结合 DenseNet 与通道注意力机制的空对地目标检测算法

王文庆[1]，丰林[1]，刘洋[1*]，杨东方[2]，张萌[2]
[1]西安邮电大学自动化学院，陕西 西安 710121；
[2]火箭军工程大学导弹工程学院，陕西 西安 710025

摘要 空对地环境下成像视角单一，且需要依靠深层网络提供强特征表达能力。针对深层网络存在的计算量大、收敛速度慢等问题，在稠密连接网络(DenseNet)框架下，提出了一种用通道差异化表示的目标检测网络模型。首先，用 DenseNet 作为特征提取网络，并用较少的参数加深网络，以提高网络对目标的提取能力；其次，引入通道注意力机制，使网络更关注特征层中的有效特征通道，重新调整特征图；最后，用空对地目标检测数据进行了对比实验。结果表明，改进模型的平均精度均值比基于视觉几何组(VGG16)的单步多框检测算法高 3.44 个百分点。

关键词 图像处理；目标检测算法；特征提取；通道注意力机制；有效特征；密集连接
中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP 57.221010

Air-to-Ground Target Detection Algorithm Based on DenseNet and Channel Attention Mechanism

Wang Wenqing[1], Feng Lin[1], Liu Yang[1*], Yang Dongfang[2], Zhang Meng[2]
[1]*College of Automation, Xi'an University of Posts & Telecommunications, Xi'an, Shaanxi* 710121, *China*;
[2]*College of Missile Engineering, Rocket Force University of Engineering, Xi'an, Shaanxi* 710025, *China*

Abstract In the air-to-ground environment, the imaging perspective is single, and it is necessary to rely on deep network to provide stronger feature representation capabilities. Aiming at the problems of large amount of calculation and slow convergence speed brought by deep network. Under the framework of densely connected network (DenseNet), a target detection network model expressed by channel differentiation is proposed. First, this article uses DenseNet as a feature extraction network, and uses fewer parameters to deepen the network to improve the ability to extract objects. Second, channel attention mechanism is introduced to make the network pay more attention to the effective feature channels in the feature layer and readjust the feature map. Finally, a comparative experiment is carried out by using the air-to-ground object detection data. The results show that the mean average precision of the improved model is 3.44 percentage points higher than that of single shot multibox detection algorithm based on visual geometry group (VGG16).

Key words image processing; target detection algorithm; feature extraction; channel attention mechanism; effective feature; dense connection
OCIS codes 100.2000; 150.1135; 150.0150; 280.0280

1 引 言

空对地的目标检测广泛应用于多个领域中，如对地面的侦察与精确打击、海面的巡航与搜救、地面交通状况的监控、自然灾害救援。近年来，计算机视觉的研究推动了人工智能技术的快速发展，赋予了机器“看”的能力，而目标检测是计算机视觉领域中一个长期和具有挑战性的任务，也是解决更复

收稿日期：2020-02-07；**修回日期**：2020-03-25；**录用日期**：2020-04-13
基金项目：国家自然科学基金(61673017，61403398)、陕西省自然科学基金(2017JM6077)、陕西省科技厅重点项目(2018ZDXM-GY-039)

*E-mail：yyangbrand@163.com

杂视觉问题的基础。可将对目标检测的研究分为两种类型[1-2]，一种是对特定实例的检测，包括检测单个类别(如面孔、行人)或少数特定类别。第二种是通用目标检测，可检测不同的预定义目标类别(如人、飞机、汽车)。自深度学习系统在 2012 年 ImageNet 举办的图像分类竞赛中获得冠军后，卷积神经网络(CNN)在图像领域取得了更高的分类精度[3]，多数计算机视觉的相关研究都集中在深度学习方法上。近年来，基于深度学习的目标检测取得了巨大进步[4]，主要算法框架分为一步(One-stage)目标检测算法与两步(Two-stage)目标检测算法。代表性的 Two-stage[5-7] 算法有 R-CNN(Region-CNN)、Fast R-CNN、Faster R-CNN[8] 等，这些算法放弃使用滑动窗口进行区域选择，采取有建议性的区域生成算法，如选择性搜索(SS)算法，减少了冗余候选区域的生成；特征提取也从传统的手工设计变为利用深度卷积网络[9-10]学习和提取图像中目标的深层特征，最后将特征向量进行分类和精确定位。Two-stage 系列检测算法在检测精度上取得了很好的效果，但仍然无法实现端对端的实时检测。One-stage 检测算法[11-12]可直接对输入图像进行类别预测与定位，比 Two-stage 算法的精确度略低，但检测速度有很大的提升，可实现端对端的检测，很好地平衡了检测精度和检测速度，如 YOLO(You only look once)[12] 和单步多框检测(SSD)[11] 算法。

SSD 算法是基于多尺度图像的检测，在理想条件下对小目标的检测效果较好，但在成像视角单一的空对地场景下，拍摄的目标尺寸较小、结构信息较少，且容易受到复杂背景的干扰，对 SSD 算法的特征提取网络性能要求较高。针对传统 SSD 算法的主干网络——视觉几何组(VGG16)网络[13] 参数量大、检测率低的问题，本文提出了一种结合密集连接网络(DenseNet)和压缩激励(SE)的单步多框检测(SE-DenseSSD)算法。用 DenseNet 作为 SSD 算法的主干网络，DenseNet 具有参数少和特征复用的特点，能提取到更丰富有效的特征；考虑到空对地目标易受到背景干扰，一些背景信息可能分布在特征图的某些通道上，在生成候选区域的不同尺度特征图中引入通道注意力机制。在训练过程中，可根据各通道特征的有效程度赋予不同的权值，将权值大的有效信息进行重点表达，权值小的特征通道作为冗余通道，从而减少背景信息的干扰。

2 算法描述

用 SSD 算法作为基本框架，用 DenseNet 作为 SSD 算法的特征提取网络，同时引入通道注意力学习机制，提出了一种适用于空对地任务的目标检测算法。

2.1 特征提取网络 DenseNet

相比常规图像，空对地图像的目标特征较少，VGG 网络的连接方式与浅层特征提取网络对目标特征的提取能力有限，而直接增加网络层数会导致参数量变大，容易出现过拟合问题，因此，需要对网络进行改进。传统改进方法通过加深网络层数或加宽网络结构使网络能更好地挖掘输入数据，如 Google 公司提出的 Inception 系列网络[14]。He 等[10]提出的残差神经网络(ResNet)通过引入残差块[15]，有效提高了网络性能。但随着网络层数的加深，网络提取的特征越抽象，导致模型丢失了中间层部分的细节信息，同时增加了运行成本。DenseNet 可以缓解中间层信息随网络深度增加和变宽带来的问题，主要由多个 Dense block 组成，每个 Dense block 内部的连接如图 1 所示。其中，$\boldsymbol{X}_0$ 为输入，$\boldsymbol{H}$ 为批量标准化(BN)、线性整流函数(ReLU)、卷积三种操作对应的非线性转化函数，$\boldsymbol{H}_1$ 的输入为 $\boldsymbol{X}_0$，$\boldsymbol{H}_2$ 的输入为 $\boldsymbol{X}_0$ 和 $\boldsymbol{X}_1$，$\boldsymbol{X}_1$ 为 $\boldsymbol{H}_1$ 的输出，经 l 次的非线性转化函数后，最终得到的 $\boldsymbol{X}_l$ 可表示为

$$\boldsymbol{X}_l=\boldsymbol{H}_l(\boldsymbol{X}_{l-1})=\boldsymbol{H}_l[\boldsymbol{X}_0,\boldsymbol{X}_1,\cdots,\boldsymbol{X}_{l-1}]。\quad(1)$$

设增长速率为 k，即经过 $\boldsymbol{H}_l$ 后可得到 k 个通道特征图，设置合适的 k 可以保持通道维度适中，防止网络过宽。两个 Dense block 之间包括一个由 BN+Conv(1×1)+Average pooling(1×1)组成的 transition layer，其中，大小为 1×1 的卷积可进行降维处理，提高计算效率。

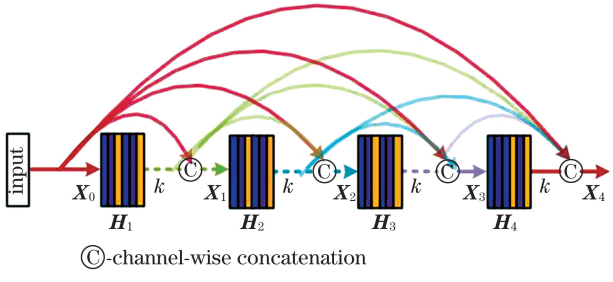

图 1 Dense block 的结构图

Fig. 1 Structure diagram of Dense block

传统 CNN 中，如果网络层数为 L，则有 L 个连接，但在 DenseNet 中，有 $L(L+1)/2$ 个连接，因此也被称为密集连接(dense connection)。网络中每层的输入都是前面所有层输出结果的叠加，不仅解决了梯度消失的问题，还在加深网络层数的同时减少了网络参数。

2.2 通道注意力机制

在空对地场景中，一些复杂的背景干扰信息会分布在特征图的某些通道上，从而在训练过程中阻碍检测精度的提升，导致误检率与漏检率上升。因此，用注意力机制重新调整特征通道，不再平均考虑特征图中所有通道的特征信息，而是赋予特征图各通道不同的权值。根据权值大小对特征通道进行筛选，使模型聚焦于更有效的特征，从而减轻背景对目标的干扰。为了通过学习得到每个通道的权值，引入压缩激励网络(SENet)[16]，并将其简化为一般形式，如图 2 所示。

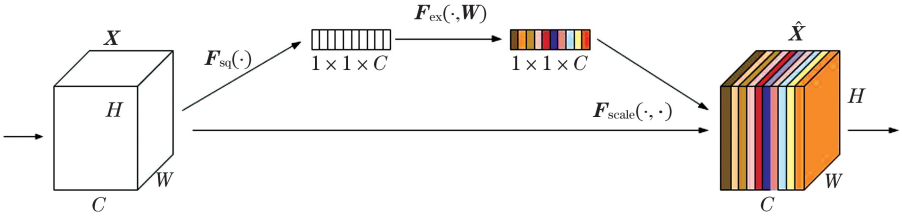

图 2 SENet 的结构

Fig. 2 Structure of SENet

图 2 中，$\boldsymbol{X}$ 为输入特征图，$\boldsymbol{F}_{\mathrm{sq}}$ 为压缩函数，$\boldsymbol{F}_{\mathrm{ex}}$ 为激励函数，$\boldsymbol{F}_{\mathrm{scale}}$ 为加权函数，$\hat{\boldsymbol{X}}$ 为输出特征图，其中，$\boldsymbol{X}\in\boldsymbol{R}^{H\times W\times C}$，共包含 C 个通道。各个通道的权值都是由特征图在空间维度 $H\times W$ 上压缩得到，即对每个通道的全局信息进行软编码，经压缩通道得到的权值矩阵可表示为

$$\boldsymbol{Z}=\boldsymbol{F}_{\mathrm{sq}}(\boldsymbol{X}_c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{X}_c(i,j),\quad(2)$$

式中，c 为变量，取值范围为 $1\sim C$，(i,j) 为对该通道值的索引。

为了自适应生成各通道的注意力权值，引入一个具有隐含层的多层感知机。隐含层的神经元个数为 C/r，r 为超参数，合适的 r 可以提高计算效率。用 Sigmoid 函数得到最终的权值矩阵为

$$\boldsymbol{S}_c=\boldsymbol{F}_{\mathrm{ex}}(\boldsymbol{Z},\boldsymbol{W})=\sigma[g(\boldsymbol{Z},\boldsymbol{W})]=\sigma[\boldsymbol{W}_2\delta(\boldsymbol{W}_1\boldsymbol{Z})],\quad(3)$$

式中，$g(\boldsymbol{Z},\boldsymbol{W})$ 为两层的神经网络，$\boldsymbol{W}$ 为整个网络的权重，$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 分别为网络中第一、第二隐层的权重，δ 为 ReLU 激活函数，σ 为 Sigmoid 激活函数，$\boldsymbol{S}_c$ 的维度为 $1\times1\times C$。

用 $\boldsymbol{S}_c$ 对输入通道进行调整，通道注意力加权后的输出图像可表示为

$$\hat{\boldsymbol{X}}_c=\boldsymbol{F}_{\mathrm{scale}}(\boldsymbol{X}_c,\boldsymbol{S}_c)=\boldsymbol{X}_c\otimes\boldsymbol{S}_c,\quad(4)$$

式中，$\otimes$ 为逐元素相乘符号。

2.3 SE-DenseSSD 算法的网络化描述

实验使用的检测算法结构如图 3 所示，用

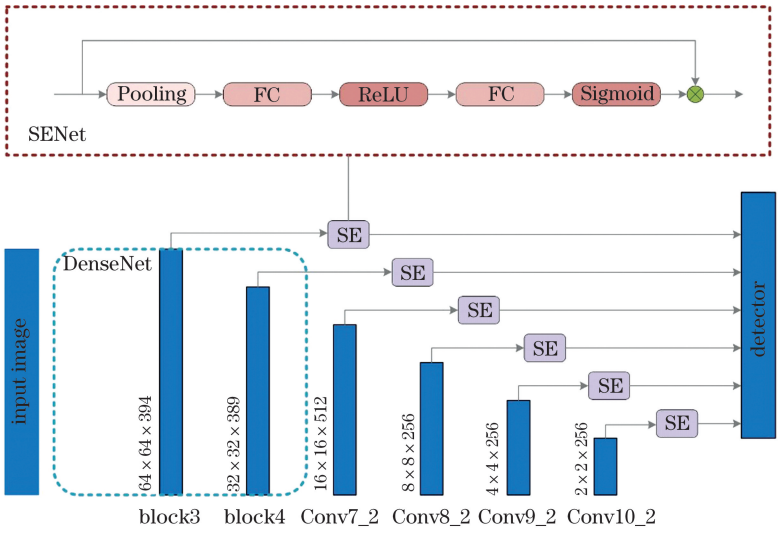

图 3 SE-DenseSSD 算法的结构图

Fig. 3 Structure diagram of SE-DenseSSD algorithm

DenseNet 作为 SSD 算法的主干网络，共设置 4 个 block。其中，block3、block4 的输出层分别对应原 VGG16 的 Conv4_3 与转换为全连接层的 Conv7，即 SSD 算法的基本框架由 DenseNet 的 block3、block4 以及 4 个递减的卷积层 Conv7_2、Conv8_2、Conv9_2、Conv10_2 组成。SSD 算法可在特征图中通过不同大小、纵横比的目标候选框对不同尺寸的目标类别和位置进行回归[17]。输入图像的尺寸为 512 pixel×512 pixel，经 4 个 Dense block（$\boldsymbol{H}_l$ 中的 l 分别取 3、6、12、8 组成的 DenesNet）。为了防止网络增长过快，将 k 设置为 24。在 block 中，利用密集连接运算使上一层的特征图与本层产生的特征图进行合并，最后一层会得到所有特征图的并集，有效缓解了梯度消失问题。block3、block4 的最终输出层与额外层（Conv7_2、Conv8_2、Conv9_2、Conv10_2）作为用于检测的特征图。其次，将每个 Dense block 与额外层作为 SENet 的输入 $\boldsymbol{X}$，经过全局平均池化（Pooling）与两个全连接层（FC），其中，第一个 FC 的缩放系数 $r=2$，可有效提高计算效率；再通过激活函数 Sigmoid 得到各通道的权值，与原输入 $\boldsymbol{X}$ 对应相乘得到最终用于检测的特征图，如图 2 中的虚线框所示。利用深层 DenseNet 高效的特征提取能力，得到不同尺度的特征图，经过通道注意力机制得到特征显著的特征图，然后再进行目标检测。

3 实验分析

为验证本算法的性能，在空对地数据集上进行测试，并与基于 VGG16 的 SSD 算法进行对比分析。

3.1 实验数据集

常用于深度学习目标检测算法研究的数据集，如 Pascal VOC[18]、ILSVRC[19]、COCO[20] 数据集，虽然可为目标检测算法提供大量的训练资源，但这些数据集获取的大多是常规视角下的内容，难以满足空对地背景下的目标检测。空对地检测的数据集往往需要无人机等飞行器进行采集，因此，公开的空对地目标检测数据集相对较少，且规模和样本质量差异较大。针对该问题，实验通过无人机对道路进行拍摄，共获得 7000 张样本图像。由于成像传感器的焦距、分辨率等内参和高度会影响空对地的成像结果，从而对目标检测的设计和实际检测效果产生影响。因此，根据目标在整张图像中所占的像素，衡量实验获得的数据集。在标注过程中，目标占用的像素应不少于 10 pixel。共标定了 5 类检测目标，分别为汽车、卡车、厢式货车、公交车、人。图 4 为数据集中某张图像的一部分区域，可以看出，在该视角下，不同车辆的信息都有所损失，有的受到遮挡，有的所占像素较少，且不同时段道路场景中车辆的亮度也不同，在这些更真实的数据下，挖掘深层的目标特征尤为重要。

3.2 实验参数设置

数据集中的图像尺寸为 1920 pixel×1080 pixel，为了验证本算法的有效性，用 SSD512 算法直接对原始图像进行训练。训练阶段，按 7∶1∶2 的比例将数据集划分为训练集、验证集和测试集，设置模型的初始学习率为 0.001，batch size 为 16，训练次数（epoch）为 800 次。

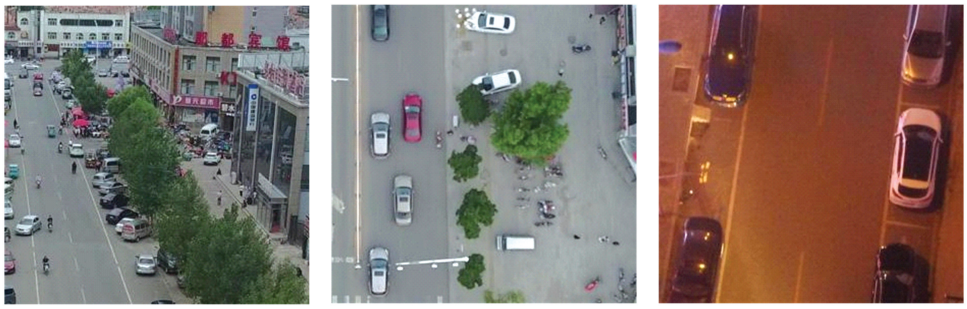

图 4 数据集中的样本图像

Fig. 4 Sample images in the data set

3.3 评价指标和结果分析

为了对比本算法和 VGG-SSD 系列算法的检测性能，用平均精度均值（mAP）作为性能评价指标。平均精度（AP）能衡量检测算法在每个类别上的性能好坏，可表示为

$$X_{\mathrm{AP}}=\int_0^1 P(R)\,\mathrm{d}R, \tag{5}$$

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}}, \tag{6}$$

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}},\tag{7}$$

式中，P 为精确率（Precision），R 为召回率（Recall），N_{TP} 为正确识别的目标数量，N_{FP} 为将非目标识别为目标的数量，N_{FN} 为检测中未识别出的目标数量，$N_{\mathrm{TP}}+N_{\mathrm{FP}}$ 为检测中识别出的目标数量，$N_{\mathrm{TP}}+N_{\mathrm{FN}}$ 为图像中标注的目标数量。

mAP 可评价检测算法在多目标类别检测中的性能，可表示为

$$X_{\mathrm{mAP}}=\frac{1}{N}\sum_{i=1}^{N}X_{\mathrm{AP}}^{i},\tag{8}$$

式中，N 为检测的类别数，X_{AP}^{i} 为类别 i 的检测精度。

为了验证本算法用 DenseNet 作为特征提取网络与引入通道注意力学习机制的有效性，在相同数据规模下，用基于 VGG16 的 SSD 算法、只使用 DenseNet 作为特征提取网络的 SSD（DenseSSD）算法、本算法分别进行训练，得到三种算法的 mAP 和模型参数大小如表 1 所示。可以发现，VGG-SSD 算法、DenseSSD 算法的 mAP 分别为 36.53%、38.36%，本算法的 mAP 为 39.97%，相比 DenseSSD 和 VGG-SSD 算法分别提高了 1.61 和 3.44 个百分点，这表明改进后的模型可在提升目标检测算法精度的同时减少参数量。

表 1 不同算法得到模型参数和 mAP

Table 1 Model parameters and mAP obtained by different algorithms

Algorithm	mAP/%	Parameter/ MB
VGG-SSD	36.53	101
DenseSSD	**38.36**	**17**
SE-DenseSSD	**39.97**	**21**

图 5 为不同算法训练得到的 mAP，可以发现，在相同的训练次数下，本算法中网络模型的 mAP 均高于原网络，在较少的训练资源下，可以实现较快的训练收敛速度，提升训练性能。

不同算法的可视化结果如图 6 所示，可以发现，由于目标特征较少或复杂背景等因素的干扰，VGG-SSD 算法得到的模型容易出现漏检现象，而 SE-DenseSSD 算法得到的模型能有效降低所有类别的漏检率，从而更好地学习目标对象的特征，有效减弱复杂背景的干扰，提高空对地目标检测算法的准确性和鲁棒性。

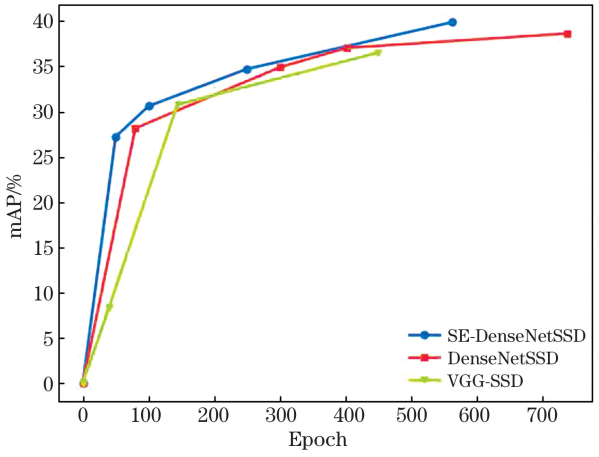

图 5 不同算法的训练曲线

Fig. 5 Training curves of different algorithms

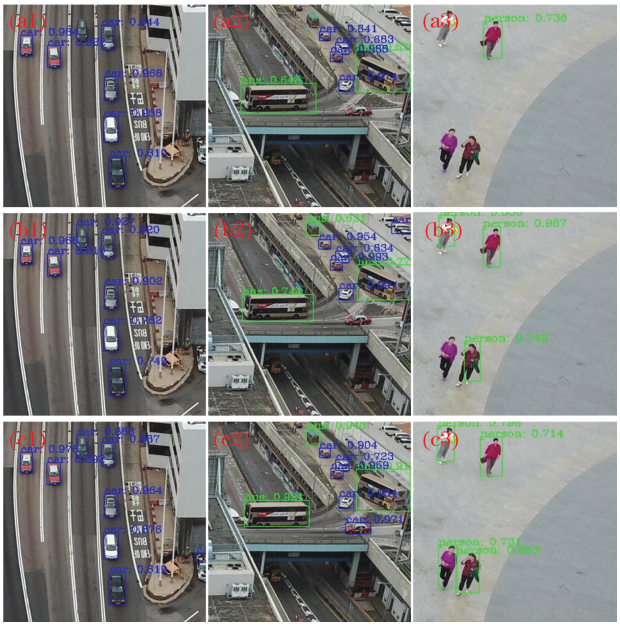

图 6 不同算法的检测结果。(a) VGG-SSD；(b) DenseSSD；(c) SE-DenseSSD

Fig. 6 Test results of different algorithms. (a) VGG-SSD; (b) DenseSSD; (c) SE-DenseSSD

4 结 论

针对空对地场景中目标信息量少且容易被复杂背景干扰的问题，基于 SSD 算法的检测框架提出了 SE-DenseSSD 算法。通过引入 DenseNet 的密集连接方式，在加深网络的同时有效防止过拟合、提高训练过程的收敛速度；通过引入通道注意力 SE 机制，构建了卷积特征通道之间的相互依赖性，自适应地校准通道权值，减少复杂背景特征的响应。实验结果表明，本算法能有效提高对空对地目标的检测精度。

参 考 文 献

[1] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[2] Grauman K, Leibe B. Visual object recognition[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2011, 5(2): 1-181.

[3] Xu L X, Chen X J, Ban Y, et al. Method for intelligent detection of parking spaces based on deep learning[J]. Chinese Journal of Lasers, 2019, 46(4): 0404013.
徐乐先, 陈西江, 班亚, 等. 基于深度学习的车位智能检测方法[J]. 中国激光, 2019, 46(4): 0404013.

[4] Cui J H, Zhang Y Z, Wang Z, et al. Light-weight object detection networks for embedded platform[J]. Acta Optica Sinica, 2019, 39(4): 0415006.
崔家华, 张云洲, 王争, 等. 面向嵌入式平台的轻量级目标检测网络[J]. 光学学报, 2019, 39(4): 0415006.

[5] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386-397.

[6] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[EB/OL]. [2020-01-25]. http://de.arxiv.org/pdf/1506.01497.

[7] 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.

[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] Howard A G, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. [2020-01-30]. https://arxiv.org/abs/1704.04861.

[10] 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, 2016: 770-778.

[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, 9905: 21-37.

[12] 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, 2016: 779-788.

[13] Zhao H, An W S. Image salient object detection combined with deep learning[J]. Laser & Optoelectronics Progress, 2018, 55(12): 121003.
赵恒, 安维胜. 结合深度学习的图像显著目标检测[J]. 激光与光电子学进展, 2018, 55(12): 121003.

[14] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 2818-2826.

[15] Wu P, Lin G Q, Guo Y R, et al. Self-learning sparse DenseNet image classification method[J]. Journal of Signal Processing, 2019, 35(10): 1747-1752.
吴鹏, 林国强, 郭玉荣, 等. 自学习稀疏密集连接卷积神经网络图像分类方法[J]. 信号处理, 2019, 35(10): 1747-1752.

[16] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[EB/OL]. [2020-01-31]. https://arxiv.org/abs/1709.01507.

[17] 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.

[18] Everingham M, van Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[19] 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.

[20] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[EB/OL]. [2020-01-27]. https://arxiv.org/abs/1405.0312.