第 58 卷 第 2 期/2021 年 1 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

基于注意力和特征融合的遥感图像目标检测模型

汪亚妮，汪西莉*

陕西师范大学计算机科学学院，陕西 西安 710119

摘要 针对环境背景复杂且包含小目标的遥感图像难以进行精准目标检测的问题，在单阶段检测(SSD)模型的基础上，提出了一种基于注意力和特征融合的单阶段目标检测模型，该模型主要由检测分支和注意力分支组成。首先，在检测分支 SSD 中加入注意力分支，注意力分支的全卷积网络通过逐像素回归得到待检测目标的位置特征；其次，采用对应元素相加的方法对检测分支和注意力分支进行特征融合，获得细节信息和语义信息更丰富的高质量特征图；最后，用软非极大值抑制(Soft-NMS)进行后处理，进一步提高目标检测的准确性。实验结果表明，本模型在 UCAS-AOD 和 NWPU VHR-10 数据集上的平均精度均值分别为 92.52%和 82.49%，相比其他模型，检测效率更高。

关键词 图像处理；遥感图像；注意力分支；特征融合；目标检测

中图分类号 O436 **文献标志码** A **doi**：10.3788/LOP202158.0228003

Remote Sensing Image Target Detection Model Based on Attention and Feature Fusion

Wang Yani, Wang Xili*

School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi 710119, *China*

Abstract Aiming at the problem that remote sensing images with complex environmental backgrounds and small targets are difficult to perform accurate target detection, based on the single-stage detection model (SSD), a single-stage target detection model based on attention and feature fusion is proposed in this paper, which is mainly composed of detection branch and attention branch. First, the attention branch is added to the detection branch SSD. The fully convolutional network (FCN) of the attention branch obtains the location characteristics of the target to be detected through pixel-by-pixel regression. Second, by using the method of adding corresponding elements to the detection branch and attention branch, the feature fusion of detection branch and attention branch are carried out to obtain high-quality feature image with more detailed information and semantic information. Finally, soft non-maximum suppression (Soft-NMS) is used as a post-processing part to further improve the accuracy of target detection. Experimental results show that the mean average accuracy of the model on the UCAS-AOD and NWPU VHR-10 data sets are 92.52% and 82.49%, respectively. Compared with other models, the detection efficiency of the model is higher.

Key words image processing; remote sensing images; attention branch; feature fusion; target detection

OCIS codes 100.2000; 100.3008; 100.1830

1 引 言

计算机视觉技术的快速发展使遥感图像的目标检测取得了巨大进步。相比普通光学图像，遥感图像具有尺寸大、拍摄角度特殊、目标尺度变化大、目标尺寸小、目标被云层遮挡等特点。目前，针对多尺度物体的检测，可采用图像金字塔模型将来源相同的原始图像按照尺度的大小进行排列，组成金字塔。

收稿日期：2020-07-06；**修回日期**：2020-07-24；**录用日期**：2020-08-13

基金项目：国家自然科学基金(41471280，61701290，61701289)

* **E-mail**：wangxili@snnu.edu.cn

金字塔底部的尺度最大，顶部的尺度最小，不同尺度的图像可生成不同尺度的特征，但该方法时间成本高。为了节省时间成本，空间金字塔池化网络（SPP Net）[1]、快速区域卷积神经网络（Fast R-CNN）[2]等模型直接用最后一层对待测物体进行分类和预测，但这类模型单层特征图中有限的信息限制了检测结果的提升。为了充分利用特征图中的有效信息，常采用特征金字塔模型提取不同尺度的特征图，以预测目标类别和检测框的位置，如单阶段检测（SSD）[3]模型的浅层特征用于检测小目标，深层特征用于检测大目标。

遥感图像目标检测中小目标在遥感图像中的尺寸较小，且骨干网络的下采样操作会使小目标在特征图中所占的像素更小，导致模型的检测效果较差。现有的小目标检测方法可分为三类，第一类是直接对检测模型中的候选区域、检测框和后处理模块进行优化和修改。如 Sommer 等[4]对区域生成网络（RPN）进行修改，提出适合遥感图像小目标检测的 RPN^2。Long 等[5]用无监督检测框模型对候选区域检测到的边界框进行优化，从而精确定位待测目标，提高小目标的检测精度。陈立里等[6]提出用可变最低阈值方法优化后处理模块，以减少冗余的检测框、降低漏检率。第二类是采用特征融合的方法。在特征金字塔中，浅层特征图包含的语义信息较少，但目标位置准确；深层特征图包含的语义信息较多，但目标位置比较模糊，不利于小目标的检测。特征融合方法将浅层特征与深层特征融合，使特征图同时具备丰富的细节信息和语义信息，包括通道维度拼接和特征图对应元素相加的融合方式。Tian 等[7]采用通道合并融合的方式，将遥感图像中来源不同的区域特征和场景特征进行融合，产生更精确的分类和定位结果。Zhang 等[8]提出了一种新的多尺度特征融合模型，先通过双线性差值对高层特征图进行反卷积操作，再将反卷积后的特征图和上一层特征图进行对应元素相加的融合操作，得到带有丰富细节信息和语义信息的特征图，提高了小目标的检测精度。第三类是基于注意力的方法。注意力机制通过掩码获取新的权重，使深度神经网络学到图像中需要关注的区域，从而标识出图像中关键的特征。可通过加入额外的神经网络[9-10]，硬性选择输入神经网络中的部分信息或自适应地分配给不同信息不同的权重，得到位置注意力或通道注意力。

针对遥感图像中小目标难检测的问题，本文提出了一种基于注意力和特征融合的单阶段检测（AFFSSD）模型。首先，通过位置注意力选择性地关注输入图像中需要被关注的区域，以提取更丰富、准确的位置特征，增强特征图的信息表达能力，过滤复杂的背景特征。在原始 SSD 模型中引入位置注意力以提取目标潜在的位置，与空间变换网络（STN）[11]的位置注意力结构不同，本模型中的位置注意力为全卷积网络（FCN）[12]，可自动学习特征图的空间位置信息，节省空间开销。其次，采用特征融合策略，加强检测模型中浅层特征图的语义信息，用特征融合技术将深层特征图中的位置特征融合到原有检测模型的浅层特征图中，以提升小目标的检测性能。最后，采用软非极大值抑制（Soft-NMS）[13]进行后处理，对遥感图像中相邻目标的检测框进行调整而非彻底抑制，降低了检测时的误删情况。

2 AFFSSD 模型

2.1 模型结构

实验选择经典的 SSD 模型作为基本框架，通过引入位置注意力，提出了改进的 AFFSSD 模型。AFFSSD 模型包含检测分支和注意力分支两部分，具体结构如图 1 所示。其中，两个分支并行排列，分别由大小不同的卷积层组成。上排是由 conv 卷积层构成的检测分支，下排是由 att_conv 构成的注意力分支，检测分支中的虚线框表示注意力分支的卷积层结构。注意力分支通过卷积层的运算得到待测目标的位置特征，然后经过两个反卷积操作分别与检测分支中的 conv8_2 和 conv4_3 进行特征融合。

2.1.1 检测分支

检测分支 SSD 模型是一个端到端的模型，检测速度较快。SSD 模型的特征金字塔结构利用多层特征图，有利于进行多尺度检测，可适应于遥感图像中目标尺度多样性和目标大小分布不规律等问题。在 SSD 模型中，输入遥感图像的大小为 512×512，将大规模视觉几何组 VGG16 网络（深层卷积网络）的卷积层作为主干网络进行特征提取。为了得到检测过程中所需的特征图，将 VGG16 网络中原有的全连接层 fc6、fc7 分别改成大小为 3×3、1×1 的卷积层，之后用大小为 1×1×128 的卷积增加 conv8_1、conv8_2、conv9_1、conv9_2、conv10_1、conv10_2、conv11_1、conv11_2、conv12_1、conv12_2 卷积层，最后用平均池化层在 conv12_2 层得到大小为 1×1 的输出。整个过程中 conv4_3、fc7、conv8_2、conv9_2、conv10_2、conv11_2、conv12_2 层的特征图都进

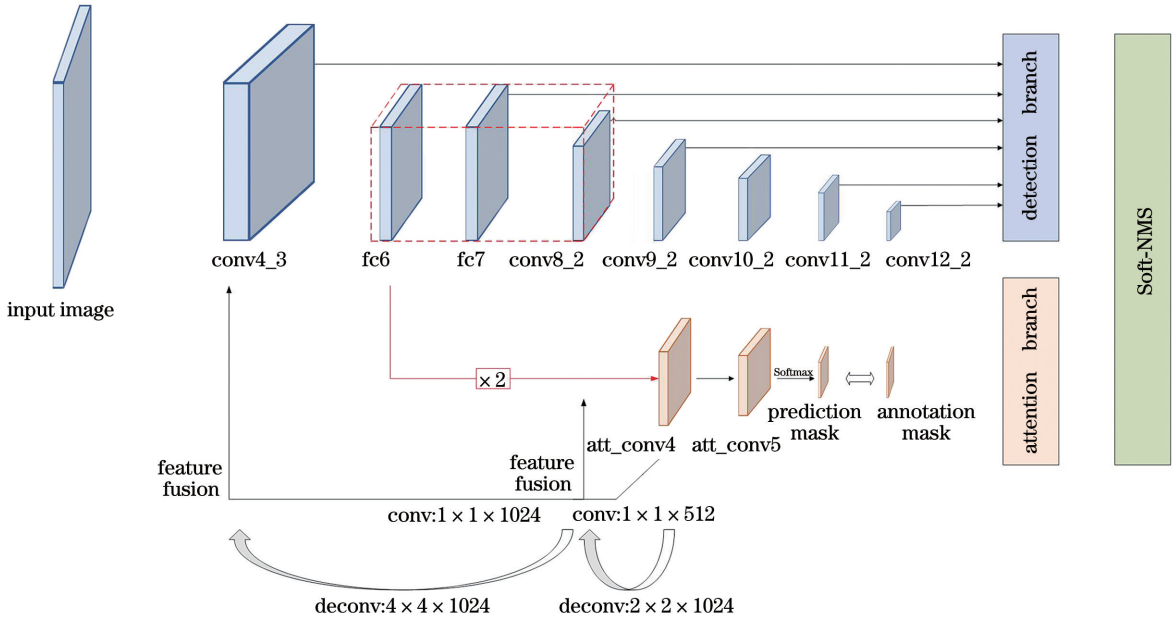

图 1　AFFSSD 模型的结构

Fig. 1　Structure of the AFFSSD model

行类别和位置的回归操作，大小分别为 64×64、32×32、16×16、8×8、4×4、2×2、1×1。

2.1.2　注意力分支

为了获得待测目标的位置信息，在原始 SSD 模型的基础上增加注意力分支。注意力分支通过添加权重的方式标识出遥感图像中存在待检测目标概率大的位置，使模型在学习过程中学到每张图像需要关注的区域。注意力分支采用 FCN 模型，将 CNN 最后一层全连接层替换成卷积层，通过上采样操作，输出逐像素分类结果图。FCN 模型的优点：1）输入和输出的图像大小不变，在进行检测任务时，不用单独对图像的大小进行调整，节省时间的同时又能保留检测图像的真实信息；2）最后输出的是每个像素的类别信息，反映了存在待检测目标概率大的位置。注意力分支将与图 1 中虚线方块相同的结构加在检测分支 conv5_4 层之后，具体参数如表 1 所示。FCN 模块的注意力分支中，采用 Softmax 函数预测待测目标位置的注意力权重。

表 1　AFFSSD 模型的具体参数

Table 1　Specific parameter of the AFFSSD model

Branch	Layer name	Output size	Operation of convolution
Detection branch	conv4_3	64×64×512	[3×3×1024]
	fc6	32×32×1024	[1×1×1024]
	fc7	32×32×1024	$\begin{bmatrix}1\times1\times256\\3\times3\times512\end{bmatrix}$
	conv8_2	16×16×512	$\begin{bmatrix}1\times1\times128\\3\times3\times256\end{bmatrix}$
	conv9_2	8×8×256	$\begin{bmatrix}1\times1\times128\\3\times3\times256\end{bmatrix}$
	conv10_2	4×4×256	$\begin{bmatrix}1\times1\times128\\3\times3\times256\end{bmatrix}$
	conv11_2	2×2×256	$\begin{bmatrix}1\times1\times128\\3\times3\times256\\3\times3\times256\end{bmatrix}$
	conv12_2	1×1×256	–
Attention branch	att_conv4	8×8×256	$\begin{bmatrix}1\times1\times128\\3\times3\times2\end{bmatrix}$
	att_conv5	8×8×2	–

2.2　特征融合

卷积神经网络（CNN）中不同层次的特征图带有不同的信息，深层特征图中语义信息丰富，能很好地表示出目标的位置信息；浅层特征图包含丰富的细节信息，如纹理、边缘信息。为了充分利用浅层特征图的细节信息，加强对小尺度目标的检测，采用特征融合方法，将注意力分支中深层特征图通过反卷积融合操作，以缩小检测分支中浅层特征图之间的语义差异。常见的融合方式有通道维度拼接、特征图空间位置相加、相乘等。实验将不同特征图对应的空间位置进行累加，可在不增加特征图维度的同时增加特征图的信息量，复杂度较低。

特征融合的具体过程：首先，将注意力分支中大

小为 8×8×256 的特征图经过 2×2 的反卷积操作，得到和检测分支 conv8_2 层大小相同的特征图；其次，将大小为 16×16×512 的特征图经过 4×4 的反卷积操作，得到和检测分支 conv4_3 层大小相同的特征图；最后，将两次反卷积操作得到的特征图分别和原检测分支 conv8_2、conv4_3 层的特征图进行特征融合。

2.3 Soft-NMS 后处理

目标检测中后处理模块是必不可少的一部分，常见的后处理是非极大值抑制(NMS)[14]。NMS 选择置信度最高的检测框并删除超过某一阈值的相邻检测框，目的是删除检测结果中重复的检测框，得到更精确的检测结果。由于遥感图像中的目标分布密集，目标排列方向无规律可循，在检测过程中 NMS 往往会将相邻目标的检测框删除，导致漏检率增加。因此，用 Soft-NMS 作为 AFFSSD 模型的后处理模块，通过降低检测框的置信度，提高检测结果的召回率，可表示为

$$s=\begin{cases}s_i, & X_{\mathrm{IOU}}(m,b_i)<t\\ s_i\left[1-X_{\mathrm{IOU}}(m,b_i)\right], & X_{\mathrm{IOU}}(m,b_i)\geqslant t\end{cases},\tag{1}$$

式中，b_i 为待处理框，i 为检测框的序号，m 为当前得分最高的框，$X_{\mathrm{IOU}}(m,b_i)$ 为 m 和 b_i 的交并比，即产生的边界框和原始边界框的重叠率，s_i 为 b_i 的检测得分，$X_{\mathrm{IOU}}(m,b_i)$ 的值越大，s_i 的得分越低，t 为阈值。可以发现，原始得分 s 会被新得分 s_i 代替，最终保存在得分集合中。

2.4 损失函数

AFFSSD 模型的损失函数由检测分支和注意力分支的损失函数组成，可表示为

$$L_{\mathrm{total}}=L_{\mathrm{detection}}+L_{\mathrm{attention}}=\frac{1}{N}(\alpha L_{\mathrm{loc}}+L_{\mathrm{conf}})+\beta\sum_i\|M-Z_i\|^2,\tag{2}$$

式中，N 为匹配的候选框数量，$L_{\mathrm{detection}}$ 为检测分支的损失函数，包括位置损失 L_{loc} 和置信度损失 L_{conf}，α 为平衡位置损失和置信度损失的超参数，$L_{\mathrm{attention}}$ 为注意力分支的损失函数，即位置掩码的损失，M 为注意力分支中的预测权重，$Z_i\in\{0,1\}$ 为标识权重，β 为平衡检测分支和注意力分支的超参数。位置损失可表示为

$$L_{\mathrm{loc}}=\sum_{i\in X_{\mathrm{pos}}}^{N}\sum_{m\in\{c_x,c_y,w,h\}}x_{ij}^{k}S_{\mathrm{L1}}(l_i^u-\hat{g}_j^u),\tag{3}$$

式中，X_{pos} 为正样本，u 为类别，l 为预测框，g 为 ground-truth，S_{L1} 为 Smooth $\mathrm{L_1}$ 损失函数，x_{ij}^{k} 为第 i 个候选框和类别 k 的真实框 j 相匹配，c_x，c_y，w，h 分别为候选框的中心坐标、宽度和高度。置信度损失可表示为

$$L_{\mathrm{conf}}=-\sum_{i\in X_{\mathrm{pos}}}y_i\log L-\sum_{i\in X_{\mathrm{neg}}}\log L^0,\tag{4}$$

式中，X_{neg} 为负样本，$y_i\in\{0,1\}$ 为第 i 个候选框的 ground-truth，L 为第 i 个候选框的置信度，L^0 为背景的置信度。

2.5 AFFSSD 模型的计算步骤

AFFSSD 模型对遥感图像中小目标的检测步骤如下。

1) 输入待检测的遥感图像，将遥感图像的大小统一缩放为 512×512。

2) 为了提取遥感图像中的目标特征，将步骤 1) 中大小为 512×512 的待测遥感图像输入到训练好的特征提取网络 VGG16，得到特征图。

3) 从检测分支中抽取 conv4_3、fc7、conv8_2、conv9_2、conv10_2、conv11_2、conv12_2 层的特征图；然后在这些特征图层的每个点上构造 9 个不同大小的检测框，分别用(3)式和(4)式对待测目标进行定位和分类，生成多个符合条件的初步检测框，同时注意力分支通过 Softmax 层得到上述 7 层特征图的位置注意力权重。

4) 为了增强检测分支中浅层特征图的细节信息，将步骤 3)中注意力分支得到的位置注意力权重经过两次反卷积操作后分别与检测分支中 conv4_3 和 conv8_2 层的特征图进行特征融合；同时将检测分支中原有的 conv4_3、conv8_2 特征图分别替换为反卷积操作后的特征图。

5) 将步骤 4)中特征图的检测框与步骤 3)中剩余特征图获得的检测框相结合，然后利用(1)式抑制掉冗余或不正确的检测框，生成精确的检测框，进而输出带有精确检测框的待测遥感图像。

3 实验结果及分析

3.1 数据集

1) UCAS-AOD 数据集

UCAS-AOD 数据集[15]包括飞机和汽车两类目标，有 1000 幅彩色飞机图像和 510 幅彩色车辆图像，共标注了 7482 个飞机样本和 7114 个车辆样本。该数据集中的类间差异小、类内差异大、目标方向分布均匀，在遥感图像的目标检测中具有很大的挑战。表 2 为 UCAS-AOD 数据集中两类目标大小的统计

结果，可以发现，两类待测目标中像素低于 200 的小目标在车辆类别中更多，占总车辆样本的 79%。

表 2　UCAS-AOD 数据集中的车辆尺寸

Table 2　Vehicle dimension in the UCAS-AOD data set

Scale /pixel	Scale1 (<100)	Scale2 (100-200)	Scale3 (200-300)	Scale4 (>300)	Total
Number(vehicle)	3704	1986	967	457	7114
Number(aircraft)	1028	1438	2593	2369	7482

2) NWPU VHR-10 数据集

NWPU VHR-10 数据集[16]来源于 Google Earth 和 Vaihingen，共有 800 幅高分辨率遥感图像，其空间分辨率为 0.5～2 m。由 715 幅 RGB (Red、Green、Blue)图像和 85 幅锐化彩色红外图像组成，其中，715 幅 RGB 图像采集自 Google Earth，空间分辨率为 0.5～2 m；85 幅全色锐化的红外图像来自 Vaihingen 数据，空间分辨率为 0.08 m。该数据集含有车辆、飞机、轮船、港口、桥梁、篮球场、网球场、棒球场、田径场和储油罐 10 个类别，共标注 3775 个待测目标，包括 477 辆汽车、757 架飞机、302 艘船只、224 个港口、124 座桥梁、159 个篮球场、524 个网球场、390 个棒球场、163 个田径场和 655 个储油罐[16]。该数据集中的图像种类多、数据量小，在遥感图像检测中有很大的难度。

3.2　评价指标

实验使用的评价指标为平均精度(AP)、平均精度均值(mAP)和每幅图像的传输时间(S)。AP 可以衡量检测任务中待测数据集每个类别的检测结果；mAP 为各类别 AP 的平均值，可以衡量检测任务中待测数据集中所有类别的检测结果；每幅图像的传输时间可以衡量检测的速度。

准确率 P 和召回率 R 可表示为

$$P=\frac{X_{\mathrm{TP}}}{X_{\mathrm{TP}}+X_{\mathrm{FP}}},\tag{5}$$

$$R=\frac{X_{\mathrm{TP}}}{X_{\mathrm{TP}}+X_{\mathrm{FN}}},\tag{6}$$

式中，X_{TP} 为真正例，X_{FP} 为假正例，X_{FN} 为假反例。AP、mAP 可表示为

$$X_{\mathrm{AP}}=\int_0^1 P(R)\mathrm{d}R,\tag{7}$$

$$X_{\mathrm{mAP}}=\frac{1}{n}\sum_{i=1}^{n}\int P(R)\mathrm{d}R,\tag{8}$$

式中，n 为待检测数据集中所有类别的数目。

3.3　实验结果及分析

实验采用 Pytorch 深度学习框架，硬件配置为 Intel(R) Xeon(R) CPU E5-2690 v3 2.6 GHz 处理器，用 12 GB 的 NVIDIA Tesla K 40c GPU 进行加速，模型参数针对实际实验情况并参考文献[3]进行设置。

针对 AFFSSD 模型的特点，输入图像的大小为 512×512，随机将两个数据集中每个类别按照 7∶3 的比例划分为训练集和测试集。训练时，用加速器进行加速，采用事先训练好的 VGG16 网络进行预训练初始化模型，Batch size 为 10，学习率为 0.0001，动量为 0.9，权值衰退率为 0.00005，gamma 为 1，迭代次数为 100000 次，损失函数中 $\alpha=1$，$\beta=0.35$。其他对比实验的参数与训练过程中的参数相同。

3.3.1　UCAS-AOD 数据集

图 2 为 UCAS-AOD 部分测试集在 SSD 模型和 AFFSSD 模型中的飞机检测结果，可以发现，SSD 模型和 AFFSSD 模型都能检测不同尺度的待

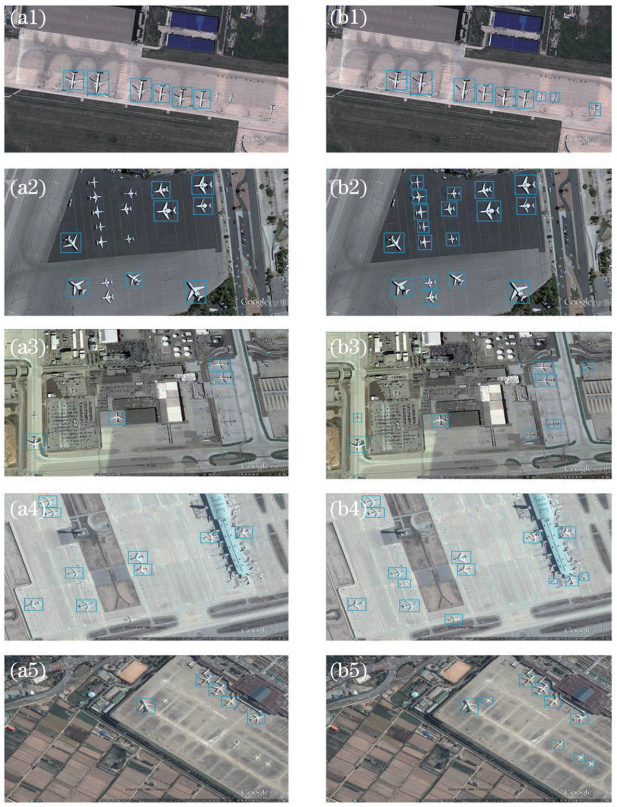

图 2　不同模型的飞机检测结果。(a) SSD 模型；(b) AFFSSD 模型

Fig. 2　Test results of different models of aircraft. (a) SSD model; (b) AFFSSD model

测飞机，但 SSD 模型对飞机样本的检测结果中漏检情况较多，尤其是尺寸较小的飞机。而 AFFSSD 模型解决了 SSD 模型中的漏检问题，原因是 AFFSSD 模型有更精细的特征图，对小目标的检测优势更大。

为了验证 AFFSSD 模型的检测性能，列出了 AFFSSD 模型和其他对比方法在 UCAS-AOD 数据集中的检测结果，如表 3 所示。可以发现，AFFSSD 模型的 mAP 最高，为 92.52%，每幅图像的传输时间为 0.26 s，检测结果和检测速度均达到了最优。这表明 AFFSSD 模型在一定程度上能提高遥感图像的检测性能。

表 3　不同方法在 UCAS-AOD 数据集中的检测结果

Table 3　Detection results of different methods in UCAS-AOD data set

Method	AP of plane /%	AP of small-vehicle /%	mAP /%	S /s
SSD	88.13	85.09	86.61	0.36
Ref.[17]	90.66	88.17	89.41	0.34
AFFSSD	**93.70**	**91.34**	**92.52**	**0.26**

文献[17]中的方法用 Darknet-19 作为主干网络，同时引入转移层得到更精细的特征图，采用的融合操作在通道维度上对特征图进行拼接。相比文献[17]中的方法，AFFSSD 模型的 mAP 提高了 3.11 个百分比，每幅图像的传输时间降低了 0.08 s，这表明 AFFSSD 模型的融合方式更有优势。

图 3 为不同模型在 UCAS-AOD 数据集中的车辆检测结果，可以发现，对于小尺度车辆的检测，两种模型的检测结果有明显的差异；且 AFFSSD 模型没有漏检车辆，检测效果优于 SSD 模型。

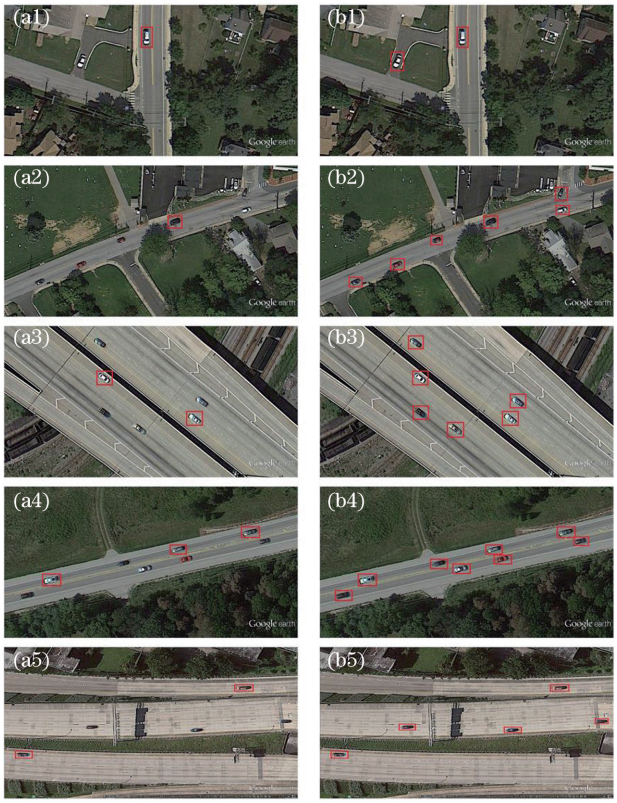

图 3　不同模型的车辆检测结果。(a) SSD 模型；(b) AFFSSD 模型

Fig. 3　Test results of different models of vehicle. (a) SSD model; (b) AFFSSD model

表 4 为 SSD 模型和 AFFSSD 模型对不同尺度目标的检测结果，其中 scale1～scale4 对应表 1 中不同的尺度。可以发现，两种模型都能检测出不同尺度的目标，也能很好地检测大尺度目标。相比 SSD 模型，在尺度为 200～300 的 scale3 和尺度大于 300 的 scale4(大尺度)检测中，AFFSSD 模型的 mAP 分别提升了 3.63 和 1.62 个百分点。在尺度小于 100 的 scale1 和尺度为 100～200 的 scale2(小尺度)检测中，AFFSSD 模型的 mAP 分别提升了 4.96 和 6.22 个百分比。原因是 SSD 模型进行小目标检测的浅层特征图表征不强，而 AFFSSD 模型中引入的注意力分支将深层的位置信息融合到低层的特征图中，在获得待测目标位置信息的同时提升了小目标的检测效果；同时 AFFSSD 模型处理图像的速度更快，有效缩短了测试时间，这表明 AFFSSD 模型对小目标的检测是有效的。

表 4　两种模型对不同尺度目标的检测结果

Table 4　Detection results of the two models on different scale targets

Method	SSD	AFFSSD
AP of scale1/%	57.11	**62.07**
AP of scale2/%	64.09	**70.31**
AP of scale3/%	69.02	**72.65**
AP of scale4/%	69.71	**71.33**
S /s	38.90	**38.10**

3.3.2　NWPU VHR-10 数据集

图 4 为两种模型在 NWPU VHR-10 数据集上的飞机、车辆、网球场、油罐、篮球场和船检测结果。

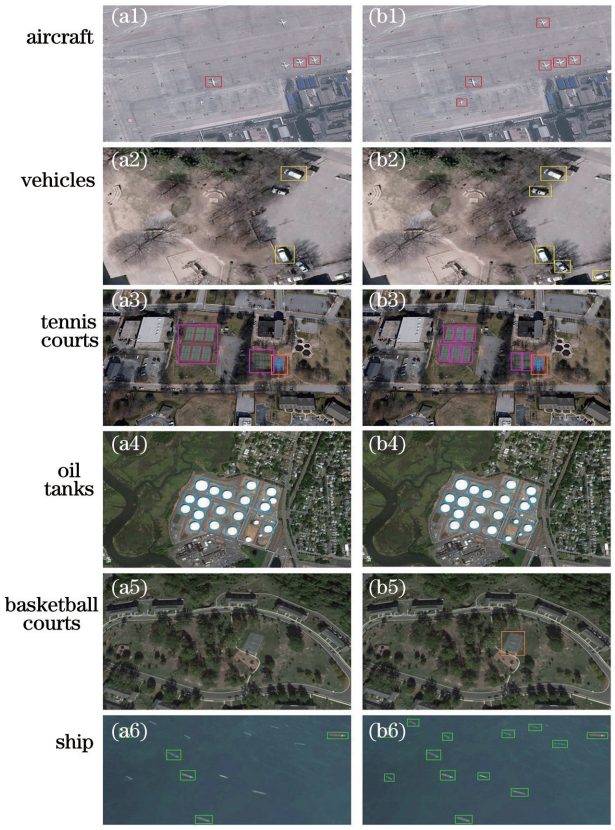

图 4　两种模型的部分检测结果。(a) SSD 模型；(b) AFFSSD 模型

Fig. 4　Partial test results of the two models. (a) SSD model; (b) AFFSSD model

可以发现，对于飞机和车辆这种待测目标分布无规律且像素占比少的类别，SSD 模型的检测效果欠佳，飞机和船只均存在漏检情况，如图 4(a1)、图 4(a6)所示；但这种待测目标在 AFFSSD 模型中的表现较好，没有漏检情况，如图 4(b1)、图 4(b6)所示，这表明 AFFSSD 模型对小目标的检测是有效的。对于车辆和篮球场这种待检测目标和周围环境背景相近，不易区分的类别，SSD 模型的检测结果不理想，有漏检情况，如图 4(a2)、图 4(a5)所示，但 AFFSSD 模型能检测出被树枝遮挡的车辆，如图 4(b2)所示，也能检测出和周围环境相差不大的篮球场，如图 4(b5)所示，这表明环境背景复杂的图像对 AFFSSD 模型的影响不大。对于飞机和车辆这种待测目标分布密集的类型，SSD 模型和 AFFSSD 模型都能检测出网球场，但检测结果的表现形式不同。SSD 模型的检测结果粗糙，不理想，如图 4(a3)所示；而 AFFSSD 模型可以检测出每个不同的网球场，如图 4(b3)所示；对于分布密集的油罐，AFFSSD 模型比 SSD 模型的检测效果更好，如图 4(a4)、图 4(b4)所示，这表明 AFFSSD 模型能检测出分布密集的目标。

表 5 为不同模型对 NWPU VHR-10 数据集中 10 个类别的检测结果，其中，RICNN[18] 为基于 CNN 的旋转不变模型，在训练过程中通过正则化约束优化了一个新的目标函数，以实现遥感图像目标检测的旋转不变性。SSD 模型[3] 利用回归思想，采用端到端的模型，利用不同尺度的特征图，进行多尺度的检测，是 one stage 模型。反卷积 SSD(DSSD)模型[19] 将 VGG 替换为残差网络(ResNet101)，并在分类回归前引入残差模块，将深层的语义信息融合到浅层网络的特征中，以此提高对小目标的检测效果。AFFSSD 模型和 DSSD 模型均采用反卷积的方法进行特征融合，不同之处在于 AFFSSD 模型中引入了位置注意力模块；李红艳等[20] 在特征提取

表 5　不同模型在 NWPU VHR-10 数据集中的检测结果

Table 5　Detection results of different models in the NWPU VHR-10 data set　　unit: %

Model	RICNN[18]	SSD	DSSD[19]	Ref. [20]	Deformable R-FCN[21]	Faster R-CNN[22]	AFFSSD
Aircraft	88.35	84.32	86.50	**95.20**	87.30	94.60	87.02
Ship	77.34	62.90	65.40	79.70	81.40	82.30	**83.50**
Oil tank	85.27	78.25	**90.30**	73.70	63.60	65.32	80.69
Baseball diamond	88.12	89.33	89.60	96.40	90.40	95.50	**96.02**
Tennis court	40.83	79.41	**85.10**	71.60	81.60	81.90	80.32
Basketball court	58.45	87.69	80.40	72.10	74.10	89.70	**90.10**
Ground track field	86.73	80.61	78.20	**99.70**	90.30	92.40	81.36
Harbor	68.60	71.37	70.50	73.20	75.30	72.40	**75.80**
Bridge	61.51	65.35	68.20	57.00	71.40	57.50	**72.03**
Vehicle	71.10	62.30	74.20	72.00	75.50	77.80	**78.01**
mAP	72.63	76.15	78.84	79.06	79.09	80.94	**82.49**

时加入注意力模块(CBAM),将空间注意力和通道注意力按照并行顺序的方式组合在一起,得到空间和通道两个维度的注意力权重。Ren 等[21]为了解决传统 CNN 模型在遥感目标几何变化方面的局限性,用一种端到端的可变形卷积神经网络模型 Deformable R-FCN 在卷积层中添加偏移量,以学习遥感目标的几何信息,增强对形状大小不一遥感目标的检测效果。同时 Deformable R-FCN 模型也提出了新的后处理方法 arcNMS(aspect ratio constrained NMS)。不同于本方法中的 Soft-NMS,arcNMS 可用于删除可变形卷积中假区域的检测框。Ren 等[22]在 Faster R-CNN 中设计的区域生成网络(PRN)直接提取候选区域,使卷积层的特征可被整个模型共享,大幅度地提升了检测速度。可以发现,AFFSSD 模型的检测结果比其他模型的检测结果更好,其 mAP 为 82.49%,这表明一定程度地融合策略和加入注意力模块可以提升对遥感图像目标的检测效果。

由表 5 可知,RICNN 的检测结果最差,原因是其他模型都是基于深度学习的检测算法,检测性能提升明显,这表明深度学习的端到端模型对于遥感图像的目标检测更有优势。DSSD 模型和 AFFSSD 模型都采用特征融合的策略,但 AFFSSD 模型的总体检测结果更好,这表明特征融合策略对于提升遥感图像的检测效果是有意义的。文献[21]的检测模型和 AFFSSD 模型采用不同的注意力模块,前者主要关注空间和通道的注意力,后者主要关注待测目标位置的注意力,而 AFFSSD 模型的检测结果更好,这表明本模型中的位置注意力对 NWPU VHR-10 遥感数据集的检测更有效。Deformable R-FCN 模型和 AFFSSD 模型都考虑用后处理方法进一步过滤冗余的检测框,但 Deformable R-FCN 模型更关注形变目标的检测框,两者检测结果相差 3.4 个百分点,AFFSSD 模型的检测结果更好,这表明对于目标种类多且目标中几何形状类别不多的 NWPU VHR-10 数据集来说,本模型中的后处理方法对检测结果的提升更明显。SSD、Faster R-CNN 和 AFFSSD 模型的检测结果差距明显,且 AFFSSD 模型的检测结果最优,这表明本模型中的三个改进对遥感图像的目标检测更有优势。

DSSD 模型对油罐和网球场这两类密集目标的检测 AP 均是最优的,分别为 90.30% 和 85.10%。而 AFFSSD 模型在这类密集目标的检测中结果表现不佳的原因:1)待测目标在定位过程中有很多参数的变化,不同应用场景中最优值的表现可能不同;2)检测过程中两个模型的检测框大小不一,为了包含各个角度且大小不同的待测目标,检测框的尺度通常较大,但在密集区域一个检测框可能包含多个目标。文献[21]中的检测模型对飞机和田径场这两类的检测效果最好,AFFSSD 模型对这两类目标检测表现不佳的原因是待检测的飞机和操场外形具有一定的几何规则且环境背景复杂,图像中干扰信息较多,而本模型仅关注位置注意力的检测模型。

4 结 论

提出了一种基于注意力和特征融合的单阶段遥感图像目标检测模型,本模型首先在检测分支的基础上加入注意力分支,通过注意力分支引导检测分支关注潜在的目标信息,抑制其他无用信息,为检测特征图提供待检测目标的位置信息;然后将两个分支的特征进行融合,得到语义信息更丰富的大尺度特征图,进一步提升了小目标的检测效果。实验结果表明,本模型在 UCAS-AOD 和 NWPU VHR-10 数据集上表现优异,与其他模型相比,检测结果均有明显的提升。接下来的工作中,还需对遥感图像的密集目标检测进行深入研究,进一步提升本模型的检测效果。

参 考 文 献

[1] 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 & Machine Intelligence, 2014, 37(9): 1904-1916.

[2] Girshick R. Fast R-CNN[C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1440-1448.

[3] 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.

[4] Sommer L W, Schuchert T, Beyerer J. Fast deep vehicle detection in aerial images[C] // 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), March 24-31, 2017, Santa Rosa, CA, USA. New York: IEEE, 2017: 311-319.

[5] Long Y, Gong Y P, Xiao Z F, et al. Accurate object localization in remote sensing images based on convolutional neural networks[J]. IEEE Transactions

on Geoscience and Remote Sensing, 2017, 55(5): 2486-2498.

[6] Chen L L, Zhang Z D, Peng L. Real-time detection based on improved single shot MultiBox detector[J]. Laser & Optoelectronics Progress, 2019, 56(1): 011002.
陈立里, 张正道, 彭力. 基于改进 SSD 的实时检测方法[J]. 激光与光电子学进展, 2019, 56(1): 011002.

[7] Tian Z Z, Wang W, Zhan R H, et al. Cascaded detection framework based on a novel backbone network and feature fusion[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(9): 3480-3491.

[8] Zhang W H, Jiao L C, Liu X, et al. Multi-scale feature fusion network for object detection in VHR optical remote sensing images[C]//IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, July 28-August 2, 2019, Yokohama, Japan. New York: IEEE, 2019: 330-333.

[9] Ji Z, Kong Q K, Wang J. Object detection algorithm guided by dual attention models[J]. Laser & Optoelectronics Progress, 2020, 57(6): 061008.
冀中, 孔乾坤, 王建. 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.

[10] Zhang M, Wang S C, Yang D F. 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. 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.

[11] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[EB/OL]. [2020-06-25]. https://arxiv.org/abs/1506.02025.

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

[13] Bodla N, Singh B, Chellappa R, et al. Soft-NMS: improving object detection with one line of code[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 5562-5570.

[14] Neubeck A, van Gool L. Efficient non-maximum suppression[C]//18th International Conference on Pattern Recognition (ICPR′06), August 20-24, 2006, Hong Kong, China. New York: IEEE, 2006: 850-855.

[15] Zhu H G, Chen X G, Dai W Q, et al. Orientation robust object detection in aerial images using deep convolutional neural network[C]// 2015 IEEE International Conference on Image Processing (ICIP), September 27-30, 2015, Quebec City, QC, Canada. New York: IEEE, 2015: 3735-3739.

[16] Cheng G, Han J W, Zhou P C, et al. Multi-class geospatial object detection and geographic image classification based on collection of part detectors[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2014, 98: 119-132.

[17] Xia G S, Bai X, Ding J, et al. DOTA: alarge-scale dataset for object detection in aerial images[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 3974-3983.

[18] Cheng G, Zhou P C, Han J W. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(12): 7405-7415.

[19] Fu C Y, Liu W, Ranga A, et al. DSSD: deconvolutional single shot detector[EB/OL]. [2020-06-25]. https://arxiv.org/abs/1701.06659.

[20] Li H Y, Li C G, An J B, et al. Attention mechanism improves CNN remote sensing image object detection [J]. Journal of Image and Graphics, 2019, 24(8): 1400-1408.
李红艳, 李春庚, 安居白, 等. 注意力机制改进卷积神经网络的遥感图像目标检测[J]. 中国图象图形学报, 2019, 24(8): 1400-1408.

[21] Ren Y, Zhu C R, Xiao S P. Deformable faster R-CNN with aggregating multi-layer features for partially occluded object detection in optical remote sensing images[J]. Remote Sensing, 2018, 10(9): 1470.

[22] 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.