Computer Science and Application 计算机科学与应用, 2021, 11(10), 2538-2545
Published Online October 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.1110257

Hans汉斯

一种用于遥感图像目标检测的特征融合检测模型

刘晓东，王卓薇，徐　超，张　鳌，陈海源

广东工业大学计算机学院，广东 广州

收稿日期：2021年9月20日；录用日期：2021年10月17日；发布日期：2021年10月25日

摘　要

遥感图像和自然图像的差异导致目标检测在遥感图像中效果不佳，因此本文将注意力机制运用到特征提取的过程中，提高特征提取的能力。并使用自注意力机制对所提取各层级特征信息进行融合，基于集成特征做后续的目标检测，提高在目标多尺度问题上的表现。本文在DIOR数据集验证网络模型的可行性，并与当前常见的目标检测模型进行对比验证，取得了67.1%的全类平均精度。实验表明，该模型对比于其他常见目标检测模型在遥感图像上的检测性能有显著地提升。

关键词

遥感图像目标检测，注意力机制，特征融合，深度学习

A Feature Fusion Detection Model for Object Detection in Remote Sensing Images

Xiaodong Liu, Zhuowei Wang, Chao Xu, Ao Zhang, Haiyuan Chen

School of Computer Science, Guangdong University of Technology, Guangzhou Guangdong

Received: Sep. 20th, 2021; accepted: Oct. 17th, 2021; published: Oct. 25th, 2021

Abstract

The difference between remote sensing image and natural image leads to poor target detection effect in remote sensing image. Therefore, this paper applies the attention mechanism to the feature extraction process to improve the feature extraction ability. And make the self-attention mechanism to fuse the extracted feature information of each level, and do the follow-up target detection based on the integrated features to improve the performance on the target multi-scale prob-

文章引用: 刘晓东, 王卓薇, 徐超, 张鳌, 陈海源. 一种用于遥感图像目标检测的特征融合检测模型[J]. 计算机科学与应用, 2021, 11(10): 2538-2545. DOI: 10.12677/csa.2021.1110257

lem. This paper verifies the feasibility of the network model on the DIOR data set, and compares and verifies it with the current common target detection model, and achieves an average accuracy of 67.1% for all classes. Experiments show that compared with other common target detection models, the detection performance of this model on remote sensing images has been significantly improved.

Keywords

Remote Sensing Image Object Detection, Attention Mechanism, Feature Fusion, Deep Learning

1. 引言

遥感作为对地观测的重要工具，已广泛应用于生态环境监测、天气预报等军事和民用领域。近年来，随着遥感技术的不断发展，遥感图像数量急剧增加，传统的人工解译已经成为一项艰巨的任务。目标检测是指识别感兴趣目标的对应类别并标记其位置，对于遥感图像的图像理解有着重要的作用。

近年来，许多基于卷积神经网络的目标检测模型取得了显著的效果[1]。基于深度学习的目标检测算法大致可分为两阶段检测法[2]和一阶段检测法[3] [4]。两阶段目标检测算法包括两个阶段：方案的生成阶段和方案的回归分类阶段。在第一阶段，基于骨干网络提取的特征地图，通过区域提案网络(RPN)为每幅图像生成一系列提案框。在第二阶段，根据方案对特征图进行裁剪，然后调整到相同的大小，并进一步利用裁剪后的特征图对对象类别进行预测，并对方案的位置进行微调，使预测更加准确。在一阶段的目标检测算法中，目标检测被看作是一个回归问题。利用多个比例和长宽比在图像的不同位置设置密集采样锚盒，然后利用卷积神经网络进行特征提取，直接对目标进行分类和预测。整个过程只需要一步，检测速度比两步法快。然而，一阶段目标检测算法有一个明显的缺陷。密集采样锚盒会使目标探测器的正负样本极不平衡，且负样本的数量将远远大于阳性样本的数量，导致后续的预测效果不佳。

但目标检测在航空图像中的进展却比较缓慢[5]。这主要是因为遥感图像与自然图像之间存在着巨大的差异。遥感图像通常由卫星或者飞行器在高空俯视拍摄，包含较大的场景，但目标物体体积占比小且尺寸变化大、背景复杂。自然图像与遥感图像之间的差异导致了目标检测算法在遥感图像中表现不佳。

针对于遥感图像的特点，本文提出了一种适用于遥感图像的目标检测模型。本文的主要贡献分为两个部分：

1) 将注意力机制运用在特征提取的过程中，解决遥感图像目标体积小、背景复杂的问题。

2) 利用自注意力机制对特征提取网络各层级特征图进行融合，提高模型对于目标多尺度问题的表现。

本文结构如下：第二章介绍特征提取与目标多尺度问题的相关工作，第三章介绍提出的方法，第四章展示实验，第五章总结本文。

2. 相关工作

VGG [6]网络是一种常用的目标特征提取网络。该网络使用了更小的卷积核(都使用了 3 × 3 的核)，并加深了模型结构，从而获得了更好的特征提取效果。在一定程度上深化模型结构有助于提高网络性能，但是随着网络层数的增加，网络将变得难以训练，并且可能发生梯度退化，从而错过最佳收敛并导致更

大的学习误差。针对梯度退化问题，后期开发的特征提取网络 ResNet [7]引入了跨层连接，可以解决增加层数带来的退化问题。与之前建立的许多其他网络相比，ResNet 网络的特征提取效率大大提高；尽管如此，ResNet 在复杂背景的遥感照片中目标检测的性能有限。目前，许多网络设计者已经开始引入注意力机制[8] [9]来解决这个问题。在特征提取结构中部署了注意力机制来调整特征的权重。这使得特征提取网络能够专注于目标物体区域而不是背景噪声，选择性地增强目标物体的特征信息，实现目标特征信息的高效提取并抑制背景噪声的干扰。

多尺度问题是目标检测任务中的一个重要问题。在经典的物体检测算法 Fast R-CNN 和 Faster R-CNN 中，只有最后一层特征图用于预测。然而，这种方法对于小物体来说是一个巨大的挑战，会导致大量的小物体被遗漏。对于这个问题，一个比较原始的解决方案是图像金字塔。输入图像依次缩放，然后训练检测器。这种方法对多尺度问题有很好的效果，但代价是带来了巨大的计算开销。在 SSD 中，同时选取不同层级的特征图进行预测，汇总不同层级的预测结果。但是这种做法存在一定问题，深层特征图语义信息丰富，感受野较大，但空间信息相对缺乏，对小目标物体的预测会造成大量漏检。浅层特征图空间信息丰富，分辨率高，但语义信息不足会导致大量误检测。Lin 提出了特征金字塔(FPN) [10]，构建了一条自顶向下的道路，将多个层次特征图的语义信息和空间信息融合，得到多个融合的特征图。特征金字塔对于目标检测具有重要意义，大大提高了在多尺度问题上的表现。这表明浅层次的信息和深层次的信息是互补的，对于后续的目标检测效果来说非常重要。但 FPN 的特征融合方式存在些问题[11]，他更多的是关注相邻层级的特征信息，非相邻层级的特征信息在每次融合中都会被稀释一次。

3. 本文方法

本文在 Faster RCNN 的基础上进行改进。将注意力机制应用在特征提取的过程中，使得特征提取网络能够专注于目标物体区域而不是背景噪声，选择性地增强目标物体的特征信息。以此来解决遥感图像中，目标物体体积小、背景复杂和噪音干扰严重的问题。使用非局部信息统计的注意力机制[12]对特征提取网络的各层级信息进行融合，使得最终的集成特征是来自于每个层级的特征信息而不是只关注与相邻层级的特征信息，整体网络结构如图 1 所示。

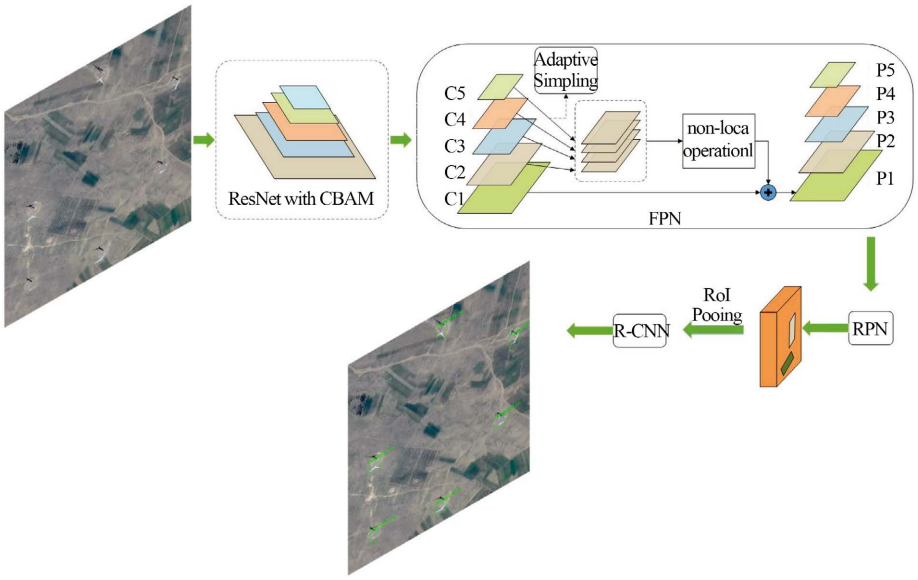


Figure 1. Network structure
图 1. 网络结构

3.1. 基于注意力机制改进的特征提取

本文在特征提取过程中应用注意力机制来解决复杂背景带来的问题，缓解特征提取过程中小物体携带的语义信息减少甚至消失的问题。卷积注意力模块(CBAM)，是一种用于前馈卷积神经网络的简单而有效的注意力模块。CBAM 模块整体结果如图 2 所示，首先会将特征图通过平均值池化操作和最大值池化操作在空间维度进行压缩成，得到一个一维矢量，进行训练，得到各通道的权重，然后与输入特征图做乘法操作来得到加权结果。接下来，会将通道注意力的结果作为输入，通过平均值池化操作和最大值池化操作在通道维度进行压缩，得到一个二维的空间注意力图，进行训练，得到空间上各位置的权重，然后与输入特征图做乘法操作来得到最终特征。

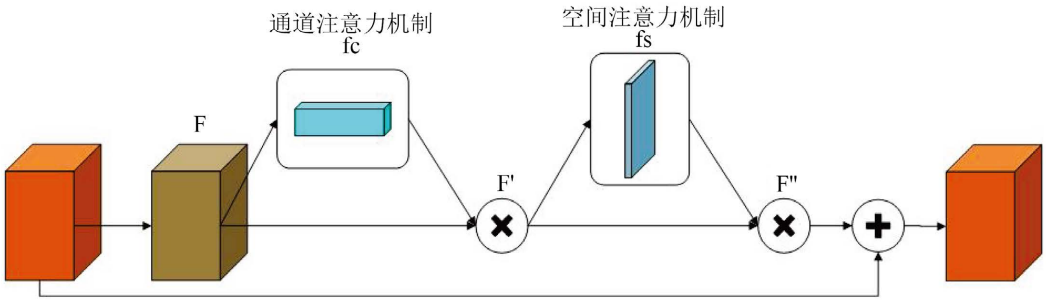


Figure 2. Network structure
图 2. 网络结构

将 $F \in \mathbb{R}^{C\times H\times W}$ 定义为输入的特征图， $M_C \in \mathbb{R}^{C\times 1\times 1}$ 定义为一维的通道注意力图， $M_S \in \mathbb{R}^{1\times H\times W}$ 定义为二维的空间注意力图，CBAM 模块的处理逻辑如等式(1)、(2)所示。

$$F' = M_C(F) \otimes F \tag{1}$$

$$F'' = M_S(F') \otimes F' \tag{2}$$

其中， $\otimes$ 表示乘法操作， F' 为通道注意力机制处理的结果， F'' 为经过空间注意力机制处理得到的最终结果。通道注意力机制与空间注意力机制处理逻辑如等式(3)、(4)所示。

$$M_C = \sigma\left(W_1 * \left(W_0\left(F_{\text{avg}}^C\right)\right) + W_1 * \left(W_0\left(F_{\max}^C\right)\right)\right) \tag{3}$$

$$M_S = \sigma\left(f^{7\times 7}\left(\left[F_{\text{avg}}^S; F_{\max}^S\right]\right)\right) \tag{4}$$

其中， W_0 、 W_1 为多层感知器的参数， σ 为 RELU 激活函数， $f^{7\times 7}$ 为 7 × 7 大小的卷积核。

在这里，我们选择残差网络 Resnet-101 作为我们的主干，并将 CBAM 插入主干的 Stages 3 到 5。首先，给定任意大小的遥感图像作为输入，在残差网络中，快捷连接可以解决由于层数较深而引起的退化问题。在残差网络的第 3 到第 5 阶段，将 CBAM 模块插入到每个残差模块中。通过以上操作，在提取图像特征的过程中调整特征权重。这使得特征提取网络能够专注于目标物体区域而不是背景噪声，从而选择性地增强目标物体的特征信息。

3.2. 基于特征融合改进的特征金字塔

特征金字塔将特征提取网络所提取的浅层级信息和深层级信息进行融合，并基于融合后的集成特征进行目标检测，大大提高了目标多尺度的问题。但是特征金字塔对于特征提取网络各层级的信息所采用的融合方式更多关注的是相邻层级的特征信息，非相邻层级的特征信息在每次融合中都会被稀释一次。

致使最终的集成特征是不平衡的。因此本文将改进特征金字塔的融合方式，来得到一个质量更高的集成特征，以此来提高目标检测的准确性。

将C_l定义为第 l 层的特征图。为了整合多级特征并同时保留它们的语义层次，我们首先将多级特征$\{C_2,C_3,C_4,C_5\}$的大小调整为中间大小，即与C_2相同的大小，并使用插值和最大值分别池化。大小调整完以后，通过简单的平均获得平衡的语义特征，如等式所示。

$$C=\frac{1}{L}\sum_{l=l_{\min}}^{l_{\max}}C_l \tag{5}$$

在获得平衡的语义特征之后，使用非局部信息统计的注意力机制对其进行细化，使其更具有辨别力。非局部信息统计的注意力机制可以建立远程依赖，建立图像上两个有一定距离的像素之间的联系。通过计算任意两个位置之间的交互直接捕捉远程依赖，而不用局限于相邻点，其相当于构造了一个和特征图谱尺寸一样大的卷积核，如等式所示。

$$y_i=\frac{1}{C(x)}\sum_{\forall j}f\left(x_i,x_j\right)g\left(x_j\right) \tag{6}$$

其中，i 是要计算其响应的输出位置的索引，j 是枚举所有可能位置的索引。x 是输入特征，y 是与 x 大小相同的输出信号。成对函数 f 计算 i 和所有 j 之间的关联。一元函数 g 计算位置 j 处的输入特征的表示。响应由系数$C(x)$进行归一化。

在这里，我们采用的是高斯非局部信息统计，最终的运算逻辑如等式(7)、(8)、(9)、(10)所示。

$$f\left(x_i,x_j\right)=\mathrm{e}^{x_i^{\mathrm{T}}x_j} \tag{7}$$

$$C(x)=\sum\nolimits_{\forall j}\mathrm{e}^{x_i^{\mathrm{T}}x_j} \tag{8}$$

$$g\left(x_j\right)=W_g x_j \tag{9}$$

$$y_i=\frac{1}{\sum_{\forall j}\mathrm{e}^{x_i^{\mathrm{T}}x_j}}\mathrm{e}^{x_i^{\mathrm{T}}x_j}W_g x_j \tag{10}$$

其中，输入是 x，输出是 y，i 和 j 分别代表输入的某个空间位置，W_g 是一个需要学习的权重矩阵。

4. 实验与分析

4.1. 数据集

本次实验采用 DIOR 遥感数据集[13]。DIOR 是一个大型的遥感数据集，包含 23,463 张遥感图片，其中共有 20 种不同的类别。该数据集分为三部分：训练集、验证集和测试集。训练集包含 5862 张图片，验证集包含 5863 张图片，测试集包含 11,738 张图片。

4.2. 评估标准

为了评估目标检测模型的性能，本文采用全类平均精度(*mAP*)来作为评估指标，如果预测的目标与真实框(*GT*)的交并比(*IoU*)大于 0.5，则认为预测结果为真正例。全类平均精度定义见公式。

$$mAP=\frac{1}{B}\sum\nolimits_{i=1}^{B}(AP)_i \tag{11}$$

$$AP=\sum_{k=1}^{N}P(k)\Delta r(k) \tag{12}$$

$$IoU = \frac{A \cup B}{A \cap B} \tag{13}$$

其中，B 为类别总数，N 为判断正负例的阈值，取值范围为[0,1]，$P(k)$为每一阈值下的精度值，而则表示阈值从 $k-1$ 变化到 k 时召回率的变化情况，A 为预测的预测框，B 为真实框。

4.3. 实验结果

本文研究是基于 Faster RCNN with FPN 所改进的适用于遥感图像的目标检测模型。为了测试该模型的有效性，我们将 SSD、YOLOV3、Faster RCNN with FPN 和 RetinaNet 等主流的目标检测模型与本文模型进行实验对比，具体实验结果如表 1 所示。观察表 1 可得，本文模型在 DIOR 数据集上的全类平均精度为 67.1%，相比于原来的 Faster RCNN with FPN 的检测模型，其检测精度提高了 2.0%，相比于 RetinaNet 也有 1.0%的精度提升。我们的模型在网球场、体育场、地面田径场、烟囱、篮球场、棒球场、储存罐和飞机等 8 个类别上取得了最好的表现，并且其他类别也维持着较高的水平。这些类别具有遥感图像中目标的特点。网球场、体育场、地面田径场、烟囱、篮球场和棒球场等从高空拍摄，特征不明显并且背景复杂。储存罐和飞机等目标，体积小，并且聚集密集也是遥感图像中的一大特点。这充分表明本文提出的模型适用于遥感图像。

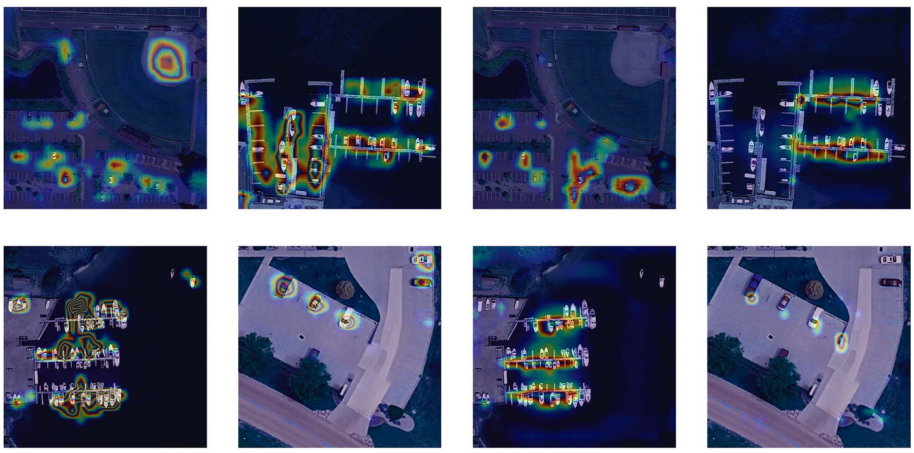
Table 1. Comparison of the detection accuracy of each target detection model in each category of the DIOR data set
表 1. 各目标检测模型在 DIOR 数据集各类别的检测精度对比

	SSD	YOLOv3	Faster RCNN with FPN	RetinaNet	Ours
Airplane	59.5	72.2	54.0	53.3	**80.0**
Airport	72.7	29.2	74.5	**77.0**	75.1
Baseball field	72.4	74.0	63.3	69.3	**78.1**
Basketball court	75.7	78.6	80.7	85.0	**85.8**
Bridge	29.7	31.2	**44.8**	44.1	37.1
Chimney	65.8	69.7	72.5	73.2	**78.7**
Dam	56.6	26.9	60.0	**62.4**	57.0
Expressway service area	63.5	48.6	75.6	**78.6**	71.8
Expressway toll station	53.1	54.4	62.3	**62.8**	52.9
Golf course	65.3	31.1	76.0	**78.6**	73.4
Ground track field	68.6	61.1	76.8	76.6	**78.5**
Harbor	49.4	44.9	46.4	**49.9**	48.1
Overpass	48.1	49.7	57.2	**59.6**	57.0
Ship	59.2	87.4	71.8	**71.1**	70.8
Stadium	61.0	70.6	68.3	68.4	**74.8**
Storage tank	46.6	68.7	53.8	45.8	**61.6**
Tennis court	76.3	87.3	81.1	81.3	**86.5**
Train station	55.1	29.4	**59.5**	55.2	54.1
Vehicle	27.4	**48.3**	43.1	44.4	41.7
Wind mill	65.7	78.7	81.2	**85.5**	79.2
mAP	58.6	57.1	65.1	66.1	**67.1**

为了进一步的验证模型的检测效果，我们将模型目标检测的结果进行一个可视化，展示了在 DIOR 数据集上的一些目标检测示例，具体结果如图 3 所示。

Figure 3. Example of target detection of this model in the DIOR dataset
图 3. 本文模型在 DIOR 数据集的目标检测示例

同时，我们使用加权梯度类激活映射(Grad-CAM)可视化了模型提取到的特征。图 4 中，对比观察到使用了 CBAM 模块的方法提取的特征与原始 ResNet-101 提取的特征，可以得出使用 CBAM 模块所提取到的特征更为显著、更为全面。这表明特征提取的过程中使用 CBAM 模块可以增强特征提取的能力。因此，实验结果表明了我们引入的 CBAM 模块的有效性。



(a) (b)

Figure 4. Use Grad-CAM to visualize the features extracted from the model. (a) is the feature extracted by the model in this paper, (b) is the feature extracted by ResNet-101
图 4. 使用 Grad-CAM 可视化模型所提取的特征。(a)为本文模型所提取的特征，(b)为 ResNet-101 所提取的特征

综上所述，在 DIOR 大型遥感数据集上的实验表明我们的模型提高了遥感图像的目标检测精度，并且在 8 个类别上取得了最好的表现。但是众多目标检测模型在在桥梁、火车站和港口取得效果不佳，我们的模型也不例外。原因可能是这些类别的目标形状都较为细长，长宽比比较极端，导致锚框拟合的效果不佳。后续可以对数据集中目标形状做抽样统计，基于统计结果来设计锚框比等超参数，或者可以尝试基于关键点的目标检测方法去自适应拟合目标的形状。

5. 结语

遥感图像与自然图像之间的差异，导致目标检测算法在遥感图像中效果不佳。本文提出一个适用于遥感图像的目标检测模型，将注意力机制运用在特征提取中，使得特征提取网络能够专注于目标物体区域而不是背景噪声，从而选择性地增强目标物体的特征信息。运用自注意力机制对特征金字塔各层级的特征图进行融合，更好地解决遥感图像中的目标多尺度问题。该模型相较于其他常见的目标检测模型，在 DIOR 数据集上精度有较大的提升。

基金项目

广东省基础与应用基础研究基金项目(2020A1515011409)；广东省信息物理融合系统重点实验室(2020B1212060069)；大南海区域广东高分大数据平台与应用示范(83-Y40G33-9001-18/20)；“农业区块链”共性关键技术研发创新团队(2019KJ147)。

参考文献

[1] Jiao, L., *et al.* (2019) A Survey of Deep Learning-Based Object Detection. *IEEE Access*, **7**, 128837-128868. https://doi.org/10.1109/ACCESS.2019.2939201

[2] Ren, S., He, K., Girshick, R. and Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[3] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. arXiv:1804.02767 [cs.CV]

[4] Liu, W., Anguelov, D., Erhan, D., *et al.* (2016) SSD: Single Shot MultiBox Detector. arXiv:1512.02325 [cs.CV] https://doi.org/10.1007/978-3-319-46448-0_2

[5] Fatima, S.A., Kumar, A., Pratap, A., *et al.* (2020) Object Recognition and Detection in Remote Sensing Images: A Comparative Study. 2020 *International Conference on Artificial Intelligence and Signal Processing* (*AISP*), Amaravati, 10-12 January 2020, 1-5. https://doi.org/10.1109/AISP48273.2020.9073614

[6] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

[7] He, K., Zhang, X., Ren, S., *et al.* (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[8] Woo, S., Park, J., Lee, J.Y., *et al.* (2018) CBAM: Convolutional Block Attention Module. *European Conference on Computer Vision*, Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[9] Jie, H., Li, S., Gang, S., *et al.* (2017) Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2011-2023. https://doi.org/10.1109/TPAMI.2019.2913372

[10] Lin, T.Y., Dollar, P., Girshick, R., *et al.* (2017) Feature Pyramid Networks for Object Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[11] Oksuz, K., Cam, B.C., Kalkan, S., *et al.* (2020) Imbalance Problems in Object Detection: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 3388-3415. https://doi.org/10.1109/TPAMI.2020.2981890

[12] Wang, X., Girshick, R., Gupta, A., *et al.* (2017) Non-Local Neural Networks. arXiv:1711.07971 [cs.CV] https://doi.org/10.1109/CVPR.2018.00813

[13] Li, K., Wan, G., Cheng, G., *et al.* (2019) Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. arXiv:1909.00133 [cs.CV] https://doi.org/10.1016/j.isprsjprs.2019.11.023