第 58 卷 第 14 期/2021 年 7 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

基于注意力编码的轻量化语义分割网络

陈小龙[1*]，赵骥[1,2]，陈思溢[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 为了解决自注意力机制的注意力图计算复杂度高、内存占用大等问题，同时提高语义分割网络的性能，提出了一种基于注意力编码的轻量化网络。该网络用自适应位置注意力模块和全局注意力上采样模块分别对长距离语义依赖关系进行编码和解码，在计算注意力图时先用自适应位置注意力模块排除无用的基组，再获取上下文信息；全局注意力上采样模块用全局上下文信息引导低层特征重构高分辨率图像。实验结果表明，本网络在 PASCAL VOC2012 验证集上的分割精度达到 84.9%，相比分割精度相近的双路注意力网络，本网络的每秒浮点运算次数降低了 16.9%，占用的 GPU 内存减少了 12.9%。

关键词 图像处理；语义分割；自注意力模块；轻量化网络；编解码结构

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.1410012

Lightweight Semantic Segmentation Network Based on Attention Coding

Chen Xiaolong[1*], Zhao Ji[1,2], Chen Siyi[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 To address the issues of high computational complexity and large memory footprint of the attention map of the self-attention mechanism and to improve the performance of the semantic segmentation network, we propose a lightweight network based on attention coding. The network uses an adaptive positional attention module and global attention upsampling module to encode and decode long-range dependency information, respectively. When calculating the attention map, adaptive positional attention module excludes useless basis sets and context information is obtained. A global attention upsampling module uses global context information to guide low-level features to reconstruct high-resolution images. Experimental results show that the segmentation accuracy of the network on the PASCAL VOC2012 verification set reaches a value of 84.9%. Compared with dual attention network, which has a similar segmentation accuracy, the giga floating-point operations per second and the GPU memory of the network are reduced by 16.9% and 12.9%, respectively.

Key words image processing; semantic segmentation; self-attention module; lightweight network; encoder-decoder structures

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图像分割和目标识别两个任务，其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 1 张具有像素语义标注的图像。目前，

收稿日期：2020-09-16；**修回日期**：2020-10-13；**录用日期**：2020-11-14

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

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[1-3]。

近年来，深度学习技术得到了迅速发展，卷积神经网络(CNN)[4]的提出使基于神经网络的图像分类算法层出不穷，尤其是 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)逐渐在各类视觉任务中占据了主流地位。Long 等[7]提出了一种基于全卷积网络(FCN)的语义分割算法，用 FCN 将普通分类网络的全连接层替换为对应尺寸的卷积层，再通过上采样恢复成原始输入图像的尺寸。但卷积操作固有的几何特性，使基于 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，类别区分能力较低。为了解决 FCN 不能充分利用上下文信息的缺点，Zhao 等[8]提出了一种金字塔池化模块，可聚合不同区域的上下文信息；Chen 等[9-11]进一步提出了空洞空间金字塔池化(ASPP)模块，通过聚合不同区域的上下文信息大幅提高了网络的分类精度。ASPP 模块使用多个不同膨胀率的卷积层对神经网络的特征进行提取，膨胀率越大，表明上下文信息范围越大，越有利于获取不同尺度范围的上下文信息，从而提高网络的语义分割性能。Peng 等[12]采用尺寸较大的卷积核获得了较大范围的上下文信息。Wang 等[13]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，从而得到更丰富的上下文信息特征表示。Fu 等[14]使用两个注意力机制模块分别对通道维度和空间维度上的依赖信息进行提取。Yuan 等[15]使用金字塔目标语义模块去除空间上相隔较远像素间的影响，并加强了相隔较近像素间的影响。上述算法都是基于非局部神经网络[13]的自注意力机制语义分割算法，需要生成一个巨大的注意力图，其时间和空间复杂度均为 $O[(H\times W)\times(H\times W)]$，计算复杂度高、GPU 内存占用大。其中，$H$ 和 W 分别为特征图的长和宽。自注意力算法中非局部模块的每个像素都融合了其他所有位置的特征信息，但某些位置的特征信息是冗余的。因此，计算注意力图时可自适应剔除没有价值的位置特征信息，从而大幅降低计算复杂度和 GPU 的占用内存。此外，虽然高层特征在类别分类方面的表现较好，但对原始图像的重构效果较差。因此，人们提出了一些 U 型网络，如 SegNet[16]、U-net[17]、Refinenet[18]，但这些网络都采用复杂的解码器模块恢复高层特征的细节信息，耗时较长，且对 GPU 的内存占用大。

针对上述问题，本文提出了一种基于全局注意力上采样(GAU)模块和自适应位置注意力模块(APAM)的网络。GAU 模块可在计算量较小的情况下将全局上下文特征信息作为引导，对低层特征信息进行加权。APAM 通过结合每个像素与其他有效位置的特征信息聚合远距离像素的上下文信息，将网络的时间和空间复杂度从 $O[(H\times W)\times(H\times W)]$ 降到原始网络的 1/64，即 $O[(H/8)\times(W/8)\times(H\times W)]$。在 PASCAL VOC2012[19]验证集上的实验结果表明，本网络的分割精度可达到 84.9%，相比分割精度相近的双路注意力网络(DANet)，本网络的每秒浮点运算次数(GFLOPS)降低了 16.9%，GPU 占用内存减少了 12.9%。

2 本文工作

2.1 基于注意力编码的轻量化网络

基于注意力编码的轻量化网络整体框架如图 1 所示。输入图像先通过一个 DCNN，即 FCN，为了产生更细致和高效的稠密特征图，该网络移除了最后两个下采样操作，并在随后的卷积层中使用了空洞卷积。因此，输出特征 $\boldsymbol{X}$ 的宽和高均为输入图像的 1/8。

获得特征 $\boldsymbol{X}$ 后，首先，用一个卷积层(Conv)接收特征信息，同时将特征信息送入 APAM 中聚合长距离的上下文信息并编码成一个新的特征 $\boldsymbol{M}$。然后，通过卷积操作将特征 $\boldsymbol{M}$ 降维成 $\boldsymbol{M}'$，并将 DCNN 在进行空洞卷积之前的特征进行下采样，得到特征 $\boldsymbol{P}$。其次，将 $\boldsymbol{P}$ 和 $\boldsymbol{M}'$ 同时送入 GAU 模块中进行解码，用高层语义引导低层特征重构出更细致的特征 $\boldsymbol{Q}$。最后，通过卷积层将解码器的输出特征进行聚合映射，获得用于像素级预测的特征表达。

2.2 自适应位置注意力模块

长距离的上下文依赖关系信息对于语义分割是必不可少的，自注意力机制模块可以很好地捕捉长距离的上下文依赖关系，但其注意力图计算复杂度较高且 GPU 内存占用大。因此，提出了一种 APAM，其结构如图 2 所示。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 $\boldsymbol{B}$ 和 $\boldsymbol{E}$，其中 $\{\boldsymbol{B},\boldsymbol{E}\}\in\boldsymbol{R}^{C\times H\times W}$。其次，为了在计算注意力图时自适应过滤掉那些无用的基组，采用自适应池化的方式将基组 $\boldsymbol{B}$ 过滤为 $\boldsymbol{D}$，其中 $\boldsymbol{D}\in\boldsymbol{R}^{C\times H'\times W'}$。将 $\boldsymbol{D}$ 重建为 $\boldsymbol{R}^{C\times N'}$，$N'=H'\times W'$ 为特

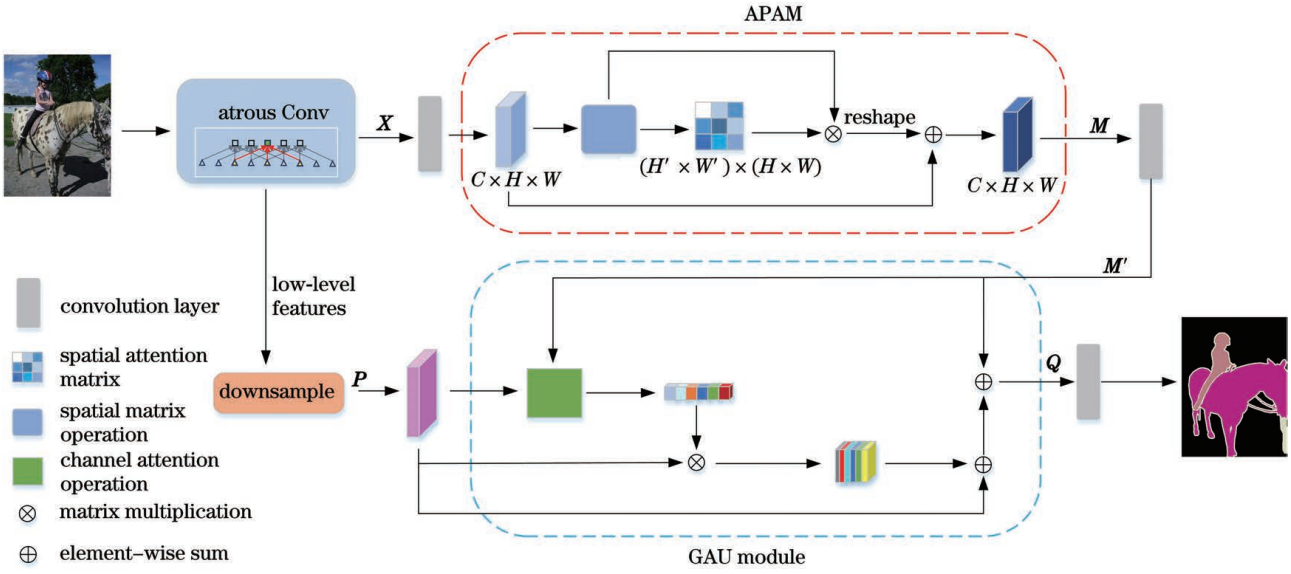

图 1　基于注意力编码的轻量化网络

Fig. 1　Lightweight network based on attention coding

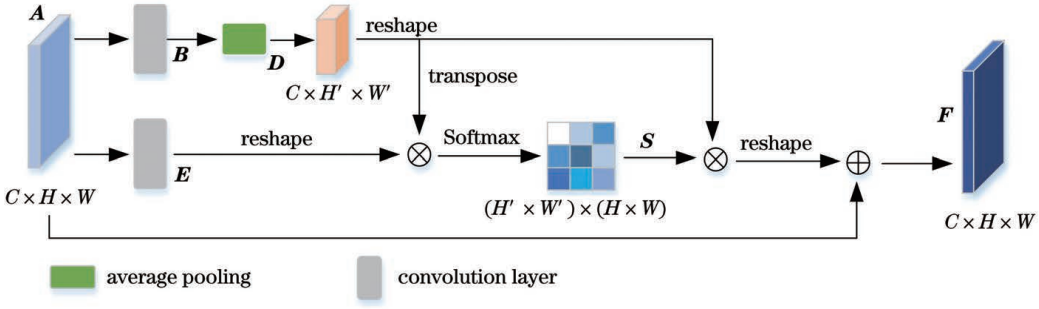

图 2　APAM 的结构

Fig. 2　Structure of the APAM

征的数量且 $N' \ll H \times W$。同时，将 $\boldsymbol{E}$ 重建为 $\boldsymbol{R}^{C\times N}$，$N=H\times W$ 为特征的数量。最后，将 $\boldsymbol{D}$ 的转置与 $\boldsymbol{E}$ 相乘，再与参数 α 相乘后送入 Softmax 层计算空间注意力图 $\boldsymbol{S}\in\boldsymbol{R}^{N'\times N}$，可表示为

$$\boldsymbol{S}_{ji}=\frac{\exp(\alpha \boldsymbol{E}_i\boldsymbol{D}_j)}{\sum_{i=1}^{N'}(\alpha \boldsymbol{E}_i\boldsymbol{D}_j)},\tag{1}$$

式中，$\boldsymbol{S}_{ji}$ 为第 i 个位置对第 j 个基组的影响。

将 $\boldsymbol{D}$ 和 $\boldsymbol{S}$ 相乘后的结果重建为 $\boldsymbol{R}^{C\times H\times W}$，并乘以一个调整参数 β，然后与特征 $\boldsymbol{A}$ 对应元素相加，得到最终的输出结果 $\boldsymbol{F}\in\boldsymbol{R}^{C\times H\times W}$，可表示为

$$\boldsymbol{F}_j=\beta\sum_{i=1}^{N'}(\boldsymbol{S}_{ji}\boldsymbol{D}_i)+\boldsymbol{A}_j,\tag{2}$$

式中，α 和 β 被初始化为 1.0 并通过学习逐渐分配更合适的权重。可以发现，$\boldsymbol{F}$ 中每个位置的特征均为 $\boldsymbol{D}$ 和 $\boldsymbol{S}$ 相乘并重建后的特征与原始特征的加权和，从而使相似的语义特征相互增强，也增加了类内的紧凑性和语义的一致性。通过滤除用于计算注意力图的无用基组，一定程度上减小了计算的复杂度和 GPU 的占用内存。

2.3　全局注意力上采样模块

语义分割任务中的解码结构对于恢复高分辨率图像是非常重要的，FCN、金字塔场景解析(PSP)网络、DeepLab v3+等均采用线性插值上采样方式获取高分辨率图像，缺乏对图像细节信息的恢复。而 SegNet、U-net、Refinenet 等网络均采用复杂的解码器模块帮助高层特征恢复细节信息，需要耗费大量的时间和 GPU 内存。为了解决这些问题，提出了一种基于全局信息引导低层特征恢复高分辨率图像的高效上采样模块，其结构如图 3 所示。GAU 模块先改变低层特征每个通道的权重，以增强类内特征的一致性，并将低层特征 $\boldsymbol{A}$ 和高层特征 $\boldsymbol{B}$ 结合进行上采样，以获取高分辨率图像。首先，GAU 模块结合低层、高层特征并通过全局池化获得全局信息。然后，将全局信息分别经过带有 ReLU 和 Sigmod

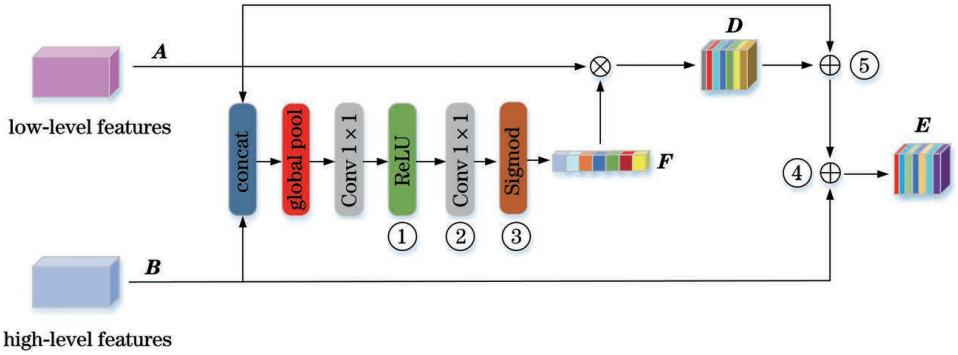

图 3　GAU 模块的结构

Fig. 3　Structure of the GAU module

的 1×1 卷积层，获取低层特征的通道权重向量 $\boldsymbol{F}$，并将其与低层特征 $\boldsymbol{A}$ 相乘后得到加权特征 $\boldsymbol{D}$。最后，将 $\boldsymbol{D}$ 与 γ 倍的 $\boldsymbol{A}$、δ 倍的 $\boldsymbol{B}$ 相加，得到最终的输出 $\boldsymbol{E}$，其中 $\{\boldsymbol{A},\boldsymbol{B},\boldsymbol{D},\boldsymbol{E}\}\in\boldsymbol{R}^{C\times H\times W}$，可表示为

$$\boldsymbol{E}_{mn}^{k}=\gamma\boldsymbol{A}_{mn}^{k}+\delta\boldsymbol{B}_{mn}^{k}+\boldsymbol{D}_{mn}^{k},\tag{3}$$

式中，$k\in\{1,\cdots,K\}$ 为通道数，m、n 为像素的位置，参数 γ 和 δ 都被初始化为 1.0 并通过逐步学习分配更合适的权重。可以发现，最终的输出 $\boldsymbol{E}$ 结合了低层特征的细节信息和高层特征的语义信息，使上采样恢复的图像分辨率更高。

3　实验结果与分析

3.1　数据集与实现细节

为了验证本算法的有效性，在 PASCAL VOC2012 数据集上进行了大量实验。PASCAL VOC2012 数据集包含 20 个前景目标类别和 1 个背景类别，初始数据集包含 1464 张用于训练的图像、1449 张用于验证的图像以及 1456 张用于测试的图像。Cityscapes 数据集[20-21]包含 5000 张精细标注的图像，包括 2975 张训练图像、500 张验证图像和 1525 张测试图像。城市风光数据集的图像分辨率均为 2048 pixel×1024 pixel，每个像素点对应一个类别标签，共分为 19 类（包括车、建筑和行人等）。实验将验证集的图像尺寸均裁剪为 512 pixel×512 pixel，使用的评价指标为单尺度输入 21 个类的平均像素交并比（mIoU）。

实验的硬件环境：GPU 为 GeForceGTX 1080Ti，框架为 Pytorch 框架。用批量随机梯度下降算法进行训练，为了更公平地对比不同算法的性能，采用了多元学习策略。其中，初始学习率在每次迭代后都乘以因子 $(1-I/M_{\text{iter}})^{P}$，以减小学习率。其中，$I$ 为当前迭代次数，M_{iter} 为总迭代次数，$P=0.9$。为保证实验结果的公平性，网络训练过程中的超参数采用了 Fu 等[14]的设置。初始学习率为 0.0001，动量系数为 0.9，权重衰减系数为 0.0001。由于计算资源受限，最小批量为 8。将类别中每个像素位置的交叉熵损失之和作为本算法的损失函数，对于 PASCAL VOC2012 数据集训练 50 轮，并在 APAM 的末端使用了辅助监督。训练过程中采用随机裁剪成尺寸为 512 pixel×512 pixel 的图像，随机左右翻转和在 0.5～2.0 之间进行随机缩放进行数据增广。

3.2　消融实验

将 ResNet101[22]最后两个下采样层改为空洞卷积层，然后上采样到原始图像尺寸的网络设为 Baseline。在 APAM 中，先排除一些无价值的基组，然后再计算注意力图，并捕获长距离的依赖关系信息。表 1 为计算注意力图所需的基组数量，其中，Ours-1～Ours-5 表示在 Baseline 的基础上增加不同尺寸的 APAM。可以发现，增加 APAM 后，语义分割图像的 mIoU 至少提升了 13.5 个百分点，这充分证明了 APAM 提取的长距离依赖关系对网络的影响，也验证了该模块的有效性。此外，随着基组数

表 1　APAM 基组数量对分割精度的影响

Table 1　Influence of APAM basis set number on segmentation accuracy

Algorithm	Base size	mIoU /%
Baseline	–	69.8
Ours-1	36	83.3
Ours-2	49	83.4
Ours-3	64	83.5
Ours-4	81	83.6
Ours-5	100	83.6

量的增加，网络的分割精度也有一定的增加，这表明有效基组数量的增加可提升语义分割的效果。相比 Ours-4，Ours-5 的基组数量虽然有所增加，但其分割精度并没有增加，这表明有效基组数量有一个饱和值，在饱和值之下分割精度随有效基组数量的增加而增加；在饱和值之上分割精度随有效基组数量的增加没有变化。相比非局部神经网络模块，APAM 降低了计算复杂度和 GPU 的占用内存，且捕获的长距离依赖关系更适用于图像分割。

为了验证 APAM 的性能，将 APAM 与现有注意力模块的性能进行了对比，结果如表 2 所示。其中，SGE 为空间分组增强模块，MHA 为混合高阶注意力模块。受限于计算资源，实验在以 ResNet50 为主体的网络上进行训练和测试。在 APAM 中，挑选基组的方式有 max 和 average 两种，对比发现，采用 average 方式挑选的基组分割精度比 max 方式提升了 1.06 个百分点。还可以发现，相比已有的注意力模块，本算法的分割精度最高，最大可提升 12.09 个百分点。相比基于非局部神经网络的位置注意力模块（PAM）和通道注意力模块（CAM），采用 average 挑选方式的 APAM 分割精度至少提升了 0.28 个百分点，这表明 APAM 排除无效基组的干扰后，不仅可以降低计算复杂度，还可以提升语义分割的准确率。

表 2 不同注意力模块的分割结果

Table 2 Segmentation results of different attention modules

Algorithm	Base size	mIoU /%
MHA[23]	–	59.95
SGE[24]	–	58.90
PAM	–	70.71
CAM	–	70.53
Ours(average)	64	70.99
Ours(max)	64	69.93

确定 APAM 的基组数量和选基方式后，为进一步提高网络性能，再次对 GAU 模块的结构进行分析。首先，在以 ResNet101 为主体的网络末端添加了 APAM-Our（average）进行编码；然后，用 GAU 模块进行解码，实验结果如表 3 所示。其中，CF 为不同通道的计算方式，LCF 为图 3 中第④步的融合方式。相比图 3 中的结构 F5，F1 没有步骤①、②，且步骤⑤中没有与 ***A*** 融合；F2 没有步骤①、②；F3 没有步骤①；F4 的步骤⑤中没有与 ***A*** 融合；F6 没有步骤①、②且步骤③采用 Softmax 层，步骤⑤中没有融合 ***A***，步骤④中没有融合 ***B***。对比发现，添加 APAM 后，本算法的语义分割精度从 83.6% 提升到了 84.9%，这表明融合低层细节信息的上采样模块更有利于重构高分辨率图像。而在 APAM 中，使用对应元素相加的融合方式分割精度比串联方式提高了 3.1 个百分点，且占据的 GPU 内存更少。虽然 F5 与 F6 的分割精度相同，但 F6 中的 Softmax 操作会增加网络的计算复杂度，因此 APAM 采用了 F5 的上采样方式。

表 3 APAM 的结构分析

Table 3 Structural analysis of the APAM

CF	LCF	mIoU /%
F1	concat	84.6
F2	concat	83.2
F3	concat	83.3
F4	concat	81.4
F4	sum	84.5
F5	sum	84.9
F6	sum	84.9

3.3 实验对比

为了验证网络的整体性能，将本算法与已有算法的计算复杂度、参数量、GPU 内存及分割精度进行了对比。本算法用以 ResNet101 为主体、输入图像尺寸为 512 pixel×512 pixel、最小批次为 1 的超参设置对现存的语义分割网络进行测试，结果如表 4 所示。可以发现，相比 FCN，本算法虽然计算复杂度和参数量略有增加，但分割准确率提升了 15.1 个百分点。相比 U-net 与 SegNet，本算法的分割精度分别提升了 14.1 和 13.8 个百分点。相比分割精度相近的DANet，本网络的GFLOPS降低了

表 4 不同算法的性能参数

Table 4 Performance parameters of different algorithms

Algorithm	GFLOPS	Params /M	Memory /G	mIoU /%
FCN	216.0	54.0	7.7	69.8
U-net	262.2	34.5	8.7	70.8
SegNet	244.6	32.4	8.9	71.1
DeepLab v2	251.7	44.6	8.0	71.6
PSP	254.7	67.6	8.1	80.9
DANet	275.4	68.5	9.3	85.1
Ours	228.8	67.7	8.1	84.9

16.9%，占用的 GPU 内存减少了 12.9%。这表明对于高分辨率图像的恢复，采用 GAU 比复杂的上采样方式更有效。相比其他分割网络，本算法重构的高分辨率图像效果也更好，这也验证了基于注意力编码的轻量化网络的高效性和有效性。

用尺寸为 512 pixel×512 pixel 的图像训练网络，然后在 PASCAL VOC2012 验证集中测试不同算法的分类性能，结果如表 5 所示。可以发现，本算法对不同类别的 mIoU 最高，可达到 84.9%。对于轮廓更精细的物体类别，如 boat、car、chair、cow、table、soft 和 plant，本算法的语义分割精度有显著提升，验证了本算法中捕获长距离依赖关系信息的 APAM 以及重构生成高分辨率图像的 GAU 模块的有效性。不同算法在 Cityscapes 验证集上的分类结果如表 6 所示，可以发现，本算法对所有类的 mIoU 最高，可达到 71.6%，这也验证了本算法的有效性。

表 5 不同算法在 PASCAL VOC2012 验证集上的分类结果

Table 5 Classification results of different algorithms on the PASCAL VOC2012 validation set unit: %

Algorithm	FCN	DeepLab v2	DPN[25]	DeepLab v3+	PSP	ResNet38[26]	EncNet[27]	Ours
aero	82.4	84.4	87.7	88.0	87.4	**94.4**	94.1	91.6
bike	47.4	54.5	59.4	56.3	56.3	**72.9**	69.2	57.9
bird	81.2	81.5	78.4	86.3	85.7	94.9	**96.3**	90.0
boat	68.6	63.6	64.9	69.4	79.4	68.8	76.7	**85.5**
bottle	75.3	65.9	70.3	72.2	73.8	78.4	**86.2**	82.5
bus	81.3	85.1	89.3	90.3	92.3	90.6	**96.3**	95.0
car	79.9	79.1	83.5	85.7	87.3	90.0	90.7	**90.8**
cat	81.6	83.4	86.1	89.6	92.3	92.1	**94.2**	94.0
chair	33.7	30.7	31.7	28.9	53.3	40.1	38.8	**53.8**
cow	68.4	74.1	79.9	85.9	90.4	90.4	90.7	**93.7**
table	52.3	59.8	62.6	59.3	75.2	71.7	73.3	**78.0**
dog	76.4	79.0	81.9	84.2	87.3	89.9	90.0	**93.0**
horse	64.9	76.1	80.0	80.2	85.9	**93.7**	92.5	91.4
mbike	73.4	83.2	83.5	84.2	83.8	**91.0**	88.8	87.1
person	81.2	80.8	82.3	82.8	84.5	**89.1**	87.9	88.2
plant	56.7	59.7	60.5	56.0	68.1	71.3	68.7	**74.1**
sheep	69.7	82.2	83.2	78.5	87.0	90.7	**92.6**	91.3
sofa	50.9	50.4	53.4	51.6	73.0	61.3	59.0	**75.8**
train	78.5	73.1	77.9	84.5	91.1	87.7	86.4	**92.8**
tv	70.1	63.7	65.0	69.6	71.5	78.1	73.4	**80.6**
mIoU	69.8	71.6	74.1	75.1	80.9	82.5	82.9	**84.9**

为了定性分析本算法与基础算法的分割效果，用两种算法对不同的图像进行分割，结果如图 4 所示。图 4(a)为 PASCAL VOC2012 验证集中的原始图像，图 4(b)为 PASCAL VOC2012 验证集的语义标签，图 4(c)和图 4(d)分别为基础算法和本算法的语义分割的结果。可以发现，相比基础算法，本算法的分割精度更高，如图像中的桌子部分、男士的衣服处，原因是基础算法只使用了长距离的依赖信息，缺少局部细节特征，而本算法采用低层细节特征和高层语义特征共同进行上采样，可增强网络对物体轮廓细节的分割性能。且基础算法将图像中自行车的一部分误分类成摩托车，原因是自行车的外观特

表 6 不同算法在 Cityscapes 验证集上分类结果

Table 6 Classification results of different algorithms on the Cityscapes validation set unit: %

Algorithm	FCN	PSP	DenseASPP[28]	DANet	Ours
road	95.1	96.4	97.3	97.2	**97.4**
sidewalk	67.8	74.4	78.1	77.8	**79.4**
building	88.5	89.1	89.5	89.8	**90.0**
wall	50.5	52.9	56.0	56.1	**57.0**
fence	44.6	47.9	48.5	48.6	**51.1**
pole	35.6	39.9	40.3	40.8	**43.5**
traffic light	47.0	51.9	52.8	53.0	**53.4**
Traffic sign	60.4	62.4	65.3	65.2	**66.5**
vegetation	88.6	89.4	89.6	89.7	**89.8**
terrain	55.6	57.6	60.5	60.7	**60.9**
sky	91.4	92.0	92.3	92.4	**92.7**
person	68.8	70.4	71.4	71.9	**72.8**
rider	47.9	49.9	52.0	52.2	**53.4**
car	90.3	91.4	92.1	**92.4**	92.4
truck	73.8	73.9	**83.0**	82.8	79.3
bus	73.6	75.8	80.2	79.4	**81.9**
train	62.8	66.4	70.4	70.8	**74.3**
motocycle	51.7	55.0	58.4	58.9	**58.6**
bicycle	63.1	63.6	65.5	65.8	**66.6**
mIoU	66.2	68.4	70.7	70.8	**71.6**

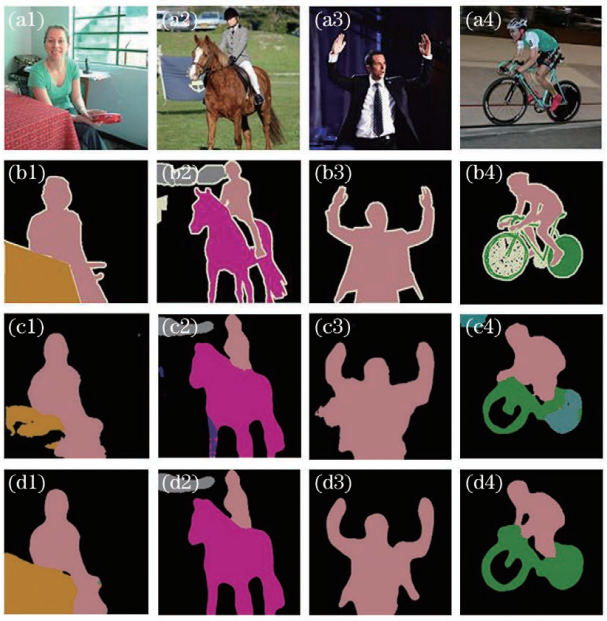

图 4 不同算法的分割结果。(a)原始图像;(b)真实语义标签;(c)基础算法;(d)本算法

Fig. 4 Segmentation results of different algorithms. (a) Original image; (b) real semantic label; (c) basic algorithm; (d) our algorithm

征与摩托车类似,基础算法使用长距离的依赖关系信息时相似类别存在信息干扰。

4 结 论

在基于注意力编码的轻量化网络中,使用了高效的注意力机制捕获长距离依赖关系信息和全局信息上采样重构高分辨率分割图像,不仅降低了算法的计算复杂度与 GPU 的占用内存,还增强了网络对类别的区分能力,从而提高了网络的语义分割性能。在 PSCAL VOC2012 数据集上的实验结果表明,基于注意力编码的轻量化网络分割精度可达到 84.9%,比基础算法提高了 15.1 个百分点;相比分割精度相近的 DANet,本网络的 GFLOPS 降低了 16.9%,占用的 GPU 内存减少了 12.9%,这表明该网络可在未来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] Goodfellow I, Bengio Y, Courville A. Deep learning [M]. The MIT Press, 2016.

[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) [2020-09-02]. 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] 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] 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.

[15] Yuan Y H, Huang L, Guo J Y, et al. OCNet: object context network for scene parsing[EB/OL]. (2018-09-16) [2020-09-02]. https//arxiv. org/abs/1809.00916.

[16] 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.

[17] 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.

[18] Lin G S, Milan A, Shen C H, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5168-5177.

[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] 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.

[23] Chen B H, Deng W H, Hu J N. Mixed high-order attention network for person re-identification[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 371-381.

[24] Li X, Hu X L, Yang J. Spatial group-wise enhance: improving semantic feature learning in convolutional networks[J]. (2018-11-28)[2020-09-02]. https://arxiv.org/abs/1905.09646.

[25] 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.

[26] 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.

[27] 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.

[28] Yang M K, Yu K, Zhang C, et al. DenseASPP for semantic segmentation in street scenes[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: 3684-3692.