文章编号:1005-0523(2024)01-0061-09

基于Transformer的交通标志检测模型研究

严丽平[1],张文剥[1],宋 凯[2],蔡 彧[1],王 静[1],徐嘉悦[1]

(1. 华东交通大学软件学院,江西 南昌 330013; 2. 华东交通大学信息工程学院,江西 南昌 330013)

摘要:【目的】为了解决在复杂环境下,对小目标特征困难以及对小目标检测效果不佳等问题,提出了一种基于Transformer的交通标志检测基干模型。**【方法】**通过充分利用卷积和Transformer的优势,构建了一种注意力融合的多尺度特征提取基干模型,能够使基干网络以全局上下文信息为支撑,有选择地增强有用信息的特征,并抑制不重要的特征。此外,为了在增强特征融合的同时防止网络退化,还加入了类池连接。最后,在TT100K数据集上进行实验。**【结果】**实验结果表明,以该模型为骨干的元体系结构取得了最高84%的mAP,与基线模型相比mAP最大提升约7%。**【结论】**模型在提高特征提取效果的同时,也为交通标志检测提供了一种新的思路。

关键词:交通标志检测;自动驾驶;Transformer;注意力融合

中图分类号:TU391.41;U463.6 **文献标志码:**A

本文引用格式:严丽平,张文剥,宋凯,等. 基于Transformer的交通标志检测模型研究[J]. 华东交通大学学报,2024,41(1):61-69.

Research on Traffic Sign Detection Model Based on Transformer

Yan Liping[1], Zhang Wenbo[1], Song Kai[2], Cai Yu[1], Wang Jing[1], Xu Jiayue[1]

(1. School of Software, East China Jiaotong University, Nanchang 330013, China;
2. School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract:【Objective】 In order to solve the difficulties such as small target feature extraction, a transformer-based traffic sign detection model was proposed.**【Method】**Through fully utilizing the advantages of convolution and Transformer, a multi-scale feature extraction backbone model was established with attention fusion, which could enable the backbone network to selectively enhance the features of useful information and suppress the unimportant ones with the support of global context information. In addition, pooling-like connection are incorporated in order to prevent network degradation while enhancing feature fusion. Finally, experiments were conducted on the TT100K dataset.**【Result】**The experimental results show that the meta-architecture with this model as the backbone achieves the highest mAP of 84%, and the maximum improvement of mAP is about 7% compared with the baseline model.**【Conclusion】** The model provides a new idea for traffic sign detection while improving feature extraction.

Key words: traffic sign detection; automatic driving; Ttransformer; attention fusion

Citation format: YAN L P, ZHANG W B, SONG K, et al. Research on traffic sign detection model based on Transformer[J]. East China Jiaotong University, 2024, 41(1): 61-69.

收稿日期:2023-10-24

基金项目:国家自然科学基金项目(62362031,62262022);江西省自然科学基金项目(20224BAB202021)

【研究意义】作为自动驾驶和高清地图环境感知的关键技术之一，交通标志检测对于为车辆提供道路信息判断和实时安全预警具有重要意义。由于道路条件和自然环境的不同，交通标志检测的结果受到光线变化、恶劣天气和运动模糊等诸多因素的限制，大大增加了这项任务的难度。

【研究进展】大多数传统的交通标志检测方法都依赖于人工从颜色信息[1]和几何形状[2]中提取特征。但是，由于传感器在运动中传输，交通标志区域的比例变化、遮挡等问题阻碍了这些方法的实际应用。

为了在准确性和效率之间取得平衡，先进的物体检测算法开始使用深度卷积神经网络（CNN）[3]代替人工特征提取。经典的两阶段检测模型如Faster R-CNN[4]、Cascade R-CNN[5]等，虽然检测精度高，但其复杂的结构导致检测效率低下。与两阶段模型相比，单阶段模型，如RetinaNet[6]、SSD[7]系列以及YOLO[8]系列的结构相对简单，其检测效率较高，但检测精度却不尽如人意。

最近，基于Transformer的新模型表明端到端的标准转换器可以执行目标检测[9]、分类[10]、分割等任务[11]。如ViT、PVT[12]等在各种计算机视觉任务中取得了令人鼓舞的成果并迅速成为基干模型，这是因为Transformer拥有强大的建模能力。

【创新特色】然而，Transformer将图像视为序列，在对局部窗口中的视觉特征以及尺度变换进行建模时，缺乏获取通道维度信息的能力，随着网络深度的加深，导致每个通道之间的信息逐渐丢失，因此Transformer无法直接用于复杂环境下的小目标特征提取。然而卷积却可以为Transformer提供必要的通道维度信息。

【关键问题】本文提出了基于类池化连接的注意力融合转换器（transformer based on attention fusion with pooling-like connection, AFPC-T），通过将可缩放的卷积注意力模块（scalable convolutional attention block, SCAB）嵌入到标准的Transformer中构建双注意力融合模块（dual attention block, DAB），并通过类池化连接（pooling-like connection, PC）模块来加强特征融合，然后通过高度集成的PAB（pooling-like attention block, PAB）模块建立分层式网络基干模型，来解决在复杂的交通环境下对小目标特征提取困难等问题。

1 AFPC-T整体架构

图1展示了高度集成的AFPC-T网络架构及其组件。AFPC-T是四阶段特征提取基干模型，即输入1幅图像输出4张不同尺度的特征图用于后续的分类和回归。在第一阶段开始之前，需要对输入图像进行特征编码（Patch embedding）。例如，给定一幅大小为 $H\times W\times 3$ 的二维图像特征，其中 H 为特征高度，W 为特征宽度，3为通道数，将其划分为每个大小为 $4\times 4\times 3$ 的 $\frac{HW}{4^2}$ 个一维序列。然后，对一维序列进行线性投影，得到大小为 $\frac{HW}{4^2}\times C_1$ 的序列，其中 C_1 为映射维度，之后序列进入第一阶段。嵌入的序列在PAB模块中进行特征提取后，一方面通过维度转换（Reshape），得到大小为 $\frac{H}{4}\times\frac{W}{4}\times C_2$ 的二维特征图F1，其中 C_2 为通道维度，一方面通过特征融合后得到最终的一维序列。同样，使用前一阶段的序列映射作为输入，可以得到以下特征输出：F2，F3，F4。它们相对于输入图像的步长分别为8、16像素和32像素。最终的4个特征图{F1，F2，F3，F4}，其大小分别为 $\frac{H}{4}\times\frac{W}{4}\times C_2$，$\frac{H}{8}\times\frac{W}{8}\times 2C_2$，$\frac{H}{16}\times\frac{W}{16}\times 4C_2$，$\frac{H}{32}\times\frac{W}{32}\times 8C_2$，再经过多尺度特征融合后，最终作为分类和检测模型的输入。

2 PAB整体结构

PAB结构主要包括双注意力模块(DAB）和嵌入DAB的可缩放的卷积注意力模块(SCAB）。随后，DAB、PC和其他模块被集成到PAB中，以减少结构冗余。因此，高度集成的PAB模块构建AFPC-T四阶段分层网络结构以生成不同尺度特征图用于不同任务。

2.1 SCAB结构

卷积在深度学习中被广泛应用，因为它能够有效地捕捉数据的空间结构，提取重要特征。所以，为了增强网络的表征能力，SCAB模块被用来模拟通道之间的关系。具体方法是先压缩空间信息，再通过激励生成标量值来代表每个通道的重要性。图1（d）展示了这一过程，输入的图像特征先通过通道全局平均池化（global average pooling, GAP）来进行空间压缩，再通过一维卷积（conv1d, Conv）进行局部跨通道交互以取代原来的多层感知机（multi-

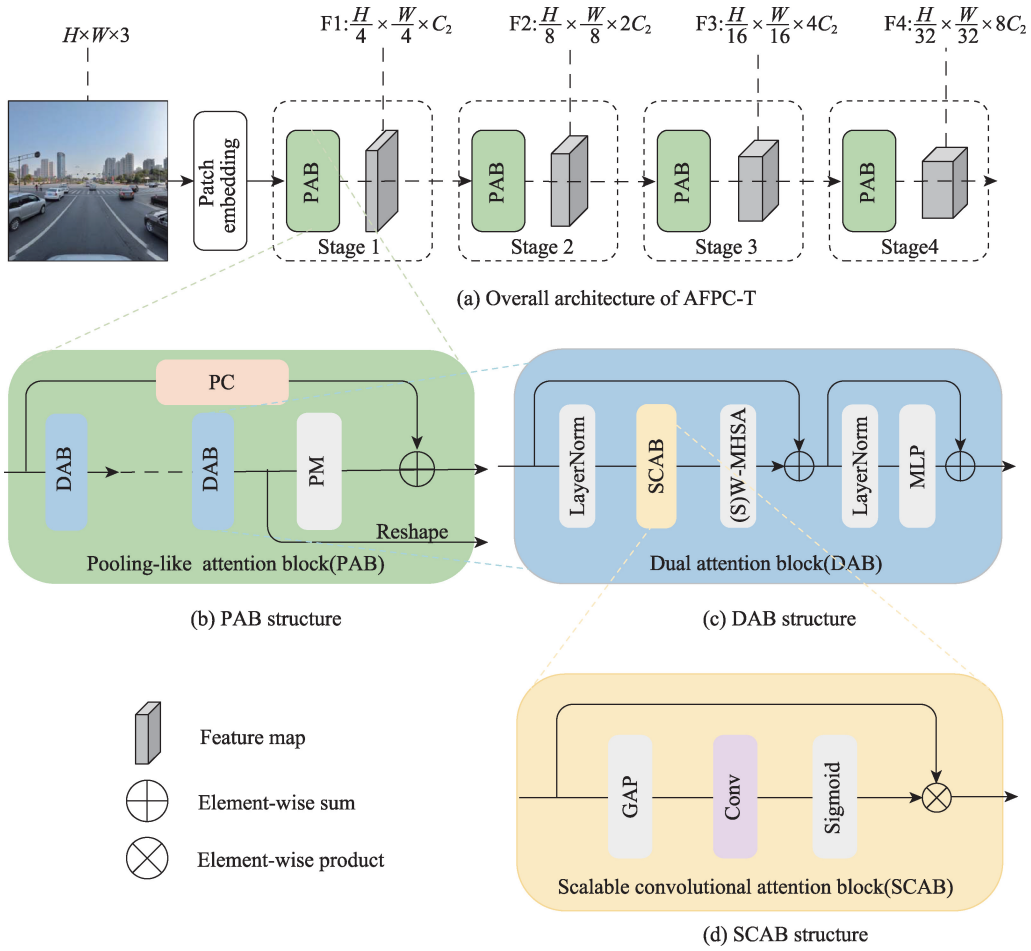


图1　AFPC-T整体架构及其组件结构

Fig. 1　AFPC-T overall architecture and its component structure

layer perceptron, MLP)来降低模型的复杂度。设 $x\in\mathbb{R}^{H\times W\times C_2}$ 为输入的图像特征，则通道权重可以写成

$$\omega=\alpha\left(\sigma\left(\mathrm{Conv}\left(g(x)\right)\right)\right) \tag{1}$$

式中：$g(x)=\frac{1}{WH}\sum_{i=1,j=1}^{W,H}x_{ij}$ 是GAP；σ 是Sigmoid激活函数；α 是平衡因子，用于平衡通道注意力的影响，其值设为0.1。设 $y=g(x)$，那么Conv可以写成

$$\mathrm{Conv}(y)=C1D_S(y) \tag{2}$$

式中：$C1D_S$ 表示一维卷积，S 表示一维卷积的卷积核大小，其值设为3。通过激活通道权重 ω，对特征 x 的每个通道下的特征映射 $x_{C_2}\in\mathbb{R}^{H\times W}$ 进行重新缩放，从而得到最终输出。具体表达式可写为

$$\tilde{x}_{C_2}=\mathrm{Fscale}\left(x_{C_2},\omega_{C_2}\right)=\omega_{C_2}x_{C_2} \tag{3}$$

式中：$\tilde{X}=\left[\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_{C_2}\right]$ 是通道注意的输出，即带有通道注意力的特征图并且 $\tilde{X}\in\mathbb{R}^{H\times W\times C_2}$；$\mathrm{Fscale}\left(x_{C_2},\omega_{C_2}\right)$ 是标量 ω_{C_2} 与特征图 $x_{c_2}\in\mathbb{R}^{H\times W}$ 之间的通道乘法。小目标通常具有相对较低的信噪比，可能会被背景干扰，而SCAB通过学习每个通道的权重，可以使网络在处理小目标时更灵敏，更有针对性地捕捉小目标的特征。

2.2　DAB结构

为了改进模型的表示，本文在标准转换器(图2(a))中添加了一个基于通道注意力的模块SCAB。在这种改进的架构中(图2(b))，在LN模块之后，输入的特征先通过SCAB模块得到带有通道注意力的特征后，再进入(S)W-MHSA模块中得到通道注意力与空间注意力融合后的特征，之后进入随后的LN模块和MLP模块。此外，需要注意的是，如图2(c)所示，每个由W-MHSA模块组成的DAB之后都必须带有一个由SW-MHSA模块组成的DAB。因此，对于给定的输入特征 x，连续的DAB可以精确描述如下

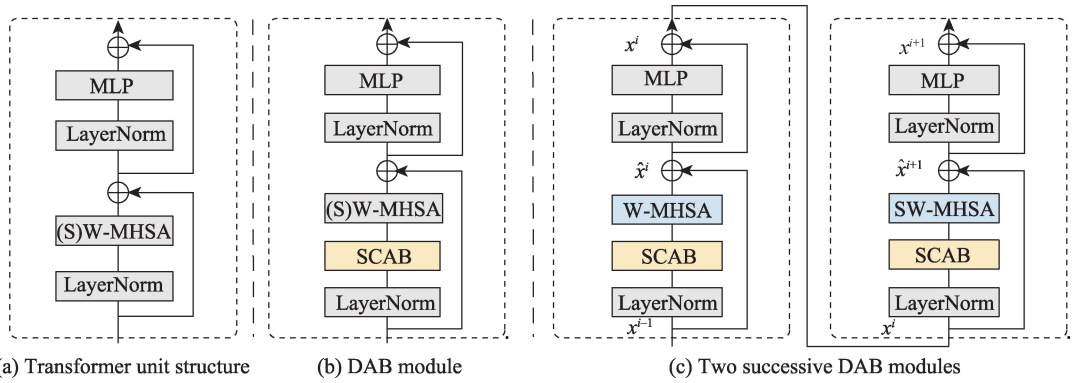


图2 不同单元结构对比

Fig. 2 Comparison of different unit structur

$$\begin{cases}\hat{x}^i=\mathrm{W-MHSA}\left(\mathrm{SCAB}\left(\mathrm{LN}\left(x^{i-1}\right)\right)\right)+x^{i-1}\\ x^i=\mathrm{MLP}\left(\mathrm{LN}\left(\hat{x}^i\right)\right)+\hat{x}^i\\ \hat{x}^{i+1}=\mathrm{SW-MHSA}\left(\mathrm{SCAB}\left(\mathrm{LN}\left(x^i\right)\right)\right)+x^i\\ x^{i+1}=\mathrm{MLP}\left(\mathrm{LN}\left(\hat{x}^{i+1}\right)\right)+\hat{x}^{i+1}\end{cases}\tag{4}$$

式中：i 表示第 i 个 DAB 模块；$\hat{x}^i$ 和 x^i 分别表示 (S)W - MHSA 和 SCAB 融合后的特征输出以及 MLP 的输出特征；W - MHSA 和 SW - MHSA 分别表示使用常规和滑动窗口的多头自注意力。对于给定的输入特征 $x\in\mathbb{R}^{H\times W\times C_2}$，将其划分为大小为 $M\times M$ 的 $\frac{HW}{M^2}$ 个局部窗口，其中 M 为窗口大小，值为7。然后，在每个局部窗口内计算自注意力，对于某个局部窗口特征 $x_w\in\mathbb{R}^{M^2\times C_1}$，使用 MHSA（multi-head self-attention, MHSA）对其依赖关系建模，则基于窗口的自注意力可以表示为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{SoftMax}\left(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{D_h}+B\right)\boldsymbol{V}\tag{5}$$

式中：查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$ 由线性映射计算得出，即 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}=x_w\boldsymbol{W}_Q,x_w\boldsymbol{W}_K,x_w\boldsymbol{W}_V$。其中 $\boldsymbol{W}_Q$，$\boldsymbol{W}_K$，$\boldsymbol{W}_V\in\mathbb{R}^{C_1\times D_h}$ 分别表示查询矩阵、键矩阵和值矩阵。D_h 通常设为 D/h，D 表示序列映射维度，h 是自注意力的头数。单头自注意力重复 h 次，并将 h 个头的输出特征沿通道维度串联起来，形成多头自注意力。SoftMax 表示Softmax激活函数，自注意力的实现采用了相对位置编码，B 表示相对位置编码，T 表示转置。SCAB 模块产生的带有通道注意力特征是作为(S)W-MHSA模块的输入，所以经过(S)W-MHSA模块建模后的特征即是最终的双注意力融合的特征 x_d。x_d 在LN模块之后进入MLP模块，MLP模块由两层神经网络组成，其精准表达为

$$\mathrm{MLP}\left(LN\left(x_d\right)\right)=\mathrm{W}_2GELU(\mathrm{W}_1\mathrm{LN}(x_d)+b_1)+b_2\tag{6}$$

式中：$W_1\in\mathbb{R}^{C_1\times D_{\mathrm{mlp}}}$，$W_2\in\mathbb{R}^{D_{\mathrm{mlp}}\times C_1}$ 是学习到的线性变换；D_{mlp} 表示线性映射维度；GELU 是激活函数。最后，双注意力融合的特征映射在经过特征交互后，进入下一个DAB模块或一方面通过Reshape操作得到一个二维的特征图输出，一方面进入序列合并（patch merging, PM）。

Transformer虽然拥有对每个元素间的相似性权重建模的强大能力，但是却缺少了建模通道间的关系。而本文将可以建模通道间关系的SCAB模块嵌入到(S)W-MHSA模块之前，利用Transformer和卷积优势互补，形成了双注意力融合的特征映射，使得模型一方面可以利用全局上下文信息，加强对小目标特征增强的同时减少背景信息的干扰，另一方面双注意力映射可以同时关注不同通道不同位置的特征，以提高模型对小目标的敏感性。

2.3 PAB结构

如图1(b)所示，PAB集成了3个主要组件：DAB模块、PM模块和PC模块。特征提取是通过在PAB中堆叠一个或多个DAB模块来实现的。DAB提取后的特征映射，一方面通过Reshape操作获得二维特征图用于后续的多尺度特征融合，一方面通过PM模块进行下采样以实现分层式结构，最后加入PC模块以增强特征提取。PC模块用序列合并取代了原始卷积操作，在增强特征融合的同时，还避免了融合不同结构特征的问题。AFPC-T可以提供不同尺度的特征，以帮助完成分类和回归任务。在进入第一阶段之前，给定的二维图像特征 $x\in\mathbb{R}^{H\times W\times 3}$ 经过序列编码和映射后变成大小为

$x\in\mathbb{R}^{(H_1\times W_1)\times C_1}$ 的一维序列，其中 $H_1=H/4$，$W_1=W/4$，C_1 是映射维度，默认为96。那么PAB模块可以被写为

$$y_j=\mathrm{PM}\left(\mathrm{DAB}_l\left(x_j\right)\right)+\mathrm{PC}\left(x_j\right) \tag{7}$$

式中：PM 和 PC 均为 Patch merging 操作；y_j 表示第 j 个 PAB 模块的一维特征输出；x_j 表示第 j 个 PAB 模块的一维特征输入，这里 j 等于1，DAB_l 表示含有 l 个 DAB。输入序列 $x\in\mathbb{R}^{(H_1\times W_1)\times C_1}$ 通过 DAB 获得用于注意力融合的特征映射，然后经过一个或多个 DAB 后进入 PM。在 PM 中，一维序列将被重塑为二维特征，然后将宽度和高度减半，维度变为原来的2倍，即 x 的大小变为 $H_2\times W_2\times 2C_2$，其中 $H_2=H_1/2$，$W_2=W_1/2$。之后，二维特征被转换回一维序列。最后，转换的序列与 PC 生成的特征融合，然后进入下一个PAB模块。

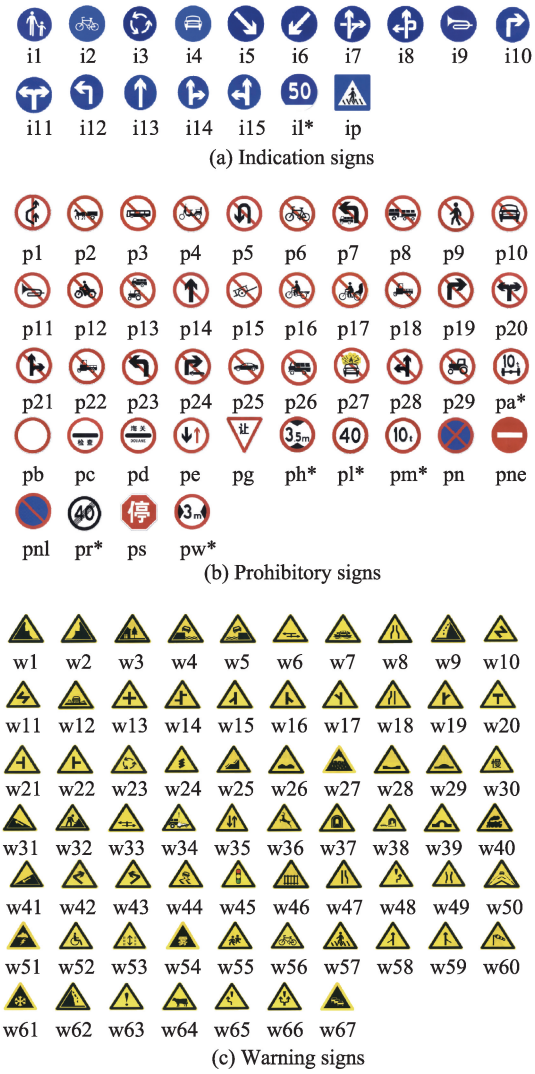
3　实验数据预处理及实施细节

TT100K数据集是最受欢迎的交通标志数据集之一，它包含各种场景下的交通标志，更能反映真实的交通状况。该数据集包含3个大类，共221种，基本覆盖中国所有交通标志。如图3所示展示了部分交通标志：指示标志，禁止标志，警告标志。照片数量超过100 000张，分辨率为2 048×2 048像素。

为了提高检测效果，本文从数据集中删除了未标记和重复的交通标志图像，并选取了42种交通标志类别，每个类别的图像都大于100张，其中有6 105张训练图像和3 071张测试图像。

此外，为了提高模型的预测性能，还采用了数据增强技术来扩展数据集。如图4所示，通过4(b)亮度变化，4(c)添加噪声和4(d)翻转等至少一种或多种效果，使得每个类别都超过500个实例。经过数据扩充后，最终的训练数据集包含17 704幅图像。表1显示了最终的训练和测试图像数量。需要注意的是，训练集以及测试集中都包含了各种各样的环境以及场景，本文没有专门区分特定的场景。

本文使用了Microsoft COCO基准中提到的相同检测指标，这有助于发现检测器对不同大小物体的检测能力。包括小型物体(面积小于32×32像素)、中型物体(面积大于32×32像素小于96×96像素)和大型物体(面积大于96×96像素)。平均精度(mean average precision, mAP)和每秒帧数(frames per second, FPS)也用于衡量每种方法的性能。对于多类



(Note: Class 42 includes: ['i2′, 'i4′, 'i5′, 'il100′, 'il60′, 'il80′, 'io′, 'ip′, 'p10′, 'p11′, 'p12′, 'p19′, 'p23′, 'p26′, 'p27′, 'p3′, 'p5′, 'p6′, 'pg′, 'ph4′, 'ph4.5′, 'pl100′, 'pl120′, 'pl20′, 'pl30′, 'pl40′, 'pl5′, 'pl50′, 'pl60′, 'pl70′, 'pl80′, 'pm20′, 'pm30′, 'pm55′, 'pn′, 'pne′, 'po′, 'pr40′, 'w13′, 'w55′, 'w57′, 'w59′])

图3　3类标志

Fig. 3　Three signs

检测，mAP表示所有类别中AP的平均值。此外，本文在计算mAP时使用的IoU(intersection over union, IoU)值为0.5和0.75。

本文使用了3种具有代表性的元架构和ResNet-101作为基线来评估AFPC-T的性能。元架构主要包括两个两阶段模型Faster R-CNN和Cascade R-CNN，以及一个单阶段模型RetinaNet。具体来说，使用AFPC-T构建这些框架的骨干，所有以AFPC-T为骨干的模型的深度均为[2，2，6，2]，均使

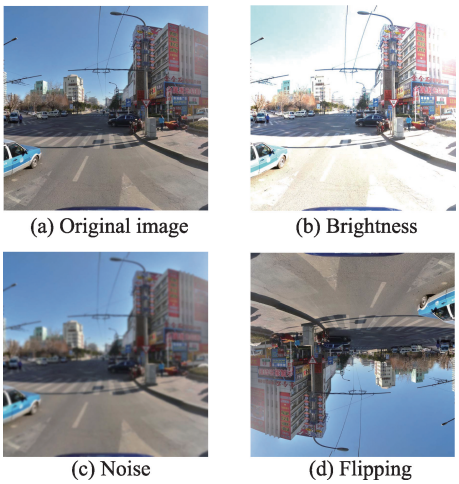


图4 数据增强

Fig. 4 Data enhancement

表1 最终的图像数量

Tab.1 Number of final images

Dataset	Size	Number
Train	2 048×2 048	17 704
Test		3 071

用AdamW优化器。对于Faster R-CNN和Cascade R-CNN，初始学习率设定为0.000 1，对于RetinaNet，初始学习率设定为0.000 05。所有模型的预热迭代次数都设定为1 000，在第8次和第11次迭代时学习率递减为前学习率的0.1倍。此外，所有模型都加载了默认的预训练权重以减少训练时间。

所有实验均在Ubuntu 20.04系统上进行，GeForce RTX 3 090 ti GPU配有24 GB内存，使用编程语言Python 3.8、深度学习框架PyTorch 1.12和MMdetection框架进行实验和评估。由于TT100K中的图像均为2 048×2 048像素，不便于训练，因此使用默认的图像缩放为(1 333, 800)。此外，每个骨干模型都提供了四阶段特征输出，经过特征融合后进入分类和回归模型。

4 实验分析

4.1 结果分析

如图5所示，展示了以AFPC-T为基干模型在训练集上训练12个epoch的Loss图像，可以看到在迭代次数为25 000次左右(即12 epoch)Loss趋于稳定，之后将训练好的元架构用于测试集测试。

通过在3种元架构中添加不同模型作为骨干进行了对比实验，表2报告了在测试集上测试的实验结果，其中mAP_{50}和mAP_{75}分别表示0.5和0.75的IoU，S、M和L分别表示对应于小、中和大型物体群的mAP。从这些结果中可以看出，在所有元架构方法中，以AFPC-T为骨干的模型都优于基线模型，且FPS没有明显下降。与基线模型相比，以RetinaNet为模型的mAP_{50}的最大提升幅度约为7%。此外，其AP_{small}提高了约3%，AP_{medium}提高了约6%，AP_{large}提高了约8%。虽然RetinaNet有了显著提高，但CascadeRCNN取得了最佳结果。在只训练了12个epoch的情况下，其mAP_{50}达到了84.0%，而mAP_{75}达到了78.7%。实验结果表明，在略微降低FPS的情况下大大提高了不同物体尺寸下的mAP，在一定程度上体现了检测精度和推理速度之间的平衡。

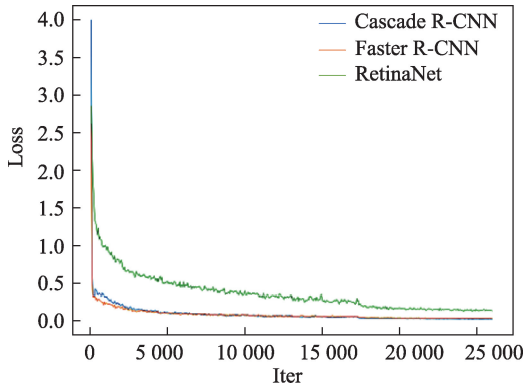


图5 以AFPC-T为基干的三元架构模型Loss图

Fig. 5 Meta-architecture Loss diagram with AFPC-T as backbone

4.2 消融分析

消融实验进一步验证AFPC-T的有效性，通过将通道注意力模块(CA)和类池连接(PC)逐一添加到基线模型中，以证明它们的效果。表3报告了消融实验的结果，+CA表示在Swin-T中添加SCAB。+CA,+PC表示在Swin-T中加入SCAB的同时加入PC。通过添加CA模块来激活更多重要维度，Faster R-CNN和Cascade R-CNN以及RetinaNet的性能得到了显著提高，尤其是在大中小型范围内。Cascade R-CNN+CA使其mAP_{50}、mAP_{75}、S、M分别从83.8%、78.5%、45.1%、74.3%提高到85.0%、79.8%、47.2%、75.1%。在FPS仅从19.8下降到18.2的情况下，CA的有效性得到了证明。

为了探索每个模块的作用，还对PC的效果进行了评估。如表3所示，PC在一定程度上提高了检

表2　性能对比

Tab.2　Performance comparison

Method	Backbone	mAP_{50}/%	mAP_{75}/%	S/%	M/%	L/%	FPS
Faster R-CNN	ResNet−50	76.9	69.8	35.5	69.8	69.5	50.2
	ResNet−101	75.4	68.7	35.6	68.1	70.6	39.6
	PVT-T	76.6	71.0	36.8	70.7	69.9	41.0
	Swin-T	78.6	73.4	37.2	71.9	74.9	37.4
	AFPC-T	80.4	75.3	40.7	73.2	75.6	35.9
RetinaNet	ResNet−50	61.1	55.5	30.8	55.3	52.7	52.7
	ResNet−101	64.3	57.4	32.4	57.9	55.6	40.4
	PVT-T	69.1	61.1	35.9	60.7	57.9	45.5
	Swin-T	67.3	60.9	34.2	61.1	58.5	37.8
	AFPC-T	71.4	63.8	35.7	63.6	63.5	36.4
Cascade R-CNN	ResNet−50	79.6	73.9	38.6	72.4	73.0	27.1
	ResNet−101	77.4	71.6	37.1	70.6	73.4	21.3
	PVT-T	83.4	77.8	44.9	73.8	75.2	19.0
	Swin-T	83.8	78.5	45.1	74.3	79.4	19.8
	AFPC-T	84.0	78.7	46.3	74.4	78.8	18.1

表3　消融对比

Tab.3　Ablation comparison

Method	Backbone	mAP_{50}/%	mAP_{75}/%	S/%	M/%	L/%	FPS
Faster R-CNN	Swin-T(Baseline)	78.6	73.4	37.2	71.9	74.9	37.4
	+CA	79.6	74.3	38.7	73.0	76.5	36.0
	+CA,+PC	80.4	75.3	40.7	73.2	75.6	35.9
RetinaNet	Swin-T(Baseline)	67.3	60.9	34.2	61.1	58.5	37.8
	+CA	69.0	62.4	35.0	61.8	64.2	36.5
	+CA,+PC	71.4	63.8	35.7	63.6	63.5	36.4
Cascade R-CNN	Swin-T(Baseline)	83.8	78.5	45.1	74.3	79.4	19.8
	+CA	85.0	79.8	47.2	75.1	78.4	18.2
	+CA,+PC	84.0	78.7	46.3	74.4	78.8	18.1

测器的性能。采用Faster R-CNN+CA+PC后，其mAP_{50}、mAP_{75}、S、M和L分别从78.6%、73.4%、37.2%、71.9%和74.9%提高到80.4%、75.3%、40.7%、73.2%和75.6%。实验结果表明，CA和PC都提高了AFPC-T的性能，而且它们的组合达到了最佳性能。为了进一步探讨PC的影响，制作了Epoch和损失之间的关系图。如图6所示，加入PC后，在相同损失范围内，训练次数略微减少，证明了PC的有效性。

4.3　可视化分析

为了探索双重注意力融合对特征的具体影响，本文对部分特征图进行了可视化处理，以便对AFPC-T进行定性检查。图7展示了3种元架构的特征可视化结果。每个元架构中的上组均为基线模型Swin-T，下组基于AFPC-T。可以看出，在这3种元架构中，AFPC-T比基线模型Swin-T能更准确地覆盖图像中的单个或多个物体，而对背景的关注较少。观察结果表明，引入通道注意力有助于AFPC-T聚焦更重要的物体。显然，通道注意力和空间注意力可以分别帮助模型更好地关注图像的重要特征和位置信息。将这两种注意力结合起来可以进一步提高模型的性能。

为了检测AFPC-T在实际交通场景中的效果，本文对部分实验结果进行了可视化展示。如图8所示，Cascade R-CNN，Faster R-CNN和RetinaNet都使用AFPC-T作为TT100K数据集上部分检测结

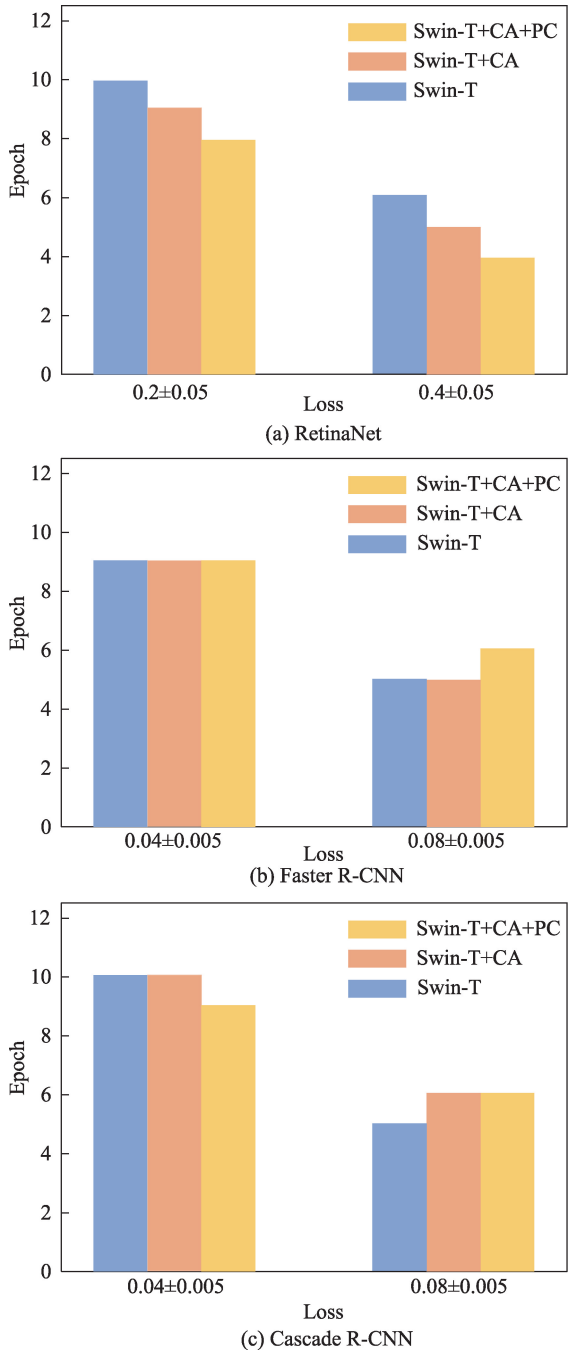


图6 Epoch与Loss之间的影响

Fig. 6 Effect between Epoch and Loss

图7 特征可视化

Fig. 7 Feature visualization

果的骨干。检测结果包括交通标志的类别和置信度。图8的放大图像部分显示，AFPC-T能够很好地适应多分类问题(42个类别)，并能准确检测小尺寸目标。

5 结论

本文通过对TT100K数据集进行扩充，对不同大小的交通标志进行检测得出以下结论。

1)本文提出一种基于Transformer交通标志检测模型。在特征提取阶段通过将全局通道注意力

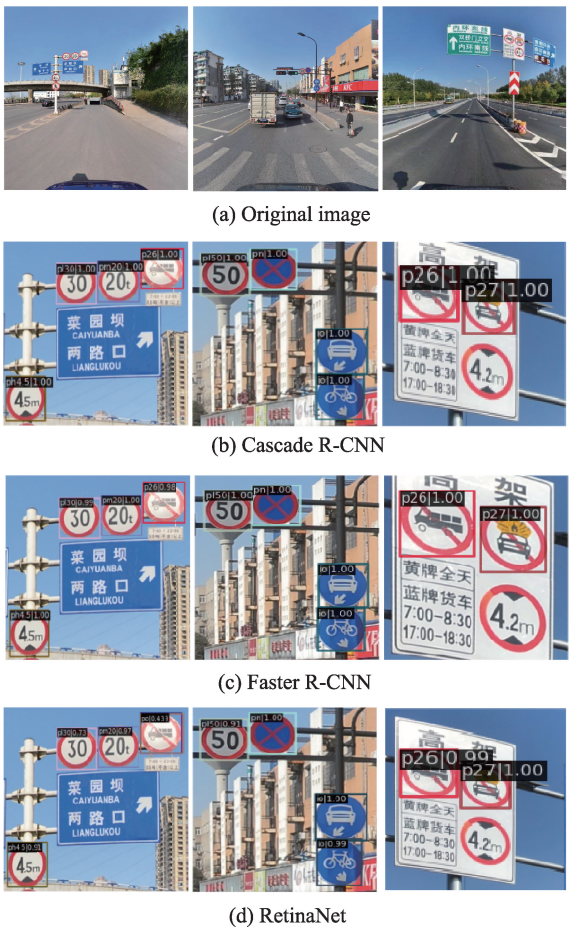


图8　检测可视化
Fig. 8　Detection visualization

引入到Transformer中，使网络学会利用全局信息，选择性地增强包含有用信息的特征，抑制不重要的特征。

2）实验结果表明，在较少epoch的训练下以AFPC-T为基干的模型均取得最高的*mAP*。其中以Cascade R-CNN为元架构的AFPC-T的mAP_{50}精度达到了84.0%，与基线模型相比提高了约7%。

参考文献：

[1] HUANG Z, YU Y, GU J, et al. An efficient method for traffic sign recognition based on extreme learning machine[J]. IEEE Transactions on Cybernetics, 2016, 47(4): 920-933.

[2] PANG Y, YUAN Y, LI X, et al. Efficient HOG human detection[J]. Signal Processing, 2011, 91(4): 773-781.

[3] QIN Z, ZHANG P, WU F, et al. Fcanet: Frequency channel attention networks[C]//Montreal: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[4] 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.

[5] CAI Z, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C]//Salt Lake: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[6] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[7] LIM J S, ASTRID M, YOON H J, et al. Small object detection using context and attention[C]//Jeju Island: 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2021.

[8] CHEN Y, WANG J, DONG Z, et al. An attention based YOLOv5 network for small traffic sign recognition[C]//Anchorage: 2022 IEEE 31st International Symposium on Industrial Electronics (ISIE), 2022.

[9] CHU X, TIAN Z, WANG Y, et al. Twins: Revisiting the design of spatial attention in vision transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 9355-9366.

[10] HUANG G, WANG Y, LYU K, et al. Glance and focus networks for dynamic visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 4605-4621.

[11] CHEN X, WANG X, ZHOU J, et al. Activating more pixels in image super-resolution transformer[C]//Vancouver: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[12] WANG W H, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Montreal: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

通信作者：严丽平（1980—），女，副教授，博士，硕士生导师，研究方向为智能交通、人工智能。E-mail: csyanliping@163.com。

（**责任编辑：**吴海燕）