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

匹配多尺度特征与预测任务的实时目标检测

杜鸿杰*，孙汉卿，曹家乐，庞彦伟

天津大学电气自动化与信息工程学院，天津 300072

摘要 在基于卷积神经网络的目标检测算法中，浅层高分辨率特征包含更多细节信息，有助于抽象特征完成精确的定位任务；深层特征包含抽象的语义信息，更适合目标存在性预测任务。研究发现，现有的不基于先验框的检测方法直接在同一特征图上预测所有任务时，并没有匹配上述特征与预测任务，而这一特征与任务不匹配的问题限制了检测精度。为解决这一问题，提出了一种匹配目标多尺度特征与预测任务的实时目标检测算法，简称 MFT 检测器。以 CenterNet 检测器为基础，同时完成浅层细节特征与精确定位任务的匹配，多尺度多感受野抽象特征与目标存在性预测任务的匹配。实验结果表明，所设计的 MFT 检测器缓解了特征与预测任务不匹配的问题，从而显著提高了检测精度，且检测速度保持在 94.5 frame/s，能够保证检测实时性。

关键词 图像处理；实时目标检测；卷积神经网络；多尺度特征；匹配

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.1210014

Matching Multi-Scale Features and Prediction Tasks for Real-Time Object Detection

Du Hongjie*, Sun Hanqing, Cao Jiale, Pang Yanwei

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

Abstract In object detection algorithms based on convolutional neural networks, high-resolution features from lower levels contain more detailed information, which can help the abstract features complete the accurate positioning task; deep-level features contain abstract semantic information, which is more suitable for target existence prediction task. When the most existing anchor-free detection method directly predicts all tasks on the same feature map, it does not match the above features and prediction tasks, which limits the detection accuracy. To this end, the MFT detector, a real-time object detection algorithm, is proposed to match multi-scale features and prediction tasks of targets. MFT detector is based on CenterNet detector, which can match shallow detail features with accurate positioning task, and match multi-scale, multi receptive field abstract features with target existence prediction task. Experimental results show that the proposed MFT detector alleviates the mismatch between features and prediction tasks, and significantly improves the detection precision while maintaining a high speed of 94.5 frame/s, which meets the requirement of a real-time vision system.

Key words image processing; real-time object detection; convolutional neural network; multi-scale feature; match

OCIS codes 100.4996; 150.1135; 070.5010

1 引 言

近几年，深度卷积神经网络(CNN)发展迅速，在当前精度最高的目标检测器中占据了重要地位[1-6]。但现有的基于深度 CNN 的检测器计算代价较大[7]，检测延时，导致很难实际应用于机器人、无人机、自动驾驶等实时视觉系统。为了加快神经网络在系统中的计算速度，已有研究提出 SSD[8]、

收稿日期：2020-09-25；**修回日期**：2020-10-14；**录用日期**：2020-10-21

基金项目：国家自然科学基金(61906131)

*E-mail：duhongjie@tju.edu.cn

YOLO[9]等实时检测器，但这些检测器的精度仍然难以满足视觉系统的实际需求。为了解决这一问题，本文提出了一种以低计算成本提高轻量级检测网络精度的实时目标检测算法，简称 MFT 检测器。

CNN 中虽然深层特征图抽象程度高，语义信息丰富，但其空间分辨率较低，位置信息相对匮乏，相对而言难以完成预测目标精确位置的任务[10-12]，这一特征属性与任务属性不匹配的问题限制了现有检测网络的精度。ConRetinaNet[11]发现 RetinaNet[13]中存在用于分类的先验框特征与分类任务不匹配的问题，所以同时使用先验框特征和预测框特征进行目标分类。Cao 等[12]发现单阶段检测器中用来预测目标类别属性的包围框与最终回归得到的包围框不一致的问题，提出了 HSD 网络，该网络先预测位置属性，再预测目标类别。

与上述方案不同，本文通过研究现有不基于先验框的检测算法[14-15]，创新地提出了 MFT 实时检测算法，同时完成浅层特征与定位任务的匹配、多尺度多感受野特征与目标存在性预测任务的匹配，解决了现有目标检测方法中存在的特征属性与任务属性不匹配的问题，提高了实时目标检测精度。本文主要贡献如下。

1）为解决不匹配问题，提出匹配多尺度特征属性与目标检测任务属性的实时目标检测器，该 MFT 检测器同时匹配了浅层特征与精确定位任务、多尺度多感受野特征与目标存在性预测任务，能够获取更高的检测精度，满足实时视觉系统的低延时需求。

2）为匹配多尺度多感受野特征与目标存在性预测任务，设计了多尺度和多感受野目标存在性预测模块（简称 MSH 模块和 MRFH 模块），分别在多分辨率多层级特征和小分辨率多感受野特征上进行目标存在性预测。进一步地，为融合 MSH/MRFH 的目标存在性预测结果，设计了一种自学习权重的目标存在性融合机制。三种策略有效地提高了目标存在性的预测精度。

3）为匹配特征属性与目标定位任务，高效地复用 CNN 中位置信息丰富的浅层特征，引入高效互补特征复用模块（简称 ECF 模块），该模块利用浅层特征增强深层特征。不同层级特征的对应融合有效地提高了目标检测分类和定位任务的精度。

2 MFT 实时目标检测算法

对比了现有的两类目标检测方案[如图 1(a)和(b)所示]与所提 MFT 算法[如图 1(c)所示]，图 1 中每个子图左列表示基础网络的特征图，中间列表示抽象多尺度特征图，右列表示网络预测任务。具体地，以图 1(a)的 CenterNet 检测算法[14]为例，展示在最深层特征图上同时完成目标存在性热图（Heatmap）、中心点偏移（Offset）、包围框尺寸（Size）三个预测任务；图 1(b)所示的算法[7,15]利用多尺度特征并对每组特征都进行了全部预测，没有考虑小分辨率特征不适宜精确定位的问题；图 1(c)所示的所提 MFT 算法利用多尺度特征进行预测任务的同时，考虑到一些抽象层特征与定位任务不匹配的问题，最终匹配地进行多尺度目标检测。

所提 MFT 网络结构主要包括基础卷积网络部分、ECF 模块、MSH 模块和 MRFH 模块、自学习权重的目标存在性融合机制。MFT 整体网络架构如图 2 所示，以 CenterNet[14]为基础；为了匹配浅层特征的位置信息和定位任务，引入 ECF 模块增强主干网络特征（图 2 三个正方块和三个圆圈部分）；设计 MSH/MRFH 模块和自学习权重的目标存在性融合机制来匹配特征尺度特性与目标存在性预测任务，获取高质量预测结果（Heatmap）。将 ECF 模块增强后的特征分别经过所设计的 MSH 模块和 MRFH 模块，分别在多分辨率多层级特征和小分辨率多感受野特征上进行目标存在性预测；最后，为了融合 MSH 模块和 MRFH 模块的预测结果，利用自学习权重的目标存在性融合机制，得到最终的目标存在性预测。将所得到的高

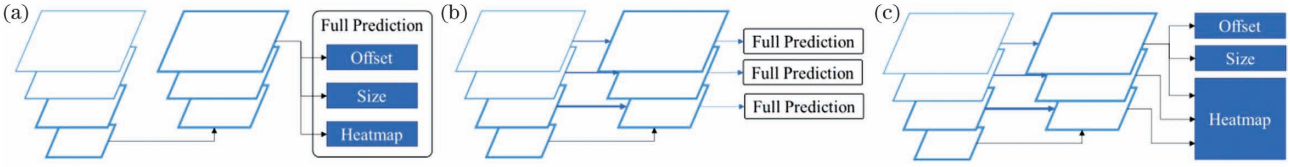

图 1 现有利用卷积特征完成预测任务的目标检测算法方案。(a)CenterNet 检测模型；(b)利用多尺度特征进行全部任务预测的检测模型；(c)所提 MFT 检测模型

Fig. 1 Existing object detection algorithms using convolution feature to complete prediction task. (a) CenterNet detection model; (b) detection model based on multi-scale feature for total task prediction; (c) proposed MFT detection model

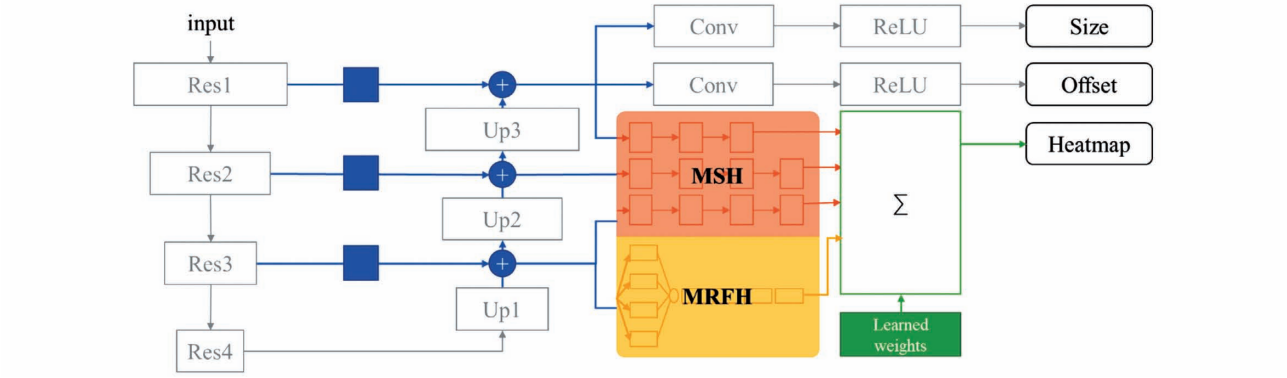

图 2　所提 MFT 网络结构

Fig. 2　The proposed MFT network structure

质量目标存在性预测结果与包围框大小(Size)、目标中心点偏移量(Offset)结合，得到目标类别属性与位置属性。

2.1　CenterNet

CenterNet[14]是一种先进的不基于先验框的单阶段目标检测器，使用目标中心点位置的特征向量预测目标所有的属性，用关键点估计的方法找到目标中心点，从而确定该位置目标的类别和位置属性。基础网部分使用常见的卷积分类网络，如 ResNet[16]、DLA[17]、Hourglass[18]等。以 ResNet 为例，输入图像首先经过 ResNet，得到下采样 32 倍的最小分辨率特征；随后使用 3 个 2 倍上采样反卷积模块，得到分辨率为输入图像分辨率 1/4 的预测特征图。

目标存在性(Heatmap)的预测表示方式为一个关键点特征图 $\hat{\boldsymbol{Y}}\in[0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$，其中 R 是特征下采样倍数($R=4$)，C 是数据集中物体的类别数目，$W\times H$ 是输入图像的分辨率。$|\hat{\boldsymbol{Y}}_{x,y,c}|=1$ 表示检测到目标关键点，$|\hat{\boldsymbol{Y}}_{x,y,c}|=0$ 表示检测到背景。$\boldsymbol{Y}_{x,y,c}$ 为未归一化高斯方法生成的热图真实值[19]，$\boldsymbol{Y}\in[0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$。目标存在性的训练损失函数使用 Focal loss[13,19]的变体形式：

$$L_{\mathrm{k}}=\begin{cases}-\dfrac{1}{N}\displaystyle\sum_{xyc}(1-|\hat{\boldsymbol{Y}}_{x,y,c}|)^{\alpha}\log(|\hat{\boldsymbol{Y}}_{x,y,c}|), & |\boldsymbol{Y}_{x,y,c}|=1\\ -\dfrac{1}{N}\displaystyle\sum_{xyc}(1-|\boldsymbol{Y}_{x,y,c}|)^{\beta}(|\hat{\boldsymbol{Y}}_{x,y,c}|)^{\alpha}\log(|1-\hat{\boldsymbol{Y}}_{x,y,c}|), & |\boldsymbol{Y}_{x,y,c}|\neq 1\end{cases},\tag{1}$$

式中：N 为关键点的个数；α，β 均为 Focal loss 的超参数，取 $\alpha=2$，$\beta=4$。

目标中心点偏移量(Offset)是预测每个目标中心点位置沿 x，y 两个方向的偏移量 $\hat{\boldsymbol{O}}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$，每个类别的关键点 $\boldsymbol{p}\in\mathbb{R}^2$ 都对应一个低分辨率的关键点真实值 $\tilde{\boldsymbol{p}}=\mathrm{floor}\left(\dfrac{\boldsymbol{p}}{R}\right)$。在训练过程中使用 L1 损失函数对存在目标的特征点得目标中心偏移量进行监督，表达式为

$$L_{\mathrm{off}}=\frac{1}{N}\sum_{\boldsymbol{p}}\left|\hat{\boldsymbol{O}}_{\tilde{\boldsymbol{p}}}-\left(\frac{\boldsymbol{p}}{R}-\tilde{\boldsymbol{p}}\right)\right|。\tag{2}$$

包围框大小(Size)是预测目标 k 在关键点 $\boldsymbol{p}$ 处的包围框尺寸 $\hat{\boldsymbol{S}}_{\boldsymbol{p}_k}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$，分别预测包围框的长度与宽度。目标 k 的包围框大小真实值为 $\boldsymbol{s}_k$，同样使用 L1 损失函数对目标中心偏移量进行监督，表达式为

$$L_{\mathrm{size}}=\frac{1}{N}\sum_{k=1}^{N}|\hat{\boldsymbol{S}}_{\boldsymbol{p}_k}-\boldsymbol{s}_k|。\tag{3}$$

训练模型的总损失函数记为 $L_{\mathrm{det}}=L_{\mathrm{k}}+\lambda_{\mathrm{size}}L_{\mathrm{size}}+\lambda_{\mathrm{off}}L_{\mathrm{off}}$，其中 $\lambda_{\mathrm{size}}=0.1$，$\lambda_{\mathrm{off}}=1$。整个检测网络在特征图中每个位置的输出是 $C+4$ 维向量，其中 C 维特征表示预测的目标存在性，包围框大小与目标中心偏移量分别用两维特征表示。CenterNet 所预测的 C 类特征图为目标存在性热图，不仅包含目标类别信息，也包含目标中心位置的粗略信息；而预测的包围框的长宽、中心点位置的偏移可以精调包围框的位置，获得目标更精细的位置信息。最终网络通过预测三个目标属性进行目标检测。

2.2　高效互补特征复用模块

卷积神经网络中的深层特征作为检测网络的预测特征起着至关重要的作用，但随着卷积神经网络

的加深，深层网络会丢失图像的位置和细节信息；而分辨率较高的浅层特征，带有丰富的局部和位置信息，有助于定位任务。为解决深层特征属性与位置预测任务不匹配的问题，受 FPN[7] 的启发，设计 ECF 模块，丰富抽象特征的细节信息，复用浅层高分辨率特征。

原始的 CenterNet 直接使用分辨率最大的抽象特征预测所有的检测任务（目标存在性、包围框大小与目标中心偏移量），但该特征在卷积神经网络的最深层位置，其位置信息匮乏的特性与目标定位任务需求不匹配，因此设计了 ECF 模块，对浅层高分辨率特征进行 1×1 卷积后与深层抽象特征进行融合，以极小的计算代价，提供细节信息，补充抽象特征，将浅层特征与位置预测任务二者的特点匹配起来。

2.3　多尺度目标存在性预测模块

MSH 模块是所提检测算法的重要组成部分之一。MSH 模块使用不同分辨率特征进行目标存在性检测，组成多分辨率目标存在性金字塔，其中大-中-小三级金字塔分别着重预测小-中-大三种尺度目标的存在性，降低了漏检率，显著提高了各尺度物体的检测精度。

FPN[7]、SSD[8]、FCOS[15] 等方法虽然也采用了多分辨率特征预测目标，在每个特征图上同时进行目标类别与位置预测两个任务，但不同特征图的定位能力、细节信息、抽象程度不同，这种特征属性与任务属性不匹配的问题限制了现有检测网络的精度[10,12]。在设计的 MSH 模块中，深层大分辨率特征预测任务保持与 CenterNet 一致，即预测目标存在性、包围框大小、目标中心偏移量三个任务；其余分辨率的特征只预测目标存在性，即只预测目标的类别属性与目标粗略的位置属性，缓解了特征属性与任务属性不匹配的问题。

为了获得多尺度目标存在性，复用多个分辨率的抽象特征，设计了图 3 所示的 MSH 模块。MSH 模块的输入为三个不同分辨率的深层特征图 $\boldsymbol{F}_1$，$\boldsymbol{F}_2$，$\boldsymbol{F}_3$，输出为三个同分辨率的目标存在性预测特征 $\boldsymbol{P}_1$，$\boldsymbol{P}_2$，$\boldsymbol{P}_3$。

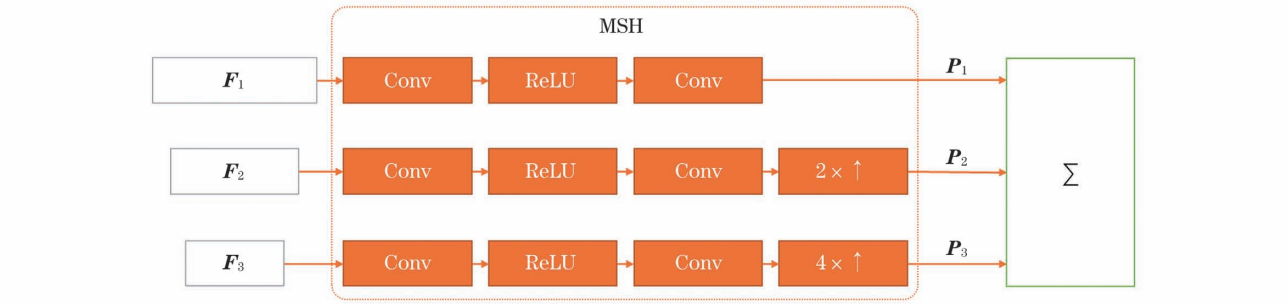

图 3　MSH 模块结构

Fig. 3　MSH module architecture

常用的多尺度预测方案分别在 $\boldsymbol{F}_1$，$\boldsymbol{F}_2$，$\boldsymbol{F}_3$ 上进行相同多个目标属性的预测，并将所有目标级的预测结果融合。但这种方案受到多尺度特征属性与定位任务属性不匹配问题的影响，不能完全发挥出所复用的各个尺度特征的优势。具体地，$\boldsymbol{F}_2$，$\boldsymbol{F}_3$ 虽然融合了一部分浅层特征，补充了特征中缺少的位置信息，但其分辨率较小（分别为 8 倍、16 倍下采样特征），难以完成精确的包围框大小与目标中心偏移量预测，只能胜任粗略的目标存在性估计。针对特征的这一属性，MSH 模块中匹配地使用特征 $\boldsymbol{F}_2$，$\boldsymbol{F}_3$，只对目标存在性这一属性进行预测。$\boldsymbol{F}_2$，$\boldsymbol{F}_3$ 经过一个 3×3 卷积、1×1 卷积后，得到了不同分辨率的存在性特征，再用双线性插值法上采样特征图，得到与 $\boldsymbol{P}_1$ 空间尺寸一致的目标存在性预测特征 $\boldsymbol{P}_2$，$\boldsymbol{P}_3$。每一个特征经 MSH 模块后得到的目标存在性预测可表示为

$$\boldsymbol{P}_i = H_i(\boldsymbol{F}_i),\ i=1,2,3, \tag{4}$$

式中：H_i 为 MSH 模块的第 i 个分支；$\boldsymbol{P}_i$ 为第 i 层特征经过 H_i 处理后得到的目标存在性预测特征。

MSH 模块以较小的计算代价捕获到了基础网络中不同尺度目标的语义特征，复用多尺度特征预测目标存在性，缓解了特征属性与任务属性不匹配的问题，进而提高实时目标检测网络的检测性能。MSH 模块用到三个不同分辨率的抽象特征，其中最大分辨率特征处在特征提取网络的最深层，最小分辨率特征处在抽象程度不够深的位置，深度上的差异使得三个特征图 $\boldsymbol{F}_1$，$\boldsymbol{F}_2$，$\boldsymbol{F}_3$ 抽象能力不匹配。为了弥补这一特征抽象能力不匹配造成的语义鸿沟问题，进一步提出 MRFH 模块。

2.4　多感受野目标存在性预测模块

为了获得目标多感受野特征，缓解 MSH 模块

中 $\boldsymbol{F}_1$，$\boldsymbol{F}_3$ 抽象能力不匹配的语义鸿沟问题，受语义分割方法捕获目标不同尺寸的上下文信息启发，借鉴 SPP[20]、ASPP[21] 的思路，设计了图 4 所示的 MRFH 模块。

MRFH 模块对 $\boldsymbol{F}_3$ 特征进行多感受野的语义信息挖掘，获得目标多感受野特征，采用不同膨胀率的膨胀卷积对分辨率较小的特征图进行多尺度语义提取，得到不同感受野的特征后，将它们拼接到一起，充分提取目标的上下文信息，同时减小了与最深层特征之间的语义鸿沟。MRFH 模块使用 4 个具有不同膨胀率的膨胀卷积 D_1，D_2，D_3，D_4，通过调整滤波器大小来调整感受野，使用的膨胀卷积在不增加参数量的情况下灵活扩大感受野，增加了小尺度特征的信息丰富度。特征 $\boldsymbol{F}_3$ 经 MRFH 模块后，得到的目标存在性预测特征可表示为

$$\boldsymbol{P}_4 = H_4\{\mathrm{cat}[D_1(\boldsymbol{F}_3), D_2(\boldsymbol{F}_3), D_3(\boldsymbol{F}_3), D_4(\boldsymbol{F}_3)]\}, \tag{5}$$

式中：cat(•) 为特征拼接；$D_{i'}$ 为第 i' 个膨胀卷积（$i'=1,2,3,4$），其膨胀率分别为 $d_1=1$，$d_2=6$，$d_3=12$，$d_4=18$；H_4 为预测分支。

MRFH 模块使用不同膨胀率的膨胀卷积充分挖掘 $\boldsymbol{F}_3$ 特征，预测多感受野的目标存在性，既增强了小尺寸特征信息丰富度，又减小与 $\boldsymbol{F}_1$ 特征之间的语义鸿沟，同时与所设计的 MSH 模块有互补。为了融合 MRFH 和 MSH 两个模块的预测结果，进一步提出了自学习权重的目标存在性融合机制。

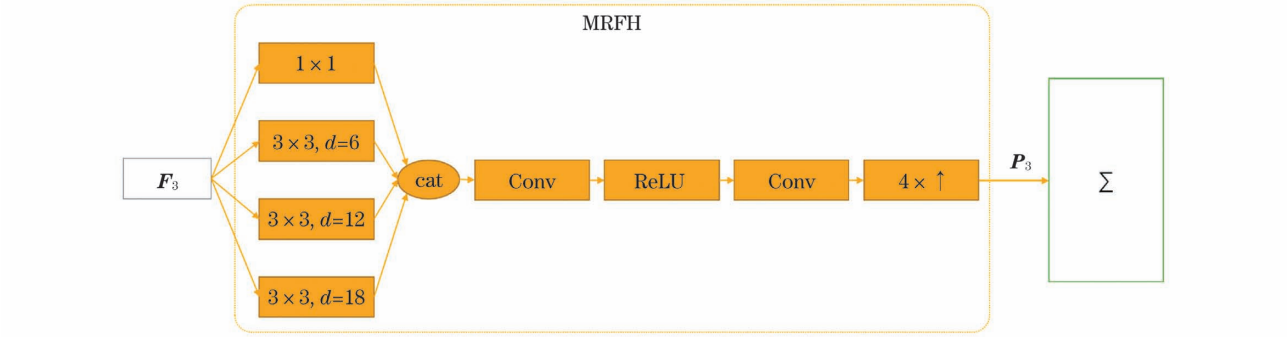

图 4　MRFH 模块结构

Fig. 4　MRFH module architecture

2.5　自学习权重的目标存在性融合机制

MFT 模型中的两个模块 MSH 和 MRFH 分别从多种分辨率、多种感受野的特征预测了目标存在性，为有效地融合这些预测结果，提出自学习权重的目标存在性融合机制。

一种简单的融合策略，如图 5(a)所示，分别根据所有预测结果计算目标属性信息，然后再通过 NMS 算法[22]将重复的目标包围框删除。这种策略虽然简单，但是 NMS 算法难以在 GPU 等计算设备上进行并行优化，因此不能满足实时检测系统的要求。

所提自学习权重的目标存在性融合机制如图 5(c)所示，通过自学习权重的方式对 MSH 模块和 MRFH 模块得到的不同特征图进行有机融合，避免图 5(a)方案的 NMS 后处理步骤的同时，弥补按照经验设置权重[如图 5(b)所示]而无法广泛适用各种应用场景的缺陷，且计算量小，适合在实时视觉系统中使用。所提自学习权重的目标存在性融合机制可表示为

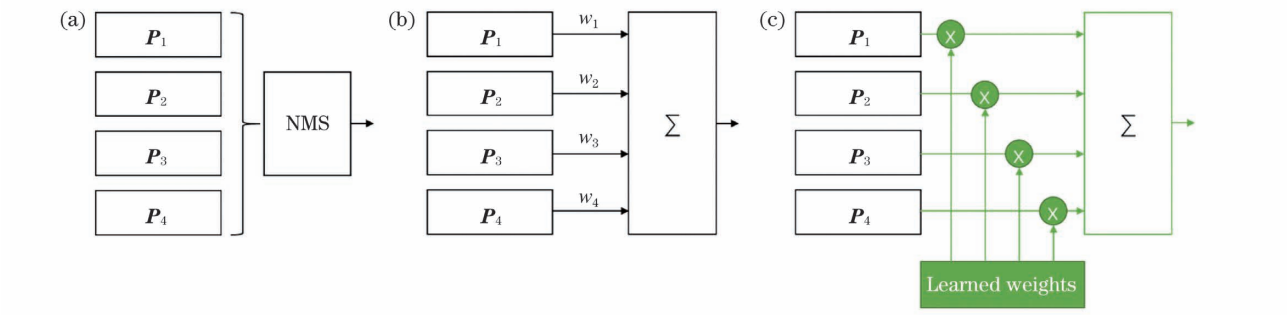

图 5　几种不同的特征融合方法结构

Fig. 5　Architectures of different feature fusion methods

$$\boldsymbol{P}=\sum_{i'=1}^{4}w_{i'}\boldsymbol{P}_{i'},\tag{6}$$

式中：$w_{i'}$为所学习到的融合权重。最终通过加权的方式融合得到目标存在性(Heatmap)预测结果$\boldsymbol{P}$，与包围框大小(Size)、目标中心偏移量(Offset)结合，得到目标类别属性和精确的位置属性，完成检测任务。

3 实验与分析

在 PASCAL VOC[23] 和 COCO[24] 两个数据集上对所提 MFT 检测器进行实验验证。MFT 结构有多种可选的基础网络，本文旨在提高轻量级网络检测精度，解决特征属性与任务属性不匹配的问题，因此实验部分，MFT 所选择的主干网络为规模较小的 18 层与 50 层的 ResNet。

3.1 实验配置及实验数据集

MFT 主干网络使用在 ImageNet 数据集[25] 预训练的权重，使用 Xavier[26] 对模型反卷积层权重进行随机初始化。训练过程中对输入数据采用随机翻转、随机放缩、随机剪裁等数据增强方法，采用 Adam 优化器[27] 进行优化。

PASCAL VOC 数据集包括 VOC2007 和 VOC2012，均包含 20 类目标。其中 VOC2007 数据集中训练集图像有 5011 张，测试集有 4952 张；VOC2012 训练集中图像有 5727 张，验证集中图像有 5823 张，测试集包含 10911 张图像。消融实验是在 PASCAL VOC 数据集上展开的，实验数据集划分方式延用文献[28]：VOC2007 的训练集、VOC2012 的训练集和 VOC2012 的验证集共有 16551 张图像，使用 PASCAL VOC2007 的测试集。采用预测框与真实值交并比(IoU)阈值为 0.5 的多类别平均检测精度(mAP)进行检测性能评估。

COCO 是计算机视觉中常用的数据集，比 PASCAL VOC 规模更大、更富挑战性，共包含 80 类目标。使用的 COCO 2017 数据集包含 118000 张训练图像和 5000 张验证图像，其中对比实验是在 COCO 数据集上展开的，实验数据集划分方式延用文献[28]：训练集使用 COCO 的训练集，测试集使用 COCO 的验证集，采用 IoU 阈值(0.5∶0.95∶0.05)下的 mAP 作为检测器检测精度的指标。

3.2 对比实验

训练 MFT 检测器的方法延用 CenterNet 的训练策略，共训练 140 轮，初始学习率为 0.00025，在第 90 轮和第 120 轮时学习率下降 1/10。实验模型的训练与测试均使用两块 NVIDIA GTX 1070 显卡，批大小(batch_size)为 50。表 1 为 MFT 检测器和几个先进的检测器在 COCO 验证集上的检测结果，对检测速度进行比较时使用 NVIDIA GTX 1080Ti。通常来说，工业摄像机的拍摄速度为 15～30 Hz，因此检测速度(V)在大于 30 frame/s 的情况下认为算法满足实时性。但实际系统的处理时间不只包含检测算法的运算时间，考虑到数据的读取、预处理、输出等时间，在实践中取 60 frame/s 为分界线划分检测算法实时性。从表 1 可以看出：主干网络为 ResNet-18 时，MFT 检测器的检测精度比基线方法 CenterNet 精度高 3.4 个百分点，精度和速度超过其他多尺度预测方法，如 FCOS[15]，SSD[8]；主干网络为 ResNet-50 时，与 YOLOv3[29] 相比，速度提升近 1 倍，MFT 检测器的检测速度仍可以满足实时检测要求，同时精度甚至高于一些主干网络为 ResNet-101、VGG16 等大型深层网络的检测器。

表 1 COCO 数据集上不同目标检测算法的检测性能对比

Table 1 Comparison of different object detection algorithms on the COCO dataset

Condition	Method	Backbone	Size	V / (frame·s^{-1})	mAP /%	AP_S/%	AP_M/%	AP_L/%
V>60 frame/s	SSD[8]	VGG16	300×300	60.6	23.2	5.3	23.2	39.6
	SSD[8]	MobileNetV2	512×512	110.7	22.1	5.8	16.9	43.6
	CenterNet[14]	Res18	512×512	128.5	28.1	10.1	31.5	42.6
	TTFNet[30]	Res18	512×512	112.3	28.1	11.8	29.5	41.5
	MTF	Res18	512×512	94.5	31.5	14.9	35.3	44.3
V<60 frame/s	FCOS[15]	Res18	1330×800	20.8	26.9	13.9	28.9	36
	CenterNet[14]	Res101	512×512	45.1	34.6	10.1	31.5	42.6
	SSD[8]	VGG16	512×512	23.4	26.8	9.0	28.9	41.9
	YOLOv3[29]	D53	608×608	30.3	33.0	18.3	25.4	41.9
	EfficientDet[28]	EfficientNet	512×512	47.1	33.8	12.4	34.7	54.4
	MTF	Res50	512×512	54.9	35.3	12.9	34.3	44.3

图 6 与图 7 为 MFT 检测器与 CenterNet 检测结果的对比图，检测结果均是在 ResNet-18 为主干网络的前提下进行的。从图 6 可以看出，对于同一类别、尺度多变的目标（图 6 中的类别“人”），MFT 检测器的漏检现象明显少于原有方法，对于重叠、遮挡目标也有更好的适应性。而对于一张测试图片中出现的多种尺度的物体，如图 7 所示，MFT 检测器可以同时检测到大尺度物体（如球类运动员）和小尺度物体（如球拍和球）。这些有益效果来源于所设计的多尺度特征属性与任务属性匹配的检测器结构，具有多尺度、多感受野的特征能够承担各种尺度物体的预测任务，显著降低了漏检，提高了各个尺度物体的检测精度。

图 6　COCO 数据集上 CenterNet 与 MFT 检测器的视觉效果对比

Fig. 6　Visual effect comparison of CenterNet and MFT detector on COCO dataset

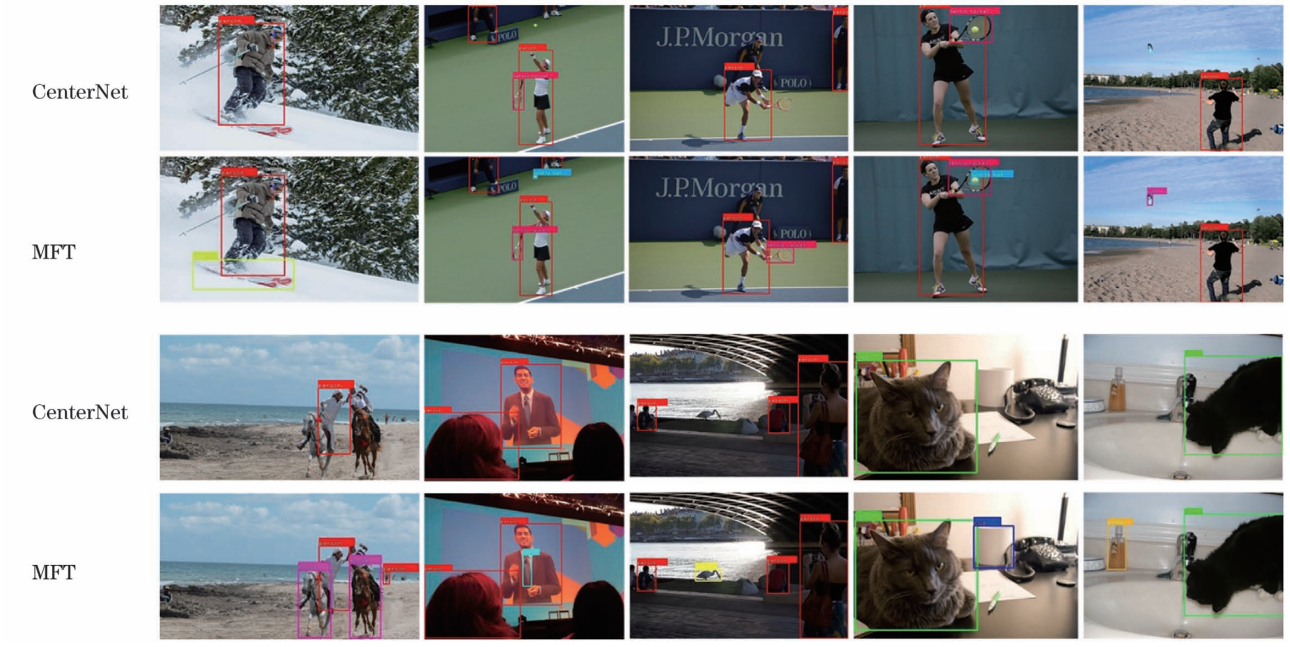

图 7　COCO 数据集上 CenterNet 与 MFT 检测器的视觉效果对比

Fig. 7　Visual effect comparison of CenterNet and MFT detector on COCO dataset

3.3 MFT 消融实验

为更好地验证 MSH、MRFH、ECF、自学习权重融合模块的有效性，在 PASCAL VOC 数据集上对各个模块进行消融实验，共训练 70 轮，初始学习率为 0.000125，在第 45 轮和第 60 轮后学习率下降 1/10。

以 CenterNet 为基础，逐个添加 ECF、MSH、MRFH、自学习权重融合模块（LWS），实验结果如表 2 所示。可以看出：和基于 ResNet-18 的 CenterNet 相比，ECF 通过复用基础网络互补特征，匹配浅层特征与定位任务，提高了 1.64 个百分点；MSH 模块和 MRFH 模块匹配多尺度、多感受野特征与目标存在性预测任务，两个模块同时预测目标存在性时，可以捕获到更多尺度的目标，提高了 1.58 个百分点，验证了二者的互补关系；当同时添加 MSH、MRFH、ECF、LWS 四个模块时，模型检测精度提高了 3.45 个百分点。

表 2　PASCAL VOC 数据集上的各模块消融实验结果

Table 2　Ablation results of different proposed modules on the PASCAL VOC dataset

Module	ECF	MSH	MRFH	LWS	mAP /%
CenterNet (baseline)					70.64
+ECF	√				72.28
+MSH	√	√			73.16
+MRFH	√	√	√		73.86
MFT	√	√	√	√	74.09

不同分辨率特征图的定位能力、细节信息、抽象程度不同，特征属性与任务属性不匹配时会导致检测精度下降，MSH 模块和 MRFH 模块分别从多尺度、多感受野特征预测目标存在性。为证明这一假设，设计了 MSH、MRFH 模块的对比实验。在 MFT 基础上逐个增加多尺度与多感受野的包围框大小、目标中心点偏移量两个属性的预测。由表 3 可知：当 MSH 模块同时预测目标存在性、包围框大小、中心点偏移量三个属性时，检测精度下降 0.17 个百分点；当 MSH 与 MRFH 都进行目标三个属性的预测时，检测精度下降 0.85 个百分点。实验结果不但证明了特征属性与任务属性不匹配问题的不利影响，还验证了所涉及的 MSH 与 MRFH 能够有效匹配多尺度、多感受野的特征属性与任务属性，从而提高检测精度。

表 3　MSH 和 MRFH 模块的合理性实验

Table 3　Reasonability of MSH module and MRFH module

Module	MSH	MRFH	MSH-mismatch	MRFH-mismatch	mAP /%
All-mismatch	√	√	√	√	73.01
MSH-mismatch	√	√	√		73.18
MFT	√	√			73.86

在 MSH 的基础上，提出 MRFH 模块用于缓解 MSH 中用于预测相同 $\boldsymbol{F}_1$ 与 $\boldsymbol{F}_3$ 特征抽象能力不匹配的问题。为了证明 MRFH 的有效性，也为说明 MRFH 与 MSH 的互补关系，进行对比实验，结果如表 4 所示。由表 4 可知，同时使用 MSH 和 MRFH 模块预测目标存在性时，物体检测精度共提高 1.58 个百分点，检测效果优于单独使用其中任意一个模块的检测效果。

表 4　MRFH 与 MSH 的互补性实验

Table 4　Complementary experiment of MRFH and MSH

MSH	MRFH	mAP /%	ΔmAP
×	×	72.28	0
√	×	73.16	+0.88
×	√	73.44	+1.16
√	√	73.86	+1.58

MSH 模块复用多尺度特征，使用不同分辨率特征进行目标存在性检测，提高检测精度。为验证所复用的 $\boldsymbol{F}_1$，$\boldsymbol{F}_2$，$\boldsymbol{F}_3$ 三个特征的有效性，将多尺度复用特征按分辨率的大小分为大、中、小三种等级的特征，依次验证所复用特征的有效性。表 5 中 Large 对应大分辨率特征（$\boldsymbol{F}_1$），Medium 对应中尺度特征图（$\boldsymbol{F}_2$），Small 对应小尺度特征（$\boldsymbol{F}_3$）。由表 5 可知，MFT 模型中 MSH 模块只复用大分辨率特征时不能覆盖多尺度目标存在性，当复用多尺度特征时，可以获得更多感受野的目标，检测精度提高 0.42 个百分点。

为有效地融合 MSH 和 MRFH 预测的目标存在性结果，提出自学习权重融合模块，对比了简单平均和自学习权重两种融合方式下的检测结果，如表6

表 5 MSH 模块所复用的不同尺度特征的消融结果

Table 5 Ablation results of different scale features reused by MSH module

Module	Large	Medium	Small	mAP/%
MSH-L	√			73.44
MSH-LM	√	√		73.51
MSH	√	√	√	73.86

表 6 不同特征融合方式的实验

Table 6 Experiment of different feature fusion methods

Fusion method	w_1	w_2	w_3	w_4	mAP /%
Simple average	1/4	1/4	1/4	1/4	73.86
Learned weight	0.2894	0.2551	0.3269	0.3307	74.09

所示，自学习权重相比简单平均融合提高 0.23 个百分点。

4 结 论

特征属性与任务属性不匹配问题的存在对现有目标检测器的精度造成了不利影响，因此提出一种匹配多尺度特征与任务的实时目标检测算法，即 MTF 检测器，同时完成了浅层细节特征与精确定位任务的匹配，多尺度多感受野抽象特征与目标存在性预测任务的匹配。通过引入 ECF 模块匹配浅层特征位置信息与目标定位任务；设计 MSH 和 MRFH 时，分别利用多尺度多感受野的特征匹配目标存在性预测任务；最终通过自学习权重的融合模块高效地融合这些结果。MTF 检测器匹配了特征属性与任务属性，提高了检测精度，且计算代价较小，为解决实时目标检测网络中检测精度低的问题提供了一种有效的解决方案。所提 MSH、MRFH、ECF、自学习权重融合模块适用于多数检测网络，与现有的检测算法兼容，通过合理嵌入所提模块到检测网络中，可以实现较小的计算代价，提高检测精度。

参 考 文 献

[1] Yang L, Su J, Huang H, et al. SAR ship detection based on convolutional neural network with deep multiscale feature fusion[J]. Acta Optica Sinica, 2020, 40(2): 0215002.
杨龙, 苏娟, 黄华, 等. 一种基于深层次多尺度特征融合 CNN 的 SAR 图像舰船目标检测算法[J]. 光学学报, 2020, 40(2): 0215002.

[2] 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.

[3] Ji Z, Kong Q K, Wang J, et al. Object detection algorithm guided by dual attention models[J]. Laser & Optoelectronics Progress, 2020, 57(6): 061008.
冀中, 孔乾坤, 王建, 等. 一种双注意力模型引导的目标检测算法[J]. 激光与光电子学进展, 2020, 57(6): 061008.

[4] Zhou B, Li R X, Shang Z H, et al. Object detection algorithm based on improved Faster R-CNN [J]. Laser & Optoelectronics Progress, 2020, 57(10): 101009.
周兵, 李润鑫, 尚振宏, 等. 基于改进的 Faster R-CNN 目标检测算法[J]. 激光与光电子学进展, 2020, 57(10): 101009.

[5] Ju M R, Luo J N, Wang Z B, et al. Multi-scale target detection algorithm based on attention mechanism[J]. Acta Optica Sinica, 2020, 40(13): 1315002.
鞠默然, 罗江宁, 王仲博, 等. 融合注意力机制的多尺度目标检测算法[J]. 光学学报, 2020, 40(13): 1315002.

[6] Yang Q L, Zhou B H, Zheng W, et al. Dim and small target detection based on fully convolutional recursive network[J]. Acta Optica Sinica, 2020, 40(13): 1310002.
杨其利, 周炳红, 郑伟, 等. 基于全卷积递归网络的弱小目标检测方法[J]. 光学学报, 2020, 40(13): 1310002.

[7] 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.

[8] 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, 9905:

21-37.

[9] 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.

[10] Oksuz K, Cam B C, Kalkan S, et al. Imbalance problems in object detection: a review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020: 99.

[11] Kong T, Sun F C, Liu H P, et al. Consistent optimization for single-shot object detection[EB/OL]. (2019-01-19)[2020-09-24]. https://arxiv.org/abs/1901.06563v2.

[12] Cao J L, Pang Y W, Han J G, et al. Hierarchical shot detector[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 9704-9713.

[13] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2999-3007.

[14] Zhou X Y, Wang D Q, Krähenbühl P, et al. Objects as points[EB/OL]. (2019-04-25)[2020-09-24]. https://arxiv.org/abs/1904.07850.

[15] Tian Z, Shen C H, Chen H, et al. FCOS: fully convolutional one-stage object detection[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 9626-9635.

[16] 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.

[17] Yu F, Wang D Q, Shelhamer E, et al. Deep layer aggregation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 2403-2412.

[18] Newell A, Yang K Y, Deng J, et al. Stacked hourglass networks for human pose estimation[M]// Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 483-499.

[19] Law H, Deng J. CornerNet: detecting objects as paired keypoints[J]. International Journal of Computer Vision, 2020, 128: 642-656.

[20] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[21] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[22] Rothe R, Guillaumin M, Gool L, et al. Non-maximum suppression for object detection by passing messages between windows[M]//Cremers D, Reid I, Saito H, et al. Computer vision-ACCV 2014. Lecture notes in computer science. Cham: Springer, 2015, 9903: 290-306.

[23] Everingham M, Eslami S M A, Gool L, et al. The pascal visual object classes challenge: a retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[24] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[25] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[26] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[J]. Journal of Machine Learning Research, 2010, 9: 249-256.

[27] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. (2017-01-30)[2020-09-24]. https://arxiv.org/abs/1412.6980.

[28] Tan M X, Pang R M, Le Q V, et al. EfficientDet: scalable and efficient object detection[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10778-10787.

[29] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-08-08)[2020-09-24]. https://arxiv.org/abs/1804.02767.

[30] Liu Z L, Zheng T, Xu G D, et al. Training-time-friendly network for real-time object detection[EB/OL]. (2019-11-24)[2020-09-24]. https://arxiv.org/abs/1909.00700v2.