文章编号　1004-924X(2023)12-1859-11

结合多尺度上下文信息的唐卡小样本目标检测

胡文瑾[1,2*]，唐慧媛[1]，乐超洋[1]，宋华飞[1]

(1. 西北民族大学 中国民族语言文字信息技术教育部重点实验室，甘肃 兰州 730030；
2. 西北民族大学 数学与计算机科学学院，甘肃 兰州 730030)

摘要：通过对图像中感兴趣的对象进行分类与定位，能够帮助人们理解唐卡图像丰富的语义信息，促进文化传承。针对唐卡图像样本较少，背景复杂，检测目标存在遮挡，检测精度不高等问题，本文提出了一种结合多尺度上下文信息和双注意力引导的唐卡小样本目标检测算法。首先，构建了一个新的多尺度特征金字塔，学习唐卡图像的多层级特征和上下文信息，提高模型对多尺度目标的判别能力。其次，在特征金字塔末端加入双注意力引导模块，提升模型对关键特征的表征能力，同时降低噪声的影响。最后利用Rank & Sort Loss替换交叉熵分类损失，简化模型训练的复杂度并提升检测精度。实验结果表明，所提出的方法在唐卡数据集和COCO数据集上的10-shot实验中，平均检测精度分别达到了19.7%和11.2%。

关　键　词：唐卡；小样本目标检测；上下文信息；多尺度特征；双注意力机制

中图分类号：TP391　　**文献标识码**：A　　**doi**：10.37188/OPE.20233112.1859

Few-shot object detection on Thangka via multi-scale context information

HU Wenjin[1,2*], TANG Huiyuan[1], YUE Chaoyang[1], SONG Huafei[1]

(1. *Key Laboratory of China's Ethnic Languages and Information Technology Ministry of Education, Northwest Minzu University, Lanzhou* 730030, *China*;
2. *School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou* 730030, *China*)
* *Corresponding author, E-mail: wenjin_zhm@126.com*

Abstract: Classifying and locating objects of interest in Thangka images can help people understand the rich semantic information of Thangka and promote cultural inheritance. To address the problems of insufficient Thangka image samples, the complex background, the occlusion of detection targets, and the low detection accuracy, this paper proposes a few-shot object detection algorithm for Thangka images that combines multi-scale context information and dual attention guidance. First, a new multi-scale feature pyramid is constructed to learn the multi-level features and contextual information of Thangka images and improve the ability of the model to discriminate multi-scale targets. Second, a dual attention guidance module is added at the end of the feature pyramid to improve the ability of the model to represent key features while reducing the impact of noise. Finally, Rank&Sort Loss is used to replace the cross-entropy classifi-

收稿日期：2022-08-22；**修订日期**：2022-10-28.

基金项目：国家自然科学基金(No. 62061042；No. 61862057)；国家民委创新团队计划资助(No. 2018[98]号)；西北民族大学“双一流”和特色发展引导专项资金资助项目

cation loss, which simplifies the model training process and increases the detection accuracy. Experimental results indicate that the proposed method achieved a mean average precision of 19.7% and 11.2% in 10-shot experiments using a Thangka dataset and the COCO dataset, respectively.

Key words: thangka; few-shot object detection; contextual information; multi-scale feature; dual attention mechanism

1 引　言

唐卡是我国珍贵的非物质文化遗产,拥有上千年的历史,以刺绣、绘画或雕刻等形式流传至今,其内容涉及到人民日常生活、传统神话故事和历史政治事件等多个领域,被誉为藏族的“百科全书”。作为一种具有鲜明特色的艺术形式,其融入了千百年来藏族人民对家乡的热爱之情,是研究藏族文明和历史发展的重要资料。将人工智能技术应用于唐卡研究中,不仅为唐卡的保护、存储和传播提供了一种智能且有效的手段,而且对传承中国传统文化和带动区域经济发展都有非常重要的意义。唐卡主尊的头饰、台座、手持物等目标的检测,将为人们理解唐卡蕴含的深度语义信息提供一种更便捷的途径,也为后续唐卡的图像描述、缺损信息的修补、图像标注等研究提供技术支撑。

目标检测是计算机视觉领域核心问题之一,主要任务是精确地定位图像中的物体并判断其类别。近年来,随着深度学习的蓬勃发展,陆续出现了R-CNN[1],Fast R-CNN[2],Faster R-CNN[3],SSD[4],YOLO系列[5-9]等多种目标检测方法。常用的检测方法都需要人工标注的数据集对模型进行训练,然而在实际应用场景中想要获得高质量的大型标注样本集往往需要耗费昂贵的人力和物力。受小样本学习的启发,参考了迁移学习、度量学习[10]以及元学习[11]的思想,研究者们提出了诸多小样本目标检测方法。2018年,Chen等人集成了Faster R-CNN算法和SSD算法的优点提出了小样本的迁移检测器(Low-Shot Transfer Detector,LSTD)[12],通过知识迁移和背景抑制正则化分别利用源域和目标域的知识,提高了模型对小样本数据的泛化能力。2019年,Kang等人[13]将单阶段目标检测算法YOLOv2与元学习思想相结合,利用支持集数据的特征对查询集特征进行重加权,使检测器能够快速适应新类别,提高了模型对新类别的识别能力。2020年,Wang等人以Faster R-CNN为基础提出了两阶段微调算法(Two-Stage Fine-tuning Approach, TFA)[14],在预训练阶段利用基类样本进行训练得到基模型;在微调阶段冻结主干网络和区域选取网络(Region Proposal Network, RPN),在整个特征提取器不变的情况下,将新类随机初始化的权值分配给box预测网络,进而微调分类和回归网络,有效地防止了小样本学习中的过拟合问题。2021年Sun等人首次将对比学习引入目标检测,提出了基于对比建议编码的小样本目标检测算法(Few-Shot Object Detection via Contrastive Proposal Encoding,FSCE)[15]。该方法冻结了Faster R-CNN中的主干特征提取器,增加了一个与分类回归分支平行的对比分支。然后通过建议框对比编码损失计算特征之间的相似性,使相同类别的实例形成一个更加紧密的簇,不同类别实例之间的边界扩大。模型使用丰富的基类数据进行训练,然后利用一部分随机采样的基类和新类实例联合微调特征金字塔网络(Feature Pyramid Networks,FPN)、bounding box分类器、回归器和对比分支。实验表明FSCE有效地降低了TFA算法检测失败率。

FSCE算法在公共数据集上检测效果良好,但是对于唐卡目标的检测效果并不理想。这主要是因为:唐卡图像中目标对象形态多样(例如:莲花座形状清晰,在画中占比较大;经书类对象形状较小且常被主尊的手遮挡),部分对象之间相互重叠(如:莲花座与须弥座,经书类对象与自然类对象)。此外,唐卡数据集中尤其是画面清晰、完整的图像数量较少,大多数图像分辨率较低,这些均导致传统的目标检测方法应用到唐卡图像上容易出现目标识别不到或错误分类的现象。本文在深入挖掘唐卡图像视觉和语义特点

的基础上，提出了一种结合多尺度上下文信息和双注意力引导的唐卡小样本目标检测算法（Few-Shot Object Detection on Thangka via Multi-scale Context Information and Dual Attention Module，FSMC），主要创新和贡献总结如下：

（1）设计新的特征提取网络，提高模型的准确性。采用了由多个空洞卷积层组成的多尺度特征提取网络，使模型更好地适应不同尺寸目标的同时还捕获了图像全局上下文信息。通过结合唐卡的场景信息和目标的局部特征，提高了模型对目标的识别能力；

（2）加入双注意力引导模块，增强网络对重叠、遮挡目标的表征能力。利用空间注意力关注唐卡目标和周围区域的相关性，增加通道注意力获取目标更具判别力的精确信息，提升模型对重叠、遮挡目标的检测能力；

（3）设计损失函数，降低相似目标的误检率，有效提高在唐卡数据集上的检测性能。引入RSLoss[16]损失函数，将RoI特征分为正、负样本，经过排序后再计算损失，不仅能够加快网络的收敛速度而且保证分类损失值更小。

2　模型介绍

2.1　模型整体结构设计

本文以Faster R-CNN作为唐卡小样本目标检测的基本框架，采用两阶段训练方法，第一阶段的训练集是大量标注的基类数据，第二阶段使用部分采样的基类实例和全部新类实例对模型进行微调。FSMC模型由特征提取网络、FPN、上下文特征提取模块（Contextual Feature Extraction Module，CFM）、双注意力引导模块（Dual Attention Module，DAM）、RPN和分类、预测、对比分支组成，网络结构如图1所示。其中：CFM中能由多个不同扩展率的空洞卷积在提取不同感受野信息的基础上通过密集连接进行特征聚合，从而得到唐卡图像中丰富的上下文信息的特征表达。DAM是由空间注意力和通道注意力并联组成的双注意力引导模块，使模型更加关注有用信息，减少背景信息对结果的干扰。此外，采用RSLoss替换了交叉熵分类损失，可减少分类时的误差，降低误分类现象出现的概率。

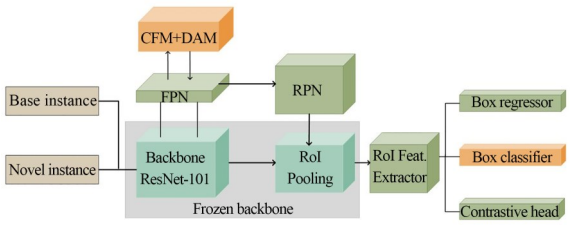


图1　FSMC网络模型

Fig. 1　FSMC network model

2.2　多尺度上下文特征提取模块

唐卡图像中的目标众多，大小尺度不一，因此唐卡小样本目标检测器需要关注到不同尺度的对象。FPN[17]是由Lin等人在2017年提出的解决物体检测中多尺度问题的网络，通过简单的网络连接的改变，获得不同大小的感受野来检测不同尺度的物体，大幅度提升了检测小物体的性能。然而FPN仅将不同尺度的特征图经过自上而下的路径进行相加合并，不同感受野捕捉到的信息之间缺乏交流，对于唐卡图像中大量语义信息之间的相关性并未充分的挖掘。

为此，本文设计了多尺度上下文特征提取模块CFM，它由多层空洞卷积层通过密集连接组成，网络结构如图2所示。为了提高对不同尺度目标的识别能力，CFM中多层空洞卷积块是由3个rate=[2,3,5]的小扩张率卷积层和3个rate=[9,12,18]的大扩张率卷积层堆叠而成。扩张率小的卷积层可实现小目标的特征提取，扩张率大的卷积层完成大目标的特征提取。利用空洞卷积来有组织地聚合多尺度的上下文信息是因为空洞卷积能够指数级地扩大感受野而不损失分辨率或者信息质量。图3显示了在特征图相同的情况下，空洞卷积根据其扩张率能够得到相应的比普通卷积更大的感受野，且不会损失图像信息。

常用的空洞卷积模块一般采用rate=[6,12,18,24]等扩张率，以获取大范围的全局信息。然而在唐卡图像中存在着大量的具有丰富语义的对象，诸如手持物类别包括有：嘎巴啦、自然类、经书、法器、兵器、庄严器具等，如果扩张率过大，会导致采样过于稀疏而损失局部特征，不利于此类小目标的识别。由此，本文对扩张率进行了优化，使得CFM模块能捕获更多细节化的上下文特征。此外，多空洞卷积层通过密集连接将学习

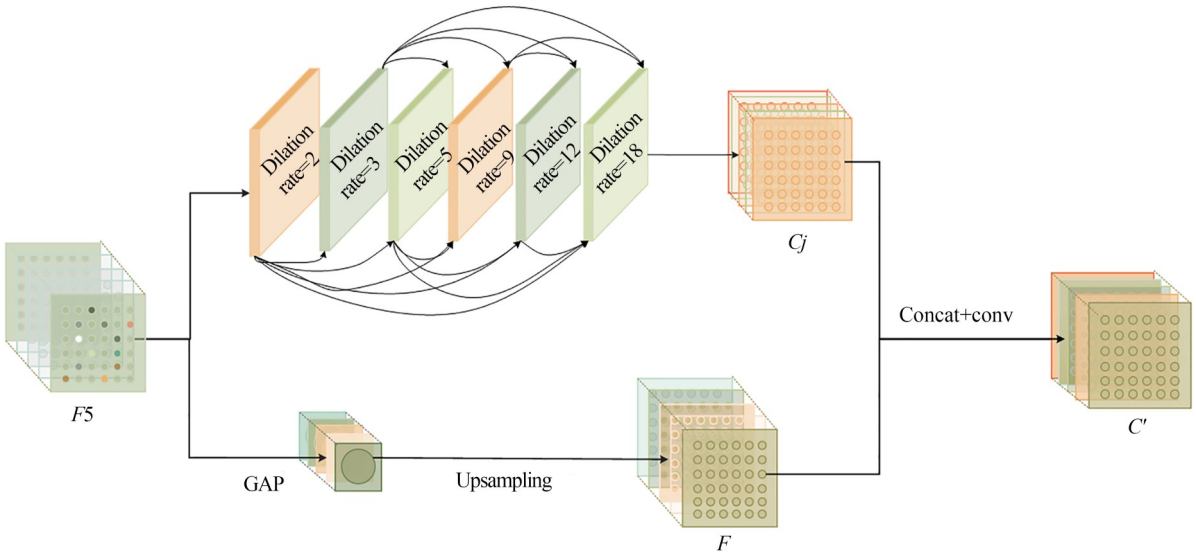
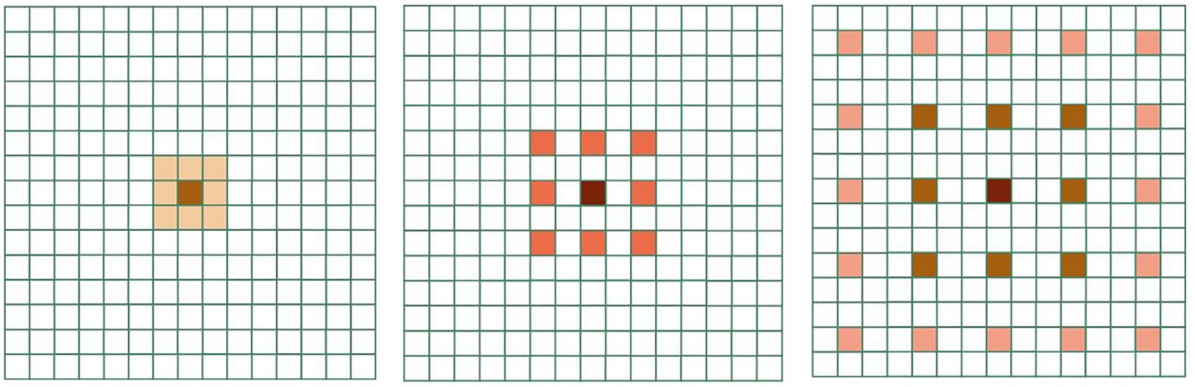


图 2　多尺度上下文特征提取模块示意图

Fig. 2　Contextual feature extraction module



(a) 普通卷积
(a) Ordinary convolution

(b) Dilation rate=2空洞卷积
(b) Dilated convolution with rate=2

(c) 叠加dilation rate=3空洞卷积
(c) Stacked dilated convolution with rate=2

图 3　普通卷积和空洞卷积

Fig. 3　Ordinary convolution and dilated convolution

到的多尺度特征进行融合，使得密集块中的每一个卷积层的输出都与其前面卷积层的输出进行拼接再传递给下一层。CFM 通过密集连接不仅加强了不同层级间特征的传递，利用了层级间的互补信息，优化了参数的表达，而且还能很好地避免了叠加空洞卷积产生的网格效应，如图 3(c)。由此可以看出：在不显著增加计算量的前提下，CFM 不仅能解决传统的深度学习方法获取图像特征尺度单一、提取精度较低等问题，还能更好地获得唐卡中丰富的上下文信息，从而最终实现利用目标区域之间的相关性提高模型的定位、分类目标的能力。

CFM 的计算过程如下：首先，将图像经过 FPN 自下而上的路径生成的特征图 $F5$ 输入到 CFM 中，得到多尺度上下文特征 C_j。然后，为了保持初始输入的粗粒度信息，将原始输入特征 $F5$ 经过全局平均池化和上采样得到粗粒度特征 F。最后将 F 与 C_j 经过 1×1 卷积进行特征融合输出为 C'。计算过程如下：

$$C_j=\varphi_j(\varphi_{j-1}(x)\oplus x), \tag{1}$$

$$F=\phi[Avgpool(F5)], \tag{2}$$

$$C'=f^{1\times1}\{F\oplus C_j\}, \tag{3}$$

其中：$\oplus$ 为特征融合 concat 操作，$\varphi(\cdot)$ 表示空洞卷积运算，$\phi(\cdot)$ 表示双线性插值上采样方法，

$f^{1\times1}(\cdot)$表示卷积核为1×1的卷积运算。

2.3　双注意力引导模块

上述CFM学习到的多尺度上下文信息有助于模型适应不同尺寸的对象，在目标检测任务中起到了重要的作用。为了进一步发掘出特征间的相似性，突出不同对象的特征表示，避免重叠或遮挡目标之间的相互影响，本文设计了一个由空间注意力和通道注意力并联组成的双注意力引导模块。利用该模块学习特征空间和通道上的相互依赖性，使模型能够在空间和通道两个维度上聚合相似特征，实现对特征的进一步甄别。最后结合全局上下文信息，提升对重叠、遮挡对象的检测率，如图4所示。

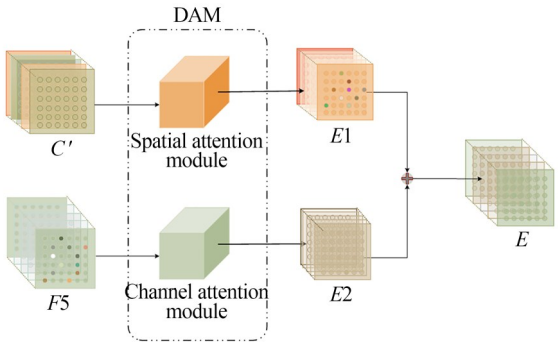


图4　双注意力机制模块

Fig. 4　Dual attention module

其中，空间注意力模块能够根据特征内部相关性，捕获特征图中各个子区域之间的空间依赖关系，即计算出特征之间的关系权值，特征越相似，权值越大。通过该权值，网络能够在全局特征图中锁定到同一类别特征的不同位置，有选择地聚合上下文信息，从而辅助模型更精准地分辨出重叠或遮挡的不同对象，结构如图5所示。具体计算过程如下：对于输入特征C'，首先经过两个卷积层C_a，C_b，得到特征图$F_a\in\boldsymbol{R}^{C\times H\times W}$，$F_b\in\boldsymbol{R}^{C\times H\times W}$。然后将这2个特征图进行reshape得到$F'_a\in\boldsymbol{R}^{C\times N}$，$F'_b\in\boldsymbol{R}^{C\times N}$，其中$N=H\times W$。再将$F'_a$与$F'_b$的转置矩阵相乘后输入到softmax中，输出结果即空间注意力图$S\in\boldsymbol{R}^{N\times N}$。接着我们将$C'$输入到另一个卷积层中生成新的特征$F_c\in\boldsymbol{R}^{C\times H\times W}$，$S$与$F_c$经过维度变换后相乘再reshape为$C\times H\times W$大小，最后将输出结果与输入$C'$相融合得出空间注意力特征$E1$。

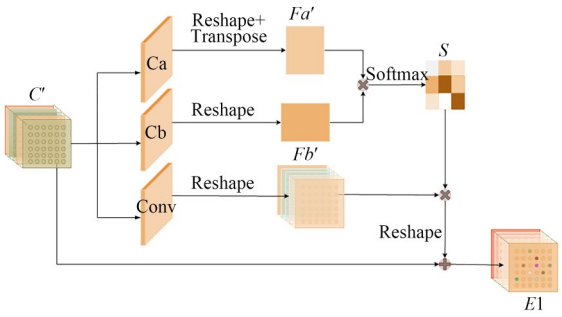


图5　空间注意力模块

Fig. 5　Spatial attention module

在高级语义特征中，每一个通道都可以被看作是对于某一特定类的响应。通道注意力模块能够挖掘出特征图上各通道之间的依赖关系，利用权重对特征进行重标定，突出重要的通道信息，使模型关注到目标的类别信息，更精确地识别出重叠、遮挡对象的类别。由于浅层网络的粗粒度特征分辨率更高，能够保留更原始的通道关系，故该通道注意力直接作用于CFM的输入特征$F5$。通道注意力特征的计算过程如图6所示：$F5\in\boldsymbol{R}^{C\times H\times W}$经过全局平均池化获得全局特征，再经过一个卷积层和批量归一化层进行变换后输入到sigmoid函数中生成各通道的注意力权重，利用该权重对原输入特征进行重校准得到通道注意力特征向量$E2$。

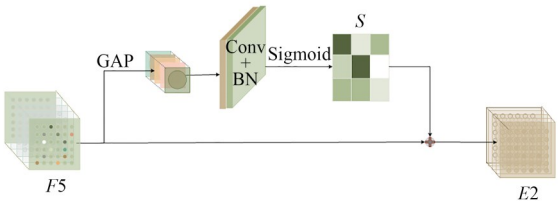


图6　通道注意力模块

Fig. 6　Channel attention module

最后将空间注意力特征$E1$与通道注意力特征$E2$进行元素相加求和，得到最终的注意力特征E：

$$E=E1+E2. \tag{4}$$

2.4　损失函数

利用上述网络提取特征，能够提高模型的检测性能，但检测任务中新类样本数量较少，即使有上下文信息的辅助，分类器仍然会对唐卡中的

相似对象产生混淆(诸如莲花座与须弥座,嘎巴啦与法器中的钵)出现误分类。为了提升模型在检测此类目标的正确率,我们采用RSLoss来监督分类器,RSLoss包括Rank与Sort两个任务。其中Rank任务可根据Classification Lofits区分出样本的正负,使正样本排名高于所有负样本;Sort任务根据IoU对正样本进行降序排列。作为一种基于排名的损失函数,RSLoss与经典的基于评分的函数(如交叉熵损失和焦点损失)相比,根据排名优先考虑正样本,显著的简化了训练,具有更好的鲁棒性。而且RSLoss不需要调整其他任务平衡系数,除学习率以外,无需其他超参数调优,能够避免次优的超参数影响模型性能的情况出现。

对于给定logits和连续的Ground Truth标签,RSLoss为所有正样本当前$l_{RS}(i)$和目标的$l^{*}_{RS}(i)$的RS error的平均值,故RSLoss定义为:

$$l_{RS}:=\frac{1}{|p|}\sum\nolimits_{i\in p}(l_{RS}(i)-l^{*}_{RS}(i)), \tag{5}$$

$l_{RS}(i)$是rank error和sort error的总和:

$$l_{RS}(i)=\underbrace{\frac{N_{FP}(i)}{\mathrm{rank}(i)}}_{(i)\text{: Current ranking error}}+\underbrace{\frac{\sum_{J\in P}H(x_{ij})(1-y_j)}{\mathrm{rank}(i)}}_{l_s(i)\text{: Current sorting error}}, \tag{6}$$

其中:N_{FP}代表所有负例中大于该正例logit的数量,rank代表所有正例和负例中不小于该正例的数量,$H(x_{ij})$在区间$[-\delta_{RS},\delta_{RS}]$中可看作$\frac{x_{ij}}{2\delta_{RS}}+0.5$,$y_i$是真实IoU标签,$P$是正数集合,特别的,当$i\in P$的排名比$i\in N$高时,$N_{FP}(i)=0$,目标排名损失$l^{*}_{RS}(i)$也为0。

通过CFM与DAM得到丰富的上下文信息提高了模型对目标的定位能力,扩充了模型的对目标进行分类的依据,再使用RSLoss代替原交叉熵函数,利用模型定位的质量监督分类器,可以有效提高模型优化的速率和准确度,模型总损失为:

$$L=L_{\mathrm{cls}(RS)}+L_{\mathrm{reg}}+L_{\mathrm{CPE}}, \tag{7}$$

其中:$L_{\mathrm{cls}(RS)}$表示RSLoss分类损失,L_{reg}表示检测损失,L_{CPE}表示对比损失。

在相同实验设置下RSLoss作为分类损失函数时,收敛更快损失值更小,效果如图7所示。

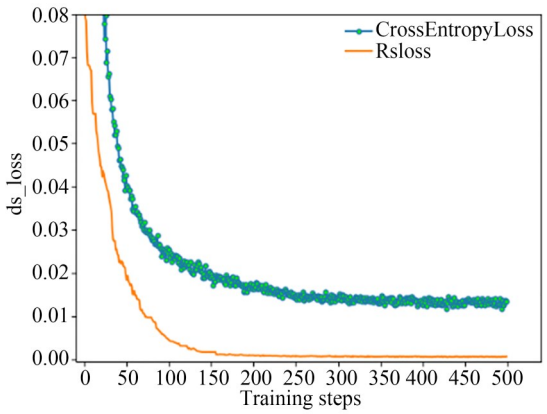


图7　损失函数曲线对比图

Fig. 7　Comparison of loss function curves

3　实验与结果分析

3.1　数据集

为了准确科学地评测FSMC模型在小样本目标检测任务中的性能,本文使用相同的实验环境与参数设置,分别在唐卡数据集与COCO2014[18]数据集上进行实验,验证模型的有效性与普适性。其中唐卡数据集包含883张图片,2 879个目标。检测目标类别共12类,包含头饰3类:宝冠(coronet)、僧帽(mitral)、发髻(chignon);手持物6类:嘎巴啦(gabala)、自然类(nature)、经书(scriptures)、法器(magicInstrument)、兵器(arms)、庄严器具(solemnInstrument);台座3类:莲花座(padmasana)、须弥座(sumeruSeat)、坐垫(cushion)。唐卡数据集中存在数据不均衡的情况,例如:台座对象中莲花座的数据远远大于坐垫,手持物中自然类数量远大于嘎巴啦。为了减少对象数据量对检测结果的影响,我们对图片进行旋转、加随机噪声、镜像等操作实现了数据集的增广与平衡。COCO2014数据集是一个大型的、公开的数据集,涵括了超过33万张自然场景图片,共计80个类别。

3.2　实验环境及参数设置

实验硬件环境:Intel Core i7-7800X CPU,NVIDIA GeForce RTX 3080 Ti GPU,软件环境:Ubuntu18.04操作系统,PyTorch 1.5.0深度学习框架,Python 3.6编程语言,Cuda 10.1环境。

实验中输入图像尺寸设置为1 024×1 024,batch_size设置为16,epoch设置为150,使用随机

梯度下降(Stochastic Gradient Descent,SGD)优化器,动量为0.9,权重衰减为1e-4。

3.3　评价指标

假设检测目标共两类:正样本(Positive)和负样本(Negative),一般通过计算预测框与真实标注框之间的交并比(Intersection Over Union,IoU)对样本进行预测分类。两类目标会出现4种预测结果:正样本并且预测结果为正的样本个数TP(True Positives);负样本但预测结果为正的样本个数FP(False Positives);正样本但预测结果为负的样本数FN(False Negatives);负样本且预测结果为负的样本个数TN(True Negatives)。

准确率(Precision)表示预测正确的正样本在所有被预测为正样本数中的占比,计算表达式为:

$$Precision=\frac{TP}{TP+FP}. \tag{8}$$

召回率(Recall)表示预测正确的正样本在所有预测正确的样本数中的比例,计算表达式为:

$$Recall=\frac{TP}{TP+FN}. \tag{9}$$

一般来说,如果召回率越高,准确率就会越低。AP(Average Precision)指每一类目标的平均准确率。

$$AP=\int_0^1 Precision(R)\mathrm{d}R. \tag{10}$$

对于多目标检测任务,目前最常用的评价标准是平均精准度均值(mean Average Precision,mAP),对所有类的AP值取平均就是mAP。本文采用mAP,AP50和AP75对模型进行性能评估。AP50表示IoU=0.5,AP75表示IoU=0.75。

3.4　实验结果分析

本文在唐卡数据集上构建6-way K-shot的实验机制对数据进行测试,其中$K\in\{10,20\}$。根据表1可以看出,LSTD算法使用迁移学习的通用框架,在唐卡数据集上检测效果较其他算法略差,MPSR[19]算法利用正样本多尺度细化解决了数据集中尺度分布稀疏的问题,其在唐卡数据集上的检测精度高于LSTD与TFA方法,但是效果并没有达到最优。FSCE方法与我们的FSMC方法的检测精度要远高于以上几种算法。在10-shot实验中,FSMC的AP50,AP75,mAP分别为34.6%,19.6%,19.7%,相较于FSCE分别提高了5.0%,1.6%,3.0%。并且随着K的增大,检测性能提升越明显。在20-shot实验中,FSMC的AP50,AP75,mAP比FSCE方法提高了6.2%,6.5%,4.3%。

表1　不同算法在唐卡数据集的检测结果对比

Tab.1　Comparison of detection results of different algorithms in Thangka dataset

Shot	Methods	AP50	AP75	mAP
10	LSTD[12]	11.3	9.7	10.3
	TFA[14]	14.2	9.7	10.8
	MPSR[19]	15.5	13.6	13.2
	FSCE[15]	29.6	18.0	16.7
	FSMC	**34.6**	**19.6**	**19.7**
20	LSTD[12]	14.5	11.7	12.3
	TFA[14]	16.	11.5	14.3
	MPSR[19]	21.7	16.1	14.6
	FSCE[15]	35.6	19.2	18.1
	FSMC	**41.8**	**25.7**	**24.4**

这表明更多的样本能够使模型学习到更多信息,网络的检测效果将更好。若能继续扩充唐卡数据集的规模,本文方法的性能将会得到进一步提升。

为了更加直观地看出FSCE与FSMC算法性能的对比,图8是检测结果可视化效果图。图8(a)是FSCE模型的结果图,可以看出FSCE算法对头饰类目标可以精准的识别,但对手持物类、台座类目标存在漏检现象。如a组中,唐卡中的庄严器具与须弥座目标未被识别出来,同时莲花座的检测精度远低于宝冠。图8(b)是FSMC模型的检测效果图。观察b组图片,可以明显看出改进后的算法整体检测结果明显提高,被FSCE漏检的庄严器具和须弥座也能够准确的识别,说明该模型对漏检现象的改善效果良好。由此说明FSMC方法加入上下文的特征表征并改良分类损失后,对重叠目标中容易被忽略的小目标和易被错误分类的相似目标均能达到较高的识别率,模型整体检测性能明显优于FSCE

模型。

表 2 记录了 FSCE 与 FSMC 两种方法对各类测试目标的检测结果。由表中数据可知 FSCE 模型对头饰类目标的识别率较高，手持物与台座类目标检测效果较差。造成该差异的主要原因在于唐卡本身的构图特点：手持物类对象通常尺寸差别大且相互重叠；莲花座与须弥座两类对象形状相近且往往同时出现，这样的结构导致手持物和台座类对象的检测难度高于头饰类对象。本文在 FSMC 模型中利用图像自身的上下文信息辅助检测，由表 2 数据与图 8 结果可以看出，该方法对重叠目标出现的漏检现象改进效果良好。与 FSCE 模型相比，所有目标的检测准确率均有明显提升，充分说明了改进工作的有效性与必要性。但是改进后的算法在计算时无法彻底规避相邻实例的影响，所以手持物、台座类目标的检测结果与头饰目标的检测结果依然存在差异。

表 2　不同类别目标的检测平均准确率(AP)

Tab. 2　Detection results of different categories

shot	methods	sumeruSeat	padmasana	nature	solemnInstrument	chignon	coronet
10	FSCE	11.7	10.0	17.6	14.2	31.0	35.8
	FSMC	**17.3**	**14.6**	**18.5**	**19.3**	**32.1**	**37.1**
20	FSCE	13.9	13.6	22.4	19.2	33.1	36.3
	FSMC	**21.4**	**18.1**	**22.5**	**25.5**	**33.5**	**36.5**

(a) FSCE方法的检测结果
(a) Experimental results of FSCE

(b) FSMC方法的检测结果
(b) Experimental results of FSMC

图 8　检测效果图

Fig. 8　Experimental results

为验证 FSMC 方法的普适性，在 COCO2014 数据集上执行进一步的实验。使用与 FSCE 原文中相同的训练数据和测试数据，数据集中前 60 个类别被选为训练集，其余 20 个类别被选为测试集。在表 3 中，展示了 FSCE 和 FSMC 在 20-way 10-shot 和 20-way 20-shot 实验机制下的实验结果。与 FSCE 模型相比，本文模型在 10-shot 的实验机制下 AP50，AP75，mAP 上分别提高了 1.7%，3.4%，2.3%；在 20-shot 机制下分别提高了 3.1%，1.7%，3.4%，证明该模型对公共数据集的目标检测仍然有较好的性能，也进一步说明本文方法鲁棒性较高，能够适应不同的数据集。

3.5　消融实验

这里，在唐卡数据集上进一步分析各组件对

表 3　COCO 数据集的检测结果对比

Tab. 3　Comparison of detection results of COCO dataset

Shot	Methods	AP50	AP75	mAP
10	FSCE	10.4	7.2	8.9
	FSMC	**12.1**	**10.6**	**11.2**
20	FSCE	12.5	12.7	11.5
	FSMC	**15.6**	**14.4**	**14.9**

表 4　唐卡数据集上的消融实验结果

Table 4　Ablation experiment results on Thangka dataset

Shot	Methods	AP50	AP75	mAP	Parameters/10^4(M)	Average detection time/s
10	FSCE	29.6	18.0	16.7	1.12	0.050
	FSCE+CFM	32.6	18.2	18.0	—	—
	FSCE+CFM+DAM	33.9	19.1	19.4	—	—
	FSCE+CFM+DAM+RSLoss (FSMC)	**34.6**	**19.6**	**19.7**	**1.17**	**0.093**

提升模型性能的贡献。在这一部分中依次应用CFM、DAM与RSLoss,在相同数据集和实验设置条件下,逐步分析各模块的作用,验证各模块的有效性,结果如表4所示。

可以看出在嵌入各模块后,均能获得比基线模型更好的性能。其中,加入CFM后AP50,AP75,mAP分别提升了3.0%,0.2%,1.3%。这主要是因为:CFM在特征提取过程中采用不同扩张率的卷积层提取上下文信息,提高了模型对不同尺度对象的处理能力,降低了小目标被忽略的概率。此外,采用密集连接使网络浅层的特征在网络深层被复用,通过特征融合的方式充分地利用了粗粒度与细粒度特征之间的互补信息,这样捕获到的大量上下文信息帮助模型提高了对目标的定位和分类能力,但同时也引入了一些无用的信息从而影响了平均检测精度。加入双注意力机制DAM后,检测结果提高了1.3%,0.9%,1.4%。通过双注意力引导模块,网络聚焦于对象的特征信息,抑制了噪声对结果的干扰,使得模型对重叠、遮挡对象的识别能力进一步提升。将全部模块聚合后检测效果达到最佳,检测精度提升效果相当明显,AP50提升了5%。但与此同时模型的参数量与平均检测时间也发生了改变,改进后的模型由于增加了密集连接的多层卷积网络,实现有效特征最大化利用的同时导致模型参数量提高了4.4%。平均检测时间变化也较明显,主要原因是FSMC所使用的RSLoss的计算过程比原来的交叉熵损失函数更加复杂,需要的处理时间也更多。但是利用性能更优的处理器与不同的参数设置能够有效降低算法检测效率之间的差距。经过对各项指标的综合分析,本文提出的FSMC以牺牲少量的处理时间为代价提升了检测精度,更加适用于小样本目标检测,特别针对构图复杂,目标有重叠的图像检测效果更优。

4　结　论

将目标检测方法引入唐卡图像理解的研究,对于传统文化的传承与非物质文化遗产的数字化保护工作具有重要意义。本文在深入分析唐卡图像特点的基础上,提出了结合多尺度上下文信息与双注意力引导的小样本目标检测算法。首先,在特征提取阶段构建了一种新的能够提取到丰富上下文特征的多尺度特征金字塔,具体为:采用空洞卷积获取到不同大小的感受野,提升模型对多尺度对象的泛化能力,利用上下文信息提高模型的定位能力,并为模型提供更多的分类依据。其次,在特征金字塔末端引入双注意力引导模块,在空间与通道两个维度上对特征之间的依赖关系进行建模,增强重要特征的表示,从而提升检测准确率,同时进一步降低冗余信息对检测结果的影响。最后,利用RSLoss代替交叉熵损失函数监督分类器,有效地提升网络的分类性能。在唐卡数据集和COCO数据集上的10-shot实验机制下mAP分别达到了19.7%,11.5%;在20-shot实验机制下,mAP分别达到了24.4%,14.9%,实验结果表明本文提出的方法能够适应各类数据集并且性能提升明显。消融实验证明了改进的各模块的有效性,但该方法依然具有较大的提升空间。后续工作可以通过提升唐卡数据集质量,利用对比学习的知识优化对比分支等多种方式进一

步提升检测性能。此外，本文引入注意力机制与RSLoss后增大了训练模型的计算量，后续可在保障检测性能的前提下进一步研究如何提升检测效率。

参考文献：

[1] 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*. 23-28, 2014, *Columbus*, *OH*, *USA*. IEEE, 2014: 580-587.

[2] GIRSHICK R. Fast R-CNN[C]. 2015 *IEEE International Conference on Computer Vision* (*ICCV*). 7-13, 2015, *Santiago*, *Chile*. IEEE, 2016: 1440-1448.

[3] 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.

[4] LIU W, ANGUELOV D, ERHAN D, *et al*. *SSD*: *Single Shot Multibox Detector*[M]. Computer Vision - ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[5] 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*). 27-30, 2016, *Las Vegas*, *NV*, *USA*. IEEE, 2016: 779-788.

[6] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 6517-6525.

[7] REDMON J, FARHADI A. YOLOv3: An Incremental Improvement [EB/OL]. 2018: *arXiv*: 1804.02767. https://arxiv.org/abs/1804.02767

[8] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection [EB/OL]. 2020: *arXiv*: 2004.10934. https://arxiv.org/abs/2004.10934

[9] CHEN Q, WANG Y M, YANG T, *et al*. You Only Look One-Level Feature[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 13034-13043.

[10] KARLINSKY L, SHTOK J, HARARY S, *et al*. RepMet: Representative-Based Metric Learning for Classification and Few-Shot Object Detection[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 5192-5201.

[11] YAN X P, CHEN Z L, XU A N, *et al*. Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27-*November* 2, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 9576-9585.

[12] CHEN H, WANG Y L, WANG G Y, *et al*. LSTD: A Low-Shot Transfer Detector for Object Detection[C]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, 32(1).

[13] KANG B Y, LIU Z, WANG X, *et al*. Few-Shot Object Detection via Feature Reweighting[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27 - *November* 2, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 8419-8428.

[14] WANG X, HUANG T E, DARRELL T, *et al*. Frustratingly Simple Few-Shot Object Detection [EB/OL]. 2020: *arXiv*: 2003.06957. https://arxiv.org/abs/2003.06957

[15] SUN B, LI B H, CAI S C, *et al*. FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 7348-7358.

[16] OKSUZ K, CAM B C, AKBAS E, *et al*. Rank & Sort Loss for Object Detection and Instance Segmentation[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 2989-2998.

[17] 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*). 21-26, 2017,

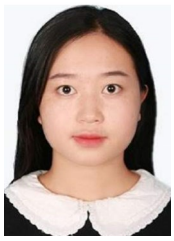
Honolulu, *HI*, *USA*. IEEE, 2017: 936-944.

[18] LIN T Y, MAIRE M, BELONGIE S, *et al*. Microsoft COCO: Common Objects in Context[EB/OL]. 2014: *arXiv*: 1405.0312. https://arxiv.org/abs/1405.0312

[19] WU J X, LIU S T, HUANG D, *et al*. *Multi-Scale Positive Sample Refinement for Few-Shot Object Detection*[M]. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 456-472.

作者简介:

胡文瑾(1981—),女,甘肃庆阳人,博士,教授。2015年3月毕业于兰州理工大学,获工学博士学位。现为西北民族大学数学与计算机科学学院教授,主要从事图像处理、计算机视觉等方面的研究。E-mail: wenjin_zhm@126.com



唐慧媛(1998—),女,甘肃武威人,硕士研究生,就读于西北民族大学中国民族信息技术研究院。主要从事图像处理、计算机视觉等算法研究。E-mail: tanghy981213@163.com