激光与光电子学进展 先进成像

全局采样空间注意力机制及其在图像分类与小目标检测识别中的应用

卢镜宇[1,2,3]，张海洋[1,2,3*]，王文鑫[1,2,3]，赵长明[1,2,3]

[1]北京理工大学光电学院，北京 100081；

[2]光电成像技术与系统教育部重点实验室，北京 100081；

[3]信息光子技术工业和信息化部重点实验室，北京 100081

摘要 注意力机制的出现和应用在一定程度上改善了神经网络对全局信息应用不足的缺陷，但常见的注意力机制模块也同样存在感受野小无法关注全局信息的问题，而某些全局注意力机制模块则计算成本过高。为此，提出一种基于卷积、池化、对比方法的轻量化注意力模块，即全局采样空间注意力模块。对于深度网络推理过程中部分模块输出的中间特征图，该注意力模块通过对比差值的形式获取所需要的空间注意力图。全局采样空间注意力模块是一种轻量化的通用模块，能够直接置入卷积神经网络中，增加的成本几乎可以忽略不计，并且其能够与网络一同进行端到端训练。主要在随机抽取的部分 ImageNet-1K 数据集和团队自制的“低慢小”无人机数据集中对模块进行了验证。实验结果显示，相比其他模块，所提模块在图像分类和小目标检测识别任务中具备 1 百分点～3 百分点的性能提升效果，证明了所提模块的性能与其在小目标检测方面的适用性。

关键词 注意力机制；全局采样；轻量化；图像分类；小目标探测

中图分类号 TP751.1 **文献标志码** A **DOI**：10.3788/LOP231933

Global-Sampling Spatial-Attention Module and its Application in Image Classification and Small Object Detection and Recognition

Lu Jingyu[1,2,3], Zhang Haiyang[1,2,3*], Wang Wenxin[1,2,3], Zhao Changming[1,2,3]

[1]*School of Optics and Photonics, Beijing Institute of Technology, Beijing* 100081;

[2]*Key Laboratory of Optoelectronic Imaging Technology and Systems, Ministry of Education, Beijing* 100081;

[3]*Key Laboratory of Information Photonics Technology, Ministry of Industry and Information Technology, Beijing* 100081

Abstract The emergence and application of attention mechanisms have addressed some limitations of neural networks concerning the utilization of global information. However, common attention modules face issues with the receptive field being too small to focus on overall information. Moreover, existing global attention modules tend to incur high computational costs. To address these challenges, a lightweight, universal attention module, termed “global-sampling spatial-attention module”, is introduced herein based on convolution, pooling, and comparison methods. This module relies on the comparison methods to derive spatial-attention maps for intermediate feature maps generated during deep network inference. Moreover, this module can be directly integrated into convolutional neural networks with minimal costs and can be end-to-end trained with the networks. The introduced module was primarily validated using a randomly selected subset of the ImageNet-1K dataset and a proprietary low-slow-small drone dataset. Experimental results show that compared with other modules, this module exhibits an improvement of approximately 1–3 percentage points in tasks related to image classification and small object detection and recognition. These findings underscore the efficacy of the proposed module and its applicability in small object detection.

Key words attention mechanism; global sampling; lightweight; image classification; small target detection

收稿日期：2023-08-18；**修回日期**：2023-09-14；**录用日期**：2023-10-09；**网络首发日期**：2023-11-07

通信作者：*ocean@bit.edu.cn

1　引　　言

深度学习是一系列计算机视觉任务的重要方法。深度网络作为深度学习的重要工具，拥有数据导向特性和大范围平行计算能力，在众多领域中拥有出色的表现[1]。其中使用频繁的卷积神经网络（CNN）在不同任务中的表现优劣除了取决于网络自身属性（深度、宽度及超参数的制定），还依赖于网络结构以外的算法设计（优化方法、训练策略等）。VGGNet[2]、ResNet[3]、EfficientNet[4]、DarkNet53[5]等一系列网络结构是网络自身属性方面的创新；Adam优化算法[6]、SGDR算法[7]、余弦退火重启算法[8]则是网络后处理层面的改进。

此外，在处理信息量较大时，研究人员往往更加关心其中某些独特的信息，这种关注程度的不同在深度学习中使用“注意力”来描述[9]。为了能够更好地将CNN的重心转移到正确的位置，研究者们引入了注意力机制模块。注意力机制模块最先应用于机器翻译领域，只使用注意力机制模块的网络结构Transformer能够得到优秀的翻译效果[10]。二维形式的注意力机制BAM[1]与CBAM[11]通过对特征图进行额外卷积处理，分别获取空间注意力图和通道注意力图；可变形卷积网络[12]通过自适应的异形卷积实现对更大或更小范围内像素的感知，从而提升注意力效果；非局部神经网络[13]通过对视频特征图序列中每一幅特征图上每一点数值与其他所有点数值进行对比，获取真正意义上的全局注意力；图像中的Transformer[14]将图像输入不同的网络分支，根据各分支结果计算像素之间的联系；Coordinate Attention[15]将位置信息嵌入通道注意力，扩充了注意力的来源，避免了更大的计算开销。

大部分注意力机制模块在CNN原有基础上扩充了图像各点的关注范围，但仍然属于局部注意力，没有很好地结合图像全局；而非局部注意力机制尽管实现了全局注意力分配，过程中大量的矩阵运算又带来了很高的计算成本。为此，本文提出了一种新的注意力模块，即全局采样空间注意力模块（GSSAM）。该模块利用池化操作较为均匀地在特征图整体中抽取像素值并以此为基准生成注意力图。与现有方法相比，GSSAM有3项优势：注意力图的生成涉及特征图整体，其能够较好地考虑整体信息；使用池化采样的方式提取整体信息，大幅度减少了全局注意力图生成过程中的运算量；注意力模块由卷积层、池化层和基础运算组成，能够直接参与训练，无须进行额外处理。在部分ImageNet-1K数据集上进行的图像分类实验表明，多种经典网络结构在插入所提GSSAM后，网络都能够获得一定分类精度的提升，并且在插入网络结构前后，网络参数量、运算时间的增加相较于基础量而言基本可以忽略。同时，实验利用“低慢小”无人机数据集对附加当前性能较为优秀的YOLOx目标检测识别网络[16]前后的GSSAM进行训练与验证，最终证实了GSSAM在小目标感知方面具备优秀的性能。

2　GSSAM算法原理

GSSAM通过对输入模块的特征图进行卷积、池化、对比处理，分别完成特征图中像素点感受野的扩张、特征图全局范围内基准点的选取、特征图所有点与基准点之间的求差操作，是在借鉴BAM与CBAM形式、非局部神经网络、可变形卷积思想的基础上，考虑操作可行性、轻量化设计得到的注意力模块。

2.1　参考网络

BAM[1]与CBAM[11]是两种相似性较高的注意力模块，结构如图1所示。参考两种模块结构，GSSAM采用类似的形式，在网络主干外并行附加一系列操作生成注意力图，然后将其与原本特征图相乘，得到附带注意力权重的输出特征图。

非局部神经网络[13]形式如图2所示，大方框部分表示获取每个点对之间关联的过程。GSSAM在该网络的基础上，使用采样的方式获取全局信息，通过各点

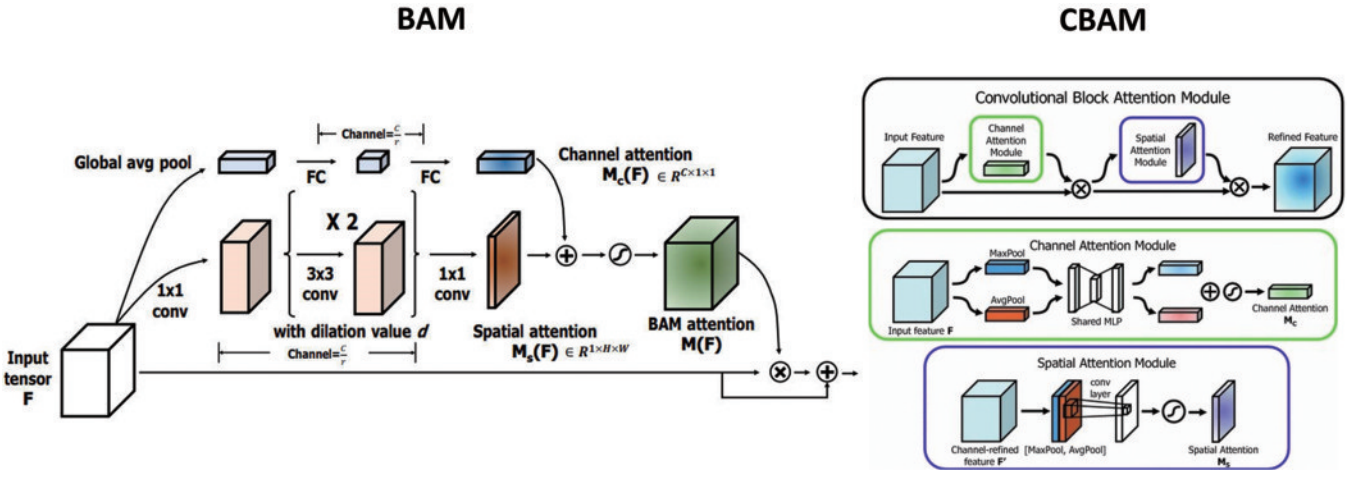

图1　BAM结构[1]与CBAM结构[11]

Fig. 1　BAM architecture[1] and CBAM architecture[11]

与全局信息对比的方式充分考虑了特征图上每个点与图像全局的关系，实现了全局性的注意力机制，并尽可能地降低了计算成本。

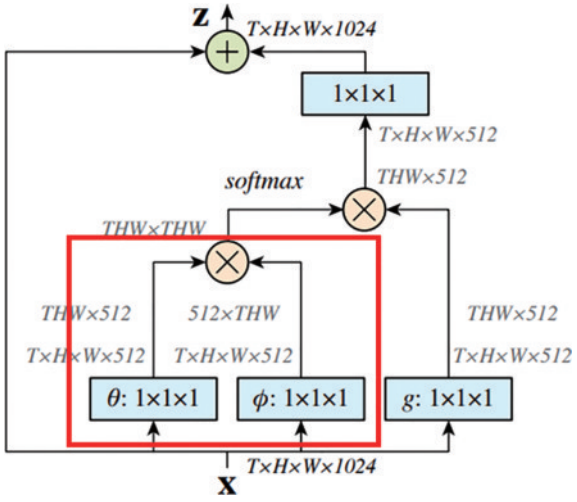

图 2　非局部神经网络注意力模块的结构[13]

Fig. 2　Architecture of non-local neural network attention module[13]

可变形卷积网络[12]使用的可变形卷积模块结构如图 3 所示。GSSAM 参考其自适应选择信息的形式筛选图像基准点，将其表示背景信息，更加针对性实现对单幅图像的背景环境信息提取。

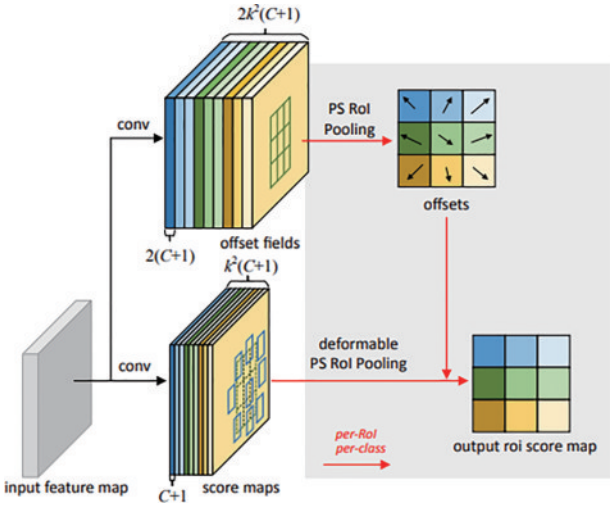

图 3　3×3 可变形卷积模块的结构[12]

Fig. 3　Architecture of 3×3 deformable convolution module[12]

2.2　GSSAM

在综合考虑上述参考注意力机制优势与缺陷的基础上，GSSAM 通过在输入模块的特征图全局范围内进行卷积和采样，并与图上各点进行差值对比来获取注意力图。模块先通过 1×1 卷积进行特征图降维，再使用多个 3×3 卷积实现信息收集，利用最大池化完成采样筛选，获取具有全局代表性的特征点，最终将这些特征点与所有像素点进行对比，根据对比结果差异生成注意力图，以乘法的方式在输入特征图上分配计算权重。

GSSAM 考虑了注意力来源的全局性、自适应性与轻量化，在对图像全局进行考察的同时保证了较低的运算成本。由于本文设计 GSSAM 的原始目标是尝试通过注意力机制在图像中区分背景与目标，从而将网络的关注重心向目标偏移，减弱背景对图像分类的影响，因此 GSSAM 是一种空间注意力模块，暂时没有附加通道注意力机制。GSSAM 的详细结构如图 4 所示。

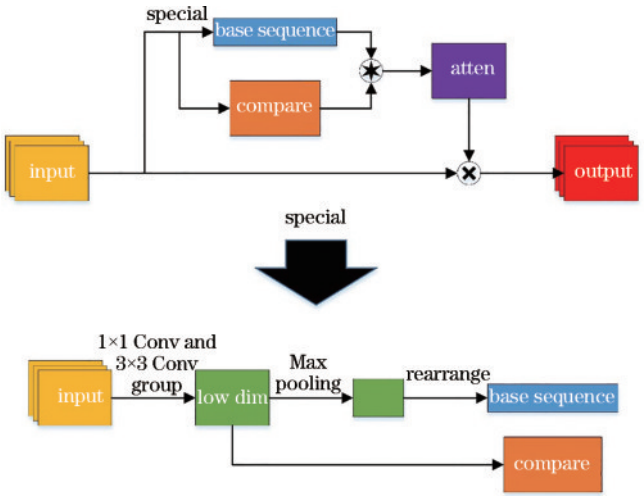

图 4　GSSAM 的详细结构

Fig. 4　Detailed architecture of the GSSAM

记输入 GSSAM 的特征图为 $\boldsymbol{F}\in\mathbb{R}^{C\times H\times W}$，$H$、$W$、$C$ 分别为特征图长宽尺寸以及通道数量。针对该输入特征图，GSSAM 将生成一个对应的二维注意力图 $\boldsymbol{A}\in\mathbb{R}^{H\times W}$，该注意力图长宽尺寸与输入特征图一致。带有注意力权重的特征图 $\boldsymbol{F}_{\text{atten}}\in\mathbb{R}^{C\times H\times W}$ 的计算公式为

$$\boldsymbol{F}_{\text{atten}}=\boldsymbol{F}\otimes\boldsymbol{A},\tag{1}$$

式中：$\otimes$ 代表二维注意力图中每一个位置的权重与三维特征图中各个通道上对应位置像素值直接相乘的操作。

GSSAM 的核心思想是：找出特征图中与其他像素点相比更加突出的点集，赋予其大注意力权重，使得网络关注重心向这些点偏移。而在图像输入到网络结构再输出的一系列过程中，特征图往往具备通道数多的特征，想要对比两个具备多重数值的像素点比较困难，并且容易导致计算量的大幅度上升。为此，GSSAM 使用一个 1×1 卷积对输入特征图进行降维处理，在尽可能保证原有通道信息不丢失或少丢失的情况下，将特征图从高维降到二维，即特征图每个像素只对应一个数值，如图 5 所示。该“单层”特征图在模块中代表输入特征图的信息总和，其上各点数值作为输入特征图对应像素点的代表参与后续运算，其经过一系列不改变特征图大小的 3×3 卷积得到新特征图，将新特征图称为待对比图。待对比图的点相较于

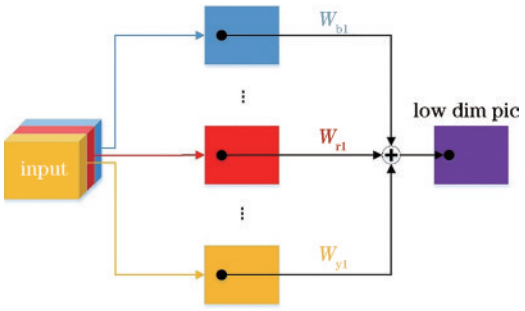

图 5　输入特征图降维

Fig. 5　Dimension reduction of input feature map

输入特征图上的点集中了部分“环境信息”，作为参与全局对比的对象，能够更好地体现输入图像中全局与局部的差异。

为了实现图像全局的注意力，非局部神经网络为中间特征图上所有点对之间建立联系，而 GSSAM 则采取最大池化的形式，将输入模块的特征图分块，从每一个像素块中取出最具有代表性的像素提供给网络。为便于后续论述，将这些取出的像素点称为基准点。

最大池化过程筛选出的是当前特征图中像素值最大的像素点，但输入模块的特征图存在的两个不利因素影响了基准点的选取：1）输入特征图通道数多，只是用最大池化得到的结果将是通道数不变的特征图，并且其上每一点的各个像素值很可能来自输入特征图中不同的像素；2）注意力应当来源于输入特征图，但最大池化操作实质上是在之前网络的基础上进行的筛选，其不能很好地代表输入特征图自身像素点之间的信息。为此，选择直接在待对比图的基础上进行池化操作，以实现降维与信息整合的目的。待对比图获取过程中使用多层卷积，通过加深卷积层次实现感受野的放大，避免了大卷积核的直接使用导致的参数量增加。各层卷积之间增加 BN 层和 ReLU 激活函数，用于提升训练效率，增强网络表达能力和增加网络稀疏性[17]。基准点选取过程如图 6 所示。

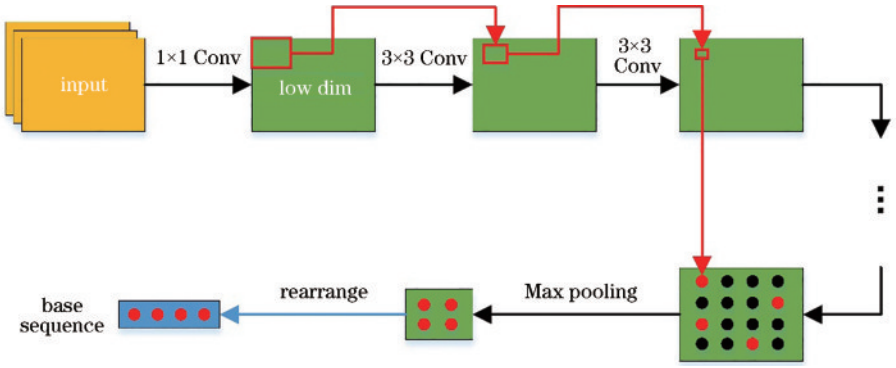

图 6　基准点选取过程示意图

Fig. 6　Schematic of benchmark selection process

基准点的选取过程在实质上与可变形卷积神经网络中的可变形卷积操作类似，都是通过对图像信息进行搜集、使用来调整用于生成注意力的图像范围的。但可变形卷积是小范围内的像素位置调整过程，且由于其使用了 RoI 池化操作生成卷积时每个位置的偏移量，因此无法直接进行网络训练。而 GSSAM 采用卷积加最大池化的形式，能够在图像全局范围内采集需要像素的同时保证深度网络训练的顺利进行，可以直接插入任何已有的卷积神经网络中。GSSAM 选取基准点时，所有卷积层带来的感受野范围应当大于或等于最大池化使用的算子大小，以保证池化时每一个像素块中各个像素能够充分考虑像素块其他像素的情况。

GSSAM 将获取的基准点排列成一个 $1\times1\times N$ 的向量（N 代表基准点个数），该向量与待对比图相减，通过 Python 软件提供的传播机制，形成 $H\times W\times N$ 的三维矩阵。$H\times W$ 个像素各自包含的 N 个像素值代表待对比图中对应像素与各个基准点之间的差值。对该向量以第三维为轴进行绝对值求和，得到大小为 $H\times W$ 的二维矩阵，各点像素值表征该点在特征图整体情况中的突出程度，数值越大，该点越显著，越有可能是网络需要关注的位置。而数值较小的点通常属于图中的背景或是物体中不重要的部分。由于在网络训练过程中，一个像素数值的剧烈提升（上述差值对比结果中最小值和最大值的差距往往远大于输入特征图本身）很可能会导致训练过程中的梯度爆炸，因此 GSSAM 对该二维矩阵进行归一化操作，将其中数值限制在 0 和 1 之间，得到最终输出的注意力图。该过程在图 4 中以带圈的星号描述，具体细节如图 7 所示。

综上所述，GSSAM 首先对输入特征图进行多重卷积处理，实现降维与信息融合，再将最后一层卷积获得的特征图作为待对比图，并从待对比图中以最大池化的形式选择合适的样本点，最终将样本点与图上各点进行对比，获取最终的注意力图。该过程可以描述为

$$\boldsymbol{c}=f_n^{3\times3}(\boldsymbol{F}),\tag{2}$$

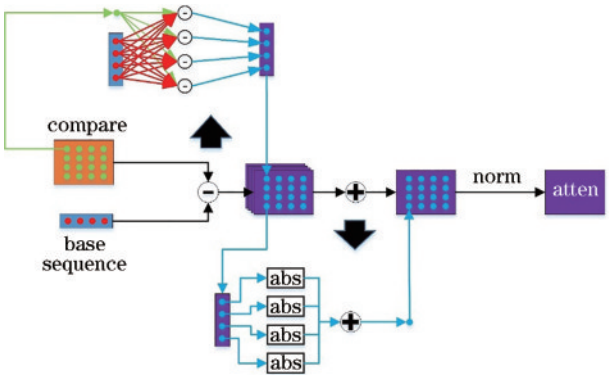

图 7　差值对比获取注意力图的过程

Fig. 7　Process of obtaining attention map by difference comparison

$$\boldsymbol{A}=\mathrm{norm}\left\{\mathrm{sum}_2\left\{\mathrm{abs}\left\{\boldsymbol{c}-\mathrm{flatten}\left[\mathrm{Max\ pooling}^{r\times r}(\boldsymbol{c})\right]\right\}\right\}\right\},\tag{3}$$

式中：norm(·)代表对二维矩阵范围内所有元素进行整体的归一化；$\mathrm{sum}_k(\cdot)(k\in 0,1,2)$代表以第 k 维为轴对向量进行绝对值求和操作；abs(·)代表向量中对所有元素取绝对值；$f_n^{s\times s}(\cdot)$代表卷积次数为 n、卷积核大小为 $s\times s$ 且对于任何通道数的输入特征图，输出特征图通道数都为 1 的卷积操作；flatten(·)代表将一个二维向量排列成一维向量的操作；$\mathrm{Max\ pooling}^{r\times r}(\cdot)$代表算子大小为 $r\times r$ 的最大池化操作。其中，n 与 r 的数值根据网络中输入 GSSAM 的特征图情况决定。

以大小为 $H\times W$ 的特征图为例，全局逐点对比所需运算次数为 $(H\times W)^2$，对比结果占用空间同样为 $(H\times W)^2$，此时需要对每一点对之间的区别进行保存，但当特征图较大时很容易导致运算速度过慢、存储空间不足等问题。且使用 GPU 进行并行计算的过程中，大量的显存占用会妨碍多线程的运行，影响运算速度。而全局采样性质的对比所需运算次数与占用空间都为 $(H\times W)^2/k_1^{\ 2}$，k_1 代表采样步长，当特征图较大时选取较大的 k_1 值就能够有效减少运算量和减小存储空间所需占比，提升算法实时性。

得到的注意力图各点数值代表了网络需要向输入特征图各个位置提供的注意力，使得网络能够将计算重心向更有可能是物体而非背景的位置移动，消除背景导致的某些不利影响。利用式(2)和式(3)计算出输入特征图对应的注意力图后，利用式(1)将输入特征图与注意力图相乘，即可获得 GSSAM 的输出特征图。

3　实验结果分析与讨论

为验证 GSSAM 自适应分配注意力权重带来的效果，在多种经典图像分类网络以及当前较为先进的 YOLOx 目标检测识别网络的基础上进行了实验。实验在统一的硬件和软件环境条件下进行，显卡型号为 RTX3090，CPU 型号为 i9-12900K，PyCharm 版本为 PyCharm Community Edition 2022.1，Python 版本为 3.8，PyTorch 版本为 1.11.0。

为了能够较为全面地说明 GSSAM 在计算机视觉任务中起到的作用，本课题组将 ImageNet-1K 数据集中抽取的部分数据和团队自制的“低慢小”无人机数据集用于实验。首先将模块分别置入不同的基础网络中，在 ImageNet-1K 数据集上进行图像分类，以分类准确率来表现模块在不同图像分类网络中的效果，说明 GSSAM 在基础网络中广泛的适用性；随后利用 Grad-CAM 方法[18]将网络中最终输出的特征图权重可视化，说明 GSSAM 注意力分配的结果；最后将 GSSAM 嵌入 YOLOx 目标识别检测网络，在“低慢小”无人机数据集中开展目标检测识别实验，说明 GSSAM 在小目标检测方面的性能。

3.1　ImageNet-1K 图像分类

实验选用 ImageNet-1K 作为体现 GSSAM 作用的数据集。ImageNet-1K 是一种在图像分类任务中常用的数据集，包含图像类别 1000 种，每种类别的训练集数量超过 1000 张，验证集和测试集图片各有 50 张，是大规模视觉识别挑战赛(ILSVRC)2012—2017 图像分类和定位数据集，因此也被称为 ImageNet ILSVRC 数据集。受到计算资源和训练时间的限制，本实验仅从数据集中随机抽取一部分类别进行图像分类，目前抽取类别数为 10。由于抽取的随机性，数据集的大幅度缩减一般不影响实验结果的普适性。

为了能够充分体现 GSSAM 在卷积神经网络中具备的广泛适用性，选取 VGG19 网络、一系列 ResNet 残差网络、DarkNet53 深度网络作为基础网络进行测试。分类准确率用测试集图片中分类正确图片的占比表示。测试中，基础网络、在基础网络附加不同模块(CBAM 和 GSSAM)得到的网络都在同样的训练集上进行 1000 个 epoch 的训练，每经历 10 epoch 进行一次验证，最终选取在验证集上分类准确率最高的模型作为输出模型，将输出模型在测试集得到的分类准确率作为最后的评判指标。

以 VGG19 网络为例，图 8～10 展示了原始网络、附加 CBAM 的网络、附加 GSSAM 的网络 3 种结构在整体训练过程中的损失、分类准确率变化情况。

从图 8 能够看出，随着训练次数的增加，3 种网络对训练集的处理效果逐渐上升，最终都达到了几乎相同的损失与准确率。对于在训练网络过程中使用的数据集，训练集本身准确率不具备过高参考价值，因此在不考虑训练时长的前提下，训练集的验证结果不具备太大的参考价值。图 9 与图 10 分别体现了 3 种不同网络在验证集中实时的作用效果变化。由于每 10 个 epoch 对原始数据进行一次验证，且验证集相较于训练集对网络而言存在较大不确定性，因此获取的结果离散性较强，很难直观看出网络之间的差异，如图 9(a)

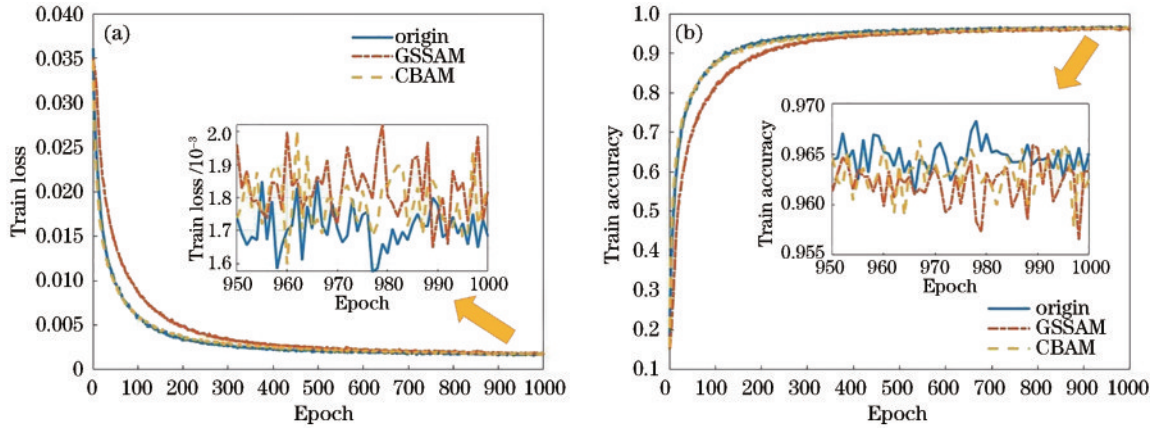

图 8　训练集上损失和准确率变化。(a) 损失变化;(b) 准确率变化

Fig. 8　Loss and accuracy changes on training set. (a) Loss change; (b) accuracy change

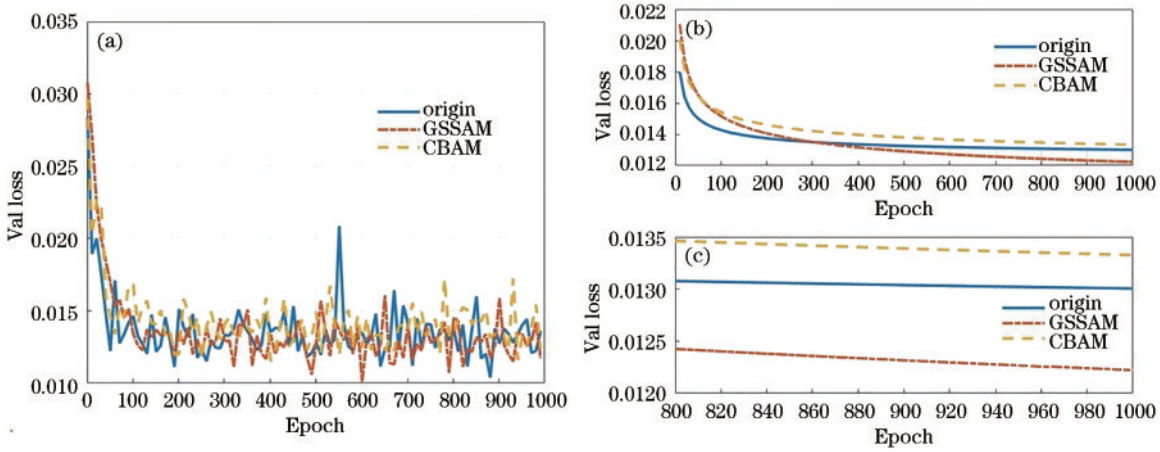

图 9　验证集上不同网络的损失变化。(a) 原有损失变化;(b)损失拟合;(c) 局部对比

Fig. 9　Loss change on validation set. (a) Original loss change; (b) loss fitting; (c) local comparison

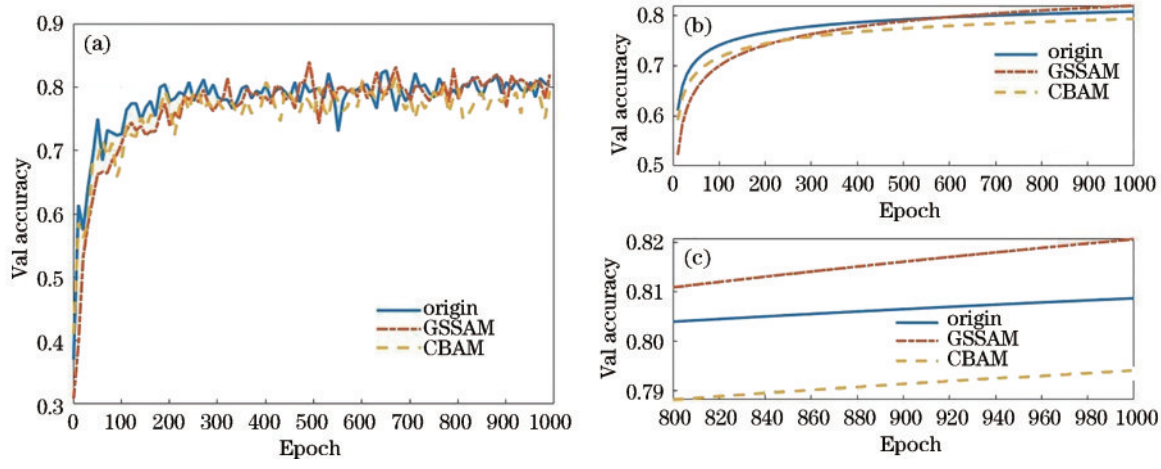

图 10　验证集上不同网络的准确率变化。(a) 原有准确率变化;(b)准确率拟合;(c) 局部对比

Fig. 10　Accuracy change on validation set. (a) Original accuracy change; (b) accuracy fitting; (c) local comparison

和图 10(a)所示。为此,对结果中的离散点进行指数形式拟合,最终在损失与准确率两方面各自得到了 3 条较为平滑的曲线。能够发现:置入 GSSAM 的 VGG19 在训练 400 个 epoch 后,在验证集中得到的损失与准确率都已经超越了原始网络与置入 CBAM 的网络;较为简单的 CBAM 反而带来了一定的负面效果,即损失曲线高于原始网络,准确率曲线低于原始网络,从一定程度上证明了 GSSAM 为网络带来了更好的图像分类效果和泛化性能。

图 11 与图 12 展示了 VGG19 与 ResNet50 两种不同网络结构在原始网络、置入 CBAM 与置入 GSSAM 情况下在部分 ImageNet-1K 数据集中验证集上的准确

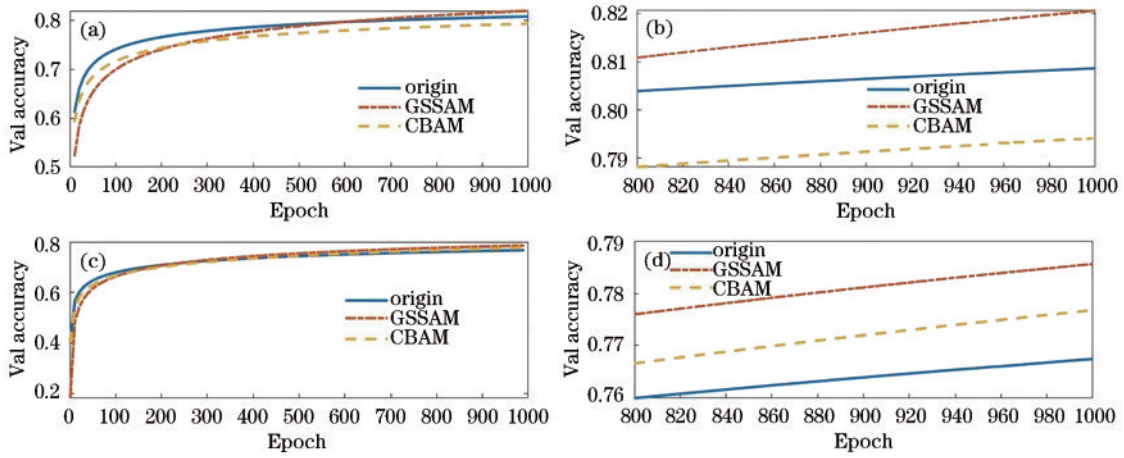

图 11　不同网络的准确率拟合曲线对比。(a) VGG19 拟合曲线；(b) VGG19 局部拟合曲线；(c) ResNet50 拟合曲线；(d) ResNet50 局部拟合曲线

Fig. 11　Comparison of accuracy fitting curves of different networks. (a) VGG19 fitting curves; (b) VGG19 local fitting curves; (c) ResNet50 fitting curves; (d) ResNet50 local fitting curves

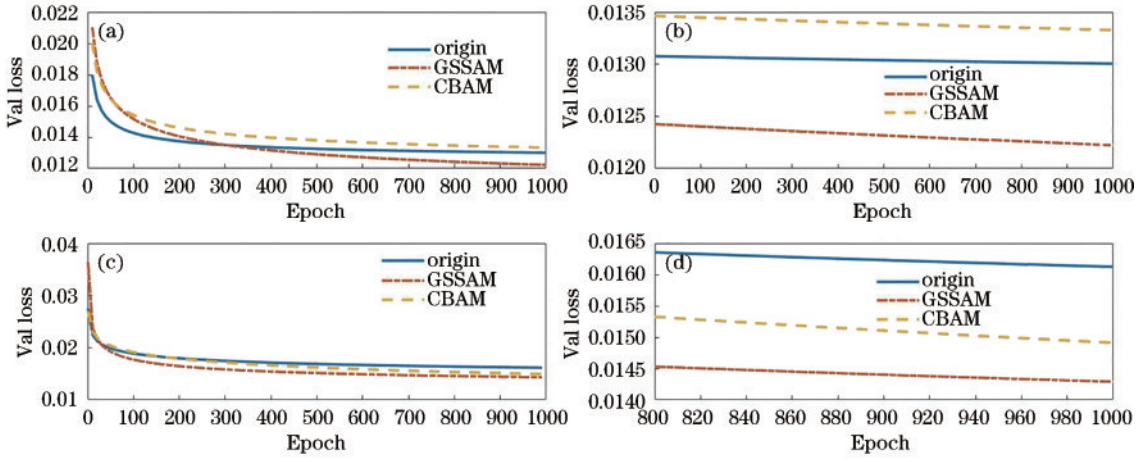

图 12　不同网络的损失拟合曲线对比。(a) VGG19 拟合曲线；(b) VGG19 局部拟合曲线；(c) ResNet50 拟合曲线；(d) ResNet50 局部拟合曲线

Fig. 12　Comparison of loss fitting curves of different networks. (a) VGG19 fitting curves; (b) VGG19 local fitting curves; (c) ResNet50 fitting curves; (d) ResNet50 local fitting curves

率和损失变化情况，在经过一定次数的训练后，相比其他网络，置入 GSSAM 的两种网络的准确率与损失都更好。

除了对比不同网络、同一网络增加 CBAM 以及 GSSAM 前后分类准确率的差异外，还需要对比各自网络模型大小和浮点运算数计算量（FLOPs），用以判断网络模型的复杂程度，证明 GSSAM 的轻量化特征。表 1 总结了在图像分类测试中各个网络结构的具体表现。

表 1 证实了 GSSAM 在图像分类任务中具备的有效性。相较于原始网络，附加了 GSSAM 的网络在通常情况下能够获得更加良好的图像分类结果，在原始网络基础上有 1 百分点～2 百分点的分类准确率提升。相较于 CBAM，GSSAM 带来的准确率提升在部分原始网络中同样具备一定的优势，并且尽管 CBAM 的结构、操作流程相比 GSSAM 更加简单，在总参数量方面 GSSAM 反而更加具有竞争力。在 FLOPs 方面，由于 GSSAM 操作过程中包含有更多的传播操作（二维特征图与一维向量相减时 Python 自动运行的特征图复制、向量复制），因此存在一定的不足，约有 0.01×10^9 至 0.03×10^9 的增加，可能会导致网络推理时间的少许增加，但与网络自身 FLOPs 基数相比，增加的时间成本几乎可以忽略。

从实验结果中能够得出结论：GSSAM 作为一种与 CBAM 类似的网络模块，能够非常简便地插入到多种用于实现图像分类任务的基础卷积神经网络中，能直接进行训练并且能够带来一定程度的分类准确率提升。在具备较为稳定的提升分类准确率的能力同时，GSSAM 所带来的网络总参数量和 FLOPs 增长并不明显，实现了模块的轻量化。

表 1 部分 ImageNet-1K 数据集上不同网络对图像的分类结果

Table1 Classification results of different networks for images on some ImageNet-1K data sets

Model	Params	FLOPs /10^9	Accuracy /%
ResNet34	25,856,586	11.26	80.00
ResNet34+CBAM	25,908,788	11.26	76.80
ResNet34+GSSAM	25,864,728	11.27	**81.00**
ResNet50	23,528,522	5.19	78.20
ResNet50+CBAM	24,225,844	5.19	**80.80**
ResNet50+GSSAM	23,551,292	5.21	79.80
ResNet101	42,520,650	10.05	79.60
ResNet101+CBAM	43,217,972	10.05	**80.00**
ResNet101+GSSAM	42,543,456	10.07	79.40
ResNet152	58,164,298	14.85	77.80
ResNet152+CBAM	58,861,620	14.85	77.20
ResNet152+GSSAM	58,187,068	14.87	**79.40**
DarkNet53	40,613,034	8.99	80.00
DarkNet53+CBAM	40,788,116	9.00	81.20
DarkNet53+GSSAM	40,624,742	9.01	**81.60**
VGG19	139,622,218	23.02	82.60
VGG19+CBAM	139,698,996	23.03	81.80
VGG19+GSSAM	139,632,920	23.04	**84.00**

3.2 Grad-CAM 实现网络权重可视化

Grad-CAM 算法[18]是在 CAM 算法[19]的基础上设计的一种用于评估深度网络在图像分类过程中在输入图像各个区域对图像分类做出贡献大小的算法。该算法利用一幅图像输入网络后在网络中生成的最后一幅特征图(即最后一层卷积层输出的结果),以权重形式附加图像分类正确时网络后向传播过程中为最后一个卷积层提供的特征图梯度,再进行归一化和 ReLU 函数激活,输出用于表征在分类过程中输入图片各个区域受重视程度的类激活图。将类激活图叠加到原始输入图像中,就能够将网络更加“重视”的区域以高亮的形式表现出来。

为了能够更好地观察具有 GSSAM 的深度网络在图像分类任务中寻找有效物体、排除背景的效果,使用 Grad-CAM 算法对 VGG19、附加了 GSSAM 的 VGG19、附加了 CBAM 的 VGG19 处理的结果进行可视化操作。从 10 种类别的测试集中分别随机选取一幅被三种网络同时识别正确的图像,将图像输入三种网络后获取类激活图,将原始图像与类激活图相叠加得到所需图像。三种网络针对不同类别物体生成的不同结果如图 13 所示。

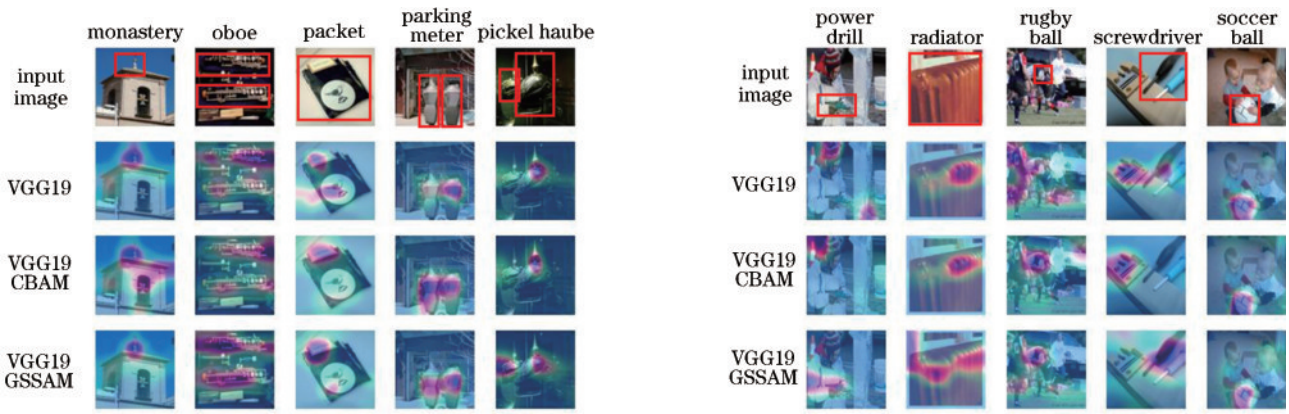

图 13 Grad-CAM 可视化结果对比

Fig. 13 Comparison of Grad-CAM's visualization results

当输入图像中目标类别物体体积较小、所占像素数量较少时,原始网络存在无法准确定位合适的关注点、关注重心向背景乃至其他物体偏移的情况,例如目标物体为橄榄球时,原始网络生成的高亮区域(即网络关注重心)包含了极大一部分属于人的图像区域,而橄榄球只占其中的小部分;当图像中目标类别数量大于 1 时,原始网络甚至是附加了 CBAM 的网络在部分图像中存在无法完全锁定每一个目标类别物体所在区域的问题,例如在黑管类别的图像中,原始网络和附加 CBAM 的网络的重心大多数集中于上方的黑管而忽视了下方的黑管。相较于原始网络,在大部分输入图像中,GSSAM 都能够让网络在图像中更加容易地找到目标类别物体所在的区域,特别是部分小目标物体,从而使得图像分类的准确率得到一定程度的提升。而相比附加 CBAM 的网络,附加 GSSAM 的网络也能够在很多图像中表现出小目标和多目标定位、注意力权重赋予的优势,具备良好的应用价值。

3.3 GSSAM 在可见光图像中的小目标感知应用

目前较为常见的针对小目标的检测识别主要通过图像低级边缘轮廓信息与高级语义信息相结合来实现。例如:2021 年提出的非对称语境调制模块(ACM)[20]使用高级信息调制低级注意力、低级信息调制高级注意力的方式实现更加优秀的红外弱小目标检测;DCP[21]将可变形卷积与图像金字塔相结合,构建了用于水下小目标探测的 USTD 网络结构;同年,SSD 与特征扩散的方式相结合被证实也能够提升对可见光小目标的检测识别准确率[22];GSSAM 的全局对比在实质上同样实现了图像全局高级信息与局部低级信息的融合。

由于GSSAM对特征图的处理包含了一系列卷积操作，再由池化操作获取多个基准点，因此基准点包含的信息实质上来源于特征图中的一个个像素块，即基准点序列代表了特征图上多个局部区域的信息。对特征图上各点与这些被选取的区域信息进行插值对比，得到各点与绝大部分区域的相似性值。由此可以推测，当目标图像具备背景重复度较高、目标较小的特性时，GSSAM同样具备出色的性能。

图14描述了GSSAM对包含小目标、背景重复图像作用的过程。其中：黄色区域代表在推理过程中GSSAM获取的基准点覆盖的感受野，称为基准区域；红色区域是目标所在区域；绿色区域代表背景区域，即与基准区域类似的区域；三者指向的圆点代表区域经过卷积形成的特征图像素。

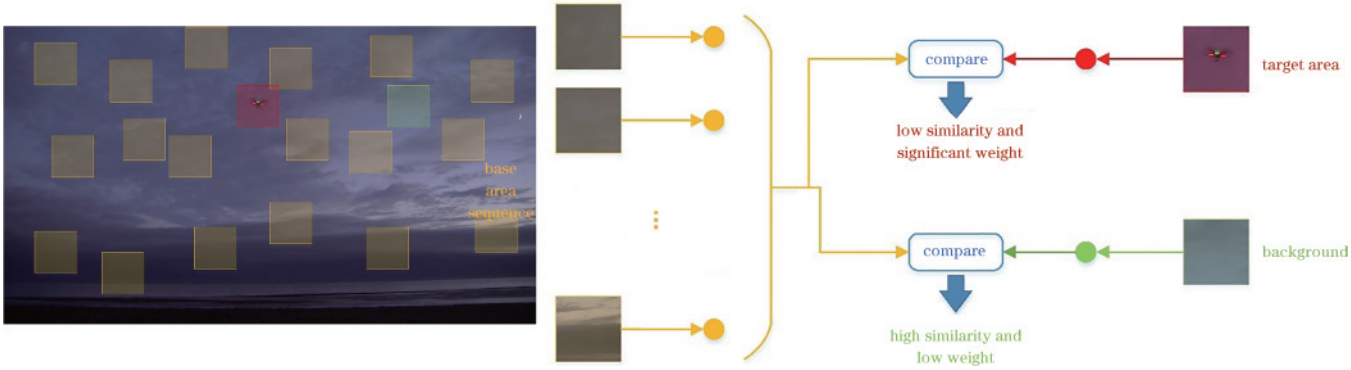

图14　GSSAM在小目标图像中作用的示意图

Fig. 14　Schematic of the role of GSSAM in small target images

为了验证上述推测，将GSSAM与团队长期研究的“低慢小”无人机目标检测识别领域的相关技术相结合并进行了实验。实验使用YOLOx目标检测识别网络[16]作为基础网络，由于该网络在目标识别检测领域已经取得了优秀的成果，具备较高的网络复杂度，且其识别精度、准确率都十分优秀，因此相较于在图像分类中使用的VGG19、ResNet等简单网络，YOLOx作为基础网络进行对比实验能够更好地展示GSSAM在小目标检测方面的性能。图15描述了实验中YOLOx与GSSAM结合的方式。GSSAM1到GSSAM5是5个输入特征图尺寸不同的GSSAM，将它们分别放置到YOLOx原有的5个深度模块前后，形成模块与模块之间的注意力。

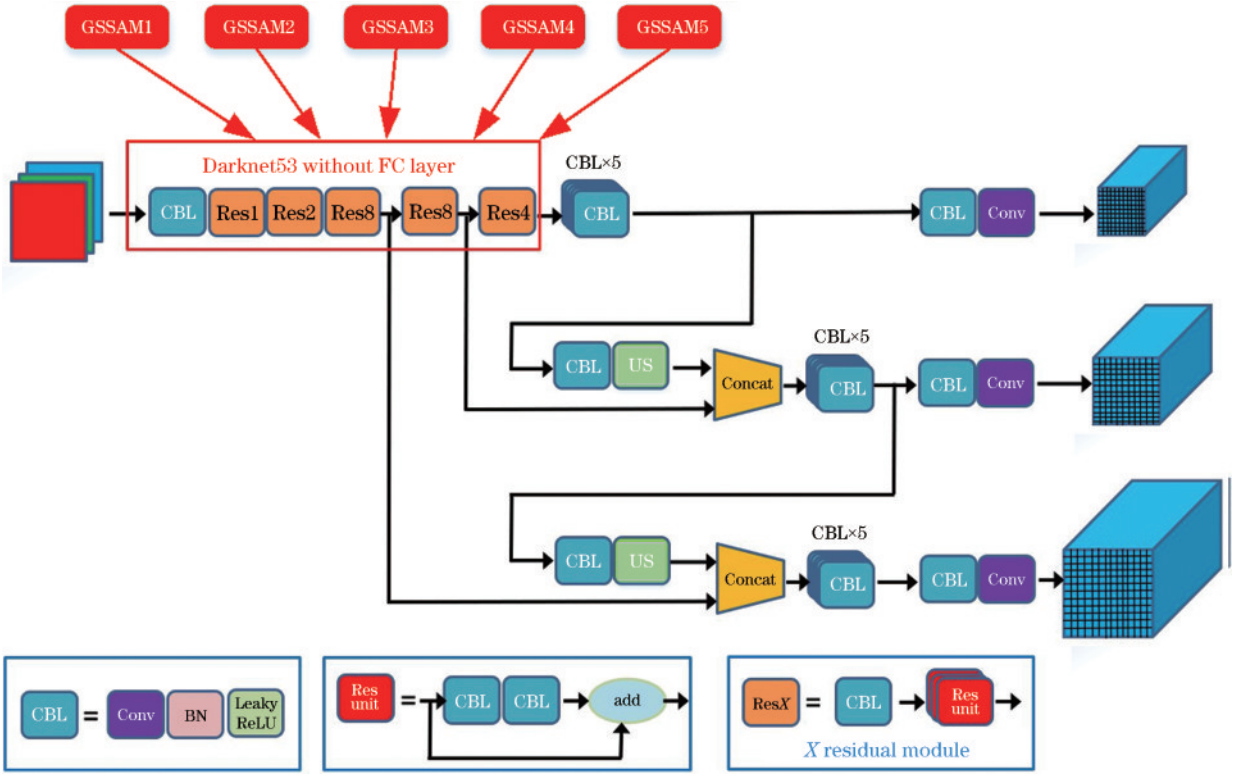

图15　GSSAM嵌入YOLOx的示意图

Fig. 15　Schematic of GSSAM embedding YOLOx

由于YOLOx提取的多个深度模块的输出作为主干网络的输出，GSSAM放置在输出提取前或输出提取后同样会影响YOLOx的识别结果，为此，实验设计的过程中需要将GSSAM放置位置作为变量纳入考虑范围。

实验在团队自制的“低慢小”无人机数据集上进行。该数据集包含不同地理环境、不同光照条件、多种无人机类型的图像约20000张，标注完成4000余张。选择其中80%作为训练集用于对附加GSSAM前后的YOLOx进行训练；10%作为验证集，在训练过程中参与每个epoch的泛化性测试，同时保证网络尽可能避免过拟合现象；剩余10%作为测试集，在训练结束后使用网络进行推理并计算准确率，这些参数作为网络效果的最终评判指标。数据集中部分图像如图16所示。

图16　部分“低慢小”无人机图像

Fig. 16　Partial “low slow small” UAV images

YOLOx作为“一步式”的目标检测识别算法，在深度网络卷积的过程中除了需要实现对目标类别的评价判断外，还需要对目标的位置信息做出预测，因此其生成的特征图与图像分类网络处理的特征图包含的信息存在一定的差异，这种差异在GSSAM计算过程中的归一化部分尤其明显。在图像分类网络中使用常规的标准化算法来实现注意力图的归一化，即线性地将注意力图数值压缩到0与1区间内。由于图像分类过程实质上是对图像各点进行“评分”的过程，对该评分的标准化加权实际上增大了各点分数的差异，因此能够突出“更加需要”关注的部分；而YOLOx网络的中间特征图上各点信息除包含分数，还携带了目标位置信息，如果直接使用标准化强行为该点赋予权重，会导致位置信息的较大损失。因此，在YOLOx附加的GSSAM中尝试选用Sigmoid函数实现权重的归一化，同时增加模块的非线性，以实现更好的目标检测识别效果，而使用标准化算法实现归一化的GSSAM则同样作为对照。此外，考虑到注意力权重的引入可能对原有网络的输出内容存在一定程度的破坏，因此尝试在GSSAM中引入残差概念（即输入与经过处理的输出相加作为输出）进行额外实验。残差形式的GSSAM如图17所示。

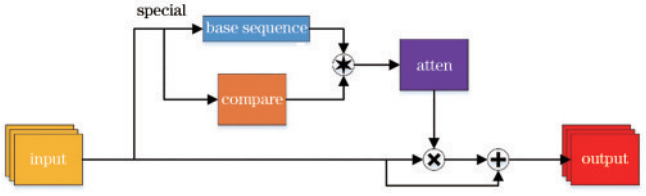

图17　残差GSSAM结构示意图

Fig. 17　Schematic of residual GSSAM structure

汇总的实验结果如表2所示。为了能够较为简便地区分YOLOx网络中GSSAM置入的位置，将GSSAM置入各层输出提取前的形式称为位置1（position 1），置入各层输出提取后的形式称为位置2（position 2）；使用GSSAM standardization代表以标准化实现归一化的GSSAM，GSSAM Sigmoid表示使用Sigmoid函数实现归一化的GSSAM，GSSAM Sigmoid Res则代表具有残差形式、以Sigmoid函数实现归一化的GSSAM。

表2　实验结果统计

Table 2　Statistical of experimental results

Model		Time /ms	$mAP_{50:95}$ /%
Origin YOLOx		20.96	58.67
YOLOx-GSSAM Sigmoid	position 1	22.46	60.64
	Position 2	22.47	60.50
YOLOx-GSSAM standardization	position 1	24.96	59.50
	Position 2	24.43	58.96
YOLOx-GSSAM Sigmoid Res	position 1	22.31	61.44
	position 2	21.99	61.01

表2中：Time是推理1000幅图像所需的单幅图像对应的时间均值；$mAP_{50:95}$是一项专门针对目标检测识别使用的普遍判别指标，即平均精度均值，能够较好地表示目标检测识别算法在目标位置、大小、类别上的效果，其中50:95指交并比（IoU）阈值范围为0.50～0.95，步长为0.05。实验数据显示：相较于原始网络，附加任意GSSAM的YOLOx目标检测识别网络的$mAP_{50:95}$都获得了较为明显的提升；置入标准化GSSAM的网络的

$mAP_{50:95}$提升相对较小，在 position 1 与 position 2 置入形式中分别将 YOLOx 原本的检测识别效果分别提升 0.83 百分点与 0.29 百分点；置入 Sigmoid 实现归一化的 GSSAM 的网络提升 1.97 百分点与 1.83 百分点；置入残差形式且以 Sigmoid 实现归一化的 GSSAM 的网络提升 2.77 百分点与 2.34 百分点。该结果证实了 GSSAM 在已有目标检测识别网络上的使用确实能够为小目标检测识别带来一定程度的改进，具备较好的小目标感知能力。同时，GSSAM 的置入给 YOLOx 原有网络带来的计算成本与时间成本并不高，通常导致对单幅图像的推理时间增加 1～3 ms，不会导致算法实时性的过分衰减，存在可观的应用价值。

GSSAM 在对输入特征图的处理过程中除了有作为采样和全局对比使用的池化操作以及通过“广播”机制实现的减法对比外，还存在占比较高的卷积操作。而其作为“全局采样空间注意力机制”，有必要利用一定程度的对比实验来证明采样与全局对比的重要性。以效果较好的使用 Sigmoid 函数实现归一化的模块为例，在使用残差结构和不使用残差结构的情况下去除 GSSAM 的采样、对比结构，并将其重新置入 YOLOx 网络进行训练与测试。去除采样对比的注意力模块结构如图 18 所示，其结果与原始 GSSAM 获得的结果的对比情况如表 3 所示。

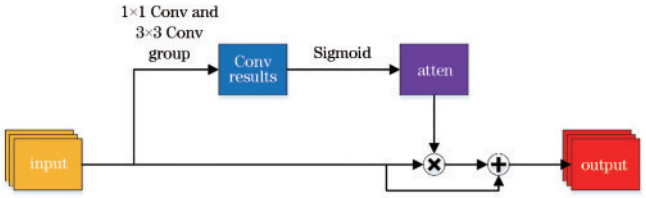

图 18　无采样和对比的模块结构示意图

Fig. 18　Schematic of module structure without sampling and comparison

表 3　采样、对比效果验证

Table 3　Verification for sampling and comparison effects

Parameter	GSSAM		GSSAM Res		Conv		Conv Res	
	position 1	position 2	position 1	position 2	position 1	position 2	position 1	position 2
$mAP_{50:95}$ /%	60.64	60.50	61.44	61.01	59.56	59.49	60.01	60.06

事实证明，采样与对比为模块带来了更加良好的提升效果。单纯以卷积获取注意力的形式与 CBAM 和 BAM 结构类似，尽管能够带来一定程度的提升，但提升效果逊色于增加了采样和对比操作的 GSSAM，并且根据部分未展示实验结果，当卷积层数不足时容易给网络带来非正向的影响，说明采样与对比在 GSSAM 中占据着重要的地位，也是 GSSAM 提出的重要思想来源。

4　结　　论

针对计算机视觉任务中背景导致的有效目标被忽略或是关注程度不够的问题，提出了一种基于卷积、池化、对比方法的注意力机制模块 GSSAM。该模块先通过对神经网络推理过程中产生的一系列特征图进行额外的卷积操作在各点处汇总附近区域信息，再使用池化、对比的方式获取各点与整体图像之间的差异，最终生成注意力图，实现网络在图像中的注意力分配。为了验证模块的性能与适用性，首先通过比较置入 GSSAM 前后不同图像分类网络在部分 ImageNet-1K 数据集上表现的差异，证明 GSSAM 为深度网络在图像分类任务中带来了一致性的提升效果；随后，利用 Grad-CAM 算法对实验中使用的各个网络进行权重可视化，以直观的方式表现 GSSAM 在深度网络中分配注意力的良好性能；最后，将 GSSAM 以不同形式置入 YOLOx 目标检测识别网络，在“低慢小”无人机数据集中开展实验，验证了 GSSAM 较好的小目标感知能力。

由于 GSSAM 由基础的卷积、池化等操作组成，因此模块计算复杂度较低，带来的额外计算成本较低。同时，相比 CBAM、BAM 等简单结构，GSSAM 又具备较大感受野、全局对比的功能，能够实现图像全局信息向局部区域的会聚。模块使用采样的方式，以代表性较强的一点代表一块图像区域，大幅度减少了全局信息的使用带来的高额计算资源占用，为需要进行图像整体与局部对比的部分网络和算法提供了一种可行的选择。

考虑到 GSSAM 的设计理念与结构，GSSAM 在背景重复度较高、目标较小的情况下具备更好的适用性，但当前模块对比部分为简单的相减求和，存在一定的局限性，很可能限制 GSSAM 在更广泛领域中的使用和性能，因此未来将尝试利用不同的对比方式实现全局与局部的比较，探索性能更好、适用性更强的模块结构。同时，对于 GSSAM 的插入带来的少许损失收敛速度下降的问题，也将在日后的研究中进行进一步的探索和改进。

参　考　文　献

[1] Park J, Woo S, Lee J Y, et al. BAM: bottleneck attention module[EB/OL]. (2018-07-17) [2023-06-05]. https://arxiv.org/abs/1807.06514.

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2023-05-06]. https://arxiv.org/abs/1409.1556.

[3] 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 Press, 2016: 770-778.

[4] Tan M X, Le Q V. EfficientNetV2: smaller models and faster training[EB/OL]. (2021-04-01)[2023-06-05]. https://arxiv.org/abs/2104.00298.

[5] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2023-05-06]. https://arxiv.org/abs/1804.02767.

[6] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22)[2023-03-05]. https://arxiv.org/abs/1412.6980.

[7] Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts[EB/OL]. (2016-08-13)[2023-05-06]. https://arxiv.org/abs/1608.03983.

[8] Loshchilov I, Hutter F. Decoupled weight decay regularization[EB/OL]. (2017-11-14)[2023-05-06]. https://arxiv.org/abs/1711.05101.

[9] Niu Z Y, Zhong G Q, Yu H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. New York: ACM Press, 2017: 6000-6010.

[11] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[12] Dai J F, Qi H Z, Xiong Y W, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 764-773.

[13] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7794-7803.

[14] Han K, Xiao A, Wu E H, et al. Transformer in transformer[EB/OL]. (2021-02-27)[2023-05-06]. https://arxiv.org/abs/2103.00112.

[15] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 13708-13717.

[16] Ge Z, Liu S T, Wang F, et al. YOLOX: exceeding YOLO series in 2021[EB/OL]. (2021-07-18)[2023-05-06]. https://arxiv.org/abs/2107.08430.

[17] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [EB/OL]. (2015-02-11)[2023-05-06]. https://arxiv.org/abs/1502.03167.

[18] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 618-626.

[19] Zhou B L, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2921-2929.

[20] Dai Y M, Wu Y Q, Zhou F, et al. Asymmetric contextual modulation for infrared small target detection [C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV), January 3-8, 2021, Waikoloa, HI, USA. New York: IEEE Press, 2021: 949-958.

[21] Qi S H, Du J J, Wu M Y, et al. Underwater small target detection based on deformable convolutional pyramid [C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 23-27, 2022, Singapore, Singapore. New York: IEEE Press, 2022: 2784-2788.

[22] Gao Q, Hong R F, Chen Y T, et al. Research on detection algorithm of foreign object debris and small targets in airport runway based on SSD[C]//CONF-CDS 2021: The 2nd International Conference on Computing and Data Science, January 28-30, 2021, Stanford, CA, USA. New York: ACM Press, 2021.