引用格式：范红超,李万志,章超权.基于Anchor-free的交通标志检测[J].地球信息科学学报,2020,22(1):88-99. [Fan H C, Li W Z, Zhang C Q. Anchor-free traffic sign detection[J]. Journal of Geo-information Science, 2020,22(1):88-99.] **DOI:**10.12082/dqxxkx.2020.190424

基于Anchor-free的交通标志检测

范红超[1],李万志[2],章超权[1,2]

1. 挪威科技大学,特隆赫姆 7491;2. 武汉大学,武汉 430072

Anchor-Free Traffic Sign Detection

FAN Hongchao[1*], LI Wanzhi[2], ZHANG Chaoquan[1,2]

1. Norwegian University of Science and Technology, Trondheim 7491, Norway; 2. Wuhan University, Wuhan 430072, China

Abstract: Traffic signs are essential elements in High Definition (HD) maps and hence very important for vehicles in autonomous driving. Real-time and accurate detection of traffic signs from street level images is of great significance for the development of autonomous driving. Conventional algorithms detect traffic signs based on image color and shape features, and can only work for specific kinds of traffic signs. Algorithms based on image feature and machine learning classifier need artificial designed features, and the detection speed is slow. To date, many approaches using deep learning methods have been developed based on anchor boxes, which introduce extra hyper parameters in network design. When switching to a different detection task, anchor boxes need to be redesigned. Anchor-based methods also generate massive redundant anchor boxes during model training, which easily cause imbalance between positive and negative samples. Inspired by the idea of anchor-free and YOLO, this paper proposed a real-time traffic sign detection network called AF-TSD, which regresses object boundary directly. AF-TSD adopts an effective convolution module named deformable convolution to enhance the feature expression ability of convolutional neural networks. This module adds 2D offsets to the regular grid sampling locations in the standard convolution. It also modulates input feature amplitudes from different spatial locations/bins. Both the offsets and amplitudes are learned from the preceding feature maps, via additional convolutional layers. In addition, AF-TSD introduces attention mechanism. It is inserted after fusion of the feature pyramid, and adaptively recalibrates channel-wise feature responses by explicitly modeling the interdependencies between channels. This module first squeezes global spatial information into a channel descriptor. Then the excitation operator maps the input-specific descriptor to a set of channel weights. The attention mechanism in this paper is lightweight and imposes only a slight increase in model complexity and computational burden. To test the superiority of AF-TSD, extensive comparative experiments were carried out. We first evaluated the influence of different modules on detection precision. The experimental results show that the deformable convolution and attention mechanism can help extract features of traffic signs. Then, AF-TSD was compared with mainstream detection networks, including Faster R-CNN, RetinaNet, and YOLOv3. Our proposed AF-TSD traffic sign detection network achieved 96.80% of mAP on GTSDB traffic sign detection dataset, which was superior to mainstream detection algorithms. The average detection speed was 32ms per

收稿日期:2019-08-05;**修回日期:**2019-11-27.

基金项目:国家自然科学基金项目(41771484)。[**Foundation item:** National Natural Science Foundation of China, No.41771484.]

作者简介:范红超(1977—),男,湖北襄阳人,博士,教授,主要从事众源地理信息数据挖掘与分析研究。E-mail: hongchao.fan@ntnu.no

images, which can meet the requirements of real-time detection.

Key words: VGI data; traffic sign detection; convolutional neural networks; deformable convolution; attention mechanism; anchor-free; AF-TSD

***Corresponding author:** FAN Hongchao, E-mail: hongchao.fan@ntnu.no

摘 要：交通标志检测是自动驾驶中的重要研究方向，实时准确地从街景图像中检测交通标志对实现自动驾驶及智慧城市的发展具有重要意义。传统的算法基于颜色、形状特征进行检测，只能提取特定种类的交通标志，算法无法同时检测不同类型的交通标志。基于图像特征+机器学习分类器的算法需要人工设计特征，算法速度较慢。主流的基于深度学习的方法多基于先验框，在网络设计上引入了额外的超参数，且在训练过程中产生过量的冗余边界框，容易造成正负样本不平衡。本文受Anchor-free思想的启发，引用YOLO检测器直接回归物体边界框的思路，提出一种基于Anchor-free的实时交通标志检测网络AF-TSD（Anchor-free Traffic Sign Detection）。AF-TSD摒弃了先验框的设计，并引入自适应采样位置可变卷积与注意力机制，大大提高网络的特征表达能力。本文开展大量对比实验，实验结果表明本文提出的AF-TSD交通标志检测网络速度接近主流算法，但精度优于主流算法，在德国GTSDB交通标志检测数据集上取得了96.80%的精度，检测速度平均单张图片32 ms，达到实时检测的要求。

关键词：众源地理信息数据；交通标志检测；卷积神经网络；可变形卷积；注意力机制；Anchor-free；AF-TSD

1 引言

交通标志检测是自动驾驶中的重要研究方向，实时准确地从街景图像中检测交通标志对实现自动驾驶及智慧城市的发展具有重要意义。当前交通标志检测算法主要包括基于颜色、形状特征的传统图像处理算法、基于图像特征+机器学习分类器的算法和基于深度学习的目标检测算法，传统图像处理算法[1−6]利用交通标志显著的颜色特征或形状特征提取出特定种类的交通标志，算法无法同时检测不同类型的交通标志。基于图像特征+机器学习分类器的算法中比较具有代表性的方法是Harr特征+Adboost分类器[7−9]和HOG特征+SVM分类器[10−11]，然而这些算法需要人工设计特征，且算法速度较慢。随着卷积神经网络在计算机视觉领域的不断发展，基于深度学习的目标检测算法以及强大的特征表征能力及快速的检测速度得到学术界和工业界的青睐。

基于深度学习的检测算法大致分为2类：① FAIR（Facebook AI Research）研发的R-CNN系列为代表的两阶段检测算法，如R-CNN[12]，Faster R-CNN[13]，R-FCN[14]，Cascade R-CNN[15]。该系列算法首先从图像中预测高质量的区域候选框，然后分别连接分类和回归的子网络判断区域候选框的类别并矫正其位置。② 以YOLOv3[16]、SSD[17]为代表的单阶段检测算法，该系列算法在预测候选框的同时，进行物体类别的分类和位置的回归。

然而无论是两阶段检测算法还是单阶段检测算法，均基于先验框（Anchor Boxes），在离散的图像空间中产生大量的anchor boxes以求尽可能的覆盖感兴趣的目标。虽然anchor boxes思想促进了深度学习目标检测的发展，但会带来以下3个问题：

（1）Anchor boxes引入了额外的超参数。Anchor boxes的尺度、宽高比和数量都是需要考虑的因素，在RetinaNet[18]中，anchor boxes超参数的优化甚至使得算法在COCO数据基准[19]中的精度提升了4%。然而，如何最优地设计这些超参数将是研究者所需要面临的问题。

（2）基于anchor boxes的算法无法在不同的应用场景之间切换。预定义的先验框在遇到形状变化较大的数据时将无法检测到物体，例如交通标志检测anchor boxes更倾向于方形的先验框，而行人检测则更倾向于细长的矩形先验框。在切换应用场景时，往往需要根据不同的数据重新设计anchor boxes。

（3）冗余框非常多。大量的anchor boxes密集分布在图像上，然而只有很小一部分被标记为正样本，其余anchor boxes被标记为负样本，这会造成样本不平衡。同时，过量的anchor boxes会增加计算量和显存的消耗。

近两年，改进anchor boxes设计过程的方法逐渐在学术界出现。旷视科技提出MetaAnchor[20]，通过建模anchor函数动态生成anchor boxes。该方法对先验知识的要求已大大降低，anchor boxes也更加灵活，解决了anchor boxes超参数设定和尺度固定两个问题，但anchor boxes的冗余并未解决。密歇根大学提出CornerNet[21]，摒弃了anchor机制，通过预测物体边界框的左上角和右下角的2个关键点

完成定位，将检测任务转化为key-point任务来处理。然而CornerNet需要定位的角点语义信息并不充足，与人脸的关键点、人体的骨骼点不同，后者特征表达更为明显。德克萨斯大学科研人员提出ExtremeNet[22]，同样基于关键点，引入了目标的上、下、左、右4个方向上的极值点进行预测，特征表达意义明确，且在检测和实例分割任务中均取得较好的表现，然而检测速度非常慢。

相比之下，人眼视觉系统感知物体的位置和大小时并没有利用预定义的anchor boxes，同时也没有刻意寻找关键点来确定物体的位置和大小，仅仅通过聚焦物体并感知其宽高。DenseBox[23]、UnitBox[24]和FCOS[25]直接在特征图上预测空间点到物体框4个边界的距离，摆脱了对anchor boxes的依赖，FoveaBox[26]通过建模一个从特征图到物体边界坐标值的映射实现Anchor-free。受到Anchor-free成功思想的启发，本文引用YOLO[27]直接回归物体边界框的思路，提出一种基于Anchor-free的实时交通标志检测网络AF-TSD（Anchor-free Traffic Signs Detection）。YOLO将图像分成7×7的网格，在每个网格内预测2个边界框，每个边界框的信息包括其中心点偏离所在格网点的距离及其宽高，预测流程非常快速。然而YOLO的输出为全连接层结构，这限制了多尺度的图像输入。且7×7的特征图分辨率信息不足，极其容易损失交通标志的特征，不利于交通标志的检测。针对这些不足，本文提出的AF-TSD网络具有如下特性：

（1）AF-TSD为全卷积网络，适用于不同尺度的图像输入，基础网络为带有BatchNorm层的VGGNet[28]。此外引入自适应采样位置可变卷积DCN（Deformable Convolutional Networks[29-30]），提高网络表达交通标志特征的能力。

（2）考虑到交通标志尺度小、特征极其丢失，AF-TSD引入特征金字塔结构[31]，确保特征图具备充足的语义信息与分辨率信息。并于金字塔结构特征图融合处，利用注意力机制[32]对交通标志的特征进行过滤，增强积极的信息。

（3）本文提出的AF-TSD网络对街景图像中交通标志检测具有很强的适用性，并且基于Anchor-free，直接在特征图上回归出交通标志的中心点及其宽高。在德国交通标志检测数据集GTSDB[33]上的mAP（mean Average Precision）达到96.80%，平均单张街景图片检测速度为32 ms，满足实时检测的要求。

2 相关研究

交通标志检测本质上是一种实例化的目标检测。进入深度学习时代以来，目标检测的算法研究不断取得突破，并广泛应用于实际生活中。目标检测算法主要有以下4类：

2.1 两阶段检测器

两阶段检测器以FAIR的R-CNN系列算法为代表，随后衍生出Faster R-CNN、Mask R-CNN[34]、Cascade R-CNN等优秀算法。Faster R-CNN是最经典的两阶段检测算法，由Fast R-CNN[35]演化而来。在Fast R-CNN中，检测速度较R-CNN有了显著提升，但如果算法“选择搜索”算法生成区域候选框的时间，速度提升并不是特别明显，且“选择搜索”算法是在CPU上完成的，无法充分利用GPU的性能。因此，Faster R-CNN提出RPN（Region Proposal Network），代替了费时的“选择搜索”算法，利用共享的卷积网络来生成区域候选框，大幅提升了目标检测的速度，而且这个过程是在GPU上完成的。Mask R-CNN则针对Faster R-CNN中RoI池化进行了修改，使得区域候选框不会因池化时的量化操作而产生一部分信息的丢失。RoI池化操作之前，需要将候选边界框缩放至当前特征图上，在这里会进行第一次量化，造成部分信息的丢失。进行池化时，需要将候选区域等分成$K\times K$（如3×3）的区域，这里需要进行第二次量化，特征信息将会再次丢失。针对此缺陷，Mask R-CNN提出的RoI Align通过双线性保留了必要信息。为了更进一步提升检测的精度，Cascade R-CNN通过级联多个检测网络达到优化的结果，其中不同检测网络设置不同的阈值以确定正负样本。

虽然两阶段算法精度较高，但检测速度较慢，难以满足实时目标检测的需求，消耗的时间成本较高。相对之下，单阶段检测器速度更快，受到工业界的青睐。

2.2 单阶段检测器

单阶段检测器以SSD和YOLOv3为代表。SSD网络性能强、检测速度快，可以做到实时检测。基于SSD的变体也非常多，如DSSD[36]。SSD在网络结构上采用不同尺度的特征图进行检测，这里的多尺度特征均取自不同卷积层输出的特征，可以提供高级的语义信息和足够的分辨率信息。SSD与两

阶段目标检测器如Faster R-CNN等最大的不同在于取消了RPN的结构，直接进行类别的预测和位置的回归，无需先生成区域候选框。然而SSD对小目标的检测效果较差，YOLOv3则通过多尺度特征图融合改善这类问题。此外，YOLOv3设计了一套高性能基础网络Darknet53网络，借鉴了ResNet的残差结构，但速度却远超ResNet。

单阶段检测器虽然速度较快，但精度上往往不如两阶段检测器。

2.3 基于关键点的检测器

基于关键点的检测算法在严格意义上是属于单阶段目标检测器，但由于其检测方法的特殊性，因此单独归为一类。两阶段检测器在RPN阶段产生过多的先验框，容易造成正负样本不平衡。CornetNet摒弃了以往的先验框思路，提出通过预测物体边界框的左上角和右下角的两个顶点来确定物体的位置。ExtremNet同样是基于关键点的检测算法，基础网络是2个Hourglass网络，经过特征提取后生成4个极值点（上、下、左、右）热点图和1个中心点热点图。此外生成个位置偏移图，分别对应4个极值点的位置偏移，用于纠正预测的极值点位置。

虽然CornerNet和ExtremeNet在COCO数据集上精度非常高，但其局限性在于检测速度慢，甚至低于两阶段检测器。

2.4 Anchor-free检测器

Anchor-free检测器出现较早，但时至今日算法才得到再次发展。YOLO是第一个Anchor-free检测器，检测物体的过程中并未涉及到anchor。YOLO将图像分成7×7的网格，在每个网格内预测2个边界框，每个边界框的信息包括其中心点偏离所在格网点的距离及其宽高，预测流程非常快速，然而召回率非常低。2019年的FCOS以每个像素预测的方式进行目标检测，直接在特征图上预测空间点到物体框4个边界的距离。该算法同样不需要anchor，完全避免了与anchor相关的复杂计算。FoveaBox则是通过建模一个从特征图到物体边界坐标值的映射获取目标的位置，思路上与YOLO类似，同样也摒弃了anchor机制。

3 方法设计

3.1 AF-TSD网络结构

AF-TSD是一个端对端的交通标志检测网络，网络结构如图1所示。基础网络采用带有可变形卷积的VGG16_BN，并引入特征金字塔结构。在特征金字塔模块，本文巧妙设置特征尺度筛选，将不同尺度的交通标志分离至不同的特征层，用于检测特定尺度的交通标志。并于特征融合后，利用注意力机制抑制次要信息，增强实际需要的特征。

算法流程如下：

（1）输入街景图像，在维持图像宽高比不变的前提下将图像的宽缩放至608，并用黑色像素将缩放后图像的高填充至608，构成608像元×608像元的输入图像，如图2所示。

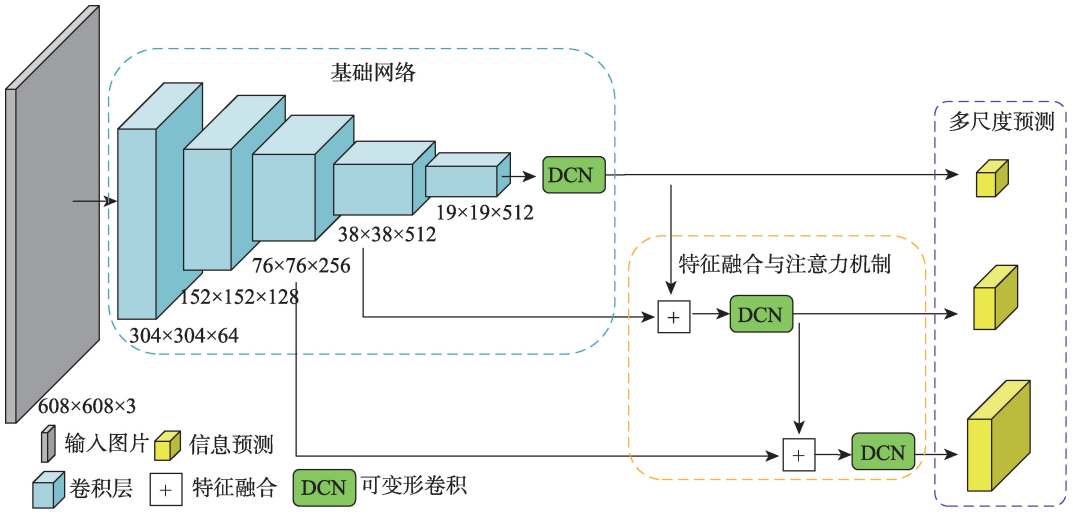


图1　AF-TSD网络结构

Fig. 1　AF-TSD network structure

原图（1360像元×800像元） 输入图像（608像元×608像元）

图2 输入图像预处理

Fig. 2 Pre-processing of the input image

（2）通过基础网络对输入图像进行特征提取，并于基础网络及特征融合模块中引入可变形卷积。

（3）在基础网络表达的高阶特征图上进行Anchor-free交通标志检测，结果记为S1。

（4）将第（3）步中的高阶特征图与基础网络提取特征过程中的低阶特征图进行融合，并利用注意力机制对融合后的次要特征进行抑制。在输出的特征图上进行Anchor-free交通标志检测，结果记为S2。

（5）将第（4）步中产生的特征图与基础网络提取特征过程中更为低阶的特征图进行融合，同样利用注意力机制对融合后的次要特征进行抑制。在输出的特征图上进行Anchor-free交通标志检测，结果记为S3。

（6）将S1、S2与S3尺度的预测结果进行融合，并执行非极大值机制去除冗余的检测框，输出最终交通标志的位置。

3.2 可变形卷积

传统的卷积方式是卷积核与图像上对应的区域进行乘积运算，其表达式为：

$$y(p)=\sum_{k=1}^{K}w_k\cdot x(p+p_k) \tag{1}$$

式中：x和y分别表示输入特征图和输出特征图；p为特征图上像素的位置；K为卷积核的元素数量；w_k表示第k个位置处的权重；p_k则表示预定义的采样偏移。假设K=9，则$p_k\in\{(-1,-1),(-1,0),\cdots,(1,1)\}$。

传统的卷积方式只能提取规则区域内的特征，对于尺度和形态变化较大的物体则具有局限性，容易受背景信息所影响。如一个展开双臂的舞者和一个站着军姿的士兵在形态上差异较大，展开双臂的舞者所拥有的边界框会包含更多的背景信息，在很大程度上会影响网络提取前景的特征。对于交通标志的形状规则，通常有三角形、圆形、多边形（如八边形的“STOP”标志）。对于三角形的交通标志，规则的卷积核所采样的特征中更易融入无关的背景信息，不利于前景的判断与边界框的回归。此外在同一层特征图中，不同尺度的物体共享相同大小的卷积核并不符合规则。如街景图像中距离拍摄位置较近的交通标志其尺度较大，距离拍摄位置较远的交通标志其尺度较小。卷积核在采样特征时应考虑到尺度上的差异，对于尺度较小的交通标志，卷积核采样范围应当适应性地缩小。

针对传统卷积方式的局限性，并为了更好的表达街景图像中交通标志的特征，本文引入自适应采样位置可变形卷积，其表达式为：

$$y(p)=\sum_{k=1}^{K}w_k\cdot x(p+p_k+\Delta p_k)\cdot\Delta m_k \tag{2}$$

式中：x和y分别表示输入特征图和输出特征图；p为特征图上像素的位置；K为卷积核的元素数量；w_k表示第k个位置处的权重；p_k则表示预定义的采样偏移；Δp_k为可学习的偏移量；Δm_k为可学习的缩放因子，取值范围$\in[0,1]$。假设K=9，则$p_k\in\{(-1,-1),(-1,0),\cdots,(1,1)\}$。

Δp_k与Δm_k均通过额外的卷积操作获得，输出通道为$3K$，其中$2K$为卷积核的坐标偏移，$1K$为缩放因子。由于卷积核发生了坐标偏移，其采样位置由整型转为浮点型，因此需要对浮点数位置上的特

征做双线性插值。可变形卷积的实现流程如图3所示，具体过程如下：

（1）生成offset特征图。假设输入特征图的维度为 $H\times W\times C$ ，其中 H 和 W 分别表示特征图的高和宽，C 表示特征图通道数。利用27个 3×3 的卷积核对输入特征图进行卷积操作可以得到维度为 $H\times W\times27$ 的offset特征图，该特征图每3个通道负责记录 3×3 卷积核其中一个元素的采样权重和采样偏移量 Δx 与 Δy 。

（2）可变形卷积。由第（1）步可得到卷积核进行采样时的坐标偏移量与采样权重，因此利用这些信息对输入特征图进行特征提取的卷积核称为可变形卷积核。如图3所示，可变形卷积核对输入特征图左上角区域进行采样，采样过程中从offset特征图同名区域上获取位置偏移量与采样权重，即可变形卷积核1号位置的元素需要从offset特征图1号位置上的同一颜色特征图（粉色，右数第1个）处获取偏移量与权重；可变形卷积核7号位置的元素需要从offset特征图7号位置上的同一颜色特征图（金黄色，右数第7个）处获取偏移量与权重。利用可变形卷积的方式依次对输入特征图所有区域进行采样，最终得到输出特征图。

可变形卷积在一定上程度上可以解决传统卷积存在的问题，通过学习的方式改变采样位置，从而提取感兴趣的前景特征以适应物体的形变，可以提高交通标志的检测精度。

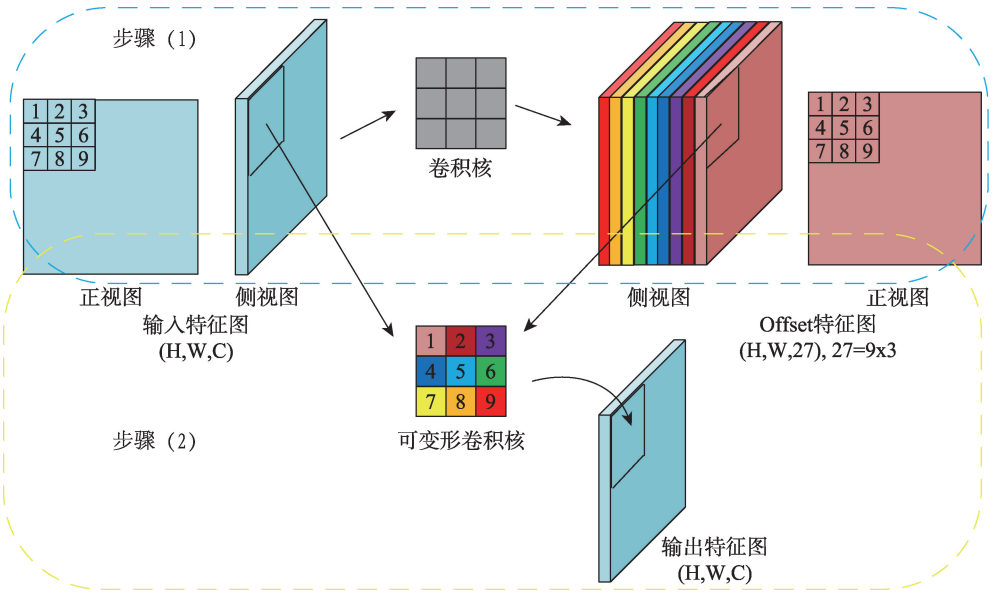
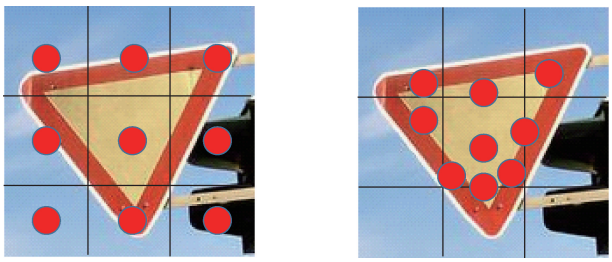


图3　可变形卷积实现流程

Fig. 3　Pipeline of deformable convolution

如图4所示，假设交通标志的区域图像已被 3×3 的中间特征图所表达，接着用 3×3 的卷积核对中间特征图进行采样，其中红色点表示卷积核的采样位置。传统的卷积将无法避免地采样背景干扰信息，而自适应采样位置可变的卷积则会根据提取到的特征自适应地调整采样位置。

除了得到有效的采样位置，在采样的过程中，同一片感受野区域下的特征的重要程度也并非是等同的。有些特征也许并不重要，对于无关紧要的特征，甚至可以为0。对此，可变形卷积设置了可自适应学习的缩放因子用于调节特征的权重，在一定程度上提升了网络的特征提取能力。



(a)传统卷积　　(b)可变形卷积

图4　传统卷积与可变形卷积

Fig. 4　Traditional convolution and deformable convolution

3.3 注意力机制

卷积核作为卷积神经网络的核心，在局部感受野上将空间上的信息和特征通道上的信息进行信息聚合。许多研究从空间维度层面提升网络的性能，如Inception[37-40]结构采用多路径的卷积方式聚合多种不同感受野上的特征，以此获得性能增益。然而旷视科技的学者则认为，高性能的网络应降低碎片化程度[41]。基于前人的研究，本文从特征通道的层面，利用Momenta提出的注意力机制结构来提高网络对交通标志的特征表达能力。

注意力机制结构如图5所示。给定输入特征图，进行特征压缩。首先根据空间维度进行特征压缩，将二维的特征通道压缩为1×1的特征单元。该特征单元具有全局的感受野，表达特征通道上响应的全局分布。其次通过通道压缩与通道扩大的方式为每个特征通道生成权重，通过这两个步骤可以显式地建模特征通道间的相关性，并且大大减少了建模的参数量。最后对输入特征图进行加权。通过网络学习得到的通道权重用于表示输入特征图中不同特征通道的重要性，完成在通道维度上对原始特征的重标定，可以有效抑制冗余信息，并相对地增益积极的特征。

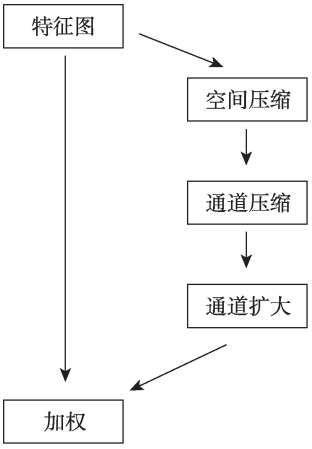


图5 注意力机制实现流程

Fig. 5 Pipeline of the attention mechanism

注意力机制在一定程度上有利于交通标志的特征表达，对于从复杂街景图像中检测交通标志，其检测精度得到一定的提升。

3.4 Anchor-free算法

目标检测算法发展已久，进入深度学习时代后，主流的目标检测算法如Faster R-CNN、SSD、YOLOv3等都是基于anchor boxes的算法。Anchor-based算法从图像中设置大量的尺度、宽高比不同的先验框，以求尽可能地覆盖感兴趣的目标，从而辅助预测潜在物体的类别和位置。FAIR研究的R-CNN系列网络通常采用2个阶段检测目标，首先通过RPN网络获得潜在的区域候选框。该边界框在类别上只有前景和背景之分，回归的位置也并不是非常准确，因此需要进行二次分类和位置的回归。单阶段检测网络SSD取消了RPN的结构，直接进行类别的预测和位置的回归，无需生成区域候选框。预测阶段也由固定尺度的特征度改为多尺度特征图预测。颇受工业界青睐的YOLOv3检测算法则加入了多尺度特征融合，对富有高级语义信息的高阶特征图与富有分辨率信息的低阶特征图进行了信息拼接。然而Anchor-based算法需要面临超参数最优设计、无法在不同应用场景之间切换等问题，并且冗余现象严重，容易造成样本不平衡。

本文受Anchor-free思想逐渐兴起的趋势启发，引用YOLO关于无先验框的检测方法，并设计出基于Anchor-free的实时交通标志检测网络AF-TSD。在AF-TSD网络中，街景图像经过带有自适应采样位置可变卷积的VGG16_BN的特征表达后，形成特征金字塔。本文对特征金字塔中不同尺度的特征图进行特征融合，并利用注意力机制抑制冗余信息、增益积极信息，最终在3种不同尺度的特征图上检测潜在的交通标志。

假设街景图像在输入层被缩放为608像元×608像元，则本文分别在1/8、1/16、1/32特征图处进行交通标志检测。交通标志的中心在不同特征图中会落入不同的格网点中，则该格网点负责预测该交通标志。对于真值的设置，Anchor-free不再以anchor boxes为基准，而是以图像的宽高为参考。关于交通标志检测的真值包括t_{obj}、t_x、t_y、t_w与t_h，其表达式如下：

$$t_x = \frac{x}{S} - \left\lfloor \frac{x}{S} \right\rfloor \quad (3)$$

$$t_y = \frac{y}{S} - \left\lfloor \frac{y}{S} \right\rfloor \quad (4)$$

$$t_w = \frac{w}{img_w} \quad (5)$$

$$t_h = \frac{h}{img_h} \quad (6)$$

$$t_{obj} = \begin{cases} 1 & \text{有交通标志} \quad (7) \\ 0 & \text{其他} \quad (8) \end{cases}$$

式中：x、y、w、h分别表示真实交通标志的中心坐标及宽高；S表示网络下采样的步长；img_w和img_h分

别表示图像的宽和高。

此外在进行多尺度预测前，本文巧妙设置特征尺度选择，将不同尺度的交通标志分离至不同的特征图。特征尺度选择在一定程度上减少了网络的计算量，并采用分级训练的方式获得较好的收敛效果。特征尺度选择的表达式如下：

$$\begin{cases} w \leqslant (S\times 3) \text{ 或 } h \leqslant (S\times 3) & S=8 \\ S \leqslant w \leqslant (S\times 4) \text{ 和 } S \leqslant h \leqslant (S\times 4) & S=16 \\ w \geqslant \left(3\times\dfrac{S}{2}\right) \text{ 和 } h \geqslant \left(3\times\dfrac{S}{2}\right) & S=32 \end{cases} \tag{9}$$

式中：w和h分别表示交通标志的宽和高；S表示网络的下采样步长。

4　实验与分析

4.1　数据集

本文选择的数据集为德国公开的交通标志检测数据集GTSDB[33]，如图6所示。GTSDB检测数据集中总共包含900张街景图片，总计1213个交通标志，由车载相机在自然场景下拍摄得到，图片的分辨率为1360像元×800像元。训练集包含600张图片，测试集包含300张图片。GTSDB数据集采集场景包括城市、高速公路和郊区，天气状况跨度非常全面，有艳阳高照的晴天，也有昏暗的阴天。时间上跨越了白天与傍晚，拍摄环境也跨越了正常光线下的拍摄与逆光拍摄。

以图像最短边分辨率800像元作为标准，本文对600张训练数据中的交通标志做K均值聚类（令K=9），得到9个聚类中心，分别为{(20,20)，(24,24)，(28,28)，(33,33)，(40,39)，(47,47)，(60,58)，(77,74)，(109,106)}。聚类中心的表示形式为：(交通标志的宽，交通标志的高)。根据聚类结果，发现街景图片中交通标志的尺度集中为小尺度，以20个像素至47个像素为聚类中心的交通标志占据聚类中心的2/3，这反应出街景图片中约有2/3交通标志可视为小目标。9个聚类中心中，仅有一个聚类中心大于100个像素，说明大尺度的交通标志在此数据集中出现的概率非常小。

图6　GTSDB数据集

Fig. 6　GTSDB dataset

4.2　评价指标

交通标志检测在计算机视觉上属于目标检测任务。对于目标检测，其综合精度评价指标为mAP（mean Average Precision）。mAP是所有类别平均精度的均值，其中每一类的平均精度为AP（图7）。AP

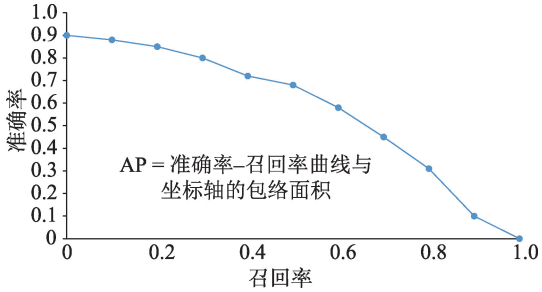


图7　交通标志检测准确率相对于召回率的变化曲线

Fig. 7　Curve of precision relative to recall rate in traffic sign detection

的表达形式如下：

$$AP=\int P(R)\cdot \mathrm{d}R \tag{10}$$

式中：R表示召回率；P表示准确率。AP的数学含义为P-R曲线与坐标轴包围得到的面积。对于本论文中的交通标志检测，由于只涉及一个类别，因此AP即mAP。

在P-R曲线中，准确率的含义是预测的交通标志中，预测正确的交通标志数量占总预测数量的比值，因此准确率也叫查准率。召回率的含义是预测正确的交通标志数量占交通标志总数量的比值，因此召回率也叫查全率。对于目标检测任务，召回率越高，往往准确率则越低，二者呈负相关。

4.3 实验结果与分析

本实验所采用的设备配置为单张英伟达GTX 1080Ti 11G显卡，数据集为GTSDB，其中训练集600张，测试集300张。

在训练过程中本文使用如下数据增强手段：

（1）镜像变换：将街景照片沿竖直中心线随机左右翻转。

（2）仿射变换：包括旋转（±5°），平移（±10%，水平和竖直方向），缩放（±10%），错切（±2°，水平和竖直方向）。

（3）颜色空间变换：将照片的颜色空间由RGB转换至HSV。

训练初始阶段采用warm-up策略，即训练一开始，以一个极小的学习率开始学习。随着迭代次数的增加逐渐升高到初始学习率0.001，从第3个epoch开始，利用余弦学习率衰减不断调整学习率的大小。

图8为训练时损失值随迭代次数（前150次迭代）的变化曲线，从图中可以看出，交通标志检测的任务收敛速度较快。经过前100次迭代，训练损失与测试损失便已经下降到非常低的位置。

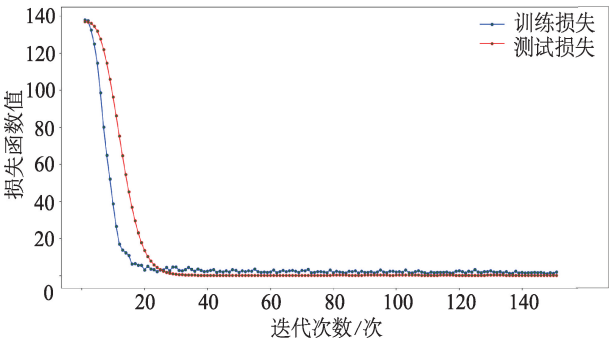


图8 AF-TSD网络训练损失与测试损失的变化曲线（前150次迭代）

Fig. 8 Curve of training loss and test loss in AF-TSD network (the first 150 iterations)

本文将街景图片缩放至固定尺度608像元×608像元作为网络的输入，总类别只有1类，即交通标志。由于交通标志为小目标，因此本文在开展的所有对比实验中规定将IoU（Intersection of Union）阈值统一设置为0.3。除此以外，规定设置的置信度阈值需大于等于0.5。经过300个epoch的迭代训练，本文提出的AF-TSD网络最终在测试集300张街景图片上的mAP达到96.8%，网络预测时间平均单张街景图片仅需32 ms，达到实时检测的标准。

本文针对网络结构上的设计，开展了必要的对比实验验证网络设计上的优越性，实验结果如表1所示。

表1 AF-TSD网络结构设计对比

Tab 1 Comparison of network structure designs in AF-TSD

基础网络	输入图像尺寸	mAP/%
VGG	608×608	95.40
VGG-DCN	608×608	96.29
VGG-DCN-Attention	608×608	96.80
VGG-Attention	608×608	96.02

根据表1，基础网络为VGG16_BN时，AF-TSD在测试集上的mAP为95.40%。当引入了自适应采样位置可变卷积（DCN），mAP提升了0.89%。自适应采样位置可变卷积通过学习的方式改变采样位置，从而适应性地提取感兴趣的前景特征以适应物体的形变。除了更改采样位置，DCN设置了可自适应学习的缩放因子用于调节特征的权重。通过实验表明，DCN在一定程度上可以提升交通标志的检测精度。

注意力机制的引入同样在测试集上精度提升了0.62%。本文将注意力机制应用于特征融合之后，通过在空间维度上进行特征压缩获得特征的全局分布，并通过为每个特征通道生成权重，完成在通道维度上对原始特征的重标定，可以有效抑制冗余信息，并相对增益积极的交通标志信息。通过实验证明，注意力机制对于从复杂街景图像中检测交通标志具有一定程度的提升。

本文将自适应采样位置可变卷积和注意力机制加入到AF-TSD中，检测精度提升了1.4%，大大提高网络的特征表达能力，最终在测试集上mAP达到96.80%。

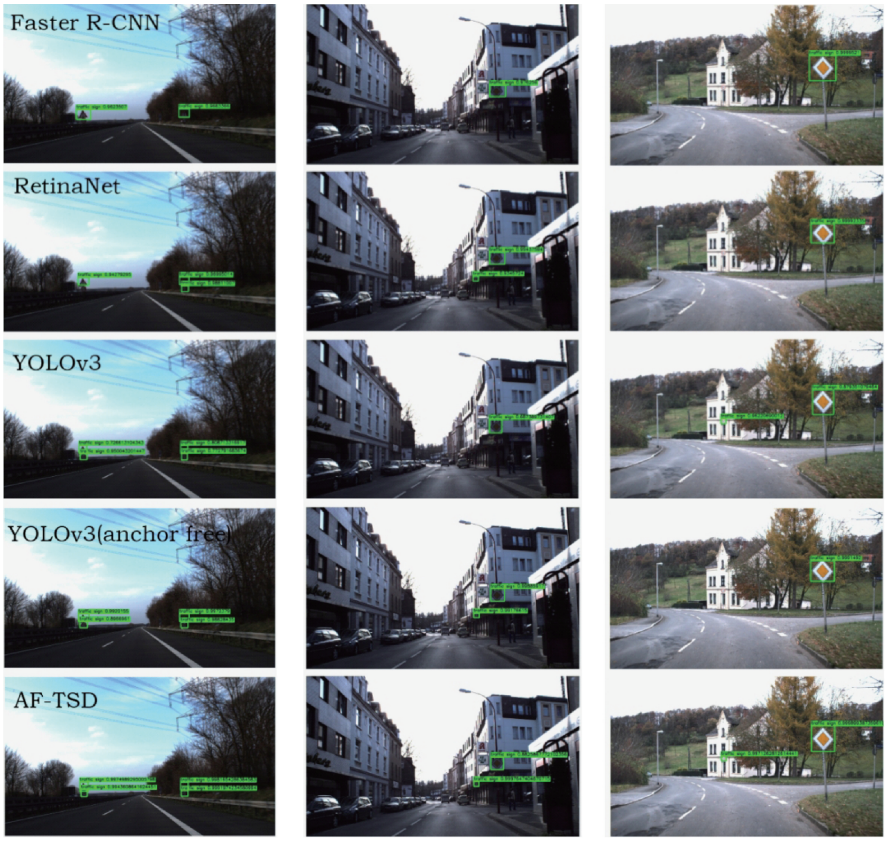
除了验证设计细节上的提升，本文同样开展实验横向对比了其他网络，如Faster R-CNN, RetinaNet和YOLOv3，实验结果如表2所示。根据表2，Faster R-CNN在测试集上mAP达到88.50%，检测速度为120 ms/img。Faster R-CNN首先通过RPN网络获取区域候选框，然后分别连接用于分类和回归的子网络得到交通标志的位置。通过对比实验可以发现，利用两个阶段检测交通标志导致其检测速度远远慢于其他网络。RetinaNet针对正负样本不平衡引入了focal loss，同时针对anchor boxes进行了超参数的优化，在GTSDB测试集上mAP达到了92.43%，然而检测速度依然较慢。YOLOv3是一个高性能单阶段检测网络，检测速度非常快，达到了24 ms/img。

表2　AF-TSD与Faster R-CNN、RetinaNet、YOLOv3及YOLOv3（Anchor-free）之间的性能对比

Tab. 2　Performance comparison of AF-TSD with Faster R-CNN, RetinaNet, YOLOv3, and YOLOve（Anchor-free）

方法	输入图像尺寸 像元×像元	mAP/%	s/每张图
Faster R-CNN	608×608	88.50	0.120
RetinaNet	608×608	92.43	0.094
YOLOv3	608×608	93.54	0.024
YOLOv3(Anchor-free)	608×608	94.92	0.026
AF-TSD	608×608	96.80	0.032

本文在实验中对YOLOv3做了改进，将原本基于anchor的检测算法进行了Anchor-free化。YOLOv3在特征图上生成3个尺度、宽高比不同的先验框，在思想上集成了滑动窗口的理念（图9）。与滑动窗口探测物体的不同之处在于，YOLOv3在感受野较大的特征图上进行检测，相比于滑动窗口大大减少了先验框的冗余与计算量。然而基于先验框的算法仍然不可避免产生了较多的冗余框，从而造成训练正负样本不平衡，同时研究者需要面临如何最优设计额外超参数的问题。本文在YOLOv3中利用Anchor-free直接预测交通标志的中心坐标及其宽高，不仅免除额外超参数的设计，并且在训练过程中不会产生过量



注：从上至下依次为Faster R-CNN, RetinaNet, YOLOv3, YOLOv3（Anchor-free）, AF-TSD。

图9　交通标志检测结果

Fig. 9　Traffic signs detection results

的冗余框，同时还将减少网络的计算量。实验结果表明，Anchor-free YOLOv3较Anchor-based YOLOv3在测试集上mAP提升了1.38%，并且检测速度仅仅慢了2 ms。这表明，Anchor-free算法可以有效提升街景图像中交通标志的检测精度。

本文将AF-TSD与上述网络做对比，在测试集上取得96.80%的mAP，检测速度为32 ms/img，依然达到实时检测的范围。实验结果再次证明，本文提出的基于Anchor-free的交通标志检测网络AF-TSD在交通标志检测任务上是可行和可靠的。

图9为表2中各个算法在GTSDB测试集上的部分检测结果，其中当道路场景中存在多个交通标志时（图9左列），Faster R-CNN与RetinaNet存在较为严重的漏检测情况，YOLOv3、YOLOv3（Anchor-free）与AF-TSD的召回率则更高。而当场景中存在较小目标或目标距离过远时（图9中间列与右列），YOLOv3与YOLOv3（Anchor-free）发生不同程度的漏检测，YOLOv3在光线偏暗的城市道路路口未能检测到远处的交通标志，而YOLOv3（Anchor-free）在光线明亮的郊区道路上漏检测了远处的目标。相比之下，本文提出的AF-TSD网络能更好地适应这些场景。

检测结果表明，在光线足够且非夜晚的道路环境下，提出的AF-TSD网络在街景图像上检测交通标志具有较好的表现，并且边界框回归的位置较为准确。

5 结论与讨论

本文受到Anchor-free思想的启发，引用YOLO直接回归物体边界框的思路，提出一种基于Anchor-free的实时交通标志检测网络AF-TSD。本文设计的网络为全卷积网络，有效适应不同尺度的图像输入。网络结构引入自适应采样位置可变卷积与注意力机制，检测精度在原有基础上提升1.4%，大大提高网络对交通标志的特征表达能力。

本文开展不同角度的实验，按照实验结果导向不断优化网络设计。除此之外，本文开展多个对比实验与主流检测网络进行对比。实验结果证明，本文设计的AF-TSD在街景图像交通标志检测上速度接近主流算法，但精度优于主流算法，在精度与速度上取得较优的平衡。

参考文献(References):

[1] De l E A, Moreno L E, Salichs M A, et al. Road traffic sign detection and classification[J]. IEEE Transactions on Industrial Electronics, 1997,44(6):848-859.

[2] Ellahyani A, El Ansari M, El Jaafari I, et al. Traffic sign detection and recognition using features combination and random forests[J]. International Journal of Advanced Computer Science and Applications, 2016,7(1):683-693.

[3] Miura J, Kanda T, Shirai Y. An active vision system for real-time traffic sign recognition[C]// Intelligent Transportation Systems, 2000. Proceedings, IEEE, 2000:52-57.

[4] 徐迪亮,唐炉亮.基于颜色和标志边缘特征的交通标志检测[J].武汉大学学报·信息科学版,2008,33(4):433-436. [Xu D H, Tang L L. A pyramid-based cracks statistical model fot massive pavement images[J]. Geomatics and Information Science of Wuhan University, 2008,33(4):433-436.]

[5] 张静,何明一,戴玉超,等.多特征融合的圆形交通标志检测[J].模式识别与人工智能,2011,24(2):226-232. [Zhang J, He M Y, Dai Y C, et al. Mutil-feature fusion based circular traffic sigh detection[J]. Patten Recognition and Artifitial Intelligence, 2011,24(2):226-232.]

[6] 贾永红,胡志雄,周明婷,等.自然场景下三角形交通标志的检测与识别[J].应用科学学报,2014,32(4):423-426. [Jia Y H, Hu Z X, Zhou M T, et al. Detection and recognition of triangular traffic signs in natural scenes[J]. Journal of Applied Sciences, 2014,32(4):423-426.]

[7] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[J]. Computer Vision and Pattern Recognition, 2001,1(511-518):3.

[8] Jiao J, Zhong Z, Park J, et al. A robust multi-class traffic sign detection and classification system using asymmetric and symmetric features[C]// IEEE International Conference on Systems, Man and Cybernetics. IEEE Press, 2009:3421-3427.

[9] Liu C, Chang F, Chen Z. Rapid multiclass traffic sign detection in high-resolution images[J]. IEEE Transactions on Intelligent Transportation Systems, 2014,15(6):2394-2403.

[10] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]// Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005:886-893.

[11] Xie Y, Liu L F, Li C H, et al. Unifying visual saliency with HOG feature learning for traffic sign detection[J]. Intelligent Vehicles Symposium IEEE, 2009:24-29.

[12] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic seg-

mentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 580-587.

[13] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems, 2015: 91-99.

[14] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[C]//Advances in neural information processing systems, 2016:379-387.

[15] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018:6154-6162.

[16] Huang L, Yang Y, Deng Y, et al. DenseBox: Unifying landmark localization with end to end object detection[J]. Computer Science, 2015(2):12-19.

[17] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016:21-37.

[18] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision, 2017:2980-2988.

[19] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014:740-755.

[20] Yang T, Zhang X, Li Z, et al. Metaanchor: Learning to detect objects with customized anchors[C]//Advances in Neural Information Processing Systems. 2018:320-330.

[21] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018:734-750.

[22] Zhou X, Zhuo J, Krahenbuhl P. Bottom-up object detection by grouping extreme and center points[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:850-859.

[23] Huang L, Yang Y, Deng Y, et al. Densebox: Unifying landmark localization with end to end object detection[J]. arXiv preprint arXiv:1509.04874, 2015.

[24] Yu J, Jiang Y, Wang Z, et al. Unitbox: An advanced object detection network[C]//Proceedings of the 24th ACM international conference on Multimedia. ACM, 2016: 516-520.

[25] Tian Z, Shen C, Chen H, et al. FCOS: Fully Convolutional One-Stage Object Detection[J]. arXiv preprint arXiv: 1904.01355, 2019.

[26] Kong T, Sun F, Liu H, et al. FoveaBox: Beyond anchor-based object detector[J]. arXiv preprint arXiv:1904.03797, 2019.

[27] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2016:779-788.

[28] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014(2):34-46.

[29] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2017:764-773.

[30] Zhu X, Hu H, Lin S, et al. Deformable convnets v2: More deformable, better results[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:9308-9316.

[31] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:2117-2125.

[32] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:7132-7141.

[33] http://benchmark.ini.rub.de/?section=gtsdb&subsection=news

[34] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017:2980-2988.

[35] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[36] Fu C Y, Liu W, Ranga A, et al. Dssd: Deconvolutional single shot detector[J]. arXiv preprint arXiv:1701.06659, 2017.

[37] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:1-9.

[38] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C]//International Conference on International Conference on Machine Learning, 2015:423-434.

[39] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:2818-2826.

[40] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[41] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018:116-131.