计算机科学与探索
Journal of Frontiers of Computer Science and Technology
ISSN 1673-9418,CN 11-5602/TP

《计算机科学与探索》网络首发论文

题目：　注意力与跨尺度融合的 SSD 目标检测算法
作者：　李青援，邓赵红，罗晓清，顾鑫，王士同
网络首发日期：　2021-03-25
引用格式：　李青援，邓赵红，罗晓清，顾鑫，王士同．注意力与跨尺度融合的 SSD 目标检测算法．计算机科学与探索.
https://kns.cnki.net/kcms/detail/11.5602.TP.20210323.1748.013.html

计算机科学与探索
Journal of Frontiers of Computer Science and Technology doi: 10.3778/j.issn.1673-9418.2102001

注意力与跨尺度融合的 SSD 目标检测算法

李青援[1]，邓赵红[1,2,3,+]，罗晓清[1]，顾　鑫[4]，王士同[1]

1 江南大学 人工智能与计算机学院，江苏 无锡 214122
2 复旦大学计算神经科学与类脑智能教育部重点实验室，上海 200433
3 张江实验室，上海 200120
4 江苏北方湖光光电有限公司，江苏 无锡 214035
+ 通讯作者 E-mail: dengzhaohong@jiangnan.edu.cn

摘　要:为了进一步提升 SSD（Single Shot MultiboxDetector）算法的性能，解决 SSD 算法在进行多尺度预测时特征图信息不平衡和小目标识别难的问题。本文设计了即插即用的模块，充分融合不同尺度特征图包含的信息并建模特征图内的重要性关系，来增强特征图的表示能力。首先，本文设计了一种新颖的特征融合方法来解决跨尺度特征融合存在的信息差异问题。其次，根据池化金字塔的思想设计了一种深度特征提取模块来提取不同感受野的信息，从而来提高模型对不同尺寸目标的检测能力。最后，为了进一步优化特征图，突出特征图对当前任务有效的信息，并建立全局像素点之间的长距离关系和各通道之间的重要性关系，提出了一种轻量级的注意力模块。通过上述机制，本文修改了 SSD 模型的架构，有效地提升了 SSD 算法的检测精度和鲁棒性。在 PASCAL VOC 数据集上设计了丰富的实验，验证了所提方法的有效性，其中，在 PASCAL VOC2007 测试集上本文方法比 SSD 算法提高了 2.9%的平均精确度(mean average precision, mAP)，同时还保留了实时检测的能力。

关键词:目标检测；特征融合；注意力机制;深度学习
文献标志码：A　　**中图分类号：**TP391.41

SSD object detection algorithm with attention and cross-scale fusion

LI Qingyuan[1], DENG Zhaohong[1,2,3+], LUO Xiaoqing[1], GU Xin[4], WANG Shitong[1]

1. School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu 214122, China
2.Key Laboratory of Computational Neuroscience and Brain-like Intelligence, Ministry of Education, Fudan University, Shanghai 200433, China
3.Zhang Jiang Laboratory, Shanghai 200120, China
4.Jiangsu North Huguang Photoelectric Co. Ltd, Wuxi, Jiangsu 214035, China

Abstract: In order to further improve the performance of the SSD (Single Shot Multibox Detector) algorithm, the problem of unbalanced feature map information and difficulty in small target recognition during multi-scale prediction of the SSD algorithm can be solved.In this paper, plug-and-play modulesare designed to fully integrate the in-

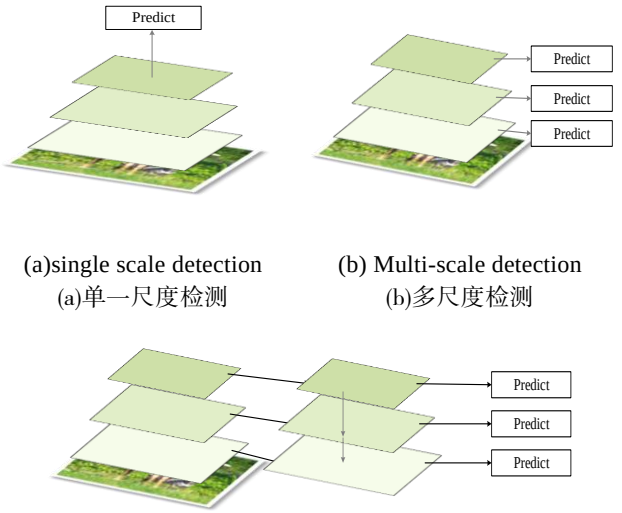
*The National Natural Science Foundation of China under Grant No. 61772239 (国家自然科学基金面上项目),The Shanghai Municipal Science and Technology Major Project under Grant No.2018SHZDZX01(上海市“脑与类脑智能基础转化应用研究”市级重大科技专项).

formation contained in feature maps of different scales and model the relationships within feature maps to enhance the representation ability of feature maps.Firstly, a novel feature fusion method is designed to solve the problem of information disparity in cross-scale feature fusion.Secondly, according to the idea of pooling pyramid, a depth feature extraction module is designed to extract the information of different receptive fields, so as to improve the detection ability of the model to object of different sizes.Finally, in order to further optimize the feature map, highlight the effective information of the feature map for the current task, and establish the global long-distance relationship between pixels and the importance relationship between each channel, a lightweight attention module is proposed.Through the above mechanism, the structure of SSD model is modified in this paper, which effectively improves the detection accuracy and robustness of SSD algorithm. Extensive experiments have been conducted on PASCAL VOC datasets to verify the efficiency of the proposed method. On Pascal VOC2007 test datasets, our method improves 2.9% over SSD algorithm, while maintaining the ability of real-time detection.
Key words:object detection; feature fusion; attentional mechanism; deep-learning

目标检测是计算机视觉的一个关键任务，其任务是给出一张图片，检测出图片中目标物体的边界框，并给出目标的类别。近年来随着深度学习的蓬勃发展，深度卷积网络在目标检测方面取得了显著成功。当前主流的目标检测框架主要有两个分支：两阶段检测方法，包括 R-CNN(regions with CNN features)[1]， FastR-CNN[2]， Faster R-CNN[3]，RefineNet[4]等。一阶段检测方法，包括 YOLO(you only look once)[5-7]，SSD(singleshot multibox detector)[8]，RetinaNet[9]等。两阶段检测方法首先在第一阶段通过一个简单的提议网络产生候选对象位置的稀疏集，然后在第二阶段对候选位置进行分类和回归得到最后的检测结果。一阶段检测方法，通过预先定义一些不同尺度和长宽比的默认框，然后直接对默认框进行分类和回归得到检测结果。由于两阶段检测方法经历了两次分类和回归，其检测精度相对于一阶段检测算法更高，但检测速度远远低于只进行一次分类和回归的一阶段检测方法。

检测尺寸跨度很大的目标是目标检测任务的一大挑战。一些检测器只采用一个尺度的特征图检测，很难检测不同尺寸的目标（如图 1a）。为了实现尺度不变性，图像金子塔和特征金子塔等方法被提出。图像金字塔是将输入图片调整为不同的分辨率，然后将这些图片分别送到网络去检测，这种手工设计特征的方法在传统的检测方法如 DPM[10]中被广泛应用，图像金字塔在一定程度上可以解决尺度变换问题，但其是在每个尺寸图像上分别检测，资源消耗巨大。特征金字塔是将深度卷积模型产生的不同尺度的特征图构成金字塔，然后分别进行检测，这种方法可以避免对图像进行重复运算，极大减少了资源消耗，被主流检测模型广泛采用。SSD 模型是最早尝试将特征金子塔用于目标检测的算法之一（图 1b）。它运用深度卷积网络前向传播产生的不同尺度的特征图进行检测，使用浅层特征图预测小目标，深层特征图预测大目标。

基于自底向上的方法产生的特征图，浅层特征图包含语义信息不足，深层特征图缺少细节信息，结果造成对小目标识别较差。为了解决这一问题，多种特征图融合方法如 FPN(feature pyramid network)[11](图 1c)，PAFPN(path aggression FPN)[12]被提出。它们以自顶向下或自底向上的方法依次将深层特征图与浅层特征图进行融合。这些融合方法是简单有效的，但是由于不同尺度特征图存在较大的信息差异，直接采用相加或通道维度拼接的方法忽略这些差异，容易产生冗余信息和噪声信息,因此这种融合方法是次优的，仍有很大的提升空间。同时，由于小目标在图像中占有较小的像素空间，外观等细节信息模糊，检测较为困难。因此检测小目标的关键因素是采用分辨率较高、包含明显外观细节信息的特征图。另外结合不同感受野的语义信息，利用其所处环境帮助识别也是重要的。



(a)single scale detection
(a)单一尺度检测

(b) Multi-scale detection
(b)多尺度检测

（c）Multi-scale fusion detection
(c)多尺度融合检测

Fig.1　Different types of detection methods
图 1　不同类型检测方法

结合上述分析，本文提出了新的特征增强方法对不同尺度的特征图进行有效融合来增强特征的表示能力。其主要包含如下三个方面的工作。

（1）本文设计了一种特征融合方法。该方法将两个不同尺度的特征图采样到同一维度，然后通过乘法融合和卷积操作产生一个包含两个特征图信息的中间层，该中间层相当于一个模板特征图，其可表示网络学习到的不同尺度特征图之间存在的信息差异。最后，将采样后的特征图与模板特征图再一次融合，从而避免不同尺度特征图直接融合产生冗余信息，并缓解不同尺度特征图的信息差异给融合带来的负面影响。

（2）本文根据池化金字塔[13]的思想设计了一个深度特征提取模块来捕获不同感受野的信息，充分利用局部和全局信息提高对不同尺寸目标的检测能力。为了有效地融合不同尺度的特征图，本文对池化金字塔做了相应的修改。首先，为了减少信息的损失，本文保留了每个分支原始的通道维度，并且在池化操作后经过3×3的卷积以减少池化操作带来的负面影响。其次，不同于池化金字塔模块，该特征提取模块需要指定特征图输出的尺度。各分支的特征图进行池化，并经过上采样以达到目标尺寸，同时，对原特征图进行适应性池化以达到目标尺寸。最后，在相同尺寸下对原特征图和分支产生的特征图进行拼接。

（3）本文设计了一种轻量级的注意力模块，将特征图像素点之间的相似关系与各通道之间的重要性关系进行有效融合，以进一步提升特征图的表示能力，从而帮助模型获取有用信息，并抑制无用信息。

在实验中我们使用 SSD[8]作为基线网络，适当的更改了 SSD 网络的架构并将提出的方法应用到改进的 SSD 框架上，以解决 SSD 算法各预测特征图之间信息不平衡的问题。为了评估方法的性能，我们在 PASCAL VOC[14]数据集上训练并测试了模型。本文主要贡献如下：

首先，设计了轻量级的，有效的深度特征融合模块和非局部通道注意力模块，可应用与任何基于卷积的网络。

其次，把提出的模块应用于 SSD 框架有效地克服了其不足。

最后，在 PASCAL VOC 基准数据集上有效地验证了所提方法的有效性。

本文剩余章节组织如下：第一节，介绍相关工作；第二节，详细介绍本文改进的 SSD 架构以及提出的特征增强模块；第三节，给出实验结果和分析；第四节，对本文进行了总结和展望。

1　相关工作

1.1 目标检测

目标检测包含目标定位和分类。从传统的基于手工设计特征的方法（如 SIFT[15]和 HOG[16]）到基于深度卷积网络的方法，目标检测技术有了巨大发展。最近基于深度卷积网络的检测算法取得了显著成功，一般可以分为两类：基于区域提议的两阶段检测算法和基于回归导向的一阶段检测算法。

两阶段检测算法由两个步骤组成：产生提议区域和对提议区域进行细化调整。首先尝试在目标检

测方面使用深度学习的算法是 R-CNN[1]。R-CNN 使用选择性搜索算法产生提议区域，然后使用卷积网络对每个区域进行特征提取，最后使用线性支持向量机（support vector machine, SVM）预测目标可能存在的位置并对目标进行分类。然而由于 R-CNN 对每一个提议区域都进行卷积，其速度较慢。Fast R-CNN[2]只进行一次特征提取，所以速度比 R-CNN 更快。但是这两个方法仍然把区域提议划分成了单独的阶段。Faster R-CNN[3]将区域提议阶段、特征提取阶段、边界框分类回归阶段整合到一个模型，并可以进行端到端的训练。尤其是区域提议网络（Region Proposal Network，RPN）的提出，进一步提高了检测的速度和精度。两阶段检测算法对于目标特征的学习是十分有效的，但是它们计算效率普遍不高。

不同于两阶段的检测算法，一阶段的检测算法遗弃了区域提议阶段，因此检测速度更快。YOLO[5]提出使用单个卷积网络同时预测多个边界框以及它们的类别概率。YOLO 的速度虽然很快但其检测精度远远低于两阶段检测算法。相对于 YOLO 中直接预测目标中心点的位置，YOLOv2[6]采用地锚框机制更利于检测，极大提高了检测精度。不同于 YOLO 采用单一尺度的特征图进行预测，SSD 算法在主干网络的顶部建立了特征金字塔，利用不同尺度的特征图检测不同尺寸的目标。相对于 YOLOv2，SSD 算法取得了更好的性能。基于 SSD 算法，DSSD(deconvolutional SSD)[17]算法采用编码-解码的方式融合特征图，提升了 SSD 算法的检测精度，但是引入了大量的计算。FSSD(feature fusion SSD)[18]在 SSD 特征金字塔的底部插入一个融合模块以提升 SSD 检测精度，在保证 SSD 检测速度的同时轻微的提升了检测精度。其他的工作，像 RefineDet[4]通过多阶段不断调整锚框的位置来提高检测精度。DSOD(Deeply Supervised Object Detector)[19]探索了如何从零训练一个检测器，并且设计了基于 DenseNet[20]的架构来提高参数利用效率。

1.2 多尺度特征预测

特征金字塔是最近目标检测算法解决跨尺度检测问题的关键技术。SSD 是其中一个最先尝试使用多个不同尺度的特征图分别预测目标的类别和边界框的算法。FPN 通过自顶向下路径和侧面路径循序的结合两个相邻特征图。这种连接有效地增强了特征描述，并且共享了深层特征图包含的丰富语义信息。类似 FPN，PAFPN[12]在 FPN 的基础上添加了一个自底向上的分支，进一步增强了特征描述。Libra R-CNN[21]整合了所有尺度的特征以产生信息更平衡的特征图。ION(Inside-Outside Net)[22]，Hypernet[23]，和 Hypercolumn[24]将不同尺度的特征图进行拼接以提高检测性能。

1.3 视觉注意力网络

注意力在人类感知系统中扮演着重要的角色。人类视觉系统的一个重要特性是，不会尝试同时处理整个场景，而是选择性的聚焦于突出部分，以便更好的捕捉视觉结构。深度学习中的注意力机制可以广义理解为专注于解决特定任务的部分输入，即从众多信息中选择出对当前任务更关键的信息。最近，也有很多工作尝试将注意力机制整合到卷积网络以提高其性能。Hu 等人[25]提出了 Squeeze-and-Excitation 模块，他们运用全局平均池化到特征图以计算每个通道的重要程度，建模通道之间的关系。CBAM(Convolutional Block Attention Module)[26]提出了空间注意力模块和通道注意力模块，结合平均池化和最大池化操作处理特征图，来更好地获取目标的显著特征，并捕获不同空间位置和不同通道之间的重要性。Non-Local[27]，Global Contextlock[28]通过在查询像素点与全局像素点建立关系，来建模像素点之间的长距离关系。

长距离关系可以理解为查询点与其他像素点之间的关系，在卷积神经网络中建立长距离关系的主要方式是通过堆叠卷积层以扩大查询点的感受野从而建立其与感受野内像素的关系。然而这种方式计算效率不高且难以优化。Non-local 网络使用自注意力机制来建模长距离关系，但是因为 Non-Local

模块需要计算每个查询位置的注意力图，随着查询点数量的增多，其计算复杂度会呈二次增长。Global Context Block 通过实验分析证明，No-local 注意力模块在每个查询位置的注意力图几乎是相同的。所以所有查询位置共享同一个注意力图是一种有效的简化方法。

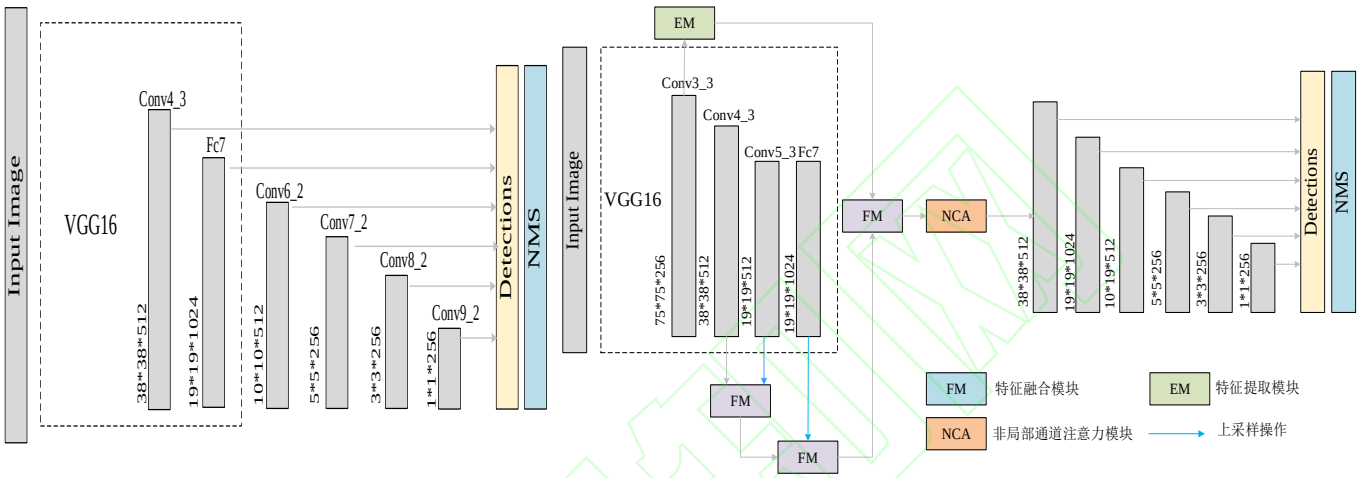


Fig.2 The overall framework of SSD
图 2 SSD 算法整体框架图

Fig.3 The overall framework of improved SSD
图 3 改进的 SSD 算法整体架构图

1.4SSD 算法

SSD 算法是一种十分有效的一阶段检测算法，SSD 算法使用 VGG16[29]作为主干网络，用不同尺度的特征图分别进行检测。SSD 算法预先在不同尺度的特征图上定义不同尺寸和长宽比的锚框，从浅层特征图到深层特征图锚框的尺寸逐渐变大，即用浅层特征图预测小目标，用深层特征图预测大目标，以此来解决检测中存在的目标尺度变化问题。最后，SSD 算法使用 NMS(non-maximum suppression)算法处理不同尺度特征图的检测结果。SSD 算法的整体框架如图 2 所示。

SSD 算法采用不同尺度的特征图分别进行检测，存在浅层特征图语义信息不足，深层特征图细节信息缺失的问题。现有的跨尺度特征图融合方法往往将不同尺度的特征图采样同一尺度，然后采用对应特征图元素相加或者在通道维度拼接的方式融合。其忽略了不同尺度特征图之间存在的信息差异，融合后的特征图可能包含冗余信息或者噪音信息。现有的注意力机制，大多是空间注意力机制与通道注意力机制的结合，空间注意力经常采用拥有较大感受野的卷积层获取局部特征。这不能有效利用像素点之间的关系。针对上述方法存在的优缺点，本文设计了新的特征增强模块。不同于上述提到将相邻两层特征图循序相加融合的方法，本文提出的特征融合模块是以一种选择融合的方式获取两个特征图之间的互补信息，来进行更深度的融合。从而来有效缓解不同尺度特征图的信息差异带来的融合负面影响。同时，为了扩大特征图空间位置的感受野，利用不同感受野的信息提升对小目标的识别能力。本文根据池化金子塔的思想设计了深度特征提取模块，通过多个分支产生不同感受野的特征图并进行整合，以使每个空间位置都有不同的感受野，有效缓解了跨尺度预测问题。进一步地，提出了非局部通道注意力模块，将空间注意力和通道注意力整合为一个轻量级模块，可有效捕获通道之间的重要性关系，并在每个查询点与全局像素点之间建立了长距离关系。

2 本文方法

卷积神经网络可以有效的提取图片的金字塔形式特征，从浅层特征到深层特征语义信息越来越丰富。SSD 算法直接使用不同尺度的特征图进行检测，这使得 SSD 算法不能有效地将包含丰富细节信息的低级特征图与包含丰富语义信息的高级特征图进行融合，从而 SSD 算法的检测结果是次优的。尤其是对于小目标，小目标在图像中占有很少的像素，因此小目标的识别需要充分结合其外观细节信息和语义信息。所以有效融合不同尺度的特征对于提高基于卷积网络的检测器的性能是十分有必要的。

本文改进了传统的 SSD 算法，主要思路是通过运用提出的方法产生一个具有丰富表征能力的特征图，然后将该特征图通过一个特征金字塔提取层，以产生多尺度的特征图来检测不同尺寸的目标。改进的 SSD 架构可以提高浅层特征图语义信息不足的缺点同时丰富了深层特征图的细节信息，极大提高了检测性能。

改进的整体架构如图 3 所示。具体改进方法如下，我们设置 Conv4_3 层为优化层，使用其他层来增强 Conv4_3 层的表示能力。首先使用特征融合模块，将 VGG 主干网络产生的 Conv4_3 层特征图与上采样后的 Conv5_3 层特征图融合，融合后的特征图再与上采样后的 Fc7 层融合，以进一步提高 Con4_3 层特征图的语义信息。同时，为了充分利用浅层特征图的细节信息，本文利用特征提取模块对 Conv3_3 层进行特征提取。设置特征提取模块的目标尺寸为两倍下采样大小，使用特征融合模块将 Conv3_3 层提取到的特征图与融合后的 Con4_3 层进一步融合，以产生更具语义和细节的特征图。然后将该特征图经过非局部通道注意力模块，来突出有用信息，抑制噪音等无效信息。最后经过一个特征金字塔层（如图 7 所示）。提取不同尺度的特征图以进行不同尺寸目标的检测。

各模块的详细描述将在下面小节介绍。

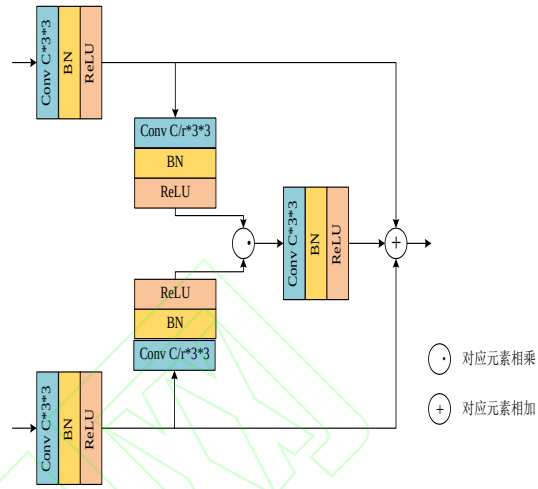


Fig.4 Feature mapfusion module

图 4 特征融合模块

2.1 深度特征融合模块

融合不同尺度的特征图已经被证明可以有效地提高检测性能，如 FPN[11]，DSSD[17]，PAFPN[12]。由于不同尺度的特征图有不同大小的感受野，这些不同尺度的特征图存在大的信息差异，常规地融合手段是将两个不同尺度的特征图采样到同一尺度然后采取直接相加或在通道维度拼接的方法对两个特征图进行融合，这种做法忽略了不同尺度特征图之间存在的信息差异，会产生次优的融合结果。

针对上述问题，我们设计了一个跨尺度特征融合模块，有效地融合深层特征图$f_h \in R^{H\times W\times C}$和浅层特征图$f_l \in R^{h\times w\times c}$。由于浅层特征图具有较小的感受野，$f_l$包含丰富的细节信息，同时也保留了一些背景噪音。相反由于多次下采样操作，f_h丢失了许多细节信息但却有着丰富的语义信息。这两种特征图之间存在大的信息差异。跨尺度特征融合模块首先将两个特征图采样到同一尺度，然后再将这两个特征图经过各自的降维模块以减少后续融合时的计算量。降维之后的两个特征图求哈达玛积并经过特征迁移层，产生一个对两个特征图有全局综合描述的中间模板特征图，然后将其与原始的特征图f_l，f_h采用按元素相加的方式融合。f_l，f_h会根据模板特征图的信息进行互补融合，吸收彼此之间有用的信息来产生互补作用，即f_l的噪音将被抑制，f_h的

语义将被增强。相比于之前方法直接将两个特征图按元素相加或者拼接，我们的融合方法可以避免引入一些对检测结果带来负面影响的冗余信息。该跨尺度融合模块包含两个分支，如图 4 所示，一条分支用来细化f_l，另一条分支用来细化f_h，细化后的两个特征图求哈达玛积后再经过特征迁移层可以产生一个包含两者信息的中间层。网络会学习到两个特征图之间的差异信息，最后原始特征图与模板特征图的融合相当于互补选择的过程。

整个过程可以用下列公式描述：

$$f_h' = T_1(f_h) \#(1)$$

$$f_l' = T_2(f_l) \#(2)$$

$$f = T_3\left(D_1(f_h') \circ D_2(f_l')\right) + f_h' + f_l' \#(3)$$

f_h'表示经过特征迁移后的高级特征图，f_l'表示经过特征迁移后的低级特征图，$\circ$表示哈达玛积，f表示融合后的特征图，$T(\cdot)$表示卷积层、BN 层和 ReLU 层结合的特征迁移层，$D(\cdot)$表示降维比例为 r 的特征迁移层。

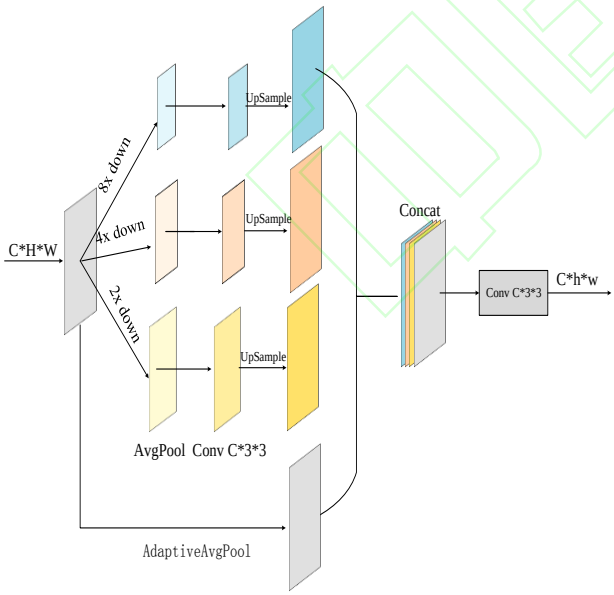


Fig.5　Feature mapextraction module

图 5　特征提取模块

2.2 深度特征提取模块

语义信息对复杂场景以及小目标的识别是重要的。例如，当我们通过目标的形状等外观信息无法判断目标是什么类别的时候，可以结合其所处的语义环境来帮助判断。在深度卷积网络模型中，可以用感受野的尺寸粗略描述模型利用了多少语义信息。Zhou 等人[30]证明了 CNN 网络的实际感受野远远小于理论感受野，尤其是在更深的卷积层中。因此为了获取不同感受野的语义信息，更进一步提取丰富的特征描述，我们根据池化金字塔的思想设计了一个深度特征提取模块，通过该模块可以让特征图的每一个空间位置看到不同大小空间的语义信息，进一步扩大特征图的感受野。深度特征提取模块架构如图 5 所示。其包含三个分支，每一个分支先经过一个指定大小池化核的平均池化操作，并对下采样后的特征图经过 3×3 的卷积以减少池化操作产生的信息偏差。然后将三个不同尺度的特征图采用双线性插值的方法上采样到目标尺寸，并将输入的特征图也采样到目标尺寸。最后将这些特征图在通道维度进行拼接，拼接后的特征图再经过 3×3的卷积层使信息充分融合。深度特征提取模块分支的数量以及池化核大小都可以更改。本文中三个分支的池化核大小各自为 2，4，8。

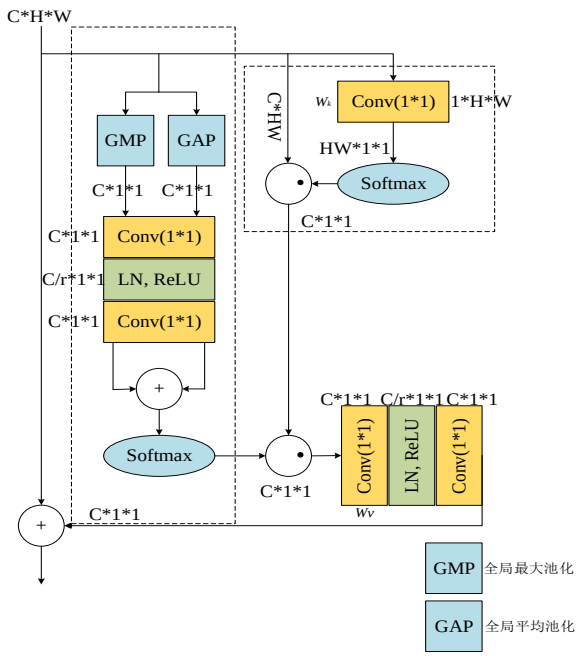


Fig.6　Non-Local channel attentional mechanism module

图 6　非局部通道注意力机制模块

2.3 非局部通道注意力机制

空间注意力机制可以基于特征图全局的关系，得到特征图中每个位置的相关性。强调网络感兴趣的部分，抑制背景等无用部分。通道注意力机制，可以结合特征图通道之间的关系，建模各通道之间的重要程度，通过通道注意力机制网络可以获得更多通道间的信息。非局部通道注意力详细结构如图 6 所示。该模块主要包含两部分，一部分用来建模各通道之间的重要性关系，另一部分用来建模像素点之间的长距离关系。其中长距离关系分支采用 Global Context Block 的全局注意力池化部分。它采用嵌入高斯$w_{ij}=\frac{\exp(<w_q x_i, w_k x_j>)}{\sum_m \exp(<w_q x_i, w_k x_m>)}$计算像素点之间的相似度。非局部通道注意力机制主要分为以下步骤，对于输入的特征图$F \in R^{C\times W\times H}$：(a)在像素点间的长距离关系部分，对 F 先采用1×1的卷积W_k和 SoftMax 函数得到注意力权重$F_w \in R^{HW\times 1\times 1}$,然后再与键值项特征$F_k \in R^{C\times HW}$相乘得到长距离关系的全局特征$z \in R^{C\times 1\times 1}$。(b)$F$ 经过全局最大池化和全局平均池化产生特征图F_M, F_A；F_M, F_A经过同一个迁移层网络产生F_M', F_A'。F_M', F_A'对应元素相加后经过 Sigmoid 激活产生通道注意力图。(c)将步骤 a 和步骤 b 产生的注意力图进行对应元素相乘融合，然后经过特征迁移层后得到非局部注意力特征图。(d)将原始的特征图与非局部注意力特征图相加。以上过程可以用下面公式表示：

$$u = T\left(\sigma\left(M(x)+G(x)\right)\right)\#(4)$$

$$z = W_v \sum_{j=1}^{N_p} \frac{\exp(W_k x_j)}{\sum_{m=1}^{N_p}\exp(W_k x_m)} x_j \#(5)$$

$$f' = x + T(u \circ z)\#(6)$$

其中x是输入的特征图，u是产生的通道注意力图，z是产生的长距离关系注意力图，$N_p = H\cdot W$是特征图中像素点的数量。$T(\cdot) = RelU(BN(conv2d(x)))$是特征迁移层。$\sigma = \frac{1}{1+e^{-x}}$是 sigmoid 激活函数$C(\cdot)$，$M(\cdot)$，$G(\cdot)$分别表示卷积，全局最大池化和全局平均池化。$\circ$表示哈达玛积。

全局最大池化可以提取特征图最显著的内容捕获目标特征的差异信息。全局平均池化可以捕获特征图全局的综合信息。融合全局最大池化和全局平均池化产生的特征图对建立特征图通道之间的关系有重要意义。

通道注意力图可以指导检测器更应该关注哪一个通道。长距离关系注意力图可以有效地建立查询点与全局像素点的联系，以获取更全面的空间相关信息。通过融合通道注意力图和空间注意力图可以提高原始特征图的表征能力，让每个空间位置和每个通道产生联系，有效地提高检测能力。

Fig.7 Pyramid feature generators layers
图 7 特征金字塔产生层

Table 1 Comparison of detection accuracy of 20 categories in the PASCAL VOC2007test dataset
表 1 PASCAL VOC2007test 数据集下 20 类别检测精度对比

模型	mAP/%	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mike	person	plant	sheep	sofa	train	tv
SSD	77.5	79.5	83.9	76.0	69.6	50.5	**87.0**	85.7	88.1	60.3	81.5	77.0	86.1	87.5	84.0	79.4	51.7	77.9	79.5	**87.6**	76.8
DSSD	78.6	81.9	84.9	**80.5**	68.4	53.9	85.6	86.2	88.9	61.1	83.5	78.7	86.7	**88.7**	**86.7**	79.7	51.7	78.0	80.9	87.2	79.4
ION	79.2	80.2	85.2	78.8	70.9	**62.6**	86.6	86.9	**89.8**	61.7	**86.9**	76.5	**88.4**	87.5	83.4	80.5	52.4	78.1	77.2	86.9	**83.5**
Our	**80.4**	**84.9**	**87.0**	79.5	**75.5**	59.5	86.7	**87.4**	89.0	**67.5**	85.0	**80.0**	86.6	88.0	86.2	**81.9**	**57.1**	**79.1**	**81.1**	86.3	79.7

Table 2 Comparison of detection accuracy of 20 categories in the PASCAL VOC2012test dataset
表 2 PASCAL VOC2012test 数据集下 20 类别检测精度对比

模型	mAP/%	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mike	person	plant	sheep	sofa	train	tv
SSD512	76.7	88.8	84.8	77.0	61.0	56.3	82.6	82.4	92.6	58.4	80.7	61.4	90.4	**87.2**	86.9	85.0	53.1	81.2	65.9	**86.4**	72.0
Our512	**78.5**	**91.0**	**87.9**	**79.8**	**63.6**	**60.3**	**84.6**	**83.5**	**92.8**	**60.9**	**82.2**	**64.3**	**91.2**	86.8	**88.3**	**87.1**	**57.1**	**85.1**	**66.1**	84.0	**73.8**

Table 3 Comparison of detection speed and accuracy under the PASCAL VOC2007test dataset
表 3 PASCAL VOC2007test 数据集下检测速度和检测精度对比

算法	网络	检测速度(fps)	GPU	锚框个数	输入尺寸	mAP/%
Faster RCNN[3]	VGG-16	7	Tian X	6000	~600 × 1000	73.2
Faster RCNN[3]	ResNet-101	2.4	K40	300	~600 × 1000	76.4
R-FCN[31]	ResNet-50	-	-	300	~600 × 1000	77.0
R-FCN[31]	ResNet-101	5.8	K40	300	~600 × 1000	79.5
YOLOv2[6]	Darknet-19	81	Tian X	-	352 × 352	73.7
SSD300[8]	VGG-16	92	2080Ti	8732	300 × 300	77.5
FSSD300[18]	VGG-16	65.8	1080Ti	8732	300 × 300	78.8
RefineDet320[4]	VGG-16	12.9	K80	6375	320 × 320	79.5
RSSD300[32]	VGG-16	35	Tian X	8732	300 × 300	78.5
DSSD321[17]	ResNet-101	9.5	Tian X	17080	321 × 321	78.6
ASSD300[33]	VGG-16	11.8	K40	8732	300 × 300	80.0
SSD512[8]	VGG16	45	2080Ti	24564	512 × 512	79.5
DSSD513[17]	ResNet-101	5.5	Tian X	43688	513 × 513	81.5
FSSD512[18]	VGG16	35.7	1080Ti	24564	512 × 512	80.9
RSSD512[32]	VGG16	16.6	Tian X	24564	512 × 512	80.8
ASSD512	VGG-16	3.4	K40	24564	512 × 512	81.6
RefineDet512[4]	VGG-16	5.6	K80	16320	512 × 512	81.2
Our300	VGG-16	44.8	2080Ti	8732	300 × 300	80.4
Our512	VGG-16	22.5	2080Ti	24564	512 × 512	82.2

3 实验结果与分析

为了评估提出方法的有效性，我们在 PASCAL VOC 数据集上设计了丰富的实验进行验证。在 PASCAL VOC 数据集上，如果预测框与真实框的交并比（IOU）大于 0.5 则预测结果是正确的。我们采用平均精确度（11 point mAP）作为评价指标。

3.1 数据集

PASCAL VOC 数据集包括 20 个类别，即 aero、bike、bird、boat、bottle、bus、car、cat、chair、cow、table、dog、horse、mbike、person、plant、sheep、sofa、train、tv。我们使用 PASCAL VOC2007trainval 和 VOC2012 trainval 训练模型，使用 PASCAL VOC2007 test 和 PASCAL VOC2012test 数据集测试模型。训练集一共 16551 张图片，VOC2007 测试集一共 4952 张图片，PASCAL VOC2012 测试集包含 10991 张图片。

3.2 实验设置

本文应用提出的方法到 SSD 框架，并基于 Pytorch 框架实现了模型。本文使用 VGG16 作为主干网络，实验中所有上采样操作后会经过 3×3 卷积和 ReLU 激活，Fc7 层上采样后的卷积层通道数调整为 512 以便后续融合。深度特征融合模块和非局部注意力模块瓶颈层的降维比率均设为 1/4。训练策略，数据增策略以及损失函数和锚框参数均与原始 SSD 一致。本文使用 RTX-2080Ti 显卡进行实验。对于 300×300 输入分辨率图片，Batch size 设为 16，初始学习率设为 0.001，120000 次和 140000 次迭代后学习率依次下降十倍，迭代 180000 次得到最终的网络模型。对于 512×512 输入分辨率图片，Batch size 设为 8，初始学习率 0.001，迭代 140000 次和 160000 次后学习率依次下降 10 倍，总共迭代 240000 次得到最终模型。

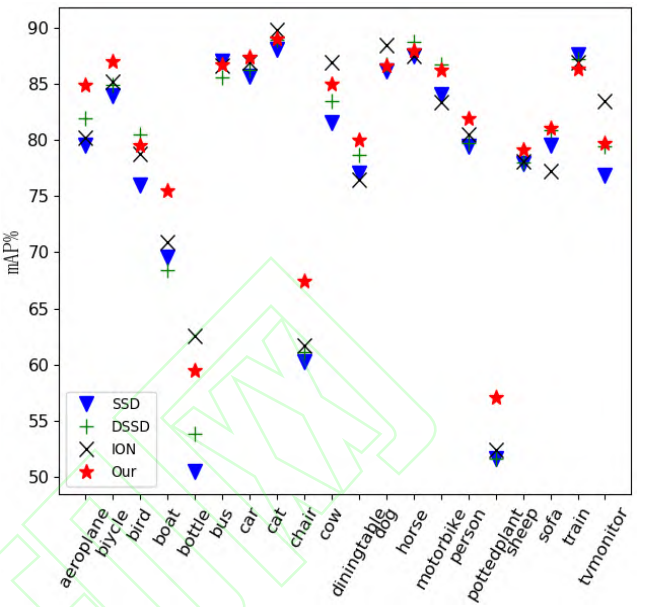


Fig.8 Comparison of four detection algorithms MAP on PASCAL VOC 2007 test dataset

图 8 PASCAL VOC 2007test 数据集上四种检测算法 mAP 对比

3.3 实验结果与分析

本文在各个类别的检测结果上与 SSD300、DSSD321、ION300 算法进行了对比。各类别检测精度详细结果如表 1 所示。比较四种算法在 PASCALVOC2007 测试集上的检测结果可以看出，应用本文提出的模块对 SSD 算法进行改进可以极大提高检测精度，各个类别相对与其他算法有了明显的精度提升，mAP 比原始的 SSD 算法提高了 2.9%。对比结果如图 8 所示。同时，为了进一步评估所提算法的有效性，在相同的实验环境下，将 SSD512 模型与本文模型对 VOC2012 测试集的预测结果分别提交至 PASCAL VOC 官方评测网站进行评测，评测结果如表 2 所示。从表中可以看出在 PASCALVOC2012 测试集上本文方法在各类别检测精度上相对于原始 SSD 算法有了显著提升。

为了比较提出方法与其他主流检测算法的差异，我们进一步在 PASCAL VOC2007 测试集上对比了多种算法的检测精度和检测速度，具体比较结果如表 3 所示。从表中可以看出本文算法与其它算法相比，在输入分辨率相近的情况下检测精度更好，且能达到实时的检测效果。

为了更直观地评价本文算法，图 9 给出了 SSD 模型和本文模型在 PASCALVOC 2007 测试集下部分图片检测结果对比。两个模型均在 VOC07+12 数据集上训练，输入图片分辨率均为300 × 300。通过对比可看出本文算法在复杂场景下鲁棒性更好，对于 SSD 算法没有检测出的目标，本文算法可有效检测，且对遮挡目标以及小目标的识别优于原始 SSD 算法。

3.4 消融实验

为了验证各模块的有效性，本文设计了一系列控制变量实验，测试在添加或不添加提出的方法下模型检测精度和检测速度的差异。实验环境与 4.2 节相同。我们在 PASCAL VOC2007 测试集下进行了实验，实验结果如表 4 所示。SSD*表示只改变原始 SSD 的架构不使用任何本文提出的方法。实验中 Conv3_3 层直接经过一个 3 × 3 卷积调整维度到 512 并下采样到 38 × 38。Conv5_3 上采样后直接与 Conv4_3 相加融合然后 Fc7 层经过上采样后再与 Conv4_3 相加融合，最后再与调整维度后的 Conv3_3 相加融合。SSD*+EM 表示在 SSD*的基础上 Conv3_3 层使用的特征提取模块。+FM 表示相加融合改为所提的特征融合。+NCA 表示融合后的特征图经过非局部通道注意力模块。通过观察表 4 可以看出在改进版的 SSD 框架中逐步加入提出的方法可以明显提高检测精度，尤其是使用深度特征融合模块代替相加融合操作可以显著提高检测精度。非局部注意力模块在略微损失检测速度的前提下提升了 0.7%的检测精度。当将三个模块全部加到改进版的 SSD 框架中可达到最好的检测精度。

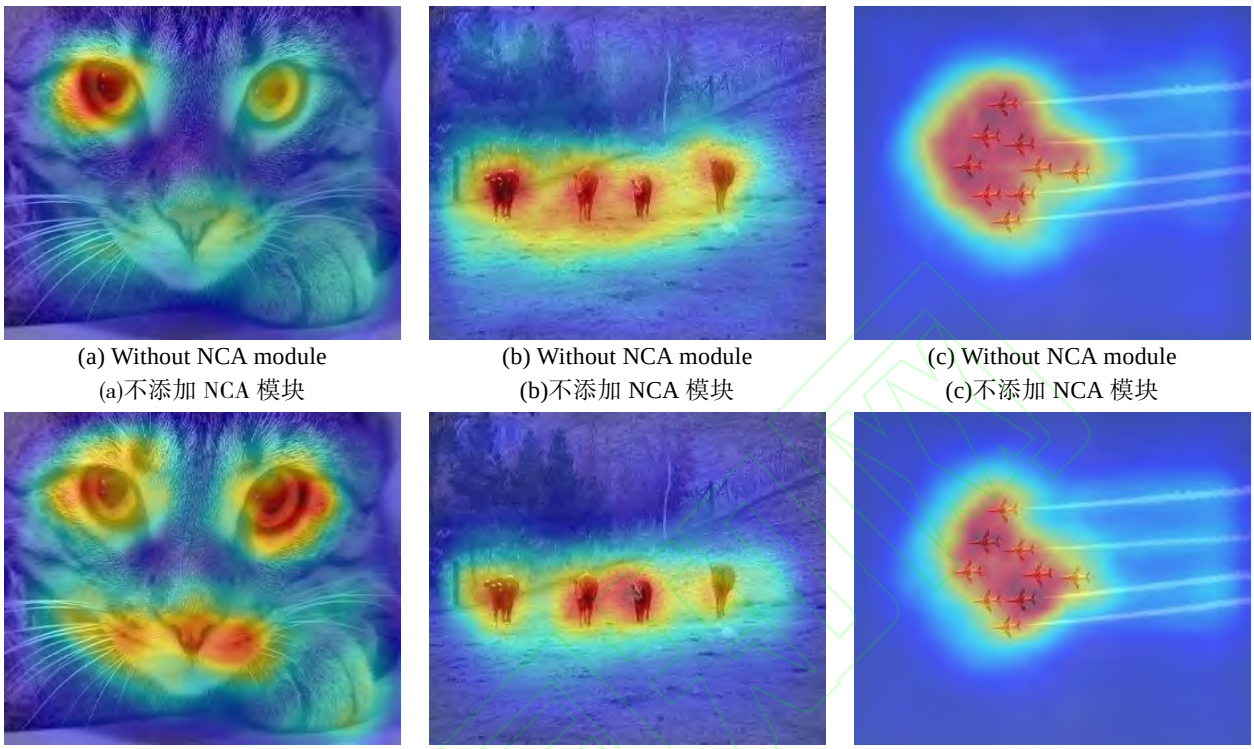
Table 4　Comparative results of ablation experiments
表 4　消融实验对比结果

方法	检测速度（FPS）	mAP/%
SSD	92	77.5
SSD*	77	78.1
SSD*+EM	69.3	78.5
SSD*+EM+FM	46.7	79.7
SSD*+EM+FM+NCA	44.8	80.4

(a)The detection result of SSD　(b)The detection result of SSD　(c)The detection result of SSD　(d)The detection result of SSD
(a)SSD 检测结果　(b)SSD 检测结果　(c)SSD 检测结果　(d)SSD 检测结果

(e)The detection result of ours　(f)The detection result of ours　(g) The detection result of ours　(h) The detection result of ours
(e)本文检测结果　(f)本文检测结果　(g)本文检测结果　(h)本文检测结果

Fig.9　The detection result comparaed of SSD and ours
图 9　本文与 SSD 算法检测结果对比



(a) Without NCA module
(a)不添加 NCA 模块

(b) Without NCA module
(b)不添加 NCA 模块

(c) Without NCA module
(c)不添加 NCA 模块

(d)With NCA module
(d)添加 NCA 模块

(e)With NCA module
(e)添加 NCA 模块

(f)With NCA module
(f)添加 NCA 模块

Fig.10 Visualization of attention maps
图 10 注意力图可视化

为了更直观地说明非局部通道注意力模块的有效性，本文使用 Grad-CAM[34]可视化了 PASCAL VOC2007 测试集的部分图片，如图 10 所示。Grad-CAM 是最近提出的一种可视化方法，它使用梯度来计算特征图空间位置的重要性。颜色越深说明该区域对类别识别的影响越大。通过观察图 10 可以发现，非局部通道注意力机制可以进一步突出对当前目标任务有利的信息，并且有效抑制了背景等对检测没有帮助的信息，因此可以推断非局部通道注意力机制对检测精度的提升有很大帮助。

4 结论和展望

针对传统 SSD 算法存在的多尺度独立预测造成的各尺度信息不平衡，以及鲁棒性差，小目标识别差等问题。本文设计了新的架构以解决这些问题。首先设计了一种新的特征图融合方法，有效地融合了不同尺度特征图，同时缓解了不同尺度特征图之间的信息差异带来的融合负面影响。其次为了增大特征图的感受野，充分利用不同感受野下的语义信息，进一步提高特征图的表示能力，本文根据池化金字塔的思想设计了深度特征提取模块，利用不同尺寸的池化核获取不同大小的感受野，以使特征图可以利用不同感受野的语义信息，从而提高对目标的识别能力。本文还设计了一种非局部通道注意力机制，将像素点的长距离关系注意力和通道关系注意力融合到一个轻量级模块，突出特征图对当前任务有效的信息，抑制背景等无效信息，进一步提升了检测精度。最后基于上述方法本文改进 SSD 算法的架构，改进的 SSD 算法有了大幅度的检测精度提升，同时还保留了实时性。本文设计的所有模块均可用于基于卷积的神经网络，即插即用。未来我们将继续改进各个模块的架构以在保持精度提升的前提下，进一步减少模块的计算量，以使用较少的速度损失获取更多的精度提升。

References：

[1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [M]. CVPR. 2014.

[2] GIRSHICK R. Fast R-CNN [J]. Computer Science, 2015,

[3] 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(6): 1137-49.

[4] RAJARAM R N, OHN-BAR E, TRIVEDI M M. RefineNet: Iterative refinement for accurate object localization [M]. IEEE International Conference on Intelligent Transportation Systems. 2016.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection [M]. Computer Vision & Pattern Recognition. 2016.

[6] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger [M]. IEEE Conference on Computer Vision & Pattern Recognition. 2017.

[7] REDMON J, FARHADI A. YOLOv3: An Incremental Improvement [J]. arXiv e-prints, 2018,

[8] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector [M]. ECCV. 2016.

[9] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal Loss for Dense Object Detection [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99): 2999-3007.

[10] FELZENSZWALB, PEDRO, F., et al. Object Detection with Discriminatively Trained Part-Based Models [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 32(9): 1627-45.

[11] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature Pyramid Networks for Object Detection [M]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[12] LIU S, QI L, QIN H, et al. Path Aggregation Network for Instance Segmentation [J]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 8759-68.

[13] ZHAO H, SHI J, QI X, et al. Pyramid Scene Parsing Network [J]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 6230-9.

[14] EVERINGHAM M, ESLAMI S, GOOL L, et al. The Pascal Visual Object Classes Challenge: A Retrospective [J]. International Journal of Computer Vision, 2014, 111(98-136.

[15] LOWE D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[16] DALAL N, TRIGGS B. Histograms of Oriented Gradients for Human Detection [M]. IEEE Computer Society Conference on Computer Vision & Pattern Recognition. 2005.

[17] FU C-Y, LIU W, RANGA A, et al. DSSD : Deconvolutional Single Shot Detector [J]. ArXiv, 2017, abs/1701.06659(

[18] LI Z, ZHOU F. FSSD: Feature Fusion Single Shot Multibox Detector [J]. ArXiv, 2017, abs/1712.00960(

[19] SHEN Z, LIU Z, LI J, et al. DSOD: Learning Deeply Supervised Object Detectors from Scratch [J]. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, 1937-45.

[20] HUANG G, LIU Z, WEINBERGER K Q. Densely Connected Convolutional Networks [J]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 2261-9.

[21] PANG J, CHEN K, SHI J, et al. Libra R-CNN: Towards Balanced Learning for Object Detection [M]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.

[22] BELL S, ZITNICK C L, BALA K, et al. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks [M]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[23] KONG T, YAO A, CHEN Y, et al. HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection [J]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 845-53.

[24] HARIHARAN B, ARBEL áEZ P, GIRSHICK R B, et al. Hypercolumns for object segmentation and fine-grained localization [J]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, 447-56.

[25] HU J, SHEN L, ALBANIE S, et al. Squeeze-and- Excitation Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2011-23.

[26] WOO S, PARK J, LEE J-Y, et al. CBAM: Convolutional Block Attention Module [M]. ECCV. 2018.

[27] WANG X, GIRSHICK R B, GUPTA A, et al. Non-local Neural Networks [J]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 7794-803.

[28] CAO Y, XU J, LIN S, et al. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond [J]. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, 1971-80.

[29] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science, 2014,

[30] ZHOU B, KHOSLA A, LAPEDRIZA À, et al. Object Detectors Emerge in Deep Scene CNNs [J]. CoRR, 2015, abs/1412.6856(

[31] DAI J, LI Y, HE K, et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks [J]. ArXiv, 2016, abs/1605.06409(

[32] JEONG J, PARK H, KWAK N. Enhancement of SSD by concatenating feature maps for object detection [J]. ArXiv, 2017, abs/1705.09587(

[33] YI J, WU P, METAXAS D. ASSD: Attentive Single Shot Multibox Detector [J]. Comput Vis Image Underst, 2019, 189(

[34] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-59.

LI Qingyuan was born in 1997. He is a master candidate at School of Artificial Intelligence and Computer Science, Jiangnan university. His research interests isdeep learning.
李青援(1997-)，男，山东潍坊人，江南大学人工智能与计算机学院研究生，主要研究领域为深度学习。

DENG Zhaohong was born in 1981. He is a professor at School of Artificial Intelligence and Computer Science, Jiangnan University. His research interests include uncertainty artificial intelligence and its applications.
邓赵红(1981-)，男，安徽蒙城人，江南大学人工智能与计算机学院教授，主要研究领域为不确定性人工智能及其应用。CCF 高级会员。

LUO Xiaoqing was born in 1980. She is an associate professor at school of Artificial Intelligence and Computer Science, Jiangnan University. Her research interests are image fusion, pattern recognition, and other problems in image technologies.
罗晓清(1980-)，女，江西南昌人，江南大学人工智能与计算机学院副教授，主要研究领域为图像融合、模式识别和图像处理等。

GU Xin was born in 1979. He is a senior engineer at Jiangsu North Huguang Opto-Electronics Co.Ltd.His research interests include pattern recognition and artificial intelligence image processing technology and itsapplication.
顾鑫（1979-），男，江苏张家港人，江苏北方湖光光电有限公司高级工程师、博士，主要研究领域为模式识别、人工智能图像处理技术研究与应用。

WANG Shitong was born in 1964. He is a professor at School of Artificial Intelligence and Computer Science, Jiangnan University. His research interests include artificial intelligence and pattern recognition, etc.
王士同(1964-)，男，江苏扬州人，江南大学人工智能与计算机学院教授，主要研究领域为人工智能和模式识别等。CCF 高级会员。