第 58 卷 第 12 期/2021 年 6 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于深度学习的深层次多尺度特征融合目标检测算法

刘鑫，陈思溢***，陈小龙**，杜鑫浩*

湘潭大学自动化与电子信息学院，湖南 湘潭 411105

摘要 基于对 SSD(Single Shot MultiBox Detector)目标检测算法的分析，提出了一种基于深度学习的深层次多尺度特征融合目标检测(DMSFFD)算法。首先将 SSD 的特征层与相邻特征层进行融合，在融合之后的特征图中加入尺寸为 3 pixel×3 pixel 的卷积层，以减小上采样的混叠效应。之后进行更深层的特征融合，分别对较小的三个卷积层进行上采样操作，然后对 4 个特征层进行 concate 操作，以生成语义信息更加丰富的特征图，从而实现多尺度的小目标检测。为了节省计算资源，提高算法的实时性，基础网络选用 VGG16。融合后的算法虽然相对于 SSD 较为复杂，但实时性基本得到了保证，而且 DMSFFD 算法能够成功检测大部分 SSD 网络漏检的小目标，检测精度相对于 SSD 也有较大提升。

关键词 图像处理；计算机视觉；卷积神经网络；目标检测；特征融合

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202158.1210029

Deep Multi-Scale Feature Fusion Target Detection Algorithm Based on Deep Learning

Liu Xin, Chen Siyi***, Chen Xiaolong**, Du Xinhao*

School of Automation and Electronic Information, Xiangtan University, Xiangtan, Hunan 411105, *China*

Abstract Based on the analysis of the single shot multibox detector (SSD) target detection algorithm, we propose a deep multi-scale feature fusion target detection (DMSFFD) algorithm based on deep learning. The SSD feature layer and its adjacent layer are first fused and the 3 pixel×3 pixel convolution layer is added into the feature map after fusion to reduce the aliasing effect of upsample. Then the deeper feature fusion is conducted and the upsample operation is performed respectively for three small convolution layers. Subsequently the concate operation is performed for four feature layers to generate feature maps with richer semantic information, and thus the multi-scale small target detection is realized. In order to save computing resources and improve the real-time performance of the algorithm, VGG16 is selected as the basic network here. Although the fused algorithm is more complex than SSD, its real-time performance is basically guaranteed. Moreover, the DMSFFD algorithm can successfully detect the small targets missed by most SSD networks, and its detection accuracy is also greatly improved compared with that of SSD.

Key words image processing; computer vision; convolutional neural network; target detection; feature fusion

OCIS codes 100.200; 100.3008; 100.499

收稿日期：2020-07-14；**修回日期**：2020-09-08；**录用日期**：2020-10-12

E-mail**：973151308@qq.com；*E-mail**：540536315@qq.com；*****E-mail**651972992@qq.com

1 引 言

随着计算机视觉技术的快速发展，大量学者开始专注于图像分类[1-3]、语义分割[4-6]和目标检测[7-14]等领域的研究。目标检测是计算机视觉中最基本和最具挑战性的问题之一，其用于检测特定目标的位置，已被广泛应用于自动驾驶[15-16]、人脸检测[17-19]等重要领域。

现有的目标检测方法可分为传统方法和基于深度学习的方法。基于深度学习的目标检测方法不需要人工设计特征，具有更好的检测结果，已经超越传统的检测方法，成为当前流行的方法。与传统检测算法相比，基于深度学习的目标检测方法主要是通过卷积神经网络来提取图像特征，应用广泛的特征提取网络有 VGG[1]、ResNet[2] 和 GoogleNet[3] 等。在目标检测领域，主流方法之一为双阶段目标检测算法，如 R-CNN[4]、Fast R-CNN[5]、Faster R-CNN[6] 和 FPN[7] 等，该类方法基于分类的目标检测框架，首先利用 Selective Search[20] 和 RPN[6] 等算法生成待检测物体的候选区域，再对这些候选区域进一步进行分类和定位，最终得到检测结果。该类方法获得了较高的检测精度，但其无法满足实时性要求。

除了上述介绍的双阶段目标检测方法，还有一种主流方法是单阶段目标检测算法，如 YOLO[10] 和 SSD (Single Shot MultiBox Detector)[11] 等。该类方法利用回归的方法，建立目标检测框架，省去了双阶段网络中生成候选区域这一操作，大幅提升了算法的实时性。YOLO (You Only Look Once) 作为单阶段检测算法，极大提高了检测的效率，但仍存在识别准确率不高、泛化能力差等问题。Liu 等[11] 提出了 SSD 模型。在 SSD 网络中，多尺度特征图由不同的下采样生成，类似于 Faster R-CNN，利用 anchor 机制生成一组宽高比不同的真实标签框，再对目标对象的位置偏移量和类别置信度进行回归，从而得到初步的预测结果，最后结合非极大值抑制，得到最终的检测结果。在目标检测过程中，SSD 模型能够有效表达图像的特征信息，相对于 YOLO，SSD 的整体性能更好。但是 SSD 模型仍然存在不足之处，在小目标检测方面，检测精度低，主要原因是小目标在高层没有足够的上下文信息。因此，在双阶段检测过程中，取得高检测精度的同时保证单阶段检测过程中的速度已成为研究热点。近年来许多研究人员提出了基于 SSD 的多尺度特征融合方法[21-25]，在这些方法中，网络结构较为复杂，实时性有待改善。所以需要一种准确率高、实时性强的方法解决上述问题。

本文提出了一种基于深度学习的深层次多尺度特征融合目标检测 (DMSFFD) 算法，通过多尺度特征融合的方式构造一组语义信息差异较小的特征图，语义信息更加丰富，分布更加均匀，极大提高了目标检测性能。

2 SSD 基本原理

SSD 原理图如图 1 所示，该模型基于前馈卷积网络，生成一个固定大小的边界框集合，并给出框中对象类别的置信度，其中输入图像的尺寸为 300×300，特征图分别表示为 Conv4_3、Conv6、Conv7、Conv8_2、Conv9_2、Conv10_2 和 Conv11_2，对应的尺寸分别为 38 pixel×38 pixel、19 pixel×19 pixel、19 pixel×19 pixel、10 pixel×10 pixel、5 pixel×5 pixel 和 1 pixel×1 pixel，通道数依次为 512、1024、1024、512、256、256 和 256。

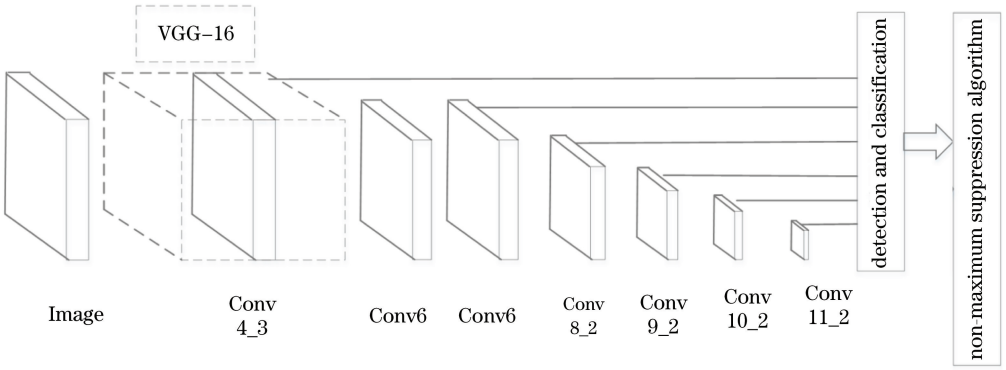

图 1 SSD 算法框架

Fig. 1 Algorithm framework of SSD

SSD 网络模型主要由两部分构成。一是位于网络前面的基于图像分类的 VGG 网络，其将 FC6 和 FC7 层转换为卷积层，被称为基础网络，用于提取低尺度的特征图。二是基础网络后面添加的辅助层，称其为辅助网络，用于提取高尺度的特征图。将辅助网络添加至被截断的基础网络的尾端。输入图像在进入网络之后不断向前传播，在此过程中生成了 6 个不同分辨率的特征图，分别为 conv4_3、conv7（由 VGG 网络中的 FC7 转换而来）、conv8_2、conv9_2、conv10_2 和 conv11_2，这些特征图尺寸逐渐减小。SSD 分别在这些特征图上，针对每一点构造 6 个不同尺度大小的默认框，默认框的个数分别为 4、6、6、6、4、4，将这些默认框连接到最后的检测分类层进行回归处理，用大特征图检测小目标，用小特征图检测大目标。首先生成多个初步符合条件的默认框，然后通过 NMS 算法过滤掉一部分重复检测或者不符合条件的默认框，生成最终的检测结果。

3　DMSFFD 算法

3.1　算法框架

通过对 SSD 目标检测算法的研究分析，本文提出了一种深层次多尺度特征融合目标检测与识别方法即 DMSFFD 算法，框架如图 2 所示，算法主要步骤如下。

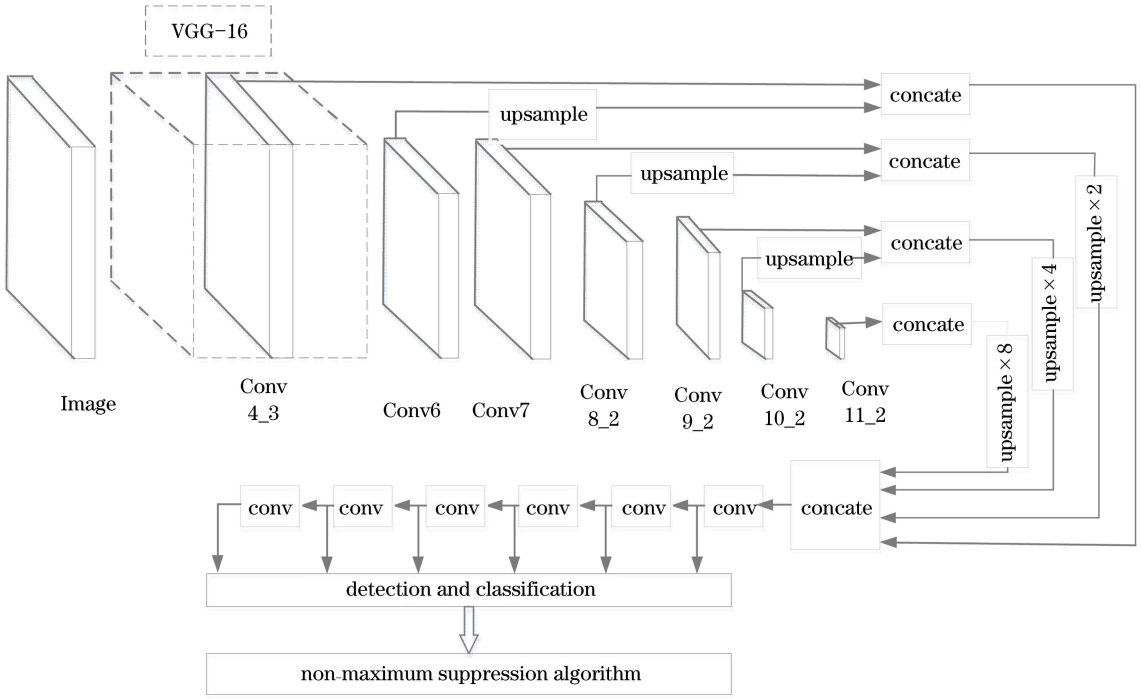

图 2　DMSFFD 算法框架

Fig. 2　Algorithm framework of DMSFFD

1）输入图像。以基础分类网络 VGG-16 作为模型的主框架，后端添加辅助网络，初步生成尺度不同的多个特征图。

2）引入一种特征融合模块。为了及时捕获特征图的语义信息，从 Conv4_3 层开始对相邻两层特征图进行融合，特征融合后得到的三个特征图含有丰富的语义信息。因 Conv4_3 处于网络高层，语义信息丰富，故保留其原有模型不变。在融合之后的特征图中加入 3 pixel×3 pixel 卷积层以减小上采样的混叠效应，将融合之后的三个特征图与尺寸为 1 pixel×1 pixel 的特征图一起输入到后面的网络中。

3）对初步融合后的特征图进行深层次的特征融合。对尺寸较小的三张特征图进行上采样操作，并将其与尺寸大的特征图进行 concate 操作，通过卷积操作，提取新的 6 张特征图。

4）将提取到的 6 张新特征图输入到检测器中进行训练，在训练过程中，设定一定大小的迭代次数，运用回归的方法，计算回归损失，不断自动调整模型参数，直至得到最好的参数模型，生成初步结果。

5）最后针对初步生成的边框，使用 NMS 算法过滤掉重复检测和不符合条件的边框，最终得到检测结果。

3.2 特征融合模块

一般情况下，在较为复杂图像的检测过程中，小目标物体只有很小的像素值，特征提取网络能提取到的语义信息是非常有限的。在特征提取过程中，浅层特征具有较高的分辨率，得到的位置信息更明确，而深层特征具有更加丰富的语义信息，但位置信息不够明确。为了解决上述问题，本文对提取到的不同特征进行融合，以丰富上下文的语义信息，结构如图 3 所示。

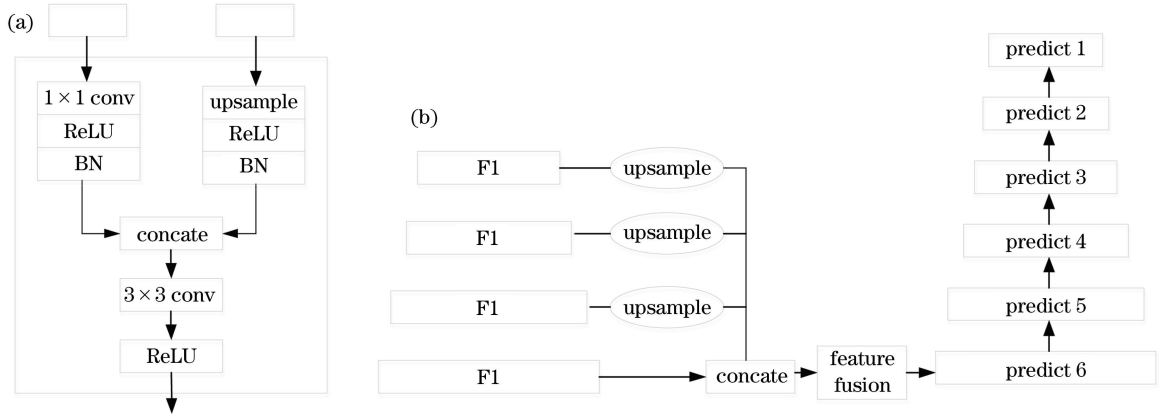

图 3　特征融合结构。(a)第一次特征融合；(b)第二次特征融合

Fig. 3 Feature fusion structure. (a) First feature fusion; (b) second feature fusion

第一次特征融合结构如图 3(a)所示，对 Conv6、Conv8_2 和 Conv10_2 进行上采样操作，Conv6、Conv8_2 和 Conv10_2 的尺寸分别与 Conv4_3、Conv7 和 Conv9_2 相同。进行融合的两个特征图的维度具有差异，因此采用 1 pixel×1 pixel 卷积层统一特征图的维度，使得进行融合的特征图具有相同的特征维度。在第一次特征融合过程中，将维度统一为 512。融合完成之后加入 3 pixel×3 pixel 的卷积层以减小上采样的混叠效应，对应的特征图尺寸分别为 38 pixel×38 pixel、19 pixel×19 pixel、5 pixel×5 pixel 和 1 pixel×1 pixel。

图 3(b)所示为第二次特征融合结构。在第一次融合得到的四个新的特征图 F1、F2、F3 和 F4 中，对较小的三个特征图 F2、F3 和 F4 进行上采样操作，然后对四个特征图进行深度融合，得到一个新的特征图 Fusion Feature Map，最后生成 6 个新的特征图进行预测。融合之后通过引入 3 pixel×3 pixel 卷积层，减小融合后的混叠效应。特征图细节信息如表 1 所示。

表 1　特征图信息表

Table 1　Feature map details

SSD		DMSFFD		Size /(pixel×pixel)
Feature map	Dimention	Feature map	Dimention	
Conv4_3	512	$F1_{conv}$	256	38×38
Conv7	1024	$F2_{conv}$	256	19×19
Conv8_2	512	$F3_{conv}$	256	10×10
Conv9_2	256	$F4_{conv}$	256	5×5
Conv10_2	256	$F5_{conv}$	256	3×3
Conv11_2	256	$F6_{conv}$	256	1×1

3.3 默认框选择

在 SSD 框架内，默认框不需要对应每层的实际感受野。我们采用多尺度的方法得到多个不同尺度的特征图，对于每个特征图，按照不同的大小和长宽比生成 k 个先验框。假设有 m 层特征图，每层特征图的大小为 s，则第 k 个特征图中的默认框尺寸为

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1), k \in [1, m], \quad (1)$$

式中：s_k 表示先验框大小相对于图片的比例，其中比例的最小值 $s_{\min}$ 为 0.2，比例的最大值 $s_{\max}$ 为 0.9。在中间层中，(1)式递增，对默认框施加不同的长宽比，表示为

$$\alpha_r \in \{1, 2, 3, 1/2, 1/3\}。 \quad (2)$$

在默认情况下，每个特征图均有一个 $\alpha_r = 1$，从而可以得到每个默认框的宽度与高度：

$$\begin{cases} w_k = s_k \sqrt{\alpha_r} \\ h_k = s_k / \sqrt{\alpha_r} \end{cases}。 \quad (3)$$

对于长宽比为 1 的默认框，即 $w_k = h_k$，默认框的宽度与高度分别为

$$w_k = h_k = \sqrt{s_k s_{k+1}}。 \quad (4)$$

对多个特征图所在位置的不同尺寸和长宽比的所有默认框进行预测，通过组合这些预测结果，得到涵盖各种输入物体尺寸和形状的集合。将一组默认边界框与每个特征映射单元相关联，以用于网络顶部的多个特征图。默认框以卷积的方式平铺特征映射，以便每个框相对于其对应单元的位置是固定的。在

每个特征映射单元中，预测相对于单元格中的默认框形状的偏移，并指出这些框中每类详细类别的评分。具体来说，对于在给定位置的 k 个框中的每个框，我们计算 c 类分数和相对于原始默认框的 4 个偏移量。因此，在特征图中的每个位置，总共需要 $(c+4)\times k$ 个预测值，尺寸为 $m_1\times n$ 的特征图产生 $(c+4)\times k\times m_1\times n$ 个输出。默认框类似于 Faster R-CNN 中使用的 anchor boxes，但我们将其应用于不同分辨率的特征图中，在多个特征图中使用不同的默认框形状，可以有效地离散可能的输出框形状空间。

3.4　损失函数

检测框架的整体目标损失函数由置信度损失(L_{conf})和位置损失(L_{loc})的加权和表示：

$$L(x,c,l,g)=\frac{1}{N}\left[L_{\text{conf}}(x,c)+\alpha L_{\text{loc}}(x,l,g)\right], \tag{5}$$

式中：N 表示先验框的正样本数目；x 表示默认框与真实标签框的匹配结果，$x=0$ 表示匹配失败，$x=1$ 表示匹配成功；c 是 Softmax 函数对每个类别的置信度；α 是权重；l 表示预测框；g 表示真实标签框。位置损失是预测框与真实标签框参数之间的平滑 L1 损失，可表示为

$$L_{\text{loc}}(x,l,g)=\sum_{i\in P_{\text{os}}}^{N}\sum_{m'\in\{c_x,c_y,w,h\}}x_{ij}^{K}\,\text{smooth}_{\text{L1}}(l_i^{m'}-\hat{g}_j^{m'}), \tag{6}$$

式中：i 为预测框序号；j 为真实框序号；；K 为类别编号；$\hat{g}_j^{m'}$ 为第 j 个预测框相对于真实标签框的位置；$l_i^{m'}$ 为第 i 个预测框的位置；$\text{smooth}_{\text{L1}}(\cdot)$ 为 smooth L1 损失函数；$x_{ij}^{K}\in\{1,0\}$，x_{ij}^{K} 取 1 时表示第 i 个预测框与第 j 个预测框的 IoU 大于阈值，此时真实框中的类别为 p，x_{ij}^{K} 取 0 时表示背景；P_{os} 为正样本；c_x、c_y、w 和 h 分别表示预测框的中心坐标及其宽、高。类似于 Faster R-CNN，以平移量 $d_i^{c_x}$、$d_i^{c_y}$ 和尺度缩放因子 d_i^{w}、d_i^{h} 获取真实标签的近似回归预测框 $\hat{g}_j^{m'}$。由于 x_{ij}^{p} 的存在，因此位置误差仅针对正样本进行计算。先对真实框(ground truth)的位置参数 g 进行编码，得到 $\hat{g}$。因为预测值也是编码值，若设置布尔参数为 True，则超参数 variance 被包含在预测值中，若设置布尔参数为 False，则需要手动设置超参数 variance，以对预测框的四个值进行缩放，编码时要加上 variance：

$$\begin{cases}\hat{g}_j^{c_x}=(g_j^{c_x}-d_i^{c_x})/d_i^{w}\\ \hat{g}_j^{c_y}=(g_j^{c_y}-d_i^{c_y})/d_i^{h}\\ \hat{g}_j^{w}=\ln\left(\dfrac{g_j^{w}}{d_i^{w}}\right)\\ \hat{g}_j^{h}=\ln\left(\dfrac{g_j^{h}}{d_i^{h}}\right)\end{cases}, \tag{7}$$

式中：$g_j^{c_x}$ 为第 j 个真实框的横坐标 x 的偏移量；$g_j^{c_y}$ 为第 j 个真实框的纵坐标 y 的偏移量；g_j^{w} 为第 j 个真实框的坐标宽度 w 的偏移量；g_j^{h} 为第 j 个真实框的坐标高度 h 的偏移量。类别置信度损失采用 softmax loss 函数，定义为

$$\begin{cases}L_{\text{conf}}(x,c)=-\sum_{i\in P_{\text{os}}}^{N}x_{ij}^{p}\ln(\hat{c}_i^{p})-\sum_{i\in N_{\text{eg}}}\ln(\hat{c}_i^{0})\\ \hat{c}_i^{p}=\dfrac{\exp(c_i^{p})}{\sum_{p}\exp(c_i^{p})}\end{cases}, \tag{8}$$

式中：$\hat{c}_i^{0}$ 表示第 i 个预测框预测为背景的概率值；N_{eg} 表示负样本；$x_{ij}^{p}\in\{1,0\}$，x_{ij}^{p} 取 1 时表示第 i 个搜索框和第 j 个类别框的重叠度(IOU)大于阈值，此时真实框中的对象类别为 p，x_{ij}^{p} 取 0 时，表示为背景；c_i^{p} 第 i 个预测框预测类别 p 的概率值。

通过 anchor 机制，生成一组宽高比不同的真实标签框，再对目标对象的位置偏移量和类别置信度进行回归处理以得到预测结果，最后结合非极大值抑制方法，得到最终的检测结果。

4　实验结果与分析

本文实验基于深度学习框架 PyTorch 实现，所使用的计算机硬件平台是 CPU Intel(R)Core(TM) i7-8700 CPU @ 3.20 GHz，Nvidia GeForce GTX 1080 Ti 显卡。

实验所用数据集为 PASCAL VOC2007 和 PASCAL VOC2012，二者均包含 20 个类别，我们在 VOC2007 和 VOC2012 合并而成的数据集上进行训练。

4.1　VOC2007 数据集中的实验

在 VOC2007 测试集中检测结果。基于 SSD 模型，输入图像的分辨率尺寸为 300 pixel×300 pixel，在训练过程中，权重衰减因子(weight_decay)设置为 0.0001，每次送入网络训练的样本数量(batch size)设置为 32，动量(momentum)设置为 0.9，初始学习率为 0.001，在迭代到 40000 和 50000 次时，分别将学习率衰减为原来的 1/10，在迭代到 120000

次时训练结束，得到最终模型。

训练结束后，采用 VOC2007 的测试数据集，选取 Faster R-CNN、SSD 和 DSSD 算法进行对比，实验结果如表 2 所示。在 20 个物体类别中，DMSFFD 算法均取得了最好的检测结果，其中有 12 个种类的平均准确率(mAP)超过 90%，19 个种类的 mAP 超过 80%。

表 2 VOC2007 数据集中的测试结果

Table 2 Test results on VOC2007 dataset

unit: %

Image	DSSD	SSD	Faster	DMSFFD
Aero	89.9	88.4	76.5	**90.7**
Bike	87.9	86.0	79.0	**89.7**
Bird	85.5	78.9	70.9	**90.3**
Boat	78.4	75.8	65.5	**88.0**
Bottle	53.9	48.8	52.1	**70.3**
Bus	88.6	86.8	83.1	**90.7**
Car	86.2	84.1	84.7	**90.0**
Cat	91.9	90.9	86.4	**90.9**
Chair	71.1	69.1	52.0	**87.1**
Cow	89.5	88.0	81.9	**90.9**
Table	78.7	78.4	65.7	**89.0**
Dog	91.3	90.5	84.8	**90.8**
Horse	89.6	89.0	84.6	**90.6**
Motor	88.4	86.8	77.5	**90.6**
Person	79.2	76.2	76.7	**84.5**
Plant	61.8	57.0	38.8	**81.7**
Sheep	78.0	72.7	73.6	**82.7**
Sofa	89.9	88.3	73.9	**93.9**
Train	93.2	92.0	83.0	**97.0**
TV	84.4	83.4	72.6	**90.6**

各个模型之间的 mAP 对比如表 3 所示。与 Faster R-CNN 相比，DMSFFD 的 mAP 提升了 15.3 个百分点；与 SSD 相比，DMSFFD 的 mAP 提升了 8 个百分点；与 DSSD 相比，DMSFFD 的 mAP 提升了 7.4 个百分点。

表 3 各算法在 VOC2007 测试集中的 mAP 对比

Table 3 mAP comparison of algorithms on VOC2007 test set

Algorithm	DSSD	SSD	Faster	DMSFFD
mAP/%	81.4	80.5	73.2	**88.5**

在 PASCAL VOC2007 测试集中进行测试，DMSFFD、DSSD 和 SSD 的检测速度如表 4 所示。与 DSSD 相比，本文网络层数相对较少，因此检测速度有明显优势。由于本文算法在 SSD 的基础上增加了特征融合模块，因此检测速度相比 SSD 稍有差距，但总体来说，实时性满足要求。

表 4 检测速度的对比

Table 4 Detection speed comparison frame/s

Algorithm	DSSD	SSD	DMSFFD
Detection time	9.5	63	**38**

4.2 PASCAL VOC2012 数据集中的实验

将 PASCAL VOC2007 测试集和训练集及 PASCAL VOC2012 训练集合并为训练集。由于训练数据增加，将迭代次数增加到 140000。初始学习率设置为 0.001，且分别在 80000、100000 和 120000 迭代次数时降为原来的 1/10，直至训练结束，其他设置均与 PASCAL VOC2007 数据集中的实验设置相同。

表 5 显示了不同算法在 PASCAL VOC2012 测试集中的检测结果。可以看出，DMSFFD 算法与 Faster R-CNN 和 SSD 相比，所有类别的准确率(AP)值均有提升；与 DSSD 相比，DMSFFD 算法在 18 个类别的目标上都获得了最高的准确率，其中有 12 个类别的 AP 值超过 90%，18 个类别的 AP 值超过 80%。从 mAP 来看，本文算法优于 DSSD，因此验证了本文算法的优势。

表 5 VOC2012 测试集中的测试结果

Table 5 Test results on VOC2012 test set %

Image	DSSD	SSD	Faster	DMSFFD
Aero	87.3	87.0	84.9	**90.7**
Bike	84.3	83.8	79.8	**90.3**
Bird	79.4	78.8	74.3	**90.0**
Boat	69.6	68.0	53.9	**84.0**
Bottle	56.8	55.4	49.8	**71.9**
Bus	86.7	84.0	77.5	**90.6**
Car	76.5	75.0	75.9	**84.1**
Cat	**92.9**	90.8	88.5	**90.9**
Chair	69.5	65.0	45.6	**83.9**
Cow	81.3	79.7	77.1	**90.0**
Table	74.3	72.6	55.3	**85.1**
Dog	**91.5**	90.3	86.9	**90.9**
Horse	88.6	88.2	81.7	**90.7**
Motor	88.6	86.8	80.9	**90.5**
Person	82.1	79.5	79.6	**86.3**
Plant	60.3	59.4	40.1	**78.5**
Sheep	79.6	77.8	72.6	**87.3**
Sofa	79.7	79.5	60.9	**90.1**
Train	88.2	88.1	81.2	**90.8**
TV	79.9	78.8	61.5	**90.6**

各模型在 PASCAL VOC2012 测试集中的 mAP 对比如表 6 所示。本文算法的 mAP 达到

87.4%，相比于 Faster R-CNN、SSD 和 DSSD，分别提高了 17、9 和 7.5 个百分点。

表 6　各算法在 VOC2012 测试集中的 mAP 对比

Table 6 mAP comparison of algorithms on VOC2012 test set　　unit: %

Algorithm	DSSD	SSD	Faster	DMSFFD
mAP	79.9	78.4	70.4	**87.4**

4.3　可视化对比实验

为了更直观地对 DMSFFD 算法进行展示，图 4～7 所示为本文所提算法与 SSD 模型可视化的检测效果对比。从图 4(a)、(b)可知，对于有遮挡的物体，SSD 算法不能检测出来，而本文所提算法能够进行有限检测。从图 5(a)、(b)可知，对于目标定位不准确的问题，DMSFFD 算法能够较好修正，而 SSD 算法对目标对象存在漏检，本文所提算法能够准确检测。从图 6(a)、(b)可知，SSD 算法对于小目标的检测效果不佳，不能准确检测图中的瓶子，而本文所提算法成功地检测出了更多的小目标，在一定程度上改善了漏检的情况，但仍然存在没有检测出来的小目标。从图 7(a)、(b)可知，对于严重遮挡的多目标，本文算法能够更准确检测。

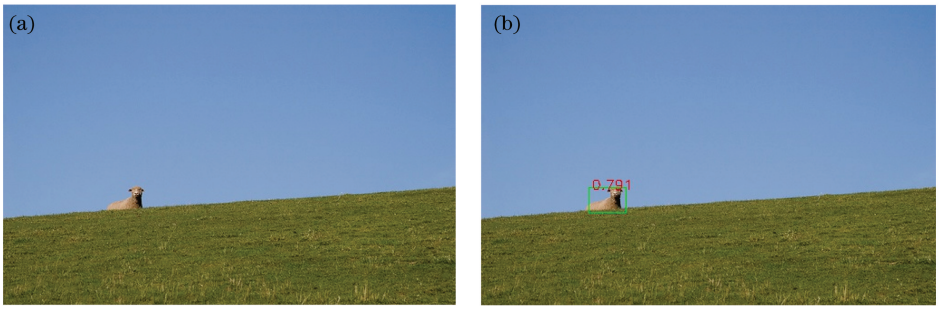

图 4　不同算法对遮挡物体的检测结果。(a) SSD；(b) DMSFFD

Fig. 4　Test result of occluded object by each algorithm. (a) SSD; (b) DMSFFD

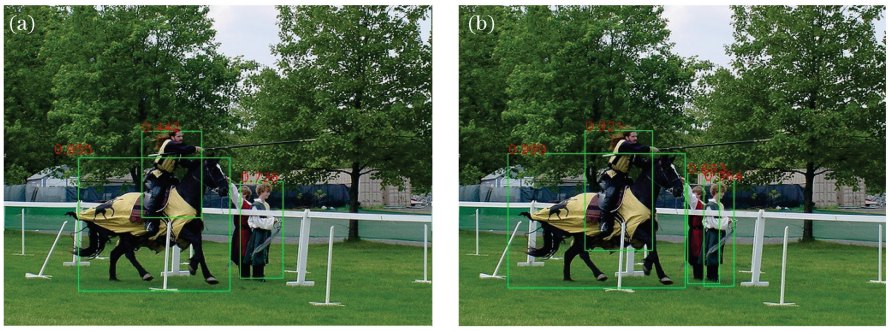

图 5　不同算法对目标的定位结果。(a) SSD；(b) DMSFFD

Fig. 5　Positioning result of target by each algorithm. (a) SSD; (b) DMSFFD

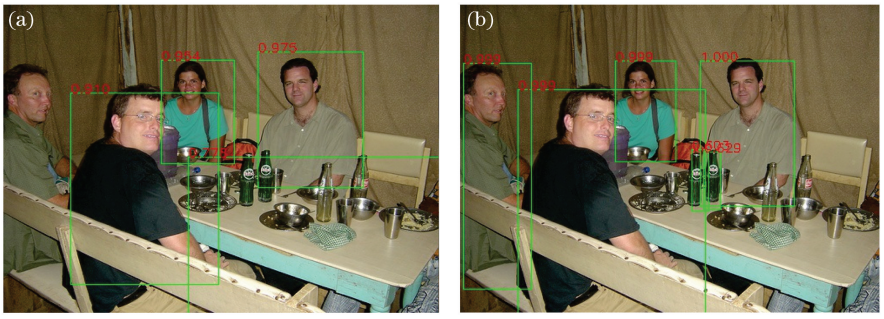

图 6　不同算法对小目标的检测结果。(a) SSD；(b) DMSFFD

Fig. 6　Detection result of small target by each algorithm. (a) SSD; (b) DMSFFD

图 7 不同算法对多目标遮挡的检测结果。(a) SSD;(b) DMSFFD

Fig. 7 Detection result of multiple occluded objects by each algorithm. (a) SSD;(b) DMSFFD

5 结 论

针对 SSD 目标检测算法存在的对小目标检测效果不佳、对目标存在漏检、对遮挡目标的检测性能不佳等问题，提出了一种深层次多尺度特征融合目标检测模型（DMSFFD）。分别在 PASCAL VOC2007 和 PASCAL VOC2012 数据集中进行实验，所得 mAP 分别为 88.5% 和 87.4%。由于 DMSFFD 在 SSD 模型的基础上引入了特征融合模块，模型变得更加复杂，相比于 SSD 模型，检测速度稍逊色，但检测精度有明显提升。与 DSSD 模型相比，DMSFFD 的检测速度更快，已达到实时性要求，同时 DMSFFD 的检测精度也展现了更明显的优势。总的来说，DMSFFD 的整体性能更优。同时，各项对比实验表明，DMSFFD 对小目标的检测能力较好，但无法对图像中所有小目标进行准确检测，接下来将探索性能更优的特征融合方式，进一步提升模型对小目标的检测能力。

参 考 文 献

[1] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]// International Conference on Learning Representations, May 7-9, 2015, San Diego, USA. New York: Cornell University Library, 2015: 1-14.

[2] 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 Press, 2016: 770-778.

[3] 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 Press, 2015: 1-9.

[4] Shelhamer E, Long J, Darrell T, et al. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[5] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[6] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018: 833-851.

[7] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

[8] 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.

[9] 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 Press, 2017: 936-944.

[10] 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 Press, 2016: 779-788.

[11] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016: 21-37.

[12] Zhang M, Wang S C, Yang D F, et al. Air-to-ground target detection algorithm based on attention learning in key areas[J]. Laser & Optoelectronics Progress, 2020, 57(4): 041006.
张萌, 王仕成, 杨东方, 等. 重点区域注意力学习的空对地目标检测算法[J]. 激光与光电子学进展, 2020, 57(4): 041006.

[13] Wang M R, Xu G M, Yuan H W, et al. Object detection by deep sparse feature learning of salient polarization parameters[J]. Laser & Optoelectronics Progress, 2019, 56(19): 191101.
王美荣, 徐国明, 袁宏武, 等. 显著性偏振参量深度稀疏特征学习的目标检测方法[J]. 激光与光电子学进展, 2019, 56(19): 191101.

[14] Ren Z J, Lin S Z, Li D W, et al. Mask R-CNN object detection method based on improved feature pyramid[J]. Laser & Optoelectronics Progress, 2019, 56(4): 041502.
任之俊, 蔺素珍, 李大威, 等. 基于改进特征金字塔的 Mask R-CNN 目标检测方法[J]. 激光与光电子学进展, 2019, 56(4): 041502.

[15] Ge Y Y, Xu Y J, Zhao S, et al. Detection of small and dense traffic signs in self-driving scenarios[J]. CAAI Transactions on Intelligent Systems, 2018, 13(3): 366-372.
葛园园, 许有疆, 赵帅, 等. 自动驾驶场景下小且密集的交通标志检测[J]. 智能系统学报, 2018, 13(3): 366-372.

[16] Li Y P, Hou L Y, Wang C, et al. Moving objects detection in automatic driving based on YOLOv3[J]. Computer Engineering and Design, 2019, 40(4): 1139-1144.
李云鹏, 侯凌燕, 王超, 等. 基于 YOLOv3 的自动驾驶中运动目标检测[J]. 计算机工程与设计, 2019, 40(4): 1139-1144.

[17] Ke X, Li J P, Guo W Z, et al. Dense small face detection based on regional cascade multi-scale method[J]. IET Image Processing, 2019, 13(14): 2796-2804.

[18] Fu J L, Bajić I V, Vaughan R G, et al. Datasets for face and object detection in fisheye images[J]. Data in Brief, 2019, 27: 104752.

[19] Bose P, Bandyopadhyay S K. Facial spots detection using convolution neural network[J]. Asian Journal of Research in Computer Science, 2020: 71-83.

[20] Uijlings J R R, van de Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[21] Jiang W T, Zhang C, Zhang S C, et al. Multiscale feature map fusion algorithm for target detection[J]. Journal of Image and Graphics, 2019, 24(11): 1918-1931.
姜文涛, 张驰, 张晟翀, 等. 多尺度特征图融合的目标检测[J]. 中国图象图形学报, 2019, 24(11): 1918-1931.

[22] Tan H C, Li S H, Liu B, et al. Feature enhancement SSD for object detection[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(4): 573-579.
谭红臣, 李淑华, 刘彬, 等. 特征增强的 SSD 算法及其在目标检测中的应用[J]. 计算机辅助设计与图形学学报, 2019, 31(4): 573-579.

[23] Shan Q W, Zheng X B, He X H, et al. Fast object detection and recognition algorithm based on improved multi-scale feature maps[J]. Laser & Optoelectronics Progress, 2019, 56(2): 021002.
单倩文, 郑新波, 何小海, 等. 基于改进多尺度特征图的目标快速检测与识别算法[J]. 激光与光电子学进展, 2019, 56(2): 021002.

[24] Song Y L, Pang Y W. Backbone network for object detection task[J]. Laser & Optoelectronics Progress, 2020, 57(4): 041021.
宋雅麟, 庞彦伟. 针对目标检测任务的基础网络[J]. 激光与光电子学进展, 2020, 57(4): 041021.

[25] Liu W J, Gao M Y, Qu H C, et al. Light-weight multi-object detection network based on inverted residual structure[J]. Laser & Optoelectronics Progress, 2019, 56(22): 221003.
刘万军, 高明月, 曲海成, 等. 基于反残差结构的轻量级多目标检测网络[J]. 激光与光电子学进展, 2019, 56(22): 221003.