第2卷 第2期 2020年4月
Vol.2 No.2 Apr.2020

ISSN 2096-742X
CN 10-1649/TP

文献DOI：
10.11871/jfdc.issn.
2096-742X.2020.
02.010

文献PID：
21.86101.2/jfdc.
2096-742X.2020.
02.010

页码：120-135

开放科学标识码
（OSID）

基于深度学习的小目标检测与识别

冷佳旭[1,2]，刘莹[1,2*]

1. 中国科学院大学计算机科学与技术学院，北京 100089
2. 中国科学院大学数据挖掘与高性能计算实验室，北京 101400

摘　要：【目的】目前，现有的基于深度学习的检测算法针对小目标的检测效果较差。本文旨在通过充分考虑小目标的特点来提升小目标的检测与识别性能。【方法】本文从不同方面来提升小目标检测与识别，其中包括特征融合、上下文学习和注意力机制。针对小目标特征难以提取问题，提出一种双向特征融合的方法。另外，鉴于小目标特征不明显问题，提出一种利用上下文信息来提升检测性能的方法。更进一步，为了更好地识别小目标的类别，提出一种注意力转移的方法。【结果】实验结果表明，我们提出的方法在公共数据集上均显著地提高了小目标的检测和识别性能。【结论】研究特征融合、上下文利用和注意力机制的方法对于提升小目标检测与识别是非常有价值的。

关键词： 小目标检测；特征融合；上下文学习；注意力机制

Small Object Detection and Recognition Based on Deep Learning

Leng Jiaxu [1,2], Liu Ying [1,2*]

1. School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100089, China
2. Data Mining and High Performance Computing Lab, University of Chinese Academy of Sciences, Beijing 101400, China

Abstract: [**Objective**] In this paper, we aim to improve the detection performance for small objects by considering the characteristics of small objects under deep learning-based detection frameworks. [**Methods**] This paper improves small object detection and recognition performance from different aspects, including feature fusion, context learning and attention mechanism. Since the features of the small object are not evident, a bidirectional feature fusion method is proposed to improve the feature expression capability for small objects. In addition, a novel method is proposed to improve the detection performance by using the context information of small objects. Furthermore, to better identify the categories of small objects, an attention transfer method is proposed to improve the recognition rate. [**Results**] Experimental results show that the three proposed methods can significantly improve the detection and recognition performance for small objects on public datasets. [**Conclusions**] The research on feature fusion, context utilization and attention mechanism is very valuable for improving small object detection in complex scenes.

Keywords: small object detection; feature fusion; context learning; attention mechanism

基金项目：国家自然科学基金(71671178，91546201)；中国科学院大学优秀青年教师科研能力提升重点项目

*通讯作者：刘莹（E-mail: yingliu@ucas.ac.cn）

引言

目标检测是计算机视觉领域的一个重要研究方向，几十年来也一直都是一个活跃的研究难题。如图 1 所示，给定一张图像，目标检测的任务是找出图像中感兴趣的区域，确定目标的位置和大小，并且判断目标所属类别。作为图像理解和计算机视觉的基石，目标检测是解决分割、场景理解、目标追踪、图像描述、事件检测和活动识别等更复杂更高层次的视觉任务的基础。随着深度学习的快速发展，目标检测算法 [1-5] 也取得了重大突破。目标检测在人工智能和信息技术的许多领域都有广泛的应用，包括机器人视觉、消费电子产品、安保、自动驾驶、人机交互、基于内容的图像检索、智能视频监控和增强现实等。

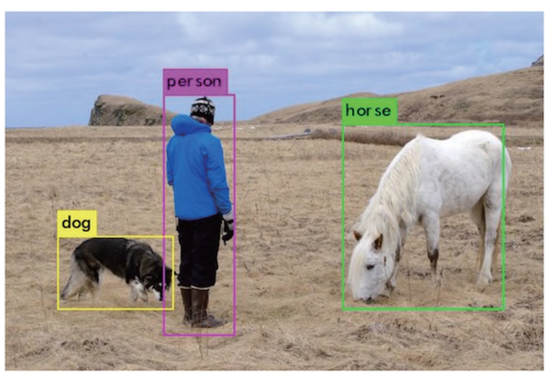

图 1 目标检测

Fig.1 Object detection

尽管目标检测算法已经取得了不错表现，但小目标（通常定义像素 30×30 以下的目标为小目标）检测仍然是一个急需解决的问题。小目标通常特征不明显，可利用信息较少，并且受到光照、遮挡等因素的干扰。然而，小目标的检测是极其重要和极具价值的。例如，准确地检测出机场跑道上微小物体（螺帽、螺钉、垫圈、钉子、保险丝等）将避免重大的经济损失；准确地检测出监控区域的小目标和对其行为进行分析将避免突发事件的发生，从而提高安防系数。可见，研究小目标的检测是非常有意义的。

为了提升小目标的检测与识别性能，一些基于深度学习的方法已纷纷被提出。所有针对小目标检测与识别的方法大致可以归纳为以下几类：

- 特征融合 [6-10]：通过融合卷积神经网络中不同层的特征图来增强小目标的特征表示；
- 上下文利用 [11-15]：利用小目标周围和图像的全局信息来辅助小目标的检测和识别；
- 注意力机制 [16-20]：通过模拟人类的注意力机制，提取出小目标中具有鉴别力的特征，从而提高模型的识别能力。

通过以上方式可以大幅度提升小目标的检测与识别性能，但这些方法仍有不足。比如，特征融合通常是单向的，上下文信息的利用并不充分，注意力的学习不够准确等，这些问题严重影响了小目标的检测和识别准确率。因此，本文有针对性的对这三类方法分别进行改进，具体研究内容如下：

（1）一种双向特征融合方法。在经典的单级目标检测算法 SSD 的基础上，通过特征融合的方式，将不同层之间的特征图进行融合。不同于现有的特征融合方法，本文中融合方式是双向的，不仅从深层向浅层进行信息传递，也从浅层向深层进行信息传递。

（2）上下文学习网络。通过设计神经网络来捕捉图像中物体与物体、物体和场景的关系，包括了局部上下文信息和全局上下文信息。

（3）注意力转移模型。为了更好地捕捉图像中具有鉴别力的特征，通过迭代的方式来逐步地定位图像中有利于目标识别的区域。在每一次迭代中，都会生成对应的注意力图，并将其作用于下一次迭代。也就意味着，本文的注意力是在不断转移的，并且注意力的转移不是随机的，而是与上一次的注意力息息相关的。

为了证明提出方法的有效性，本文将提出的方法融入到现有的目标检测框架中，并在公共数据集 PASCAL VOC 进行了实验验证。实验结果表明，改进后的方法大幅度提升了目标的检测性能，尤其是小目标的检测性能。

1 相关工作

目前基于深度学习的方法已经在目标检测领域占领了主导地位。目标检测算法大致可以分为两大类：两级目标检测 [1-3] 和单级目标检测 [5,10]。两级目标检测算法将检测任务拆解为目标定位和目标识别，首先在图像上生成大量的候选框，然后对候选框进行分类识别。单级目标检测算法将检测任务简化为回归任务，直接在图像上回归出目标所在位置以及对应的类别。相比较而言，两级目标检测在检测准确率上有优势，而单级目标检测在检测速度上有明显的优势。为了提升小目标在复杂场景下 [21] 的检测性能，研究学者从不同方面对小目标检测算法进行了改进。

1.1 特征融合

特征融合是提升小目标检测的一种重要手段。许多基于深度学习的检测算法也尝试了通过融合神经网络中不同层的特征来提升小目标的特征表达能力。文献 [22] 提出一种 Inside-Outside Network (ION) 方法。该方法首先从卷积神经网络的不同层中裁剪出候选区域特征，然后通过 ROI Pooling 将不同尺度的特征区域进行尺度归一化，最后将这些多尺度特征进行融合，从而提升区域特征表达能力。

HyperNet[23] 提出了一种类似于 ION 思想的方法。该方法精心设计了高分辨率的超特征图，通过整合中间层和浅层特征来生成候选区域和目标检测。该方法中通过利用反卷积层来向上采样深层特征图，并通过批标准化层来对输入特征图进行标准化。构建的超特征图还可以隐式地对来自不同层的上下文信息进行编码。文献 [24] 受到细粒度分类算法的启发，这些算法集成了高阶表示，而不是利用候选目标的简单一阶表示。该方法提出了一种新的多尺度位置感知和表示框架，该框架能够有效地捕获候选特征的高阶统计量，并生成更具区分性的特征表示。组合特征表示更具描述性，为分类和定位提供了语义和空间信息。FCN[25] 使用跳跃连接方式来融合浅层和深层的特性，以获得更好的特征表达。目前，FPN[26] 是最流行的利用多尺度特征的网络，它引入了一种自底向上、自顶向下的结构，将相邻层的特征结合起来以提高性能。该方法结构可以分为三个部分：自底向上（图 2 左）、自顶向下（图 2 右）和横向连接。自底向上就是一个前向的过程，生成一些不同尺度的特征图。自顶向下就是一个上采样的过程，通过横向连接将上采样的结果和自底向上生成的相同大小的特征图进行融合。通过这种方式，将深层特征和浅层特征进行了有效的融合，从而提高特征表达能力。与 FPN 类似，在单级目标检测 SSD 的框架下，文献 [27] 提出一种类似彩虹连接的方法来实现特征融合。

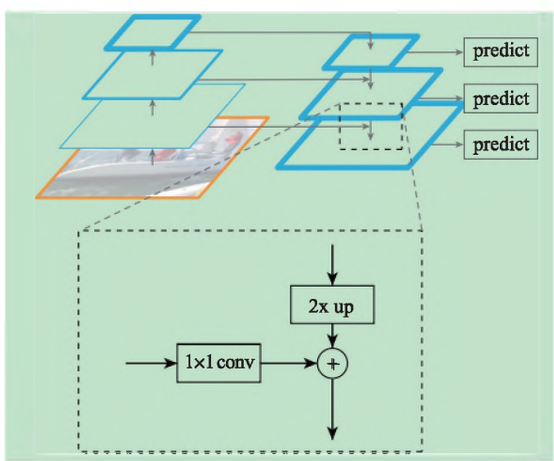

图 2 FPN 的网络结构

Fig.2 The network structure of FPN

1.2 上下文学习

上下文信息是我们理解目标特征信息的一种重要的补充信息，充分利用上下文信息将帮助我们提升小目标的检测性能。在深度学习之前，已有研究[28-30]证明通过对上下文进行合适的建模可以改进目标检测算法。随着深度神经网络的广泛应用，许多研究[31-33]也试图将目标周围的上下文集成到深度神经网络中。通常，它们利用手动设计的上下文窗口中的附加上下文特征来增强特征表示。上下文窗口通常比相应的候选区域稍大或稍小。通过提取上下文窗口中的特征信息，并将这些上下文信息用于增强对应候选区域的特征表示。但是，由于上下文窗口通常是通过手动设计的，这种方式严重限制了上下文信息提取的范围，很有可能丢失一些重要的上下文信息。一些研究[34-35]试图使用递归神经网络来编码上下文信息，而不是使用上下文窗口。文献[22]提出了一种没有使用上下文窗口的方法，该方法在特征图上沿左、右、上、下四个方向进行上下文信息传输，以捕获有价值的上下文。然而，该方法使模型变得复杂，并且在训练时需要仔细初始化参数。文献[36]提出了一种空间记忆网络，该网络通过多次记忆迭代有效地对实例级上下文进行建模。在此之后，文献[37]提出了一个迭代视觉推理框架，以此来捕捉场景中目标的上下文关系。由于空间和语义推理被集成到框架中，文献[37]中的迭代视觉推理在具有挑战的数据集COCO[21]（该数据集包含难以检测的对象，例如小的、被遮挡的和变形的目标）上获得了非常不错的检测性能。

1.3 注意力机制

深度学习中的注意来源于人类视觉系统的注意机制。人脑在接收到视觉信息、听觉信息等外部信息时，并不是对所有信息进行处理和理解，而是只关注一些重要或有趣的信息，这有助于滤除干扰信息，从而提高信息处理效率。

受到人类视觉注意力机制的启发，研究学者提出了许多算法来模拟人类的注意机制。最近，人们尝试性地将注意力应用到深层神经网络中[38-44]。深度玻尔兹曼机[45]在训练阶段，通过其重构过程包含了自上而下的注意力。注意机制也被广泛应用于递归神经网络（RNN）和长期短期记忆（LSTM）[46]中，来处理顺序决策任务[47-49]。注意力机制有多种实现形式，大致可分为软注意和硬注意。其中最具代表性的基于硬注意力的是递归注意力模型（RAM）[50]，它按时间顺序处理输入，并在图像中定位注意区域。该模型减少了不必要信息的干扰和噪声的影响，同时降低了计算成本。由于基于硬注意力的识别模型需要对焦点区域进行预测，因此在训练中通常采用强化学习，这会导致收敛困难。基于软注意的可微模型可以通过反向传播进行训练。考虑到软注意易于训练的优点，提出了许多基于软注意的识别算法[51-52]。两级注意网络（TLAN）[52]使用DNN将视觉注意应用于细粒度分类问题。全卷积注意网络（FCAN）[53]提出了一种基于强化学习的全卷积注意定位网络，用于自适应地选择多个任务驱动的视觉注意区域。

1.4 其他方法

GAN及其变体[57-58]在许多领域显示出了不错的效果，并在目标检测中得到了成功的应用。Li等人提出了一种专门针对小目标检测的感知GAN方法[59]，该方法通过生成器和鉴别器相互对抗的方式来学习小目标的高分辨率特征表示。具体来说，感知GAN的生成器将低分辨率的小区域特征转换为高分辨率特征，并与能够识别真正高分辨率特征的鉴别器竞争。最后，生成器学会了为小目标生成高质量特征的能力。进一步地，针对目标遮挡和形变问题，Wang等人提出了一种基于Fast R-CNN的改进检测模型[60]，它是由生成的对抗样本训练而成的。为了增强对遮挡和形变的鲁棒性，该模型中

引入了自动生成包含遮挡和变形特征的网络。通过对区域特征的遮挡和形变处理，检测模型可以接收到更多的对抗样本，从而使得训练的模型具有更强的能力。

此外，一些方法也尝试通过摆脱锚框的约束来提升小目标的检测性能。Law 等人提出了一种基于关键点的目标检测方法 CornerNet[61]。CornerNet 不再需要通过锚框来预测目标的位置，而是将目标建模为一对角点（目标的左上角和右下角）。在不依赖手工设计锚框来匹配目标的情况下，CornerNet 在公共数据集上取得了不错的表现。然而，由于角点对的错误匹配，CornerNet 会预测出大量错误的边界框。为了进一步提升检测精度，Duan 等人在 CornerNet 的基础上提出了一种基于中心点的目标检测框架 CenterNet[62]。CenterNet 首先预测两种类型的角点（左上角和右下角）和中心点，然后通过角点匹配确定边界框，最后利用预测的中心点来过滤角点不匹配引起的边界框。

2　一种双向特征特征融合方法

SSD 是一种主流的单级目标检测方法，该方法能够在保证检测速度的同时，还能保证较高的检测准确率。图 3 展示了 SSD 的网络结构图。通过图 3 可以发现，尽管 SSD 充分利用了不同尺度的特征图来进行目标检测，但是不同层之间是相互独立的，并没有充分利用不同特征图之间的相关性。这严重约束了 ssd 的目标检测性能，尤其是对于可视化特征较少的小目标。

2.1　基于双向特征融合的 SSD

事实上，不同尺度特征图上包含的特征是不相同的。浅层的特征图中通常包含有丰富的细节特征，而深层的特征图中包含有丰富的语义特征。为了充分利用浅层和深层特征，本文提出了一种双向特征融合方法，通过由深层到浅层和由浅层到深层的特征信息传递，使得用于目标检测的特征图既包含丰富的细节特征，又包含丰富的语义特征。更加特别之处在于，本文提出的双向特征融合方法能够使得每个层都包含有其它层的特征信息，从而大大提高特征表达能力。

图 4 展示了 ESSD（改进版 SSD) 的架构图。通过双向特征融合的方法增强小目标的特征表达，从而提高最终的小目标检测性能。图 4 中灰色部分为原始 SSD 中的操作，其他带颜色的为 ESSD 增加的操作。其中黄色箭头表示深层向浅层进行特征传递的过程，紫色箭头表示浅层向深层进行特征传递的过程，蓝色部分为双向特征融合后新生成的特征图。如后文表 1 中的检测结果所示，通过利用融合后的特征图构 成的特征金字塔，小目标可以被准确地检测出来。

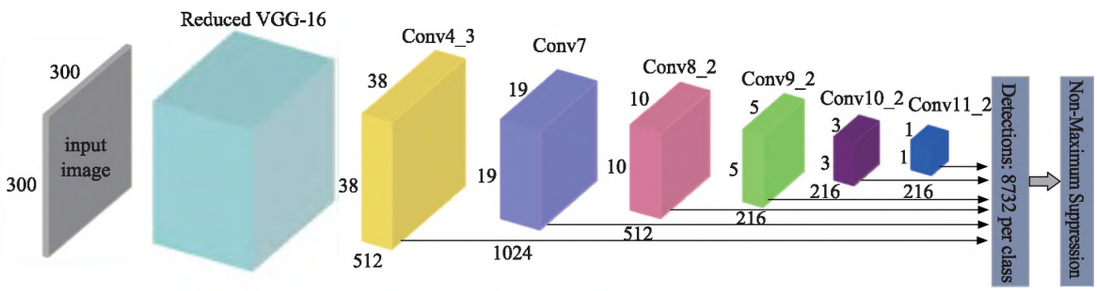

图 3 SSD 的网络结构图

Fig.3 The network structure of SSD

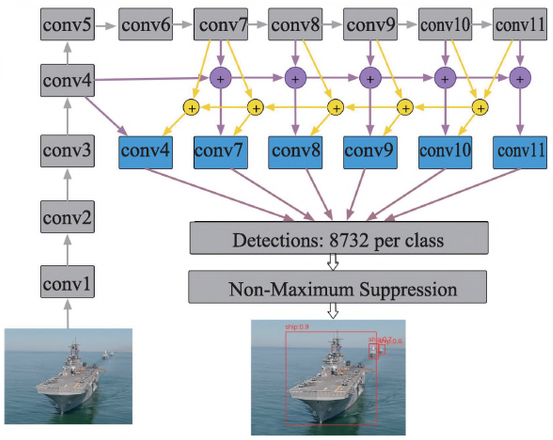

图 4 ESSD 的结构图

Fig.4 The framework of ESSD (Enhanced SSD)

2.2 双向特征融合细节

为了进一步说明本文是如何对浅层和深层特征进行融合的，图 5 中给出了特征融合的具体细节。如图 5 所示，中间层为目标层，目标是将第一层（具有高分辨率的浅层）特征和第三层（具有低分辨率的深层）特征融合到目标层中。为了实现特征融合，第一步是实现特征变为与目标层特征图相同的大小 2H×2W。之后，通过 1×1 的卷积操作来统一特征图的通道数，即将第一层的 2H×2W×C 特征图的 C 变为与目标层相同的通道数 512。同样地，第三层的 2H×2W×D 的特征图的 D 变为 512。考虑到每一层特征图中特征值的分布是非常不同的，因此在融合之前统一特征值的分布是非常有必要的。在图 5 中，通过 batch normalization 来实现不同特征图中特征值的分布统一。最后，融合来自不同层并且经过特殊处理的特征图，并生成新的具有更强表达能力的特征图。

如图 5 所示，特征融合过程包括降采样、上采样和融合。下采样和上采样有多种方法，如最近邻插值、双线性插值和三次插值。最大池法和反卷积可分别用于下采样和上采样。为了避免复杂度的增加，在实验中选择了最大池和双线性插值的降采样和上采样。此外，融合模式也是可选择，如逐元素求和、逐元素求积和 1×1 卷积操作。在实验中，通过 1×1 卷积来融合特征图。采用这种策略可以使网络自主学习加权求和的系数，从而实现更加有效的特征融合。

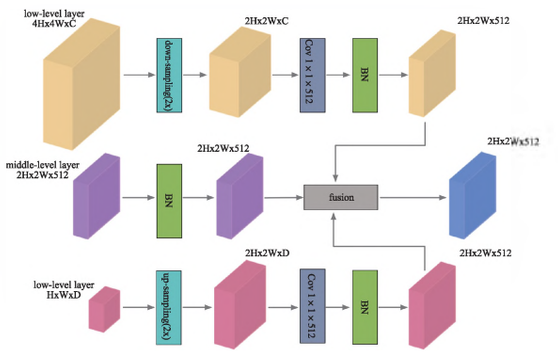

图 5 特征融合细节

Fig.5 Details of feature fusion

3 上下文学习网络

检测一个物体需要大量的信息，包括物体自身的细节信息及其周围的环境信息（上下文信息）。当目标较大或者特征较为明显时，单纯依靠目标自身特征就能很好的完成定位和识别。然而，当目标较小（30×30 像素以下）或者特征不明显时，单纯依靠目标自身特征就很难完成检测任务，而此时图像中的上下文信息成为了重要补充信息来源。目前主流的检测算法主要是利用目标自身特征信息来进行目标检测，这种方式严重限制了目标检测的准确率。本文提出了一个上下文学习网络，该网络的目标是捕捉对象之间的成对关系和每个对象的全局上下文。该网络由两个子网络组成：三层感知机和两层卷积神经网络。首先，为了捕获“成对”目标之间的上下文关系，本文设计了三层的感知机。然后，通过两层的卷积神经网络对成对的上下文关系进行聚合，进一步学习全局上下文。最后，得到具有丰富上下文信息的上下文特征图，这些信息对于准确的目标检测是非常有价值的。

本文所提出的上下文学习网络是轻量级的，并且易于嵌入在任何现有的网络中用于目标检测框架中。在本文中，将其嵌入 Faster R-CNN 的框架中。

3.1 上下文学习网络

当目标处于简单场景或者相似场景，并且目标外观不存在严重变化时，单纯依靠目标自身特征就能很好地完成定位和识别。但是，当目标的可视信息被损坏、模糊或者不完整（例如：一幅图像中包含噪声、不良照明条件或者目标被遮挡或截断），单纯依靠目标自身特征就很难完成检测任务，而此时可视上下文信息就成为了信息的重要来源。通常地，某些目标类经常出现在特定的情况下（比如，飞机出现在天空、盘子出现在桌面上），或者经常与其他类别的目标同时出现（比如，棒球和棒球棒）。鉴于上下文信息的重要，我们提出了一种上下文学习网络，该网络通过学习局部和全局的上下文信息来增强卷积特征图的表达能力。为了更好地说明上下文信息的作用，本文通过数学表达式来进行阐述。假设在图像 I 中有一些物体 $0=[0_1,0_1,\cdots,0_N]$，其中 N 是物体的总个数。我们的目标是检测出图像中的所有物体，这个过程可以通过以下公式来描述：

$$\text{argmax}L=\log P(O_{1:N}|M,I) \tag{1}$$

其中，M 是最大化对数似然估计 L 的模型，$O_{1:N}$ 表示 N 个物体 $[0_1,0_1,\cdots,0_N]$。为了利用物体之间的关系，对公式（1）进行等价变化，

$$\begin{aligned}\text{argmax}L&=\log P(O_{1:N}|M,I)\\&=\sum_{k=1:N}\log P(O_k|O_{1:N},M,I)\end{aligned} \tag{2}$$

进一步地，在公式（2）的基础上，本文增加上下文学习模型到目标函数中，目标函数近似为：

$$\arg_{M,S}\max^{L}\approx\sum_{k=1:N}\log P(O_k|S,M,I) \tag{3}$$

其中，上下文模型 S 和目标检测模型 M 联合进行优化。公式（3）表明，可以通过设计复杂的网络来提取物体自身的细节特征，以此提升目标的检测性能，还可以通过挖掘物体之间的上下文关系来协助目标的检测。

基于以上考虑，本文提出一种上下文学习网络，该网络致力于学习图像中物体与物体和物体与场景之间的关系。本文提出的上下文学习网络主要包括一个三层的感知机和两层的卷积神经网络。上下文学习网络的计算量主要集中于三层感知机中。三层感知机主要学习物体与物体之间的关系。该模块学习的是物体两两之间的关系。因此，可以通过使用 GPU 并行计算来提速。在学习物体两两之间的关系以后，将其通过设计的两层卷积神经网络来学习场景与每个物体间的关系。由于我们的输入和输出在维度上没有发生任何改变，因此，上下文学习网络可以作为一个基础模块，灵活地应用于任何存在的网络。原则上来说，我们的方法是现有基于卷积神经网络的目标检测方法的补充。在本节的余下部分，本文将详细介绍提出的上下文学习网络。

图 6 展示了上下文学习网络的所有细节信息。上下文学习网络的输入是通过卷积原始图像得到的特征图。假设我们获得了 $d\times d\times k$ 的特征图，其中 $d\times d$ 表示特征图的大小，k 表示特征图的个数。在 $d\times d$ 的特征图中，每一个 k 维特征向量对应一个坐标，以揭示其相对空间位置。由于在不同的图像中，物体个数是不相同的，并且我们很难知道哪些图像特征构成一个物体。因此，上下文学习网络将 $d\times d$ 特征图中的每个 k 维的特征向量当作一个物体，如图 6 所示。这也就意味着一个物体可以是背景、真实物体、物体之间的合并、物体与背景之间合并等。这种设计方式使得我们的模型在学习过程中具有更大的灵活性。

为了学习所有两两成对物体之间的关系，本文设计了一个三层的感知机，其中每一层的神经元个数为 512，并且随后紧跟非线性激活函数 ReLU。其中对于每一个物体（k 维特征向量），我们将其与其它物体两两连接构成一个 2 倍长的特征向量，并通过设计的感知机学习两两之间的关系。在通过感知

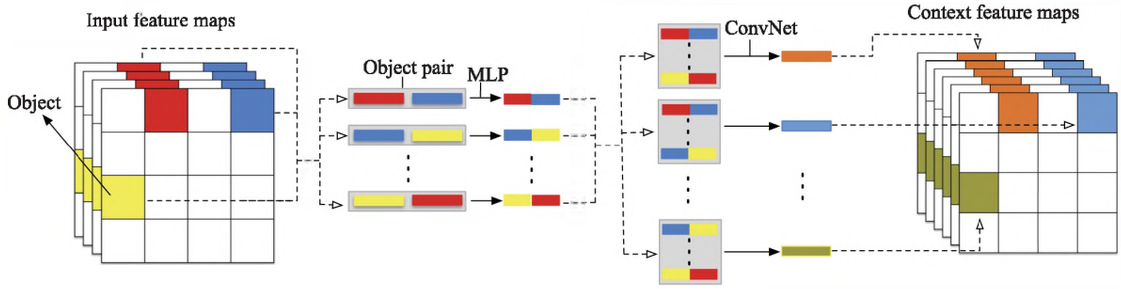

图 6 上下文学习网络

Fig.6 Context learning network

机后，我们得到了 $N\times(N-1)/2$ 个 512 维的特征向量，其中每一个特征向量表示了物体两两之间关系。这种方式只是考虑了局部物体级别的上下文，而忽略了全局图像级别的上下文。因此，我们设计了一个两层的卷积神经网络来学习物体与全局场景之间关系。对于每一个物体，我们将其与其它物体的关系特征向量进行融合，从而学习到其与整个场景之间的关系。考虑到场景不同位置的不同物体对指定物体的类别判断的影响程度是不相同的，我们对不同物体赋予不同的权重，该权重也是通过网络学习而来的。

对于每一个物体，上下文学习网络首先将其通过感知机得到的 $N-1$ 个 512 维特征向量进行串联。之后，将其通过两个卷积核大小为 1×3 的卷积层，其中卷积核个数分别为 256 和 512。最后，通过一个 1×1 的卷积操作将其进行通道融合，从而使得输出的 512 维特征向量包含有丰富的全局上下文信息。可以发现，输出的上下文特征图的每一个位置与输入的卷积特征图是相对应的。因此，我们可以轻易地融合卷积特征图和上下文特征图，从而得到一个更具特征表达能力的特征图。

3.2 基于上下文学习的 Faster R-CNN

本文提出的上下文学习网络是一个通用的模块，它可以应用到任何现有的深度卷积神经网络中。在本小节，我们将提出的上下文学习网络嵌入到两级目标检测算法 Faster R-CNN 中，使得 Faster R-CNN 具有感知上下文的能力，从而提高对小目标的检测性能。

图 7 展示了本文如何将上下文学习网络应用于 Faster R-CNN 的检测框架中。首先，通过 VGG16 进行特征提取。然后，通过 RPN 生成候选区域，并基于卷积特征图生成具有局部和全局上下文信息的上下文特征图。之后，使用 RoI 池化分别为 conv5_3 特征图和上下文特征图中的每个候选区域生成一个固定长度的特征描述符，并对每个描述符对进行批量规范化、串联和降维（1×1 卷积）以生成最终的描述符。每个生成的描述符紧接着由两个完全连接（fc）层进行处理，最终得到两个输出：一个 K 类预测和一个对边界框的调整。

数据输入：本文选择 VGG16 作为特征提取网络。VGG16 由 13 个卷积层、5 个最大池化层和 2 个全连接层构成。通过利用最后全连接层得到的特征向量，可预测出目标的类别和位置。给定一张分辨率为 $w\times h$ 的图像，将其通过 VGG16，从而得到用于目标检测的 conv5_3 特征图。该特征图的尺寸为 $w'\times h'$，是原始输入图像的 1/16。conv5_3 特征图后续将作为上下文学习网络的输入。

上下文学习：对于输入的 conv5_3 特征图，可以得到 $n=w'*h'$ 个“物体”，这些物体将作为我们上下文学习网络的输入。这也就意味着，通过设计的三层感知机将学习到 $n*(n-1)/2$ 个关系，每一个关系隐含

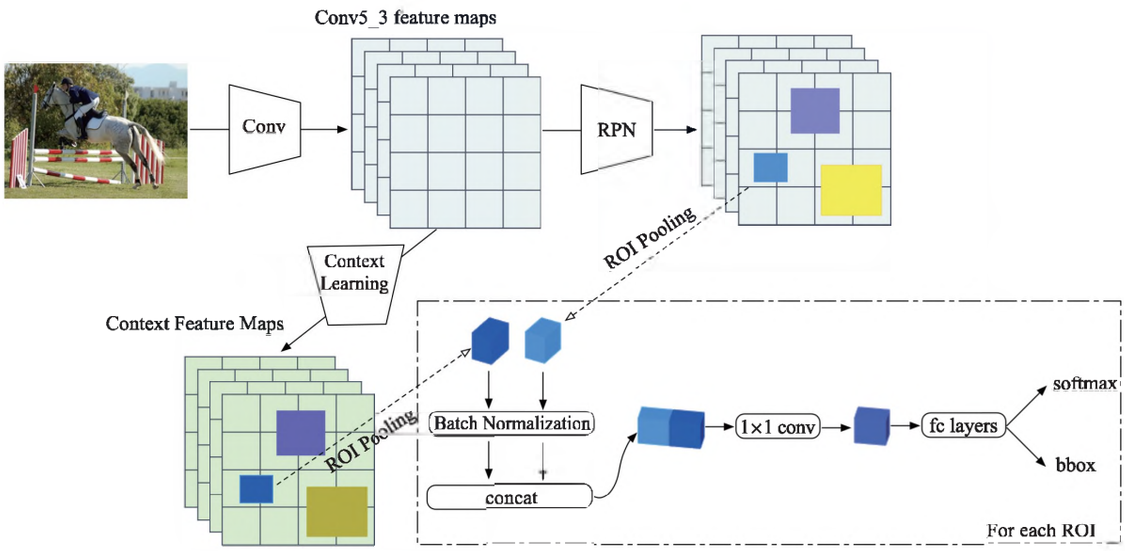

图 7 具有上下文意识的 Faster R-CNN

Fig.7 Context-aware Faster R-CNN

了物体对之间的上下文信息。对于每一个物体，我们将学习到 $n-1$ 个特征向量，每一个特征向量表示了这个物体和另一个物体之间的关系。之后，利用设计的卷积网络将这些特征进行融合，得到单个物体与其他所有物体，或者说单个物体与整个场景的上下文关系。最后，输出与输入特征图尺寸相同的上下文特征图。

特征融合：通过 RPN 生成网络，生成一系列的候选区域，在 conv5_3 中获得它们对应的卷积特征。此外，本文还通过提出的上文学习网络获得每个位置对应的上下文特征。由于 conv5_3 的卷积特征图和上下文特征图具有相同的大小，因此将所有特征进行融合以生成新的特征图是非常容易的。通过融合卷积特征和上下文特征，将使得新特征图具有更强的特征表达能力，既包含有目标细粒度特征，又包含有丰富的上下文特征。为了实现特征融合，本文利用 ROI 池化操作使得卷积特征向量和上下文特征向量具有相同的大小。然后，对每个候选区域的两个特征向量进行归一化、串联和通道融合（1×1 卷积），最后得到一个新的特征表示向量。通过利用新生成特征向量，就可以实现更加准确的目标定位和类别判断。

4　注意力转移模型

图像的精准识别是一件极具挑战的事情。目前，存在的方法通过利用深度卷积网络已经取得了不错的分类结果。但是，这些方法在面临图像中目标区域占比较小时将会失效。其原因在于，现有方法在特征提取的过程中是平等考虑图像中的每个位置的特征信息的。当图像中目标区域较小时，将会忽略目标区域本身的特征，从而丢失了一些有利于识别的关键特征信息。为了提升对小目标的识别，本文提出一种用于图像识别的注意力转移模型（ATM）（模拟人类视觉注意力机制），该网络通过迭代的方式能够有效地捕捉图像中的关键特征。该网络不再是对全图进行处理，而是通过迭代的方式生成不同的注意力区域。在每一次迭代中，我们都会生成对应的注意力图，并将其作用于下一次迭代。也就意味着，我们的的注意力是在不断转移的，并且注意力的转移不是随机的，而是与上一次的注意力息息相关的。最后，我们综合考虑多个注意力区域实现精确的图像分类。

在观察一幅图或者一个场景的时候，人类不会把注意力均匀的分布在每个区域。通常，人类首先会快速定位一些显著性区域，然后基于这些区域，

不断扩散和转移注意力。为了模拟人类的这种视觉机制，本文设计了一种注意力转移模型（ATM），该模型通过多次迭代生成不同的注意力图，每次生成的注意力图都包含了不同的焦点区域，并且每次迭代生成的注意力图不是相互独立的，而是相互制约和关联的，当前生成的注意力图是基于上一次注意力图转移而来的。也就是说，每次迭代我们关注不同的焦点区域，并且焦点区域之间存在推理关系。如图 8 所示。本网络主要包括卷积特征提取、生成注意力图、注意力转移和分类四个模块。该网络通过迭代的方法在图像中生成不同的焦点区域，然后将这些焦点区域合并生成最终的注意力图，最后将生成的注意力图作用于输入网络的特征图，从而提高模型的特征提取能力。首先，利用一个全卷积来生成注意力图。具体地，我们的输入是通过特征提取网络得到的特征图，输出是与输入同等大小的特征图（单通道）。网络结构包括收缩路径和扩张路径。在收缩路径中，包括三组卷积层，每组卷积层包含有两个同样大小的特征图。此外，在每组卷积之后，紧跟一个 2×2 MaxPool。在扩张路径中，通过反卷积操作以实现上采样，生成与搜索路径对称的有同样大小的特征图。最后，通过一个 1×1 conv + sigmoid 实现通道融合，输出特征图。在生成单个特征图以后，我们还需要通过多次迭代来生成更多的焦点区域。因此，基于当前状态我们需要预测生成新的注意力图，即注意力转移。为了使得每次迭代关注不同的焦点区域，我们需要对上一次迭代生成的焦点区域进行抑制，具体操作如下：

$$F_{i+1}(x)=F_i(x)\cdot(1-A_i(x)) \tag{4}$$

其中，$F_i(x)$ 表示第 i 次迭代的输入，$A_i(x)$ 表示第 i 次迭代生成的注意力图。通过将 $1-A_i(x)$ 方式获得上一次迭代的非关注区域，并将其重新作用于上次迭代的输入（点乘操作），从而得到当前迭代的输入。此外，为了使得我们网络有类似循环神经网络的记忆功能，本文还将上一次迭代中生成的特征图转移到当前迭代中，并将其与当前迭代中生成的特征图进行通道融合。最后通过多次迭代，生成不同的特征图，并对其进行融合（逐像素相加操作），从而获得最终的注意力图。图 8 中展示了三次迭代生成的焦点以及转移过程。

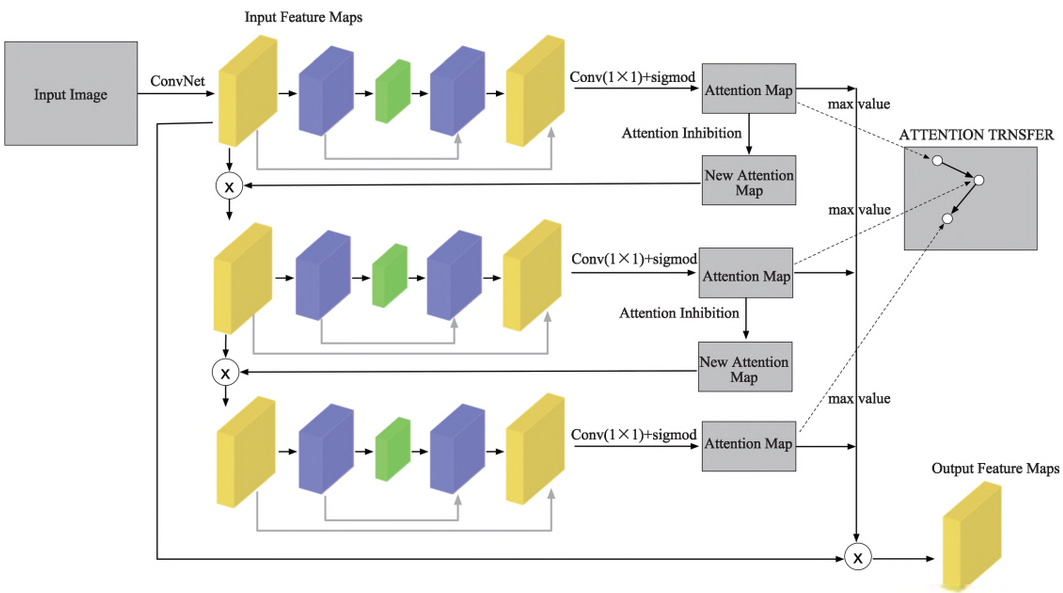

图 8 注意力转移模型的细节

Fig.8 Details of the proposed Attention Transfer Module (ATM)

表 1 ESSD 和目前主流方法在 PASCAL VOC 2007 上的检测结果

Table 1 Detection results of our ESSD and state-of-the-art detectors on PASCAL VOC 2007

方法	输入	训练数据	测试数据	mAP	FPS
YOLO	448	VOC2007＋2012	VOC2007	63.4	45
YOLOV2	416	VOC2007＋2012	VOC2007	76.8	67
Faster R-CNN		VOC2007＋2012	VOC2007	73.2	5
R-FCN		VOC2007＋2012	VOC2007	80.5	5.9
SSD	300	VOC2007＋2012	VOC2007	77.7	61
DSSD	321	VOC2007＋2012	VOC2007	78.6	9
ESSD	300	VOC2007＋2012	VOC2007	79.2	52
SSD	512	VOC2007＋2012	VOC2007	79.8	25
DSSD	513	VOC2007＋2012	VOC2007	81.5	6
ESSD	512	VOC2007＋2012	VOC2007	82.4	18

为了进一步观察本文注意力是如何转移的，可视化了注意力的转移过程，如图 9 所示。通过图 9 可以发现，ATM 以不断迭代的方式逐步地定位图像中的注意力区域（具有鉴别力的区域），最后将这些注意力合并在一起，构成我们关注的所有区域。

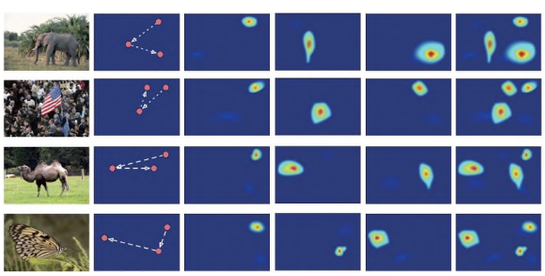

图 9 注意力转移过程

Fig.9 The attention transfer process

5 实验评估

5.1 数据集介绍

为了验证本文所提出方法的有效性，我们在目标检测 PASCAL VOC 2007 和识别数据集 CIFAR-100, Caltech-256 和 CUB-200 进行了实验验证。

PASCAL VOC 2007: 该数据集是目标检测中一个经典公开数据集，共计包含 9 963 张图像和 21 个种类的目标。其中，5 011 张图像用于训练和 4 952 张图片用于算法测试。

CIFAR-100：这个数据集共有 100 个类，每个类包含 600 张图像。每班有 500 张训练图片和 100 张测试图片。CIFAR-100 中的 100 个类被分为 20 个超类。每个图像都带有一个精细标签（它所属的类）和一个粗标签（它所属的超类）。

Caltech-256：该数据集是加利福尼亚理工学院收集整理的数据集，该数据集选自 Google Image 数据集，并手工去除了不符合其类别的图片。在该数据集中，图片被分为 256 类，每个类别的图片超过 80 张。

CUB-200：该数据集包含 11 788 张图片，分为 200 种鸟类。所有目标都使用边界框、局部位置和属性标签进行注释。这些注释信息将有助于验证注意力生成是否合理。

5.2 实验平台

本文所提出方法均是在深度学习框架 Tensorflow 下实现的。模型训练使用内存为 32G 的 Xeon 服务器，GPU 是 NVIDIA TITAN X，CUDA 版本为 8.0 和 cuDNN 5.1。

5.3 实验结果

（1）双向特征融合方法

表 1 展示了在 PASCAL VOC 数据集上的实验结果。与传统的 SSD 算法相比较，ESSD 在 mAP (mean average precision) 上大约有 3 个百分点的提升，与此同时保证较高的检测速度。此外，与两级目标检测算法 Faster R-CNN 相比较，我们的方法无论是在速度还是检测准确率上都有明显的优势。

（2）上下文学习网络

本文在公共数据 PASCAL VOC 上进行了实验。我们将上下文学习网络嵌入到 Faster R-CNN 中，命

名为 Context-Aware Faster R-CNN。

通过表 2 可以发现，在均使用 VGG16 作为基础网络时，与 Faster R-CNN 相比较，Context-Aware Faster R-CNN 在 mAP 上有 8.9% 的提升。在均使用 Residual-101 作为基础网络时，与 Faster R-CNN 相比较，Context-Aware Faster R-CNN 在 mAP 上有 8.4% 的提升。

表 2 Context-Aware Faster R-CNN 在 PASCAL VOC 2007 测试集上的实验结果

Table 2 Experimental results of Context-Aware Faster R-CNN on PASCAL VOC 2007 test set

方法	基础网络	mAP
Faster R-CNN	VGG16	73.2
Faster R-CNN	Residual-101	76.4
YOLOv2	Darknet-19	78.6
DSSD	Residual-101	81.5
Context-Aware Faster R-CNN	VGG16	82.1
Context-Aware Faster R-CNN	Residual-101	84.8

（3）注意力转移模型

为了证明注意力转移机制对小目标检测的有效性，本文在公共数据集上进行了实验，包括 CIFAR-100[54],Caltech-256[55] 和 CUB-200[56] 三个数据集。

表 3 展示了在三个数据集上分别的实验结果，并与目前基于注意力的方法（TLAN[52], FCAN[53], RA-CNN[20]）进行了比较。在 CIFAR-100, Caltech-256 和 CUB-200 三个数据集上，ATM 分别取得了 82.42%, 80.32% 和 86.12% 的准确率。

表 3 ATM 在 CIFAR-100, Caltech-256 和 CUB-200 三个数据集上的识别结果

Table 3 The recognition results of ATM on CIFAR-100, Caltech-256 and CUB-200

方法	CIFAR-100	Caltech-256	CUB-200
TLAN	72.88	68.82	77.90
FCAN	95.80	76.40	82.04
RA-CNN	97.21	79.24	85.31
ATM	97.68	80.32	86.12

上述实验表明，本文所提出的双向特征融合方法、上下文学习网络和注意力转移模型对小目标的检测是有效的。实际上，这三个算法是相辅相成的，可以集成为一个整体的网络模型。首先，通过双向特征融合方法提取到较好的目标特征表示；然后，通过上下文学习网络来学习上下文信息，并将上下文信息作为目标检测的补充信息；最后，通过注意力转移的方式来提升目标的识别性能。

5.4 错误分析

本文提出方法均是基于锚框机制的，因此检测性能严重依赖于锚框尺寸和数量的设计。当检测目标与设计锚框差异较大时，检测性能将大幅度下降。此外，本文提出方法对于稠密目标的检测性能较差，会将多个小目标检测为一个目标。其原因在于，两个（多个）目标的水平边界框的重叠比过大，从而导致检测框被 NMS 消冗。图 10 展示了部分较差的检测结果。

图 10 部分较差的检测结果

Fig.10 Some poor detection results

6 总结与未来工作

针对小目标检测和识别方法存在的问题，本文从特征融合、上下文学习和注意力生成三个角度来对现有算法进行了改进。具体地，本文首先提出了

一种双向特征融合方法，通过前向和后向的传递不同层的特征信息，从而使得新生成的特征图同时包含有丰富的细粒度特征和语义特征。接下来，为了充分利用目标的上下文信息，提出了一种上下文学习网络，通过学习成对物体之间的上下文关系和单个物体与整个场景直接的关系来辅助我们目标检测和识别。最后，为了更好地识别物体的类别，提出了一种注意力转移网络，通过不断迭代的方式来生成关注不同区域的特征图，从而使得用于分类的特征更加具有鉴别力。为了证明提出方法的有效性，本文在公共数据集上进行了大量的实验，并将实验结果与目前主流方法进行了比较。实验结果表明，本文所提出的方法在针对小目标的检测和识别性能上均有明显的优势。

后续的研究主要包括以下两方面：（1）将这三个算法融入到一个目标检测框架中，使之成为一个完整的小目标检测与识别的网络模型；（2）由于目前的方法都是基于锚框机制，这些方法的检测性能严重依赖于锚框的预定义，因此后续的研究将尝试利用关键点检测来替代边界框的回归。

利益冲突声明

所有作者声明不存在利益冲突关系

参考文献

[1] Z. Cai and N. Vasconcelos. Cascade r-cnn: delving into high quality object detection[C]. in *IEEE CVPR*, 2018.

[2] K. He, G. Gkioxari, P. Dolla´r, and R. Girshick. Mask r-cnn[C]. in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 1137-1149, 2017.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. -Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector[J]. in *European conference on computer vision*. Springer, 2016, pp. 21-37.

[5] J. Redmon and A. Farhadi.Yolo9000: better, faster, stronger[C]. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[6] T. Kong, A. Yao, Y. Chen, and F. Sun. “Hypernet: Towards accurate region proposal generation and joint object detection[C]. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 845-853.

[7] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better[J]. *arXiv preprint arXiv:1506.04579*, 2015.

[8] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation[C]. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[9] T. -Y. Lin, P. Dolla´r, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection[C]. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117-2125.

[10] J. Jeong, H. Park, and N. Kwak. Enhancement of ssd by concatenating feature maps for object detection. 2017.

[11] K. He, X. Zhang, S. Ren, J .Sun. Deep residual learning for image recognition[C]. in: CVPR, 2016.

[12] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C. -C. Loy, et al.. Deepid-net: Deformable deep convolutional neural networks for object detection[C] in: CVPR, 2015.

[13] W. Chu, D. Cai. Deep feature based contextual model for object detection[J]. in: Neurocomputing, 2018.

[14] Y. Zhu, R. Urtasun, R. Salakhutdinov, S. Fidler. segdeepm:

Exploiting segmentation and context in deep neural networks for object detection[C]. in: CVPR, 2015.

[15] X. Chen, A. Gupta. Spatial memory for context reasoning in object detection[C]. in: ICCV, 2017.

[16] K.Hara,M.-Y.Liu,O.Tuzel,andA.-m.Farahmand. Attentionalnetwork for visual object detection[J]. arXiv preprint arXiv:1702.01478, 2017.

[17] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan. Attentive contexts for object detection[J]. IEEE Transactions on Multimedia, 19(5):944-954, 2017.

[18] K.He, X.Zhang, S.Ren, and J.Sun. Identity mappings in deep residual networks[J]. In European conference on computer vision, pages 630-645. Springer, 2016.

[19] X. Liu, T. Xia, J. Wang, and Y. Lin. Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. CoRR, abs/1603.06765, 2016.

[20] Fu J, Zheng H, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]//CVPR. 2017, 2: 3.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dolla´r, and C. L. Zitnick. Microsoft coco: Common objects in context[J]. In European conference on computer vision, pages 740–755. Springer, 2014.

[22] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks[C]. in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2874–2883.

[23] T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: Towards accurate region proposal generation and joint object detection[C]. in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 845-853.

[24] Wang H, Wang Q, Gao M, et al. Multi-scale location-aware kernel representation for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1248-1257.

[25] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation[C]. in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[26] T.-Y. Lin, P. Dolla´r, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection[C]. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[27] J. Jeong, H. Park, and N. Kwak. Enhancement of ssd by concatenating feature maps for object detection. 2017.

[28] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection[C]. In CVPR 2009. IEEE Conference on, pages 1271–1278. IEEE, 2009.

[29] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild[J]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 891–898, 2014.

[30] R. Yu, X. Chen, V. I. Morariu, and L. S. Davis. The role of context selection in object detection[J]. arXiv preprint arXiv:1609.02948, 2016.

[31] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model[C]. In Proceedings of the IEEE International Conference on Computer Vision, pages 1134–1142, 2015.

[32] W. Ouyang, K. Wang, X. Zhu, and X. Wang. Learning chained deep features and classifiers for cascade in object detection[J]. arXiv preprint arXiv:1702.07054, 2017.

[33] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang, et al. Crafting gbd-net for object detection[J]. IEEE transactions on pattern

analysis and machine intelligence, 40(9):2109– 2123, 2018.

[34] Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.. Natural language object retrieval[C]. In: CVPR. (2016).

[35] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.. Generation and comprehension of unambiguous object descriptions[C]. In: CVPR. (2016).

[36] X. Chen and A. Gupta. Spatial memory for context reasoning in object detection[J]. arXiv preprint arXiv:1704.04224, 2017.

[37] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta. Iterative visual reasoning beyond convolutions[J]. arXiv preprint arXiv:1803.11189, 2018.

[38] Ji Y, Zhang H, Wu QMJ .Salient object detection via multi-scale attention CNN[J]. Neurocomputing 322:130–140,2018.

[39] Zhang H, Ji Y, Huang W et al. Sitcom-star-based clothing retrieval for video advertising: a deep learning framework[J]. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3579-x.2018.

[40] Xu K, Ba J, Kiros R et al . Show, attend and tell: Neural image caption generation with visual attention[C]. In: International conference on machine learning, pp 2048–2057.2015.

[41] Chen L, Zhang H, Xiao J et al . SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5659–5667,2017.

[42] Seo PH, Lin Z, Cohen S et al . Progressive attention networks for visual attribute prediction[J]. arXiv preprint arXiv:1606.02393 .2016.

[43] Das D, George Lee CS. Sample-to-sample correspondence for unsupervised domain adaptation[J]. Eng Appl Artif Intell 73:80–91.2018.

[44] Das D, George Lee CS. Unsupervised domain adaptation using regularized hyper-graph matching[C]. In: 2018 25th IEEE international conference on image processing (ICIP).

[45] Larochelle H, Hinton GE .Learning to combine foveal glimpses with a third-order Boltzmann machine[J]. In: Advances in neural information processing systems, pp 1243–1251, 2010.

[46] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Comput 9(8):1735–1780,1997.

[47] Kim JH, Lee SW, Kwak D et al. Multimodal residual learning for visual QA[J]. In: Advances in neural information pro- cessing systems, pp 361–369, 2016.

[48] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation[C]. In: Proceedings of the IEEE interna- tional conference on computer vision, pp 1520–1528,2015.

[49] Srivastava RK, Greff K, Schmidhuber J .Training very deep networks[J]. In: Advances in neural information processing systems, pp 2377–2385,2015.

[50] Mnih V, Heess N, Graves A et al.Recurrent models of visual attention[C]. In: NIPS.2014.

[51] Jaderberg M, Simonyan K, Zisserman A .Spatial transformer networks[J]. In: Advances in neural information processing systems, pp 2017–2025,2015.

[52] Xiao T, Xu Y, Yang K et al. The application of two-level attention models in deep convolutional neural network for fine- grained image classification[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 842–850,2015.

[53] Zhang Y, Qiu Z, Yao T, et al. Fully convolutional adaptation networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6810-6818.

[54] R. Yu, X. Chen, V. I. Morariu, and L. S. Davis. The role of context selection in object detection[J]. arXiv preprint

arXiv:1609.02948, 2016.

[55] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dolla r. A multipath network for object detection[J]. arXiv preprint arXiv:1604.02135, 2016.

[56] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang, et al. Crafting gbd-net for object detection[J]. IEEE transactions on pattern analysis and machine intelligence, 40(9):2109- 2123, 2018.

[57] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[58] Brock A, Donahue J, Simonyan K. Large scale gan training for high fidelity natural image synthesis[J]. arXiv preprint arXiv:1809.11096, 2018.

[59] Li J, Liang X, Wei Y, et al. Perceptual generative adversarial networks for small object detection[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1222-1230.

[60] Wang X, Shrivastava A, Gupta A. A-fast-rcnn: Hard positive generation via adversary for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2606-2615.

[61] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 734-750.

[62] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 6569-6578.

收稿日期：2020 年 1 月 10 日

冷佳旭，博士生，目前就读于中国科学院大学。主要研究方向包括：计算机视觉、深度学习、目标检测、目标跟踪和双目立体视觉。

本文主要负责算法设计与实验验证部分。

Leng Jiaxu is currently pursuing his Ph.D. degree in School of Computer Science and Tecnology in University of Chinese Academy of Sciences. His current research interests include computer vision, deep learning, object detection, object tracking, and stereo vision.

In this paper, he is responsible for the design and experimental analysis of the proposed algorithms.

E-mail: lengjiaxu17@mails.ucas.ac.cn

刘莹，中国科学院大学教授，中国科学院数据挖掘与高性能计算实验室负责人。主要研究方向包括数据挖掘、人工智能、并行计算等。

本文中完成了论文的国内外现状分析、方法原理和结论展望。

Liu Ying is currently a professor of School of Computer Science and Tecnology in University of Chinese Academy of Sciences, and the Dean of the Data Mining and High Performanle Computing Lab. Her research interests include data mining, artificial intelligence, parallel computing, etc.

In this paper, she is responsible for the literature review, principles and conclusions.

E-mail: yingliu@ucas.ac.cn

引文格式：冷佳旭,刘莹.基于深度学习的小目标检测与识别[J]. 数据与计算发展前沿,2020,2(2):120-135.DOI:10.11871/jfdc.issn.2096-742X.2020.02.010.PID:21.86101.2/jfdc.2096-742X.2020.02.010.

Leng Jiaxu, Liu Ying. Small Object Detection and Recognition Based on Deep Learning [J].Frontiers of Data & Coputing,2020,2(2):120-135.DOI:10.11871/jfdc.issn.2096-742X.2020.02.010.PID:21.86101.2/jfdc.2096-742X.2020.02.010.