JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
07
VOL.27

ISSN1006-8961
CN11-3758/TB

舰船目标检测 *P2078*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第7期（总第315期）
2022年7月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

2022年7月 第27卷 第7期（总第315期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

深度学习背景下视觉显著性物体检测综述(第2112页)

提取全局语义信息的场景图生成算法(第2214页)

融合弱监督目标定位的细粒度小样本学习(第2226页)

Volume 27, Number 7
Published July 16, 2022

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Review of deep learning based salient object detection(P2112)

Global semantic information extraction based scene graph generation algorithm(P2214)

Weakly-supervised object localization based fine-grained few-shot learning(P2226)

©中国图象图形学报版权所有

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2022)07-2157-14

论文引用格式:Li Z X, Zhang J, Wu J L and Ma H F. 2022. Semi-supervised adversarial learning based semantic image segmentation. Journal of Image and Graphics,27(07):2157-2170(李志欣,张佳,吴璟莉,马慧芳. 2022. 基于半监督对抗学习的图像语义分割. 中国图象图形学报,27(07):2157-2170)[DOI:10.11834/jig.200600]

基于半监督对抗学习的图像语义分割

李志欣[1*],张佳[1],吴璟莉[1],马慧芳[2]

1. 广西师范大学广西多源信息挖掘与安全重点实验室, 桂林　541004; 2. 西北师范大学计算机科学与工程学院, 兰州　730070

摘　要:**目的**　将半监督对抗学习应用于图像语义分割,可以有效减少训练过程中人工生成标记的数量。作为生成器的分割网络的卷积算子只具有局部感受域,因此对于图像不同区域之间的远程依赖关系只能通过多个卷积层或增加卷积核的大小进行建模,但这种做法也同时失去了使用局部卷积结构获得的计算效率。此外,生成对抗网络(generative adversarial network, GAN)中的另一个挑战是判别器的性能控制。在高维空间中,由判别器进行的密度比估计通常是不准确且不稳定的。为此,本文提出面向图像语义分割的半监督对抗学习方法。**方法**　在生成对抗网络的分割网络中附加两层自注意模块,在空间维度上对语义依赖关系进行建模。自注意模块通过对所有位置的特征进行加权求和,有选择地在每个位置聚合特征。因而能够在像素级正确标记值数据的基础上有效处理输入图像中广泛分离的空间区域之间的关系。同时,为解决提出的半监督对抗学习方法的稳定性问题,在训练过程中将谱归一化应用到对抗网络的判别器中,这种加权归一化方法不仅可以稳定判别器网络的训练,并且不需要对唯一的超参数进行密集调整即可获得满意性能,且实现简单,计算量少,即使在缺乏互补的正则化技术的情况下,谱归一化也可以比权重归一化和梯度损失更好地改善生成图像的质量。**结果**　实验在 Cityscapes 数据集及 PASCAL VOC 2012(pattern analysis, statistical modeling and computational learning visual object classes)数据集上与9种方法进行比较。在 Cityscapes 数据集中,相比基线模型,性能提高了2.3%～3.2%。在 PASCAL VOC 2012 数据集中,性能比基线模型提高了1.4%～2.5%。同时,在 PASCAL VOC 2012 数据集上进行消融实验,可以看出本文方法的有效性。**结论**　本文提出的半监督对抗学习的语义分割方法,通过引入的自注意力机制捕获特征图上各像素之间的依赖关系,应用谱归一化增强对抗生成网络的稳定性,表现出了较好的鲁棒性和有效性。

关键词:半监督学习;卷积神经网络(CNN);图像语义分割;生成对抗网络(GAN);自注意机制;谱归一化

Semi-supervised adversarial learning based semantic image segmentation

Li Zhixin[1*], Zhang Jia[1], Wu Jingli[1], Ma Huifang[2]

1. Guangxi Key Laboratory of Multi-source Information Mining and Security, Guangxi Normal University, Guilin　541004, China;
2. College of Computer Science and Engineering, Northwest Normal University, Lanzhou　730070, China

Abstract: **Objective** Deep learning network training models is based on labeled data. It is challenged to obtain pixel-level label annotations for labor-intensive semantic segmentation. However, the convolution operator of the segmentation network

收稿日期:2020-10-10;**修回日期**:2021-03-15;**预印本日期**:2021-03-22

*通信作者:李志欣 lizx@gxnu.edu.cn

基金项目:国家自然科学基金项目(61966004,61866004,61762015,61762078);广西自然科学基金项目(2019GXNSFDA245018,2018GXNSFDA281009)

Supported by:National Natural Science Foundation of China (61966004, 61866004, 61762015, 61762078)

has single local receptive field as the generator, but the size of each convolution kernel is very limited, and each convolution operation just cover a tiny pixel-related neighborhood. The height and width of long-range feature map is dramatically declined due to the constraints of multi-layer convolution and pooling operations. The lower the layer is, the larger the area is covered, which is via the mapped convolution kernel retrace to the original image, which makes it difficult to capture the long-range feature relationship. It is a challenge to coordinate multiple convolutional layers to capture these dependent parameter values in detail via optimization algorithm. Therefore, the long-range dependency between different regions of the image can just be modeled through multiple convolutional layers or the enlargement of the convolution kernel, but this local convolution structural approach also loses the computational efficiency. In addition, another generative adversarial network (GAN) challenge is the manipulation ability of the discriminator. The discriminator training is equivalent to training a good evaluator to estimate the density ratio between the generated distribution and the target distribution. The discriminator based density ratio estimation is inaccurate and unstable in related to high-dimensional space in common. The better the discriminator is trained, the more severed gradient returned to the generator is ignored, and the training process will be cut once the gradient completely disappeared. The traditional method proposes that the parameter matrix of the discriminator is required to meet the Lipschitz constraint, but this method is not detailed enough. The method limit the parameter matrix factors, but it is not greater than a certain value. Although the Lipschitz constraint can also be guaranteed, the structure of the entire parameter matrix is disturbed due to the changed proportional relationship between the parameters. **Method** The semi-supervised adversarial learning application can effectively reduce the number of manually generated labels to semantic image segmentation in the training process. Our segmentation network is used as the generator of the GAN, and the segmentation network outputs the semantic label probability map of a targeted image. Hence, the output of the segmentation network is possible close to the ground truth label in space. The fully convolutional neural network (CNN) is used as the discriminator. When doing semi-supervised training, the discriminator can distinguish the ground truth label map from the class probability map predicted by the segmentation network. The discriminator network generates a confidence map that it can be used as a supervision signal to guide the cross-entropy loss. Based on the confidence map, it is easy to see the regions in the prediction distribution that are close to the ground truth label distribution, and then use the masked cross-entropy loss to make the segmentation network trust and train these credible predictions. This method is similar to the probabilistic graphical model. The network does not increase the computational load because redundant post-processing modules are not appeared in the test phase and discriminator is not needed in the inference process. We extend two layers of self-attention modules to the segmentation network of GAN, and model the semantic dependency in the spatial dimension. The segmentation network as a generator can precisely coordinate the fine details of each pixel position on the feature map with the fine details in the distance part of the image through this attention module. The self-attention module is optional to aggregate the features at each location via a weighted sum on the features of multifaceted locations. Therefore, the relationship between widely discrete spatial regions in the input image can be effectively processed based on pixel-level ground truth data. A good balance is achieved between long-range dependency modeling capabilities and computational efficiency. We carry out spectral normalization to the discriminator of the adversarial network during the training process. This method introduces Lipschitz continuity constraints from the perspective of the spectral norm of the parameter matrix of each layer of neural network. The neural network beyond the disturbance of the input image and make the training process more stable and easier to converge. This is a more refined way to make the discriminator meet Lipschitz connectivity, which limits the violent degree of function changes and makes the model more stable. This weighted normalization method can not only stabilize the training of the discriminator network, but also obtain satisfactory performance without intensive adjustment of the unique hyper-parameters, and is easy to implement and requires less calculation. When spectral normalization is applied to adversarial generation networks on semantic segmentation tasks, the generated cases are also more diverse than traditional weight normalization. In the absence of complementary regularization techniques, spectral normalization can even improve the quality of the generated image better than other weight normalization and gradient loss. **Result** Our experiment is compared to the latest 9 methods derived from the Cityscapes dataset and the pattern analysis, statistical modeling and computational learning visual object classes (PASCAL VOC 2012) dataset. In the Cityscapes dataset, the performance is improved by 2.3% to 3.2% compared to the baseline model. In the PASCAL VOC 2012 dataset, the performance is improved by 1.4% to

2.5% over the baseline model. Simultaneously an ablation experiment is conducted on the PASCAL VOC 2012 dataset. **Conclusion** The semantic segmentation method of semi-supervised adversarial learning proposed uses the introduced self-attention mechanism to capture the dependence between pixels on the feature map. The application of spectral normalization to stabilize the adversarial generation network has its qualified robustness and effectiveness.

Key words: semi-supervised learning; convolutional neural network (CNN); semantic image segmentation; generative adversarial network (GAN); self-attention mechanism; spectral normalization

0 引　言

图像语义分割是图像处理和计算机视觉领域的一项重要工作,像素级的分割任务通常称为图像语义分割。随着卷积神经网络(convolutional neural network, CNN)的发展,图像语义分割取得了显著进展(Chen 等,2018;Long 等,2015;Oliver 等,2019;Yu 和 Koltun,2016),广泛应用于自动驾驶(Geiger 等,2012)、图像编辑(Tsai 等,2017)等领域。然而在实际的图像语义分割应用中,在进行全监督 CNN 训练时,通常需要大量像素级标注的真实标记(ground truth, GT)传达对象边界及其组成部分之间的关系,这些数据通常是人工获取的,需要付出巨大代价。为了减少训练过程中使用人工生成标记的数量,最常见的是在图像语义分割中采用半监督或弱监督的训练方法。

半监督学习方法的关键是其使用的弱标记数据仅表示某个对象类的存在,不提供对象位置或边界的 GT 信息。显然,这些注释比像素级的标记弱,且在大量的可视数据中很容易获得,或者说能以相对较低的成本手工获得。因此,半监督学习为训练具有有限标记数据和大量未标记数据的图像语义分割模型提供了一种有吸引力的方法。针对图像语义分割,目前提出了多种半监督训练方法。Kalluri 等人(2019)将半监督学习和无监督域适应结合起来。Stekovic 等人(2019)实现了 3 维场景的多个视图之间的几何约束。一致性正则化(Oliver 等,2019)代表了一类用于训练深度神经网络分类器的半监督学习算法,它也被开发用于产生最先进的半监督分类结果,这些结果在概念上很简单,而且通常容易实现。Kingma 和 Ba(2017)注意到在图像语义分割中使用图像级标注依赖于分类网络获得的位置图来拟合图像级标注和像素级标注之间的差距。然而,这些特征图只关注物体的一小部分,没有精确的边界表示。因此,目前已有的采用半监督学习的图像分割方法与全监督网络训练方法相比效果较差。

在卷积过程中的优化算法可能无法适当地协调多个卷积层以捕获这些依赖性的参数值,因此可能会妨碍远程依赖性的学习。虽然增大卷积核的大小可以提高网络的表达能力,但这样做也会失去使用局部卷积结构获得的计算效率。上下文依赖关系已经在许多方面得到解决,例如,学习上下文已经被证明依赖于局部特征,并且有助于特征表示。Shuai 等人(2018)使用递归神经网络(recursive neural network, RNN)创建有向无环图模型,以捕获丰富的上下文依赖关系。Zhao 等人(2018)提出 PSANet(point-wise spatial attention network),通过卷积和空间维的相对位置信息捕获像素级关系。此外,Zhang 等人(2018)提出的 EncNet(context encoding network)引入了一个通道注意机制来捕获全局上下文。注意模块对远程依赖关系的建模能力已经得到了证明,也已经在许多任务中得到了广泛的应用。同时,自注意力机制在计算机视觉领域的应用也越来越广泛。Vaswani 等人(2017)利用一种自注意机制训练更好的分类生成器。然而,这些工作目前并没有有效地应用于半监督图像语义分割。因此,本文尝试将自注意机制应用于半监督图像语义分割任务中,并获得了很好的效果。

生成对抗网络(generative adversarial network, GAN)(Goodfellow 等,2014)的发展使得半监督和弱监督学习在图像语义分割中的应用取得了显著进展。判别器的性能控制是提出的半监督对抗学习图像语义分割在训练过程中面临的另一个挑战。在这里,高维空间中判别器的密度比估计在训练中往往是不准确且不稳定的。优化的目的是获得一个能够很好地区分生成分布和目标分布(Arjovsky 和 Bottou,2017)的判别器。然而一旦获得这样一个判别器,生成器的训练就完全停止了。为了提高 GAN 训练的稳定性,众多研究者做了各种努力。Radford 等

人(2016)为了从体系结构设计的角度寻找一套更好的网络架构设置,开发了 DCGAN(deep convolutional GAN)模型,在图像生成领域进行了广泛的实验验证。Wasserstein GAN 模型(Arjovsky 等,2017)通过引入 Wasserstein 距离的概念,从理论角度解决了 GAN 训练不稳定问题。在 Wasserstein GAN 模型中,判别器的参数矩阵必须满足 Lipschitz 约束。但是采用的约束方法相对简单粗暴,直接约束参数矩阵中的元素,使其不大于给定值。尽管此方法可以保证 Lipschitz 约束,但也破坏了参数之间的比例关系。本文使用的谱归一化是一种既满足 Lipschitz 条件,又不破坏矩阵结构的方法,仅需使每层网络的网络参数除以该层参数矩阵的谱范数即可满足 Lipschitz 等于 1 的约束,因此实现起来也较为简单。

本文提出一种稳定的自注意半监督对抗性学习方法,使用的基础分割网络是基于 DeepLabv2 框架(Chen 等,2018)和在 MSCOCO(Microsoft common objects in context)(Lin 等,2014)数据集上预训练的 ResNet-101(residual neural network)模型。其中利用一个全卷积的判别器,产生一个像素级的置信度图以区分生成器生成的数据和 GT 分割图。置信图中的每个像素可以用一个简单的阈值分割成 0 或 1,1 表示可信预测结果,0 表示结果不可信。将置信图作为掩膜,将分割预测看做假标记,用于训练分割网络。分割网络可以不断学习不同的不可见模式,以寻求和改进优化,然后在不可见的类别中识别新的模式。本文的贡献主要有:1)在 GAN 的分割网络中引入一种自注意机制,在基于像素级 GT 数据的半监督对抗训练中,通过计算特征图中任意两个位置之间的相互作用,直接捕获其远程依赖关系;2)采用谱归一化(Miyato 等,2018)稳定判别器网络的训练,不需要对超参数进行大范围的调整即可达到较好的判别器训练效果,与其他常用方法相比,这种归一化技术计算量更小,更容易集成到当前的实现方法中;3)在数据集 PASCAL VOC 2012(pattern analysis, statistical modeling and computational learning visual object classes)(Everingham 等,2010)和 Cityscapes(Cordts 等,2016)上进行实验评估,与当前先进的半监督和全监督图像语义分割方法相比,本文方法具有更好的性能。实验以 Hung 等人(2018)提出的 AdvSemiSeg 方法作为半监督基线模型,以 DeepLabv2 网络为全监督方法基线模型。此外,本文给出了在不应用谱归一化的情况下(即仅应用自注意模块)获得的性能。

1 相关工作

深度学习在图像分类中的一些突破性方法已用于图像语义分割任务,但图像语义分割任务的核心是如何将分割与分类两项任务结合起来。很多分割方法都采用迁移学习,通常以 ResNet(He 等,2016)和 VGG(Visual Geometry Group)(Simonyan 和 Zisserman,2015)分类网络的卷积层作为骨干。Long 等人(2015)提出的将卷积 21 路分类器与 VGG-16 骨干网相连接的全卷积网络(fully convolutional network,FCN)的应用,证明了深度神经网络在图像语义分割中的有效性。Chen 等人(2018)将空洞卷积应用于 VGG-16 网络的后几层,在保持接收域的同时提高预测的空间分辨率。编解码器架构(Ding 等,2018)在图像语义分割中也得到了应用。编码器是一种特殊的神经网络,用于特征提取和数据降维,生成具有语义信息的特征图像。解码器网络的作用是将编码器网络输出的低分辨率特征图像映射回输入图像的大小,进行逐像素分类。U-Nets(Ronneberger 等,2015)使用置换卷积层增加分辨率,其跳跃连接带有完整的特征图。

以上方法都表现出了非常优越的性能,但在训练过程中都需要大量的标记数据,通常要在有像素级注释的大型数据集上进行训练,例如数据集 PASCAL VOC 2012 和 Cityscapes 等,获取这些标记数据非常耗时且昂贵。一些研究采用半监督方法,即只使用部分标记数据的训练处理这个问题。在本文工作中考虑的半监督方法,可以分为仅使用图像级标记和仅使用边界框(Sun 和 Li,2019)的弱注释数据的方法,或者只是对部分数据进行标记而另外部分数据完全未标记的方法。Luc 等人(2016)首先将对抗学习引入图像语义分割,Souly 等人(2017)为其在半监督学习中的应用开辟了道路。Liu 等人(2019)在基于对抗学习的方法中采用全卷积判别器,试图在像素级区分预测概率图和 GT 分割分布。这些工作都是只标记部分数据集,另外未标记的数据来自相同的数据集,并且与标记的数据共享相同的域数据分布。

Goodfellow 等人(2014)重新引入图像生成任务

的对抗性学习的概念，并使用 GAN 成功地从随机噪声中生成了手写数字和人脸等图像。然而，随机噪声和有意义的图像显然来自不同的数据域，且分布不一致。因此，GAN 模型可以解决不同数据域间分布不一致的问题。大多数生成器从噪声矢量生成图像。Liu 等人（2019）提出使用 GAN 生成低显示类别的真实图像以增强数据，从而平衡标记分布。Hung 等人（2018）提出使用对抗网络促进小规模数据集中的语义分割，当给定一个特定图像时，判别器用来输出语义标记的置信度图，经过这样调整可以强制分割预测，使参数在空间上更接近 GT，之后该生成器可以在半监督设置下提高分割精度。

自注意（Vaswani 等，2017）最初的目的是为了解决机器翻译问题，在随后的工作中进一步提出了非局部神经网络（Wang 等，2018），用于视频分类、目标检测和实例分割等一系列任务。Hu 等人（2018）还应用自注意机制对对象之间的关系进行建模，以实现更好的对象检测。最近的一些工作（Zhang 等，2019）将类似的机制应用于语义分割并取得了良好的分割性能。本文的工作与上述工作密切相关，处理高分辨率输入，通过关注所有输入位置，计算每个输出位置的上下文信息，对远程依赖关系进行建模，通过配备自注意机制的单层模型模拟输入特征图中任何位置之间的依赖关系。

尽管 GAN 在改善数据驱动的生成模型的样本质量方面非常成功（Brock 等，2019；Karras 等，2018），但对抗训练也导致了 GAN 的不稳定性，已有的工作（Arjovsky 等，2017）表明，GAN 的这种不稳定性是由于梯度爆炸和梯度消失导致的。一个标准的抗扰训练方案涉及使用抗扰样本拟合判别器（Szegedy 等，2014），目的是产生一个训练有素的判别器，该判别器对测试样本的攻击具有更好的鲁棒性。为了提高 GAN 的稳定性，学者提出许多方法，包括利用不同的体系机构（Radford 等，2016）、采用正则化技术（Salimans 和 Kingma，2016）和梯度惩罚（Gulrajani 等，2017）等。谱归一化技术（Miyato 等，2018）是最好的方法之一，本文通过在 GAN 结构的判别器中引入谱归一化，达到了控制判别器的 Lipschitz 常数的效果，缓解了梯度消失问题，提高了 GAN 训练的稳定性。

2　模型概述

提出的半监督图像语义分割的方法框架主要由两个子网络构成，包括分割网络 G 和判别器 D，如图 1 所示。其中分割网络 G 输出类别概率图，SA（self-attention）表示自注意力模块，SN（spectral normalization）表示应用谱归一化技术，判别器网络 D 输出置信度图，L_{ce} 是基于 GT 图像的标准交叉熵损失，L_{adv} 是 D 的对抗损失，L_{semi} 是掩膜交叉熵损失。分割网络中输入第 n 个图像 $\boldsymbol{X}_n$ 的尺寸为 $H \times W \times 3$。G 中的特征图通过引入的两层自注意模块，首先应用卷积层获取降维特征，然后将输入自注意模块的特征生成一个空间注意矩阵，对特征图的任意两个像素之间的空间关系进行建模。接下来，在自注意矩阵和原始特征之间执行矩阵乘法。最后，对上面相乘的结果和原始特征进行逐元素的求和运算，获得远程上下文的表示。这使生成器可以基于局部特征对丰富的上下文关系进行建模，从而在生成图像时可以很好地协调每个位置和远端的细节。输出是维度 $H \times W \times C$ 的类概率图，其中 C 为语义类的个数。

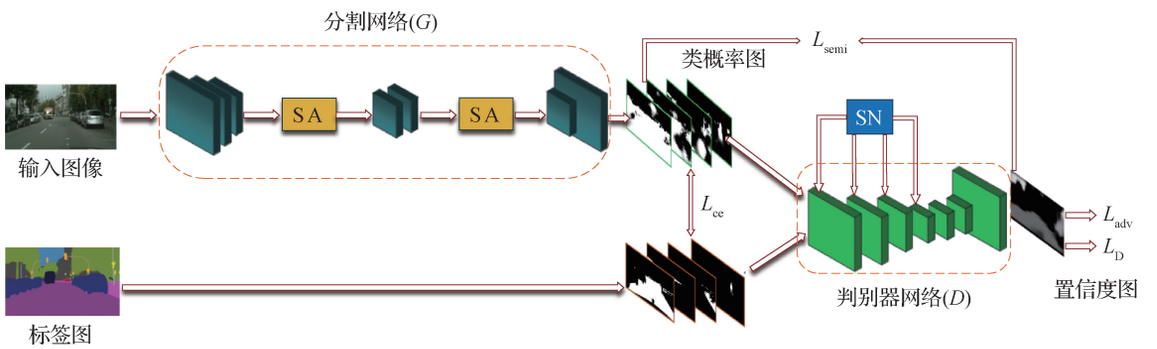


图 1　半监督语义分割方法框架图

Fig. 1　Framework of semi-supervised semantic segmentation method

通过使用空间交叉熵损失 L_{D} 训练基于全卷积的判别器网络。判别器 D 可以接受不同大小的输入,其由 G 输出的类概率图($G(\boldsymbol{X}_n)$)或一个独热编码的标记图 $\boldsymbol{I}_n$ 作为输入,最终输出一个尺寸为 $H\times W\times 1$ 的置信度图。这里,对于置信度图的每个像素 i,如果来自分割网络 G,则设为 0;如果来自标记图,则设为 1。因此,置信图表示 G 的概率预测输出更接近 GT 分布的区域。同时在 D 内应用谱归一化,保证其映射函数满足 Lipschitz 约束。

使用未标记图像和标记图像进行半监督训练。在整个训练过程中,将未标记的数据应用于训练 G,而附加的自注意模块则有效地解决了输入图像中广泛分离的空间区域之间的关系。当使用标记数据时,G 的训练将同时根据基于 $\boldsymbol{I}_n$ 的标准交叉熵损失 L_{ce} 和从 D 获得的对抗损失 L_{adv} 进行监督。然后,根据置信度图给出的可信预测,以自学习的方式将置信度图和掩膜交叉熵损失 L_{semi} 一起用做训练 G 的监督信号。

3 半监督损失

3.1 损失函数

训练分割网络通过最小化多任务损失函数实现,具体为

$$L_G = L_{\mathrm{ce}} + \lambda_{\mathrm{adv}} L_{\mathrm{adv}} + \lambda_{\mathrm{semi}} L_{\mathrm{semi}} \tag{1}$$

式中,λ_{adv} 和 λ_{semi} 是权重,用于最小化多任务损失函数。对于式(1)中的第 1 个损失分量,标准交叉熵损失定义为

$$L_{\mathrm{ce}} = -\sum_{h,w}\sum_{c\in \boldsymbol{U}} \boldsymbol{I}_n^{(h,w,c)} \log\left(G(\boldsymbol{X}_n)^{(h,w,c)}\right) \tag{2}$$

式(2)是为了将离散的 GT 信息映射转换为一个 c 通道概率映射。GT 信息的映射采用独热编码方案,其中,$\boldsymbol{U}$ 为所有语义类集合,如果 $\boldsymbol{X}_n^{(h,w,c)}$ 中的像素属于 c 类,则 $\boldsymbol{I}_n{}^{(h,w,c)}$ 取 1,否则取 0。式(1)中第 2 个损失分量,对抗损失定义为

$$L_{\mathrm{adv}} = -\sum_{h,w} \log\left(D(G(\boldsymbol{X}_n))^{(h,w)}\right) \tag{3}$$

式中,$D(G(\boldsymbol{X}_n))^{(h,w)}$ 是 $\boldsymbol{X}_n$ 在位置 (h,w) 处的置信度图。因为未标记的数据不包含 GT,所以未标记的数据不会产生与 L_{ce} 相关的损失,此时只需要判别器网络,即此时对抗损失 L_{adv} 仍然适用。最后,使用指标函数 $F(\cdot)$ 和阈值 T_{semi} 定义式(1)中的第 3 个损失分量,以对置信度图进行二值化,更好地显示可信区域。第 3 个损失分量,掩膜交叉熵损失可表示为

$$L_{\mathrm{semi}} = -\sum_{h,w}\sum_{c\in \boldsymbol{U}} F\left(D(G(\boldsymbol{X}_n))^{(h,w)} > T_{\mathrm{semi}}\right) \cdot \hat{\boldsymbol{I}}_n^{(h,w,c)} \log\left(G(\boldsymbol{X}_n)^{(h,w,c)}\right) \tag{4}$$

令 $c^* = \underset{c}{\operatorname{argmax}}\, G(\boldsymbol{X}_n)^{(h,w,c)}$,则自学习的独热编码的标记图 $\hat{\boldsymbol{I}}_n$ 由 $\hat{\boldsymbol{I}}_n^{(h,w,c^*)} = 1$ 逐元素设置。在训练过程中,自学习目标 $\hat{\boldsymbol{I}}_n$ 和 $F(\cdot)$ 的乘积视为一个常数。实验表明,$T_{\mathrm{semi}} = 0.2$ 时,在训练过程中具有良好的鲁棒性。

通过最小化空间交叉损失函数 L_{D} 训练判别器网络,具体为

$$L_{\mathrm{D}} = -\sum_{h,w} (1-y_n)\log\left(1-D(G(\boldsymbol{X}_n))^{(h,w)}\right) + y_n \log\left(D(\boldsymbol{I}_n)^{(h,w)}\right) \tag{5}$$

式中,如果判别器输入为 $G(\boldsymbol{X}_n)$,则 $y_n = 0$;如果判别器输入为 $\boldsymbol{I}_n$,则 $y_n = 1$,而 $D(\boldsymbol{I}_n)^{(h,w)}$ 是 $\boldsymbol{I}_n$ 在位置 (h,w) 处的置信度图。

3.2 自注意模块

传统的 GAN 网络使用小的卷积核很难发现图像中的依赖关系,但使用大的卷积核就丧失了卷积网络参数与计算的效率。尤其在语义分割这种多类别的数据集上训练时,卷积 GAN 网络对某些图像类的建模比其他图像类的建模更困难。在本文提出的半监督图像语义分割框架的分割网络 G 中,每个卷积核的尺寸均有限,每次卷积操作只能覆盖像素点周围很小一块邻域,对距离较远的特征不容易捕获,因为多层的卷积和池化操作使得特征图的宽和高变得越来越小,越靠后的卷积层,卷积核覆盖区域映射回原图时对应的面积也就越大。自注意通过直接计算图像中任意两个像素点之间的关系,获取图像的全局几何特征,通过关注特征图所有位置,并在嵌入空间中取其加权平均值表示特征图中某位置处的响应。简单来说就是在前一层的特征图上加入注意力机制,使得 GAN 在生成时能够区别不同的特征图。

给定一个像素点,为了计算特征图上所有像素点对这个点的影响,需要用一个函数,针对特征图 $\boldsymbol{Q}$ 中的某一个位置,计算特征图 $\boldsymbol{K}$ 中所有位置对它的影响。这个函数可以通过学习得到,因此考虑对这两个特征图分别做卷积核为 1×1 的卷积,且卷积核的权重可以学习得到。

本文提出的两层自注意模块的框架如图 2 所示,此处符号⊗表示矩阵对应元素相乘。该自注意

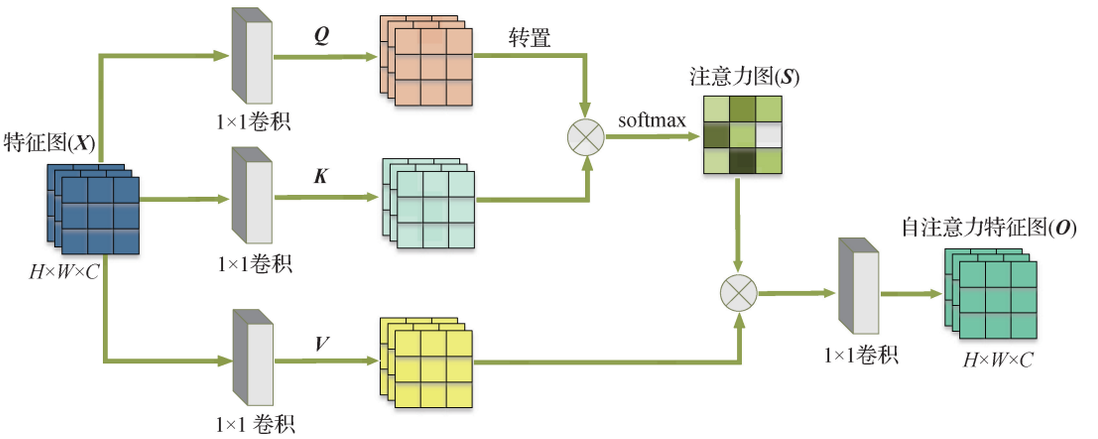


图 2　自注意力模块框架图

Fig. 2　Framework of self-attention module

模块以上一层的特征图 $\boldsymbol{X}\in\mathbf{R}^{H\times W\times C}$作为输入，并生成两个特征图 $\boldsymbol{Q}$、$\boldsymbol{K}\in\mathbf{R}^{H\times W\times C}$，在对 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 进行转置后执行矩阵乘法，并使用 softmax 层计算注意力图 $\boldsymbol{S}\in\mathbf{R}^{N\times N}$，其中 $N=H\times W$ 是像素数。这里，特征图 $\boldsymbol{S}$ 的元素表达了第 i 个像素对第 j 个像素的依赖性度量，即

$$\boldsymbol{S}_{ji}=\frac{\exp(\boldsymbol{Q}_i\cdot\boldsymbol{K}_j)}{\sum_{i=1}^{N}\exp(\boldsymbol{Q}_i\cdot\boldsymbol{K}_j)} \tag{6}$$

式中，$\boldsymbol{Q}_i$ 和 $\boldsymbol{K}_j$ 分别表示特征图 $\boldsymbol{Q}$ 的第 i 个位置的像素和特征图 $\boldsymbol{K}$ 中的第 j 个位置的像素所对应的 C 维向量。为了学习更多的参数，在原始特征图 $\boldsymbol{X}$ 中加入卷积映射，以获得新的特征图 $\boldsymbol{V}\in\mathbf{R}^{H\times W\times C}$，并将 $\boldsymbol{S}$ 和 $\boldsymbol{V}$ 进行转置以及矩阵乘法。$\boldsymbol{V}$ 可以看成对原特征图多加了一层卷积映射，这样可以学习到的参数更多，否则 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的参数太少。然后将其乘以比例参数 α。最后，自注意模块逐渐学习将注意加权特征图添加到原始特征图 $\boldsymbol{X}$ 中，即

$$\boldsymbol{O}_j=\alpha\sum_{i=1}^{N}(\boldsymbol{S}_{ji}\boldsymbol{V}_i)+\boldsymbol{X}_j \tag{7}$$

式中，$\boldsymbol{O}_j$ 表示第 j 个位置上的 C 维结果特征向量。$\boldsymbol{V}_i$ 和 $\boldsymbol{X}_j$ 分别表示特征图 V 的第 i 个位置的像素和原始特征图 $\boldsymbol{X}$ 的第 j 个位置的像素所对应的 C 维向量。α 初始化为 0，并且 α 通过自学习方式为非局部特征分配更多权重。最终特征图 $\boldsymbol{O}$ 是所有位置的特征与原始特征的加权总和。因此，对特征图之间的远程语义依赖性进行建模，有助于提高特征的可分辨性。

3.3　谱归一化

原始 GAN 网络的目标函数是为了优化真实数据分布与生成数据分布之间的 JS（Jensen-Shannon）散度。但存在的问题是判别器训练得越好，生成器的梯度消失得越严重。即当近似得到最优判别器时，最小化生成器的损失等价于最小化生成数据分布与真实数据分布之间的 JS 散度。可生成数据分布和真实数据分布几乎不可能有不可忽略的重叠，因此无论生成数据分布与真实数据分布相距多远，JS 散度都是常数，这也导致生成器的梯度最终会近似为 0，即梯度消失。

本文提出的半监督对抗学习图像语义分割方法存在的训练困难是如何控制判别器网络 D 的稳定性，因为在目标分布和生成分布分开的情况下，可以存在一个判别器能够完美地将生成数据和真实数据完全区分开。如果输入的真实图像没有归一化到[-1，1]，而生成的数据均在[-1，1]区间，那么在训练过程中，将会导致生成器 G 的梯度消失近似为 0。接下来再训练 G 时，生成的图像质量就很难提升。因为这两个分布差异很大，D 很容易区分，所以达到了最优化。

相对常规的 GAN，谱归一化后的 GAN 引入了新的正则项，该正则项防止权重矩阵的列空间在训练中只关心一个特定的方向，同时其防止 D 中每层的转换对某一个方向敏感。与 Wasserstein GAN 模型只对判别器的参数矩阵中的元素直接限制不同，谱归一化方法以一种温和的方式使判别器满足 Lipschitz 连续性，限制了判别器函数的变化剧烈程度，使模型更稳定。

对于标准 GAN，判别器 D 的最佳形式为

$$D_G^*(\boldsymbol{x})=\frac{q_{\text{data}}(\boldsymbol{x})}{q_{\text{data}}(\boldsymbol{x})+p_G(\boldsymbol{x})}=\text{sigmoid}(f^*(\boldsymbol{x})) \tag{8}$$

式中，q_{data}是数据 $\boldsymbol{x}$ 的分布，p_G 是对应的 $\boldsymbol{x}$ 的生成模型的分布，该模型是通过对抗性最小最大优化过程学习的，且 $f^*(\boldsymbol{x}) = \log q_{\text{data}}(\boldsymbol{x}) - \log p_G(\boldsymbol{x})$，其导数为

$$\nabla_x f^*(\boldsymbol{x}) = \frac{1}{q_{\text{data}}(\boldsymbol{x})} \nabla_x q_{\text{data}}(\boldsymbol{x}) - \frac{1}{p_G(\boldsymbol{x})} \nabla_x p_G(\boldsymbol{x}) \tag{9}$$

然而，这一导数项是无界的，甚至是不可计算的，在实践中必须加上常规的限制。因此，需要一种机制来定义 $f^*(\boldsymbol{x})$ 的导数。由此注意到，如果忽略 D 的每一层的偏置，则可以确定 $f^*(\boldsymbol{x})$ 的上限，具体为

$$\begin{aligned} \|f\|_{\text{Lip}} &\leqslant \|(\boldsymbol{h}_L \rightarrow \boldsymbol{W}^{L+1}\boldsymbol{h}_L)\|_{\text{Lip}} \times \|\alpha_L\|_{\text{Lip}} \times \\ &\|(\boldsymbol{h}_{L-1} \rightarrow \boldsymbol{W}^{L}\boldsymbol{h}_{L-1})\|_{\text{Lip}} \times \cdots \times \|\alpha_1\|_{\text{Lip}} \times \\ &\|(\boldsymbol{h}_0 \rightarrow \boldsymbol{W}^{1}\boldsymbol{h}_0)\|_{\text{Lip}} = \\ &\prod_{l=1}^{L+1} \|\boldsymbol{h}_{l-1} \rightarrow \boldsymbol{W}^{l}\boldsymbol{h}_{l-1}\|_{\text{Lip}} = \prod_{l}^{L+1} \sigma(W^l) \end{aligned} \tag{10}$$

式中，$\boldsymbol{h}$ 是对输入 $\boldsymbol{x}$ 具有的扰动向量，$\|\cdot\|_{\text{Lip}}$代表 Lipschitz 范数，谱归一化通过严格约束每层 $g: h_{\text{in}} \rightarrow h_{\text{out}}$控制判别函数 f 的 Lipschitz 常数。$\{\boldsymbol{W}^1, \cdots, \boldsymbol{W}^L, \boldsymbol{W}^{L+1}\}$ 是学习参数集，α_1 是非线性激活函数。$\sigma(\boldsymbol{W})$表示 $\boldsymbol{W}$ 的二范数，视为常数。根据线性性质，f 的上界是 1。据此，给出了矩阵 $\boldsymbol{W}$ 的谱归一化方法，即

$$\overline{W}_{SN}(\boldsymbol{W}) = \boldsymbol{W}/\sigma(\boldsymbol{W}) \tag{11}$$

然后，在式(10)的不等式中，将每个 $\boldsymbol{W}$ 代入式(11)。若对判别器 D 的各层权值 $\boldsymbol{W}$ 进行如上所示的谱归一化处理，则判别器 D 可视为隐式 f 的函数，其 Lipschitz 范数可约束为小于 1。这达到了限制判别器 D 的 Lipschitz 范数的效果。

谱归一化的简单表述是每层的权重 $\boldsymbol{W}$ 在更新后都除以 $\boldsymbol{W}$ 的最大奇异值。但是奇异值的分解计算是很耗时的，因而采用幂迭代的方式获得近似的最大奇异值的解。

4 实验结果分析

4.1 数据集与实验设置

PASCAL VOC 2012 数据集包含 21 个对象类，利用分割边界数据集(Hariharan 等，2011)的额外注释图像，共得到 10 582 幅图像用于训练，测试集包括 1 449 幅图像。Cityscapes 数据集包含 19 个类，其中训练集、验证集和测试集分别包含 2 975、500 和 1 525 幅图像。将平均交并比(mean intersection-over-union，mIoU)作为评估指标。随机抽取 1/8、1/4、1/2 等不同比例的标记数据，其余为未标记数据进行训练，并对模型的图像分割性能进行评估。在训练过程中未标记数据和标记数据均随机抽取，对所有基线使用相同的数据分割。

在 PASCAL VOC 2012 数据集的训练过程中，采用尺寸为 321 × 321 像素的随机缩放和裁剪操作。批处理大小为 8。对于 Cityscapes 数据集，将输入图像尺寸调整为 512 × 1 024 像素，没有随机裁剪/缩放，批处理大小为 2。在半监督训练中，随机抽取无标记和有标记的数据。对判别器网络和分割网络进行联合训练。在每次迭代中，只使用包含 GT 的数据训练判别器。

本文使用 PyTorch 框架在一个具有 11 GB 内存的 NVIDIA 1080TI GPU 上训练的模型，采用随机梯度下降(stochastic gradient descent，SGD)优化器，动量为 0.9，权值衰减为 10^{-4}。初始学习速率为 2.5×10^{-4}，并随着多项式衰减以 0.9 次方减小。判别器的训练采用 Adam 优化器，学习率设置为 10^{-4}。使用未标记和标记数据进行训练时，设置 λ_{adv} 为 0.001，λ_{semi}为 0.1，T_{semi}为 0.2。

4.2 在 Cityscapes 数据集上的实验结果

为了验证本文方法的性能，在 Cityscapes 数据集上使用不同比例标记数据进行实验，并将本文方法与当前具有代表性的半监督和全监督方法进行分割性能对比。对比方法包括 FCN-8s(Long 等，2015)、Dilation10(Yu 和 Koltun，2016)、CowMix(French 等，2020a)、DST-CBC(dynamic self-training and class-balanced curriculum)(Feng 等，2021)、Sawatzky 等人(2021)、CutMix(French 等，2020b)、Mittal 等人(2021)、DeepLabv2(Chen 等，2018)和 AdvSemiSeg(Hung 等，2018)。表 1 给出了 Cityscapes 数据集的半监督和全监督评估结果。

从表 1 可以看出，与作为基线的 AdvSemiSeg 模型相比，仅使用自注意情况下，性能提高了 1.4%～2.4%，加入谱归一化后，性能提高了 2.3%～3.2%。与基线相比，使用 1/8 标记数据训练的分割性能增加了 3.2%。由此推测在低标记数据条件下，两阶段 GAN 训练效果较差，判别器仅根据标记样本进行更新。这减少了在训练中看到的数据量，容易导致过拟合。

4.3　在 PASCAL VOC 2012 数据集上的实验结果

表 2 列出了在 PASCAL VOC 数据集上的半监督和全监督训练的平均 mIOU 性能结果。在仅使用自注意的情况下，性能比基线 AdvSemiSeg 模型提高了 0.6% ～ 2.1%，添加谱归一化后，性能提高了 1.4%（标记数据为 1/4 比例）～2.5%（标记数据为 1/2 比例）。

图3给出了GT图像与本文方法在训练期间以

表 1　在 Cityscapes 数据集上随机抽取不同比例的标记数据进行训练的图像分割性能结果(mIoU)

Table 1　Image segmentation performance with randomly selected different proportions of labeled data for training on the Cityscapes dataset(mIoU)

/%

方法	标记数据			
	1/8	1/4	1/2	全部
FCN-8s(Long 等,2015)	–	–	–	65.3
Dilation10(Yu 和 Koltun,2016)	–	–	–	67.1
CowMix(French 等,2020a)	60.5	64.1	–	69.0
DST-CBC(Feng 等,2021)	60.5	64.4	–	66.9
Sawatzky 等人(2021)	63.3	65.4	66.1	66.3
CutMix(French 等,2020b)	63.4	65.2	67.7	–
DeepLabv2(Chen 等,2018)	55.5	59.9	64.1	66.4
AdvSemiSeg(Hung 等,2018)	58.8	62.3	65.7	67.7
本文(自注意)	61.2	63.7	67.5	70.4
本文(自注意 + 谱归一化)	62.0	64.6	68.3	71.1

注：“–”表示对应文献未进行该项实验。

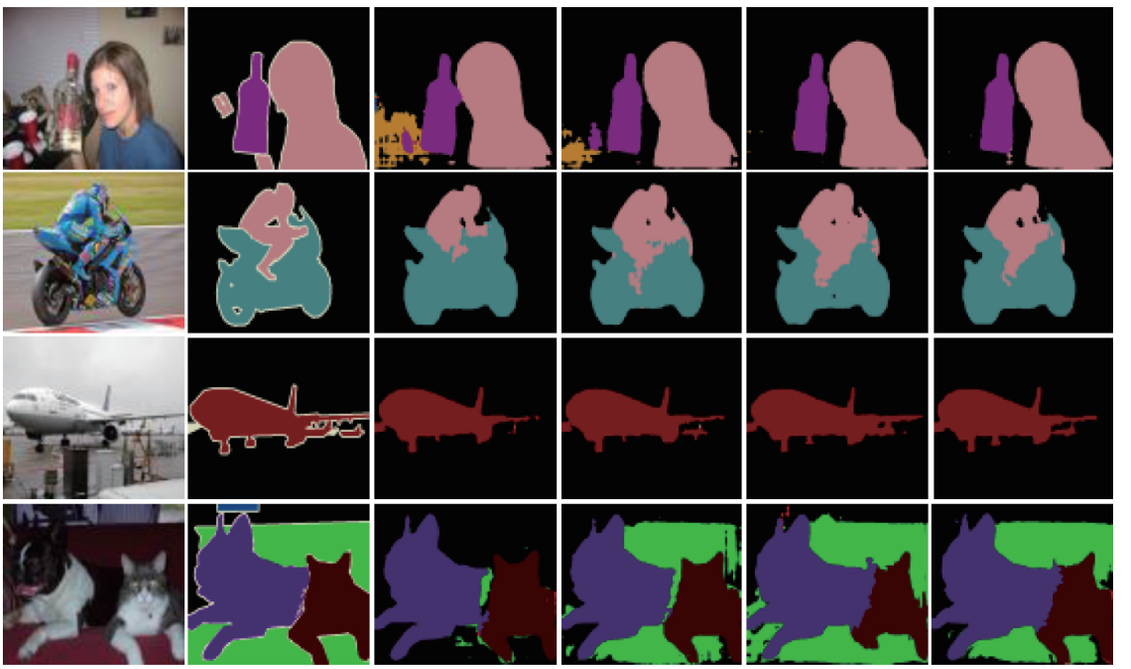
表 2　在 PASCAL VOC 2012 数据集上随机抽取不同比例的标记数据进行训练的图像分割性能结果(mIoU)

Table 2　The image segmentation performance with randomly extracting different proportions of labeled data for training on the PASCAL VOC 2012 dataset(mIoU)

/%

方法	标记数据			
	1/8	1/4	1/2	全部
FCN-8s(Long 等,2015)	–	–	–	67.2
Dilation10(Yu 和 Koltun,2016)	–	–	–	73.9
CowMix(French 等,2020a)	–	71.0	–	73.4
DST-CBC(Feng 等,2021)	70.7	71.8	–	73.5
Sawatzky 等人(2021)	71.3	72.4	73.9	75.0
Mittal(French 等,2020b)	71.4	–	–	75.6
DeepLabv2(Chen 等,2018)	66.0	68.3	69.8	73.6
AdvSemiSeg(Hung 等,2018)	69.5	72.1	73.8	74.9
本文(自注意)	70.3	72.7	75.9	78.4
本文(自注意 + 谱归一化)	71.8	73.5	76.3	78.9

注：“–”表示对应文献未进行该项实验。



(a) 原图　(b) 真值　(c) 本文(1/8)　(d) 本文(1/4)　(e) 本文(1/2)　(f) 全部

图 3　在 PASCAL VOC 2012 数据集上随机抽取不同比例标记数据获得的分割结果比较

Fig. 3　Comparison of segmentation results obtained by randomly extracting different proportions of labeled data on the PASCAL VOC 2012 dataset((a) original images; (b) ground truth; (c) ours (1/8); (d) ours (1/4); (e) ours (1/2); (f) full)

不同比例的标记数据获得的分割结果的比较。可以看出,当使用随机选取 1/2 的标记数据进行训练时,本文方法具有很好的分割效果。

表 3 列出了 PASCAL VOC 2012 数据集中具有不同比例的标记数据的每个类别的半监督和全监督训练的平均 mIoU 性能结果。其中,Adv 是基线 AdvSemiSeg 模型,SA 是本文方法只使用自注意模型,SA + SN 是本文方法同时使用自注意和谱归一化(模型)。此外,表 2 中最初报告的所有类的平均 mIoU 值都包含在最后一行中。从结果中注意到,提出的自注意模块和谱归一化显著提高了 PASCAL VOC 2012 数据集包含的 21 类图像的分割性能。在分割网络中加入自注意模块可以很好地捕捉到特征图中任意两个像素之间的远程上下文信息,提高模型的特征表示。

表 3　在 PASCAL VOC 2012 数据集上逐类分割性能结果(mIoU)

Table 3　Performance results of class-by-class segmentation on the PASCAL VOC 2012 dataset (mIoU)

类别	1/8 标记数据			1/4 标记数据			1/2 标记数据			全部标记数据		
	Adv	SA	SA + SN	Adv	SA	SA + SN	Adv	SA	SA + SN	Adv	SA	SA + SN
bkg	**0.93**	**0.93**	**0.93**	0.93	0.94	**0.94**	**0.94**	**0.94**	**0.94**	0.94	0.94	**0.95**
aero	0.82	0.85	**0.87**	0.88	0.87	**0.89**	0.88	0.88	**0.89**	0.89	0.89	**0.90**
bike	0.40	0.41	**0.42**	0.40	**0.42**	0.41	0.41	0.42	**0.42**	0.41	0.43	**0.43**
bird	0.82	0.85	**0.85**	0.85	0.86	**0.87**	0.86	**0.87**	0.85	0.87	**0.88**	0.87
boat	0.63	0.67	**0.67**	0.67	0.69	**0.70**	0.67	0.67	0.67	0.67	0.73	**0.75**
bottle	0.71	0.74	**0.76**	0.76	**0.78**	0.74	0.79	0.82	**0.83**	0.81	0.82	**0.82**
bus	0.90	0.88	**0.91**	0.91	0.92	**0.93**	0.91	**0.93**	0.92	0.91	**0.94**	0.93
car	**0.84**	0.82	0.82	0.83	0.84	**0.85**	0.85	**0.88**	0.86	0.85	**0.87**	0.86
cat	0.84	0.86	**0.87**	0.86	0.89	**0.89**	0.87	0.89	**0.89**	0.88	0.90	**0.91**
chair	0.30	0.27	**0.32**	0.31	0.32	**0.35**	0.34	**0.40**	0.39	0.36	0.41	**0.42**
cow	0.72	**0.80**	0.79	0.79	0.76	**0.82**	0.81	0.86	**0.86**	0.83	0.85	**0.87**
table	0.40	0.42	**0.45**	0.45	**0.48**	0.45	0.52	0.59	**0.61**	0.53	**0.59**	0.57
dog	0.76	0.78	**0.80**	0.79	**0.81**	0.80	0.80	0.84	**0.85**	0.82	0.86	**0.86**
horse	0.73	**0.78**	0.74	0.75	0.74	**0.83**	0.80	**0.84**	0.82	0.80	0.83	**0.86**
mbk	**0.79**	0.77	0.76	0.78	0.80	**0.80**	**0.82**	0.81	0.81	0.83	**0.86**	0.85
prsn	0.83	0.83	0.82	0.84	**0.85**	0.84	0.84	0.86	**0.86**	0.85	0.87	**0.90**
plnt	0.52	0.42	**0.52**	**0.56**	0.53	0.55	0.57	0.60	**0.60**	0.59	**0.66**	0.65
sheep	0.75	0.77	**0.82**	0.79	0.78	**0.80**	0.82	0.81	**0.88**	0.83	0.87	**0.89**
sofa	0.43	0.42	**0.45**	**0.47**	0.44	0.44	0.47	**0.48**	0.46	0.49	**0.56**	0.55
train	0.79	0.78	**0.80**	0.79	**0.82**	0.81	0.81	0.84	**0.84**	0.83	**0.88**	0.86
tv	0.71	**0.73**	0.70	0.74	0.72	**0.74**	0.73	0.75	**0.75**	0.74	0.74	**0.77**
mIoU	69.5	70.3	**71.8**	72.1	72.7	**73.5**	73.8	75.9	**76.3**	74.9	78.4	**78.9**

注:加粗字体表示各类别在各标记数据下的最优结果。

此外,在判别器中加入谱归一化处理,有利于进一步训练 GAN 网络。图 4 所示的图像分割结果进一步说明了这一点,将 GT 数据与基线 AdvSemiSeg 模型和本文模型在训练中使用 1/2 标记数据时的分割结果进行对比,该模型在引入自注意和谱归一化后的分割结果在质量上均优于 AdvSemiSeg 模型,尤其是同时引入自注意和谱归一化后的分割结果。由此可以看出,自注意模块在捕获输入图像的全局依赖关系和谱归一化稳定 GAN 方面的有效性。

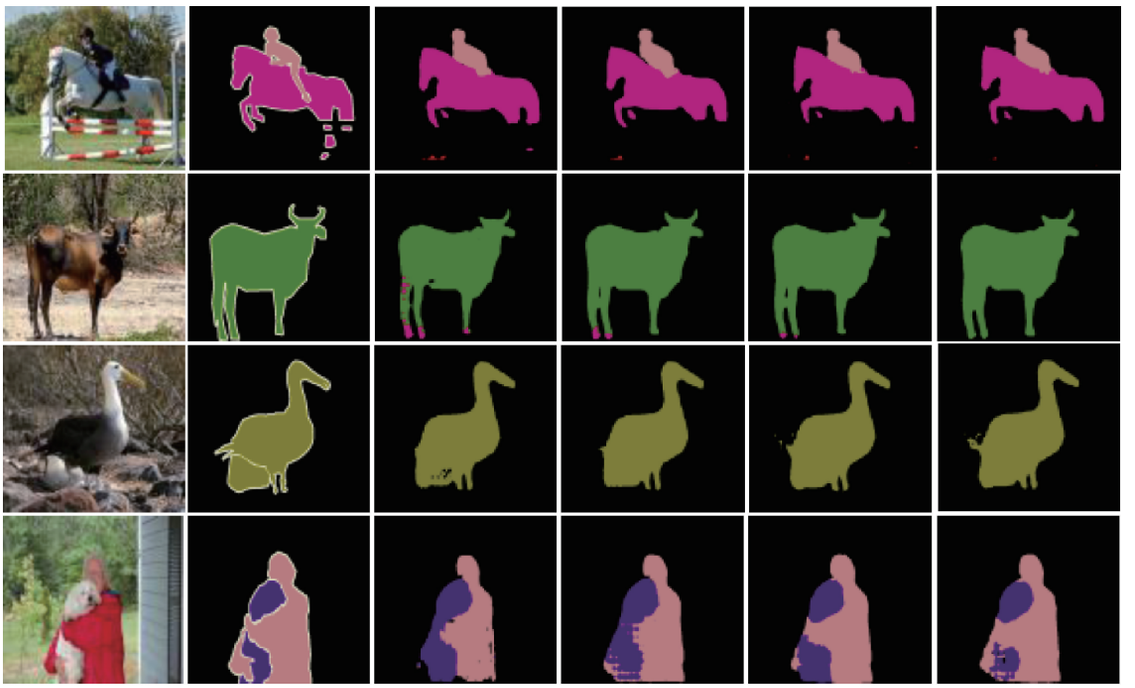


图 4　在 PASCAL VOC 2012 数据集上使用 1/2 标记数据时本文方法获得的定性结果

Fig. 4　Qualitative results obtained by our method when using 1/2 labeled data on the PASCAL VOC 2012 dataset ((a) original images; (b) ground truth; (c) AdvSemiSeg; (d) ours(SN); (e) ours(SA); (f) ours(SN + SA))

以上图像分割结果是基于 DeepLabv2 框架和在 MSCOCO 数据集上预先训练的 ResNet-101 模型。表 4 是在 PASCAL VOC 2012 数据集上使用不同主干架构和不同比例的标记数据获得的图像分割性能结果。

表 4　在 PASCAL VOC 2012 数据集上使用不同主干架构和不同比例的标记数据获得的图像分割性能结果(mIoU)

Table 4　Image segmentation performance results using different backbone architectures and different proportions of labeled data on the PASCAL VOC 2012 dataset(mIoU)

/%

方法	标记数据			
	1/8	1/4	1/2	全部
DeepLabv2	66.0	68.3	69.8	73.6
本文(DeepLabv2)	71.8	73.5	76.3	78.9
DeepLabv3	不稳定	69.4	70.9	75.2
本文(DeepLabv3)	**72.8**	**75.3**	**77.0**	**79.5**

注:加粗字体表示各列最优结果。

从表 4 可以看出,当使用 DeepLabv3 框架时,训练过程在较大比例的标记数据下是相对稳定的,在 1/8 比例的标记数据下是不稳定的。然而在使用大比例标记数据训练时,图像分割性能总比使用 Deep-Labv2 框架得到的效果好。而且通过谱归一化的应用可以很好地缓解 DeepLabv3 框架在使用 1/8 比例标记数据观察到的训练不稳定性。同时,使用 DeepLabv3 主干网络时,本文提出的半监督模型的表现更好。由此认为,基于 AdvSemiSeg 模型的方法对于少量标记样本无效的原因是其判别器网络所要施加的要求。少量的 GT 缺乏有效训练判别器从预测的分割图上区分 GT 所必需的变化,从而阻止了它有效地指导分割网络。相比之下,谱归一化可以最大程度地减少保留类扰动的预测差异,从而有效地在未标记样本之间传播标记。因此,它不会对标记的数据集的大小施加类似的要求。

为验证训练框架中包含不同组件的效果,使用 1/2 标记数据和全部标记数据对本文方法进行消融研究,评估结果如表 5 所示。

从表 5 可以看出,在训练中加入谱归一化可以明显提高模型的图像语义分割性能。此外,加入第 2 个自注意模块(SA2)似乎比第 1 个自注意模块(SA1)对分割性能有更好的结果,尽管这两个模块都确实提高了分割性能。总体而言,本文方法对改

表 5　本文方法在 PASCAL VOC 2012 数据集上的消融研究(mIoU)

Table 5　Ablation study of the proposed method on the PASCAL VOC 2012 dataset(mIoU)

/%

方法	1/2 标记数据	全部标记数据
本文	73.82	74.98
本文 + SN	74.52	76.72
本文 + SA1	75.17	77.69
本文 + SA1 + SA2	75.94	78.10
本文 + SA1 + SN	76.13	78.38
本文 + SA2 + SN	76.20	78.74
本文 + SA1 + SA2 + SN	**76.36**	**78.93**

注:加粗字体表示各列最优结果。

善半监督 GAN 网络的图像语义分割性能非常有效。

另外,本文还考虑了在使用 1/2 标记数据时,超参数 λ_{adv}、λ_{semi} 和 T_{semi} 对性能的影响。式(1)中的参数 λ_{adv} 和 λ_{semi} 是两个权重用于最小化多任务损失函数。式(4)中的参数 T_{semi} 是用于判断像素的预测是否可信的阈值。表 6 显示了将这些参数设置为不同值的效果。按照 Hung 等人(2018)方法将 T_{semi} 的值设置为 0.2,并设置不同的 λ_{adv} 和 λ_{semi},以此评估所提出方法的性能。可以看出,在(0.001,0.1)下获得最佳的 mIoU。同时,为了分析 T_{semi} 的作用,总结了将 T_{semi} 分别设置为 0.15、0.20 和 0.25 时的 mIoU,当 T_{semi} = 0.20 时,可以达到最佳的 mIoU。

表 6　在不同的超参数下 PASCAL VOC 2012 数据集上的分割性能

Table 6　Segmentation performance on the PASCAL VOC 2012 dataset under different hyperparameters

超参数			mIoU/%
λ_{adv}	λ_{semi}	T_{semi}	
0.001 25	0.125	0.2	73.9
0.001 00	0.100	0.2	**76.3**
0.000 75	0.075	0.2	74.7
0.001 00	0.100	0.15	75.6
0.001 00	0.100	0.20	**76.3**
0.001 00	0.100	0.25	74.9

注:加粗字体表示最优结果。

综上所述,通过在 Cityscapes 数据集和 PASCAL VOC 2012 数据集上的实验结果表明,本文在分割网络中利用自注意成功地捕获半监督图像语义分割中的远程上下文依赖关系。该自注意和卷积是互补的,对图像区域之间远程全局依赖进行建模,从而更好地近似原始图像分布。并且该自注意模型在中高层特征图上比在低层特征具有更好的性能。在实验中可以看出,在判别器中引入谱归一化对稳定 GAN 网络的重要性。尤其是当使用少量标记数据时,对于 GAN 网络的性能控制则更为重要。该谱归一化对标识符施加了全局正则化,应用于 GAN 网络时,生成的示例比常规的权重归一化更加多样化。相对于先前的方法获得了更好的可比较的初始分数。因此本文方法比基线模型有较大的优势,并且比其他先进的半监督语义分割方法都有更好的性能。

5　结　论

提出了一种改进的 GAN 框架半监督图像语义分割方法。首先,在分割网络中引入了自注意模块,以有效地考虑输入图像的广泛分离的空间区域之间的关系,从而捕获远程上下文信息。相比于传统方法通过增加卷积核大小或通过多个卷积层捕获这些依赖关系,更加平衡了特征图上各像素之间远程依赖性的建模能力和计算效率。其次,在判别器网络中应用了谱归一化,以增强 GAN 在训练过程中的稳定性,使得生成的样本比传统的权值归一化得到的样本更加多样化。这种方法以更细致的方式使得判别器的参数矩阵满足 Lipschitz 约束。从而使得 GAN 对输入图像的扰动不会有太大的敏感性。在 cityscapes 和 PASCAL VOC 2012 两个数据集上,与当前半监督图像语义分割方法的结果相比,提出的稳定的自注意半监督对抗性学习图像语义分割方法具有更好的性能。另外,通过实验发现,在进行半监督训练时,即使逐渐提高抽取标签数据的比例,对桌子、椅子和沙发的分割效果也没有明显提高。经过分析得出,造成此结果的原因可能是数据集中样本类别的不平衡所致。在 PASCAL VOC 2012 数据集中,针对桌子、椅子和沙发的类样本较少,这使得对这些类的训练更加困难。因此,即使添加标签数据,也很难提高这些类的分割精度。今后的工作将寻求一种比较有效的方法对易训练样本进行限制,对难训练样本进行加权,以此达到难易样本训练的平衡。

参考文献(References)

Arjovsky M and Bottou L. 2017. Towards principled methods for training generative adversarial networks [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1701.04862.pdf

Arjovsky M, Chintala S and Bottou L. 2017. Wasserstein GAN [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1701.07875.pdf

Brock A, Donahue J and Simonyan K. 2019. Large scale GAN training for high fidelity natural image synthesis [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1809.11096.pdf

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848 [DOI: 10.1109/TPAMI.2017.2699184]

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B. 2016. The cityscapes dataset for semantic urban scene understanding//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3213-3223 [DOI: 10.1109/CVPR.2016.350]

Ding H H, Jiang X D, Shuai B, Liu A Q and Wang G. 2018. Context contrasted feature and gated multi-scale aggregation for scene segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2393-2402 [DOI: 10.1109/CVPR.2018.00254]

Everingham M, Van Gool L, Williams C K I, Winn J and Zisserman A. 2010. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 88(2): 303-338 [DOI: 10.1007/s11263-009-0275-4]

Feng Z Y, Zhou Q Y, Cheng G L, Tan X, Shi J P and Ma L Z. 2021. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/2004.08514v1.pdf

French G, Aila T, Laine S, Mackiewicz M and Finlayson G. 2020a. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1906.01916.pdf

French G, Aila T, Laine S, Mackiewicz M and Finlayson G. 2020b. Consistency regularization and CutMix for semi-supervised semantic segmentation [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1906.01916v1.pdf

Geiger A, Lenz P and Urtasun R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 3354-3361 [DOI: 10.1109/CVPR.2012.6248074]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A. 2017. Improved training of wasserstein GANs [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1704.00028.pdf

Hariharan B, Arbeláez P, Bourdev L, Maji S and Malik J. 2011. Semantic contours from inverse detectors//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE: 991-998 [DOI: 10.1109/ICCV.2011.6126343]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hu H, Gu J Y, Zhang Z, Dai J F and Wei Y C. 2018. Relation networks for object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3588-3597 [DOI: 10.1109/CVPR.2018.00378]

Hung W C, Tsai Y H, Liou Y T, Lin Y Y and Yang M H. 2018. Adversarial learning for semi-supervised semantic segmentation [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1802.07934.pdf

Kalluri T, Varma G, Chandraker M and Jawahar C V. 2019. Universal semi-supervised semantic segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5258-5269 [DOI: 10.1109/ICCV.2019.00536]

Karras T, Aila T, Laine S and Lehtinen J. 2018. Progressive growing of GANs for improved quality, stability, and variation [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1710.10196.pdf

Kingma D P and Ba J. 2017. Adam: a method for stochastic optimization [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1412.6980.pdf

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu S T, Zhang J Q, Chen Y X, Liu Y F, Qin Z C and Wan T. 2019. Pixel level data augmentation for semantic image segmentation using generative adversarial networks//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: IEEE: 1902-1906 [DOI: 10.1109/ICASSP.2019.8683590]

Liu X M, Cao J, Fu T Y, Pan Z F, Hu W, Zhang K and Liu J. 2019. Semi-supervised automatic segmentation of layer and fluid region in retinal optical coherence tomography images using adversarial learning. IEEE Access, 7: 3046-3061 [DOI: 10.1109/ACCESS.2018.2889321]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference

on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Luc P, Couprie C, Chintala S and Verbeek J. 2016. Semantic segmentation using adversarial networks [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1611.08408.pdf

Mittal S, Tatarchenko M and Brox T. 2021. Semi-supervised semantic segmentation with high-and low-level consistency. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(4): 1369-1379 [DOI: 10.1109/TPAMI.2019.2960224]

Miyato T, Kataoka T, Koyama M and Yoshida Y. 2018. Spectral normalization for generative adversarial networks [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1802.05957.pdf

Oliver A, Odena A, Raffel C, Cubuk E D and Goodfellow I J. 2019. Realistic evaluation of deep semi-supervised learning algorithms [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1804.09170.pdf

Radford A, Metz L and Chintala S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1511.06434.pdf

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Salimans T and Kingma D P. 2016. Weight normalization: a simple reparameterization to accelerate training of deep neural networks [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1602.07868.pdf

Sawatzky J, Zatsarynna O and Gall J. 2021. Discovering latent classes for semi-supervised semantic segmentation [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1912.12936.pdf

Shuai B, Zuo Z, Wang B and Wang G. 2018. Scene segmentation with dag-recurrent neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6): 1480-1493 [DOI: 10.1109/TPAMI.2017.2712691]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1409.1556.pdf

Souly N, Spampinato C and Shah M. 2017. Semi and weakly supervised semantic segmentation using generative adversarial network [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1703.09695.pdf

Stekovic S, Fraundorfer F and Lepetit V. 2019. S4-Net: geometry-consistent semi-supervised semantic segmentation [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1812.10717.pdf

Sun F D and Li W H. 2019. Saliency guided deep network for weakly-supervised image segmentation. Pattern Recognition Letters, 120: 62-68 [DOI: 10.1016/J.PATREC.2019.01.009]

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I and Fergus R. 2014. Intriguing properties of neural networks [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1312.6199.pdf

Tsai Y H, Shen X H, Lin Z, Sunkavalli K, Lu X and Yang M H. 2017. Deep image harmonization//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2799-2807 [DOI: 10.1109/CVPR.2017.299]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang X L, Girshick R, Gupta A and He K M. 2018. Non-local neural networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7794-7803 [DOI: 10.1109/CVPR.2018.00813]

Yu F and Koltun V. 2016. Multi-scale context aggregation by dilated convolutions [EB/OL]. [2021-03-22]. https://arxiv.org/pdf/1511.07122.pdf

Zhang H, Dana K, Shi J P, Zhang Z Y, Wang X G, Tyagi A and Agrawal A. 2018. Context encoding for semantic segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7151-7160 [DOI: 10.1109/CVPR.2018.00747]

Zhang H, Zhang H, Wang C G and Xie J Y. 2019. Co-occurrent features in semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 548-557 [DOI: 10.1109/CVPR.2019.00064]

Zhao H S, Zhang Y, Liu S, Shi J P, Loy C C, Lin D H and Jia J Y. 2018. PSANet: point-wise spatial attention network for scene parsing//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 270-286 [DOI: 10.1007/978-3-030-01240-3_17]

作者简介

李志欣,1971 年生,男,教授,主要研究方向为图像理解、机器学习与跨媒体计算。E-mail:lizx@gxnu.edu.cn

张佳,男,硕士研究生,主要研究方向为机器学习与语义分割。E-mail:zhangjia51320@163.com

吴璟莉,女,教授,主要研究方向为生物信息学与智能优化算法。E-mail:wjlhappy@gxnu.edu.cn

马慧芳,女,教授,主要研究方向为机器学习与数据挖掘。E-mail:mahuifang@nwnu.edu.cn