文章编号：2095-2163(2023)10-0001-09 中图分类号：TP391.41 文献标志码：A

基于 DCF-DeepLab 网络的图像语义分割研究

蔡江海[1]，黄成泉[2]，杨贵燕[1]，罗森艳[1]，王顺霞[1]，周丽华[1]
（1 贵州民族大学 数据科学与信息工程学院，贵阳 550025；2 贵州民族大学 工程技术人才实践训练中心，贵阳 550025）

摘 要：针对 DeepLabv3 中存在的图像小目标信息易丢失、训练时效长、目标边界分割粗糙等问题，提出了一种基于双注意力交叉融合的 DCF-DeepLab 图像语义分割网络。该网络设计了基于注意力机制的特征融合模块，分别在 2、4、8 倍下采样特征图上使用该模块，融合浅层特征以弥补深层特征的不足；在主干特征提取网络部分引入轻量级网络 MobileNetV3-Large，减少模型推理时间；在 ASPP 模块嵌入多个注意力模块和串联结构，增强网络细节特征提取能力。通过在 PASCAL VOC 2012 数据集上的实验结果表明，该网络可有效提升图像语义分割性能，具有一定应用潜力。

关键词：DeepLabv3；特征融合；注意力机制；语义分割

Research on image semantic segmentation based on DCF-DeepLab network

CAI Jianghai[1], HUANG Chengquan[2], YANG Guiyan[1], LUO Senyan[1], WANG Shunxia[1], ZHOU Lihua[1]
(1 College of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China;
2 Engineering Training Center, Guizhou Minzu University, Guiyang 550025, China)

【Abstract】 Aiming at the problems of information loss on small target, long training time, and coarse target boundary segmentation, this paper proposes an image semantic segmentation network named DCF-DeepLab (Double Cross-attention Fusion DeepLab) based on double cross-attention fusion. In our method, a feature fusion module based on attention mechanism is designed which is used to fuse shallow features to compensate for the lack of deep features on 2, 4, and 8-fold down-sampled feature maps, respectively. Besides, the lightweight network MobileNetV3-Large is introduced as the backbone for feature extraction, which reduces the model inference time. The multiple attention modules and tandem structures are embedded in the ASPP module to enhanced network detail feature extraction. The experimental results on the PASCAL VOC 2012 dataset show that the designed model can effectively improve the image semantic segmentation performance and has application potential.

【Key words】 DeepLabv3; feature fusion; attention mechanism; semantic segmentation

0 引 言

随着计算机视觉技术的飞速发展，图像语义分割已成为该领域研究的热点之一。基于深度学习的语义分割方法相较于传统图像处理方法，性能得到了极大的提升，被广泛应用于自动驾驶、医学图像处理、人脸识别等领域。Long 等[1]提出的全卷积网络(FCN)是一种端到端的语义分割网络，可以有效应用于图像语义分割研究。然而，连续的池化和下采样操作容易引起浅层语义信息丢失，进而导致小目标信息丢失和边界分割模糊。Ronneberger 等[2]提出了 Unet 语义分割模型，该模型引入编码器-解码器结构，利用上采样和下采样过程进行跳跃连接，实现了更高精确的分割。Fu 等[3]提出了引入空间注意力和通道注意力的分割网络 DANet，有效提升了模型的性能。后续，相关研究者又陆续提出了更好的兼顾精度和速度的图像语义分割模型，如 HMANet[4]、STLNet[5] 等。Chen 等[6]在 DeepLabv1 基础上提出了 DeepLabv2，并引入 ASPP(Atrous Spatial Pyramid Pooling)模块实现多尺度的特征提取。之后又相继提出了基于 ASPP 模块的

基金项目：国家自然科学基金(62062024)；贵州省省级科技计划项目(黔科合基础-ZK[2021]一般 342)；贵州省研究生教育教学改革重点项目(黔教合 YJSJGKT[2021]018)；贵州省教育厅自然科学研究项目(黔教技[2022]015)。

作者简介：蔡江海(1999-)，男，硕士研究生，主要研究方向：统计建模与模式识别；黄成泉(1976-)，男，博士，教授，硕士生导师，主要研究方向：计算机应用技术、图像处理；杨贵燕(1997-)，女，硕士研究生，主要研究方向：统计建模与模式识别；罗森艳(1999-)，女，硕士研究生，主要研究方向：统计建模与模式识别；王顺霞(1999-)，女，硕士研究生，主要研究方向：统计建模与模式识别；周丽华(1983-)，女，硕士，副教授，主要研究方向：图像处理、模式识别。

收稿日期：2023-08-11

DeepLabv3 和采用编码器-解码器结构的 DeepLabv3+,实现了更好的图像语义分割。

到目前为止,DeepLab 系列都是在降采样 8 倍尺度上进行预测的,边界分割效果不甚理想。DeepLabv3 网络并没有包含过多的浅层特征,不仅在语义信息和位置信息的平衡上存在连续池化和下采样导致的小目标信息丢失的问题,并且由于该网络是通过多层卷积叠加而成的,存在训练时效长、目标边界分割粗糙等问题。为此,Zhu 等[7]通过注意力式可分离卷积的编码器-解码器结构,在多尺度特征上有效均衡了训练效率和分割精度。Wang 等[8]利用基于注意力机制的优势,较好地克服了因下采样导致的浅层细节信息丢失的问题,但模型参数量大,训练时效长,实用性较低。

针对以上问题,本文提出了融合多模块的 DCF-DeepLab(Double Cross-attention Fusion DeepLab)语义分割网络。首先,设计了基于双注意力交叉融合的特征融合 DAFM(Double Attention Fusion Moudle)模块,以融合浅层特征弥补深层特征的不足,并将其应用于主干特征提取网络的 2、4、8 倍下采样的特征图上,充分提取小目标特征信息,实现特征图跨模块的融合;其次,在主干特征提取部分引入轻量级网络 MobileNetV3-Large,加速整体网络的训练速率;最后,通过嵌入 DAFM 模块、注意力模块和串联结构得到 MA-ASPP(Multiple Attention ASPP)模块,实现多尺度信息编码,增强图像目标边缘的细节特征提取能力。DCF-DeepLab 语义分割网络从整体上精细了语义分割结果,提升了语义分割性能。

1 DCF-DeepLab 网络

1.1 DeepLabv3 网络

DeepLabv3 网络主要由两部分组成:

(1)在编码端使用 Resnet101[9]残差网络模型作为基本特征提取的主干网络,得到有效特征图,再利用 ASPP 模块(由 1 个 1×1 卷积、1 个全局池化层以及 3 个不同空洞率的空洞卷积共同组成)进一步提取特征得到多尺度特征图;

(2)在解码端将多尺度特征图进行拼接和 1×1 的卷积操作得到特征图(该特征图与基本特征提取主干网络下采样得到的特征图的通道数相同),最后通过上采样,将所得特征图还原回与原始图像大小相同的尺寸,得到语义分割的结果。DeepLabv3 整体结构如图 1 所示。

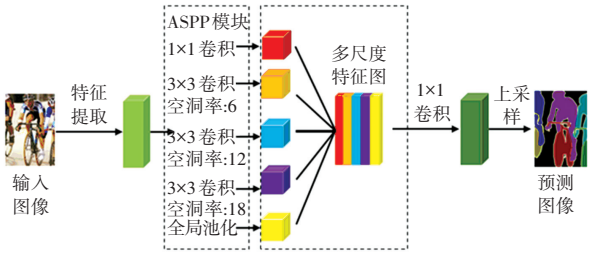


图 1 DeepLabv3 整体结构

Fig. 1 Overall structure of DeepLabv3

1.2 MobileNetV3_large 网络

为了使 DCF-DeepLab 网络高效地训练出预期的结果,提升网络训练参数的速度,解决因训练时间过长可能导致目标分割结果不佳的问题。本文采用具有高效且网络参数量、运算量小的 MobileNetV3_large 作为本文网络在编码端的主干特征提取网络。MobileNetV3_Large 网络结构见表 1。

表 1 MobileNetV3_Large 网络结构

Tab. 1 Structure of MobileNetV3_Large network

Input	Operator	Exp size	#out	SE	NL	*s*
$224^2\times3$	Conv2d	-	16	-	HS	2
$112^2\times16$	bneck, 3×3	16	16	-	RE	1
$112^2\times16$	bneck, 3×3	64	24	-	RE	2
$56^2\times24$	bneck, 3×3	72	24	-	RE	1
$56^2\times24$	bneck, 5×5	72	40	√	RE	2
$28^2\times40$	bneck, 5×5	120	40	√	RE	1
$28^2\times40$	bneck, 5×5	120	40	√	RE	1
$28^2\times40$	bneck, 3×3	240	80	-	HS	2
$14^2\times80$	bneck, 3×3	200	80	-	HS	1
$14^2\times80$	bneck, 3×3	184	80	-	HS	1
$14^2\times80$	bneck, 3×3	184	80	-	HS	1
$14^2\times80$	bneck, 3×3	480	112	√	HS	1
$14^2\times112$	bneck, 3×3	672	112	√	HS	1
$14^2\times112$	bneck, 5×5	672	160	√	HS	2
$7^2\times160$	bneck, 5×5	960	160	√	HS	1
$7^2\times160$	bneck, 5×5	960	160	√	HS	1
$7^2\times160$	conv2d, 1×1	-	960	-	HS	1
$7^2\times960$	pool, 7×7	-	-	-	-	1
$1^2\times960$	conv2d, 1×1 ,NBN	-	1 280	-	HS	1
$1^2\times1\ 280$	conv2d, 1×1 ,NBN	-	k	-	-	1

在 MobileNetV3_large 网络结构中,“Input”表示输入当前层特征矩阵的尺寸;“Operator”表示输入特征矩阵在本层中进行的操作,主要由普通二维卷

积操作、多个倒残差模块(bneck)操作以及池化操作组成，其中“NBN”表示不包含 BN 层；“Exp size”表示利用 1×1 的卷积核扩展后的通道数；“#out”表示输出特征矩阵的通道数，其中 k 表示类别数量；“SE”表示通道注意力机制；“NL”表示激活函数，其中“HS”为 H-Swish 激活函数、“RE”为 RELU 激活函数；“s”表示步距。

1.3　注意力模块

在计算机视觉领域，通道注意力机制[10]被广泛应用于各类研究。SENet(Squeeze-and-Excitation Network)模块针对不同语义信息的属性特征，在通道上实现了更好的信息获取和分割效果。SENet 模块结构如图 2 所示。

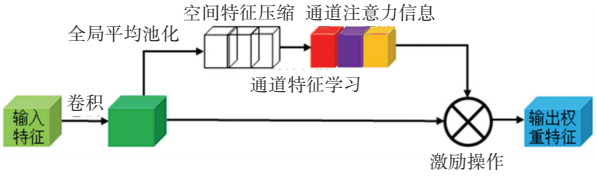


图 2　SENet 模块结构

Fig. 2　Structure of SENet module

在全局平均池化中将全局空间信息转换成通道统计信息，并执行压缩操作得到空间特征压缩量。假设输入特征统计量 $U=\{u_1,u_2,\cdots,u_k,\cdots,u_{\alpha-1},u_\alpha\}$，其中特征通道量 $u_k\in R^{H\times W}$，输出特征统计量经过全局平均池化得到，其中第 k 个元素值为

$$z_k=F_{sq}(u_k)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}u_k(i,j) \tag{1}$$

为了利用压缩操作中聚合的信息，完全捕获依赖通道的信息，执行激活操作。实现上，通过引入全连接层对输出特征统计量 Z 进行通道依赖关系编码，学习通道之间的非线性交互作用，并结合 Sigmoid 函数进行通道赋权，获取通道注意力信息[11]。最后，将得到的权重系数应用于输入特征中，得到通道注意力的输出 $\tilde{X}$ 其计算关系为

$$\tilde{X}=\sigma(P_{fcl}Z)U \tag{2}$$

其中，$\sigma(\cdot)$ 为 Sigmoid 激活函数，P_{fcl} 为全连接层参数。

CBAM(Convolutional Block Attention Module)模块分别在通道和空间维度上进行注意力操作，以获得全面的注意力信息，引导模型进行权重分配和信息指引。CBAM 模块结构如图 3 所示。

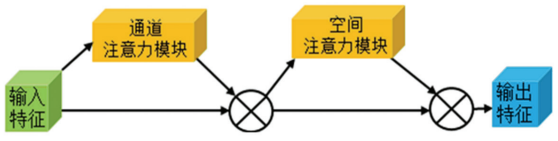


图 3　CBAM 模块结构

Fig. 3　Structure of CBAM module

假设 $F\in\mathbb{R}^{H\times W\times C}$ 表示输入的特征图，$M_c\in\mathbb{R}^{1\times 1\times C}$ 表示一维的通道注意力，$M_s\in\mathbb{R}^{H\times W\times 1}$ 表示二维的空间注意力。经过 CBAM 模块依次推导出通道注意力和空间注意力的映射，计算过程如下：

$$F'=M_c(F)\otimes F \tag{3}$$

$$F''=M_s(F')\otimes F' \tag{4}$$

式(3)中，将原始特征图 F 与其经过通道注意力模块操作后的结果进行元素相乘得到 F'；式(4)中，将输出 F' 与经过空间注意力模块做特征提取后的结果进行元素相乘，得到最终的输出结果 F''。

为了提高通道注意力模块中网络的表征能力，首先，对输入的特征进行最大池化操作(MaxPool)和平均池化操作(AvgPool)聚合特征图的空间信息；其次，将池化后的特征信息输入到共享全连接层中以生成通道注意力图[12]，其中共享全连接层由具有隐藏层的多层感知器(MLP)组成，并将输出的特征进行相加融合；最后，利用 Sigmoid 函数进行激活操作，得到输入特征层中每一个通道的权重值。通道注意力模块结构如图 4 所示。

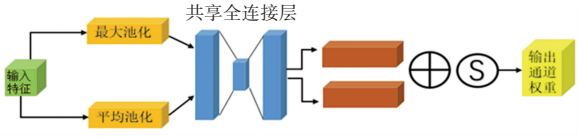


图 4　通道注意力模块结构

Fig. 4　Structure of channel attention module

假设 $F\in\mathbb{R}^{H\times W\times C}$ 表示输入的特征图，其经过通道注意力模块的计算过程为

$$\begin{aligned}M_c(F)=\sigma(&F_{MLP}(\mathrm{MaxPool}(F))+\\&F_{\mathrm{MLP}}(\mathrm{AvgPool}(F)))=\sigma(W^1(W^0(F^c_{\max}))+\\&W^1(W^0(F^c_{avg})))\end{aligned} \tag{5}$$

式中：$\sigma(\cdot)$ 表示 Sigmoid 激活函数，F_{MLP} 表示全连接层，$W_0\in\mathbb{R}^{C/r\times C}$，$W_1\in\mathbb{R}^{c\times c/r}$，其中 r 为缩减比率，$F^c_{\max}$、F^c_{avg} 分别表示最大池化特征和平均池化特征，MLP 权重 W_0、W_1 对于两个输入都是共享的。

在空间注意力模块中，为了计算空间关注度并聚合空间特征信息[13]，在通道轴上应用平均池化操作和最大池化操作，并将其连接起来生成有效的特

征描述,再利用通道数为 1 的卷积核进行降维,最后利用 Sigmoid 函数进行激活操作,获得输入特征层的每一个特征点的权重值。空间注意力模块结构如图 5 所示。

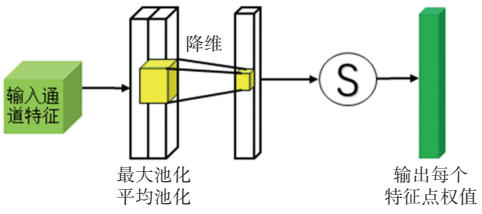


图 5 空间注意力模块结构

Fig. 5 Structure of spatial attention module

空间注意力模块的计算过程为

$$M_s(F)=\sigma(f([\mathrm{AvgPool}(F);\mathrm{MaxPool}(F)]))=\sigma(f([F^s_{avg};F^s_{\max}])) \quad (6)$$

式中:$\sigma(\cdot)$ 表示 sigmoid 激活函数,f 表示滤波器的卷积运算,$F^s_{\max}$、F^s_{avg} 分别表示最大池化特征和平均池化特征。

1.4 DAFM 模块

SENet 模块给每个特征通道上的信号都赋予了一定的权重,对特征通道上的背景信息和前景目标信息有了更明确的选择,强化了感兴趣的特征,增强了特征通道下特定语义的响应能力。CBAM 模块在通道和空间维度上对不同位置元素间的关系进行建模,其兼顾通道注意力和空间注意力的优势,获得更可靠的权重信息,增强了模型的表征能力。

计算机视觉领域常用的特征融合方式,是在同一张特征图上分别进行两种注意力机制操作,并进行结果的融合[14],其主要不同之处在于特征融合方式。考虑到语义分割任务中对图像分辨率的影响,通常情况下,分辨率低的深层特征图采取通道注意力操作,其关注点放在相关的特征通道上;分辨率高的浅层特征图采取空间注意力操作,提取特征图中关于空间位置的关键信息。因此,本文综合注意力机制对不同分辨率的深浅层特征图提取特征的优势,通过嵌入 SENet 和 CBAM 模块,得到有效融合浅层空间细节信息和深层高级语义线索的 DAFM 模块,如图 6 所示。

假设:在 DAFM 模块中,输入的低分辨率深层特征图为 U_{LR},尺寸大小为 $H_1\times W_1$;高分辨率浅层特征图的输入为 U_{HR},尺寸大小为 $H\times W$。首先,根据公式(7),将 U_{LR} 进行上采样操作得到 U'_{LR} 使得 U'_{LR} 尺寸大小与 U_{HR} 相同,均为 $H\times W$。

$$U'_{LR}=F_{UP}(U_{LR}) \quad U'_{LR}\in H\times W\times C \quad (7)$$

式中:$F_{UP}(\cdot)$ 表示采用双线性插值方法的上采样操作,其次,对 U_{HR} 进行 CBAM 注意力操作,得到 U'_{CHR}。并根据公式(2),对 U'_{LR} 通过 SENet 注意力操作得到权重 W_s,其计算公式为

$$W_s=\{\sigma(s_{1,1}),\sigma(s_{1,2}),\cdots,\sigma(s_{i,j}),\cdots,\sigma(s_{H,W})\} \quad (8)$$

再将权重 W_s 与 U'_{CHR} 相乘,即根据公式(9),得到 U'_W:

$$U'_W=\{\sigma(s_{1,1})U^{1,1}_{CHR},\sigma(s_{1,2})U^{1,2}_{CHR},\cdots,\sigma(s_{H,W})U^{H,W}_{CHR}\} \quad (9)$$

最后,将 U'_W 与 U_{HR} 相加,并进行 1×1 的卷积核降维操作,得到最终的输出特征图,即

$$U_F=c(U'_W+U_{HR}) \quad (10)$$

式中:c 表示 1 × 1 卷积操作。

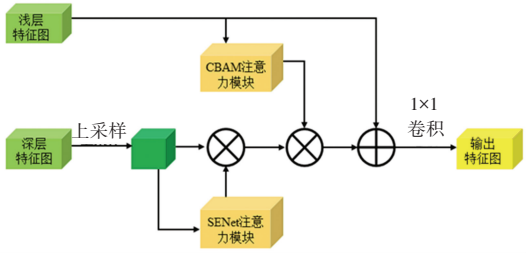


图 6 DAFM 模块结构

Fig. 6 Structure of DAFM module

在 DAFM 模块中,浅层特征图和深层特征图为 DAFM 模块的输入,深层特征图经过上采样完成,并经过 SENet 注意力模块处理后,与经过 CBAM 注意力模块处理后的浅层特征图进行像素级的相乘操作,最后经过相加以及 1×1 卷积降维操作,得到最终融合后的输出特征图。

1.5 MA-ASPP 模块

DeepLabv3 网络中的 ASPP 模块是由 1 个 1×1 卷积、1 个全局平均池化层和不同空洞率的空洞卷积[15]简单拼接而成的,存在易失去图像中被忽略的小尺度目标信息的问题,进而降低特征提取能力,导致分割精度不高。因此,在空洞率不变的前提下,提取目标多尺度信息并获得足够大的感受野变得尤为重要。

针对存在的问题,对 ASPP 模块进行了一系列的改进。首先,受 DenseASPP[16] 网络结构的启发,将 3 个不同空洞率的空洞卷积由简单的堆叠变为密集连接的形式,即在原有 3 个空洞卷积并行的基础上增加了串联结构,将空洞率较小的空洞卷积输出和主干网络的输出级联,再依次送入空洞率较大的空洞卷积中。由逐级递增的并行操作,实现更密集

化的像素级采样，增强提取细节特征的能力；其次，对于另外两个分支的卷积和全局平均池化操作，通过嵌入 CBAM 模块以获取更多浅层特征的细节信息；最后，将融合 5 个分支后的特征信息输送到 DAFM 模块中，加强对重要目标信息和细节信息的选择性注意，并结合 1×1 卷积操作，构成了具有强大特征提取能力的像素级 MA-ASPP 模块。MA-ASPP 模块结构如图 7 所示。

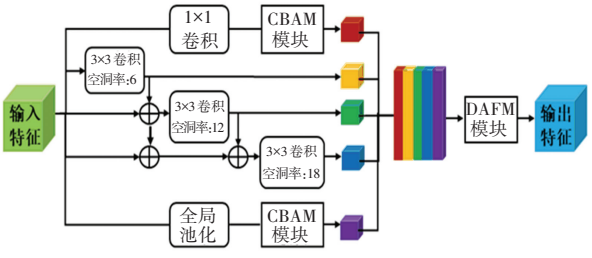


图 7　MA-ASPP 模块结构

Fig. 7　Structure of MA-ASPP module

MA-ASPP 模块中的 3 个空洞卷积分支以密集连接的方式组织，其中任意一层的空洞卷积层输出可表示为

$$y_n = H_{k,r_n}([y_{n-1}, y_{n-2}, \cdots, y_0]) \tag{11}$$

式中：k 表示卷积核的大小，r_n 表示第 n 层的空洞率，$[\cdots]$ 表示拼接操作，$[y_{n-1}, y_{n-2}, \cdots, y_0]$ 表示将所有前一层的输出拼接起来形成的输出。

密集连接的方式不仅可以获得更密集的像素级采样，还可以提供更大的感受野[17]，其计算过程如下：

$$\begin{cases} RF_n = k_n, & n = 1 \\ S_n = S_{n-1} \times s, & n \geqslant 2 \\ RF_n = RF_{n-1} + (k_n - 1) \times S_{n-1} \end{cases} \tag{12}$$

式中：RF_n 表示第 n 层感受野大小，k_n 表示第 n 层卷积核大小，S_n 表示前 n 层的总步长，S_n 表示当前层步长。

由于所采取的空洞卷积步长为 1，因此 S_n 的值恒等于 1，有：

$$RF_n = RF_{n-1} + k_n - 1 \tag{13}$$

在 DeepLabv3 网络的 ASPP 模块中，采取空洞卷积 $rates = \{6,12,18\}$ 并联连接方式的最大感受野大小为

$$RF_n = \max[k_{d=6}, k_{d=12}, k_{d=18}] = k_{d=18} = 37 \tag{14}$$

通过叠加空洞卷积，采取串联并行的连接方式，其所能获得的最大感受野大小为

$$RF_{\max} = RF_n = \sum_{n=1}^{N} k_n - (n-1) \tag{15}$$

因此，可以计算出 DA-ASPP 模块采取 $rates = \{6,12,18\}$ 的空洞卷积所对应的 $RF_{\max}$ 大小为

$$RF_{\max} = k_{d=6} + k_{d=12} + k_{d=18} - (3-1) = 72 \tag{16}$$

通过计算，由式(14)和式(16)已知，DA-ASPP 模块中的 $RF_{\max}$ 值明显大于 ASPP 模块中的 $RF_{\max}$ 值。DA-ASPP 模块通过逐层连接实现信息共享，不同空洞率的空洞卷积相互补充，使其细节信息更加丰富，并增大了感受野的范围，有利于增强特征提取能力。

1.6　DCF-DeepLab 网络结构

本文将主干特征提取网络 MobilenetV3_Large 中 2、4、8 倍下采样获得的浅层特征图与 DAFM 模块连接，并与后续通过多个模块获得的深层特征图进行融合。DCF-DeepLab 整体网络结构如图 8 所示。

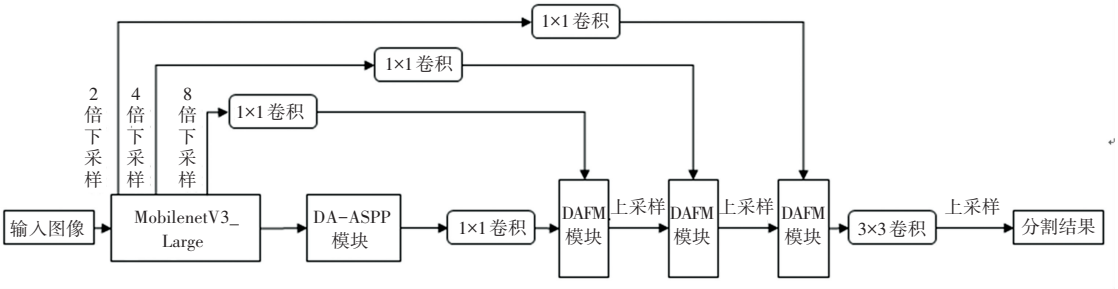


图 8　DCF-DeepLab 整体网络结构

Fig. 8　Overall network structure of DCF-DeepLab

2　实验

2.1　实验环境与数据集

本文各项实验均在 GPU 上完成，软硬件环境配置说明见表 2。

表 2　实验环境配置

Tab. 2　Experimental environment configuration

实验环境	配置
CPU	Intel Core i7-7700K @ 4.2 GHZ
GPU	NVIDIA GeForce GTX 1080Ti
编程语言	Python
操作系统	Ubuntu 16.04
深度学习框架	Pytorch

本文在公开的 PASCAL VOC 2012 数据集上进行了一系列的实验,以验证 DCF-DeepLab 网络的有效性。PASCAL VOC 2012 数据集包括人物、动物、室内外场景、交通工具等 20 个具体前景类别,外加背景共 21 类。分别采用训练集上 1 464 张训练图像、验证集上 1 449 张验证图像以及测试集上 449 张图像进行网络的性能评估。为了加速实验结果的收敛性,在训练阶段采用了基于 COCO 数据集(COCO 数据集是提供 80 个目标类别、91 个材料类别的大型常用数据集)训练得到的预训练权重,且训练时只针对和 PASCAL VOC 2012 相同的类别进行训练。

2.2 训练细节及参数设置

本文采用了在 COCO 数据集上预先训练得到的预训练模型 MobileNetV3_Large 和 ResNet_50 的模型权重,分别对网络 DCF-DeepLab 和 DeepLabv3 进行初始化,以加速收敛。

训练参数设置如下:图像预处理过程中随机水平翻转概率为 0.5,图像训练块大小为 480×480 像素,在验证阶段采取的大小为 520×520 像素;批量大小设置为 6,初始学习率为 0.001,动量设置为0.9,权重衰减为 10^{-4},模型训练的迭代次数为 500 次。

采用 Poly 学习率策略[7],其作为一种指数变换的策略,具体计算公式如下:

$$l_{poly} = lr \times (1 - \frac{iter}{\max\_iter})^{power} \tag{17}$$

式中:lr 表示初始学习率,$iter$ 表示当前迭代步数,$\max\_iter$ 表示最大迭代步数,$power$ 取 0.9。

启用混合精度训练,以减少显存占用,加快网络训练速度。使用交叉熵损失函数[18]计算主输出上的损失,并结合使用全卷积网络分割头辅助训练[19],得到的网络总输出损失为主输出和辅助分类器上的损失加权代数和,比率为 2∶1。

2.3 消融实验

为了验证 DAFM 模块、MA-ASPP 模块和特征融合模块对网络性能的影响,设置了一系列相关的消融实验。所有实验均在 PASCAL VOC 2012 数据集上进行,以语义分割中常用的标准度量平均交并比($MIoU$)作为衡量评估指标,定义为真实值和预测值的交集和并集之比,其计算公式如下:

$$MIoU = \frac{1}{N+1}\sum_{i=0}^{N}\frac{p_{ii}}{\sum_{j=0}^{N}p_{ij} + \sum_{j=0}^{N}p_{ji} - p_{ii}} \tag{18}$$

式中:N 表示前景目标类别个数,P_{ij} 表示真实值 i 被预测为 j 的数量。

2.3.1 DAFM 模块

为了验证本文所提出的 DAFM 模块对网络性能的影响,通过与 SENet 模块和 CBAM 模块基于 DeepLabv3 进行对比实验,实验结果见表 3。

表 3 注意力机制对模型性能的影响

Tab. 3 Effect of attentional mechanism on model performance

基准模型	原始网络	SENet	CBAM	DAFM	*MIoU* / %
DeepLabv3	√	×	×	×	74.2
	√	√	×	×	74.6
	√	×	√	×	74.8
	√	×	×	√	75.1

从表 2 可看出,SENet 模块和 CBAM 模块均可提升网络整体的分割性能,*MIoU* 值较原始网络分别提高了 0.4%和 0.6%。DAFM 模块结合了两者的优势,分割性能效果最好,*MIoU* 值达到 75.1%。因此,本文考虑选取多个 DAFM 模块作为网络解码部分的主体框架,以使网络达到更好的特征表达效果。

2.3.2 MA-ASPP 模块

实验采用 DeepLabv3 作为基准模型,对 ASPP 模块、MA-ASPP 模块以及 DAFM 模块进行组合实验,以验证 MA-ASPP 模块的有效性。实验结果见表 4。

表 4 不同模块组合对模型性能的影响

Tab. 4 Effect of different modules on model performance

基准模型	ASPP	MA-ASPP	DAFM	*MIoU* / %
DeepLabv3	√	×	×	72.3
	×	√	×	73.2
	×	×	√	73.0
	√	×	√	73.3
	×	√	√	74.1

从表 4 可看出,对比 ASPP 模块,MA-ASPP 模块和 DAFM 模块均可提升网络整体的分割性能,*MIoU* 值分别提高了 0.9%和 0.7%。而将 MA-ASPP 模块和 DAFM 模块同时组合进网络,融合浅层特征和深层特征,提取出了更多关键的目标特征信息,网络效果提升显著,*MIoU* 值达到 74.1%。因此,本文最终选择将 MA-ASPP 模块和 DAFM 模块同时组合进网络。

2.3.3 特征融合模块

在编码端使用不同尺度采样所得的特征图对网络解码端 DAFM 模块特征提取效果有不同程度的影响,本实验在 DAFM 模块和 MA-ASPP 模块的基

础上，对 DAFM 模块结合主干特征提取网络不同下采样率所得特征图的网络整体组合效果进行实验。实验结果见表 5。

表 5　不同尺度特征图对模型性能的影响

Tab. 5　Effect of different feature scales on model performance

基准模型	DAFM+MA-ASPP	MR(2,4)	MR(2,8)	MR(2,4,8)	*MIoU*/ %
DeepLabv3	×	√	×	×	72.1
	×	×	√	×	72.3
	×	×	×	√	72.8
	√	√	×	×	75.3
	√	×	√	×	75.5
	√	×	×	√	75.8

注：MR( )表示使用的多尺度倍率(Multiscal Rate)

从表 5 可看出，在 DeepLabv3 的基础上，对主干特征提取网络采取不同尺度的下采样，均可提升网络性能，其中使用 MR(2,4,8)对 DeepLabv3 的分割效果最优，达到 72.8%。在结合 DAFM 模块和 MA-ASPP 模块的基础上，当对主干网络使用 MR(2,4)下采样特征图进行组合时，*MIoU* 值为 75.3%；当使用 MR(2,8)下采样特征图进行组合时，*MIoU* 值为 75.5%；当使用 MR(2,4,8)下采样特征图输入 DAFM 模块时，网络的整体性能最佳，*MIoU* 值达到 75.8%。因此，本文还基于多尺度下采样倍率 MR(2,4,8)结合 DAFM 模块来改进原始网络，实现了更密集化的像素级采样，使得 DCF-DeepLab 网络具有更好的分割性能。

2.4　实验结果分析

DeepLabv3 与 DCF-DeepLab 在 PASCAL VOC2012 验证集上包括背景的 21 个类别的测试结果见表 6。从表中可看出，相比于基础语义分割网络 DeepLabv3 在 PASCAL VOC2012 验证集上的测试结果，基于 DAFM 模块的 DCF-DeepLab 在其中 18 个类别上的检测精度都有所提升，尤其在自行车、瓶子、沙发这 3 个类别上检测精度提高了 2% ~ 4%。总体上，*MIoU* 值由 69.7%提高到 70.6%，提升了 0.9%。

为了进一步验证本文提出的 DCF-DeepLab 网络的有效性，在 PASCAL VOC 2012 数据集上将 DCF-DeepLab 网络与其它经典语义分割网络的实验对比结果见表 7。

从表 7 可看出，DCF-DeepLab 在融合多尺度语义信息的基础上，*MIoU* 达到 75.8%，优于其他以 VGGNet 和 ResNet50 为主干特征提取网络的语义分割模型。同时，DCF-DeepLab 在模型参数量和分割时间上取得了较好的平衡，其模型参数量大小为 48.9 MB，每张分割时间为 0.123 s，明显优于对比网络。DCF-DeepLab 和 DeepLabv3 在 PASCAL VOC 2012 数据集上的部分可视化结果如图 9 所示。

表 6　不同类别检测性能对比

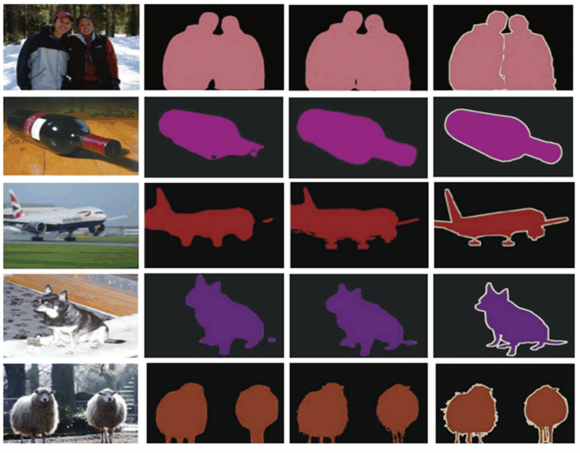
Tab. 6　Comparison of detection performance on different categories

类别	*MIoU* / %	
	DeepLabv3	DCF-DeepLab
Background	92.4	92.7
Areoplane	84.9	84.6
Bicycle	45.7	48.5
Bird	80.8	81.2
Boat	66.5	65.3
Bottle	64.2	67.8
Bus	86.5	86.8
Car	79.4	80.2
Cat	81.3	82.2
Chair	28.1	29.0
Cow	81.8	82.2
Diningtable	58.4	58
Dog	71.8	73.6
Horse	76.1	76.8
Motorbike	78.6	78.8
Person	79.8	80.2
Potted plan	53.6	55.3
Sheep	78.2	78.3
Sofa	35.8	38.7
train	82.4	82.8
Tv/monitor	57.5	59.2
MIoU / %	69.7	70.6

表 7　不同网络模型测试结果对比

Tab. 7　Test results of different network models

模型	骨干网络	*MIoU*/ %	参数量/ MB	每张分割时间/s
Fcn	VGGNet	62.2	125.5	0.153
Unet	VGGNet	69.8	153.6	0.199
U2net	ResNet50	71.2	162.5	0.174
Lraspp	ResNet50	72.2	52.6	0.134
DeepLabv3	ResNet50	74.3	198.6	0.186
DCF-DeepLab	MobileNetV3_Large	75.8	48.9	0.123



（a）原始图像 （b）DeepLabv3（c）DCF-DeepLab（d）标注图像

图 9 部分可视化结果图

Fig. 9 Visualizations of several prediction results

从图 9 可见，DCF-DeepLab 的分割性能整体上优于 DeepLabv3，尤其对精细的小目标进行分割时，其能够更好地捕捉小尺度细节语义特征，使目标边缘分割更加精细、光滑且完整。如：可视化结果中飞机的轮子和绵羊脚部位的轮廓分割等，较好地改善了 DeepLabv3 分割目标时存在的图像边界响应丢失及远距离小目标信息和目标边缘分割粗糙的问题。

3 结束语

本文针对 DeepLabv3 中存在的图像小目标信息易丢失等问题，提出了基于注意力机制的 DAFM 模块，融合浅层特征弥补深层特征的不足，实现特征图跨模块的融合。同时，为了减少训练时长，在主干特征提取网络部分引入轻量级网络 MobileNetV3_Large。针对目标边界分割粗糙等问题，通过引入注意力模块和串联结构改进 ASPP 模块，以增强局部特征提取的连贯性，进而从整体上提升语义分割性能。

消融实验表明，DAFM 模块和 MA-ASPP 模块对原始网络的语义分割性能有一定程度的提升；对比实验表明，DCF-DeepLab 网络在参数量大小和分割效率等方面都取得了一定成效，验证了本文设计网络的有效性。

在后续的研究中考虑将网络的小目标分割特性拓展到其他领域（如：苗族服饰图像分割、医学图像分割、建筑物裂缝等工程问题），以提高网络的泛化性。

参考文献

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

[2] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI Cham: Springer, 2015:234-241.

[3] FU JUN, LIU JING, JIANG JIE, et al. Scene segmentation with dual relation- aware attention network[J]. IEEE Trans on Neural Networks and Learning Systems, 2020, 32(6): 2547-2560.

[4] NIU RUIGANG, SUN XIAN, YUTIAN, et al. Hybrid multiple attention network for semantic segmentation in aerial images[J]. IEEE Trans on Geoscience and Remote Sensing, 2021, 60: 1-18.

[5] ZHU L, JI D, ZHU S, et al. Learning statistical texture for semantic segmentation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12537-12546.

[6] 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 Trans on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[7] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[8] WANG Zhimin, WANG Jiasheng, YANG Kun, et al. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+[J]. Computers and Geosciences, 2022, 158: 104969.

[9] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[10] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[11] ZHANG X, WANG T, QI J, et al. Progressive attention guided recurrent network for salient object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 714-722.

[12] ZHAO Z, CHEN K, YAMANE S. CBAM-Unet++: easier to find the target with the attention module" CBAM" [C]//2021 IEEE 10th Global Conference on Consumer Electronics (GCCE). IEEE, 2021: 655-657.

[13] CHEN B, ZHANG Z, LIU N, et al. Spatiotemporal convolutional neural network with convolutional block attention module for micro-expression recognition[J]. Information, 2020, 11(8): 380.

[14] SHI Hao, FAN Jiahe, WANG Yupei, et al. Dual attention feature fusion and adaptive context for accurate segmentation of very high-resolution remote sensing images[J]. Remote Sensing, 2021, 13(18): 3715.

[15] WANG Yanjie, WANG Guodong, PAN Zhenkuan, et al. Multi-scale dilated convolution of convolutional neural network for image denoising[J]. Multimedia Tools and Applications, 2019, 78(14): 19945-19960.

[16] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

（下转第 14 页）