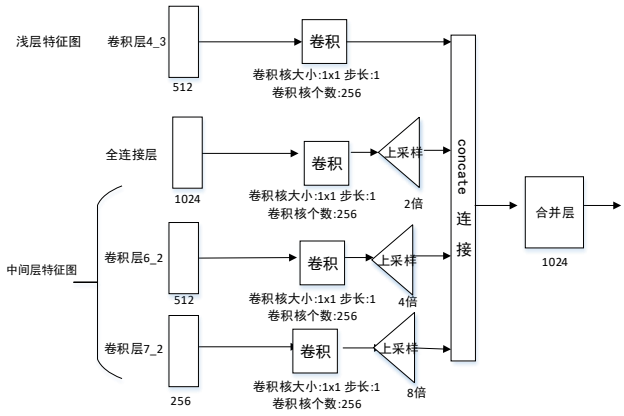


图 4 特征融合模块结构

Fig. 4 Feature fusion module structure

2.2 多尺度空洞卷积金字塔结构

在得到融合特征图后，需要生成新的检测分支。受 FPN 的启发，对融合后的特征图可以进行多尺度下采样，构成多尺度金字塔结构用于生成新的检测分支。由于主要目标是需要提升小目标的检测能力，因此，本文只生成 conv4_3 和 fc7 两个新的检测分支，用于构成小目标检测分支，如图 5 所示，新生成的检测特征图需要和原检测特征图尺寸相同，感受野相同，这样可以保证平均地检测不同尺度的目标，而标准的卷积由于感受野局限，传统做法一般采用池化操作进行下采样，但是这样容易丢失定位信息。因此，为了增大感受野的同时，又不丢失小目标定位信息，研究者们提出一种新的卷积操作：空洞卷积[23]。如图 6 所示，它是在标准卷积的基础，通过填零操作，增大了感受野的同时，而不增加学习参数，只增加了一个超参数：空洞率(dilate rate)。由于空洞卷积操作容易引起网格效应，根据 Wang 等[24]提出的空洞卷积级联参考设计准则，首先利用空洞率为 2 的空洞卷积操作增大感受野，再级联一个空洞率为 1 的标准卷积用于消除网格效应，最后利用滑动步长为 2 的 3×3 卷积进行下采样操作生成 conv7_2，在新生成的检测分支之后，均添加了一个 3×3 卷积用于整合通道内部相关性信息。新生成的小目标检测分支相比原检测分支，拥有更丰富的小目标细节特征和语义特征。

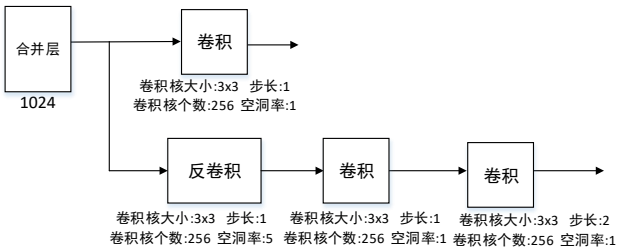
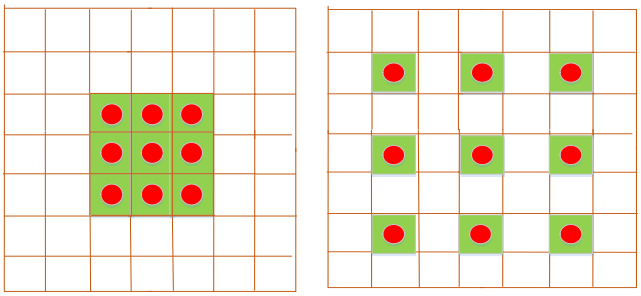


图 5 多尺度空洞卷积金字塔结构

Fig. 5 Multiscale dilate convolution pyramid structure



(a) 标准卷积核结构 (b) 空洞卷积核结构

图 6 两种卷积核结构对比

Fig. 6 Two convolution structure com[are

2.3 混合注意力模块

由于室内监控图像成像质量差，人群密度大，场景内容复杂，因此，很容易造成目标和周围背景的特征相似度高，影响网络对目标的判断。因此，要求设计的模型能够很好的区分目标和背景特征。最直接的想法是对图像进行超分辨率，然后再进行目标识别，但是，这样会造成内存占用高，计算复杂度增加，并且无法满足端到端的训练和推理，大大增加了推理和训练时间。根据 Squeeze-and-Excitation Networks(SENet)的论述[13]，神经网络不同特征图、同一特征图内不同区域对不同目标的贡献率都是不同的，如果能够只使用对关键目标贡献率高的特征图，舍弃对关键目标贡献率不高的特征图，则会大大提升对目标的定位和识别。而新近快速发展的注意力机制可以很好的实现这个功能。因此，本文设计了一种混合注意力模块用于提取关键特征，整体结构如图 7 所示，输入特征图 $x \in \mathbf{R}^{H\times W\times C}$，经过通道注意力模块提取对目标贡献率大的的通道注意力图 $F(x) \in \mathbf{R}^{1\times 1\times C}$，通过级联的方式，利用空间注意力模块提取二维的空间注意力图 $G(x) \in \mathbf{R}^{H\times W\times 1}$，得到最终的输出。整个注意力提取的过程如式（1）所示。

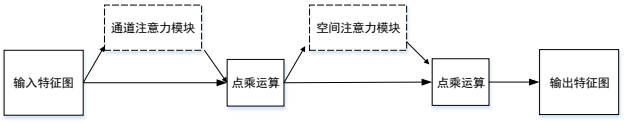


图 7 混合注意力模块结构

Fig. 7 Hybrid attention module structure

$$Z(x)=G(F(x)\otimes x)\otimes F(x) \quad (1)$$

其中：⊗是像素级点乘，在点乘过程中，注意力图被广播到不同通道、不同区域的特征图中，最终的输出 $Z(x)$ 既包含空间注意力，又包括通道注意力。如图 8 所示，输出了浅层添加注意力机制和不添加注意力机制后的部分特征图，可以看出本文设计的注意力结构很好地增强了特征图中目标区域的语义信息和细节定位信息。

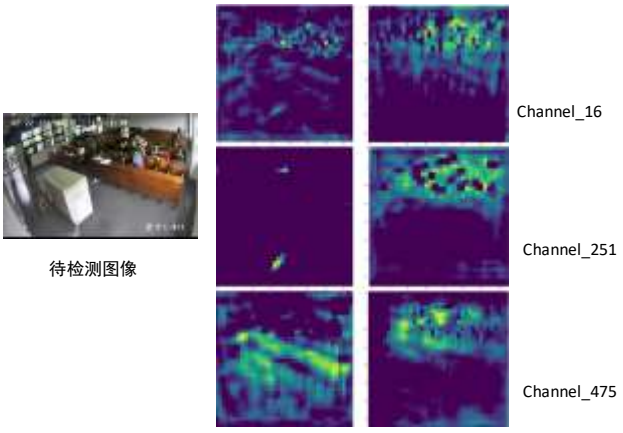


图 8 浅层特征图部分通道对比图.(a)SSD 方法检测结果 (b)本文方法检测结果

Fig. 8 Comparison of partial channel shallow feature map.(a)method of SSD (b)method of ours

2.3.1 通道注意力子模块

每一个通道特征图都可以看作是特征检测器，针对不同的目标，不同通道的特征图对关键信息的贡献率是不同的。通道注意力关注的就是不同的通道对关键信息的贡献率。因此本文设计了一种用于提取通道和目标之间内在关系的结构，如图 9 所示。

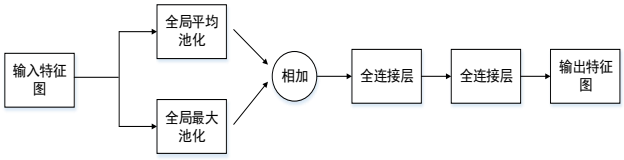


图 9 通道注意力结构

Fig. 9 Channel attention structure

为了只学习不同通道的贡献率，首先压缩空间信息，目前普遍采用全局平均池化的方法。Zhou 等提出使用全局平均池化来获得目标检测候选区域，SENet 在设计的注意力模块中使用全局平均池化统计特征图的空间信息。不同于他们的思路，本文认为全局最大池化操作可以获得目标之间差异性最大的特征，可以有助于推断更精细的通道注意力。因此，本文同时采用全局平均池化和全部最大池化两种操作。首先利用全局平均池化和全局最大池化分别生成不同的空间描述特征：$M_{ave}^{c}\in\mathbf{R}^{1\times1\times C}$，$M_{max}^{c}\in\mathbf{R}^{1\times1\times C}$。然后通过像素级相加得到融合后的通道描述特征 M_{merge}^{c}。融合后的通道描述特征送入一个多层感知机得到最终的通道注意力图。为了压缩参数，本文设置了一个压缩比（ratio），通过大量实验，最终该参数设置为 16。最后整个通道注意力提取的过程可以描述如下：

$$M_{merge}^{c}(x)=M_{ave}^{c}(x)+M_{max}^{c}(x) \quad (2)$$

$$F(x)=\sigma(W_1(\mathrm{ReLU}(W_0M_{merge}^{c}(x)))) \quad (3)$$

其中 σ 为 sigmoid 函数，因为通道注意力提取过程是获得通道特征图对关键信息的贡献率，属于广义二分类问题。多层感知机的权重：$W_0\in\mathbf{R}^{C\times C/r}$，$W_1\in\mathbf{R}^{C/r\times C}$，$W_0$ 之后~~先~~使用 ReLU 激活函数来提升网络的非线性程度。

2.3.2 空间注意力子模块

空间位置注意力主要是寻找特征图中对关键信息重要的区域，这是对通道注意力的一种补充。由于普通的卷积操作受限于卷积核的大小，只能考虑领域内的特征内在联系，无法考虑全局区域中相似特征的关联性。因此为了获取全局区域对关键信息的贡献，本文受非局部网络启发，设计了一个新颖的空间注意力结构，如图 10 所示。

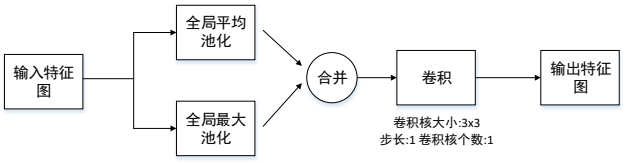


图 10 空间注意力结构

Fig. 10 Spatial attention structure

输入特征图 $x\in R^{H\times W\times C}$ 首先通过全局最大池化和全局平均池化操作，沿通道维度生成两个新的特征描述：$M_{ave}^{s}\in\mathbf{R}^{H\times W\times1}$，$M_{max}^{s}\in\mathbf{R}^{H\times W\times1}$。然后通过 concate 操作融合新的特征描述，之后通过一个标准的卷积操作激活获得最终的注意力图。整个注意力提取过程描述如下所示。

$$M_{merge}^{s}(x)=[M_{ave}^{s},M_{max}^{s}] \quad (4)$$

$$G(x)=\sigma(f^{3\times3}M_{merge}^{s}(x)) \quad (5)$$

其中：σ 为 sigmoid 函数；$f^{3\times3}$ 表示 3×3 的标准卷积操作。本文设计的空间注意力机制首先通过压缩通道维度，只留下空间位置信息，然后通过卷积操作对全局区域进行注意力学习，得到包含全局上下文信息的注意力图。通过本文设计的空间注意力模块，网络可以有效的学习到不同区域对目标的增益，从而有效的增强目标识别能力。最后，本文通过级联的方式融合了通道注意力模块和空间位置注意力模块，构成混合注意力模块。

3 损失函数

目标检测既包含分类任务又包含回归任务，因此需要构建多任务损失函数。本文的损失函数定义为定位损失和分类损失加权求和，如下所示。

$$L(x,c,l,g)=\frac{1}{N}(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)) \quad (6)$$

其中超参数α为平衡系数，用于平衡分类损失和定位损失对最终结构的影响，这里根据多次实验选取$\alpha=1$。N是匹配到的默认框数量，如果N=0，则设置损失为0。本文使用框的中心点坐标(cx,cy)和宽(ω)，高(h)四个参数定义一个目标框的图像位置。由于smoothL1相比于直接使用L2回归损失更平滑，因此使用预测框(l)和真实标签(g)之间的smoothL1损失作为定位损失，如公式(7)所示。

$$L_{loc}(x,l,g)=\sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,\omega,h\}}x_{ij}^{k}smooth_{L1}(l_i^m-\hat{g}_j^m) \quad (7)$$

$$\hat{g}_j^{cx}=(g_j^{cx}-d_i^{cx})/d_i^{\omega} \quad \hat{g}_j^{cy}=(g_j^{cy}-d_i^{cy})/d_i^{h}$$

$$\hat{g}_j^{\omega}=lg\frac{g_j^{\omega}}{d_i^{\omega}} \quad \hat{g}_j^{h}=lg\frac{g_j^{h}}{d_i^{h}}$$

分类损失使用softmax多分类损失，如式(8)所示。

$$L_{conf}(x,c)=-\sum_{i\in Pos}^{N}x_{ij}^{p}lg(\hat{c}_i^{p})-\sum_{i\in Neg}lg(\hat{c}_i^{0}) \quad (8)$$

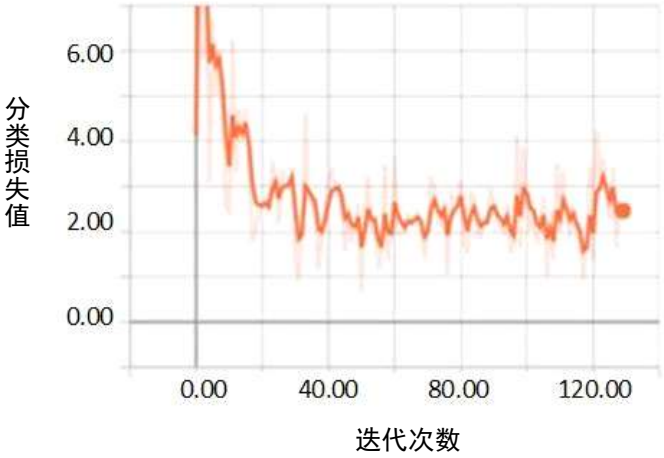
$$\hat{c}_i^{p}=\frac{\exp(c_i^{p})}{\sum_{p}\exp(c_i^{p})}$$

4 实验与结果分析

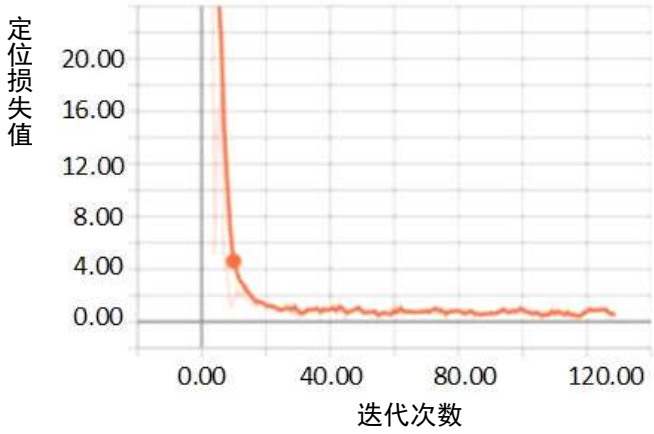
本文在公开的大学教室人群检测数据集：SCUT-HEAD[25]上进行实验。SCUT-HEAD数据集包含两个部分。PartA包含2000张大学教室监控图片，其中标记人头数67321个。PartB包含2405张互联网中下载的图片，其中标记人头数43930个。该数据集采用Pascal VOC标注标准。我们采用PartA部分训练，其中1500张用于训练，500用于测试。训练完成后本文选用查准率P(precision rate)，查全率R(recall rate)，和F1 score指标共同评估本文模型和其他的模型的性能。同时，针对特征融合模块，注意力模块的结构合理性进行了对比实验，验证结构设计的合理性。

4.1 SCUT-HEAD实验

首先进行本文提出的算法和其他常用目标检测算法的性能对比实验。数据集使用SCUT-HEAD PartA和PartB数据集。通过分析主干网络的感受野，设置default box默认尺寸如表1所示。数据增广采用了随机左右镜像、随机亮度和数据归一化三种的方式对数据进行了预处理。训练时，设备使用了1台NVIDIA P100 GPU服务器，基于VGG16作为骨干网络的SSD在MSCOCO数据上预训练的参数开始训练。采用随机梯度下降优化器(SGD)，动量设置为0.9，权重正则衰减系数设置为0.0005，初始学习率设置为1E-3，当训练80,000时，学习率设置为1E-4，当训练20,000后，学习率设置为1E-5最后训练20,000次。网络训练阶段的分类损失和定位损失曲线分别如图11所示。



(a) 分类损失曲线图



(b) 定位损失曲线图

图 11 网络损失曲线图

Fig. 11 Network loss map

表1 默认框基础尺寸设置和理论感受野

Tab. 1 Default box size setting and theoretical receptive field

检测层	步长	候选框尺寸	感受野尺寸
conv4_3	8	32	92
fc7	32	128	420
conv6_2	32	128	452
conv7_2	64	256	516
conv8_2	128	512	644
conv9_2	128	512	772

和其他方法对比结果如表 2 所示。相比于其他算法，本文算法在各个评估指标下均有很高的提升，并且各个性能指标均高于 0.9，在人群检测领域，本文算法达到了最好水平。

表 2 本文方法和其他方法的对比实验结果

Tab. 2 Comparative experimental results of the methods and other method

方法	A 部分			B 部分		
	P	R	F1	P	R	F1
Faster-RCNN	0.86	0.78	0.82	0.87	0.81	0.84
YOLOv3	0.91	0.89	0.89	0.74	0.67	0.70
SSD	0.87	0.68	0.76	0.80	0.66	0.72
R-FCN(ResNet-50)	0.87	0.78	0.82	0.90	0.82	0.86
R-FCN+FRN[17]	0.89	0.83	0.86	0.92	0.84	0.88
MFANet(proposed)	**0.94**	**0.91**	**0.92**	**0.94**	**0.93**	**0.93**

4.2 结构对比实验

本文设计实验验证特征融合模块融合浅层和中间层的合理性，设计了不同的浅层特征图和中间层特征图组合结构进行实验。数据集选用 SCUT-HEAD PartA 部分，所有实验训练配置均相同。如表 3 所示，可以发现使用浅层 conv4_3 至中间层 conv7_2 进行融合，最终性能指标最好，表明了本文特征融合模块结构设计的合理性。

本文设计实验验证新检测分支生成数量设计的合理性。设计了两种不同数目的检测分支结构：第一种是只生成新的 conv4_3 检测分支；第二种是生成新的 conv4_3，fc7 检测分支。数据集选用 SCUT-HEAD PartA 部分，所有实验训练配置均相同。如表 4 所示，可以发现，本文选取的新检测分支数量合理，可以有效的提升算法性能。

表 3 不同融合模块结构对比实验结果

Tab. 3 Comparison of experimental results of different fusion module structures

方法	融合层	P	R	F1
MFANet	conv3_3-conv7_2	0.94	0.89	0.91
MFANet	Conv4_3-conv7_2	**0.94**	**0.91**	**0.92**
MFANet	Conv4_3-conv6_2	0.90	0.88	0.89
MFANet	conv3_3-conv6_2	0.87	0.85	0.86

表 4 新检测分支设计对比实验结果

Tab. 4 New test branch design comparison experiment results

方法	融合层	P	R	F1
MFANet	new conv4_3	0.93	0.90	0.91
MFANet	new conv4_3 fc7	**0.94**	**0.91**	**0.92**

本文设计实验验证混合注意力模块结构设计的合理性，数据集选用 SCUT-HEAD PartA 部分。设计了五种不同的结构:第一种是在检测分支中不增加局部注意力模块,第二种是在检测分支中增加 SENet 的中通道注意力模块，第三种是在检测分支中增加本文设计的通道注意力模块，第四种是在检测分支中增加本文设计的空间注意力模块，第五种是在检测分支中增加本文设计的混合注意力模块。根据表 5 所示的结果可以看出，本文设计的混合注意力机制可以更好的提升网络的性能。

表 5 注意力模块结构对比实验结果

Tab. 5 Attention module structure comparison experiment results

方法	P	R	F1
MFANet（baseline）	0.88	0.86	0.87
MFANet+SEBlock	0.91	0.88	0.89
MFANet+CAM	0.92	0.87	0.89
MFANet+SAM	0.93	0.88	0.90
MFANet+CAM+SAM	**0.94**	**0.91**	**0.92**

4.3 测试结果展示

如图 11 所示，第一行展示的是小目标有遮挡的场景测试结果，第二行展示了既有多尺度目标，也有多姿态目标的一般场景，第三行和第四行分别展示的是前向和后向密集场景测试结果，最后一行展示的是在人工手动标注无法包含全部真实目标的场景下，本文及其他算法的检测结果，通过和 YOLOv3,SSD 等方法的结果对比可以看出，本文提出的 MFANet 很好地解决了目标多尺度，多姿态的检测问题和目标易与环境特征相似的检测问题，并且本文方法在密集型人群中的检测结果性能达到了领先水平

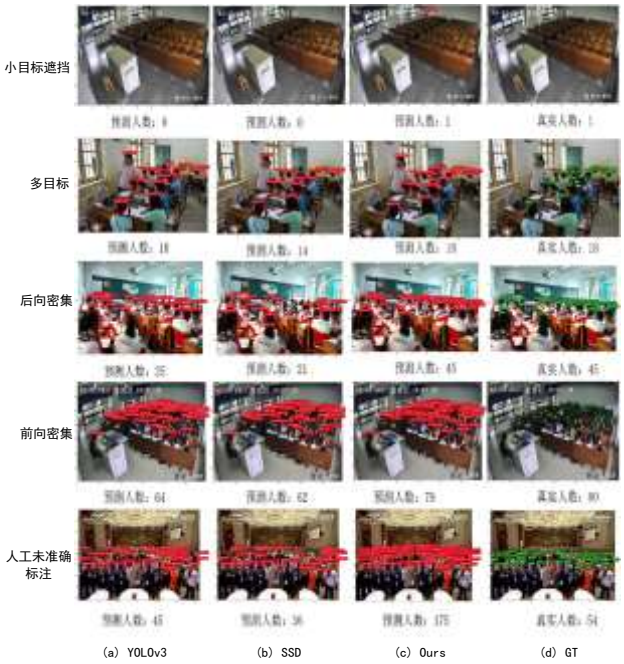


图 12 检测结果和人数统计对比

Fig. 12 Comparison of test and population statistics

5 结语

本文提出了一个新颖的人群目标检测网络 MFANet，主要是用于检测室内人群，并根据检测结果得到最终的人群计数统计。首先设计了浅层和中间层特征融合模块用于解决目标尺寸多样性的问题，然后设计了混合注意力模块用来解决目标区域和周围背景特征混淆的问题，最后采用类 SSD 的单阶段检测框架融合设计的新结构，实现了端到端的训练和预测，在 GPU 上的推理速度达到 25fps，并且在标准数据集上实现了 0.92 的 F1 score 和 0.91 的召回率。并且本文算法灵活简单，同样可以用于其他目标的检测任务中。目前只使用了以轻量级的 VGG16 作为主干网络，使用 ResNet-50，DenseNet 等性能更优的深度网络作为主干网络会更好的提升模型的性能。

参考文献

[1] WANG Q, FAN H J, SUN G, et al. Laplacian pyramid adversarial network for face completion [J]. Pattern Recognition, 2019, 88: 493-505.

[2] YIN X, LIU X M. Multi-task convolutional neural network for pose-invariant face recognition [J]. IEEE Transactions on Image Processing, 2018, 27(2): 964-975.

[3] LU J M, YUAN X, TAKASHI Y. A method of face recognition based on fuzzy clustering and parallel neural networks [J]. Signal Processing, 2006, 86(8): 2026-2039.

[4] ZHANG K, ZHANG Z, LI Z, et al. Joint Face Detection and alignment using multitask cbascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[5] CAO Y P, GUAN D Y, HUANG W L, et al. Pedestrian detection with unsupervised multispectral feature learning using deep neural networks [J]. Information Fusion, 2019, 46: 206-217.

[6] JUNG S I, HONG K S. Deep network aided by guiding network for pedestrian detection[J]. Pattern Recognition Letters, 2017, 90: 43-49

[7] NEUBECK A, GOOL L J V. Efficient non-maximum suppression[C]// Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006). Washington, DC: IEEE Computer Society, 2006:20-24

[8] SINGH B, DAVIS L S. An analysis of scale invariance in object detection-SNIP[C]// Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2018: 3578-3587

[9] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C] // ECCV 2016: Proceedings of the - 14th European Conference on Computer Vision. Berlin: Springer, 2016: 21-37.

[10] LIN T Y, DOLLAR, PIOTR, et al. Feature pyramid networks for object detection[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Washington, DC: IEEE Computer Society, 2017: 936-944.

[11] ZHOU P, NI B B, GENG C. Scale-transferrable object detection[C]// Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). Washington, DC: IEEE Computer Society, 2018: 528-537.

[12] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[C]// Advances in Neural Information Processing Systems 28. Montreal, Quebec, Canada: Neural Information Processing Systems, 2015: 2017-2025.

[13] HU J, SHEN L, SUN G. Squeeze-and-Excitation Networks[C]// Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). Washington, DC: IEEE Computer Society, 2018: 7132-7141.

[14] ZHANG H, GOODFELLOW I J, METAXAS D N, et al. Self-attention generative adversarial networks[C]// Proceedings of the 36th International Conference on Machine Learning (ICML 2019). Long Beach, California: PMLR, 2019:7354-7363.

[15] 李晓光,付陈平,李晓莉,等.面向多尺度目标检测的改进 Faster R-CNN 算法[J]. 计算机辅助设计与图形学学报,2019,31(07): 1095-1101.(LI X G, FU C P, LI X L,et al. Improved faster R-CNN for multi-scale object detection[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(07): 1095-1101.)

[16] 李静,降爱莲.复杂场景下基于 R-FCN 的小人脸检测研究[J/OL].计算机工程与应用:1-12[2019-09-22]. (LI J, JIANG A L. Face detection based on R-FCN IN Complex scenes[J/OL]. Journal of Computer Engineering and Applications:1-12[2019-09-22].)

[17] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Washington, DC: IEEE Computer Society, 2017:6517-6525.

[18] LI J N, LIANG X D, SHEN S M, et al. Scale-aware fast R-CNN for pedestrian detection[J]. IEEE Transactions on Multimedia, 2018, 20(4): 985-996.

[19] ~~SERMANET P, EIGEN D, ZhANG X, et al. OverFeat: integrated recognition, localization and detection using convolutional networks[J]. Eprint Arxiv, 2013.~~

[19] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks[J]. CoRR, abs/1312.6229.

[20] CAI Z W, FAN Q F, FERIS R S, et al. A unified multi-scale deep convolutional neural network for fast object detection[C]\\ ECCV 2016: Proceedings of the 14th European Conference on Computer Vision. Amsterdam,The Netherlands, , Part IV, 2016:354-370.

[21] ~~FU C Y, LIU W, RANGA A , et al. DSSD : Deconvolutional Single Shot Detector[J]. CoRR, 2017。~~

[21] Fu, C., Liu, W., Ranga, A., Tyagi, A., & Berg, A.C. (2017). DSSD : Deconvolutional Single Shot Detector[J]. ArXiv, abs/1701.06659.

[22] 杨康,宋慧慧,张开华.基于双重注意力孪生网络的实时视觉跟踪[J].计算机应用,2019,39(06):1652-1656.(YANG K, SONG H H, ZHANG K H. Real-time visual tracking based on dual attention Siamese network[J]. Journal of Computer Applications, 2019, 39(06): 1652-1656.)

[23] QUAN Y, LI Z X, ZHANG C L. Object detection by combining deep dilated convolutions network and light-weight network[C]// Proceedings of the 12th International Conference on Knowledge Science, Engineering and Management (KSEM 2019). Berlin: Springer, 2019:452-463.

[24] WANG P Q, CHEN P F, YUAN Y, et al. Understanding convolution for semantic segmentation[C]\\ Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV 2018). Washington, DC: IEEE Computer Society, 2018: 1451-1460.

[25] PENG D Z, SUN Z K, CHEN Z R, et al. Detecting heads using feature refine net and cascaded multi-scale architecture[C]\\ Proceedings of the 24th International Conference on Pattern Recognition (ICPR 2018). Washington, DC: IEEE Computer Society, 2018: 2528-2533.

This work is partially supported by the Shanxi Provincial Key Research and Development Plan(201803D31212-1).

SHEN Wenxiang, born in 1995, M. S. candidate. His research interests include deep learning, computer vision.

QIN Pinle, born in 1978, Ph. D., professor. His research interests include big data, machine learning, medical imaging.

ZENG Jianchao, born in 1963, Ph. D., professor. His research interests include machine learning, evolutionary calculation.

变量符号	含义	备注
		无矩阵
		无矢量（向量）

第一项基金项目种类	山西省重点研发计划项目
基金项目编号	201803D31212-1
基金项目名称	监控视频中大角度人像矫正技术研究及应用
基金项目主持人	闫寒梅
基金项目起止时间	2018.7—2020.12
基金项目主要研究内容	
本文内容与项目研究的关系	子课题
第 1 作者在项目承担人中的排名以及在该项目中承担的工作内容	
第 2 作者在项目承担人中的排名以及在该项目中承担的工作内容	排名 2 算法数据库研究
……	
最后一位作者在项目承担人中的排名以及在该项目中承担的工作内容	

非常感谢您提出如此专业和细致的审稿意见和指导！根据审稿专家和编辑老师的意见，我们逐一作了修改，具体修改情况如下：

1. 存在很多文字描述问题，已经在修订稿中标出。参考修改搞，需修改之处，已用不同颜色进行标注，供作者参考。

修改说明：对于文字表达上的错误我们深感抱歉。已经按照要求在新的稿件中改正，并删除了批注。

2. 摘要的第一句话太长，建议修改。

修改说明：在文章第 1 页摘要部分，已经将摘要按照要求重新撰写。具体如下：

（1）摘要第一句已经按“针对……问题，提出了……”的格式修改；

（2）增加“首先、然后、最后”等关键字介绍本文方法。

3. 摘要中提到了算法增强了对小目标的检测能力，这个结论应该在实验中进行验证说明。

修改说明：实验中所选用的数据集都是所鉴定的小目标，因此验证了本文算法对小目标的检测能力。

4. 第 5 页中，“W0 之后先用非线性激活函数 relu 激活”是什么意思？式（3）中并没有 relu 激活函数的作用。

修改说明：具体修改如下

（1）为表述更加清楚，已经将第 5 页“W_0 之后先用非线性激活函数 relu 激活”修改为“W_0 之后使用 ReLU 激活函数来提升网络的非线性程度”，并将 relu 更正为 ReLU。

（2）在公式（3）中，补充了 ReLU。

5. 应该给出网络训练的损失函数演变曲线

修改说明：已经在第 6 页补充了损失函数演变曲线，标为图 11 和图 12。

6. 图 11 中的最后一行，本文算法的结果误差为何那么大？

修改说明：没有表述清楚是我们的过失，已经在图 11 之前的文字描述部分作了修改：“最后一行展示的是在人工手动标注无法包含全部真实目标的场景下，本文及其他算法的检测结果”。

7. 全是英文文献，缺少对国内相关领域的了解，适当补充近几年较新的高质量中文期刊文献。

修改说明：在本文第 8 页参考文献中补充了近几年较新的高质量中文期刊文献，如文献[15][16][22]。