JIG
JOURNAL OF IMAGE AND GRAPHICS
主办：中国科学院遥感与数字地球研究所
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

2020
05
VOL.25
ISSN1006-8961
CN11-3758/TB

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第25卷第5期（总第289期）
2020年5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 中国科学院遥感与数字地球研究所
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 顾行发
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@radi.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 Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Gu Xingfa
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@radi.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

2020年5月 第25卷 第5期（总第289期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

多相图像分割Vese-Chan模型连续最大流方法(第0926页)

保持区域平滑的3维动画形状高效编辑(第1019页)

树形结构卷积神经网络优化的城区遥感图像语义分割(第1043页)

封面图片由武汉大学张良培教授团队提供

Volume 25, Number 5
Published May 16, 2020

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Continuous max-flow method for multiphase segmentation Vese-Chan model(P0926)

Efficient 3D animation shape manipulation with region smoothness preservation(P1019)

Semantic segmentation of urban remote sensing image based on optimized tree structure convolutional neural network (P1043)

中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2020)05-0913-13

论文引用格式:Zhang G M, Pan G F and Liu J X. 2020. Domain adaptation for semantic segmentation based on adaption learning rate. Journal of Image and Graphics,25(05):0913-0925(张桂梅,潘国峰,刘建新. 2020. 域自适应城市场景语义分割. 中国图象图形学报,25(05):0913-0925)[DOI:10.11834/jig.190424]

域自适应城市场景语义分割

张桂梅[1],潘国峰[1],刘建新[2]

1. 南昌航空大学计算机视觉研究所,南昌　330063;2. 西华大学机械工程学院,成都　610039

摘　要:**目的**　域自适应分割网(AdaptSegNet)在城市场景语义分割中可获得较好的效果,但是该方法直接采用存在较大域差异(domain gap)的源域数据集GTA(grand theft auto)5与目标域数据集Cityscapes进行对抗训练,并且在网络的不同特征层间的对抗学习中使用固定的学习率,所以分割精度仍有待提高。针对上述问题,提出了一种新的域自适应的城市场景语义分割方法。**方法**　采用SG-GAN(semantic-aware grad-generative adversarial network(GAN))方法对虚拟数据集GTA5进行预处理,生成新的数据集SG-GTA5,其在灰度、结构以及边缘等信息上都更加接近现实场景Cityscapes,并用新生成的数据集代替原来的GTA5数据集作为网络的输入。针对AdaptSegNet加入的固定学习率问题,在网络的不同特征层引入自适应的学习率进行对抗学习,通过该学习率自适应地调整不同特征层的损失值,达到动态更新网络参数的目标。同时,在对抗网络的判别器中增加一层卷积层,以增强网络的判别能力。**结果**　在真实场景数据集Cityscapes上进行验证,并与相关的域自适应分割模型进行对比,结果表明:提出的网络模型能更好地分割出城市交通场景中较复杂的物体,对于sidewalk、wall、pole、car、sky的平均交并比(mean intersection over union, mIoU)分别提高了9.6%、5.9%、4.9%、5.5%、4.8%。**结论**　提出方法降低了源域和目标域数据集之间的域差异,减少了训练过程中的对抗损失值,规避了网络在反向传播训练过程中出现的梯度爆炸问题,从而有效地提高了网络模型的分割精度;同时提出基于该自适应的学习率进一步提升模型的分割性能;在模型的判别器网络中新添加一个卷积层,能学习到图像的更多高层语义信息,有效地缓解了类漂移的问题。

关键词:城市场景;语义分割;生成对抗网络;域自适应;自适应学习率

Domain adaptation for semantic segmentation based on adaption learning rate

Zhang Guimei[1], Pan Guofeng[1], Liu Jianxin[2]

1. Institute of Computer Vision, Nanchang Hangkong University, Nanchang　330063, China;

2. School of Mechanical Engineering, Xihua University, Chengdu　610039, China

Abstract: **Objective** Semantic segmentation is a core computer vision task where one aims to densely assign labels to each pixel in the input image, such as person, car, road, pole, traffic light, or tree. Convolutional neural network-based approaches achieve state-of-the-art performance on various semantic segmentation tasks with applications for autonomous driving, image editing, and video monitoring. Despite such progress, these models often rely on massive amounts of pixel-level labels. However, for a real urban scene task, large amounts of labeled data are unavailable because of the high labor of annotating segmentation ground truth. When the labeled dataset is difficult to obtain, adversarial-training-based methods

收稿日期:2019-08-20;**修回日期**:2019-10-28;**预印本日期**:2019-11-04

基金项目:国家自然科学基金项目(61462065)

Supported by:National Natural Science Foundation of China(61462065)

are preferred. These methods seek to adapt by confusing the domain discriminator with domain alignment standalone from task-specific learning under a separate loss. Another challenge is that a large difference exists between source data and target data in real scenarios. For instance, the distribution of appearance for objects and scenes may vary in different places, and even weather and lighting conditions can change significantly at the same place. In particular, such differences are often called as "domain gaps" and could cause significantly decreased performance. Unsupervised domain adaptation seeks to overcome such problems without target domain labels. Domain adaption aims to bridge the source and target domains by learning domain-invariant feature representations without using target labels. Such efforts have been made by using a deep learning network like AdaptSegNet, which has gained good results in the semantic segmentation of urban scenes. However, the network is trained directly using the synthetic dataset GTA5 and the real urban scene dataset Cityscapes, which exhibit a domain gap in gray, structure, and edge information. The fixed learning rate is employed in the model during the adversarial learning of different feature layers. In sum, segmentation accuracy needs to be improved. **Method** To handle these problems, a new domain adaptation method is proposed for urban scene semantic segmentation. To reduce the domain gap between source and target datasets, knowledge transfer or domain adaption is proposed to close the gap between source and target domains. This work is based on adversarial learning. First, the semantic-aware grad-generative adversarial network (GAN) (SG-GAN) is introduced to pre-process the synthetic dataset of GTA5. As a result, a new dataset SG-GTA5 is generated, which brings the newly dataset SG-GTA5 considerably closer to the urban scene dataset Cityscapes in gray, structure, and edge information. It is also suitable to substitute the original dataset GTA5 in AdaptSegNet. Second, the newly dataset SG-GTA5 is used as input of our network. To further enhance the adapted model and handle the fixed learning rate of AdaptSegNet, a multi-level adversarial network is constructed to effectively perform output space domain adaptation at different feature levels. Third, an adaptive learning rate is introduced in different feature levels of the network. Fourth, the loss value of different levels is adjusted by the proposed adaptive learning rate. Thus, the network's parameters can be updated dynamically. Fifth, a new convolution layer is added into the discriminator of GAN. As a result, the discriminant ability of the network is enhanced. For the discriminator, we use an architecture that uses fully convolutional layers to replace all fully connected layers to retain the spatial information. This architecture is composed of six convolution layers with 4×4 kernel and a stride of 2 for the first four kernels, a stride of 1 for the fifth kernel and channel numbers of 64, 128, 256, 512, 1 024, and 1, respectively. Except for the last layer, each convolution layer is followed by a leaky ReLU parameterized by 0.2. An up-sampling layer is added to the last convolution layer for re-scaling the output to the input size. No batch-normalization layers are used because we jointly train the discriminator with the segmentation network using a small batch size. For the segmentation network, it is essential to build upon a good baseline model to achieve high-quality segmentation results. We adopt the DeepLab-v2 framework with ResNet-101 model pre-trained on ImageNet as our segmentation baseline network. Similar to the recent work on semantic segmentation, we remove the last classification layer and modify the stride of the last two convolution layers from 2 to 1, making the resolution of the output feature maps effectively 1/8 times the input image size. To enlarge the receptive field, we apply dilated convolution layers in conv4 and conv5 layers with a stride of 2 and 4, respectively. After the last layer, we use the atrous spatial pyramid pooling (ASPP) as the final classifier. Finally, the batch normalization (BN) layers are removed because the discriminator network is trained with small batch generator network. Furthermore, we implement our network using the PyTorch toolbox on a single GTX1080Ti GPU with 11 GB memory. **Result** The new model is verified using the Cityscapes dataset. Experimental results demonstrate that the presented model is capable of segmenting more complex targets in the urban traffic scene precisely. The model also performs well against existing state-of-the-art segmentation model in terms of accuracy and visual quality. The segmentation accuracy of sidewalk, wall, pole, car and sky are improved by 9.6%, 5.9%, 4.9%, 5.5%, and 4.8%, respectively. **Conclusion** The effectiveness of the proposed model is validated by using the real urban scene Cityscapes. The segmentation precision is improved through the presented dataset preprocessing scheme on the synthetic dataset of GTA5 by using the SG-GAN model, which makes the newly dataset SG-GTA5 much closer to the urban scene dataset Cityscapes on gray, structure, and edge information. The presented data preprocessing method also reduces the adversarial loss value effectively and avoids gradient explosion during the back propagation process. The network's learning capability is also further strengthened and the model's segmentation precision is improved through the presented adaptive learning rate, which is

used for different adversarial layers to adjust the loss value of each layer. The learning rate can also update network parameters dynamically and optimize the performance of the generator and discriminator network. Finally, the discrimination capability of the proposed model is further improved by adding a new convolution layer in the discriminator, which enables the model to learn high layer semantic information. The domain shift is also alleviated to some extent.

Key words: urban scene; semantic segmentation; generative adversarial network (GAN); domain adaptation; adapt learning rate

0 引　言

语义分割是计算机视觉和医学图像中的重要研究问题之一,其主要任务是对图像中每个像素进行精确分类并加以标注,如将交通场景中的道路、行人、车辆和建筑物等目标进行标注。深度学习(LeCun 等,2015)由于能够学习到高层的语义特征,故其在计算机视觉研究和医学图像领域得到广泛应用(曲仕茹 等,2018)。但其在图像语义分割中存在两大难题:1)语义分割的精度严重依赖于大规模的训练数据集以及它们的精确标注,通常对真实场景的精确标注过程繁琐且需耗费大量的人力物力; 2)分割网络模型的迁移能力较差,即在一个场景训练得到的分割网络模型往往很难泛化到不同的场景或不同的拍摄条件下的同一场景。

针对数据集获取及数据注释所带来的大量人力物力消耗问题,学者们进行了一系列研究。如 Dai 等人(2015)提出的 Boxsup 方法,在减少了对数据依赖的同时,也分别在 Pascal、Pascal Voc 2012 数据集上取得了较好的分割效果。Hong 等人(2015)提出了一种新的弱监督分割模型,在 Pascal Voc 数据集上进行测试,该模型在测试中只用到了少量的带标注信息的图片。与其他弱监督分割方法相比,分割效果得到较好提升。Khoreva 等人(2017)提出适合于语义标注和实例分割任务的弱监督分割方法,实验结果表明当给定精细的边界框输入标签时,经过一轮训练就可以获得较满意的分割效果。Papandreou 等人(2015)提出在网络模型的训练学习过程中引入最大期望(expectation maximization, EM)的方法,可较大地降低图片注释所带来的成本,仅仅利用弱标注信息就可训练得到用于语义分割的 CNN(convolutional neural network),并且取得了不错的分割精度。Pathak 等人(2015)提出了一种具有约束条件的 CNN(constrained CNN, C-CNN)方法,其主要思想是将训练目标转变成线性模型的双凸优化问题,获得了弱监督领域中语义分割的最佳效果。

以上方法都是基于现有数据集的弱监督方法,然而在大多数实际应用中,数据集的获取需要特定的硬件环境与社会环境,高质量的像素级标注通常难以获得,需要大量的人力物力(Cordts 等,2016),因而通过计算机合成虚拟数据集,并自动得到其标注,受到了学者们的热切关注。如 Johnson-Roberson 等人(2017)提出了一种将仿真引擎技术和真实感图像相结合的方法,快速生成带标注信息的虚拟数据集。Richter 等人(2017)提出了包含多个视觉任务的虚拟数据集及其标注,大大减轻了手工标记真实数据的繁琐工作量。因此通过自动生成虚拟数据集及其对应的标注,是解决人工标注数据集费时费力的有效途径。

但是,在一个数据集(称为源域)训练得到的分割网络模型,应用到其他数据集(称为目标域)测试时很难获得较好的效果,原因是来自源域的数据集和来自目标域的数据之间往往存在较大的域差异(domain gap),如 Tsai 等人(2018)方法中合成的虚拟数据集 GTA5 和真实交通场景数据集 Cityscapes。合成图像和真实图像间的灰度、结构和边缘等信息存在差异,所以使用合成的虚拟数据集训练学习得到的模型很难泛化到真实的目标数据集。针对该问题,Ganin 等人(2015)提出域自适应的分割方法,其训练数据和测试数据为具有不同特征信息分布的同一个场景,用该方法在标准数据集上进行测试,实验结果表明其达到了目前最好的域适应性能。Long 等人(2015)针对域差异越大,高层的特征可迁移性越下降的问题,提出了深度自适应网(deep adaptation network, DAN),该网络可将深度卷积神经网络(deep CNN, DCNN)拓宽到域自适应的语义分割中。Zhang 等人(2017)和 Hoffman 等人(2017)在特征空间中,通过对抗性学习对语义分割进行了像素级域自适应的研究。Chen 等人(2017)提出一种弱

监督学习方法,以适应不同城市的道路场景分割。该方法不需要采集大量感兴趣城市的标注图像,而是对分割器进行训练或微调,可以在不需要任何用户注释或交互的情况下,实现使用预训练的分割器对该城市自适应学习和分割。传统的对抗网络进行域自适应处理后,虽然源域和目标域的整体特征分布更加接近,域差异有所减少,但也可能使原来已经接近的某些类别反而差异更大。Luo 等人(2019)针对该问题,提出使用协同训练的方式获知各个类别的接近程度,协同训练两个分类器,当两个分类器的结果不同时,可以理解为该类别的特征没有靠近,从而导致两个分类器的结果不同,此时再对没有靠近的类别赋予较大的对抗损失。Hoffman 等人(2016)针对不同拍摄条件下的城市交通场景,提出一种新的基于域自适应的分割模型,并在多个大型的交通场景数据集上进行了测试,都取得了良好的分割结果。但是以上分割方法的分割预测图均是在深层特征图进行上采样,没有利用浅层的特征图,所以分割精度不够理想。Chen 等人(2017)构造了一个多层次的对抗性网络,有效地实现了不同特征层的域自适应,该方法对合成数据集和真实数据集的不同输出特征采用了对抗性学习,并用多个实验证明,提出的模型在分割精度上超越了当时同类的其他方法。但是 Chen 等人(2017)的自适应分割模型是直接对源域和目标域进行对抗训练,然而源域数据集 GTA5 和目标域数据集 Cityscapes 之间存在较大的域差异,再则此模型的对抗训练学习是在不同特征层中引入固定的学习率,从而使得分割结果仍有待提升。

本文在 Chen 等人(2017)的基础上进行了以下改进:

1)采用语义感知对抗学习网络(semantic-aware grad-GAN,SG-GAN)(Li 等,2018)对合成数据集进行风格转换,使新得到的虚拟数据集在灰度、结构和边缘等信息上更接近真实场景数据集,有效地减少了源域与目标域之间的差异,避免模型在反向训练过程中出现梯度爆炸,从而提高了分割精度;

2)构造多层次的对抗性网络,提出在网络的不同特征层采用自适应的学习率进行对抗训练学习,该学习率可以自适应地调整不同特征层的损失值,从而动态地更新模型网络的参数,增强了网络的学习能力,进一步提升了提出模型的分割性能;

3)在模型的判别器中新增了一个卷积层,从而增强了模型的学习能力,能学习到图像的更多高层语义信息,提升了模型的判别能力,有效地缓解了类漂移的问题。

1 基本理论

1.1 生成对抗网络

弱监督学习不需要大量带标注信息的样本,但是准确率上往往达不到效果。研究者们欲提高弱监督学习的精度,减少对监督学习时中带标注信息的样本的依赖。生成对抗网络(generative adversarial network,GAN)(Goodfellow 等,2014)由于能生成与真实数据集类似的新样本数据,不需要大量带标注信息的样本,所以广泛应用于图像处理的各个领域。

GAN 的基本模型:设 z 为随机噪声,x 为真实数据,生成器网络和判别器网络可以分别用 G 和 D 表示,D 可以看做一个二分类器,采用交叉熵表示为

$$\min_G \max_D V(D,G) = E_{x\sim P_{\text{data}}(x)}[\log D(x)] + E_{z\sim P_z(z)}[\log(1-D(G(z)))] \tag{1}$$

式中,第 1 项 $\log D(x)$ 表示判别器对真实数据的判断,第 2 项 $\log(1-D(G(z)))$ 表示判别器对噪声的判断,E 表示期望,$P_{\text{data}}(x)$ 表示真实数据的分布,$P_z(z)$ 表示噪声的分布。通过这样一个极大极小博弈,循环交替优化 G 和 D 训练所需要的生成器网络和判别器网络。

1.2 SG-GAN

传统的 CycleGAN(Zhu 等,2017)只能将目标域图像的整体灰度信息和纹理结构转移到源域图像上,而忽略了每个语义、每个类别的关键特征,从而使得合成的图像较模糊和失真,导致类感染。针对该问题,Li 等人(2018)在 CycleGAN 的基础上提出了梯度敏感损失函数 L-grad 以进一步优化生成网络,并提出了具有语义感知能力的判别网络,即语义感知的梯度生成对抗网络。用梯度敏感损失函数优化生成网络的目的是,无论每个语义类别的纹理如何变化,在语义类别的边界处都应该存在一些可区分的视觉差异。SG-GAN 应用边缘检测中的 sobel 算子,将 sobel 算子与图像进行卷积操作后,可以提取出图像的边缘轮廓信息。此外传统的判别网络对输入图像判别时,是对整幅图像整体判别,但理想情况是判别网络能够对一幅图像中的每个语义类别进

行判别。基于此，提出了语义感知的判别网络，它能够对图像的每个语义类别进行判别而不仅仅是对整幅图像进行判别。故 SG-GAN 能将目标域图像的灰度、结构和边缘等特征信息转移到源域的图像上，即更好地缩小源域数据集与目标域数据集之间的差异。

1.3　DeepLab-v2 网络模型

基于传统 DCNN 的语义分割方法，主要存在以下两个问题：1）不断的降采样操作使得网络的输出特征图的分辨率下降；2）连续的池化操作和卷积操作难以保证输入图像空间位置的不变性。针对这些问题，Chen 等人（2018）提出了 DeepLab 系列网络结构，针对降采样引起的分辨率下降问题，提出采用增加空洞数的卷积核；针对空间位置的不变性，提出在分割图的最后加入条件随机场以对输出的分割结果做进一步优化。

DeepLab-v2 在 DeepLab-v1 的基础上新增了空洞卷积空间金字塔池化（atrous spatial pyramid pooling，ASPP），采用不同的空洞率对图像卷积并融合。ASPP 先用不同的空洞卷积进行并行采样，再将采样的结果融合在一起，可以提升分割精度，该结构对不同尺度大小的物体仍能精确分割。其次，DeepLab-v2 使用空洞卷积的密集深度卷积神经网络，在多个标准的数据集上测试均得到较好的性能。

2　本文模型

2.1　本文的模型框架

本文模型框架如图 1 所示。输入来自源域的图像 $\boldsymbol{I}_{\mathrm{S}}$ 和来自目标域的图像 $\boldsymbol{I}_{\mathrm{T}}$，首先将源图像 $\boldsymbol{I}_{\mathrm{S}}$ 进行前向计算，通过分割网络得到特征 $\boldsymbol{Y}_{\mathrm{S}}$，并用以优化生成器网络 G；然后用优化得到的网络预测目标图像 $\boldsymbol{I}_{\mathrm{T}}$，得到分割后的输出结果 $\boldsymbol{Y}_{\mathrm{T}}$。为了使源图像的特征 $\boldsymbol{Y}_{\mathrm{S}}$ 和目标图像的特征 $\boldsymbol{Y}_{\mathrm{T}}$ 接近，使用 $\boldsymbol{Y}_{\mathrm{S}}$ 和 $\boldsymbol{Y}_{\mathrm{T}}$ 作为判别器 D 的输入，以判断输入是来自源域还是目标域。通过目标预测中的对抗损失，网络将梯度从 D 反向传播到 G，这将引导 G 使目标域的图像分割结果接近于源域的分割结果。图中，T_{S} 表示源数据集的 ground truth（标注），L_D 表示判别器网络的判别损失。L_{adv} 表示判别器网络的对抗损失，L_{seg} 表示源域的 ground truth 与分割预测图的交叉熵损失。

2.2　基于 SG-GAN 对源域数据集进行风格转换

Chen 等人（2017）提出的基于域自适应的对抗

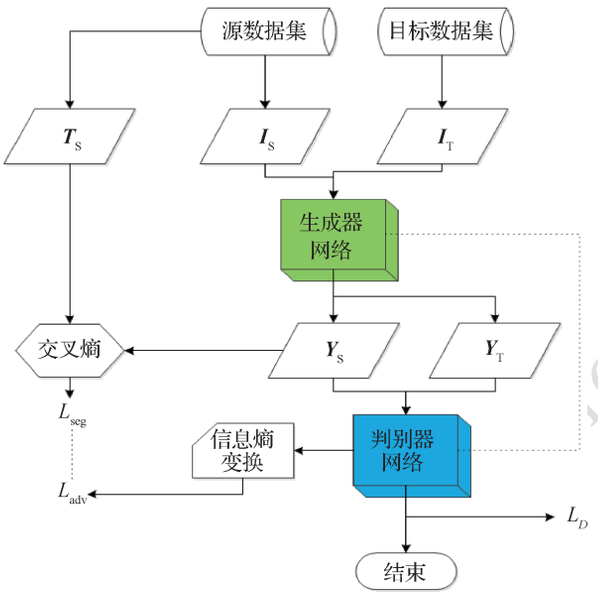


图 1　本文模型框架

Fig. 1　Framework of proposed model

性学习模型中，直接使用合成的虚拟数据集 GTA5 作为源域数据集，但是由于源域的图像和目标域的图像之间在灰度、结构和各类别的边缘信息等方面均存在较大的差异，从而易导致类漂移和类感染。针对该问题，采用 Li 等人（2018）的 SG-GAN 方法对虚拟数据集 GTA5 进行风格转换，得到新的数据集 SG-GTA5，新数据集的特征如灰度、结构和边缘等信息与目标域数据集的特征更贴近，并用新的 SG-GTA5 数据集代替 AdaptSegNet 中的源数据集 GTA5。

GTA5 数据集根据洛杉矶的视频游戏制作而成，包含 24 966 幅视频帧图像，从该视频帧图像中进行稀疏采样得到 5 000 幅图像，将其作为网络模型的源域数据集，Cityscapes 中的图像作为目标域数据集，将源数据集和目标数据集输入到 SG-GAN 进行对抗训练，最后得到优化的风格转换模型。再将 GTA5 数据集的所有图像输入到该训练好的转换模型中，从而得到与目标域数据集在灰度、结构和边缘等信息上都更加接近的新数据集，称之为 SG-GTA5 数据集。并用 SG-GTA5 代替原始的 GTA5 数据集作为网络模型的源域数据集，通过这种方法使得源域与目标域的差异得到有效减少，从而避免网络模型在反向训练过程中出现梯度爆炸。

为了验证新的数据集转换方法的有效性，本文分别将 GTA5 中的图像、SG-GTA5 中的图像与真实

场景 Cityscapes 中的图像进行相似度比较。采用视觉效果和定量的相似性进行分析和比较,其中定量的指标分别用结构相似性度量(structural similarity index,SSIM)和颜色直方图分布来评价。SSIM 分别从亮度、对比度、结构 3 个方面度量图像相似性,取值范围为 [0,1],值越大,表示两个图像越相似。颜色直方图能够描述图像中颜色的全局分布,该方法能够判断两幅图像的颜色相似度。

本次实验选用了 3 组图像,如图 2 所示。从图 2 可以看出,与图 2(a)相比,图 2(b)的道路、天空、建筑物和植被等在灰度、结构和边缘分布上与图 2(c)更接近。故视觉效果表明转换后的新数据集在灰度、结构和边缘等信息与目标域 Cityscapes 数据集更接近。

图 2 视觉效果比较

Fig. 2 Visual effect comparsion((a) GTA5; (b) SG-GTA5; (c) Cityscapes)

为进一步验证新的模型转换方法的有效性,进行定量分析,实验结果如表 1 所示。表 1 中 $S_{a\text{-}c}$ 表示源数据集 GTA5 与真实数据集 Cityscapes 的相似性度量、$S_{b\text{-}c}$ 表示采用 SG-GAN 方法生成的新数据集与真实数据集 Cityscapes 的相似性度量,即 $S_{a\text{-}c}$、$S_{b\text{-}c}$ 分别表示图 2(a)与图 2(c)、图 2(b)与图 2(c)相似度度量。根据表 1 可以发现对于图像Ⅰ,本文合成数据集 SG-GTA5 与真实数据集 Cityscapes 之间的 SSIM,比源数据集 GAT5 与 Cityscapes 之间的 SSIM 提高了 0.257 2,图像Ⅱ和Ⅲ的 SSIM 则分别提高了 0.432 6、0.384 9。而在颜色直方图上,对于各图像Ⅰ、Ⅱ和Ⅲ,本文合成数据集 SG-GTA5 与真实数据集 Cityscapes 之间的颜色直方图相比较于 GTA5 与 Cityscapes 之间的颜色直方图也分别提高了 0.271 7、0.357 9、0.498 4。因此定量实验结果进一步表明,本文得到的新数据集 SG-GTA5 与真实场景 Cityscapes 之间的相似度要大于源数据集 GTA5 与 Cityscapes 的相似度。

表 1 相似度比较

Table 1 Simialrity comparsion

编号	相比较的数据集	SSIM	颜色直方图
Ⅰ	$S_{a\text{-}c}$	0.098 0	0.260 0
	$S_{b\text{-}c}$	**0.355 2**	**0.531 7**
Ⅱ	$S_{a\text{-}c}$	0.135 8	0.254 6
	$S_{b\text{-}c}$	**0.568 4**	**0.612 5**
Ⅲ	$S_{a\text{-}c}$	0.079 8	0.148 7
	$S_{b\text{-}c}$	**0.464 7**	**0.647 1**

注:加粗字体表示每类图像的相似度最优值。

2.3　网络模型的结构

提出的模型包含两个网络，分别为生成器网络 G 和判别器网络 D_i 。

对于生成器网络，采用与 AdaptSegNet 类似的框架，即使用 DeepLab-v2 框架作为分割网络。首先，去掉最后一个分类层，并将 DeepLab-v2 中最后两个卷积层的步长从 2 改为 1。同时为了扩大感受野，分别在第 4 卷积层和第 5 卷积层采用空洞卷积，其空洞数分别取 2 和 4，并使用 ASPP 来代替标准卷积作为最终的分类器。最后，采用一个上采样层 Softmax 来匹配输入图像的大小。另外，由于使用小批量的生成器网络训练判别器网络，故本文不再使用 BN(batch normalization)层。

对于判别器网络，为了更好地保留空间信息，采用全卷积层代替传统的全连接层。不同于 AdaptSegNet 中的判别器网络总共由 5 个卷积层组成，本文判别器网络则有 6 个卷积层，前 5 个的卷积核大小均为 4 ×4，但前 4 个卷积层的步长为 2，第 5 个卷积层的步长为 1。6 个卷积层的通道数分别为(64，128，256，512，1 024，1)。前 5 个卷积层后均连接 1 个激活函数 Leaky ReLU，定义为

$$y_i = \begin{cases} x_i & x_i \geqslant 0 \\ ax_i & x_i < 0 \end{cases}$$

式中，a 为(0，1)之间的调整参数，经过实验取 $a = 0.2$ 。

2.4　多个特征层的域自适应

AdaptSegNet 在网络模型的不同特征层进行对抗训练学习，同时在不同特征层的学习中加入固定的学习率，实际上，随着网络训练次数的增加，对于模型输出的特征应该赋予不同的权因子。针对该问题，构造了一种自适应的学习率函数，并将其引入到网络模型第 4 层和第 5 层的对抗训练中，这种自适应的学习率可以动态地调整第 4 特征层和第 5 特征层对抗训练时的损失值，从而动态地更新网络模型的参数。

对于生成器网络，即分割网络，第 4、5 层的损失函数定义为

$$L(\boldsymbol{I}_S,\boldsymbol{I}_T) = [(1-c_i)\sum_{i=0}^{1}\lambda_{seg}^{i}L_{seg}^{i}(\boldsymbol{I}_S) + c_i] + \sum_{i=0}^{1} b_i L_{adv}^{i}(\boldsymbol{I}_T) \tag{2}$$

式中，L_{seg} 是源域数据集中的标注与得到的分割预测间的交叉熵损失，L_{adv} 是对抗损失，c_i 是第 4、5 特征层所对应的学习率，$i = 0$ ，$i = 1$ 分别对应网络的第 4、5 层，c_i 的定义为

$$c_i = k \cdot \left(1 - \frac{j}{n}\right)^p \tag{3}$$

式中，n 为总的训练次数，j 为第 j 次训练，k 为生成器的基本学习率，按多次实验经验取 $k = 2 \times 10^{-4}$ ，p 是固定参数，通过多次实验取 $p = 0.9$ 。式(2)中的 b_i 指对抗损失 L_{adv} 在每次训练网络的第 4 层、第 5 层对应的学习率，计算公式为

$$b_i = \begin{cases} k\left(1 - \frac{j}{n}\right)^p & i = 0 \\ 50k\left(1 - \frac{j}{n}\right)^p & i = 1 \end{cases} \tag{4}$$

基于式(2)优化生成网络模型的最大最小准则，即

$$\max_D \min_G L(\boldsymbol{I}_S,\boldsymbol{I}_T) \tag{5}$$

生成器网络的训练目标：1) 使得对源域图像的分割损失函数达到最小；2) 使得目标域的分割图尽可能地接近源域分割图。

判别器网络的损失函数为

$$L_d^i(\boldsymbol{P}) = (1-b_i)\cdot[-\sum_{h,w}(1-z)\log(D(\boldsymbol{P})^{(h,w,0)}) + z\log(D(\boldsymbol{P})^{(h,w,1)})] + b_i \tag{6}$$

式中，$\boldsymbol{P}$ 表示生成器网络的输出特征图，h，w 分别表示输出特征图的高和宽。$z = 0$ 表示输入来自于源域数据集，$z = 1$ 表示输入来自于目标域数据集。b_i 指判别器每次训练所对应的学习率。

判别器的目标是要判别输入的分割图是来自源域还是目标域数据集。

2.5　网络训练

利用少量带有标注的虚拟数据集和未带标注的目标域数据集训练本文网络模型。首先，将带有标注的虚拟数据集导入到分割网络得到概率得分 P_S，并通过式(2)中的 L_{seg} 来优化分割网络，再将未带标注的目标域数据集输入到分割网络得到概率得分 P_T；然后将 P_S 和 P_T 输入到判别器网络，并通过式(6)中的 L_d^i 优化判别网络，以提升判别网路的鉴别能力。最后，将目标域预测分割图输入到判别网络得到对抗损失 L_{adv}，输入到生成器网络中并优化生成器网络。

另外，采用随机梯度下降法(stochastic gradient descent，SGD)优化生成器网络，采用自适应 Adam

方法来优化判别器网络。

3 实验及结果分析

实验的深度学习算法框架为 PyTorch 0.3.1.post2，网络模型的训练和测试均采用 PyTorch 完成。

本文在现实场景数据集 Cityscapes 上对提出的模型进行验证，采用平均交并比（mean intersection over union，mIoU）作为性能测试的评价指标。

为了验证提出算法的有效性，选择 Cityscapes 验证集中的 500 幅带标注信息的图像作为模型的验证数据集，SG-GTA5 中的图像及其标注作为源域数据集，Cityscapes 中的图像（不使用标注信息）作为目标域数据集。本文实验从两个方面进行：1）分别分析转换源域数据集、采用自适应学习率和在判别网络增加一层卷积层对本文分割网络模型的影响；2）将提出的模型与主流的语义分割算法模型进行比较。

实验 1　为了验证本文的 3 个改进点对分割效果的影响，采用了控制变量法对网络模型进行验证，具体如下：1）仅将本文得到的 SG-GTA5 代替源数据集 GTA5，其他两个保持不变，训练模型并分析其分割结果，本文将其称为 SG-AdptSeg；2）仅采用自适应惩罚因子，其他保持不变，训练模型并分析其分割结果，本文将其称为 L-AdaptSeg；3）仅在判别网络最后一层添加卷积层，其他保持不变，训练模型并分析其分割结果，本文将其称为 One-AdaptSeg；4）综合上述 3 个变化，即源数据集替换为新得到的 SG-GTA5、采用自适应惩罚因子以及判别网络最后一层添加 1×1 的卷积层，训练模型并分析分割效果，本文将其缩写为 L-SG-One-AdaptSeg，并与 AdaptSegNet 的方法进行了对比。如表 2 所示。

表 2　不同策略的分割结果比较

Table 2　Segmentation results comparsion of different strategies

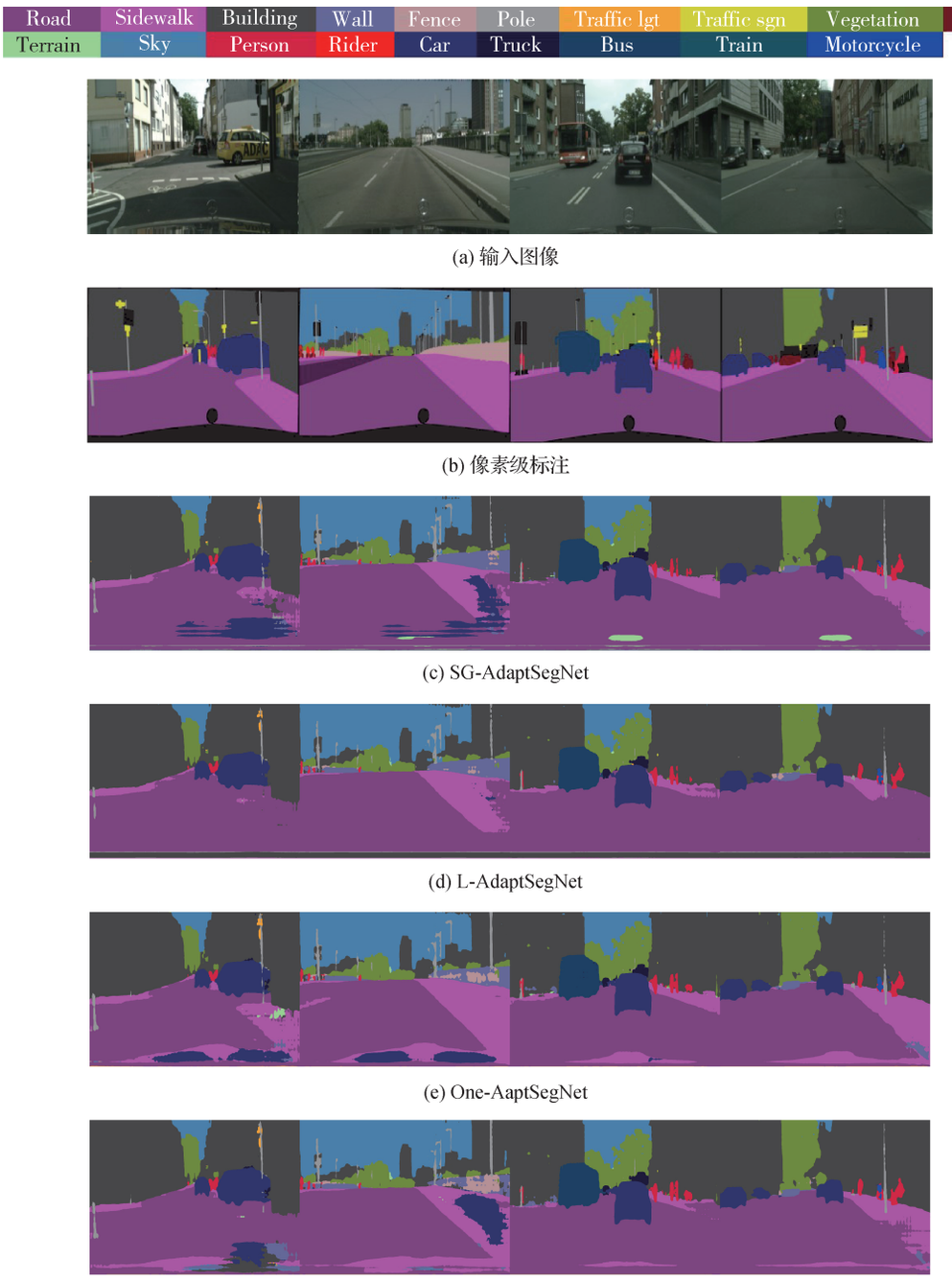
/%

类别	AdaptSegNet（multi-level）（Chen 等，2017）	L-AdaptSeg	SG-AdptSeg	One-AdaptSeg	L-SG-One-AdaptSeg（本文）
Road	86.5	87.4	87.8	88.2	**88.4**
Sidewalk	36.0	39.9	45.5	35.7	**46.6**
Building	79.9	80.1	80.6	81.6	**82.6**
Wall	23.4	28.1	28.4	26.5	**29.3**
Fence	23.1	17.4	19.3	18.2	**23.3**
Pole	23.9	26.6	29.6	30.4	**30.7**
Light	35.2	36.1	34.5	34.3	**36.2**
Sign	14.8	15.3	15.2	13.4	**15.4**
Vegetation	83.4	83.4	83.8	83.6	**83.8**
Terrain	33.3	33.5	33.5	33.3	**33.6**
Sky	75.6	79.5	79.0	80.6	**80.7**
Person	**58.5**	51.4	48.2	55.0	55.1
Rider	27.6	26.1	24.8	27.6	**27.8**
Car	73.7	81.2	73.9	81.4	**81.8**
Truck	32.5	33.5	31.2	26.9	**33.8**
Bus	**35.4**	30.0	31.1	31.2	31.4
Train	**3.9**	3.3	2.4	3.2	3.8
Mbike	**30.1**	26.8	26.2	26.8	26.9
Bike	**28.1**	16.2	5.6	5.4	8.9
mIoU	42.4	43.0	42.3	41.97	**43.2**

注：加粗字体表示每行最优值。

表 2 统计了本文 3 种策略的分割结果，以及 Chen 等人（2017）方法基于像素级中的 multi-level 的分割结果。可以看出，本文模型在结合了 3 种改进策略后的分割精度比单个改进策略均有提高，并且大多数类别的分割精度均比 AdaptSegNet 有较大提升，特别是对于较大的目标，如 Road、Sidewalk、Building、wall、Pole、Vegetation、Terrain、Sky、Bus 提升得较多。这是因为本文采用了 SG-GAN 对 GTA5 进行了特征转换，有效地缩小了源域与目标域之间的差异，这使得对抗损失值得到有效降低，从而提高了分割了精度；再则通过引入自适应的惩罚因子，该因子能够自适应地调整不同特征层的损失值，进而动态更新网络参数。此外，本文在对抗网络的判别器中增加了一层卷积层来提高网络的判别能力。

定性的视觉分割效果如图 3 所示。分别使用不同颜色示意出 Cityscapes 中 19 个不同的类别，如



(a) 输入图像

(b) 像素级标注

(c) SG-AdaptSegNet

(d) L-AdaptSegNet

(e) One-AaptSegNet

(f) L-SG-One-AdaptSegNet(本文)

图 3　不同策略的分割效果图

Fig. 3　Segmentation results of different strategies ((a) input images; (b) ground truth; (c) SG-AdaptSegNet; (d) L-AdaptSegNet; (e) One-AaptSegNet; (f) L-SG-One-AdaptSegNet(ours))

图 3(a)所示。

图 3(a)选择的图像样本为城市场景中的交叉路口,L-AdaptSegNet 与 SG-AdaptSegNet 都将 Sidewalk 分割出一部分,One-AdaptSegNet 在汽车与电线杆处的 Sidewalk 分割丢失,而 L-SG-One-AdaptSegNet 将 Sidewalk 基本上完整的分割出来;图 3(b)选择的图像样本为城市交通场景中的主干交通道路情况,L-AdaptSegNet 将 Sidewalk 较好的分割出来但存在缺失,SG-AdaptSegNet 和 One-AdaptSegNet 在 Sidewalk 分割结果中出现了与 car 的类间感染导致分割效果不理想,而 L-SG-One-AdaptSegNet 将 sidewalk 较好的分割出来;图 3(c)和图 3(d)选择的图像样本为城市交通场景中的子干交通道路场景,L-AdaptSegNet将 Sidewalk 分割出来但存在严重缺失,SG-AdaptSegNet 和 One-AdaptSegNet 在 road 分割结果中出现了与 pole 的类间感染导致分割效果不理想,而 L-SG-One-AdaptSegNet 将 Sidewalk 与 Road 较好地分割出来。该实验结果表明 L-AdaptSegNet 能较好地缓解类间感染,SG-AdaptSegNet 和 One-AdaptSegNet 能够较好地提取局部信息,而 L-SG-One-AdaptSegNet 将两者优点较好地进行结合,故分割效果整体得到有效提升。

实验 2　为了进一步验证本文算法的有效性,分别将本文模型与目前几种主流的语义分割算法模型(Hoffman 等人(2016)、Zhang 等人(2017)、Hoffman 等人(2017)、Chen 等人(2017)、Luo 等人(2019))进行了对比。Hoffman 等人(2017)和 Chen 等人(2017)两种方法均分别比较了基于特征级与像素级的域自适应效果,定量的实验结果如表 3 所示。另外还将本文的模型得到的分割图与 Chen

表 3　各种分割算法的分割精度对比

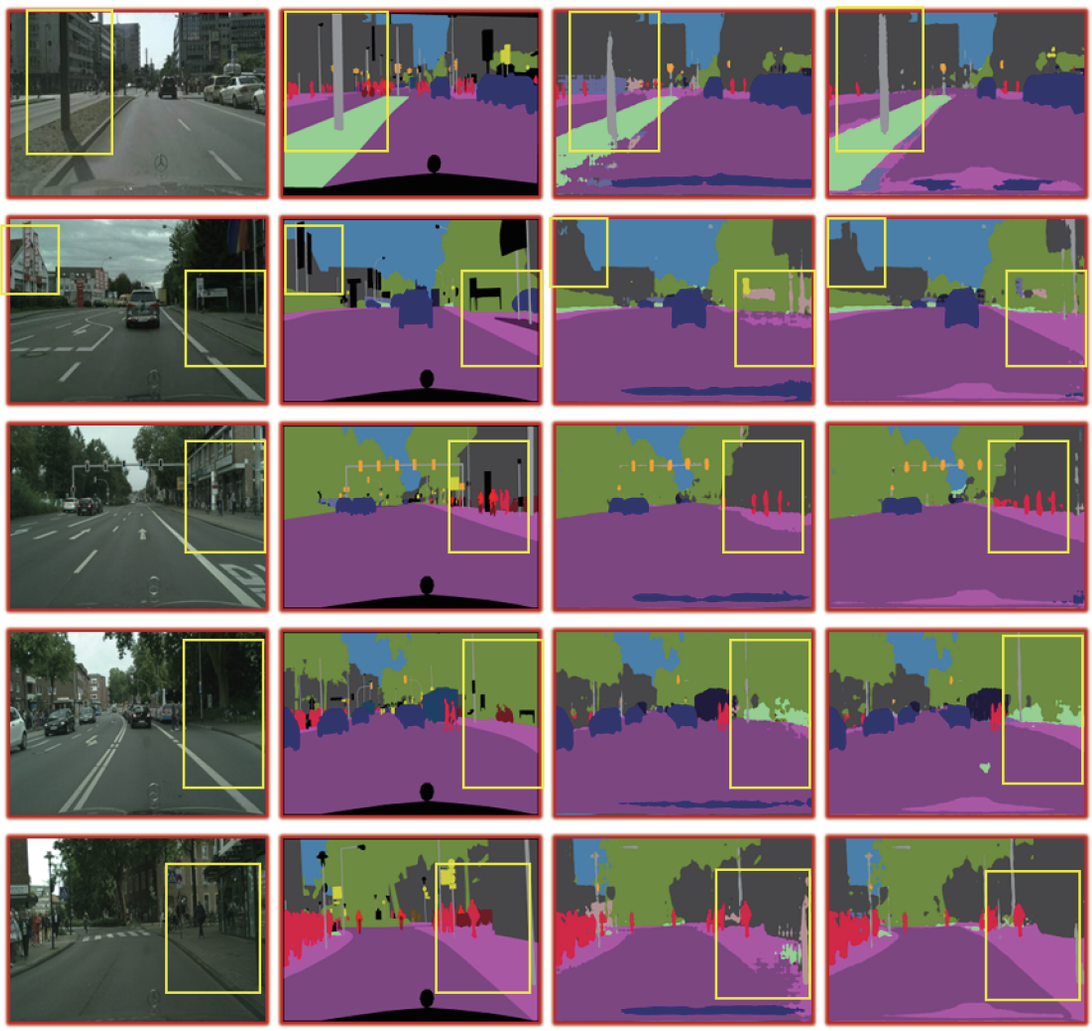
Fig. 3　Segmentation accuracy comparsion of different methods

/%

类别	FCNs in the Wild (Hoffman 等,2016)	CDA (Zhang 等, 2017)	CyCADA (feature) (Hoffman 等,2017)	CyCADA(pixel) (Hoffman 等,2017)	AdaptSegNet (feature) (Chen 等, 2017)	AdaptSegNet (multi-level) (Chen 等, 2017)	CLAN (Luo 等, 2019)	本文
Sidewalk	32.4	22.0	30.7	38.3	27.6	36.0	27.1	**45.6**
Road	70.4	74.9	85.6	83.5	83.7	86.5	87.0	**88.4**
Building	62.1	71.1	74.7	76.4	75.5	79.9	79.6	**82.6**
Pole	10.9	8.4	17.6	22.2	27.4	23.9	28.3	**30.4**
Wall	14.9	6.0	14.4	20.6	20.3	23.4	27.3	**29.3**
Sky	64.6	66.5	69.9	65.7	70.1	75.6	74.2	**80.7**
Car	70.4	55.2	72.3	74.6	72.9	73.7	76.2	**81.8**
Fence	5.4	11.9	13.0	16.5	19.9	**23.3**	**23.3**	23.1
Sign	2.7	11.1	5.8	21.9	**27.4**	14.8	24.2	15.4
Light	14.2	16.3	13.7	26.2	28.3	35.2	35.5	**36.2**
Veg	79.2	75.7	74.6	80.4	79.0	83.4	83.6	**83.8**
Terrain	21.3	13.3	15.8	28.7	28.4	33.3	27.4	**33.6**
Rider	4.2	9.3	3.5	4.2	20.2	27.6	**28.0**	27.8
Truck	8.0	18.8	16.0	16.0	22.5	32.5	33.1	**33.8**
Person	44.1	38.0	38.2	49.4	55.1	58.5	**58.6**	55.1
Train	0.0	0.0	0.1	2.0	**8.3**	3.9	6.7	3.8
Bike	0.0	14.6	0.0	0.0	23.0	28.1	**31.4**	8.9
Bus	7.3	18.9	5.0	26.6	35.7	35.4	**36.7**	31.4
Mbike	3.5	16.8	3.6	8.0	20.6	30.1	**31.9**	26.9
mIoU	27.1	28.9	29.2	34.8	39.3	42.4	**43.2**	**43.2**

注:加粗的字体为每行最优值。

等人(2017)得到的分割结果进行了视觉效果对比，如图 4 所示。在图 4 中，用不同的颜色表示场景中 19 个不同的类别目标。从图 4 可以发现，本文模型的分割性能更优。将着重对比分析方框圈出的地方，如图 4 第 1 行，提出的模型对 Sidewalk、Road、Terrain 等类别的分割都更加接近其对应的 ground truth；如图 4 第 2 行，提出的模型对 Sidewalk、Road 和 Building 都分割更准确，尤其是左上角的 Building 类，分割效果更好；同理，图 4 的第 3 行、第 4 行和第 5 行，提出的模型对 Sidewalk、Road、Pole 的分割效果仍然比 AdaptSegNet 更好。这是因为 AdaptSegNet 在上述类别的分割中受到 Terrain 类别的干扰较大，本文提出模型分割效果提升的主要原因在于，采用了学习率自适应方法可以动态调整不同特征层的损失函数，另外，本文采用 SG-GAN 方法合成具有目标域风格的源域数据也是提升分割效果的主要原因。



(a) 输入图像　(b) 像素级标注　(c) AdaptSegNet　(d) 本文

图 4　分割视觉效果

Fig. 4　Segmentation results ((a) input images; (b) ground truth; (c) AdaptSegNet; (d) ours))

从表 3 可以看出，提出模型对较大目标的分割精度提升较好，如 Sidewalk、Road、Building 等分割精度都有较好的提升，与 Chen 等人(2017)的分割模型 AdaptSegNet 比较，提出模型对典型类别的分割精度都有一定的提升，如 Sidewalk、Road、Building、Pole、Wall、Sky 和 Car 这些类别上的分割精度分别提高了 9.6%、1.9%、2.7%、6.5%、5.9%、5.1% 和 8.1%；但对于小目标如 Sign，Truck 等，其分割效果提高较小。提升效果可归功于：一方面，本文生成新的源域数据集 SG-GTA5 在灰度、结构和边缘等信息熵都与目标域数据集 Cityscapes 更加接近，缩小了域差异，提高了网络模型的分割效果；另一方面，本

文自适应地调整不同特征层的损失函数,优化了网络性能,增强了网络的学习能力,进一步提升了模型分割效果。此外,提出模型在判别器网络的最后新增了一个卷积层,使得判别网络的判别能力得到增强。但是由于本文网络模型的判别器感受野比较大,导致对于细长或者复杂目标的信息易丢失,再则新的 SG-GTA5 数据集中较复杂、小目标类别包含得较少,如自行车和路灯等,所以本文模型对复杂结构与细长类别的分割效果不理想。

4 结 论

本文提出了一种新的基于域自适应的、用于城市交通场景语义分割的对抗生成网络模型。一方面,为了减小源域数据集和目标域数据集之间的差异,本文首先采用 SG-GAN 对虚拟数据集 GTA5 进行风格转换,使得新的虚拟数据集 SG-GTA5 在灰度、结构以及边缘信息等都与目标域数据集更加接近,在一定程度上规避了模型网络在反向训练学习中出现的梯度爆炸问题,有效提升了网络的性能;另一方面,为了自适应地调整各特征层的损失值,动态地更新网络参数,本文在网络模型的不同特征层进行对抗训练学习中加入自适应的学习率,通过此方法优化了网络的性能,增强了网络的学习能力,从而有效提升了模型的分割效果;此外在判别器网络的最后新增了一个卷积层,使其能够更好地对图像的高层语义特征进行学习。

后期研究将进一步优化网络结构,考虑在源数据集中增加更多小目标、复杂结构目标类别以及其对应的标注。

参考文献(References)

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]

Chen Y H, Chen W H, Chen Y T, Tsai B C, Frank Wang Y C and Sun M. 2017. No more discrimination: cross city adaptation of road scene segmenters//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 1992-2001 [DOI: 10.1109/ICCV.2017.220]

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 (CVPR). Las Vegas, NV, USA: IEEE: 3213-3223 [DOI: 10.1109/CVPR.2016.350]

Dai J F, He K M and Sun J. 2015. BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 1635-1643 [DOI: 10.1109/ICCV.2015.191]

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M and Lempitsky V. 2015. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(1): 2096-2030 [DOI:10.1007/978-3-319-58347-1_10]

Goodfellow I, 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

Hoffman J, Tzeng E, Park T, Zhu J Y, Isola P, Saenko K, Efros A and Darrell T. 2017. CyCADA: cycle-consistent adversarial domain adaptation//Proceedings of the 35th International Conference on Machine Learning (ICML). [s.l.]: PMLR: 1994-2003

Hoffman J, Wang D Q, Yu F and Darrell T. 2016. FCNs in the wild: pixel-level adversarial and constraint-based adaptation [EB/OL]. [2019-08-15]. https://arxiv.org/pdf/1612.62649.pdf

Hong S, Noh H and Han B. 2015. Decoupled deep neural network for semi-supervised semantic segmentation//Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal, Canada: Neural Information Processing Systems Foundation: 1495-1503

Johnson-Roberson M, Barto C, Mehta R, Sridhar S N, Rosaen K and Vasudevan R. 2017. Driving in the matrix: can virtual worlds replace human-generated annotations for real world tasks? //Proceedings of 2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore: IEEE: 746-753 [DOI: 10.1109/ICRA.2017.7989092]

Khoreva A, Benenson R, Hosang J, Hein M and Schiele B. 2017. Simple does it: weakly supervised instance and semantic segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE: 876-885 [DOI: 10.1109/CVPR.2017.181]

LeCun Y, Bengio Y and Hinton G. 2015. Deep learning. Nature, 521(7553): 436-444 [DOI: 10.1038/nature14539]

Li P L, Liang X D, Jia D Y and Xing E P. 2018. Semantic-aware grad-GAN for virtual-to-real urban scene adaption [EB/OL]. [2019-08-05]. https://arxiv.org/pdf/1801.01726.pdf

Long M S, Cao Y, Wang J M and Jordan M I. 2015. Learning transferable features with deep adaptation networks//Proceedings of the 32nd

International Conference on Machine Learning (ICML). Lille, France: ACM: 97-105

Luo Y W, Zheng L, Guan T, Yu J Q and Yang Y. 2019. Taking a closer look at domain shift: category-level adversaries for semantics consistent domain adaptation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [s. l.]: [s. n.]: 2507-2516

Papandreou G, Chen L C, Murphy K and Yuille A L. 2015. Weakly - and semi-supervised learning of a deep convolutional network for semantic image segmentation//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE Computer Society: 1742-1750 [DOI: 10.1109/ICCV.2015.203]

Pathak D, Krahenbuhl P and Darrell T. 2015. Constrained convolutional neural networks for weakly supervised segmentation//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 1796-1804 [DOI: 10.1109/ICCV.2015.209]

Qu S R, Xi Y L and Ding S T. 2018. Image caption description of traffic scene based on deep learning. Journal of Northwestern Polytechnical University, 36(3): 522-527 (曲仕茹, 席玉玲, 丁松涛. 2018. 基于深度学习的交通场景语义描述. 西北工业大学学报, 36(3): 522-527) [DOI: 10.3969/j.issn.1000-2758.2018.03.017]

Richter S R, Hayder Z and Koltun V. 2017. Playing for benchmarks//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2213-2222 [DOI: 10.1109/ICCV.2017.243]

Tsai Y H, Hung W C, Schulter S, Sohn K, Yang M H and Chandraker M. 2018. Learning to adapt structured output space for semantic segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE: 7472-7481 [DOI: 10.1109/CVPR.2018.00780]

Zhang Y, David P and Gong B Q. 2017. Curriculum domain adaptation for semantic segmentation of urban scenes//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2020-2030 [DOI: 10.1109/ICCV.2017.223]

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2223-2232 [DOI: 10.1109/ICCV.2017.244]

作者简介

张桂梅，1970 年生，女，教授，主要研究方向为计算机视觉。
E-mail: guimei.zh@163.com

刘建新，通信作者，男，教授，主要研究方向为智能机器人，基于视觉的定位与导航。
E-mail: jamson_liu@163.com

潘国峰，男，硕士研究生，主要研究方向为图像处理与模式识别。E-mail: nchu_pgf@163.com