计算机科学
COMPUTER SCIENCE

变分推断域适配驱动的城市街景语义分割

金玉杰，初旭，王亚沙，赵俊峰

引用本文

金玉杰，初旭，王亚沙，赵俊峰. 变分推断域适配驱动的城市街景语义分割[J]. 计算机科学, 2022, 49(11): 126-133.

JIN Yu-jie, CHU Xu, WANG Ya-sha, ZHAO Jun-feng. Variational Domain Adaptation Driven Semantic Segmentation of Urban Scenes[J]. Computer Science, 2022, 49(11): 126-133.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于多路径特征提取的实时语义分割方法

Real-time Semantic Segmentation Method Based on Multi-path Feature Extraction

计算机科学, 2022, 49(7): 120-126. https://doi.org/10.11896/jsjkx.210500157

深度卷积神经网络图像实例分割方法研究进展

Survey Progress on Image Instance Segmentation Methods of Deep Convolutional Neural Network

计算机科学, 2022, 49(5): 10-24. https://doi.org/10.11896/jsjkx.210200038

基于深度学习的自动调制识别研究

Automatic Modulation Recognition Based on Deep Learning

计算机科学, 2022, 49(5): 266-278. https://doi.org/10.11896/jsjkx.211000085

基于国产众核处理器的深度神经网络算子加速库优化

Deep Neural Network Operator Acceleration Library Optimization Based on Domestic Many-core Processor

计算机科学, 2022, 49(5): 355-362. https://doi.org/10.11896/jsjkx.210500226

基于全局属性注意力神经过程模型的数据补全研究

Study on Data Filling Based on Global-attributes Attention Neural Process Model

计算机科学, 2022, 49(10): 111-117. https://doi.org/10.11896/jsjkx.210800038

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220500193

变分推断域适配驱动的城市街景语义分割

金玉杰[1,2]　初　旭[1,3]　王亚沙[1,4]　赵俊峰[1,2]
1 高可信软件技术教育部重点实验室(北京大学)　北京 100871
2 北京大学计算机学院　北京 100871
3 清华大学计算机系　北京 100084
4 北京大学软件工程国家工程研究中心　北京 100871
(jyj17pku@pku.edu.cn)

摘　要　街景语义分割技术旨在从图像中识别分割出行人、障碍物、道路、标志物等要素,为车辆提供道路上自由空间的信息,是自动驾驶的关键技术之一。高性能的语义分割系统非常依赖于训练时所需的大量真实标注数据,然而为图像中的每个像素进行标注成本很高,往往难以实现。一种低成本获取标注数据的方法是利用视频游戏收集逼真且标注成本低的合成图片,来帮助机器学习模型对现实世界中的图片作语义分割,这对应域适配技术。与当前基于VC维理论或Rademacher复杂度理论的主流语义分割域适配方法不同,受基于PAC-Bayes理论的兼容伪标签函数的域适配目标域Gibbs风险上界启发,考虑假设空间的平均情况而非最差情况,以避免主流方法过度约束隐空间上的领域差异,从而导致目标域泛化误差上界未能被有效估计并优化的问题。在上述思想的指导下,提出了一种变分推断语义分割域适配方法(VISA),该方法在利用Dropout变分族进行变分推断求解假设空间上的理想后验分布的同时能快速得到一个近似Bayes分类器,并通过目标域熵最小化和筛选像素点使得对风险上界的估计更加准确。在街景语义分割数据集GTA5→Cityscapes上的适配的实验结果表明,VISA方法相比基线方法平均交并比提高了0.5%~6.6%,且在行人、车辆等关键街景要素上具有较高的识别准确率。

关键词:语义分割;域适配;PAC-Bayes理论;变分推断;深度神经网络
中图法分类号　TP181

Variational Domain Adaptation Driven Semantic Segmentation of Urban Scenes

JIN Yu-jie[1,2],CHU Xu[1,3],WANG Ya-sha[1,4] and ZHAO Jun-feng[1,2]
1 Key Lab of High Confidence Software Technologies(Peking University),Ministry of Education,Beijing 100871,China
2 School of Computer Science and Technology,Peking University,Beijing 100871,China
3 Department of Computer Science and Technology,Tsinghua University,Beijing 100084,China
4 National Engineering Research Center for Software Engineering,Peking University,Beijing 100871,China

Abstract　Semantic segmentation of urban scenes aims to identify and segment persons,obstacles,roads,signs and other elements from the image,and provide information of free space on the road for vehicles. It is one of the key technologies of automatic driving. High performance semantic segmentation systems rely heavily on a large number of real annotation data required for training. However,labeling each pixel in the image is costly and often difficult to achieve. One way is to collect photo-realistic synthetic data from video games,where pixel-level annotation can be automatically generated at a low cost,to train the machine learning model to segment the images in the real world,which corresponds to domain adaptation. Different from the current mainstream semantic segmentation domain adaptation methods based on Vapnik-Chervonenkis dimension theory or Rademacher complexity theory,our method is inspired by the target domain Gibbs risk upper bound compatible with pseudo labels based on PAC-Bayes theory,and considers the average situation of the hypothetical space rather than the worst situation,so as to avoid excessively constraining the domain discrepancy in the latent space which leads to the problem that the upper bound of target domain generalization error cannot be estimated and optimized effectively. Under the guidance of the above ideas,this paper proposes a variational inference method for semantic segmentation adaptation(VISA). The dropout variational family is used for variational inference. While solving the ideal posterior distribution in the hypothesis space,an approximate Bayes classifier can be quickly obtained,and the estimation of the upper bound of risk is more accurate by minimizing the entropy of the target domain and filtering

到稿日期:2022-05-20　返修日期:2022-07-22
基金项目:国家自然科学基金(62172011)
This work was supported by the National Natural Science Foundation of China(62172011).
通信作者:王亚沙(wangyasha@pku.edu.cn)

pixels. Experiments show that the mean intersection over the union(mIoU) of VISA is 0.5% 6.6% higher than that of baseline methods, and has high accuracy in pedestrian, vehicle and other urban scene elements.

Keywords Semantic segmentation, Domain adaptation, PAC-Bayes theory, Variational inference, Deep neural network

1 引言

随着自动驾驶技术的逐渐成熟，作为其核心算法技术之一的城市街景语义分割(Semantic Segmentation of Urban Scenes)技术也受到了学术界和工业界的广泛关注。语义分割，顾名思义就是对图像中的各个像素按照其语义进行分类或分割。街景语义分割即对城市中的街景进行语义分割，从而识别分割出图像中的行人、障碍物、道路、标志物等要素，进而为车辆提供道路上自由空间的信息。车载摄像头或激光雷达探查到图像后，将其输入到神经网络中，后台计算机可以自动将图像分割归类，以使无人驾驶车辆避让行人和车辆等障碍，从而大大提高自动驾驶技术的安全性和实用性。

针对语义分割任务，研究者们设计了大量性能优良的模型[1-3]，但是这些模型非常依赖于训练时所需的大量标注数据。虽然现实中我们可以通过监控摄像等获得大量的真实街景图像，然而，对图像中的每个像素进行人工标注成本很高，这为街景语义分割技术带来了挑战。因此，设计低成本的方法来获取语义分割的数据变得越来越迫切。

目前流行的方法是从视频游戏中收集逼真的合成图像，这样，像素级的标注就能够以较低的成本自动生成。例如文献[4]构建了一个大规模的合成城市场景数据集，用于电脑游戏《侠盗猎车手 V》(Grand Theft Auto V, GTA V)的语义分割，大大降低了获取训练数据和标注的成本，如图 1 所示。虽然目前的技术可以使合成的街景图像非常逼真，但其与真实数据之间仍存在相当大的领域差异，导致将在合成数据上训练的分割模型应用于真实城市场景时性能会明显下降[5-6]。主要原因是：合成数据和真实数据之间存在领域差异，在合成数据上训练的模型更偏向于合成数据域，其中的卷积层倾向于过拟合合成数据，导致它们无法为真实图像提取有用的特征并完成分割。

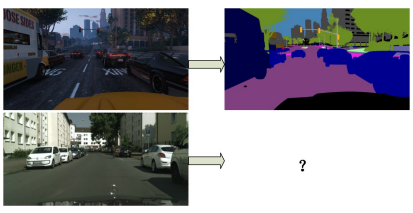


图 1 利用合成图像降低真实图像语义分割标注成本

Fig. 1 Reduce labelling cost of segmenting real images using synthetic images

域适配[7](Domain Adaptation)正是解决以上领域差异问题的一种技术，它往往涉及两个或两个以上的数据生成分布，其中被迁移知识的外部数据生成分布和目标任务数据生成分布分别称为源域(Source Domain)分布和目标域(Target Domain)分布。域适配技术探索如何利用源域大量标注样本来训练模型，从而尽可能提升目标域任务的性能，以减少领域差异对源域训练的模型在目标域性能上的负面影响。应用到街景语义分割中，源域即为合成图像，目标域即为真实图像。

目前主流的语义分割域适配方法大多基于对抗学习实现，如文献[8-9]，这些方法或从特征层和输出层通过领域对抗训练实现源域和目标域的分布对齐，或利用生成模型在输入层上实现源域和目标域的分布对齐。这种做法背后的理论依据实际上对应 Vapnik-Chervonenki 维度(VC 维)域适配理论[10]及 Rademacher 复杂度域适配理论[11]。基于 VC 维理论或 Rademacher 复杂度理论的域适配泛化误差上界表明，要使一个分类器在目标域中有较小的泛化误差，除了需要保证较小的源域经验误差外，还必须尽可能压缩领域分歧，即源域和目标域样本在某种度量下的分布距离。对抗学习的做法等效于压缩源域和目标域在隐空间的 Wasserstein 距离。然而，近期有研究指出，当源域和目标域的数据生成分布差异较大时，约束在隐空间上的领域差异会导致丢失过多目标域类别鉴别性信息[12]。此外，有研究指出，在非凸非凹的神经网络诱导的假设空间中，基于对抗学习的极小化极大优化易落入局部最优值，而非全局最优值[13]。这给估计并优化目标域泛化误差上界带来了挑战。

本文受基于 PAC-Bayes 理论的兼容伪标签函数的域适配目标域 Gibbs 风险上界[14]启发，该理论考虑假设空间中的平均 Gibbs 分类器而非单一分类器，其估计并优化的误差上界是平均情况而非最差情况，以避免主流方法中对最差情况的上界未能有效估计并优化的问题。为了解决街景语义分割域适配任务，针对语义分割任务的特点，提出了一种基于该上界的变分推断语义分割方法(Variational Inference for semantic Segmentation Adaptation, VISA)。在 VISA 中，本文使用贝叶斯卷积神经网络，利用 Dropout 变分族进行变分推断，求解假设空间上能尽量压低目标域 Gibbs 风险上界的理想后验分布，并能快速得到一个近似 Bayes 分类器。除了利用源域的有标注样本进行监督学习，为了使目标域伪标签函数更接近真实标签函数，本文使用了熵最小化的方法。为了解决语义类别出现频率分布不均匀的问题，根据每种语义类别的预测频率，为每个类别加权，形成一种类别平衡的熵最小化方法。为了能够更准确地估计目标域 Gibbs 风险上界[14]中的伪期望分歧，本文在目标域样本中逐类别地筛选出预测属于每一类中置信度较高的像素点。通过在语义分割数据集 GTA5 及 Cityscapes 上的适配实验，验证了 VISA 相比基线方法的性能提升。

本文的主要贡献如下：

(1)为了更好地估计并优化目标域泛化误差上界，提出了基于 PAC-Bayes 域适配理论的街景语义分割域适配方法框架，设计了一种变分推断域适配语义分割方法 VISA；

(2)为了更好地估计伪期望分歧，使用了逐类别筛选目标域图像像素点的方法；

(3)在街景语义分割域适配基准数据集上进行了一系列实验，验证了所提方法相比基线方法的性能提升。

本文第2节介绍城市街景语义分割及语义分割域适配的相关工作；第3节对预备知识进行介绍；第4节介绍所提出的VISA方法；第5节对实验数据集以及实验设计进行介绍，并展示与基准方法对比的实验结果；最后总结全文并展望未来。

2 相关工作

2.1 语义分割

一般地，语义分割的任务是为图像中的每个像素分配一个与其语义内容相对应的标签。研究者们已经创建了许多语义分割数据集，如城市街景语义分割数据集Cityscapes和Mapillary，以及具有深度信息的室内场景数据集NYUD-v2和SUN-RGBD。

语义分割是一个非常广泛的研究领域，为此研究者们提出了大量的方法。特别是，近年来深度学习技术研究的进展使语义分割性能获得了长足的进步。从著名的完全卷积网络(Fully Convolutional Networks，FCN)结构[15]被提出以来，研究者们提出了许多模型，如文献[1]通过金字塔池化模块和PSPNet来聚合基于不同区域的上下文信息；DRN[2]提出用扩张卷积替换卷积网络的下采样层，以增大神经元的感受野；DeepLab结构[3]通过带条件随机场(Conditional Random Field，CRF)的卷积神经网络推理图像的空间信息。

2.2 域适配

传统机器学习中的一个基本假设是训练数据和测试数据满足独立同分布条件。然而，这在现实场景中往往不成立，导致将训练的模型应用于测试数据时其性能会下降。域适配的目标是减小分布差异的负面影响，提升训练模型在目标域上的泛化能力。在计算机视觉中，图像分类中的域适配被广泛研究。目前主流的方法是通过特征匹配[16]或对抗学习[17]来减小源域和目标域的领域差异。

2.3 语义分割域适配

文献[5]最早提出了无监督域适配在语义分割中的应用，并提出了一种模型FCN in the wild，通过全局特征对齐和标签统计特征的匹配来解决语义分割域适配问题，由此激发了一系列无监督域适配在语义分割中的研究。根据适配(迁移)层次的不同，这些工作可大致分为三大类。

(1)输入层适配[18-19]：源域和目标域图像在高层的语义如场景、轮廓上可能高度相似，但域间低层次统计差异的存在仍然有可能导致在目标域样本上的预测性能下降，即使它们并不携带高层的语义信息。一类输入层适配工作的做法是通过风格迁移(Style Transfer)技术来减小源域和目标域图像边缘分布的差异。

(2)特征层适配[5,20]：这类工作的做法是学习一个领域分布对齐的隐空间，其核心思想是通过全局或逐类地进行隐空间特征对齐，使得特征提取器能够提取领域不变的特征，分割器可以从公共隐空间的表示学会正确分割。相比图像分类，语义分割的隐表示更高维也更复杂，包括全局和局部的视觉信息，而且特征层的对齐并不代表图像-标签联合分布的对齐，这可能会导致信息的丢失及误导。这类方法需要精心设计或结合其他技巧以避免上述问题。

(3)输出层适配[8-9]：这类方法在网络的输出层对齐领域分布。在保持足够复杂和丰富的语义线索的同时，来自分割网络输出的预测确定了一个低维空间，在该空间中可以有效执行适配，例如重复使用对抗策略。此外，输出层的标签统计可以很容易地从无标注的目标域数据上推断出来，从而为分割任务引入了一种自构造(Self-constructed)的弱监督形式。在适配过程中，也可以施加来自源域标签分布的先验，因为它们通常涉及不受特定域限制的高层结构特征。

这些做法背后的理论依据实际上对应基于VC维域适配理论[10]及Rademacher复杂度域适配理论[11]压缩隐空间中的领域分歧。然而，近期有研究指出，当源域和目标域的数据生成分布差异较大时，约束在隐空间上的领域差异会导致丢失过多目标域类别鉴别性信息[12]。此外，在非凸非凹的神经网络诱导的假设空间中，基于对抗学习的极小化极大优化易落入局部最优值，而非全局最优值[13]。本文的做法正是为了避免过度约束隐空间领域差异以及极小极大优化。

2.4 用合成数据学习

也有一些研究工作提出从合成数据中训练机器学习模型[21-22]。在文献[21]中，一般的目标探测器是从合成图像中训练出来的，而在文献[22]中，虚拟图像可用于提升真实环境中的行人检测性能。

3 预备知识

下面介绍一种兼容伪标签的PAC-Bayes域适配目标域风险上界[14]，作为本文方法的预备知识。

定理1(兼容伪标签函数的目标域Gibbs风险上界[14])对二分类问题及0-1损失函数，设$\tilde{y}'$是目标域伪标签函数。给定样本量为n的源域标注样本集合S和目标域无标注样本集合T，那么对于假设空间H上的所有分布Q及其对应的Gibbs分类器$\mathcal{G}_Q$，任意的$a,b>0$，$a'=\frac{2a}{1-e^{-2a}}$，$b'=\frac{b}{1-e^{-b}}$。给定置信水平$\delta\in(0,1]$，以不小于$1-\delta$的概率成立以下不等式：

$$R_{P_T}(\mathcal{G}_Q)\leqslant b'\hat{R}_{P_S}(\mathcal{G}_Q)+\frac{a'}{2}\hat{d}_Q(P_S,P_{\tilde{T}})+(a'-1)+\left(\frac{b'}{nb}+\frac{a'}{na}\right)\left(KL(Q\|\pi)+\log\frac{3}{\delta}\right)+\lambda_T^{\tilde{T}} \quad (1)$$

其中，$R_{P_T}(\mathcal{G}_Q)$和$R_{P_S}(\mathcal{G}_Q)$分别代表随机分类器$\mathcal{G}_Q$在目标域和源域上的Gibbs风险，上标$\hat{}$代表在数据集上的经验估计。$d_Q(P_S,P_{\tilde{T}})$为文献[14]定义的伪期望分歧(Expected Pseudo Discrepancy，EPD)项，用于衡量假设空间所能探测到的源域和目标域数据分布间的差异；$KL(Q\|\pi)$是假设空间上后验分布Q与某一个先验分布π的KL散度，是假设空间复杂度的度量；$\lambda_T^{\tilde{T}}$为文献[14]定义的伪标签期望分歧(Expected Pseudo Labeling Discrepancy，EPLD)项，是目标域伪标签函数和真实标签函数之间差异的度量。

由PAC-Bayes理论[23-24]可知，若能找到假设空间H上的一个后验分布Q，使得对应的随机Gibbs分类器$\mathcal{G}_Q$之Gibbs风险较低，则Q对应的Q-平均的Bayes分类器的泛化误差也可以被控制在较低水平。由定理1提出的目标域Gibbs风险上界可知，若能找到一个理想的后验分布Q，使得源域Gibbs风险的经验估计$\hat{R}_{P_S}(\mathcal{G}_Q)$、EPD项的经验估计$\hat{d}_Q(P_S,P_{\tilde{T}})$，以及

$KL(Q\|\pi)$都控制在较低水平，并且目标域伪标签函数和真实标签函数差异较小，则此后验分布 Q 具有较低的目标域 Gibbs 风险。相比文献[14]，本文的创新之处在于我们将针对图像分类的变分推断方法适配到图像的分割应用上，针对应用的特点设计了一套新颖的变分推断域适配框架。

4 VISA 方法

4.1 问题形式化

在街景语义分割域适配中，给定一个有标注源域街景数据集 $S=(\boldsymbol{x}_{s,i},\boldsymbol{y}_{s,i})_{i=1}^{n_s}$ 和一个无标注目标域街景数据集 $T=(\boldsymbol{x}_{t,i})_{i=1}^{n_t}$。其中 $\boldsymbol{x}_{s,i}\in\mathbb{R}^{H\times W\times 3}$，$\boldsymbol{x}_{t,i}\in\mathbb{R}^{H\times W\times 3}$ 为尺寸为 $H\times W$ 的彩色图片。考虑总语义类别为 C，$\boldsymbol{y}_{s,i}\in(1,C)^{H\times W}$ 是 $\boldsymbol{x}_{s,i}$ 的语义分割映射，即将 $\boldsymbol{x}_{s,i}$ 的每一个像素对应于一个独热编码向量(one-hot vector)，表示该像素的语义类别。为书写简便，本节在非必要时省略下标，且下文中的“源域”指代合成街景数据，“目标域”指代现实街景数据。

令 F 为一个语义分割网络，将一张图片 $\boldsymbol{x}$ 作为输入，并输出一个逐像素的 C 维概率分布表示预测结果，即：$\boldsymbol{p}=F(\boldsymbol{x})$，$p^{h,w,c}$ 表示预测图片 $\boldsymbol{x}$ 的像素点 (h,w) 属于第 c 类语义的概率。域适配的目标是利用 S 和 T 学习分割网络 F_ω 的参数 $\boldsymbol{\omega}$，使得其在目标域上有良好的分割性能，即：

$$\min_{\boldsymbol{x}}\mathbb{E}_{(\boldsymbol{x}_t,\boldsymbol{y}_t)\sim P_T}\ell(F_{\boldsymbol{\omega}}(\boldsymbol{x}_t),\boldsymbol{y}_t) \tag{2}$$

其中，P_T 为目标域图像与标签联合分布，ℓ 为某种选定的衡量预测结果与真实标签之差异的损失函数。

4.2 Dropout 变分推断

为了求解假设空间上的理想后验分布，本文使用以 KL 散度为优化目标的变分推断方法，使用一个 Dropout 变分族 $\tilde{Q}_\theta$ 来近似所要求解的后验 Q，估计并优化 Gibbs 风险上界，同时能够快速得到一个近似的 Bayes 分类器。假设 $\boldsymbol{\theta}=\{\boldsymbol{M}_i\}_{i=1}^{L}$ 是神经网络每一层的权重矩阵，$\boldsymbol{\varepsilon}=\{\boldsymbol{\varepsilon}_i\}_{i=1}^{L}$ 是 L 个伯努利随机向量。在第 i 层的激活层使用著名的 Dropout 正则化，相当于将 $\boldsymbol{M}_i$ 乘以 $\boldsymbol{\varepsilon}_i$ 得到一个随机矩阵 $\boldsymbol{W}_i=\boldsymbol{\varepsilon}_i\boldsymbol{M}_i$。这样，$\boldsymbol{\omega}=\{\boldsymbol{W}_i\}_{i=1}^{L}$ 可以被看作是由 $\boldsymbol{\theta}=\{\boldsymbol{M}_i\}_{i=1}^{L}$ 参数化的近似后验分布 $\tilde{Q}_\theta$ 中采样获得，这就是 Dropout 变分族(Dropout Variational Family)。从 $\boldsymbol{\varepsilon}=\{\boldsymbol{\varepsilon}_i\}_{i=1}^{L}$ 向 $\boldsymbol{\omega}=r(\boldsymbol{\theta},\boldsymbol{\varepsilon})$ 的转化被称为重参数化(Reparameterization)技巧。遵循这一技巧，本文在网络 F 中加入 Dropout 层，将网络 F 的参数重参数化为 $\boldsymbol{\omega}=r(\boldsymbol{\theta},\boldsymbol{\varepsilon})$，并默认使用权重衰减正则项及带动量的随机梯度下降优化算法。这种做法不但隐式地优化了定理 1 中的 KL 散度项[14]，而且便于得到定理 1 中源域 Gibbs 风险及伪期望分歧的经验估计，下面分别加以说明。

为了压低源域 Gibbs 风险，提升网络在源域图片上的分割性能，本文使用语义分割中常用的像素级交叉熵损失函数作为优化目标。用 $\mathbb{D}$ 表示每一维度以某一概率等于 1 的伯努利随机向量，假设随机样本 $\boldsymbol{\varepsilon}\mathbb{D}$，重参数化的网络为 $F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})}$。令 $\boldsymbol{p}_s=F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})}(\boldsymbol{x}_s)$ 为网络对源域样本 $\boldsymbol{x}_s$ 的预测输出，则：

$$L_{\mathrm{CE}}(\boldsymbol{x}_s,\boldsymbol{y}_s,F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})})=-\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}y_s^{h,w,c}\log\boldsymbol{p}_s^{h,w,c} \tag{3}$$

优化式(3)可以使得网络对合成的虚拟街景图片拟合出较好的分割性能，这是迁移到现实街景的基础。为估计并压低伪期望分歧，减少分割网络所能探测到的源域和目标域的分布差异，本文使用 EPD 损失函数作为优化目标。用 $\hat{\boldsymbol{y}}_t$ 表示对目标域样本 x_t 赋予的像素级伪标签(见 4.3 节)，类似地，令 $\boldsymbol{p}_t=F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})}(\boldsymbol{x}_t)$ 为网络对源域样本 $\boldsymbol{x}_t$ 的预测输出，定义目标域上带伪标签的交叉熵损失函数为：

$$L_{\mathrm{CE}}(\boldsymbol{x}_t,\hat{\boldsymbol{y}}_t,F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})})=-\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}\hat{\boldsymbol{y}}_t^{h,w,c}\log\boldsymbol{p}_t^{h,w,c} \tag{4}$$

则 EPD 损失函数定义为：

$$\begin{aligned}L_{\mathrm{EPD}}(S,T)=\Big|\frac{1}{m}\sum_{(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}')\sim\mathbb{D}^2}^{m}\Big[\frac{1}{n_t}\sum_{i=1}^{n_t}\big(L_{\mathrm{CE}}(\boldsymbol{x}_t,\hat{\boldsymbol{y}}_t,F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})})+\\ L_{\mathrm{CE}}(\boldsymbol{x}_t,\hat{\boldsymbol{y}}_t,F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon}')})\big)-\frac{1}{n_s}\sum_{i=1}^{n_s}\big(L_{\mathrm{CE}}(\boldsymbol{x}_s,\boldsymbol{y}_s,\\ F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})}),L_{\mathrm{CE}}(\boldsymbol{x}_s,\boldsymbol{y}_s,F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon}')})\big)\Big]\Big|\end{aligned} \tag{5}$$

优化 EPD 损失函数的意义是：对于从近似后验分布(即 Dropout 变分族)中随机采样的两个分割网络 $F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})}$ 及 $F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon}')}$，考虑它们在目标域和源域数据生成分布下的期望表现差异，既可以约束 $F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})}$ 及 $F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon}')}$ 在两个数据生成分布上表现的一致性，也可以约束它们在两个数据生成分布上表现的不一致性。一致性指，当 $F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})}$ 及 $F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon}')}$ 在目标域图片上的分割性能较好(差)时，期望它们在源域图片上的分割性能也较好(差)。不一致性指，当 $F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})}$ 及 $F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon}')}$ 在目标域图片上的分割性能差异较大时，期望它们在源域图片上的分割性能差异也较大。

4.3 目标域的伪标签构造及像素点筛选

4.3.1 熵最小化

由于整个网络优化了源域交叉熵损失函数(见式(3))，因此可以将其视为一个在源域上具有较好性能的分割网络。本文直接使用网络 F(关闭 Dropout 层，近似 Q-平均 Bayes 分类器)对每个目标域样本的预测输出，为其每个像素生成一个独热编码，将预测值最高的维度分量置为 1，其他维度置为 0，作为该像素点的伪标签，即：

$$\hat{\boldsymbol{y}}_t^{h,w,c}=\begin{cases}1, & F^{h,w,c}(\boldsymbol{x}_t)>F^{h,w,c'}(\boldsymbol{x}_t)\text{对所有}c'\neq c\\ 0, & \text{其他情况}\end{cases} \tag{6}$$

由于领域差异的存在，直接利用由源域分割网络对目标域样本的预测结果来构造伪标签可能不够准确。考虑到一个普遍现象，即在一个领域上训练的模型往往会产生过分自信的低熵(Low-entropy)预测结果。为了使得用上述方法构造的伪标签更加准确，本文约束模型对目标域样本也输出熵低的预测值，从而实现源域和目标域在网络预测值特征意义下的“一致性”。因此，本文将目标域样本预测值的熵作为优化目标，即：

$$L_{\mathrm{ent}}(\boldsymbol{x}_t,F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})})=-\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}\boldsymbol{p}_t^{h,w,c}\log\boldsymbol{p}_t^{h,w,c} \tag{7}$$

街景语义分割问题场景下经常出现类别失衡问题，即易于准确预测的类别往往在图片上有众多像素点，例如以下类别往往占据街景图像上大量像素点且易于准确预测：天空(Sky)，街道(Road)，建筑物(Building)；以下类别像素点占比较少且难以准确预测：火车(Train)，骑手(Rider)，栅栏(Fence)等。这将导致不同类别对损失函数的贡献比例失衡。一种常用的针对类别失衡问题的方法是，为每一类引入一个权重 α_c，反比于该类出现的频率[25]。在无监督域适配问题

中，由于目标域无真实标签，无法计算真实类别频率，且因为不能确保目标域和源域数据集有相同的类别频率，所以并不合适将源域的统计频率特征照搬过来。为解决该问题，本文使用一种逐图像类别平衡加权(Image-wise Class-balanced Weighting)方法。首先，在每张目标域图片上计算类别的预测频率：

$$m^{h,w,c}=\begin{cases}1, & \text{如果 } c=\underset{c'}{\operatorname{argmax}}\ \boldsymbol{p}^{h,w,c'}\\ 0, & \text{其他情况}\end{cases} \tag{8}$$

$$N_c=\sum_{h=1}^{H}\sum_{w=1}^{W}m^{h,w,c} \tag{9}$$

N_c统计了一张图片被预测为语义类别 c 的像素点个数。为了平衡类别频率对损失函数贡献的影响，文献[26]提出的加权系数为：

$$\alpha_c \propto \frac{1}{N_c^{\alpha}} \tag{10}$$

其中，α 为可调节的超参数。考虑到预测可能不准确，α 取值应介于(0,1)。本文将式(7)的熵最小化优化目标改为类别加权形式的熵最小化：

$$L_{\text{ent}}^{cb}(\boldsymbol{x}_t,F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})})=-\frac{1}{N_c^{\alpha}(HW)^{-\alpha}}\cdot L_{\text{ent}}(\boldsymbol{x}_t,F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})}) \tag{11}$$

4.3.2 按类别筛选像素点

式(5)在计算伪期望分歧损失函数时使用了目标域图片的所有像素点。考虑到目标域伪标签并非完全准确，这种做法可能欠妥。受 VIDA 的目标域样本加权(Target Instance Weighting，TIW)及类别失衡问题的启发，本文采取一种按类别筛选目标域图片像素点的方法。

具体地，给定一张目标域图片 $\boldsymbol{x}_t$，对每一语义类别 c，本文挑选出满足以下条件的像素点 (h,w)：1)预测伪标签属于类别 c，即 $c=\underset{c'}{\operatorname{argmax}}\ p^{h,w,c'}(\boldsymbol{x}_t)$。2)预测属于第 c 类的置信度，在所有满足条件 1)的像素点的置信度从大到小排序中，位于前 μ 的比例。令 $\delta^{h,w}$ 为 0-1 二值变量，为 1 表示像素点 (h,w) 满足以上条件，否则为 0。于是，计算式(5)中的 $L_{\text{CE}}(\boldsymbol{x}_t,\hat{\boldsymbol{y}}_t,F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})})$ 时将只计入 $\delta^{h,w}=1$ 的像素点的贡献。其中 μ 为随着训练进行可动态调节的参数。由于随着训练的进行期望模型对目标域的预测变得准确，因此可以将更大比例的像素点加入损失函数的计算。在实际实验中，将 μ 的初始值设为 20%，每过一个 epoch 将 μ 增大 1%，μ 的最大值不得超过 50%。采用按类别筛选像素点的方法，使得每个迭代轮次可以挑选出预测属于每一类别中置信度最高的一部分像素点加入模型训练，使伪期望分歧的估计更准确。

模型整体架构图如图 2 所示，通过端到端联合优化式(3)的源域交叉熵损失函数、式(11)的目标域类别平衡加权熵损失函数和式(5)所示的伪期望分歧损失函数来优化参数，即总优化目标为：

$$\frac{1}{n_s}\sum_{S}L_{\text{CE}}(\boldsymbol{x}_s,\boldsymbol{y}_s,F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})})+\frac{\lambda_1}{n_t}\sum_{T}L_{\text{ent}}^{cb}(\boldsymbol{x}_t,F_{r(\boldsymbol{\theta},\boldsymbol{\varepsilon})})+\lambda_2 L_{\text{EPD}}(S,T) \tag{12}$$

其中 λ_1 和 λ_2 为平衡损失函数的系数。另外，优化过程中默认使用权重衰减正则项及带动量的随机梯度下降算法。算法 1 对模型的训练过程进行了总结。

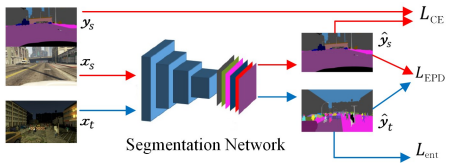


图 2 本文方法网络架构图

Fig. 2 Network framework of our method

算法 1 本文方法的训练过程

输入：源域数据集 $S=\{(\mathbf{x}_i^s,\mathbf{y}_i^s)\}_{i=1}^{n_s}$，目标域数据集 $T=\{(\mathbf{x}_i^t)\}_{i=1}^{n_t}$，带有 dropout 的语义分割网络模型 F^{ω}

输入超参数：源域批量大小 b_s，目标域批量大小 b_t，初始比例 μ_0，后验分布采样数 m，交叉熵、伪期望分歧损失函数的系数 λ_1，λ_2，更新目标域伪标签的间隔迭代次数 I，开始加上伪期望分歧损失函数的迭代次数 startIter，目标域像素点筛选比例的初始值 μ_0，步进值 μ_{step}，最大值 μ_m

输出：已学习的参数 ω

iter←0，$\mu\leftarrow\mu_0$
while 方法未收敛 do
  if iter mod I=0 then
    if iter>startIter then
      $\mu\leftarrow\min(\mu+\mu_{\text{step}},\mu_m)$
    for T 中的每一个 $\mathbf{x}_i^t$ do
      计算 $\mathbf{x}_i^t$ 的伪标签 $\hat{\mathbf{y}}_i^t$
    end for
  end if
  $\mathscr{L}_{\text{train}}\leftarrow 0$
  $\mathscr{B}_s\leftarrow$从 S 中随机采样 b_s 个样本 $\{(\mathbf{x}_i^s,\mathbf{y}_i^s)\}_{i=1}^{b_s}$
  $\mathscr{B}_t\leftarrow$从 T 中随机采样 b_t 个样本 $\{(\mathbf{x}_i^t,\tilde{\mathbf{y}}_i^t)\}_{i=1}^{b_t}$
  $\mathscr{L}_{\text{train}}\leftarrow\mathscr{L}_{\text{train}}+\frac{1}{b_s}\sum_{i=1}^{b_s}L_{\text{CE}}(\mathbf{x}_i^s,\mathbf{y}_i^s,F^{\omega})$
  $\mathscr{L}_{\text{train}}\leftarrow\mathscr{L}_{\text{train}}+\lambda_1 L_{\text{ent}}^{cb}$
  if iter>startIter then
    随机采样 m 对伯努利随机向量 $(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}')$，用于 dropout
    $\mathscr{L}_{\text{train}}\leftarrow\mathscr{L}_{\text{train}}+\lambda_2\mathscr{L}_{\text{EPD}}$
  end if
  通过带动量的随机梯度下降算法用 $\nabla_{\omega}L_{\text{train}}$ 更新 $\boldsymbol{\omega}$
  iter←iter+1
end while

5 实验验证

5.1 街景数据集

为了与主流城市街景语义分割工作对比，本文选取 GTA5[4] 作为源域数据集，Cityscapes[27] 作为目标域数据集，即适配任务为 GTA5→Cityscapes。

GTA5：该数据集来自电脑游戏 Grand Theft Auto V，均是从汽车内视角观察到的美式虚拟城市的场景。包含 24 966 张分辨率为 1 914×1 052 px 的图片。原始 GTA5 图片带有像素级别的标注，共计 33 类，本文仅使用其中与 Cityscapes 共有的 19 类进行模型训练及测试。

Cityscapes：该数据集来自欧洲 50 个国家的真实街道场景，分辨率为 2 048×1 024 px。本文遵循标准协议，在训练

模型时，使用 Cityscapes 训练集的 2975 张图片，不使用它们的真实标注；在测试模型时，使用 Cityscapes 验证集的 500 张图片及它们的真实标注。

5.2 实验设置

5.2.1 评价指标

在语义分割领域，一般使用平均交并比（Mean Intersection over the Union，mIoU）作为方法性能的评价指标，即每类交并比（IoU_c 对类别 c）的平均值，它们的定义如下：

$$IoU_c=\frac{TP_c}{FP_c+FN_c+TP_c} \tag{13}$$

$$mIoU=\frac{1}{C}\sum_{c=1}^{C}IoU_c \tag{14}$$

其中，TP_c、FP_c、FN_c 分别为对类别 c 的分类真正样本数、假正样本数、假负样本数（“样本”即像素点），C 为总类别数。

5.2.2 分割网络

为了达到更好的实际分割效果及与其他工作的实验结果进行合理公正的比较，很有必要选取一个较强的基线网络作为分割网络并验证本文方法的性能。因此，本文采用以在 ImageNet[28] 数据集上预训练的 ResNet-101[29] 网络为骨架的 DeepLab-v2[3] 模型作为分割网络 F。为了捕捉语义信息，该模型在 ResNet-101 的最后一层特征层的输出后面加上了空洞空间金字塔池化层，与 ASPP-L 模型[3] 类似，本文将 ASPP 采样率固定为{6，12，18，24}。本文在 ASPP 层的后面加了一层 Dropout 层。最后，本文采用一个无参数的上采样层（Upsampling Layer）和一个 Softmax 层得到最终的概率预测输出 $F(x)$。遵循近期语义分割的工作[8] 等的做法，本文修改了 ResNet-101 最后两个卷积层 conv4 和 conv5 的步长和空洞卷积扩张率，以获得更密集的特征映射和更大的感受野。为公平起见，用于对比的基线方法均采取基于 ResNet-101 骨架的 DeepLab-v2 模型作为分割网络。

5.2.3 实验细节及超参数选取

全部网络参数采用带动量的随机梯度下降算法进行更新，动量值取为 0.9。全部实验默认使用权重衰减 L_2 正则项，正则项强度系数设为 5×10^{-4}。训练过程中冻结 ResNet-101 骨架中所有批标准化层的参数，不对其进行梯度下降更新。ResNet-101 骨架的初始参数设为在 ImageNet 数据集上预训练得到的参数，ASPP 网络的参数随机初始化，将 ResNet-101 骨架的学习率设为 ASPP 网络学习率的 0.1 倍。本文的学习日程调度使用文献[3]的多项式退火调度：$r=r_0\cdot\left(1-\frac{iter}{max\_iter}\right)^{0.9}$，初始学习率 r_0 设为 2.5×10^{-3}。因为 GPU 显存限制，批量大小设为 2（分别两张源域及两张目标域图片）。

由于显存大小限制，在所有实验中固定源域批量大小和目标域批量大小为 2。训练时对图片采用了数据增强策略，包括随机长宽比裁切（均裁切为固定尺寸 1024×512 px）、修改亮度及饱和度和随机水平镜像翻转。测试时仅将输入图片的尺寸缩小为 1024×512 px。

在每个适配任务上，训练 150000 个迭代轮次。为了得到更准确的伪标签函数和伪期望分歧的估计，考虑到神经网络开始时需要一定的迭代轮次来拟合源域数据，因此在实验中，在第 75000 个迭代轮次后再将伪期望损失函数加入优化目标。为了使涉及伪标签的优化过程更稳定，每隔 2500 个迭代轮次更新一次目标域样本的伪标签及二值筛选变量。

实验中用多轮贪心调整的方式进行超参数选取。因为 GPU 显存限制，本文没有对采样数 m 进行调优，而是将其固定为 2。对于 μ_{step}，本文将其固定为 0.01。UISA 方法中进行调优的各个超参数，其考虑的范围及最终选取的值如表 1 所列。

表 1 本文实验考虑的超参数范围及最终选取的超参数

Table 1 Range and final selected hyperparameters in our experiments

超参数	考虑范围	最终选取
λ_1	{1，0.2，0.1，0.05，0.02，0.01}	0.2
λ_2	{1，0.1，0.01，0.005，0.001，0.0005}	0.001
α	{0，0.2，0.5，1}	0.2
(μ_0,μ_m)	{(0.2，0.5)，(0.5，0.8)}	(0.2，0.5)

5.3 结果比较

本文将 VISA 方法和一些采用基于 ResNet-101 的 DeepLab-v2 模型作为分割网络的基线方法进行了对比。这些基线方法包括：无适配方法 NonAdapt，对抗式方法 AdaStruct[8]，SIBAN[30]；生成与对抗结合式方法 DLOW[19]，Cycada[18]。在任务 GTA5→Cityscapes 上的实验结果如表 2 所列。从中可以看出，VISA 方法相比各种基线方法达到了最高的 mIoU 指标 43.2%，并在语义类别行人、汽车、卡车、公共汽车等关键要素上表现出较高的准确率。

表 2 GTA5→Cityscapes 适配实验结果

Table 2 Adaption results from GTA5 to Cityscapes

方法	mIoU	road	sidewalk	building	wall	fence	pole	light	sign	veg
NonAdapt	36.6	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3
AdaStruct[8]	41.4	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8
DLOW[19]	42.3	87.1	33.5	**80.5**	24.5	13.2	29.8	29.5	26.6	82.6
SIBAN[30]	42.6	**88.5**	35.4	79.5	**26.3**	24.3	28.5	32.5	18.3	81.2
Cycada[18]	42.7	86.7	**35.6**	80.1	19.8	17.5	**38.0**	**39.9**	**41.5**	**82.7**
VISA	**43.2**	83.0	25.7	79.3	22.2	**25.5**	24.9	34.2	22.1	82.1
方法	terrain	sky	person	rider	car	truck	bus	train	mbike	bike
NonAdapt	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	26.3	36.0
AdaStruct[8]	25.9	75.9	57.3	**26.2**	76.3	29.8	32.1	**7.2**	29.5	32.5
DLOW[19]	26.7	**81.8**	55.9	25.3	78.0	33.5	38.7	0.0	22.9	34.5
SIBAN[30]	**40.0**	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5
Cycada[18]	27.9	73.6	**64.9**	19.0	65.0	12.0	28.6	4.5	**31.1**	**42.0**
VISA	35.7	80.1	58.3	23.3	**82.6**	**43.2**	**45.9**	0.6	28.7	23.3

5.4 消融实验

为了验证本文模型优化目标中类别平衡熵损失函数和伪期望分歧损失的有效性，本文在 GTA5→Cityscapes 任务上进行了消融实验(Ablation Study)，并比较了如下几种消融变体：1)CE：模型仅用源域交叉熵损失函数训练；2)CE+Ent：模型用源域交叉熵损失函数及类别平衡熵损失函数训练；3)VISA：完整模型，即用交叉熵损失、类别平衡熵损失、伪期望分歧损失训练。4)VISA*：将 VISA 模型中的伪期望分歧用期望分歧替代，即取消目标域像素点的伪标签构造及筛选步骤，用真实标签替代(由于训练时不能获得真实标签，该实验仅用于验证 EPD 损失的意义)。

实验结果如表 3 所列。可见完整 VISA 模型的性能优于 CE 及 CE+Ent 两种消融变体的性能，且 VISA* 比 VISA 的性能更好，这说明了：1)类别平衡熵损失的有效性；2)限制随机采样的一对模型在源域和目标域数据生成分布上性能表现的一致性和不一致性，即约束期望分歧是有效的；3)研究如何使目标域的伪标签更准确，是改进 VISA 性能的关键之一。

表 3 GTA5→Cityscapes 消融实验结果

Table 3 Adaption results from Cityscapes to GTA5

方法	mIoU	road	sidewalk	building	wall	fence	pole	light	sign	veg
CE	36.6	75.8	16.8	77.2	12.5	21.0	**25.5**	30.1	20.1	81.3
CE+Ent	42.2	78.9	**25.6**	79.4	21.8	**25.6**	22.1	29.6	16.4	81.9
VISA	43.2	83.0	25.7	79.3	**22.2**	25.5	24.9	**34.2**	**22.1**	82.1
VISA*	**44.5**	**87.0**	**27.7**	**79.6**	19.6	20.8	23.0	32.4	21.4	**82.9**
方法	terrain	sky	person	rider	car	truck	bus	train	mbike	bike
CE	24.6	70.3	53.8	**26.4**	49.9	17.2	25.9	**6.5**	26.3	36.0
CE+Ent	33.3	**80.5**	**60.0**	**28.9**	80.9	33.7	37.9	0.1	**32.4**	32.9
VISA	35.7	80.1	58.3	23.3	**82.6**	**43.2**	45.9	0.6	28.7	23.3
VISA*	**36.1**	80.2	59.0	**28.9**	82.3	40.9	**46.7**	6.2	29.8	**41.1**

5.5 超参数敏感度

本文展示了 VISA 模型的性能对超参数 λ_1，λ_2，α 的敏感度，如表 4 所列。由表可知，对于 λ_1，λ_2 都不应设置过大或过小，而是在取适中值时性能最佳。

表 4 超参数敏感度实验结果

Table 4 Results of hyperparameter sensitivity

	GTA5→Cityscapes			
λ_1	0.05	0.10	0.20	0.50
	41.8	42.3	43.2	42.8
λ_2	0.0005	0.0010	0.0020	0.0050
	42.4	43.2	41.4	38.1
α	0.1	0.2	0.5	1.0
	42.5	43.2	41.7	36.6

对于类别加权因子中的参数 α，当其趋近于 0 时即趋向无加权的熵最小化，当其趋近 1 于时，由于预测值并不准确，导致对每一类别的像素点统计值不准确，这时所赋的权重非常不准，以致于严重影响性能，故 α 也需要有适中的取值。

结束语 本文为了解决街景语义分割面临的标注成本高的问题，引入了迁移学习的域适配技术，利用大量标注成本低且易于获得的合成图片，来帮助机器学习模型对现实世界中的无标注街景图片作语义分割。与当前基于 VC 维理论或 Rademacher 复杂度理论的主流语义分割域适配方法不同，本文方法受基于 PAC-Bayes 理论的兼容伪标签函数的域适配目标域 Gibbs 风险上界的启发，考虑假设空间的平均情况而非最差情况，以避免主流方法因过度约束隐空间上的领域差异，从而导致目标域泛化误差上界未能被有效估计并优化的问题。在上述思想的指导下，本文提出了一种变分推断语义分割域适配方法 VISA，此方法在利用 Dropout 变分族进行变分推断求解假设空间上的理想后验分布的同时能快速得到一个近似 Bayes 分类器，本文通过类别平衡的熵最小化方法使得目标域伪标签函数更接近真实标签函数，并通过按类别筛选预测置信度高的像素点，使得对伪期望分歧的估计更加准确。通过在语义分割域数据集 GTA5 及 Cityscapes 上的适配实验，验证了 VISA 相对于基线方法的性能提升。

本文仍有一些可补充或改进之处可作为未来工作的方向：1)本文的伪期望分歧损失函数是神经网络所探测到的源域和目标域的差异的度量。对于街景语义分割任务而言，它不仅要求模型对图像有全局性的理解，还要求对图像的内在结构特征(包括局部和全局)有着更深层次、更综合视角的理解。因此有必要探索在 PAC-Bayes 的框架下设计新的损失函数或新的方法，使之能够更全面、更深层地捕捉源域和目标域样本的全局及局部差异；2)街景图像的类别信息及其内部的相互关联在源域和目标域上应具有某种相似性。例如：语义为“天空”的像素点往往在图像上方；语义为“骑手”的像素点和“自行车”的像素点应在位置上邻近。这种关联在目前的工作中未得到充分探索。

参 考 文 献

[1] 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.

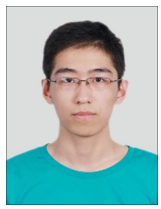
[2] YU F, KOLTUNV, FUNKHOUSER T. Dilated residual networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 472-480.

[3] 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.

[4] RICHTER S R, VINEET V, ROTH S, et al. Playing for data: Ground truth from computer games[C]// European Conference on ComputerVision. Cham: Springer, 2016: 102-118.

[5] HOFFMAN J, WANG D, YU F, et al. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation[C]// CoRR.

2016.
[6] ZHANG Y,DAVID P,GONG B. Curriculum domain adaptation for semantic segmentation of urban scenes[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:2020-2030.
[7] WANG M,DENG W. Deep visual domain adaptation:A survey [J]. Neurocomputing,2018,312:135-153.
[8] TSAI Y H,HUNG W C,SCHULTER S,et al. Learning to adapt structured output space for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7472-7481.
[9] VU T H,JAIN H,BUCHER M,et al. Advent:Adversarial entropy minimization for domain adaptation in semantic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:2517-2526.
[10] BEN-DAVID S,BLITZER J,CRAMMER K,et al. Analysis of representations for domain adaptation[C]// Advances in Neural Information Processing Systems 19:Proceedings of the 2006 Conference. MIT Press,2006.
[11] MANSOUR Y,MOHRI M,ROSTAMIZADEH A. Domain adaptation:Learning bounds and algorithms[J]. arXiv:0902.3430,2009.
[12] LIU H,LONG M,WANG J,et al. Transferable adversarial training:A general approach to adapting deep classifiers[C]// International Conference on Machine Learning. PMLR,2019:4013-4022.
[13] JIN C,NETRAPALLI P,JORDAN M. What is local optimality in nonconvex-nonconcave minimax optimization? [C]// International Conference on Machine Learning. PMLR,2020:4880-4889.
[14] CHU X. Feature Map Sharing towards High-dimensional Under-Labeled Data Analysis[D]. Beijing:Peking University,2021.
[15] LONG J,SHELHAMER E,DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition. 2015:3431-3440.
[16] LONG M,CAO Y,WANG J,et al. Learning transferable features with deep adaptation networks[C]// International Conference on Machine Learning. PMLR,2015:97-105.
[17] GANIN Y,USTINOVA E,AJAKAN H,et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research,2016,17(1):2096-2030.
[18] JUDY H,ERIC T,TAESUNG P,et al. Cycada:Cycle-consistent adversarial domain adaptation[C]// Proceedings of the 35th International Conference on Machine Learning. 2018.
[19] GONG R,LI W,CHEN Y,et al. Dlow:Domain flow for adaptation and generalization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:2477-2486.
[20] CHEN Y,LI W,VAN GOOL L. Road:Reality oriented adaptation for semantic segmentation of urban scenes[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7892-7901.
[21] SUN B,SAENKO K. From Virtual to Reality:Fast Adaptation of Virtual Object Detectors to Real Domains[C]// BMVC. 2014.
[22] VAZQUEZ D,LOPEZ A M,MARIN J,et al. Virtual and real world adaptation for pedestrian detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,36(4):797-809.
[23] GERMAIN P,HABRARD A,LAVIOLETTE F,et al. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers[C]// International Conference on Machine Learning. PMLR,2013:738-746.
[24] GERMAIN P,HABRARD A,LAVIOLETTE F,et al. PAC-Bayes and domain adaptation[J]. Neurocomputing,2020,379:379-397.
[25] LIN T Y,GOYAL P,GIRSHICK R,et al. Focal lossfor dense object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:2980-2988.
[26] CHEN M,XUE H,CAI D. Domain adaptation for semantic segmentation with maximum squares loss[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:2090-2099.
[27] 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.
[28] DENG J,DONG W,SOCHER R,et al. Imagenet:A large-scale hierarchical image database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2009:248-255.
[29] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.
[30] LUO Y,LIU P,GUAN T,et al. Significance-aware information bottleneck for domain adaptive semantic segmentation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:6778-6787.



JIN Yu-jie, born in 1999, postgraduate. His main research interests include machine learning and data mining.

WANG Ya-sha, born in 1975, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include smart city and big data analysis.

(责任编辑:何杨)