YOLOv3-A：基于注意力机制的交通标志检测网络

郭璠，张泳祥，唐瑾，李伟清
（中南大学自动化学院，湖南 长沙 410083）

摘 要：为了解决已有 YOLOv3 算法对于存在小目标问题和背景复杂问题的交通标志检测任务会有较多的误检和漏检的问题，在 YOLOv3 算法的基础上，提出了目标检测的通道注意力方法和基于语义分割引导的空间注意力方法，形成 YOLOv3-A 算法。YOLOv3-A 算法通过对检测分支特征在通道和空间 2 个维度进行重新标定，使网络聚焦和增强有效特征，并抑制干扰特征，提高了算法的检测能力。在 TT100K 交通标志数据集上的实验表明，所提算法对小目标检测性能的改善尤为明显，相比于 YOLOv3 算法，所提算法的精度和召回率分别提升了 1.9% 和 2.8%。

关键词：交通标志检测；小目标检测；注意力机制；语义分割
中图分类号：TP391.41
文献标识码：A
DOI: 10.11959/j.issn.1000−436x.2021031

YOLOv3-A: a traffic sign detection network based on attention mechanism

GUO Fan, ZHANG Yongxiang, TANG Jin, LI Weiqing
School of Automation, Central South University, Changsha 410083, China

Abstract: To solve the problem that the existing YOLOv3 algorithm had more false detections and missed detections for traffic sign detection task with small target problems and complex background, based on the YOLOv3, a channel attention method for target detection and a spatial attention method based on semantic segmentation guidance were proposed to form the YOLOv3-A (attention) algorithm. The detection features in the channel and spatial dimensions were recalibrated, allowing the network to focus and enhance the effective features, and suppress interference features, which greatly improved the detection performance. Experiments on the TT100K traffic sign data set show that the algorithm improves the detection performance of small targets, and the accuracy and recall rate of the YOLOv3 are improved by 1.9% and 2.8% respectively.

Keywords: traffic sign detection, small target detection, attention mechanism, semantic segmentation

1 引言

交通标志检测不仅能够为辅助驾驶系统提供有效的路况数据支持，而且在建立高精度地图方面可以避免烦琐易错的人工标注。因此，对交通标志检测系统进行深入研究，不仅在提高道路安全性方面具有很大的实用价值，而且能够对无人驾驶技术的发展起到推动性作用。传统的交通标志检测算法可以分为感兴趣区域（RoI, region of interest）提取和 RoI 分类 2 个阶段。在 RoI 提取阶段，通常使用不同尺度和比例的滑动窗口在整幅图像上扫描，以获得潜在的目标区域；在 RoI 分类阶段，常用 HOG

收稿日期：2020−07−16；**修回日期**：2020−09−28
基金项目：国家自然科学基金资助项目（No.61502537）；湖南省自然科学基金资助项目（No.2018JJ3681）；中南大学中央高校基本科研业务费专项基金资助项目（No.2020zzts567）
Foundation Items: The National Natural Science Foundation of China (No. 61502537), The Natural Science Foundation of Hunan Province (No.2018JJ3681), The Fundamental Research Funds for the Central Universities of Central South University (No.2020zzts567)

（histograms of oriented gradient）[1]、Gabor[2]、Haar-like[3]等人工设计特征，结合机器学习算法进行RoI的类别判断。由于存在光照、变形、遮挡等问题，传统方法在实际的交通标志检测任务中难以取得良好效果。

近年来，随着卷积神经网络在计算机视觉领域不断深入和发展，基于深度学习的交通标志检测算法也取得了很大提升。现有的检测方法可以分为两阶段方法和一阶段方法。以Faster R-CNN（region-convolutional neural network）[4]为代表的两阶段方法使用RPN（region proposal network）通过共享卷积特征的方式在特征层面生成建议框，再利用建议框区域的卷积特征进行分类和目标框的定位学习，具有精度高但速度慢的特点；以YOLO（you only look once）[5]、SSD（single shot detector）[6]为代表的一阶段目标检测方法将目标框的定位和识别任务统一按照回归的逻辑，由卷积神经网络在输出层一次性预测完成，具有速度快但精度低的特点。实时性是工业领域和实际应用场景中的关键指标，因此提高一阶段检测方法的精度更有实用价值。

目前，交通标志检测算法的主要改进方向有探索语义特征更抽象的基础网络、融合不同层级特征的特征融合方法和数据预处理方法等。Rajendran等[7]以RetinaNet[8]为基础，使用层数更深的ResNet[9]为基础网络，并在网络底层使用反卷积模块丰富特征的语义信息，最后在GTSDB（German traffic sign detection benchmark）交通数据集[10]上获得96.7%mAP的效果，这种方法会引入大量的额外参数。Yang等[11]使用多尺度的全卷积网络DMS-Net（dual multi-scale network）来检测不同尺度的交通标志，并引入在线困难样本挖掘（OHEM, online hard example mining）策略，最终在STSD（Swedish traffic signs dataset）数据集[12]上获得99.88%的准确率和96.61%的召回率。Meng等[13]在图像金字塔的基础上，将每幅图像划分为200像素×200像素的小图，送入SSD网络进行目标检测，训练得到一个对小目标敏感的SOS（small object sensitive）网络，在测试时同样需要进行图像金字塔和子图划分操作，降低了算法的实时性。上述工作从不同角度提升了交通标志检测算法的性能。

在实际的交通标志检测场景中，图像背景复杂多样，存在各种广告牌、干扰物体和其他提示标志，这些伪交通标志在外形和颜色上很容易与真实的交通标志形成混淆，容易导致误检。此外，为了提前获得道路信息，车载相机拍到的交通标志一般像素绝对尺度较小，而且占据整幅图像的相对比例也十分小。交通标志绝对尺度小，包含的有效信息少、噪声多，在模糊不清的情况下，很容易出现误检和漏检；目标相对尺度较小，意味着图像中包含更多的背景区域，更容易出现误检。注意力机制在很多计算机视觉任务中被证明可以有效地提升网络性能，该方法模拟了人类大脑提取外部信息的过程，即人类视觉系统会在图像上的某些区域产生局部聚焦，通过对聚焦区域投入更多的注意力，获得有效的细节信息。注意力机制使人类在有限的视觉感知能力下，对海量的输入信息进行合理的抑制和增强，极大地提高了人类视觉系统的信息处理能力。

YOLOv3检测算法[14]并未对卷积特征进行聚焦处理，即同等对待特征图中的每个区域，认为每个区域对最终检测结果的贡献是相同的，而且特征金字塔网络[15]（FPN, feature pyramid network）在对不同层级特征融合时，直接进行拼接处理，会存在大量的冗余信息。因此，本文以YOLOv3检测算法为基础，提出了目标检测的通道注意力（CA, channel attention）方法和基于语义分割引导的空间注意力（MGSA, mask guided spatial attention）方法，形成了YOLOv3-A（attention）算法。YOLOv3-A算法对检测分支特征在通道和空间2个维度进行重新标定，能够聚焦网络和增强有效特征，抑制干扰特征，提高神经网络对小目标的注意能力，并且抑制背景中的干扰物体。本文的主要贡献如下。

1) 在FPN融合不同层级特征时，根据目标检测特征中含有大量干扰信息的特点，本文对SENet[16]的通道注意力机制进行了改进，使用全局最大池化和全局平均池化对特征在空间维度进行压缩，并且进行维度拼接后，通过全连接网络学习每个通道的融合权重，使不同层级特征在融合时具有了区分度。

2) 本文将目标物体的标定框作为监督信息，在特征层面预测一个语义分割掩模，并将此掩模作为引导，与自带attention属性的深层卷积特征相结合，得到每个通道的空间注意力权重，对特征在空间维度进行重新标定，以精细化小目标特征，抑制背景特征，减少YOLOv3网络的漏检和误检情况。

2　YOLOv3-A 交通标志检测网络

2.1　YOLOv3-A 的检测分支结构

YOLOv3 网络在使用 FPN 方法融合不同层级特征时，将逐元素相加的特征融合方式改为在通道方向上的直接拼接，这样可以避免特征直接相加导致不同尺度特征相互影响的问题，更有利于网络对多尺度特征的利用。受此结构启发，所提网络将通道注意力机制和语义分割引导的空间注意力机制引入 YOLOv3 网络中的检测分支，形成 YOLOv3-A 网络，其结构如图 1 所示。其中，F 代表残差学习，以保证引入的注意力机制不会导致网络退化。该网络首先经过基础语义特征网络提取特征，在使用 FPN 特征金字塔方法对不同层级特征进行拼接时，引入通道注意力机制对多尺度特征进行通道间的重新标定，以达到增强有效通道特征、抑制冗余通道特征的目的。然后，经过特征融合模块（YOLOBlock）对通道注意力特征进行融合，接入语义分割引导的空间注意模块，有监督地对特征在空间维度上进行重新标定，以达到强化有效区域特征、抑制干扰区域特征的目的。最后，在得到的注意力特征上进行目标检测。由此可见，通道注意力机制和空间注意力机制是该检测分支结构的核心。

2.2　通道注意力机制

SENet 引入通道注意力机制，以绝对优势获得了 2017 年 ImageNet 竞赛中图像分类冠军。其核心思想是将特征在空间维度上压缩，去除空间位置影响，再经过全连接网络的学习和激活得到输入特征各通道的权重，完成对原始特征在通道维度上的重新标定。具体而言，首先，在空间维度上压缩特征，即将尺度为 $H\times W\times C$ 的输入特征经过平均池化，得到具有二维全局感受野的 $1\times1\times C$ 的压缩特征。然后，经过两层全连接网络对压缩特征进行编码和解码，再经过 sigmoid 函数激活后，输出与输入特征通道数一致的 $1\times1\times C$ 的注意力权重，用来反映不同通道的重要程度。最后，将通道权重与输入特征按通道相乘，得到重新校准后的通道注意力特征。SENet 是针对图像分类任务所设计的通道注意力方法，使用全局平均池化可以获得代表通道特征的响应情况。对于目标物体占据了特征图很大面积比例的图像分类任务来说，通道特征的平均值能够较好地代表该通道的响应情况。但是，对于目标检测任务来说，目标物体通常较小，在特征图上只能占据很小的区域。除此之外，目标检测的原始图像中通常包含了很多无关物体，这些物体虽然响应较小，但是数量较多，总的响应贡献依旧很大。因此，对目标检测特征在空间维度进行平均池化，并不能很好地代表网络对前景目标的响应，而每个通道的响应极值或许能够更好地反映该通道对前景目标的响应情况。

本文针对目标检测网络所改进的通道注意力模块结构如图 2 所示。该结构将 FPN 特征融合部分的多层级特征 F_1 作为待标定特征，首先对 F_1 特征在通道方向上分别进行全局最大池化和全局平均池化，并将池化结果在通道方向上进行拼接，得到 $1\times1\times2C$ 的压缩特征；然后将 $1\times1\times2C$ 的压缩特征

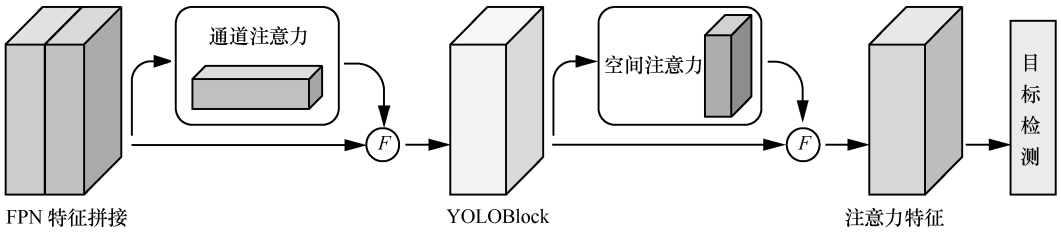


图 1　YOLOv3-A 网络结构

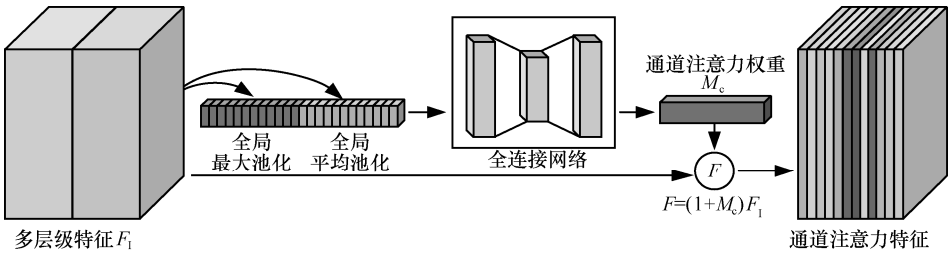


图 2　通道注意力模块结构

送入具有 3 个隐藏层的全连接网络进行特征的编解码，经过 sigmoid 激活函数得到通道注意力权重 M_c；最后将通道注意力权重 M_c 与多层级特征 F_I 进行残差连接并按通道相乘，即 $F=(1+M_c)F_I$，得到最终的通道注意力特征，以保证网络不会出现退化问题。

通道注意力模块中的全连接网络结构如图 3 所示。经过对多层级特征 F_I 在空间维度上的全局最大池化和全局平均池化得到尺度为 $1\times1\times2C$ 的特征向量，经过第一个全连接层进行特征融合和降维，把特征从 $2C$ 个通道降维到 C 个通道，并使用 ReLU 函数进行激活。第二个全连接层将 C 个通道压缩成 $\frac{C}{r}$ 个通道，进行全局特征的编码，达到降低计算量的目的，同样使用 ReLU 函数激活。最后一个全连接层将特征的通道数恢复为 C 个通道，并使用 sigmoid 函数激活，代表多层级特征 F_I 中的不同通道的重要程度。

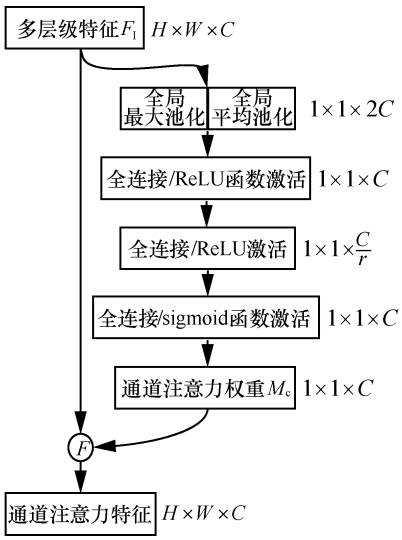


图 3 通道注意力模块全连接网络结构

综上所述，本文根据目标检测特征中含有大量干扰信息的特点，对 SENet 的通道注意力机制进行了改进，优化了 FPN 在进行不同层级特征拼接时存在的冗余通道问题，使不同层级的特征在融合时具有了区分度，抑制了对干扰信息响应较大的通道，保留了利于检测任务的有效信息。

2.3 基于语义分割引导的空间注意力机制

2.3.1 MGSA 算法结构

空间注意力机制不仅可以让网络聚焦于有效区域，而且能够对聚焦区域的特征进行改善和增强[17]。在图像分类和图像显著性检测任务中，深层特征的激活区域恰好对应目标物体最具有区分度的部分，这说明深层卷积神经网络自带 attention 效果。文献[17-19]使用深层卷积神经网络特征的 attention 属性，无监督地实现了对特征在空间维度上的聚焦和改善，并通过消融实验验证了这种弱监督注意力在图像分类和图像显著性检测任务中的有效性。在交通标志检测任务中，目标物体尺度一般较小，而且图像中会包含大量干扰物体，使深层特征的激活区域不能明显地反映目标物体的空间位置。因此，本文提出了基于语义分割引导的空间注意力机制，通过有监督的方式生成目标检测深层特征的语义分割 Mask，并将此 Mask 作为引导与自带 attention 属性的深层特征相结合，得到输入特征在空间位置上的注意力分布。

MGSA 算法结构如图 4 所示。输入图像首先经过特征提取和融合得到检测分支特征 F_d。然后检测分支特征 F_d 经过所设计的语义分割模块进行前景和背景类别的分割，其中语义分割的监督标签通过将输入图像中目标的标定框映射到特征图尺寸后得到。接着将语义分割结果 M 与深层特征 F_s 相结合得到空间注意力权重 S_w。最后将检测分支特征

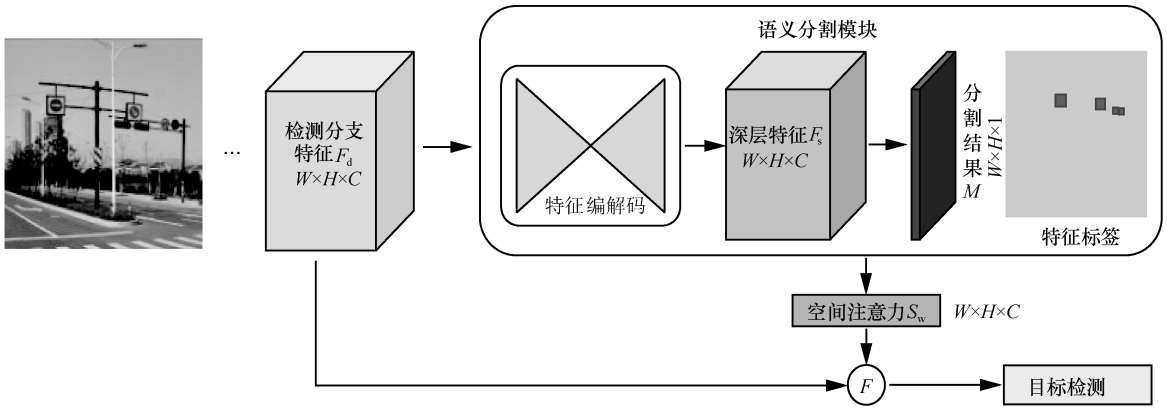


图 4 MGSA 算法结构

F_d 与空间注意力权重 S_w 通过残差注意力的方法进行结合，即 $F=(1+S_w)F_d$，得到聚焦和改善的特征后，进行后续的目标检测过程。

2.3.2　语义分割模块结构

图像语义分割任务是对输入图像的所有像素进行分类，将同类物体上的像素归为一类，因此图像语义分割任务是从像素的角度去理解图像。基于深度学习的图像语义分割方法通常可看作编码−解码模型。以 FCN（fully convolutional network）[20] 语义分割模型为例，编码过程通过若干卷积层的堆叠和池化得到大感受野且低分辨率的编码特征。解码过程是将编码特征进行反卷积上采样，得到高分辨率的特征，并预测图像中每个像素所属的类别。本文所提出的语义分割模块作为特征空间注意力的引导，是对每个特征的空间位置进行前景和背景的预测，不需要对编码特征上采样到原始图像的尺度。

Mask R-CNN[21]在同一个网络结构的不同分支中实现了目标检测任务和语义分割任务。Mask R-CNN 以 Faster R-CNN 目标检测方法为基础，通过对 RPN 提取的目标侯选区域中的特征进行 RoI Align 池化后，多分支地进行目标边界框的回归和分类学习，并且增添语义分割分支，对共享的池化特征进行编解码，同时预测图像中的物体掩膜。Mask R-CNN 不仅证明了在一个网络中进行语义分割和目标检测的多任务联合学习可以达到相互促进的效果，而且证明了深层卷积特征中含有丰富的语义信息，不同的任务可以通过共享深层卷积特征的方式，利用不同的训练标签和损失函数学习不同的分类器。本文提出的基于语义分割引导的空间注意力模块同样通过共享 YOLOv3 中检测分支的深层特征，完成特征级别的语义分割任务，并且在训练阶段通过多任务联合训练的方式提升目标检测的性能。

鉴于交通标志都为矩形、圆形或三角形等规则形状，而且是对低分辨率的特征图进行语义分割，因此在建立特征的语义分割标签时，本文利用 TT100K 数据集中目标框的标签在原始图像上进行像素的前景和背景类别划分，然后将标记过前景和背景的像素 Mask 按照相应的尺度映射到不同层级的特征图上，得到特征语义分割 Mask 标签。此过程如图 5 所示，首先根据数据集中的标签，判断图像上的像素点是否落在任意标定框中。如果某像素点落在了标定框中，那么该像素点就标记为前景类；如果某像素点未落在任意的标定框中，那么该像素点标记为背景类，得到像素级语义分割 Mask。最后根据不同检测分支上特征图 stride 的倍数，对像素级 Mask 进行尺度映射，得到特征级别上的语义分割标签。

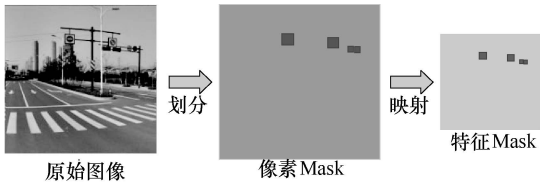


图 5　特征 Mask 标签生成过程

语义分割任务需要结合像素点的上下文信息完成对像素的类别判断，需要具有较大感受野的非局部特征。本文借鉴 Inception[22]多分支网络的思想，设计的检测特征编解码模块结构如图 6 所示。检测分支特征经过两部分完成特征的编解码：一部分是使用能够快速获得全局信息的 MaxPool 操作，以保留特征图中局部响应最大的部分，然后使用反卷积操作将低分率的池化特征恢复到高分辨率，即采用 Down-Up 的方式提取特征的非局部信息；另一部分是为了避免 MaxPool 特征在 Down-Up 的过程中出现激活信息偏移，引入了 DeepLabv3[23]中的 ASPP（atrous spatial pyramid pooling）模块。利用不同膨胀率 rate 的空洞卷积在不增加过多参数量的前提下，不断扩大特征的感受野，以获得精确的非局部特征。最后将这两部分非局部特征进行拼接融合，完成检测特征编解码。

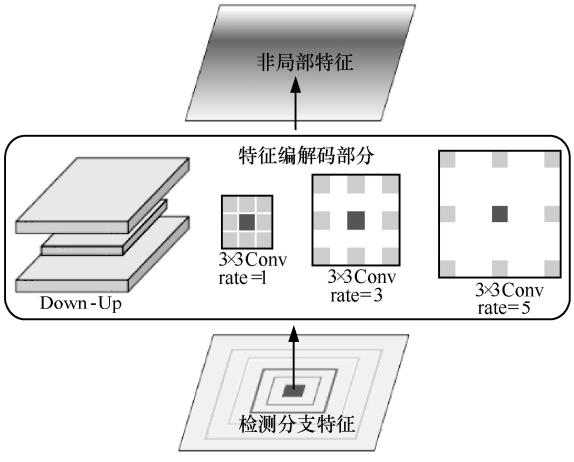


图 6　检测特征编解码模块结构

检测特征编解码模块的网络结构如图 7 所示。

Down-Up 部分首先进行两次 stride=2，卷积核 kernel 尺寸为 3×3 的最大池化，完成特征的下采样，然后使用两次卷积核 kernel 为 3×3、stride=2、padding=1 的转置卷积，将池化特征恢复到原始特征尺度。同时，所提方法还对 DeepLabv3 中的 ASPP 部分进行了改进。具体如下：首先使用 3 个 1×1 的标准卷积对检测分支特征进行通道降维，以减少计算量；然后在其中 2 个分支上使用卷积核 kernel 尺寸为 3 和 5 的标准卷积获得不同的基础感受野特征；接着在 3 个分支中分别使用 rate=1、3、5 的膨胀卷积，进一步扩大特征的感受野，卷积步长 stride 均为 1，以保证卷积过程中特征图的尺度不会发生变化；最后将 Down-Up 部分特征和改进后的 ASPP 部分特征进行拼接，经过 1×1 卷积进行特征融合和通道降维后，得到检测分支特征的非局部特征。

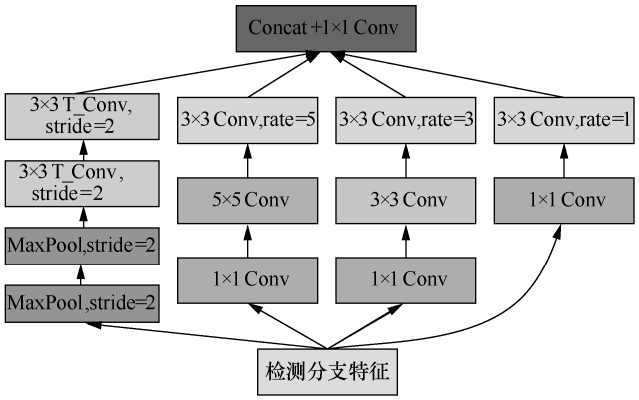


图 7 检测特征编解码模块的网络结构

在训练阶段，由于本文研究任务只有前景和背景 2 个类别，因此对网络预测的语义分割结果 M 与真实标签 M^* 使用二分类交叉熵计算损失如下

$$L_{\text{mask}} = -\sum_i \sum_j M_{ij}^* \log(M_{ij}) + \alpha(1-M_{ij}^*)\log(1-M_{ij}) \tag{1}$$

其中，i、j 分别为特征图上的横、纵坐标，α 为平衡正负样本所使用的权重。此外，在分配语义分割标签时，根据 YOLOv3 的思想，让不同层级特征负责不同尺度物体的语义分割学习。

2.3.3 空间注意力权重的形成

为了充分利用卷积神经网络自带的 attention 属性，所提方法在输出语义分割预测结果的前一层，生成了一组与检测分支特征尺度一致的深度特征 F_s 作为空间注意力的基础。因此，所提基于语义分割引导的空间注意力机制的形成如图 8 所示。该方法首先将语义分割模块产生的 $W\times H\times C$ 尺度的特征 Mask，在通道维度上进行广播复制，得到尺度为 $W\times H\times(1\times C)$的扩展特征 Mask。然后，将编、解码网络生成的深度特征 F_s 与扩展特征 Mask 按元素进行相加融合。接着，对融合后的特征使用 sigmoid 函数进行激活，将空间注意力权重的范围映射到 [0,1]，得到最终的空间注意力权重 S_w。最后，将空间注意力权重 S_w 与检测分支特征 F_d 通过残差注意力的方式进行结合，即 $F=(1+S_w)F_d$，完成对检测特征在空间位置上的聚焦和改善。

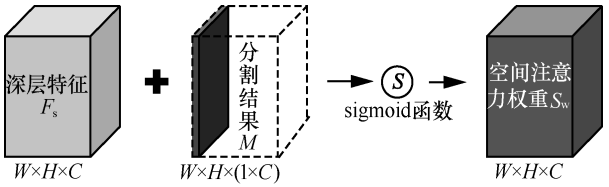


图 8 空间注意力机制的形成

2.4 网络输出结构与训练策略

2.4.1 YOLOv3-A 输出特征结构

YOLOv3-A 网络使用了基于注意力机制的 FPN 结构，以解决目标多尺度问题，并且改善 TT100K 数据集中的小目标问题和遮挡问题。该网络通过 3 个具有不同感受野的分支进行目标检测，因此在输出 head 部分共有 3 个尺度的特征图，这些特征图的长宽值相对于输入图像的下采样倍数分别为 32、16 和 8。特征图的尺度越小，其拥有的感受野就越大，因此小尺度的特征图分支用来检测大尺度物体，而大尺度特征图分支用来检测小尺度物体。YOLOv3-A 网络使用 K-means 算法对训练集中目标物体的尺度进行聚类，得到 9 个不同尺度和比例的 anchor 先验值。每个检测分支分配 3 个尺度相近的 anchor，即特征图上的每个单元格预设 3 个 anchor 框，因此输出特征图的维度为 $N\times N\times[3\times(4+1+C)]$，其中，$N\times N$ 为输出特征图的单元格数，每个 anchor 框需要预测 4 维边界框的中心点和长宽信息（x、y、w、h）、一维边界框的置信度 c 和 C 维类别概率 cls，YOLOv3-A 网络输出特征结构如图 9 所示。YOLOv3-A 网络模型的整体结构如图 10 所示，基础语义特征网络使用的是具有残差结构的 Darknet53。

由图 10 可知，特征融合的 neck 部分使用的是经过通道注意力方法和基于语义分割引导的空间注意力方法改进后的 FPN 结构，在输出 head 部分

有 P5、P4、P3 共 3 个层级分支，每个分支上输出特征图的长宽尺度分别为输入图像的$1/2^5$、$1/2^4$和$1/2^3$倍，特征图的通道数为 60×(3×5+45)，并且每个 MGSA 模块都有一个相应尺度的特征 Mask 预测输出。此外，在 FPN 特征融合部分，采用经过注意力机制的高层级特征，并且经过线性插值上采样 2 倍后与基础特征进行拼接融合。对于不需要融合其他层级特征的 P5 层级，只使用了 MGSA 模块进行空间注意力的改善。

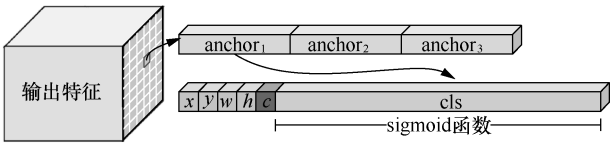


图 9　YOLOv3-A 网络输出特征结构

2.4.2　YOLOv3-A 训练策略

实验所采用的数据集中图像的分辨率为 2 048 像素×2 048 像素，若直接送入网络模型进行训练，很容易导致 GPU 内存不足。因此在训练数据预处理方面，本文以图像中每个目标的标定框为参考，随机生成 3 个 512 像素×512 像素的窗口，裁剪出含有目标的图像。同时按照 Selective Search 方法[24]在图像上裁剪出 2 个纹理和颜色丰富且只包含背景的图像，以丰富数据集中的背景样本。对训练数据进行裁剪后，随机地对图像在 HSV 颜色空间进行颜色变化处理，共得到 42 317 张 512 像素×512 像素的训练图像。通过从高分辨率图像上裁剪训练样本，不仅使交通标志位于图像上不同的位置，达到增加样本多样性的目的，而且在训练时可以适量增大 batch size，让每批训练数据都能够更好地代表样本分布。

在训练 YOLOv3-A 网络时，使用了多任务联合学习的方式，检测分支在输出特征层上直接回归出预测框的定位和分类信息，引入的 MGSA 模块预测出检测分支特征的语义分割 Mask。因此在训练时不仅需要计算每个 anchor 框的定位损失L_{ij}^{reg}、置信度损失L_{ij}^{conf}和分类损失L_{ij}^{cls}，还要计算特征语义分割损失L_r^{mask}，由此完整的损失函数计算表达式为

$$
\begin{aligned}
&\mathrm{Loss}=\sum_{i=0}^{S^2}\sum_{j=0}^{B}(L_{ij}^{\mathrm{reg}}+L_{ij}^{\mathrm{conf}}+L_{ij}^{\mathrm{cls}})+\sum_{r=0}^{S_{\mathrm{m}}^2}L_r^{\mathrm{mask}}L_{ij}^{\mathrm{reg}}=\\
&\lambda_{\mathrm{coord}}I_{ij}^{\mathrm{obj}}[(x_{ij}-x_{ij}^{*})^2+(y_{ij}-y_{ij}^{*})^2+\\
&(\sqrt{w_{ij}}-\sqrt{w_{ij}^{*}})^2+(\sqrt{h_{ij}}-\sqrt{h_{ij}^{*}})^2]\\
&L_{ij}^{\mathrm{conf}}=-I_{ij}^{\mathrm{obj}}\log(\mathrm{cof}_{ij})-\lambda_{\mathrm{noobj}}I_{ij}^{\mathrm{noobj}}\log(1-\mathrm{cof}_{ij})\\
&L_{ij}^{\mathrm{cls}}=-I_{ij}^{\mathrm{obj}}\sum_{c\in\mathrm{cls}}(p_{ij}^{c^{*}}\log(p_{ij}^{c})+(1-p_{ij}^{c^{*}})\log(1-p_{ij}^{c}))\\
&L_r^{\mathrm{mask}}=-M_r^{*}\log(M_r)+\alpha(1-M_r^{*})\log(1-M_r)
\end{aligned}
\tag{2}
$$

其中，S 为输出特征图的边长；B 为每个单元格中设置 anchor 框的数量；I_{ij}^{obj}、I_{ij}^{noobj}分别为第 i 个单元格中第 j 个 anchor 的正、负样本划分情况，当 anchor_{ij} 为正样本时 $I_{ij}^{\mathrm{obj}}=1$，$I_{ij}^{\mathrm{noobj}}=0$，当 anchor_{ij} 为负样本时，取值相反；λ_{coord} 和 λ_{noobj} 分别为平衡定位损失L_{ij}^{reg}和负样本 anchor 的置信度损失L_{ij}^{conf}的权值系数；x_{ij} 为网络输出的预测框中心点横坐标信息，x_{ij}^{*} 为真实框中心点经过编码后的标签，y_{ij}、y_{ij}^{*}、w_{ij}、w_{ij}^{*}、h_{ij}、h_{ij}^{*}与此类似；cof_{ij} 为网络预测第 i 个单元格中第 j 个预测框的置信度值；p_{ij}^{c} 为网络预测

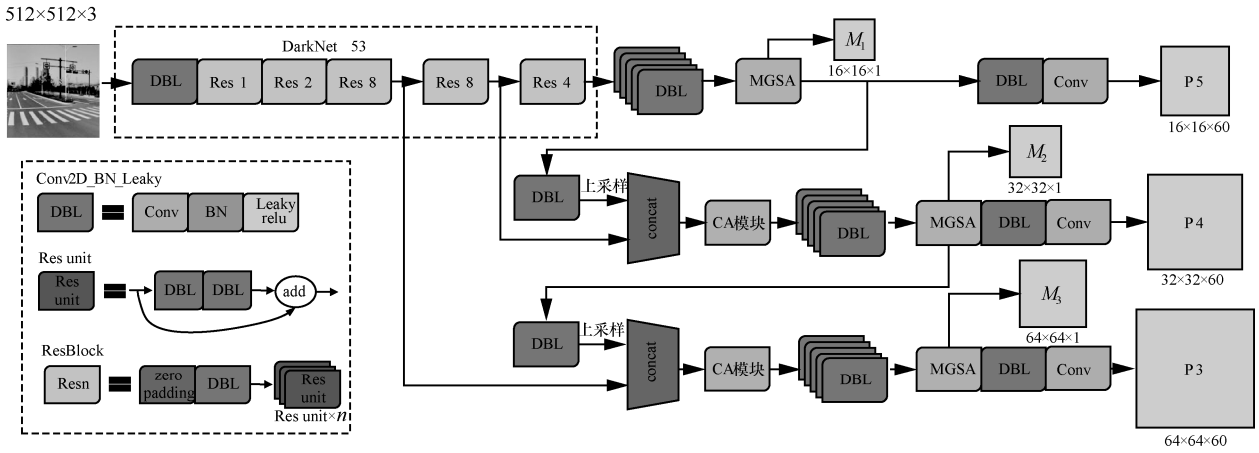


图 10　YOLOv3-A 网络整体结构

第 i 个单元格中第 j 个预测框中包含第 c 类物体的概率；p_{ij}^{c*} 为第 c 类物体的标签。在求解 L_r^{mask} 时，S_m 为语义分割输出特征的边长，M_r 和 M_r^* 分别为第 r 个特征的分割结果和标签。

本文将所提方法在深度学习框架 Pytorch 上实现，网络训练和测试方法如下。在训练阶段，使用随机梯度下降优化算法更新网络参数，设置初始学习率 $\text{lr}=1\times10^{-3}$，每训练 10 个 epoch 将学习率降低至原来的1/10，动量 Momentum=0.9，weight decay 设置为 1×10^{-4}。每个 batch 随机选取 12 个训练样本，在一块 GTX2080Ti 显卡上训练 30 个 epoch 后停止。在测试阶段，使用滑动窗口的方法在 3 071 张 2 048 像素×2 048 像素的测试图像上检测交通标志，滑动窗口大小为 512 像素×512 像素，步长设置为 256 像素，最终整合整幅图像上所有的预测框，并经过 NMS 算法后得到最终的预测框。其中，通道注意力网络中的压缩系数 $r=16$，NMS 算法中的 IOU（intersection over union）阈值设置为 0.5，目标置信度阈值设置为 0.1。

3　实验结果与分析

3.1　实验数据集

本文实验主要采用我国的 TT100K 交通标志数据集[25]，选取该实验集的主要原因在于此数据集包含的交通标志种类齐全，场景丰富。TT100K 数据集中包含的交通标志种类为 221 类，总目标个数为 26 349 个，这两项数据都大大超过了 GTSDB[10]、STS（Swedish traffic signs）[26]和 LISA（laboratory for intelligent and safe automobiles）[27]等数据集。但在 TT100K 数据集上进行目标检测具有较大的挑战性，例如该数据集中的小目标存在绝对尺度小和相对尺度小 2 个难题。目标绝对尺度小是指交通标志的真实尺度较小，即所占的像素面积较小，这就使获得的图像目标模糊不清、信息少、噪声多，导致模型检测困难。如图 11(a)所示，对图像中 2 个存在目标的区域进行放大后，可以发现目标本来就是模糊不清的，难以区别具体类别。目标相对尺度小是指交通标志在整幅图像中占据的像素面积比例小，由于 TT100K 数据集中的图像分辨率为 2 048 像素×2 048 像素，在如此高分辨率背景下，交通标志容易被其他无关物体所干扰，在不断扩大感受野的深度学习网络中，背景物体的信息也被包含进来，使目标的有效信息容易被淹没。此外，高分辨率图像包含了更多的背景信息，存在更多的潜在干扰目标，使网络易出现误检情况。如图 11(b)所示，对原始 2 048 像素×2 048 像素图像的某 2 个区域进行放大后，可以发现图像中存在多处伪交通标志，这就是高分辨率图像中存在的更多干扰背景的问题。

(a) 目标尺度绝对小　　(b)目标尺度相对小

图 11　TT100K 数据集中存在的小目标问题

本文对 TT100K 数据集中交通标志的尺度进行了统计，其尺度分布如图 12 所示。由图 12 可知，数据集中像素面积小于 32 像素×32 像素的交通标志有 10 676 个，占总目标个数的 40.5%，因此该数据集中广泛存在目标绝对尺度小的问题。同时，交

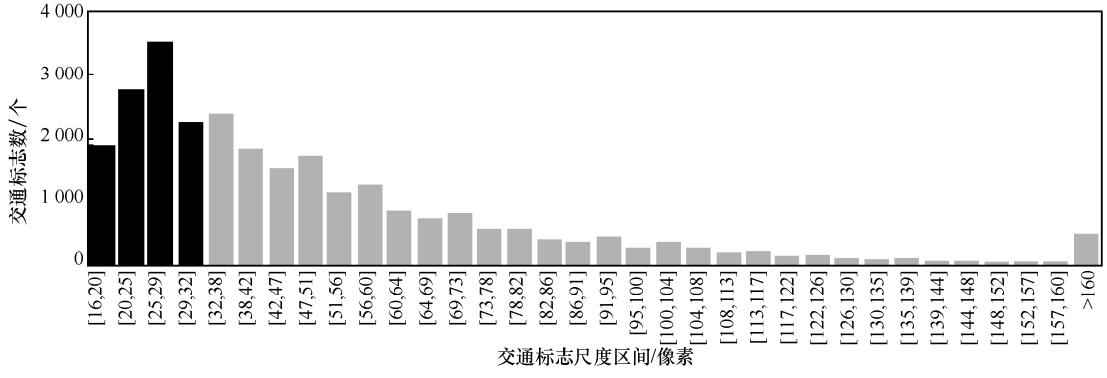


图 12　TT100K 数据集中目标尺度分布

通标志在图像中占整幅图像的像素面积比例不大于 2%的个数超过 24 970 个，占总目标个数的 94.7%，由此可见，此数据集中广泛存在目标尺度相对小的问题，即图像中包含了大量的无关背景信息。因此相比于其他公开数据集，TT100K 数据集的挑战难度较大。

3.2　评价指标

本文使用的模型评价指标与 TT100K 数据集发布者 Zhu 等[25]提供的方法保持一致，采用固定的 IOU 阈值和置信度阈值判断检测结果是否正确。然后，计算预测结果的精确率（Precision）和召回率（Recall），以衡量模型的目标分类能力和目标检测能力。此外，通过设置不同的置信度阈值，绘制模型的精确率−召回率曲线，即 P-R 曲线，直观地展示模型的检测效果。在计算模型的 Precision 和 Recall 这 2 项指标时，首先需要根据真实标签将检测结果划分为真正例（TP, true positive）、真反例（TN, true negative）、假正例（FP, false positive）和假反例（FN, false negative）4 类。

Precision 又称为查准率，通过计算检测结果中预测正确的样本数和所有预测样本数的比例得到，即正确检测到的样本数占总检出样本的比例，能够反映模型对目标的分类能力，其计算式为

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{3}$$

Recall 又称为查全率，通过计算检测结果中预测正确的样本数和所有真实样本数的比例得到，即正确检测到的样本数占真实样本数的比例，能够反映模型对目标的检测能力，其计算式为

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{4}$$

Precision 和 Recall 这 2 个指标是相互矛盾的，当设置的 IOU 阈值和物体置信度阈值较高时，计算出的 Precision 值较高，Recall 值较低。因此，为了综合对比网络性能，本文通过 P-R 曲线进行比较。P-R 曲线以 Precision 为横坐标、Recall 为纵坐标，曲线下包围的面积越大，代表模型的性能越好。

3.3　注意力机制有效性实验

本文先对所提出的通道注意力机制中不同的特征压缩方法进行了实验对比，证明了对检测特征进行全局最大池化（GMaxPool）和全局平均池化（GAvgPool）的拼接组合更利于提升算法的效果。然后通过消融实验，证明了所提的 2 种注意力机制都能够对检测结果起到正面作用，并通过特征可视化，直观地展示了 2 种注意力机制的结合对交通标志特征的聚焦和改善作用。其中，在判断预测框是否为正确检测时，设置预测框与真实框的 IOU 阈值为 0.5，类别置信度阈值为 0.5。

对于通道注意力机制的特征压缩方法的选择，本文对比了 GMaxPool、GAvgPool、全局最大池化与全局平均池化按通道相加（GMaxPool+ GAvgPool）、全局最大池化与全局平均池化在通道维度上进行拼接（Concat(GMaxPool, GAvgPool)）4 种方法。在实验时，分别将这 4 种方法用于 YOLOv3 网络之中，并且采用相同的训练和测试方法，得到在 TT100K 数据集上所有尺度目标的 Precision 和 Recall 结果，如表 1 所示。

表 1　　4 种特征压缩方法对比

特征压缩方法	Precision	Recall
GMaxPool	86.9%	89.8%
GAvgPool	86.7%	89.6%
GMaxPool + GAvgPool	87.0%	90.1%
Concat(GmaxPool, GAvgPool)	**87.2%**	**90.3%**

由表 1 可以看出，对检测特征在空间维度进行压缩时，GMaxPool 略好于 GAvgPool，Concat(GmaxPool, GAvgPool)检测效果最好。因此在所提通道注意力机制中，本文选择了对检测特征进行全局最大池化和全局平均池化后再拼接的方式，来代表检测特征的通道响应情况。

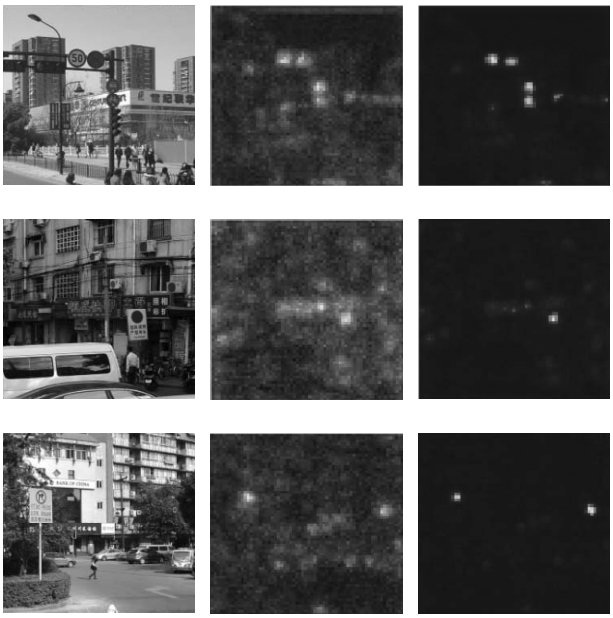
此外，针对所提的 2 种注意力方法，本文还进行了相关消融实验。首先以原始的 YOLOv3 网络为对比基准，然后分别将通道注意力机制和基于语义分割引导的空间注意力机制添加到 YOLOv3 网络的检测分支中，保持训练和测试方法一致，对比在 TT100K 数据集上所有尺度目标的 Precision 和 Recall，得到的消融实验数据如表 2 所示。

表 2　　注意力机制消融实验数据

方法	CA	MGSA	Precision	Recall
YOLOv3	—	—	86.6%	89.4%
YOLOv3-CA	√	—	87.2%	90.3%
YOLOv3-MGSA	—	√	88.0%	91.6%
YOLOv3-A	√	√	88.5%	92.2%

由表 2 可以看出，原始 YOLOv3 网络在 TT100K 数据集上的 Precision 和 Recall 值分别为 86.6%和 89.4%，融合 2 种注意力机制的 YOLOv3-A 网络的 Precision 和 Recall 值分别为 88.5%和 92.2%，分别提升了 1.9%和 2.8%。此外，将 CA 和 MGSA 分别引入 YOLOv3 网络之后，网络模型的性能均有不同程度的提高，可以看出 MGSA 模块对网格的性能改善效果更好。由此消融实验可以说明，在 YOLOv3 网络的检测分支中，加入所提出的通道注意力机制和基于语义分割引导的空间注意力机制，能够有效地提高网络的精确率和召回率，而且对召回率的改善更加明显，减少了网络的漏检和误检情况。

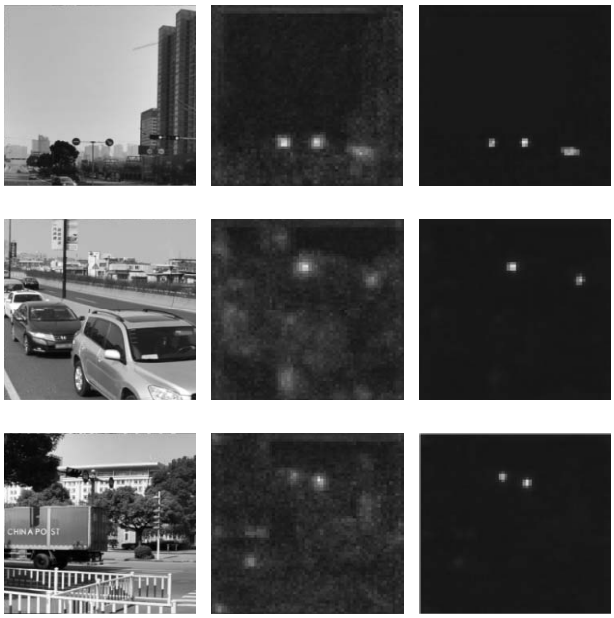
对特征可视化时，本文分别将 YOLOv3 网络和 YOLOv3-A 网络的 P3 层级上的特征在通道维度上进行平均池化，通过热图的形式进行可视化对比。复杂背景下注意力效果如图 13 所示。对于背景复杂的原始图像，YOLOv3 网络的输出特征如图 13(b)所示，除了在目标区域有较高的激活外，其他区域的特征分布杂乱，而且个别区域含有较高的激活，很容易产生误检；YOLOv3-A 网络的输出特征如图 13(c)所示，明显地只在目标区域有较高的激活，其他区域的干扰特征能够得到很好的抑制；此外，从数值上看，YOLOv3-A 网络的输出特征在目标区域的激活值更高，可以说明注意力机制能够对有效特征起到增强和改善的作用。



(a)原始图像　　(b) YOLOv3 输出特征　(c) YOLOv3-A 输出特征

图 13　复杂背景下注意力效果

小尺度目标的注意力效果如图 14 所示。对于包含小尺度目标的原始图像，YOLOv3 网络的输出特征如图 14(b)所示，在小目标周围存在其他杂乱的激活特征，这些特征再经过后层网络的卷积融合后，容易对小目标的有效特征形成干扰，使网络分类错误，造成误检情况的发生。YOLOv3-A 网络的输出特征如图 14(c)所示，在经过 2 种注意力机制之后，小目标物体所在区域的特征形成了明显聚焦，而且很好地抑制了小目标周围的干扰特征和其他无关区域的特征，这说明了所提出的注意力方法能够起到了保护小目标有效特征的作用。



(a)原始图像　　(b)YOLOv3 特征　　(c)YOLOv3-A 特征

图 14　小尺度目标的注意力效果

由此可见，本文提出的通道注意力方法和基于语义分割引导的空间注意力方法能够模拟人类的视觉选择性机制，让网络聚焦和增强有效区域信息，同时能较好地抑制干扰信息，能够对城市街道场景下交通标志检测存在的图像背景复杂、干扰物体较多和小目标问题起到良好的改善作用。

3.4　与其他方法的性能对比

为了对比 YOLOv3-A 网络与其他主流一阶段方法在不同尺度物体上的检测性能，本节实验主要按照 Zhu 等[25]提出的划分方法，将目标尺度在(0, 32]像素的物体设为小目标，尺度在(32, 96]像素的物体设为中目标，尺度在(96, 400]像素的物体设为大目标，然后调整分类置信度，分别计算模型预测结果在小目标、中目标、大目标和整体尺度(0, 400]像素

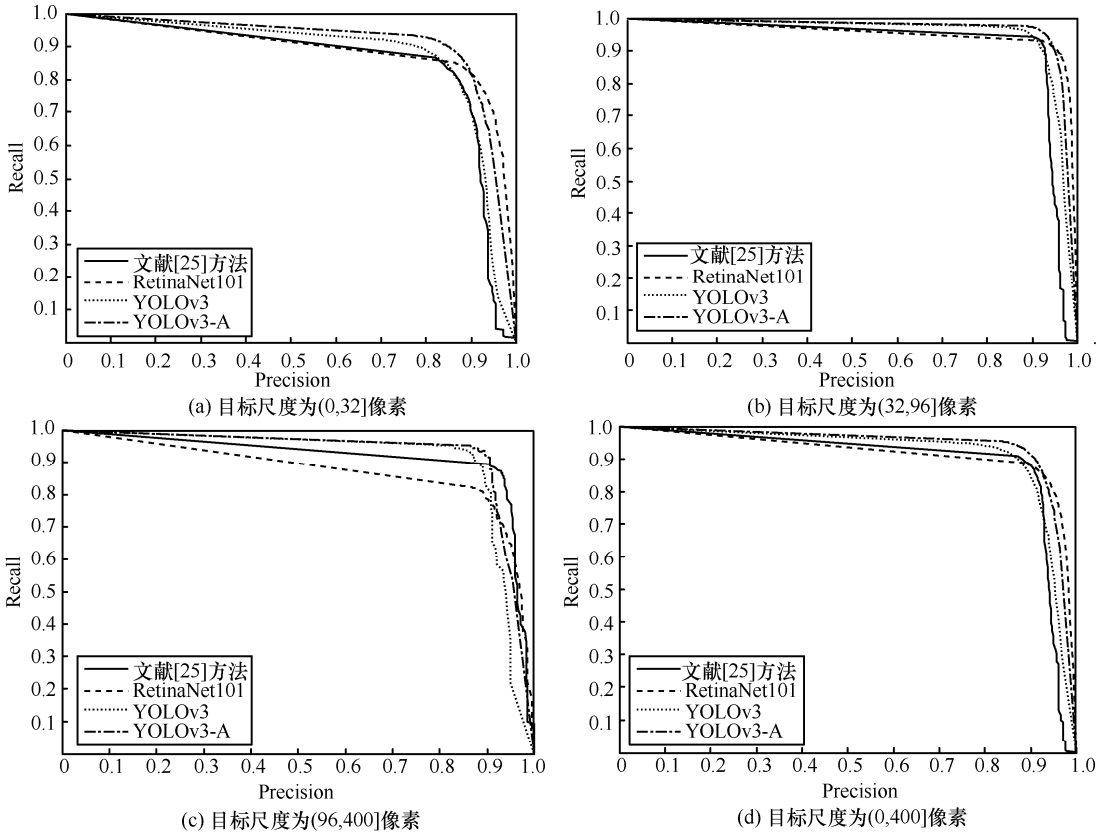


图 15　YOLOv3-A 与其他方法对比的 P-R 曲线

目标上的 Precision 和 Recall 值，绘制出的 P-R 曲线如图 15 所示，直观地展示了文献[25]、RetinaNet101[8]、YOLOv3[14]和 YOLOv3-A 等方法在 TT100K 数据集上的检测效果。文献[25]方法是在 OverFeat 网络框架[28]的基础上改进而来的，使用全卷积的方式完成目标检测和分类；RetinaNet101 算法引入 focal loss 损失函数，极大地缓解了一阶段目标检测算法中正负样本不均衡的问题，提升了一阶段目标检测算法的性能，是一阶段目标检测算法的代表。

由图 15 可以看出，代表 YOLOv3-A 网络的 P-R 曲线在不同尺度物体的检测结果中，都能包围住 YOLOv3 网络的 P-R 曲线，这说明引入的 2 种注意力机制对各个尺度物体的检测都有不同程度地提高。此外，YOLOv3-A 网络的曲线在各个尺度物体的检测结果中都可以绝大程度地包围文献[25]和 RetinaNet 方法的 P-R 曲线。

为了从数值上对比 4 种检测方法在各个尺度上检测性能的差异，本文计算了每条曲线与坐标轴围成的面积 AUC，计算结果如表 3 所示。从表 3 中可以看出，4 种方法对于中等尺度目标的检测性能均优于对小目标和大目标的检测结果，这是由于 TT100K 数据集中的中等尺度目标占据了绝大部分样本而导致的。另外，YOLOv3-A 网络在各个尺度上的检测结果都优于其他 3 种检测方法，而且相对于 YOLOv3，其在小目标检测结果中的提升最多，达到了 3.5%，说明了引入的注意力机制能够缓解目标检测方法中的多尺度问题，而且对小目标检测性能的改善尤为明显。

表 3　　**4 种检测方法的 AUC 对比**

方法	小目标	中目标	大目标	整体尺度目标
文献[25]方法	0.854	0.921	0.915	0.897
RetinaNet101[8]	0.891	0.957	0.873	0.920
YOLOv3[14]	0.876	0.952	0.914	0.917
YOLOv3-A	**0.911**	**0.965**	**0.938**	**0.941**

在运行时间方面，YOLOv3-A 的运行环境为 NVIDIA RTX 2080 Ti GPU，训练阶段所需的时间大约为 5 h。尽管因数据量较大该过程耗时较长，但由于是离线操作并不会给后续测试过程造成影响。实验测试结果表明，YOLOv3-A 仅需 0.8 s 即可求

得交通标志检测结果，因而能够满足实际应用场景的实时性要求。

4 结束语

本文主要介绍了基于注意力机制的交通标志检测网络 YOLOv3-A，分析了在实际的交通标志检测场景中普遍存在的目标绝对尺度小和相对尺度小问题对目标检测网络的影响，提出了在 YOLOv3 网络的检测分支上引入通道注意力机制和基于语义分割引导的空间注意力机制 2 种方法改善网络对目标的关注程度，提高了对小目标和遮挡变形目标的检测性能。本文对 2 种注意力机制的设计原理和网络结构进行了详细的阐述，其中，通道注意力机制结合 SENet 中的注意力方法和 FPN 特征融合的特点进行改进；基于语义分割引导的空间注意力机制以目标的标定框为监督信息，在特征层面进行语义分割的学习，并且与自带 attention 属性的深层卷积特征相结合，完成了特征的空间注意力机制。通过消融实验和特征可视化的方式，验证了这 2 种注意力机制的有效性。通过 P-R 曲线对比了所提方法与其他目标检测方法在不同尺度物体上的检测性能，表明了具有这 2 种注意力机制的 YOLOv3-A 网络在不同尺度目标上的检测能力更强。

参考文献：

[1] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2005: 886-893.

[2] LEE T S. Image representation using 2D Gabor wavelets[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1996, 18(10): 959-971.

[3] VIOLA P A, JONES M J. Rapid object detection using a boosted cascade of simple features[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2001: 511-518.

[4] REN S, HE K, 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.

[5] 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. Piscataway: IEEE Press, 2016: 779-788.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//European Conference on Computer Vision. Berlin: Springer, 2016: 21-37.

[7] RAJENDRAN S P, SHINE L, PRADEEP R, et al. Fast and accurate traffic sign recognition for self driving cars using RetinaNet based detector[C]//2019 International Conference on Communication and Electronics Systems. Piscataway: IEEE Press, 2019: 784-790.

[8] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2980-2988.

[9] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[10] HOUBEN S, STALLKAMP J, SALMEN J, et al. Detection of traffic signs in real-world images: the German traffic sign detection benchmark[C]//The 2013 International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2013: 1-8.

[11] YANG Y, LIU S, MA W, et al. Efficient traffic-sign recognition with scale-aware CNN[C]//British Machine Vision Conference. London: BMVA Press, 2017: 1-13.

[12] LARSSON F, FELSBERG M. Using Fourier descriptors and spatial models for traffic sign recognition[C]//2011 Scandinavian Conference on Image Analysis(SCIA 2011). Berlin: Springer, 2011: 238-249.

[13] MENG Z, FAN X, CHEN X, et al. Detecting small signs from large images[C]//2017 IEEE International Conference on Information Reuse and Integration. Piscataway: IEEE Press, 2017: 217-224.

[14] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arXiv Preprint, arXiv: 1804.02767, 2018.

[15] 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. Piscataway: IEEE Press, 2017: 2117-2125.

[16] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 42(8): 2011-2023.

[17] WANG F, JIANG M, QIAN C, et al. Residual attention network for image classification[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6450-6458.

[18] HOU Y, MA Z, LIU C, et al. Learning lightweight lane detection CNNs by self-attention distillation[C]//2019 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 1013-1021.

[19] ZHANG X, WANG T, QI J, et al. Progressive attention guided recurrent network for salient object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 714-722.

[20] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[21] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386-397.

[22] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015:1-9.

[23] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv Preprint, arXiv:1706.05587, 2017.

[24] UIJLINGS J R R, VAN D S K E A, GEVERS T, et al. Selective search for object recognition[J]. International Journal of Computer Vision,

2013, 104(2): 154-171.

[25] ZHU Z, LIANG D, ZHANG S, et al. Traffic-sign detection and classification in the wild[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2110-2118.

[26] LARSSON F, FELSBERG M, FORSSEN P E. Correlating Fourier descriptors of local patches for road sign recognition[J]. IET Computer Vision, 2011, 5(4):244-254.

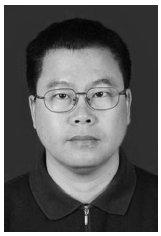
[27] MOGELMOSE A, TRIVEDI M M, MOESLUND T B. Vision based traffic sign detection and analysis for intelligent driver assistance systems: perspectives and survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(4):1484-1497.

[28] SERMANET P, EIGEN D, ZHANG X, et al. Overfeat: integrated recognition, localization and detection using convolutional networks[J]. arXiv Preprint, arXiv:1312.6229, 2013.

[作者简介]

郭璠（1982− ），女，湖南临澧人，博士，中南大学副教授、硕士生导师，主要研究方向为图像处理、计算机视觉、人工智能等。

张泳祥（1994− ），男，河南安阳人，中南大学硕士生，主要研究方向为模式识别、图像处理等。



唐琎（1966− ），男，湖南武冈人，博士，中南大学教授、博士生导师，主要研究方向为计算机视觉、机器人、嵌入式系统、智能信息处理等。

李伟清（1997− ），男，河南信阳人，中南大学硕士生，主要研究方向为医学图像处理、机器学习等。