第 58 卷 第 22 期/2021 年 11 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

分组双注意力语义分割网络

陈小龙[1*]，赵骥[1,2]，陈思溢[1**]，杜鑫浩[1]，刘鑫[1]

[1]湘潭大学自动化与电子信息学院，湖南 湘潭 411100；

[2]清华大学国家 CIMS 工程技术研究中心，北京 100084

摘要 深度学习和自注意力机制的应用，使语义分割网络的性能得到了大幅提升。针对目前自注意力机制将每个像素的所有通道看作一个向量进行计算的粗糙性，基于空间维度和通道维度提出了一种分组双注意力网络。首先，将特征层分成多组；然后，自适应过滤掉每组特征层的无效基组，从而捕获精确的上下文信息；最后，将多组加权后的信息进行融合，获得较强的上下文信息。实验结果表明，本网络在两个数据集上的分割性能均优于双注意力网络，在 PASCAL VOC2012 验证集上的分割精度为 85.6%，在 Cityscapes 验证集上的分割精度为 71.7%。

关键词 图像处理；语义分割；注意力模块；深度学习

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.2210007

Grouped Double Attention Network for Semantic Segmentation

Chen Xiaolong[1*], Zhao Ji[1,2], Chen Siyi[1**], Du Xinhao[1], Liu Xin[1]

[1] *School of Automation and Electronic Information, Xiangtan University, Xiangtan, Hunan* 411100 *China*;

[2] *National CIMS Engineering Technology Research Center, Tsinghua University, Beijing* 100084, *China*

Abstract The application of deep learning and self-attention mechanism greatly improves the performance of semantic segmentation network. Aiming at the roughness of the current self-attention mechanism that treats all channels of each pixel as a vector for calculation, we propose a grouped dual attention network based on the spatial dimension and channel dimension. First, divide the feature layer into multiple groups; then, adaptively filter out the invalid basis groups of each feature layer to capture accurate context information; finally, fuse multiple groups of weighted information to obtain stronger context information. The experimental results show that the segmentation performance of this network on the two data sets is better than dual attention network, the segmentation accuracy on the PASCAL VOC2012 verification set is 85.6%, and the segmentation accuracy on the Cityscapes verification set is 71.7%.

Key words image processing; semantic segmentation; attention module; deep learning

OCIS codes 100.4996; 100.2960; 100.5010

1 引 言

图像分类、目标检测、图像语义分割是计算机视觉的三大基本任务。其中，语义分割是最具挑战性的任务。图像语义分割融合了传统的图像分割和目标识别两个任务，目的是将图像分割成几组具有某种特定语义含义的像素区域，并识别出每个区域的类别，最终获得一张具有像素语义标注的图像。目前，图像语义分割已广泛应用于自动驾驶、卫星图像、医学图形处理等领域中[1-3]。

自卷积神经网络(CNN)[4]提出以来，采用神经网络进行图像分类的算法层出不穷。Krizhevsky 等[5]提出的 AlexNet 获得 ImageNet[6]图像分类竞赛冠军后，深度卷积神经网络(DCNN)逐渐在各类

收稿日期：2020-11-12；**修回日期**：2020-12-30；**录用日期**：2021-01-27

通信作者：*350071235@qq.com；**c.siyi@xtu.edu.cn

视觉任务中占据了主流地位。Long 等[7]基于全卷积神经网络(FCN)的语义分割模型取得了较好的分割效果,原因是 FCN 将普通分类网络的全连接层替换为对应尺寸的卷积层,再通过上采样将特征图尺寸恢复成原始输入图像的尺寸。但由于卷积层操作固有的几何特性,基于 FCN 的语义分割模型感受野较小,只能使用局部的上下文信息,类别区分能力较差。

为了解决 FCN 上下文信息使用不充分的问题,人们提出了利用多尺度信息进行融合的网络,如金字塔场景分析网络(PSPNet)[8]、DeepLab[9-11]采用金字塔池化和空洞卷积操作聚合多尺度信息。Peng 等[12]采用较大的卷积核获得了较大范围的上下文信息。U-Net[13]、SegNet[14]、DeepLabv3+采用编码-解码结构重构高分辨率分割图像。此外,部分网络还使用注意力机制捕获丰富的上下文依赖信息。如 Wang 等[15]使用自注意力机制使任意位置点的特征可接收来自其他所有位置点的特征信息,从而得到上下文信息更丰富的特征表示。Fu 等[16]使用两个注意力机制模块分别捕获 CNN 在通道维度和空间维度上的依赖信息。Yuan 等[17]使用金字塔目标语义模块去除空间上相隔较远像素的影响并加强相隔较近像素的作用。

自注意力机制计算出的注意力图越精细,越能更好地捕获长期依赖性信息。但现有基于注意力机制的语义分割方法都是通过将特征图中每个位置的所有通道看作一个向量计算注意力图,不能很好地表达像素与像素之间的关系。为了解决上述问题,本文基于空间维度和通道维度的分组注意力机制提出了一种分组双注意力网络。首先将特征图进行分组,分别提取更精准的长期依赖信息;然后结合每组提取的上下文信息,增强神经网络类别的紧凑性与区分能力,从而达到更好的语义分割性能。

2　网络框架

2.1　分组双注意力网络

分组双注意力网络的整体框架如图 1 所示。首先,使输入图像通过一个 DCNN。DCNN 采用了残差神经网络(ResNet)[18],为了产生更细致和高效的稠密特征图,移除 ResNet 最后两个下采样操作,并在随后两个残差块中使用了空洞卷积,将特征图的尺寸扩大到输入图像的 1/8,在没有额外增加参数的同时保留了更细致的信息。然后,在 ResNet 的主体网络后,并行使用分组位置注意力模块(GPAM)和分组通道注意力模块(GCAM),分别在空间与通道维度上捕获长距离的依赖信息。最后,将空间与通道维度上的特征进行融合,并将融合后的特征用于语义分割,从而提升语义分割的性能。

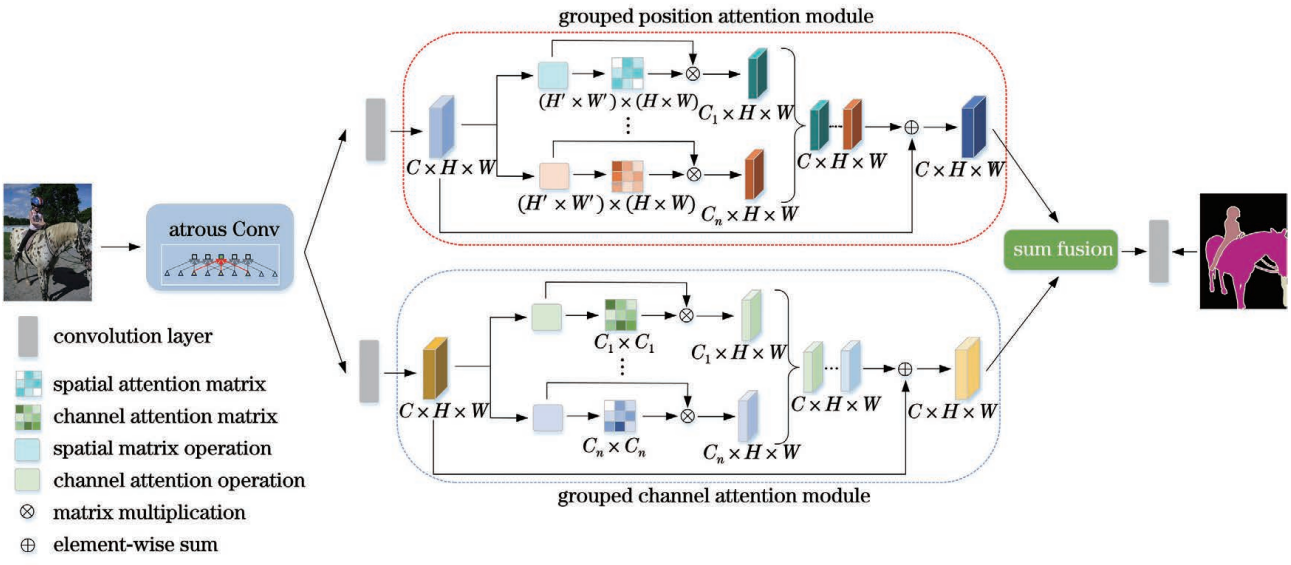

图 1　分组双注意力网络的结构

Fig. 1　Structure of the grouped double attention network

2.2　分组位置注意力模块

充分有效利用长距离的依赖信息对于语义分割任务是非常重要的。自注意力机制可以很好地捕获长距离的上下文依赖关系,但该模块将一个位置的所有通道看作一个向量进行计算,生成的注意力图非常粗糙,不能很好地表达像素之间的依赖关系。为了解决该问题,提出了一种 GPAM,其结构如图 2 所示。

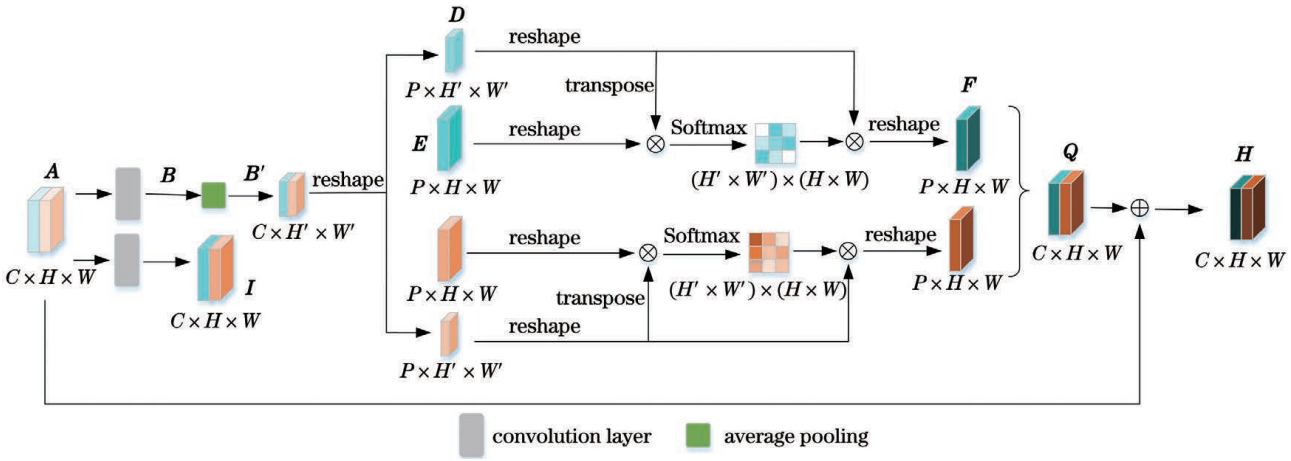

图 2　GPAM 的结构

Fig. 2　Structure of the GPAM

在以分 2 组为例的图 2 中，首先，给定一个局部特征 $\boldsymbol{A}\in\boldsymbol{R}^{C\times H\times W}$，并将其送入带有批归一化（BN）和线性整流函数（ReLU）的卷积层中分别产生特征图 $\boldsymbol{B}$ 和 $\boldsymbol{I}$，其中 $\{\boldsymbol{B},\boldsymbol{I}\}\in\boldsymbol{R}^{C\times H\times W}$。对特征 $\boldsymbol{B}$ 进行池化降维，得到特征 $\boldsymbol{B}'\in\boldsymbol{R}^{C\times H'\times W'}$。然后，将 $\boldsymbol{B}'$ 切分成 k 个特征，其中 1 个特征可表示为 $\boldsymbol{D}\in\boldsymbol{R}^{P\times H'\times W'}$；将 $\boldsymbol{I}$ 切分成 k 个特征，其中 1 个特征可表示为 $\boldsymbol{E}\in\boldsymbol{R}^{P\times H\times W}$，且 $C=P\times k$。将 $\boldsymbol{D}$、$\boldsymbol{E}$ 分别重塑为 $\boldsymbol{R}^{P\times N'}$ 和 $\boldsymbol{R}^{P\times N}$，其中，$N'=W'\times H'$、$N=W\times H$ 分别为特征的数量。将 $\boldsymbol{D}$ 的转置与 $\boldsymbol{E}$ 相乘后再通过 Softmax 层计算出每组的空间注意力图 $\boldsymbol{S}\in\boldsymbol{R}^{N'\times N}$，该特征中第 i 个位置对第 j 个基组的影响可表示为

$$\boldsymbol{s}_{ji}=\frac{\exp(\boldsymbol{D}_i^{\mathrm{T}}\cdot\boldsymbol{E}_j^{\mathrm{T}})}{\sum_{i=1}^{N'}(\boldsymbol{D}_i^{\mathrm{T}}\cdot\boldsymbol{E}_j^{\mathrm{T}})}。\tag{1}$$

计算出注意力图 $\boldsymbol{S}$ 后，将 $\boldsymbol{D}$ 与 $\boldsymbol{S}$ 相乘，生成新的特征 $\boldsymbol{F}\in\boldsymbol{R}^{P\times N}$ 并将其重塑成 $\boldsymbol{R}^{P\times H\times W}$。按相同的方法，在每组中都产生一个特征 $\boldsymbol{R}^{P\times H\times W}$，将产生的 k 组特征结合成新的特征 $\boldsymbol{Q}\in\boldsymbol{R}^{C\times H\times W}$。最后，将 α 倍的 $\boldsymbol{Q}$ 与特征 $\boldsymbol{A}$ 的对应元素相加，获得输出特征 $\boldsymbol{H}\in\boldsymbol{R}^{C\times H\times W}$，可表示为

$$\boldsymbol{F}_j=\sum_{i=1}^{N'}\boldsymbol{s}_{ji}\boldsymbol{D}_i,\tag{2}$$

$$\boldsymbol{H}_j=\alpha\boldsymbol{Q}_j+\boldsymbol{A}_j。\tag{3}$$

将权重 α 初始化为 0 并通过多次训练进行优化，参数 α 有助于对 k 组特征进行筛选并增强有效的特征，从而提高语义分割的性能。

2.3　分组通道注意力模块

每个高层特征的通道图都可以看作一个特定类别的响应，可通过挖掘通道间的相互依赖关系突出相互作用的特征，提高特定语义的特征表示。但自注意通道注意力要求提取每个通道与所有通道的相关特征，产生的注意力图尺寸较大；且会产生大量的冗余信息，使生成的注意力图不准确，最终导致部分类别分类错误。此外，该方法占用的内存较高。为了解决上述问题，提出了 GCAM，其结构如图 3 所示。

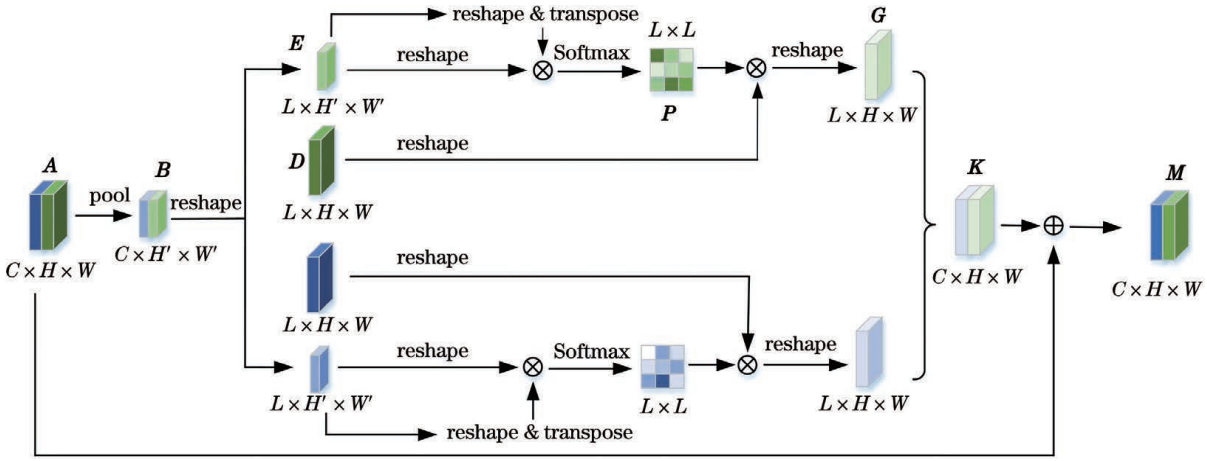

图 3　GCAM 的结构

Fig. 3　Structure of the GCAM

在 GCAM 中，首先，给定一个局部特征 $\boldsymbol{A}\in\boldsymbol{R}^{C\times H\times W}$，并对其进行池化降维，得到特征 $\boldsymbol{B}\in\boldsymbol{R}^{C\times H'\times W'}$。然后，将特征 $\boldsymbol{B}$ 拆分成 k 组，其中 1 组特征可表示为 $\boldsymbol{E}\in\boldsymbol{R}^{L\times H'\times W'}$；将特征 $\boldsymbol{A}$ 拆分成 k 组，其中 1 组特征可表示为 $\boldsymbol{D}\in\boldsymbol{R}^{L\times H\times W}$，且 $C=L\times k$。将特征 $\boldsymbol{E}$ 重塑为 $\boldsymbol{R}^{L\times N'}$；将 $\boldsymbol{E}$ 与 $\boldsymbol{E}'$ 相乘，得到一个较小的相似度矩阵 $\boldsymbol{F}$，其中 $N'=W'\times H'$。将相似度矩阵通过 Softmax 层获得注意力图 $\boldsymbol{X}\in\boldsymbol{R}^{L\times L}$，该注意力图中第 i 个通道对第 j 个通道的影响可表示为

$$\boldsymbol{x}_{ji}=\frac{\exp(\boldsymbol{E}_i\cdot\boldsymbol{E}'_j)}{\sum_{i=1}^{N'}(\boldsymbol{E}_i\cdot\boldsymbol{E}'_j)}。\tag{4}$$

计算出注意力图 $\boldsymbol{X}$ 后，将特征 $\boldsymbol{D}$ 重塑成 $\boldsymbol{R}^{L\times N}$，并将注意力图 $\boldsymbol{X}$ 与特征 $\boldsymbol{D}$ 相乘后进行通道加权，得到特征 $\boldsymbol{G}\in\boldsymbol{R}^{L\times W\times H}$。将 k 组特征进行堆叠得到 $\boldsymbol{K}\in\boldsymbol{R}^{C\times W\times H}$，并将 β 倍的 $\boldsymbol{K}$ 与 $\boldsymbol{A}$ 的对应元素相加，得到最后的输出特征 $\boldsymbol{M}\in\boldsymbol{R}^{C\times W\times H}$，可表示为

$$\boldsymbol{G}_j=\sum_{i=1}^{N'}\boldsymbol{x}_{ji}\boldsymbol{D}_i,\tag{5}$$

$$\boldsymbol{M}_j=\beta\boldsymbol{K}_j+\boldsymbol{A}_j,\tag{6}$$

式中，参数 β 初始化为 0 并通过学习分配更合适的权重。分组求取注意力图可降低注意力图占用的内存，同时减少冗余通道特征给其他通道带来的干扰，从而提高了网络的类别区分能力。

3 实验结果与分析

为了验证本方法的有效性，在 PASCAL VOC2012 数据集[19]和 Cityscapes 数据集上[20-21]进行了大量实验，并对结果进行了详细分析。首先详细介绍了数据集和实验策略，然后分别对 GPAM 和 GCAM 进行消融实验，最后将整个网络与现存网络的性能进行了对比分析。

3.1 数据集与实现细节

在 PASCAL VOC2012 和 Cityscapes 数据集上进行了语义分割实验。PASCAL VOC2012 数据集包含 20 个前景目标类别和 1 个背景类别。初始数据集含有 1464 张训练图像、1449 张验证图像和 1456 张测试图像。Cityscapes 数据集有 19 类(包括车、建筑和行人等)图像，包括从 50 个城市获取的 5000 张精细标注和 2000 张粗略标注的城市路面场景图像。实验仅采用 5000 张精细标注数据，未使用粗略标注数据。5000 张精细标注图像中包含 2975 张训练图像、500 张验证图像和 1525 张测试图像。为了保证对比的公平性，仅使用训练图像进行训练，并在验证图像或测试图像上进行测试。

基于文献[22-24]的研究结果，采用语义分割领域常用的平均交并比(mIoU)评价分割效果。mIoU 能反映预测值与真实值之间的相关度，且相关度越高，mIoU 越大，可表示为

$$X_{\mathrm{mIoU}}=\frac{f_{\mathrm{TP}}}{f_{\mathrm{TP}}+f_{\mathrm{FP}}+f_{\mathrm{FN}}},\tag{7}$$

式中，f_{TP}、f_{FP} 和 f_{FN} 分别为真阳率(标签为正，预测结果为正)、假阴率(标签为负，预测结果为正)和假阳率(标签为正，预测结果为负)。

实验采用的硬件环境：GPU 为 GeForceGTX 1080Ti，软件平台为 Pytorch 框架。使用批量随机梯度下降法进行训练，为了更公平地对比不同方法的效果，采用多元学习策略，初始学习率每次迭代后都乘以因子 $\left(1-\frac{X_{\mathrm{iter}}}{X_{\mathrm{max\_iter}}}\right)^{p}$，以减小学习率。其中，$X_{\mathrm{iter}}$ 为当前迭代次数，$X_{\mathrm{max\_iter}}$ 为总迭代次数，动量系数 p 为 0.9。网络训练过程的超参数采用了 Fu 等[16]的设置，设置的初始学习率为 0.0001，动量系数为 0.9，权重衰减系数为 0.0001，将类别中每个像素位置的交叉熵损失和作为损失函数。受计算资源的限制，将 PASCAL VOC2012 数据集的最小批量设为 8，Cityscapes 数据集的最小批量设为 4。对 PASCAL VOC2012 数据集训练 50 轮，对 Cityscapes 数据集则训练 180 轮。训练时在 GPAM 和 GCAM 的末端使用了辅助监督，训练过程采用将原始图像尺寸随机裁剪成 512 pixel×512 pixel、随机左右翻转及在 0.5 到 2.0 之间随机缩放进行数据增广。

3.2 消融实验

通过逐步分解本方法每个注意力模块的效果，研究每个注意模块中的最优参数设置，以获得更好的性能。在 PASCAL VOC2012 数据集上进行参数优化，在 PASCAL VOC2012 数据集和 Cityscapes 数据集上对比不同方法的分割结果。

为了更好地对比 GPAM、GCAM 和整个网络之间的分割性能，设置了相同的主干网络(ResNet50)。在 GPAM 中，需要对特征图进行分组操作。分组数量太多会使注意力模块的噪声太大，分组数量太少会使冗余信息太多，影响有效特征的捕捉，从而降低语义分割的类别紧凑性。因此，对 GPAM 的分组数量进行了研究。表 1 为 GPAM 的分组数量对网络性能的影响，先将

ResNet50 最后 2 个下采样层改为空洞卷积层作为主干网络，然后将上采样到原始图像尺寸的 FCN 作为 Baseline1。采用相同的主干网络并在末端加入双注意力网络(DANet)中的位置注意力模块(PAM)作为 Baseline2。在主干网络后加入 GPAM 的方法为本方法，并记为 Our。其中，GNP 为分组的数量。

表 1 GPAM 分组数量对网络性能的影响

Table 1 Influence of the number of GPAM groups on network performance

Method	Backbone	PAM	GNP	mIoU /%
Baseline1	ResNet50			69.8
Baseline2	ResNet50	√		83.2
Our1	ResNet50		1	84.1
Our2	ResNet50		2	84.9
Our3	ResNet50		4	84.4
Our4	ResNet50		8	84.2
Our5	ResNet50		16	82.9
Our6	ResNet50		64	81.1

从表 1 可以发现，相比 Baseline1 方法，使用 GPAM 后的 Our2 方法 mIoU 提高了 15.1 个百分点，验证了 GPAM 的有效性。相比 Baseline2 方法，Our1 方法使用的分组注意力模块与基于 Non-local 的注意力模块组数相同，但其语义分割性能提高了 0.9 个百分点。这表明过滤掉无效基组可减少噪声信息的干扰，从而提高语义分割的性能。相比 Our1 方法，Our2 方法的 mIoU 增加了 0.8 个百分点，这表明将特征信息分成多组后对每组进行位置加权，可捕获更细致的长距离依赖信息，有利于语义分割性能的提升。对比 Our2～Our6 的分割性能发现，随着分组数量的增加，分割性能逐渐呈降低趋势。原因是分组数量过多，每组特征的通道数过少，使 GPAM 的噪声过大，从而降低了分割效果。

为了探索基组数量对语义分割性能的影响，研究了不同基组数量下本方法的性能，结果如表 2 所示。其中，所有方法的分组数都为 2，NBP 为 GPAM 的基组数。可以发现，相比 Baseline2 方法，Our8 方法的分割 mIoU 提高了 1.8 个百分点，这表明过滤掉某些冗余位置的信息能计算出更精细的注意力图，从而增强类别的紧凑性并提高语义分割的性能。此外，随着 GPAM 中基组数量的减少，语义分割性能也有略微降低，这表明基组数量过少会使注意力模块特征表达不足，从而降低语义分割的性能。

表 2 GPAM 基组数量对网络性能的影响

Table 2 Influence of the number of GPAM basis sets on network performance

Method	Backbone	PAM	NBP	mIoU /%
Baseline2	ResNet50	√		83.2
Our7	ResNet50		32	85.0
Our8	ResNet50		16	85.0
Our2	ResNet50		8	84.9

在确定 GPAM 的分组数量和基组数量后，为了使网络的分割性能达到最优，进一步对 GCAM 的结构进行了研究。表 3 为 GCAM 分组数量对网络分割性能的影响，其中，GNC 为 GCAM 的分组数，主干网络均为 ResNet50，将主干网络后端接 DANet 中 CAM 的网络记为 Baseline3。可以发现，相比 Baseline3 方法，Our-3 方法的 mIoU 增加了 2.3 个百分点。这表明 GCAM 能捕捉更多种类的特征信息，且减少了其他无关通道对加权通道信息的干扰，从而提升了语义分割的性能。此外，随着分组数量的增加，网络的语义分割性能也有一定提升。这表明在求取加权通道信息时，采用分组注意力能解决大量冗余通道信息的干扰。

表 3 GCAM 数量对网络性能的影响

Table 3 Influence of the number of GCAM groups on network performance

Method	Backbone	CAM	GNC	mIoU /%
Baseline3	ResNet50	√		82.6
Our-1	ResNet50		8	83.9
Our-2	ResNet50		16	84.1
Our-3	ResNet50		32	84.9

图 4 为传统 CAM 和 GCAM 的注意力图，图 4(a)为 DANet 中 CAM 的注意力图，图 4(b)为分 3 组 CAM 的注意力图。可以发现，在基于自注意力机制的模块中，占用 GPU 内存最大的为注意力图。对比发现，CAM 的注意力图尺寸为 GCAM 注意图的分组数倍，从而在一定程度上减少了 GPU 内存的消耗。表 4 为传统 CAM 和 GCAM 占用的内存，可以发现，相比 CAM，GCAM 占用的内存有明显降低，验证了该方法的有效性；且随着分组数量的增加，GPU 占用内存也有所降低，原因是随着分

组数量的增加，占用 GPU 内存的注意力图尺寸逐渐减小。

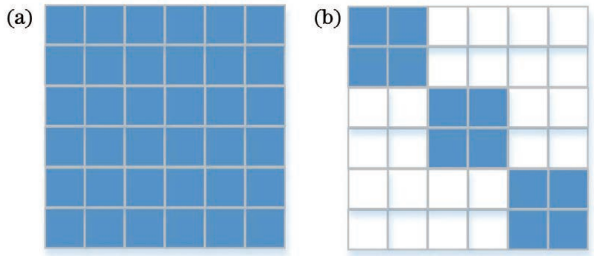

图 4　CAM 和 GCAM 的注意力图。(a)CAM;(b)GCMA

Fig. 4　Attention maps of CAM and GCAM. (a) CAM; (b) GCMA

表 4　CAM 和 GCAM 占用的内存

Table 4　Memory occupied by CAM and GCAM

Method	GNC	Memory /G	mIoU /%
CAM	–	1.00	82.6
GCAM	8	0.85	83.9
GCAM	16	0.73	84.1
GCAM	32	0.68	84.9

确定 GCAM 的分组数后，对该模块中的池化尺寸进行了研究。表 5 为 GCAM 池化尺寸对网络性能的形象，该模块的分组数均为 32，PSC 为 GCAM 的池化尺寸。可以发现，相比 Our-3 和 Baseline3 方法，Our-4 方法的分割 mIoU 分别提高 0.2 和 2.3 个百分点。随着池化尺寸的增加，网络的语义分割性能有所增加，这表明过滤掉每个通道中冗余位置的信息后再计算通道之间的相似度，可计算出更精确的注意力图。相比 Our-3 方法，随着池化尺寸的增加，Our-5 方法的分割 mIoU 降低了 0.6 个百分点。原因是 GCAM 的池化尺寸过大会过滤掉每个通道的有效信息，降低注意力图的精确度，最终导致分割性能降低。

表 5　GCAM 池化尺寸对分割性能的影响

Table 5　Influence of the size of the GCAM pooling on segmentation performance

Method	Backbone	CAM	PSC	mIoU /%
Baseline3	ResNet50	√		82.6
Our-4	ResNet50		4	84.7
Our-3	ResNet50		8	84.9
Our-5	ResNet50		16	84.3

3.3　实验结果的对比

为了验证网络的整体分割性能，将本网络与 Baseline 及每个模块的网络进行了对比实验，网络中的超参数均与基本方法一致。以 ResNet50 为主干网络，用分辨率为 512 pixel×512 pixel 的图像进行训练，并在验证集上进行测试，结果如表 6 所示。对比 Baseline2 和 Our7 方法可以发现，GPAM 比 DANet 中 PAM 的分割 mIoU 高 1.8 个百分点，这表明过滤掉无用基组能得到更精准的注意力图，且捕获的特征种类更多，从而提升语义分割的准确性。对比 Baseline3 和 Our-3 方法可以发现，GCAM 比 DANet 中 CAM 的分割 mIoU 高 2.3 个百分点，这表明分组注意力过滤掉无关通道后再进行加权使关注的特征更加突出，从而提高了语义分割的效果。表 6 中的 GDANet 将 GPAM 与 GCAM 并行连接后，再将两个结果进行融合，得到分组双注意力网络，使网络的分割 mIoU 又提升了 0.6 个百分点。相比基础网络 Baseline1，本方法(GDANet)的分割 mIoU 提高了 15.8 个百分点，进一步验证了本方法的有效性。

表 6　分组双注意力网络与 Baseline 的实验结果

Table 6　Experimental results of grouped double attention network and Baseline

Method	PAM	CAM	GPAM	GCAM	mIoU /%
Baseline1					69.8
Baseline2	√				83.2
Baseline3		√			82.6
Our7			√		85.0
Our-3				√	84.9
GDANet			√	√	85.6

为了对比本方法与现有方法的性能，将本网络中的超参数与其他网络保持一致。首先用分辨率为 512 pixel×512 pixel 的图像进行训练，用官方验证集进行测试。不同方法在 PASCAL VOC2012 验证集上的实验结果如表 7 所示，可以发现，相比其他方法，本方法的语义分割精度最高，可达到 85.6%。需要说明的是，对于轮廓更精细的物体类别，如 aero、bike、bus、cow、mbike、person 和 plant 类别，相比其他方法，本方法的语义分割精度有显著提升。不同方法在 Cityscapes 验证集上的实验结果如表 8 所示，可以发现，本方法的分割精度达到了 71.7%，优于部分现有的语义分割方法。这表明过滤掉冗余信息计算出的注意力图能提升网络的语义分割性能。

表 7 不同方法在 PASCAL VOC2012 验证集的实验结果

Table 7 Experimental results of different methods in the PASCAL VOC2012 validation set unit: %

Method	FCN	DeepLabv2	DPN[25]	DeepLabv3	PSP	DANet	Ours
Aero	82.4	84.4	87.7	88.0	87.4	90.1	**92.8**
Bike	47.4	54.5	59.4	56.3	56.3	61.8	**67.8**
Bird	81.2	81.5	78.4	86.3	85.7	91.7	**91.8**
Boat	68.6	63.6	64.9	69.4	79.4	75.6	**82.5**
Bottle	75.3	65.9	70.3	72.2	73.8	75.6	**76.7**
Bus	81.3	85.1	89.3	90.3	92.3	93.1	**95.0**
Car	79.9	79.1	83.5	85.7	87.3	88.5	**90.7**
Cat	81.6	83.4	86.1	89.6	92.3	**92.9**	92.7
Chair	33.7	30.7	31.7	28.9	53.3	53.4	**61.7**
Cow	68.4	74.1	79.9	85.9	90.4	93.3	**94.8**
Table	52.3	59.8	62.6	59.3	75.2	74.3	**81.3**
Dog	76.4	79	81.9	84.2	87.3	92	**93.5**
Horse	64.9	76.1	80	80.2	85.9	89.1	**92.4**
Mbike	73.4	83.2	83.5	84.2	83.8	85.4	**88.7**
Person	81.2	80.8	82.3	82.8	84.5	85.7	**88.3**
Plant	56.7	59.7	60.5	56.0	68.1	62.8	**70.0**
Sheep	69.7	82.2	83.2	78.5	87	91.6	**92.6**
Sofa	50.9	50.4	53.4	51.6	73	74.6	**78.1**
Train	78.5	73.1	77.9	84.5	91.1	90.2	**92.0**
Tv	70.1	63.7	65.0	69.6	71.5	73.1	**77.1**
mIoU	69.8	71.6	74.1	75.1	80.9	82.4	**85.6**

表 8 不同方法在 Cityscapes 验证集上的实验结果

Table 8 Experimental results of different methods on the Cityscapes validation set unit: %

Method	FCN	PSP	DANet	Ours	Method	FCN	PSP	DANet	Ours
Road	95.1	96.4	97.2	**97.5**	Sky	91.4	92	92.4	**92.8**
Sidewalk	67.8	74.4	77.8	**79.3**	Person	68.8	70.4	71.9	**72.9**
Building	88.5	89.1	89.8	**90.1**	Rider	47.9	49.9	52.2	**53.3**
Wall	50.5	52.9	56.1	**57.1**	Car	90.3	91.4	**92.4**	**92.4**
Fence	44.6	47.9	48.6	**51.2**	Truck	73.8	73.9	**82.8**	79.2
Pole	35.6	39.9	40.8	**43.4**	Bus	73.6	75.8	79.4	**81.9**
Traffic light	47.0	51.9	53.0	**53.5**	Train	62.8	66.4	70.8	**74.5**
Traffic sign	60.4	62.4	65.2	**66.4**	Motocycle	51.7	55.0	**58.9**	58.7
Vegetation	88.6	89.4	89.7	**89.9**	Bicycle	63.1	63.6	65.8	**66.7**
Terrain	55.6	57.6	60.7	**60.9**					
mIoU	66.2	68.4	70.8	**71.7**	mIoU	66.2	68.4	70.8	**71.7**

用定性方式对比分析了本方法与 Baseline1 方法的性能，结果如图 5 所示。可以发现，在图像中的公交车车头部分，本方法对汽车的轮廓分割比 Baseline1 更精细。在餐桌的座椅处，本方法对物体的轮廓分割更贴合实际。原因是 Baseline1 方法只使用了长距离的依赖信息，缺乏对空间维度有效信息的捕获，而本方法采用 GPAM 增强类内目标的紧凑性，使分割轮廓更细致。此外，Baseline1 方法将牛的部分位置误分类为马，原因是牛的某些外观特征与马相似，Baseline1 使用长距离的依赖关系信息，相似类别存在信息干扰，而本方法通过 GCAM 改善类别与类别之间的关系，从而增强了对类别的区分能力。

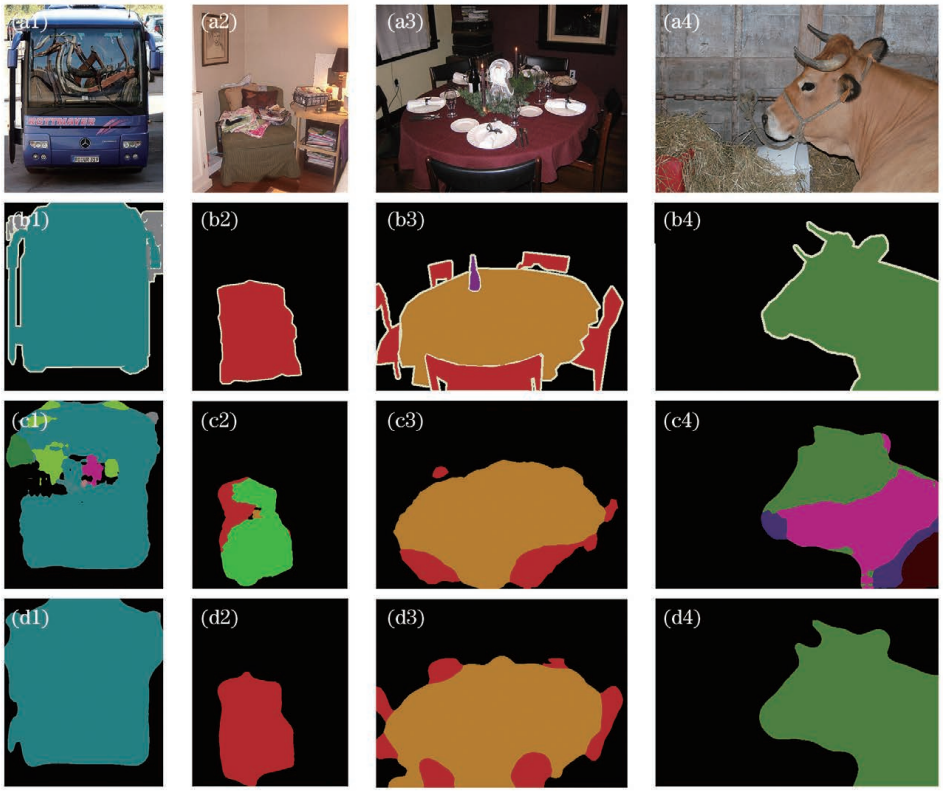

图 5　不同方法的分割结果。(a)原始图像；(b)真实的语义分割标签；(c)基础方法；(d)本方法

Fig. 5　Segmentation results of different methods. (a) Original image; (b) real semantic label; (c) basic method; (d) our method

4　结　　论

在分组双注意力网络中，GPAM 与 GCAM 分别在空间维度和通道维度上捕获长距离的依赖信息。为了避免冗余信息对注意力图的干扰，采用分组求取注意力图的方式增强网络对类别的区分能力和分割图像类内的紧凑性，从而获得更好的语义分割效果。在 PSCAL VOC2012 验证集上，分组双注意力网络的分割精度达到了 85.6%，比 Baseline1 提高了 15.8 个百分点。在 Cityscapes 验证集上，其分割精度达到了 71.7%，比 DANet 提高了 0.8 个百分点。之后的研究可集中在如何将该模型在智能驾驶等领域中进行广泛应用。

参　考　文　献

[1]　Cheng X Y, Zhao L Z, Hu Q, et al. Real-time semantic segmentation based on dilated convolution smoothing and lightweight up-sampling[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021017.
程晓悦，赵龙章，胡穹，等. 基于膨胀卷积平滑及轻型上采样的实时语义分割[J]. 激光与光电子学进展，2020，57(2)：021017.

[2]　Li L F, Hu M. Method for small-bridge-crack segmentation based on generative adversarial network [J]. Laser & Optoelectronics Progress, 2019, 56(10): 101004.
李良福，胡敏. 基于生成式对抗网络的细小桥梁裂缝分割方法[J]. 激光与光电子学进展，2019，56(10)：101004.

[3]　Cai Y, Huang X G, Zhang Z A, et al. Real-time semantic segmentation algorithm based on feature fusion technology[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021011.
蔡雨，黄学功，张志安，等. 基于特征融合的实时语

义分割算法[J]. 激光与光电子学进展, 2020, 57(2): 021011.

[4] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[5] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database [C] // 2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 248-255.

[7] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 3431-3440.

[8] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[9] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[10] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05)[2020-11-05]. https://arxiv.org/abs/1706.05587.

[11] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 833-851.

[12] Peng C, Zhang X Y, Yu G, et al. Large kernel matters: improve semantic segmentation by global convolutional network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1743-1751.

[13] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.

[14] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[15] 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.

[16] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 3141-3149.

[17] Yuan Y H, Wang J D. OCNet: object context network for scene parsing[EB/OL]. (2018-09-04)[2020-11-05]. https://arxiv.org/abs/1809.00916.

[18] 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.

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

[20] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset [C]//CVPR Workshop on the Future of Datasets in Vision, June 7-12, 2015, Boston, Massachusetts. New York: IEEE Press, 2015.

[21] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 3213-3223.

[22] Zhang H, Dana K, Shi J P, et al. Context encoding for semantic segmentation [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: 7151-7160.

[23] Wu Z F, Shen C H, van den Hengel A. Wider or deeper: revisiting the ResNet model for visual recognition[J]. Pattern Recognition, 2019, 90: 119-

133.

[24] Liu Z W, Li X X, Luo P, et al. Semantic image segmentation via deep parsing network [C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1377-1385.

[25] Chen Y P, Li J N, Xiao H X, et al. Dual path networks[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. Canada: NIPS, 2017: 4467-4475.