计算机科学
COMPUTER SCIENCE

一种基于对偶学习的场景分割模型

刘思纯, 王小平, 裴喜龙, 罗航宇

引用本文

刘思纯, 王小平, 裴喜龙, 罗航宇. 一种基于对偶学习的场景分割模型[J]. 计算机科学, 2024, 51(8): 133-142.

LIU Sichun, WANG Xiaoping, PEI Xilong, LUO Hangyu. Scene Segmentation Model Based on Dual Learning [J]. Computer Science, 2024, 51(8): 133-142.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700207

一种基于对偶学习的场景分割模型

刘思纯　王小平　裴喜龙　罗航宇
同济大学电子与信息工程学院　上海 200092
(sichun_liu@tongji.edu.cn)

摘　要　城市场景分割等复杂任务存在特征图空间信息利用率低下、分割边界不够精准以及网络参数量过大的问题。为解决这些问题,提出了一种基于对偶学习的场景分割模型 DualSeg。首先,采用深度可分离卷积使模型参数量显著减少;其次,融合空洞金字塔池化与双重注意力机制模块获取准确的上下文信息;最后,利用对偶学习构建闭环反馈网络,通过对偶关系约束映射空间,同时训练"图像场景分割"和"对偶图像重建"两个任务,辅助场景分割模型的训练,帮助模型更好地感知类别边界、提高识别能力。实验结果表明,在自然场景分割数据集 PASCAL VOC 中,基于 Xception 骨架网络的 DualSeg 模型的 mIoU 和全局准确率分别达到 81.3%和 95.1%,在 CityScapes 数据集上 mIoU 达到 77.4%,并且模型参数量减少 18.45%,验证了模型的有效性。后续将探索更有效的注意力机制,进一步提高分割精度。

关键词:场景分割;图像重建;对偶学习;注意力机制;深度可分离卷积;多层次特征融合
中图分类号　TP391

Scene Segmentation Model Based on Dual Learning

LIU Sichun, WANG Xiaoping, PEI Xilong and LUO Hangyu
School of Electronics and Information Engineering, Tongji University, Shanghai 200092, China

Abstract　For complex tasks such as urban scene segmentation, there are problems such as low utilization of feature map space information, inaccurate segmentation boundaries, and excessive network parameters. To solve these problems, DualSeg, a scene segmentation model based on dual learning, is proposed. Firstly, depthwise separable convolution is used to significantly reduce the number of model parameters Secondly, accurate context information is obtained by fusing hollow pyramid pooling and double attention mechanism modules. Finally, dual learning is used to construct a closed-loop feedback network, and the mapping space is constrained by duality, while training the two tasks of "image scene segmentation" and "dual image reconstruction", it can assist the training of the scene segmentation model, help the model to better perceive the category boundary and improve the recognition ability. Experimental results show that the DualSeg model based on the Xception skeleton network achieves 81.3% mIoU and 95.1% global accuracy on natural scene segmentation dataset PASCAL VOC, respectively, and the mIoU reaches 77.4% on the CityScapes dataset, and the number of model parameters decreases by 18.45%, which verifies the effectiveness of the model. A more effective attention mechanism will be explored in the future to further improve the segmentation accuracy.

Keywords　Scene segmentation, Image reconstruction, Dual learning, Attention mechanism, Depthwise separable convolution, Multi-level feature fusion

1　引言

场景分割是计算机视觉领域一个基本而重要的问题,与一般的图像分割有着明显不同。它专注于对自然场景、城市场景等复杂场景下的图像进行语义分割。这些图像通常具有一定的透视形变,且包含的视觉要素数量较多,因此带来了一系列挑战,包括分割粒度细、尺度变化多样和空间相关性强等。

在自动驾驶技术领域,城市场景的分割对于自动驾驶系统的道路状况判断至关重要。这种场景分割不仅需要更准确地捕捉物体边界,还需要提高小物体分割的精度,以确保对道路、车辆和行人等细小要素的准确识别。然而,当前城市场景分割存在特征图空间信息的利用率较低、分割边界不够精准以及模型参数量过大等问题。

对偶学习是一种机器学习范式,旨在通过两个相互依赖的任务之间互相学习并提供反馈信息,进而提升模型性能。在传统的场景分割任务中,通常会单独训练模型,导致其他相关任务的信息被忽略,无法充分利用其他任务的信息来提升

到稿日期:2023-07-27　返修日期:2023-11-22
基金项目:国家重点研发计划(2022YFB4300504-4)
This work was supported by the National Key Research and Development Program of China(2022YFB4300504-4).
通信作者:王小平(xpwang6510@tongji.edu.cn)

性能。但引入具有对偶性的任务可以为场景分割模型提供额外的约束和指导，这一过程有助于模型更好地感知物体边界、实例以及深度信息，从而提高对这些信息的准确识别能力。因此，本文提出一种基于对偶学习的场景分割模型 DualSeg，构建“图像场景分割”和“对偶图像重建”对偶任务，并配合其他策略提升场景分割的性能。

本文的主要贡献如下：

1）根据“图像场景分割”和“对偶图像重建”任务之间的对偶关系，设计损失函数，构建对偶学习模型，使得两个任务互相提供反馈，互相促进，从而提升场景分割和图像重建的性能。

2）针对图像场景分割分支，引入 DA 双重注意力模块，融合多层次特征融合，并结合深度可分离卷积，既增强了模型获取全局上下文信息的能力，又使得模型较为轻量化。

3）在复杂场景下，场景分割模型能够更精准地分割物体边缘和模糊区域。无论在 CityScapes 城市场景分割数据集还是在 VOC 自然场景数据集上，本文算法都具有良好的性能。

2 相关工作

2.1 场景分割

场景分割的基本思路可以分为两种：基于深度卷积神经网络的分割和基于注意力机制的分割。这两种方法在场景分割任务中各具优势，并在不同应用场景下得到广泛应用。

FCN[1]，SegNet[2]等在传统语义分割任务上效果较好，但是对于复杂场景的分割仍然难以达到理想的效果。FCN 是全卷积网络的语义分割算法，它具有动态输入图像尺寸的优点，显著降低了计算复杂度，但是下采样操作导致特征的分辨率降低，目标的边界变得模糊，分割不够精确。文献[2]提出基于 FCN 的 SegNet 网络，编码部分使用 same 卷积提取特征，保持了卷积后图像的原始尺寸，同时在解码端也采用了 same 卷积，有助于重新学习池化过程中丢失的信息，从而生成高分辨率的预测图。SegNet 和 FCN 各自都有一定的局限性，例如对尺度变化和空间相关性较强的场景分割问题的处理仍然面临挑战。为了解决这些问题，谷歌提出了 DeepLab 系列的模型，该模型在复杂场景的语义分割问题上具有一定的代表性。DeepLabV1[3]能够对语义分割结果进行后处理，以获得更加平滑和准确的分割边界，但也存在图像分辨率降低和空间不变性的限制的问题。相较于 DeepLabV1，DeepLabV2[4]通过空洞卷积结构增大图像的感受野，使得模型能够更好地保留目标的细节信息，提高了分割结果的准确性。DeepLabV3[5]基于空洞卷积，引入空间金字塔池化，能够有效捕获多尺度信息，但是预测的特征图直接双线性上采样 16 倍，导致细节信息不够。DeepLabV3＋[6]基于 DeepLabV3 进行改进，通过多次双线性插值恢复丢失信息，并且将低维度特征与高维度特征相结合。

随着注意力机制在自然语言处理领域表现出优异性能之后，越来越多的研究人员尝试将其应用于计算机视觉领域，并取得了良好的效果。ViT(Vision Transformer)[7]是一种基于 Transformer 的图像识别模型，其将 Transformer 模型成功应用于计算机视觉领域。与传统的卷积神经网络不同，ViT 完全抛弃了卷积层，将图像分割任务转化为像素级分类任务。继 ViT 之后，Swin-Transformer[8]，SERT[9]，SegFormer[10]，Mask2Former[11]，OneFormer[12]等利用 Transformer 的强大表达能力和自注意力机制，有效地捕捉图像的多尺度信息和全局上下文，从而提升了分割的准确性和对细节的捕捉能力，但存在参数量过大、显存占用高和训练时间长等缺点，主要原因是场景分割是高密度预测任务，对分割精度要求很高，训练参数量巨大，增加了训练成本。

2.2 对偶学习

对偶学习利用任务的对称属性，通过引导和增强学习过程来获得更有效的反馈。许多人工智能处理任务中都存在对偶性，两个任务可以互相提供反馈信息，这些反馈信息可以用于训练模型。一方面，基于无监督或半监督的对偶学习可以减少对数据标记的依赖性。为了减少数据对标签的依赖性，Luo 等[13]使用半监督的方式，结合对偶学习，使用二元图像分割模型从图像中预测标签，并映射重建图像，从而减少对标签的依赖。另一方面，给定一个对偶任务模型，对偶模型也可以为原始模型提供有效反馈，增强学习到的特征，对于超分辨率重建、场景分割等对精度要求较高的任务有很好的促进作用。例如，文献[14]提出一种基于渐进上采样的对偶学习算法用于图像的超分辨率重建。该方法通过对偶学习策略构建闭环反馈网络，利用对偶关系相互约束映射空间，以获取最佳重建函数。此外，Li 等[15]提出一种双重超分辨率学习方法，将超分辨率恢复和语义分割两个任务结合起来，通过对偶学习提高了语义分割的精度和细节。目前将对偶学习直接应用于复杂的场景分割任务的研究相对有限，在计算机视觉领域，主要集中在图像生成、重建以及超分辨率重建等方面。

为了提升场景分割性能并充分利用对偶任务之间的反馈信息，本文提出了一种基于对偶学习的场景分割模型 DualSeg。在 DeepLabV3＋的基础上引入对偶学习范式，构建了“场景图像分割”和“图像重建”的对偶学习模型，使其相互协作，相互提供反馈信息以增强模型性能。此外，采用轻量级的双重注意力模块，并融合多层次特征，使得模型能够更好地捕捉全局上下文信息，在 CityScapes[16]和 PASCOL VOC 2012[17]上显著提升了场景分割性能，同时减少了模型的参数量。

3 基于对偶学习的场景分割模型——DualSeg

如图 1 所示，模型主要分为两个部分：场景分割分支和对偶图像重建分支，整个模型将“图像场景分割”和“对偶图像重建”任务进行联合建模。原始任务为图像场景分割，即对原始图像进行场景分割，得到分割掩码；对偶任务为图像重建任务，即从分割掩码还原为原始图像，与原始任务相比，体现出对偶性。模型训练时两个任务同时进行训练，对偶任务主要用于辅助场景分割任务，预测时丢弃对偶图像重建分支，只进行场景分割预测。整个模型的损失包括 dual loss 以及场景分割任务预测的 loss。

场景分割分支基于 DeepLabV3＋进行改进。为使模型轻量化，主干网络主要采用 Xception[18]。将主干网络提取的特征送入 DoubleAttention 模块，包括通道注意力模块和空间注意力模块，用于增强每个通道的特征表示，并强调图像中不同位置的重要性，以更好地提取全局上下文信息，从而提高

模型的感知能力，并且不会增加网络复杂性。DoubleAttention 模块获取的全局上下文信息经过 ASPP 模块，增大了图像的感受野，ASPP 模块将得到的特征上采样 4 倍后与主干网络中的低阶特征进行融合，融合后将特征图上采样 4 倍到原图大小，最后对每个像素进行分类，得到图像分割掩码。此外，在特征融合的过程中，需要使用 1×1 的卷积修改通道数，以便进行特征拼接。

对偶图像重建分支将主干网络提取的低阶特征经过卷积转换到同一个通道后，进行多层次的特征融合，使用 4 倍上采样将其还原为原始图像大小，与场景分割的结果进行拼接，拼接后再进行卷积上采样等操作，即可得到重建后的图像，如图 1 中 Recon_Image 所示。

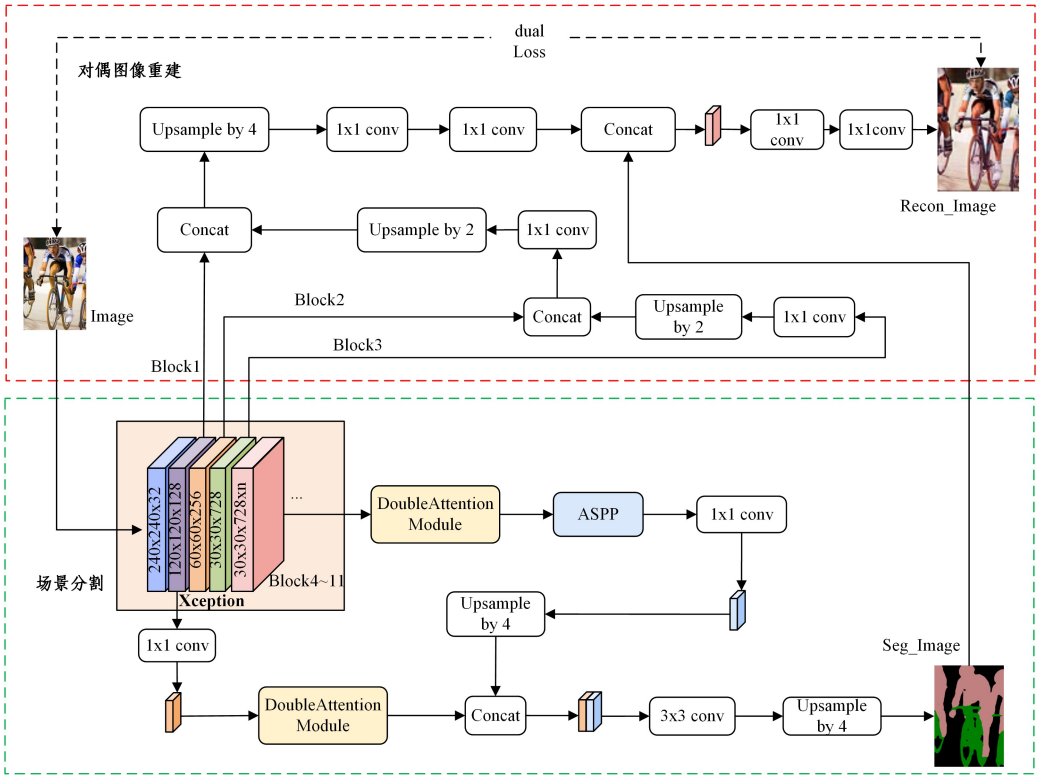
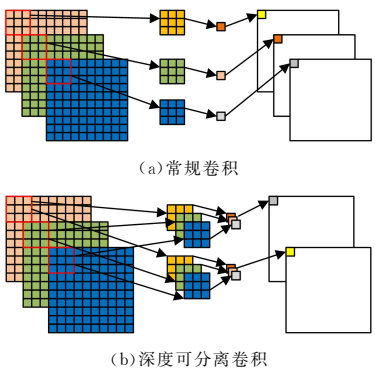


图 1 DualSeg 网络架构

Fig. 1 DualSeg network architecture

3.1 特征提取网络

实验阶段，本文构建多个主干网络，包括 Xception，HRNet 和 ResNet。Xception 引入深度可分离卷积，对不同输入通道采用不同的卷积核进行卷积，将卷积过程分解为深度卷积和逐点卷积，如图 2(b)所示。相比常规的卷积操作，其参数数量和运算成本显著降低。



(a)常规卷积

(b)深度可分离卷积

图 2 常规卷积与深度可分离卷积[13]

Fig. 2 Conventional conv and depthwise separable conv[13]

图 1 给出了以 Xception 为 Backbone 的网络结构，Xception Backbone 主要由 low-level block 和 middle-level-block 组成，其基本组成结构如图 3 所示。其中 low-level-block 由两个 *stride*=1 的深度可分离卷积和一个 *stride*=2 的深度可分离卷积堆叠，并且通过跨层连接结构与步距为 2、卷积核为 1×1 的普通卷积拼接。middle-level-block 中不存在 stride=2 的深度可分离卷积，也不包含残差结构。图 1 中的 block1，block2 和 block3 分别对应 channel 为 128，256 和 728 时的 low-level-block，而 block4－blcok11 则是 middle-level-block，最后再接入 channel 分别为 1 024，1 536 和 2 048 的深度可分离卷积，以进一步提取特征。采用这样的网络结构，能够在深度可分离卷积和跨层连接的基础上，充分利用不同层级的特征提取能力。

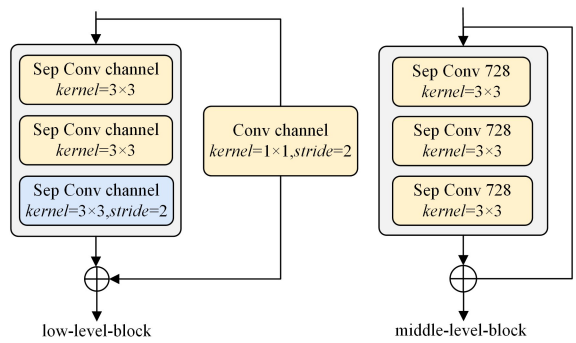


图 3 Xception Backbone 的基本组成结构

Fig. 3 Basic structure of Xception Backbone

3.2 双重注意力模块

DualSeg 模型中引入了轻量化的双重注意力模块，它可以串行连接通道注意模块和空间注意模块来改进 DeepLabV3+，使得模型能够捕捉全局上下文信息。双重注意力模块的结构如图 4 所示。

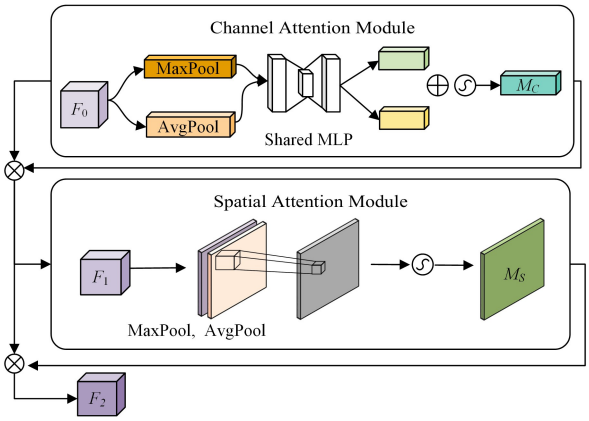


图 4 双重注意力模块

Fig. 4 Double attention module

通道注意力模块的主要作用是获取不同通道特征的注意力权重，可以压缩特征图的空间维度，放大不同目标之间的特征差异。空间注意力模块的主要作用是获取同一个通道上不同位置特征的注意力权重，可以整合多尺度空间信息，扩展复杂场景中不同尺寸物体的差异。双重注意力模块是串行结构，可以沿两个独立的维度依次推断出注意力权重，和原特征图相乘来对特征进行自适应特征优化。与并行结构相比，串行结构不会导致通道和空间维度之间的信息交互不足，可以更好地平衡这两个维度之间的关系。整个注意过程可以概括为式(1)和式(2)：

$$F_1 = M_c(F_0) \otimes F_0 \tag{1}$$

$$F_2 = M_s(F_1) \otimes F_1 \tag{2}$$

其中，$\otimes$表示逐元素乘法，M_c 为通道注意图，M_s 为空间注意图，F_0 为输入特征图，F_1 为中间特征图。通道注意力模块在空间维度上利用最大池化输出和平均池化输出，通过共享网络对特征映射进行压缩，从而对 F_0 的空间信息进行聚合。将每个特征图中对应的像素逐像素相加，得到通道注意图 M_c。将 M_c 和 F_0 逐元素相乘后，得到 F_1。通道注意力模块表示为：

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{avg}))) + W_1(W_0(F^c_{max})) \end{aligned} \tag{3}$$

空间注意模块以 F_1 为输入特征，经过两次池化操作，得到平均池化特征和最大池化特征。通过卷积层进行拼接，产生二维的 M_s。将 M_s 和 F_1 逐元相乘后，得到双重注意力模块的输出 F_2。空间注意力模块可以表示为：

$$\begin{aligned} M_s(F) &= \sigma(f^{(7\times7)}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{(7\times7)}([F^s_{avg}; F^s_{max}])) \end{aligned} \tag{4}$$

其中，σ 表示 sigmoid 函数，$f^{7\times7}$ 表示卷积核的大小。复杂的场景分割任务关注类别信息和边界信息，而通道注意模块和空间注意模块可以分别学习在通道和空间上需要注意的类别信息和边界信息。

3.3 ASPP 和多层次特征融合

场景分割模型的 ASPP 模块和多层次特征融合模块如图 5 所示。经过 Backbone 得到的特征图，经过双重注意力模块后，将通道数调整为 1 280，送入 ASPP 模块，分别使用膨胀率不同的多个并行空洞卷积层。对于每个模块，通道数均为 128，融合生成最终结果。将 ASPP 模块的输出经过 1×1 的卷积层，减少模型参数，进行 4 倍上采样后与经过双重注意力模块的 Block1 浅层特征拼接，最后再进行卷积和上采样，还原回原图大小。

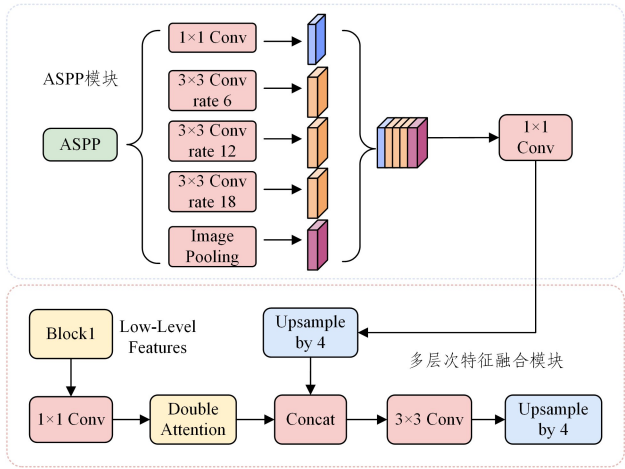


图 5 ASPP 和多层次特征融合

Fig. 5 ASPP and multi-level feature fusion

ASPP 模块中，不同扩张率的膨胀卷积能在不使用池化层且计算量相当的情况下，提供更大的感受野，获取多尺度物体信息。而不同层神经元提取和处理的特征信息强度也不一样，输出的特征语义信息具有不同强度和分辨率。城市场景分割和自然图像分割数据集上，不同层提取的特征差异如图 6 所示。

图 6 不同层次的特征图

Fig. 6 Feature maps of different levels

3.4 对偶回归网络

3.4.1 网络结构

本文将对偶学习引入图像分割任务中辅助模型的训练，对偶学习同时考虑两个任务 $f: x \to y$ 和 $g: y \to x$。原始任务 f 表示图像分割任务，对偶任务 g 表示从图像分割到原始图像的重建，具体结构如图 7 所示。场景分割任务 f 为对偶图像重建任务 g 提供反馈；同样地，对偶图像重建任务 g 也可以为场景分割任务提供反馈。

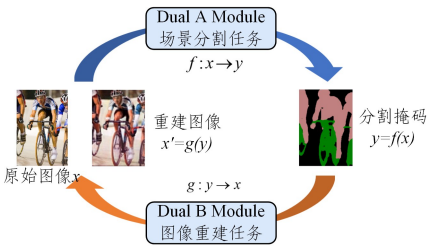


图 7 对偶学习网络

Fig. 7 Dual learning network

两个任务之间的关系可以描述为：假设原始任务从空间 χ 中抽取一个样本 x 输入空间 γ 中，对偶任务从空间 γ 中抽取一个样本 y 输入空间 χ 中，则原有任务学习参数 θ_{xy} 的条件概率为 $P(y|x;\theta_{xy})$，其对偶任务学习参数 θ_{yx} 的条件概率为 $P(x|y;\theta_{yx})$，其中 $x\in\chi$，$y\in\gamma$。

场景分割任务和对偶图像重建任务相互关联，以对称的结构形式相互约束。输入的图像 x 通过场景分割模型生成 y，得到分割预测结果。y 再由对偶图像重建模型生成 x'，通过比较原始图像 x 与重建图像 x' 之间的真实性来判断模型的优劣。原始图像与重建图像越接近，则原始任务 f 和对偶任务 g 表现越好；若相差较大，则两个模型相互反馈，优化模型的训练，进而提高模型的学习能力。

3.4.2 对偶网络训练方法

对偶网络的训练通过增加正则项来进行约束，场景分割任务和对偶图像重建任务互相关联，通过联合分布约束模型，学习两个模型的参数 θ_{xy} 和 θ_{yx} 并使其损失不断降低。通过原始任务与对偶任务之间的概率关系，促进网络的训练，使其正向反馈。场景分割任务与对偶重建任务的作用域内，$\forall(x,y)$ 应满足式(5)：

$$P(x)P(y|x;\theta_{xy})=P(y)P(x|y;\theta_{yx}) \tag{5}$$

基于条件概率对模型，f 和 g 可以定义为：

$$f(x;\theta_{xy})\overset{\triangle}{=}\arg\max_{y'\in\gamma}P(y'|x;\theta_{xy}) \tag{6}$$

$$g(y;\theta_{yx})\overset{\triangle}{=}\arg\max_{x'\in\chi}P(x'|y;\theta_{yx}) \tag{7}$$

令 $\ell_1(f(x),y)$ 和 $\ell_2(g(y),x)$ 分别为模型 f 和 g 的损失函数，网络训练过程中应满足式(8)和式(9)：

$$\min_{\theta_{xy}}\frac{1}{n}\sum_{i=1}^{n}\ell_1(f(x_i;\theta_{xy}),y_i) \tag{8}$$

$$\min_{\theta_{yx}}\frac{1}{n}\sum_{i=1}^{n}\ell_1(g(y_i;\theta_{yx}),x_i) \tag{9}$$

在训练过程中，场景分割模型和对偶重建模型二者利用概率对偶关系相互约束，通过最小化参数 θ_{xy} 和 θ_{yx} 对原始图像到图像分割空间的映射函数进行约束，得到最优的映射关系。

3.5 损失函数设计

图像分割任务中常用的损失函数包括交叉熵损失、Dice 损失等。交叉熵损失在图像分割任务中常用于像素级别的分类。多分类任务中每个类别相互独立，而单类别中只存在属于和不属于两种情况，因此对于每个类别的交叉熵的计算式为：

$$CrossLoss=-(y\log(\tilde{y})+(1-y)\log(1-\tilde{y})) \tag{10}$$

本文对 DeeplabV3+算法的损失函数进行了改进，融合了对偶学习的损失，以便网络能够捕捉待分割目标的细节信息。对偶学习分支的损失即为图像重建的损失，使用 MSE 均方误差损失。本文提出的模型的损失包含两部分，分别是图像分割掩码的损失以及对偶图像重建分支的损失。最终损失函数的表达式为：

$$Loss=\sum_{j=1}^{m}\sum_{i=1}^{n}(y\log(\tilde{y})+(1-y)\log(1-\tilde{y}))+\lambda\cdot\frac{1}{m}\sum_{j=1}^{m}(y_i-\bar{y}_i)^2 \tag{11}$$

其中，m 表示图片的数量，n 表示类别数；y 代表标签值，$\tilde{y}$ 代表图像分割模型的预测值，$\bar{y}$ 代表图像重建的结果；λ 是常数，表示对偶学习损失所占的权重。损失函数的第一部分为图像分割对于一个 batchsize 内多张图片中所有类别的交叉熵损失；第二部分为对偶学习进行图像重建分支的损失。

3.6 模型训练与预测

3.6.1 模型训练

模型训练时采用联合训练的策略，同时训练原始场景分割任务和对偶图像重建任务。通过 3.4.2 节中的对偶学习策略，原始任务与对偶任务互相促进。

在实际的任务中，真实的边缘分布 $P(x)$ 和 $P(y)$ 通常是不可获得的。作为替代，通常是利用在给定数据集上拟合出的两个经验的边缘分布 $\hat{P}(x)$ 和 $\hat{P}(y)$ 来完成式(5)、式(8)和式(9)中的约束。为了解决对偶有监督学习的问题，采用拉格朗日乘子法将有约束优化问题转化为无约束问题。式(1)中的约束会转变为一个正则项，并且真实边缘分布被替换成为两个经验边缘分布。

$$\ell_{duality}=(\log\hat{P}(x)+\log P(y|x;\theta_{xy})-\log\hat{P}(y)-\log P(x|y;\theta_{yx}))^2 \tag{12}$$

有了正则项后，即可进行对偶有监督学习算法的训练。算法伪代码描述如算法 1 所示。

算法 1 对偶学习训练过程

输入：训练集 D

参数：学习率 poly，最大迭代次数 K，拉格朗日乘子[19] λ_{xy} 和 λ_{yx}，边缘分布 $\hat{P}(x)$ 和 $\hat{P}(y)$

输出：原始网络和对偶网络的参数 θ_{xy} 和 θ_{yx}

1. for k∈{1,…,K} do

2. 随机采样出 m 个数据对 $\{(x^{(j)},y^{(j)})\}_{j=1}^{m}$；

3. 按照如下方式计算梯度：

$$G_f=\nabla_{\theta_{xy}}\frac{1}{m}\sum_{j=1}^{m}[\ell_1(f(x^{(j)};\theta_{xy}),y^{(j)})+\lambda_{xy}\ell_{duality}(x^{(j)},y^{(j)};\theta_{xy},\theta_{yx})] \tag{13}$$

$$G_g=\nabla_{\theta_{yx}}\frac{1}{m}\sum_{j=1}^{m}[\ell_2(g(y^{(j)};\theta_{yx}),x^{(j)})+\lambda_{yx}\ell_{duality}(x^{(j)},y^{(j)};\theta_{xy},\theta_{yx})] \tag{14}$$

4. 分别更新模型的参数 θ_{xy} 和 θ_{yx}：$G_f\to\theta_{xy}$ 和 $G_g\to\theta_{yx}$

5. end

6. return θ_{xy} 和 θ_{yx}

联合训练前，分布独自训练两个任务至收敛，使用该结果初始化对偶模型，以节省训练时间。在对偶有监督学习的训练过程中，式(13)和式(14)中的 λ_{xy} 和 λ_{yx} 固定为 0.01。

联合训练过程中，使用在 ImageNet 预训练的 Xception 和 ResNet101 骨架网络提取特征，HRNetV2 骨架网络提取的特征部分从零开始训练，优化器采用 Momentum，训练时

采用"poly"学习率策略，并将初始学习率设置为 0.007。使用 Xception 和 ResNet101 作为骨架训练网络时，图像的尺寸为 513×513；使用 HRNetV2 作为骨架网络训练时，需要先将其裁剪为 512×512。

3.6.2 模型预测

在验证对偶学习的有效性时，分别在场景分割任务和图像重建任务中进行验证。针对复杂场景数据集 CityScapes 和 VOC 数据集，进行模型的预测，检验性能是否提升。针对场景分割应用，可以仅保留场景分割分支，丢弃对偶重建分支，以减少模型推理的计算量。

4 实验与结果分析

4.1 实验设置

4.1.1 实验数据集

实验主要面向复杂场景分割，因此采用 CityScapes 数据集和 PASCAL VOC 2012 数据集。

1)Cityscapes 数据集是一个高质量的专注于城市街道场景理解的数据集，每张图像中都包含大量的动态物体、复杂变化的街景布局，涵盖不同时间段。在实际训练过程中，将精细语义标注数据集划分为训练集(Training)、验证集(Valuation)和测试集(Testing)，其中用于训练的图片样本 2975 张，用于验证的图片样本 500 张，用于测试的图片 1525 张。在实验过程中，仅对其中的 19 个城市街道场景常见的类别进行训练和评估验证。

2)PASCAL VOC 2012 数据集包含 1464 张训练图片、1449 张验证图片和 1456 张测试图片，包含 20 个对象类别和 1 个背景类别。由于数据量较少，实验中采用旋转、裁剪和翻转等手段进行数据增强，数据增强后的训练集为 10582 张，验证集为 1449 张，测试集为 1456 张。

4.1.2 实验环境

表 1 介绍了本文的实验环境。

表 1 实验环境

Table 1 Experimental environment

实验环境	环境配置
操作系统	Ubuntu 18.04 LTS
处理器	Intel Xeon(R)CPU@2.20GHz×24
内存	11GB
显卡	NVIDIA GTX 1080Ti ×2
深度学习框架	Pytorch 1.10.0

在训练过程中使用双 GPU 进行训练，因此可以将 batchsize 设置得较大。针对 Xception 和 ResNet-101 骨架网络，batchsize 设置为 16，预估训练 100 个 epoch 左右，若网络提前收敛，则终止训练；针对 HRNet 骨架网络，其训练参数量较大，batchsize 设置为 8，训练的轮次也设计得更多，预估 300 个 epoch 模型可以收敛，若没有收敛，则继续训练。

4.2 实验评价指标

1)图像分割评价指标

语义分割算法的主要目的是解决像素点分类的问题，因此，需要判断每一个像素点是否分类正确。常用的评价指标有 4 个：PA(也称为 Global Acc)、mPA、IoU 以及 mIoU。像素准确率 PA 表示预测类别正确的像素数占总像素数的比例。类别平均像素准确率 mPA 表示分别计算每个类被正确分类像素数的比例，累加求平均。平均交并比 mIoU 即预测区域和实际区域交集除以预测区域和实际区域的并集，这样计算得到的是单个类别下的 IoU，重复此算法计算其他类别的 IoU，最后计算其平均数。各评价指标的计算式如下：

$$Global\ Acc=\sum_i \frac{n_{ii}}{t_i} \tag{15}$$

$$mean\ Accuracy=\frac{1}{n_{\text{cls}}}\cdot\sum_i \frac{n_{ii}}{t_i} \tag{16}$$

$$mIoU=\frac{1}{n_{\text{cls}}}\cdot\sum_i \frac{n_{ii}}{t_i+\sum_j n_{ji}-n_{ii}} \tag{17}$$

其中，n_{ij} 表示类别 i 被预测成类别 j 的像素个数；$t_i=\sum_i n_{ii}$ 表示目标类别 i 的总像素个数，即真实标签；n_{cls} 表示目标类别个数，包含背景。

2)图像重建评价指标

为评价网络重建效果的有效性，本文选用两种常用的客观评价指标，即峰值信噪比(PSNR)和结构相似度(SSIM)，来定量分析图像的重建质量。两幅图像之间的 PSNR 越高，表示重建后的图像相对于原始图像的失真越小，图像的重建质量越高。SSIM 指两幅图像之间结构信息的相似程度，SSIM 值越接近 1，表明重建图像与原始图像越接近，重建效果越好。各评价指标的计算式如下：

$$MSE=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}(X(i,j)-Y(i,j))^2 \tag{18}$$

$$PSNR=10\log_{10}\left(\frac{(2n-1)^2}{MSE}\right) \tag{19}$$

$$SSIM=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)+(\sigma_x^2+\sigma_y^2+C_1)} \tag{20}$$

其中，W 和 H 分别表示重建图像 Y 的宽和高；X 表示原始图像；μ_x 和 σ_x 分别表示原始图像的灰度平均值和方差；μ_y 和 σ_y 分别表示重建的高分辨率图像的灰度平均值和方差；σ_{xy} 为原始图像与重建图像的协方差；C_1 和 C_2 为常数。

4.3 实验结果

4.3.1 对偶损失权重

本文方法损失函数包括场景分割部分的交叉熵损失和对偶图像重建部分的 MSE 损失，λ 作为超参数可以调节，表示对偶学习损失所占的权重。在 VOC 数据集上，DualSeg 以 ResNet-50，ResNet-101，Xception 和 HRNetV2 作为 Backbone 分别训练。λ 取不同值时，在测试集上模型性能 mIoU 随 λ 的变化情况如图 8 所示。

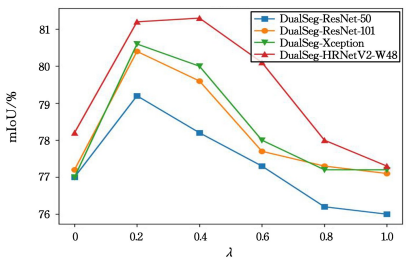


图 8 mIoU 随 λ 的取值变化

Fig. 8 mIoU changes with the value of λ

由图 8 可知，除了 HRNetV2-W48 作为 Backbone，$\lambda=0.4$ 时，模型的 mIoU 最大，其余情况下，当 $\lambda=0.2$ 时，模型的训练效果最好。当 λ 过大时，模型的整体损失值变大，模型收敛

更慢且稳定性不如 $\lambda=0.2$ 时。

4.3.2 对比实验

为衡量本文模型的各方面性能与效果，本文复现了目前主流的语义分割模型：DeepLabV3，DeepLabV3＋，HRNetV2[20] 和 PSPNet[21] 等，使用 ResNet101[22]，Xception 和 HRNetV2-W48 作为特征提取网络，Xception 在保证性能的同时能够降低计算量。以下涉及本文方法的实验均在 $\lambda=0.2$ 时完成。

1)定性分析

以 Xception 为主干网络，随机选取 PASCAL VOC2012 测试集中的一些图像进行模型预测，本文方法 DualSeg 的预测结果如图 9 所示。图 9 中，第一列是输入图像，第二列是输入图像对应的 ground truth，第三列是 DeepLabV3 的分割预测结果，第四列是 DeepLabV3＋分割预测结果，第五列是本文方法的分割预测结果，第六列是图像重建结果。可以看出：(1)本文算法总体上比 DeepLabV3 和 DeepLabV3＋网络的预测结果要好，同时对细节的预测更为精准，对细节的信息捕捉较好；(2)重建的图像保留了原始图像的主要特征和细节，并且具有良好的清晰度和准确性，能够辅助场景分割网络。

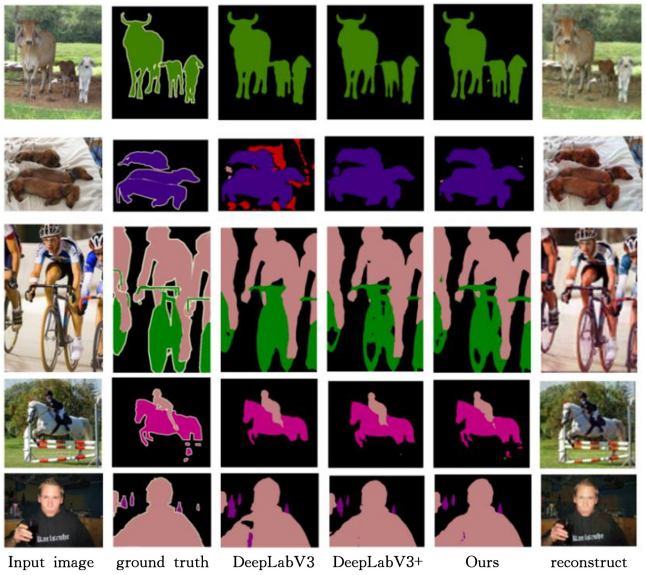


图 9 模型预测结果

Fig. 9 Model prediction results

采用 Xception 上训练的权重进行 CityScapes 数据集的预测，结果如图 10 所示。本文算法比 DeepLabV3＋的预测结果更好，能够较好地捕捉边界信息。图中黄色方框处为边界信息的分割结果，可见引入注意力机制后，本文模型对边界信息的分割较为清晰。

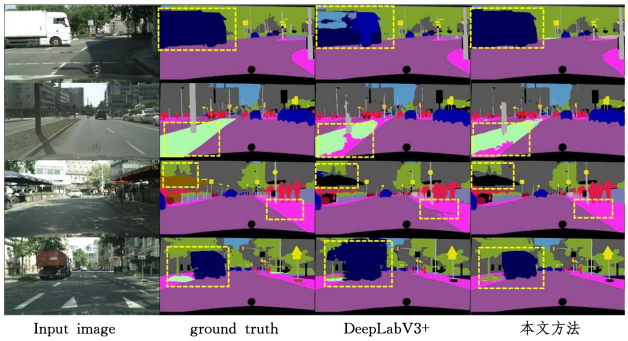


图 10 CityScapes 数据集上的预测结果

Fig. 10 Prediction results on CityScapes dataset

2)定量分析

(1)场景分割性能

将本文的算法与 DeepLabV3，DeepLabV3＋和 HRNetV2 等经典的分割算法进行实验对比，在 VOC 数据集上的实验结果如表 2 所列。对于 DeepLabV3，采用 ASPP 模块感受野较大，其 mIoU 可达到 79.3%；对于 DeepLabV3＋，基于 Xception 作为 Backbone，性能比 DeepLabV3 提升了 0.8%。HRNetV2 算法中，其 backbone 保持不同分辨率的平行分支，参数量比 DeepLabV3 和 DeepLabV3＋更大，达到 65.9×10^6，但其 mIoU 达到 80.4%。本文还将 DualSeg 的实验结果与目前主流的注意力分割模型进行了对比，SERT 和 SegFormer 的参数量比 DualSeg 大，并且性能略低于 DualSeg；而 MaskFormer 和 OneFormer 是通用图像分割任务，其性能比 DualSeg 模型略高，但是参数量远远大于 DualSeg 模型，导致训练时间较长。

表 2 VOC 数据集上不同算法的分割性能

Table 2 Segmentation performance of different algorithms on VOC dataset

方法	Backbone	*mIoU*/%	*Piexl Acc*/%	参数量
DeepLabV3	ResNet-101	79.3	94.2	58.0×10^6
DeepLabV3＋	Xception	80.1	94.3	43.5×10^6
HRNetV2	HRNetV2-W48	80.4	94.5	65.9×10^6
PSPNet	ResNet-101	79.8	94.6	63.1×10^6
SERT-PUP	T-Base	80.3	95.0	97.6×10^6
SegFormer	MiT-B5	81.2	95.2	84.7×10^6
Mask2Former	Swin-L	81.6	95.2	216.1×10^6
OneFormer	Swin-L	81.7	95.4	219.2×10^6
本文方法 (DualSeg)	Xception	80.6	95.0	47.3×10^6
	Xception＋DA	81.3	95.1	48.4×10^6
	ResNet-101	80.4	94.9	62.7×10^6
	HRNetV2-W48	81.2	95.1	68.1×10^6

本文算法引入对偶学习模块，并结合多尺度特征提取，在

不同的 Backbone 上进行训练，在 VOC 数据集上进行验证，性能均有提升。基于 Xception 的 DualSeg 模型的参数量最小，相比 DeepLabV3，其参数量减少 18.45%，且 mIoU 达到 80.6%；本文引入双重注意力模块 DA 后，mIoU 达到81.3%，且参数量仅增加 1.1×10^6，模型仍然较为轻量化。基于 HRNetV2-W48 的 DualSeg 模型的参数量最大，模型性能也接近最优，mIoU 达到了 81.2%。

为验证本文提出的 DualSeg 模型在复杂场景下的性能，在 CityScapes 城市场景分割数据集上进行了性能测试。实验结果如表 3 所列。在 CityScapes 数据集上，本文模型的性能比在自然场景数据集上差，但是与经典的算法相比，性能仍有 0.5%左右的提升，在 Xception 作为 backbone 时，模型的 mIoU 达到 76%；引入 DA 双重注意力模块后，模型的 mIoU 达到 77.4%，性能超出 SERT 1.2%，超出 SegFormer 0.5%，并且仅比 Mask2Former 和 OneFormer 低 1%左右，可见模型在 CityScapes 等复杂场景数据集上，场景分割性能依然良好。

表 3 Cityscapes 数据集上的不同算法分割性能

Table 3 Segmentation performance of different algorithms on Cityscapes dataset

方法	Backbone	*mIoU*/%
DeepLabV3	ResNet-101	74.7
DeepLabV3+	Xception	75.4
HRNet	HRNetV2-W48	76.2
PSPNet	ResNet-101	74.6
SERT-PUP	T-Base	76.2
SegFormer	MiT-B5	76.9
Mask2Former	Swin-L	78.3
OneFormer	Swin-L	78.6
本文方法 (DualSeg)	Xception	76.0
	Xception+DA	**77.4**
	ResNet-101	75.8
	HRNetV2-W48	76.9

(2)图像重建性能

峰值信噪比(PSNR)和结构相似度(SSIM)的量化结果分别如表 4 和表 5 所列，两个表的数据为在一批验证数据集上求平均的结果。由表 4 和表 5 可知，模型的 PSNR 值较大，图像重建的效果较好；SSIM 值越接近 1 越优，其值大于 0.6，基本重建出了原始图像的信息。由于 CityScapes 数据集的场景更为复杂，因此其重建效果相比 VOC 数据集略差一些。

表 4 模型在 VOC 数据集上的图像重建性能

Table 4 Image reconstruction performance on VOC dataset

Backbone	*PSNR*	*SSIM*
ResNet50	18.83	0.63
ResNet101	20.74	0.67
Xception	21.32	0.67
HRNetV2-W48	23.54	0.71

表 5 模型在 CityScapes 数据集上的图像重建性能

Table 5 Image reconstruction performance on CityScapes

Backbone	*PSNR*	*SSIM*
ResNet50	14.78	0.58
ResNet101	17.24	0.62
Xception	19.13	0.63
HRNetV2-W48	20.01	0.64

4.3.3 消融实验

1)对偶重建和多层次特征融合的有效性验证

为验证模型各模块的有效性，在 ResNet-101 和 Xception 上分别进行了消融实验，验证对偶重建分支以及多层次特征融合对场景分割的影响。消融实验的结果如表 6 所列。

表 6 模型各模块对性能的影响

Table 6 Impact of each module on performance

Method	Backbone	ASPP	Recon	Multi	Sep conv	mIoU/%
DeepLabV3	ResNet-101	√			−	79.3
	ResNet-101	√	√		−	79.7
	ResNet-101	√	√	√	−	79.8
本文方法 (DualSeg)	Xception	√			√	80.1
	Xception	√	√		√	80.3
	Xception	√	√	√	√	**80.6**
本文方法 (DualSeg)	ResNet-101	√	√			79.0
	ResNet-101	√	√	√		79.2
	ResNet-101	√	√		√	80.3
	ResNet-101	√	√	√	√	80.4

在 ResNet101-Backbone 上，对 DeepLabV3 算法进行复现，该模型引入 ASPP 模块以增大感受野，但是没有 Recon 对偶重建分支以及 Multi 多层次特征提取模块，其 mIoU 值最低，仅为 79.3%。对其引入本文的 Recon 模块，mIoU 达到 79.7%；引入 Multi 多层次特征提取模块后，mIoU 提升了 0.1%。可见，在其他模型上应用本文提出的改进模块，能够实现性能上的提升。

本文模型在 DeepLabv3+的基础上进行改进，引入了对偶重建分支来辅助图像分割任务，在 ResNet-101 上其 mIoU 为 79.0%。但是该算法的对偶重建分支进行图像重建的能力较差，因此对该部分进行改进，引入 Multi 多层次特征融合模块，在图像重建的过程中融合底层特征信息，提升模型重建的性能，并辅助图像分割任务；引入 Multi 模块的对偶图像分割网络的 mIoU 为 79.2%，性能有了进一步的提升，但是 Multi 模块的提升较少。同理，在 Xception Backbone 上进行类似的消融实验，引入 ASPP 模块、Recon 模块和 Multi 模块后，模型的 mIoU 达到 80.6%，而未引入 Recon 模块、Multi 模块时，mIoU 为 80.1%，验证了本文模型的有效性。

2)对偶学习的有效性验证

为进一步验证对偶学习的有效性，即验证场景分割任务和对偶图像重建任务的相互促进作用，设置两组实验，分别在 VOC 数据集和 CityScapes 数据集上进行验证。控制变量为是否采用对偶学习策略，若采用，即使用本文的 DualSeg 模型(不引入双重注意力机制模块)，反之只进行场景分割任务或者只进行图像重建任务。模型的验证结果如表 7 和表 8 所列。

表 7 VOC 数据集上对偶学习的有效性

Table 7 Effectiveness of dual learning on VOC dataset

Backbone/对比结果	*PSNR*	*SSIM*	*mIoU*/%	对偶学习
Xception	17.22	0.61	80.1	×
	18.83	0.63	80.6	√
Δ	1.61	0.02	0.5	↑

表 8 CityScapes 数据集上对偶学习的有效性

Table 8 Effectiveness of dual learning on CityScapes dataset

Backbone/对比结果	*PSNR*	*SSIM*	*mIoU*/%	对偶学习
Xception	18.19	0.58	75.4	×
	19.13	0.63	76.0	√
Δ	0.94	0.05	0.6	↑

模型的峰值信噪比 PSNR 值越大，说明图像重建的效果越好；而结构相似度 SSIM 越接近 1 说明模型性能越好。本文对以 Xception 作为 Backbone 的模型进行验证，可以看出，在 VOC 数据集上，没有使用对偶学习策略时，无论是场景分割的性能 mIoU，还是图像重建的性能 PSNR 和 SSIM，均不如使用对偶学习策略时的性能优，可见对偶学习策略能够使两个模型互相促进。

类似地，在 CityScapes 数据集上进行实验验证，使用对偶学习策略的场景分割和图像重建性能均优于不使用对偶学习策略的性能，验证了模型之间能够互相促进。

3）上采样倍率的有效性验证

基于表 6 中的实验，对 DualSeg 算法的上采样倍率进行验证，实验结果如表 9 所列。当 DualSeg 算法具有 ASPP 模块、对偶图像重建模块和多层次特征融合模块时，控制上采样的倍率不同，进行验证。当骨干网络采用 Xception 时，连续使用两次 2 倍上采样，模型性能比直接使用 1 个 4 倍上采样性能差 0.3%；同样，骨干网络采用 ResNet101 时进行类似验证，使用 4 倍上采样时，mIoU 值比 2 倍上采样高 0.1%。

表 9 不同上采样倍率对 mIoU 的影响

Table 9 Effect of different upsampling ratios on mIoU

Backbone	ASPP	Recon	Multi	UpSample	*mIoU*/%
Xception	√	√	√	2 * 2	80.3
	√	√	√	4 * 1	80.6
ResNet-101	√	√	√	2 * 2	80.3
	√	√	√	4 * 1	80.4

4）双重注意力模块的有效性验证

DualSeg 基于对偶学习方法，辅助场景分割任务的训练，提升场景分割性能。在上述模型的基础上，引入 DoubleAttention 模块，可以帮助模型更好地提取上下文信息，使场景分割任务能够更好地关注感兴趣区域，有利于边界的分割。

使用 Xception 轻量化主干网络，在场景复杂的 CityScapes 数据集上进行训练，控制 DoubleAttention（DA）模块的有无，训练过程中 mIoU 的变化如图 11 所示。模型最终在验证集上的性能如表 10 所列。

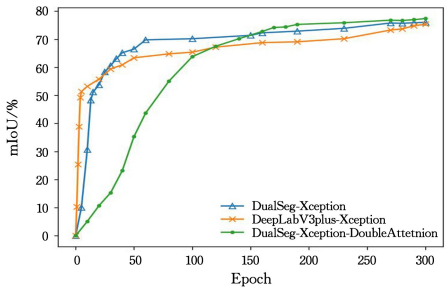


图 11 各模型训练过程中 mIoU 的变化

Fig. 11 Changes in mIoU during training of each model

表 10 双重注意力模块对性能的影响

Table 10 Effect of DA module on performance

Method	Recon	Multi	Sep conv	DA	*mIoU*/%
DeepLabV3+		√	—		75.0
	√	√	—		75.4
	√	√	—	√	75.9
本文方法（DualSeg）	√	√	√		76.0
	√	√	√	√	77.4

由图 11 和表 10 可知，与未引入双重注意力模块的模型相比，引入双重注意力模块的模型收敛速度变慢，但是最终的 mIoU 更高。

为验证模型的通用性和可视化分析模型的感兴趣区域，在 PASCAL VOC 2012 数据集上，对 DeepLabV3+，DualSeg（Without DoubleAttention）和 DualSeg（with DoubleAttention）进行训练。根据其训练权重，使用 grad-CAM[23] 工具将模型关注的区域可视化，实验结果如图 12 所示。

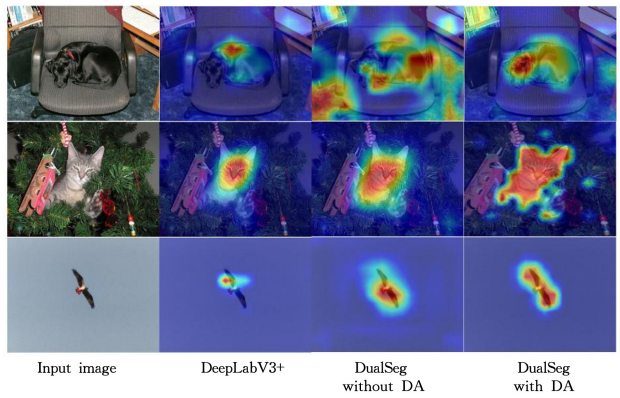


图 12 使用 grad-CAM 进行注意力可视化

Fig. 12 Attention visualization using grad-CAM

图 12 中，第一列图像为输入图像；第二列图像为 DeepLabV3+模型的可视化结果，DeepLabV3+模型能够较好地关注待分割的对象，但是边界不够清晰；第三列图像为 DualSeg 模型未引入 DoubleAttention 模块的可视化结果，该模型使用对偶学习方法，并使用多层次特征融合和空间金字塔池化模块，能够较好地关注边界信息；第四列图像为 DualSeg 模型引入 DoubleAttention 双重注意力模块的可视化结果。由图可知，引入双重注意力模块后，能够更加精确地感知图像的边界，有利于场景分割任务。

结束语 本文将对偶学习方法用于场景分割任务与图像重建任务，根据对偶学习策略，重新设计损失函数，使得两个任务互相促进。此外，提出的多层次特征融合模块和双重注意力模块在复杂场景下增强了特征图获取全局上下文的能力，提升了场景分割性能。并且，轻量化的双重注意力模块和轻量化的骨干网络减少了模型的参数量。本文提出的算法在复杂场景数据集上达到了良好的性能。但是本文的研究还存在以下不足，有待今后进一步改进：1）模型的参数量还可以继续优化，虽然采用了较为轻量的 Xception 作为主干网络，但是仍然不如 MobileNet 轻量化，后续考虑继续将模型轻量化；2）对偶学习部分，后续可以考虑引入超分辨率图像重建，使得用于反馈的网络提供的图像信息更加准确，更有利于边界的分割。

参 考 文 献

[1] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:3431-3440.

[2] 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.

[3] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv:1412.7062, 2014.

[4] 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, 2017, 40(4):834-848.

[5] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv: 1706.05587, 2017.

[6] 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.

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv:2010.11929, 2020.

[8] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[9] ZHENG S, LU J, ZHAO H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:6881-6890.

[10] XIE E, WANG W, YU Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[11] CHENG B, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:1290-1299.

[12] JAIN J, LI J, CHIU M T, et al. Oneformer: One transformer to rule universal image segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:2989-2998.

[13] LUO P, WANG G, LIN L, et al. Deep dual learning for semantic image segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:2718-2726.

[14] CHEN J L, PENG Y B, LI N. Single image super-resolution reconstruction network based on dual learning strategy [J]. Computer Application Research, 2021, 38(7):2235-2240.

[15] WANG L, LI D, ZHU Y, et al. Dual super-resolution learning for semantic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3774-3783.

[16] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:3213-3223.

[17] EVERINGHAM M, ESLAMI S M A, VAN G L, et al. The pascal visual object classes challenge: A retrospective[J]. International Journal of Computer Vision, 2015, 111:98-136.

[18] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:1251-1258.

[19] BOYD S P, VANDENBERGHE L. Convex optimization [M]. Cambridge University Press, 2004.

[20] SUN K, ZHAO Y, JIANG B, et al. High-resolution representations for labeling pixels and regions[J]. arXiv: 1904.04514, 2019.

[21] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:2881-2890.

[22] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Confe-rence on Computer Vision and Pattern Recognition. 2016:770-778.

[23] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:618-626.

LIU Sichun, born in 1998, postgraduate. Her main research interests include deep learning and computer vision.

WANG Xiaoping, born in 1965, Ph. D, professor. His main research interests include AI algorithms, deep learning and computer vision.

(责任编辑:杨雪敏)