中国科学: 技术科学 2022 年 第 52 卷 第 1 期: 26 ~ 54
SCIENTIA SINICA Technologica techcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

评 述 工业互联网专辑

基于无监督域自适应的计算机视觉任务研究进展

孙琦钰, 赵超强, 唐漾*, 钱锋

华东理工大学信息科学与工程学院, 上海 200237
* E-mail: yangtang@ecust.edu.cn

收稿日期: 2021-04-08; 接受日期: 2021-07-02; 网络版发表日期: 2021-09-02
国家自然科学基金基础科学中心项目(批准号: 61988101)、国家杰出青年科学基金(批准号: 61725301)、上海市优秀学术带头人计划项目(编号: 20XD1401300)、高等学校学科创新引智计划(编号: B17017)和中央高校基本科研业务费专项资金(编号: 222202117006)资助

摘要 作为工业互联网的典型实例之一, 车联网技术近年来飞速发展, 其核心在于信息的互联互通. 因此, 精准、可迁移的环境信息感知能力是其稳定运行的前提之一. 深度学习的进步推动了计算机视觉任务的发展, 但基于传统深度学习的方法仍存在训练过程对人工标注数据依赖强、场景泛化能力较差的弊端. 而对于计算机视觉任务来说, 训练数据的真值标签获取较难, 因此如何提升模型的迁移能力, 缓解训练对人工标注的依赖受到了学界的广泛关注. 无监督域自适应方法使用深度学习模型进行特征提取和对齐, 使得深度学习模型在不同域间迁移时仍能保证良好的性能, 在计算机视觉任务中发挥了重要作用. 因此, 本综述主要聚焦无监督域自适应方法在一些典型计算机视觉任务中的挑战和应用. 首先, 介绍了基于深度学习的无监督域自适应方法的定义、重要意义、应用难点、基本方法和相关数据集. 然后, 分别针对典型计算机视觉任务介绍了无监督域自适应方法在其中的应用. 最后, 进行了总结和展望.

关键词 无监督域自适应, 计算机视觉, 深度学习, 迁移学习, 自主系统

1 引言

工业互联网通过信息的互联互通将人、机、物紧密结合成为可靠高效的信息物理系统[1], 实现了感知、认知、决策一体化[2], 面向无人驾驶的车联网便是工业互联网的实例之一. 对于无人驾驶系统来说, 可迁移的环境信息的能力是系统安全稳定运行的前提和基础. 近年来, 计算机视觉不断发展, 逐渐成为自主体(如无人机、无人车、机器人等)感知和理解世界的重要途径[3~5]. 深度学习在很多典型计算机视觉任务如图像分类[6,7]、目标检测[8]、语义分割[9,10]、深度估计[11,12]中应用广泛, 极大提升了模型的预测精度. 一般来说, 用于计算机视觉任务的深度学习模型以有监督的方式进行训练, 需要使用大量含有真值标签的数据. 因此, 人们针对不同的视觉任务收集了大量训练数据并进行标注, 以使深度学习方法能够成功应用于各类视觉任务. 然而, 这种标注十分耗费时间和精力. 特别地, 像素级视觉任务(如语义分割和深度估计等)的真值标注时间成本非常高, 仅对常用的语义分割数据集Cityscapes中的一张图像进行语义标注就需耗时1.5小

引用格式: 孙琦钰, 赵超强, 唐漾, 等. 基于无监督域自适应的计算机视觉任务研究进展. 中国科学: 技术科学, 2022, 52: 26–54
Sun Q Y, Zhao C Q, Tang Y, et al. A survey on unsupervised domain adaptation in computer vision tasks (in Chinese). Sci Sin Tech, 2022, 52: 26–54, doi: 10.1360/SST-2021-0150

© 2021 《中国科学》杂志社

时[13], 由此, 计算机视觉任务的真值标签获取之难便可见一斑, 这也一定程度上限制了相关视觉任务分析和处理的发展.

此外, 虽然深度神经网络具有较强的拟合和表征能力, 提取的特征与手工提取的特征相比, 泛化性更强[14], 但计算机视觉任务的复杂性使得其在特定数据集上训练好的深度学习模型也很难较好地迁移到数据分布不同的新场景中[15,16]. 这就导致即使研究人员花费大量时间精力建立的有标注的视觉任务数据集, 在该数据集上训练得到的深度学习模型也仅能在此数据集上表现优良, 一旦将该模型应用到其他数据分布不同的数据集上, 其性能可能会显著下降. 这是由不同数据集之间的风格和统计数据分布的差异性造成的, 称为数据集的偏向性(Dataset bias)或域偏移(Domain shift)[17,18]. 域偏移普遍存在于实际应用场景与某一标签数据充足的数据集之间, 但若针对每个新场景都进行数据的重新采集和标注, 时间成本极高且不切合实际. 因此, 人们开始探索是否存在一种方法, 可以减轻域偏移对模型造成的影响, 使得所训练的模型在一个数据集上训练后, 可以在另一个没有或只有少量标签的数据集上也获得较好的结果, 这种方法称为域自适应(Domain adaptation, DA)[19,20]. 域自适应方法, 特别是无监督域自适应(Unsupervised domain adaptation, UDA)的发展, 极大地提升了深度学习模型的场景迁移能力. 近年来, 人们提出了许多无监督域自适应方法, 其中基于深度神经网络的域自适应方法受到了广泛关注[21].

无监督域自适应方法的发展与成熟为解决计算机视觉任务有标签训练数据稀缺的问题提供了新的思路和解决方案. 但现有的综述主要介绍了无监督域自适应在图片分类任务中的应用[19~22], 对于其他典型计算机视觉任务中使用的无监督域自适应方法的介绍还较少. 基于此, 本文对一些典型计算机视觉任务中的无监督域自适应方法的研究进展进行了总结, 包括图像分类、目标检测、语义分割和深度估计等典型计算机视觉任务, 以方便更多研究人员了解这一领域. 本文章节安排如下: 第2部分介绍了无监督域自适应的定义、重要意义、应用难点、基本方法和相关数据集; 第3~6部分分别介绍了图像分类、目标检测、语义分割和深度估计任务中无监督域自适应的进展; 第7部分对未来的研究趋势进行了展望; 最后, 对文章进行了总结.

2　无监督域自适应简述

2.1　迁移学习与域自适应

本小节主要通过参考文献[20,23,24]介绍迁移学习与域自适应的相关定义和符号表示方法.

域 域$\mathcal{D}$由两部分组成: 特征空间$\mathcal{X}$和其边缘概率分布$P(X)$, 其中, $X=\{x_1,\cdots,x_n\}$, x_i为训练数据, 则域可表示为$\mathcal{D}=\{\mathcal{X},P(X)\}$.

任务 对于某一个域$\mathcal{D}=\{\mathcal{X},P(X)\}$, 它的任务也由两部分组成, 可记为$\mathcal{T}=\{\mathcal{Y},f(\cdot)\}$. 其中, $\mathcal{Y}$为标签空间, $f(\cdot)$为映射函数. 在机器学习中, $f(\cdot)$经由数据训练学习得到. 由于机器学习一般是有监督的, 因此训练数据包含训练图片$x_i\in X$及其对应的标签$y_i\in\mathcal{Y}$, 可记为数据标签对$\{x_i,y_i\}$.

迁移学习(Transfer learning, TL) 给定一个源域$\mathcal{D}_s$和它的任务$\mathcal{T}_s$, 一个目标域$\mathcal{D}_t$和它的任务$\mathcal{T}_t$. 当$\mathcal{D}_s\neq\mathcal{D}_t$(但$\mathcal{T}_s$与$\mathcal{T}_t$相同或相关)或$\mathcal{T}_s\neq\mathcal{T}_t$ (但$\mathcal{D}_s$与$\mathcal{D}_t$相同或相关)时, 可以利用$\mathcal{D}_s$和$\mathcal{T}_s$的信息, 更好地完成$\mathcal{D}_t$的任务$\mathcal{T}_t$, 这一过程称为迁移学习. 按照$\mathcal{D}_s$与$\mathcal{D}_t$, $\mathcal{T}_s$与$\mathcal{T}_t$的关系, 综述[23]将迁移学习分为三类: 归纳式迁移学习(Inductive TL)、无监督迁移学习(Unsupervised TL)和直推式迁移学习(Transductive TL), 如表1所示. 传统机器学习不考虑域偏移和任务差异, 即$\mathcal{D}_s=\mathcal{D}_t$且$\mathcal{T}_s=\mathcal{T}_t$, 因此在训练集上训练好的模型可以直接推广应用到测试集上, 并获得不错的结果. 但在迁移学习中, $\mathcal{D}_s\neq\mathcal{D}_t$或$\mathcal{T}_s\neq\mathcal{T}_t$, 因此需要采用不同的迁移策略提升网络性能.

域自适应 域自适应属于直推式迁移学习, 即源域和目标域任务相同, 两个域的数据虽然相近但仍存在一定差异, 即$\mathcal{D}_s\neq\mathcal{D}_t$, $\mathcal{T}_s=\mathcal{T}_t$, 如表1所示. 它使用真值标签丰富的源域数据和没有真值标签(或有少数真值标签)的目标域数据一起训练模型, 使得模型能在目标域上表现良好. 域自适应旨在研究如何将模型从源域学习到的知识转移到目标域, 用以减轻域偏移的影响, 提升模型的泛化性能. 一般来说, 当有两个分布相似但不同的数据集, 可将含有较多训练数据(有真值)或更容易采集的数据集作为源域, 另一个作为目标域. 域自适应方法同时使用源域和目标域中的数据一起训练模型, 通过将源域的知识迁移到目标域上, 使得模型在

表 1 传统机器学习与各类迁移学习的关系[23]

Table 1 Relationship between traditional machine learning and various transfer learning settings [23]

类别		源域$\mathcal{D}_s$和目标域$\mathcal{D}_t$	源域任务$\mathcal{T}_s$和目标域任务$\mathcal{T}_t$
传统机器学习		相同	相同
迁移学习	归纳式迁移学习/无监督迁移学习	相同	不同但相关
		不同但相关	不同但相关
	直推式迁移学习	不同但相关	相同

标签较少甚至没有标签的目标域中也能获得不错的结果[23].

2.2 无监督域自适应

如表2所示, 域自适应根据目标域中的数据是否包含标签可分为三类: 目标域中所有数据均包含标签的有监督域自适应, 少量数据包含标签的半监督域自适应和没有数据包含标签的无监督域自适应[19,20]. 无监督域自适应旨在减小源域(有标签的数据)和目标域(无标签的数据)之间的域偏移对模型性能的影响, 使用源域中的数据提升模型在目标域的性能[25,26]. 特别地, 对于计算机视觉任务来说, 在目标域中收集有标签的数据难度大、成本高, 因此无监督域自适应方法的优势十分突出[27]. 自迁移学习兴起以来, 不断与其他机器学习方法结合, 形成了多种迁移学习框架. 域自适应作为迁移学习的一个子类, 也不断与最新的机器学习方法结合发展. 本文主要关注无监督域自适应在计算机视觉任务中的研究进展, 由于深度神经网络有着强大的层次化特征学习和数据表征能力[28], 并且在计算机视觉任务中发挥着重要作用, 因此本文主要聚焦无监督深度域自适应[29].

特别地, 在计算机视觉任务中, 对于某一源域$\mathcal{D}_s$, 定义训练样本$\mathcal{D}_s = (x_s^i, y_s^i), i = 1, \cdots, n_s$. 其中, n_s为源域中可用于训练的图片数量, x_s为源域的图片数据, y_s为x_s对应的真值标注, 其可能是用于分类任务的图片级标注、用于语义分割的像素级标注或用于其他任务的标注. 对于某一目标域$\mathcal{D}_t$, 有训练样本$\mathcal{D}_t = x_t^j, j = 1, \cdots, n_t$. 其中, n_t为目标域中可用于训练的图片数量. 注意到目标域只含有图片x_t而没有标签数据y_t, 因此希望通过域自适应在源域上训练一个从x_s到y_s的映射, 并将这个映射用于在目标域上从x_t估计y_t. 不同视觉任务数据集之间数据分布的差别造成了域偏移, 包括合成和真实数据集之间的差别, 不同的合成数据集之间的差别, 真实数据集间由于地理位置、光照、季节等的不同而引起的差别等[16]. 在无监督深度域自适应中, 深度学习网络通过共享源域和目标域之间的部分网络参数, 同时通过寻找从源域数据分布到目标域数据分布的映射, 或通过将两个不同的域映射到同一个共享空间中, 来解决域偏移带来的问题[28]. 由于目标域没有标签, 深度学习网络往往会学习如何提取一种既能较好地在源域上完成学习任务又具有域不变性的特征表示, 使模型可直接适用于目标域[29].

表 2 域自适应的分类

Table 2 The classification of different UDA methods

类别		源域标签	目标域标签
域自适应(直推式迁移学习)	有监督域自适应	大量	全部, 量少
	半监督域自适应	大量	部分, 量少
	无监督域自适应	大量	没有

2.3 无监督域自适应在计算机视觉任务中的应用概要

由于来自不同域中图片的拍摄视角、背景、光照、天气、像素质量不同, 图片中包含的物体外观和位置等也有差异, 使得一些计算机视觉任务的深度学习模型虽然在标准数据集上能够取得较好的效果, 但仍不能较好地应用于真实世界. 因此, 在计算机视觉领域, 研究人员常常使用无监督域自适应方法减小域偏移的影响[27]. 本文对一些典型计算机视觉任务中使用的无监督域自适应方法的研究进展做了归纳和分析, 包括图像分类、目标检测、语义分割和深度估计, 如表3所示. 不同的视觉任务因其特点不同, 这些任务进行域自适应的难度也不同, 表3简要介绍了这些典型

表 3　计算机视觉任务的无监督域自适应方法

Table 3　UDA methods for computer vision tasks

任务	简介	难点	方法	代表文献
图像分类	将一张图片归类到若干个类别中的某一类中.	虽然一些算法在标准数据集上取得了不错的效果, 但当将其应用于其他数据集时, 域偏移现象导致分类精度降低的幅度较大[30].	基于最小散度	[30,31,40~42,81,82,87,90~92]
			基于生成对抗	[26,29,44~47,93,94,96,97,100~103]
			基于其他方法	[14,106,107,109,110]
目标检测	包含目标定位和分类两个子任务, 旨在同时识别并定位图片中所有物体.	由于目标检测任务中目标的数目、位置、大小、外观等的不同, 目标检测的域自适应相较于图片分类更复杂[32].	基于生成对抗	[32,48,111,112,114,116]
			基于其他方法	[118,119]
语义分割	像素级分类任务, 即为图像的每个像素分配语义标签.	由于不同域图片的语义种类不同, 或不同类在不同域中的占比不同, 导致语义分割相较于图片分类的域偏移现象更加严重[27].	基于生成对抗	[27,49,74,126~149,151~154]
			基于其他方法	[155~167,170,171]
深度估计	像素级回归任务, 即为图像的每个像素分配一个深度值.	由于需要估计像素级的连续值, 域自适应难度也较大[33].	基于生成对抗	[33,50,172~176,178,179]

的视觉任务并分析了这些任务域自适应的难点. 其中, 图像分类任务中的无监督域自适应技术最为简单[30,31]. 由于目标检测需要对图片中的多个物体同时进行识别和定位[32], 而语义分割和深度估计任务需要进行像素级的分类和回归估值[27,33], 因此, 进行域自适应的难度都比图像分类大.

图像分类是指将图片归类到若干个类别中的某一类[6]. 在计算机视觉任务中, 图像分类任务相对简单, 因此图像分类任务中的无监督域自适应方法发展较早, 其基本思想也促进了其他更复杂的视觉任务的域自适应的发展[27]. 目标检测也是计算机视觉领域的一个重要且基础的问题, 它包含目标识别和定位两个子任务, 旨在同时识别并定位图片中所有物体[8]. 与图片分类任务相比, 目标检测需要关注图片中一个或多个物体的类别, 并同时实现定位. 而不同域中的目标种类、数目、大小等都会存在差异, 因此域偏移对目标检测任务的影响更大, 进行域自适应任务也更困难[32]. 语义分割和深度估计是像素级的计算机视觉任务, 语义分割指为图像的每个像素分配语义标签[9], 深度估计则是指为图像的每个像素分配一个深度值[11]. 与简单的分类与回归任务相比, 像素级视觉任务的域自适应面临着更大的挑战. 首先, 域自适应在分类任务中遇到的困难在像素级任务中都会遇到, 并且由于像素级任务需要在一个指数级的大空间中解决分类或回归问题, 使得克服域偏移变得愈加困难[27]. 对于语义分割来说, 域偏移主要有两点. 第一, 不同域的特征空间不同. 第二, 不同域的语义种类不同, 或者不同种类在不同域当中所占比例不一样. 在分类和语义分割任务中, 不同的域中不同类别的物体存在的概率分布情况往往会不一样. 对于语义分割来说, 由于不同的类可能在一张图片中出现多次, 因此这一问题更加显著[27]. 与语义分割任务不同的是, 深度估计不是分类估计, 而是回归估计, 需要估计像素点的深度真值, 因此进行无监督域自适应挑战也很大. 虽然深度估计能够使用无监督学习的方法通过图片间的几何关系构建自监督信号, 但这种自监督信号易受动态环境的干扰, 使得预测结果不准确[34].

2.4　无监督域自适应基本方法

由于计算机视觉任务类别众多, 难以一一列举, 因此本文重点介绍了图像分类等四个典型计算机视觉任务的无监督域自适应方法. 虽然其他计算机视觉任务, 例如目标跟踪、行人重识别等, 均存在域偏移的问题, 且域偏移给不同视觉任务带来的困难不尽相同, 但从整体上看, 不同计算机视觉任务中使用的无监督域自适应方法仍存在相通性. 这是因为目前完成计算机视觉任务的主流方法一般基于深度学习, 以端对端的方式完成任务, 因此不同视觉任务关联度较高. 表3对一些典型的计算机视觉任务中的域自适应方法进行了分类、归纳、总结. 对于其他计算机视觉任务而言, 无监督域自适应方法也可分为基于生成对抗网络的方法和基于其他方法两类, 因此本文介绍的方法对于其他计

算机视觉任务是域自适应也有一定借鉴作用. 表4进一步对计算机视觉任务中的无监督域自适应方法进行了总结和对比. 由于生成对抗网络的基本思想是通过两个网络的互相博弈实现较优解, 而域自适应过程也是源域和目标域的博弈, 因此生成对抗网络被广泛用在计算机视觉任务的无监督域自适应中. 此外, 其他的一些方法, 如课程学习(Curriculum learning)[35]、自监督训练(Self-training, ST)[36,37]、熵最小化(Entropy minimization)[38]、知识蒸馏(Knowledge distillation)[39]等也被应用于无监督域自适应中, 并取得了良好的结果. 特别地, 基于最小化散度的无监督域自适应方法也被用于分类任务[30,31,40]. 这通常指通过最小化某些度量来减小域偏移带来的影响, 例如最大平均偏差(Maximum mean discrepancy, MMD)[30]、相关距离(Correlation distances)[41,42]等. 而在目标检测、语义分割和深度估计任务中, 一般不采用基于最小化散度的方法进行域自适应, 这是因为这些任务的域自适应往往更加依赖图片中的纹理、上下文等信息, 仅通过最小化不同域间的数据分布的散度无法获得较好的结果[32].

基于生成对抗思想的方法一般利用生成对抗网络(Generative adversarial networks, GAN)[43]进行训练. 它包含两个子网络: 生成器和鉴别器, 通过一个零和博弈, 使生成器和鉴别器进行对抗学习. 生成器经过训练, 以一种使鉴别器混淆的方式生成数据, 而鉴别器则试图将生成器产生的数据与真实数据区分开来, 两个网络经过博弈, 最终达到平衡[43]. 网络以一种最小最大化的方式进行迭代训练, 即更新生成器使目标函数最小化, 更新鉴别器使目标函数最大化. 基于生成对抗思想的无监督域自适应方法通过使用一个域鉴别器或对抗性目标判断数据的来源, 使得网络能够提取域不变特征, 从而对齐两个域的特征分布[29,44,45]. 如图1(a)所示, 这类方法一般由一个任务网络(如图像分类网络)和一个域鉴别器组成, 两个网络通过对抗的思想进行训练. 这时, 任务网络可视为生成器[46,47], 与域鉴别器一起进行对抗训练. 任务网络通过不断更新网络参数, 使其提取的源域与目标域的特征分布尽可能相近, 从而使域鉴别器无法识别其提取的特征来自哪一域; 域鉴别器分别输入来自源域和目标域的数据, 目的是判断输入数据来自于源域还是目标域. 在此过程中, 两个网络通过对抗不断提升各自的能力, 任务网络逐渐可以习得在源域和目标域上都表现良好的域不变特征, 减小域偏移的影响. 特别地, 输入鉴别器的数据可以来自任务网络(生成器)提取的中间特征(基于特征空间的特征对齐), 也可以来自任务网络(生成器)对于特定计算机视觉任务的预测结果(基于输出空间的预测对齐). 此外,一些方法还在进行域自适应之前, 在输入空间上加入额外的GAN对训练图像进行图片风格转换从而实现数据增强, 提升域自适应效果[48~50], 如图1(b)所示. 本文根据视觉任务是否在特征空间和输出空间上使用域鉴别器(图1(a)), 以及是否在输入空间上使用额外的GAN进行数据增强(图1(b)), 按照基于特征空间、基于输出空间、基于输入空间的顺序, 对基于生成对抗思想的无监督域自适应方法进行分类

表 4 无监督域自适应各类方法的对比

Table 4 Comparisons of different UDA methods

方法			定义	优点	缺点
计算机视觉任务中的无监督域自适应	基于最小化散度的方法		最小化源域与目标域的统计量之间的散度	简单易行	效果较差
	基于生成对抗的方法	特征空间	任务网络(生成器)将提取的图片特征输入鉴别器	简单易行、应用广泛	对于语义分割、深度估计等像素级任务的效果不好
		输出空间	任务网络(生成器)将预测的结果输入鉴别器	可利用图片的空间结构、语义逻辑等信息	输出结果不准确或存在噪声
		输入空间	域自适应前, 使用风格迁移网络进行数据增强	数据增强可提升域自适应的效果	依赖额外的风格迁移网络
	基于其他方法	课程学习	由易到难逐步学习	有利于风格差异较大的域或样本之间的迁移	需要有针对性的设计或利用更多先验知识
		自监督训练	迭代生成伪标签进行自监督标签	目标域可以进行有监督训练	伪标签存在噪声
		熵最小化	最小化预测的混乱程度	减小预测噪声	对微小物体的预测不佳

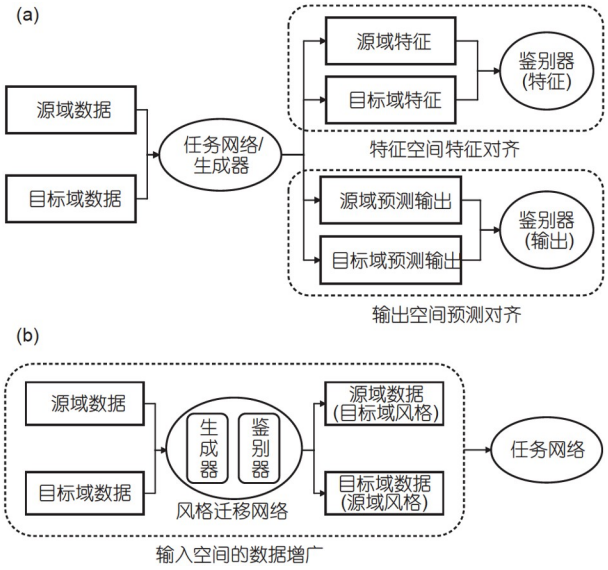

图 1　基于生成对抗的无监督域自适应方法. (a) 特征空间与输出空间; (b) 输入空间

Figure 1　Unsupervised domain adaptation methods using generative adversarial networks. (a) Feature space and output space; (b) input space.

介绍.

基于其他方法的无监督域自适应方法不依赖对抗思想, 而是通过使用课程学习[35]、自监督训练[36,37]、熵最小化[38]、知识蒸馏[39]、元学习[51]等思想进行域自适应. 课程学习是指通过模拟人的认知机理, 即首先学习简单的、普适性的知识结构, 然后逐渐增加难度, 过渡到更复杂、更专业化知识的学习; 自监督训练的方法一般适用于训练数据真值稀缺的情况, 将缺失的真值标签视为可学习的隐藏变量, 对于无监督域自适应来说, 目标域的真值是缺失的, 因此, 一些研究将自监督训练的思想用于无监督域自适应中; 熵最小化是通过计算并最小化预测输出中像素点的熵值, 来获得更可靠的目标域预测; 知识蒸馏使用一个训练好的复杂深度网络指导一个轻量级网络的训练, 以使得轻量级的网络能够习到复杂深度网络的拟合能力; 元学习也被称为学会学习, 即通过学习以往的知识和经验来引导新任务的学习. 近年来, 一些方法将元学习引入到计算机视觉任务的无监督域自适应方法中, 通过学习源域的知识以指导目标域的任务.

与其他方法相比, 基于生成对抗思想的方法的优点在于不需要在训练过程中进行复杂的抽样或推理, 但缺点在于训练可能较难达到平衡[52]. 基于课程学习的方法将源域到目标域的域自适应分为由易到难的几步进行, 这样的策略有利于风格差异较大的域或样本的迁移. 基于自监督的方法通过生成伪标签, 为目标域的训练提供伪标签, 将无监督训练转换为有监督训练, 提升训练的可靠性, 但存在伪标签置信度不高或存在噪声等问题. 基于熵最小化的方法能够有效地减小预测的噪声, 利用上下文信息. 在域自适应任务中, 可以通过组合使用不同的方法提升网络的迁移性, 在本文中, 按照是否使用生成对抗思想分类对一些典型计算机视觉任务中的无监督域自适应进行了总结和归纳.

2.5　数据集

为利用机器学习完成计算机视觉任务, 人们花费很多时间和精力收集数据并制作数据集. 从较早的简单的手写数字数据集MNIST[53], 到用于像素级视觉任务的数据集Cityscapes[13]、KITTI[54]等, 数据集规模不断扩大, 标注难度也逐渐提升. 对于图像分类、目标检测、语义分割任务来说, 首先需要收集大量图片, 然后人工为其标注真值[13]. 而深度估计任务则需要通过激光雷达或结构光深度传感器采集数据获取真值, 然后与相对应的RGB图像进行对齐和校准. 这些标注、对齐和校准的工作都十分繁琐和耗时, 且容易出现标注错误[54,55] .

近年来, 随着计算机制图和游戏引擎的不断发展成熟, 研究人员可以以较低的成本生成大规模甚至无限数量的数据集用于深度学习的训练. 这些训练数据风格多样, 涵盖内容广, 包括合成的手写数字数据集[26]、数字和图标数据集[26]、室内环境数据集[56~58]、以及道路场景数据集[59~63]等. 这些合成数据的语义真值可以自动生成, 无需人工标注. 因此, 可以使用有标签的合成数据集作为源域进行训练, 将没有标签的真实数据集作为目标域. 但由于合成数据集与真实数据集之间域偏移的存在, 需要使用域自适应算法使在合成域上训练的模型能够较好地应用到真实域. 表5归纳总结了一些常见的用于图像分类、目标检测、语义分割和深度估计中的数据集[13,25~27,53~79], 并对其进行了简要介绍.

在图像分类中, 常用数字[26,53,64,65]、图标[66,67]、不同场景[25,68,69]和物体[70~73]作为数据集. 将内容相近的数据集分别作为源域和目标域, 使用无监督的域自适

表 5 常见的用于图像分类、目标检测、语义分割和深度估计中的数据集

Table 5 The datasets used in image classification, target detection, semantic segmentation and depth estimation

数据集	年份	缩写[1)]	真实/合成	室内/外	标签	简介
SYN Numbers[26]	2016	SYNN	合成	–	类别	合成数字数据集, 由50万张图像组成, 通过改变打印字体的文本、位置、方向、背景和笔画颜色以及模糊程度生成.
MNIST-M[26]	2016	MN-M	合成	–	类别	由MNIST数据集中的数字与BSDS500数据集中的随机色块混合形成有彩色的手写数字集.
MNIST[53]	1998	MN	真实	–	类别	手写数字数据集, 由0~9十类数字组成.
USPS[64]	1990	USPS	真实	–	类别	手写数字数据集, 由0~9十类数字组成.
SVHN[65]	2011	SVHN	真实	–	类别	从谷歌街景中城市门牌号图像中提取的单个数字.
SYN SIGNS[66]	2013	SYNS	合成	–	类别	合成的标志数据集, 由维基百科上的交通信号图修改而成, 由10万幅图像和43类标志组成.
GTSRB[67]	2011	GT	真实	室外	类别	由在德国各地驾驶时拍摄的视频创建的数据集, 由约5万幅图像和43种标志组成.
Office-31[25]	2010	Of	真实	室内	类别	由来自三个不同领域(Amazon, DSLR, Webcam)的31个类别的4652张图像组成, 包含办公场景中常见对象, 如键盘、文件柜等.
Office-31+ CIFAR-10[68]	2012	O-C	真实	室外	类别	由Office-31和Caltech-256[79]两个数据集共享的10个常见类别组成.
Office-Home[69]	2017	O-H	真实	室外	类别	由来自四个不同领域的65个类别的15500张图像组成, 包含办公室和家庭中常见对象.
CIFAR-10[70]	2009	CI	真实	室外	类别	共有60000张彩色图像, 分为10个类, 每类6000张图.
STL-10[71]	2011	STL	真实	室外	类别	受CIFAR-10启发形成的数据集, 共有十类, 有100000张无标签图片用于无监督学习和少量有标签的数据.
VisDA[72]	2017	VisDA	真实、合成	室内、室外	类别、语义	由来自真实和合成域的12个类别的28万张图像组成.
ImageCLEF-DA[2)]	2014	I-DA	真实	室内、室外	类别、语义	是ImageCLEF 2014域自适应挑战的基准数据集, 由三个公共数据集(Caltech-256, ImageNet ILSVRC 2012, Pascal VOC 2012)中的12个公共类组成, 每个数据集被视为一个领域.
PASCAL VOC[73]	2010	PV	真实	室外	分类、检测、语义	PASCAL VOC 2012包含20个类别.
Cityscapes[13]	2016	C	真实	室外	检测、语义	涵盖了50个欧洲城市街道场景, 包含5000个稠密语义标注(共有34个类, 其中19个语义类别有像素级标签)和20000个稀疏标注的图像.
BDDS[27]	2016	B	真实	室外	语义	涵盖了不同条件下的驾驶场景, 如夜间的城市街景、雨天的公路场景等, 包含数千个密集注释和数十万个未标记的帧, 共有34个类别.
PBRS[56]	2017	PB	合成	室内	语义等	一个具有500K合成图像实例的数据集, 包含表面法线预测、语义分割和物体边界检测真值.
SYNTHIA[59]	2016	S	合成	室外	语义	含13类, 7个序列, 涵盖不同场景(高速公路、环岛、山路、纽约市、欧洲老城)、季节(春、夏、秋、冬)、天气(雨、小雨、雾)、光照(日落、黎明、夜晚)等, 一般使用Synthia-Rand-Cityscapes, 与Cityscapes配对.
GTA5[61]	2016	G	合成	室外	语义	由电脑游戏渲染得来, 来自位于南加州的洛杉矶的虚构城市洛斯桑托斯, 包含24966张图片, 有19个类的注释, 与Cityscapes完全兼容.
SIM 10k[61]	2016	SIM	合成	室外	检测	由10000张GTA5中渲染的图像组成, 包含58701辆车的bounding box.
Foggy Citys-capes[62]	2018	FC	合成	室外	检测、语义	使用来自Cityscapes的图像和深度图进行渲染, 在真实场景中模拟雾, 标签与Cityscapes中一样.
Synscapes[63]	2018	Y	合成	室外	语义	由25000张完全注释的RGB图像组成, 与Cityscape类似, 注释真值包含19个类别.

(续表5)

数据集	年份	缩写[1)]	真实/合成	室内/外	标签	简介
NTHU[74]	2017	N	真实	室外	语义	又称Cross-City, 包含4个域偏移较小的城市: 里约热内卢、罗马、东京和台北, 每个城市含100幅图像注释对(13类)和3200幅无注释图像.
Oxford RobotCar[75]	2017	O	真实	室外	深度、语义等	安装在车上的6个摄像头、激光雷达、GPS和INS真值采集的近2000万张照片. 数据收集了所有天气条件, 包括大雨、夜晚、阳光直射和降雪.
Mapillary Vistas[76]	2017	M	真实	室外	语义	在不同天气、季节、光照下拍摄, 包含城市、乡村、越野场等场景, 含有25000张密集注释的街道图像, 共66个类, 对其中37类进行实例标注.
KITTI[54]	2012	K	真实	室外	深度、语义等	含40000多对立体图像对及激光雷达数据, 包含市区、乡村和高速公路等场景, 可用于评测深度、光流、视觉里程计、物体检测与跟踪等任务.
NYU Depth v2[55]	2012	ND	真实	室内	深度、语义等	包含原始和处理好的样本. 原始数据集包含464个场景(训练集249, 测试集215), 处理好的数据集由1449个RGB-D样本组成, 其中深度地图被补全并与RGB图像对齐, 另外含有和894个类别标注.
SUNCG[57]	2017	SU	合成	室内	深度、语义等	用合成场景渲染单个视角的彩色深度和语义, 主要拿来做基于单帧图像的语义场景修复.
SceneNet RGB-D[58]	2017	Sc	合成	室内	深度、语义等	为场景理解(如语义分割、实例分割和目标检测)提供像素级标签, 同时也为几何计算机视觉问题(如光流、相机位姿估计和深度估计)提供真值标签.
Virtual KITTI[60]	2016	vK	合成	室外	深度、语义等	提供了对象检测与跟、场景和实例分割、深度、光流等的真值, 包含21260帧不同环境条件下生成的图像深度对.
Places365[77]	2016	P	真实	室内外	语义等	拥有1000万张场景照片, 用场景语义类别和属性标注.
Make3D[78]	2008	M3	真实	室外	深度	包含几百张图片及其对应的稀疏深度值.

1) 后文中表6~9中涉及到的数据集均采用这种缩写形式.

2) http://imageclef.org/2014/adaptation.

应方法进行训练. 目标检测常用的数据集有Cityscapes[13]、Foggy Cityscapes[62]、SIM 10k[61]、KITTI[54]等. Cityscapes与Foggy Cityscapes可以模拟不同天气情况下的域偏移, Cityscapes与SIM 10k可以模拟真实环境与虚拟环境之间的域偏移, 而Cityscapes与KITTI可以模拟不同城市之间的域偏移. 对于语义分割来说, 常用数据集有Cityscapes[13]、SYNTHIA[59]、GTA5[61]等, 一般在合成和真实数据集之间进行域自适应, 希望通过合成数据集来提升模型在真实数据集中的语义分割结果. 与语义分割一样, 深度估计的域自适应任务也可通过虚拟环境来获得有真值的合成深度的数据进行训练, 成本低且可用的训练数据量巨大, 常用的数据集有NYU Depthv2[55]、KITTI[54]、Virtual KITTI[60]等.

本节主要介绍了无监督域自适应的定义、重要意义、应用难点、基本方法和相关数据集, 以下章节将分别介绍无监督域自适应在图像分类、目标检测、语义分割和深度估计任务中的应用.

3　图像分类

本节对无监督域自适应在图像分类任务中的应用进行了归纳和总结, 将主要方法分为三类介绍: 基于最小化散度、基于生成对抗思想和基于其他方法的无监督域自适应方法, 一些典型的方法概括在表6中. 对于基于生成对抗思想的方法, 表中还特别列出了鉴别器的输入, 对于在输入空间上加入额外的风格迁移网络的方法表中特别标明了生成器的生成方向, 其中, S代表源域, T代表目标域.

3.1　基于最小化散度的无监督域自适应方法

对于图像分类任务, 早期的无监督域自适应方法通过最小化源域与目标域的统计量之间的散度对齐深度学习模型提取特征. 比较常见的散度包括最大平均偏差(Maximum mean discrepancy, MMD)[80]、相关对齐(Correlation alignment, CORAL)[41]等.

表 6 无监督域自适应方法在图像分类中的应用

Table 6 Image classification with UDA

分类		年份	方法	鉴别器输入	生成方向[1)]	数据集[2)]
最小化散度		2014	DDC[30]	–	–	Of
		2015	DAN[31]	–	–	Of, O-C
		2016	RTN[40]	–	–	Of, O-C
		2016	CORAL[41]	–	–	Of, O-C
		2016	Deep CORAL[42]	–	–	Of
		2017	JAN[81]	–	–	Of, I-DA
		2018	Rozantsev等人[82]	–	–	Of, MN, USPS, SVHN
		2018	DeepJDOT[87]	–	–	MN, USPS, MN-M, SVHN
		2018	Das等人[90,91]	–	–	USPS, MN, O-C
		2019	CAN[92]	–	–	Of, VisDA
		2020	RWOT[89]	–	–	MN, USPS, SVHN, Of, I-DA, O-H, VisDA
生成对抗	特征空间	2015	Ganin等人[26,29]	特征	–	MN, MN-M, SYNN, SVHN, SYNB, GT
		2015	Tzeng等人[44]	特征	–	Of
		2017	ADR[46]	特征	–	MN, USPS, SVHN
		2017	ADDA[93]	特征	–	MN, USPS, SVHN
		2017	WDGRL[96]	特征	–	O-C
		2018	Saito等人[47]	特征	–	MN, USPS, SVHN, GT, SYNS
		2018	Sohn等人[94]	特征	–	MN, MN-M
		2018	CDANs[97]	特征	–	Of, I-DA, O-H, MN, USPS, SVHN, VisDA
		2020	Cui等人[95]	特征	–	Of, O-H, VisDA
		2020	Xia等人[99]	特征	–	Of, O-H, I-DA
		2021	MetaAlign[100]	特征	–	Of, O-H
		2021	DDA[101]	特征	–	Of, VisDA
	输入空间	2016	CoGAN[45]	图片、特征	噪声→S, T	MN
		2017	Bousmalis等人[102]	图片	噪声, S→T	MN, MN-M, USPS
		2018	Sankaranarayanan等人[103]	图片	S, T→T	MN, USPS, SVHN, Of, VisDA
		2018	SBADA-GAN[104]	图片	S⇌T	MN, MN-M, USPS, SVHN, SYNS, GT
		2020	Ye等人[105]	图片、特征	T→T, S→S	MN, USPS, SVHN
其他		2016	DRCN[14]	–	–	MN, USPS, SVHN, CI, STL
		2016	DSNs[106]	–	–	MN, MN-M, GT, SVHN
		2016	CEAL[107]	–	–	Caltech-256
		2017	Saito等人[109]	–	–	MN, MN-M, SVHN, SYNN, SYNS, GT
		2021	PCS[110]	–	–	Of, O-H, VisDA

1) 此生成方向指在图像分类输入空间上使用的风格迁移网络中的生成器的生成方向, S代表源域, T代表目标域.

2) 数据集名称均为表5中数据集的缩写形式.

一些方法[30,31,40,81,82]使用最大平均偏差作为散度最小化两个域分布之间的距离. Tzeng等人[30]将基于MMD的域混淆损失用于一个全连接层以最小化不同域之间特征分布的差异. 随后, MMD的变体不断被提出以提升域自适应效果. Long等人提出MK-MMD[31]和JMMD[81], MK-MMD同时对分类网络中多个层进行

自适应[31]，JMMD则通过对齐不同域特定层的联合分布进行自适应[81]. 上述文献假设源域和目标域的分类网络存在直接的映射关系，而这一假设常常无法成立. 因此，Long等人[40]针对源域和目标域分别建不同的分类网络，并假设源域和目标域的分类网络之间的差别可用一个残差函数表示，同时使用MMD对齐多个网络层的输出特征. 而Rozantsev等人[82]则提出一个双流网络, 使源域和目标域的网络权重不共享, 然后通过一个损失函数关联两个域的网络权重.

相关对齐(CORAL)是另一种常用的测量域之间散度的方法，由Sun等人[41]首先提出，方法利用线性变换对齐源域和目标域特征空间的分布和特征值的二阶统计信息(协方差)以实现无监督域自适应. CORAL需要先提取特征，然后进行线性变换，之后再单独训练一个SVM分类器，因此CORAL不是一个端到端的学习. 为了能以端对端的方式进行训练，Sun等人[42]通过构造一个可微的损失函数，学习一种非线性转换，将CORAL扩展为Deep CORAL, 对齐深度神经网络中的层激活函数的统计信息. 之后，许多CORAL的变体也被陆续提出, 包括LogCORAL[83], Log D-CORAL[84]等.

近年来，其他计算不同域之间特征分布散度的方法也逐渐被提出. Courty等人[85]提出JDOT，将最优传输(Optimal transport)[86]的思想用于度量不同分布之间的距离. Damodaran等人[87]将JDOT引入到图片分类的无监督域自适应中，使用Wasserstein距离[88]度量不同域之间的散度. 然而, Damodaran等人[87]只进行了跨域匹配而忽略了域内结构，即当目标样本分布在聚类边缘或远离其对应的类中心时，很容易被源域学习的决策边界误分类. 针对此，Xu等人[89]提出一种加权最优传输策略，利用空间的典型信息和域内结构动态测量样本级域间差异，以实现精确的逐对最优传输，减少了在目标域决策边界附近的样本的负迁移. Das等人[90,91]提出基于图匹配的散度度量方法，将源域和目标域视为超图，利用图之间的一阶、二阶和三阶相似性进行类正则超图匹配. 考虑到以往的方法在最小化域差异时忽略了类的信息，这可能会使得类间对齐不对称，导致泛化性能差. Kang等人[92]提出了一种新的度量方式CDD (Contrastivedomain discrepancy)，方法明确建模了类内域偏移和类间域偏移，通过最小化类内域偏移以压缩类内样本的特征表示，最大化类间域偏移以使不同类离决策边界更远.

3.2　基于生成对抗的无监督域自适应方法

与基于最小散度的无监督域自适应方法不同，基于生成对抗思想的无监督域自适应方法通过训练一个域鉴别器进行分布对齐，一般在图像分类的特征空间和输入空间进行对抗学习, 进行域自适应.

3.2.1　图像分类特征空间

在图像分类特征空间进行对抗学习的方法中包含一个图像分类网络和一个域鉴别器. 其中, 域鉴别器用以判别分类网络提取的特征来自于源域还是目标域, 而分类网络则希望提取出域鉴别器无法判别来自某一域的特征. 两个网络通过不断对抗, 使得分类网络得以生成既能用于精准分类又不随域变化而变化的域不变特征. Ganin等人[26,29]和Tzeng等人[44]较早地将对抗思想引入图像分类任务中进行域自适应. Ganin等人[26,29]同时训练一个分类网络和一个基于逆梯度算法的鉴别网络来鉴别训练样本来自何域以进行域自适应. Tzeng等人[44]则同时利用域混淆损失和交叉熵损失，分别对齐不同域的边缘分布和各个类别. 为了学习更多的隐含信息，Tzeng等人[44]利用软标签代替具体的预测真值进行训练. 考虑到与对称映射相比，非对称映射能更好地提取不同域低层特征的差异，Tzeng等人[93]随后又提出一种分类编码网络权重不共享的对抗域自适应方法. 即与一些方法[26,29,44]在特征提取时各域分类网络权值共享不同，该方法中两个域的分类网络的编码器不共享权重，而是分别进行编码，允许对不同域间各异的特征提取.

考虑到仅使用一个域鉴别器判断提取的特征来自于何域可能会使得目标域中不同类之间的分界模糊，从而导致分类效果较差，Saito等人[46,47]对目标域的图像进行了两次不同的预测以提升分类效果. 这是通过使用分类网络的参数正则化(Dropout)来约束特征尽量远离分类的界线[46]，或使用两个分类器进行协同训练[47]来实现的. Sohn等人[94]针对源域和目标域可能不共享相同标签空间的情况，提出一个源域到目标域的特征转移网络，该网络将目标域特征与源域特征分离，并利用目标域特征与转换后的源域特征辅助域对齐. Cui等人[95]认为仅仅提取域不变特征会忽略不同域中独有的特征，因此提出一个对这种各异特征建模的模块, 分别用于生成器和鉴别器. 对于生成器, 这一模块能将源域和目标域都变换到中间域上，减小不同域直

接迁移的难度，同时减小提取的域不变特征中残留的独有特征的影响；对于鉴别器，这一模块有助于提高鉴别器鉴别能力，平衡对抗训练过程. 由于生成对抗思想与GAN有着密切的联系, 因此新的GAN方法的提出往往能为基于生成对抗思想的无监督域自适应方法中的目标函数设计提供新思路[93]. Shen等人[96]受Wasserstein GAN启发, 计算源域特征和目标域特征之间的Wasserstein距离作为鉴别器的目标函数. 而Long等人[97]受CGAN[98]的启发，提出条件域自适应方法以适应分类问题中的多模态分布.

一些方法[99,100]开始将图论和元学习方法引入域自适应中. Xia等人[99]提出一种基于图的度量方法，图的边被用来提取和对齐不同域之间的匹配关系，而节点根据学到的域间的匹配关系来约束特征提取，从而实现特征对齐. Wei等人[100]引入元学习以缓解域对齐任务和分类任务之间优化目标不相同的问题，将领域对齐任务和分类任务视为同一组样本的元训练和元测试. 之后, 一些方法开始关注实时性, Li等人[101]在低计算资源条件下以较高精度快速实现了图像分类的域自适应任务，并同时提出置信分数学习和类间平衡自监督训练策略以提升在目标域上的分类精度.

3.2.2 图像分类输入空间

基于最小散度的方法和在特征空间上进行对抗的方法都是在特征空间进行域自适应的. 随着GAN的不断发展，一些工作使用GAN进行图片生成和图片风格迁移以实现数据增强, 生成更多的训练数据. 在数据增强的过程中，鉴别器用来判断图像是真实图片还是由生成器生成的，而生成器则希望生成的图片尽可能接近真实图片, 从而能够骗过鉴别器.

一些方法使用GAN在输入空间上利用噪声和图片特征生成图片[45,102,103]或在域间进行图片风格迁移[104]. Liu等人[45]提出包含两个GAN的CoGAN, 其中, 两个生成器分别从噪声生成源域图像和目标域图像, 训练时两个GAN通过共享部分网络参数学习不同域图像的联合分布从而使得分类网络能够提取域不变特征. Bousmalis等人[102]利用GAN从源域图片和噪声生成目标域风格的图片，以此进行数据增强，用于和源域图片一起训练适用于源域和目标域的图像分类网络. Sankaranarayanan等人[103]分别提取源域和目标域图片的特征用于生成源域风格的图片，鉴别器则用来区分输入是真实的源域图片还是由特征生成的源域图片，以此进行对抗学习使得编码器能提取出域不变特征. Russo等人[104]利用GAN在源域和目标域间进行图片风格迁移，并针对目标域和源域分别训练一个分类网络，最终利用这两个分类网络同时预测来自目标域的图像类别. 以上的域自适应方法旨在学习域不变特征，这些方法通常需要根据目标域的数据更新源域分类器以适应目标域，因此有时不能同时在源域和目标域中都取得不错的效果. 针对此，Ye等人[105]不再训练分类器以适应目标域, 而是使用一种数据校准器, 在保持源域分类性能的同时提升目标域的分类性能.

3.3 基于其他方法的无监督域自适应方法

本小节中介绍的方法既不基于最小化散度，也没有利用生成对抗思想，而是使用图像重构或自监督训练的方法进行域自适应. Ghifary等人[14]和Bousmalis等人[106]通过提取图片特征并重构图片进行域自适应. 特别地，Ghifary等人[14]同时训练源域图片的分类网络和目标域图片的重构网络. 由于两个网络的编码器的权值是共享的，因此编码器需要提取域不变特征，分别适应分类和重构网络，从而起到了特征对齐的作用. Bousmalis等人[106]与Ghifary等人[14]的方法基本一致，但考虑到不同域特征的独特性，Bousmalis等人提出域分离网络，分别提取源域和目标域的共有特征和私有特征，并利用同一个解码器同时使用共有和私有特征重建输入图像.

还有一些方法将自监督训练引入到图像分类无监督域自适应中. Wang等人[107]将熵作为置信度对目标域图像进行排序，分为低置信度和高置信度两类，通过置信度从高到低逐步标注伪标签，以增量学习的方式训练分类器. 协同训练(Co-training)和三协同训练(Tri-training)使用多个分类器自动标注未标注的样本，然后将这些新标注的训练数据用于分类网络的再训练[108]. 这种方法一般假设待标定的数据来自同一个域，而域自适应问题中的数据来自不同域，针对此问题，Saito等人[109]提出了一种不对称的三协同训练方法. 网络包含三个分类网络，其中两个用于有标签的源域和有伪标签的目标域的分类，而另一个仅对有伪标签的目标域分类. 自监督训练的使用使得网络通过学习有标签的源域信息，为无标签的目标域分配伪标签进行训练. Yue等人[110]考虑到源域的有标注的样本

可能也很少，因此提出基于小样本的无监督域自适应方法，即只需要少量的有标签的源域样本和大量无标签的源域和目标域数据进行训练. 此外，Yue等人还通过提取域内典型特征的方法学习类别级的语义信息，并通过域间典型实例的自监督进行特征对齐.

图像分类任务中的无监督域自适应方法的发展相对来说已经比较成熟，其基本的思想也逐渐被应用于其他视觉任务中. 虽然不同计算机视觉任务的具体算法各异，但是对于无监督域自适应来说，不同视觉任务中的方法和经验可以相互借鉴.

4　目标检测

相比于图像分类任务，无监督域自适应方法在目标检测任务中的研究仍处于起步阶段，因此相关的文献较少. Faster R-CNN[8]是目前目标检测方向精度较高、应用广泛的方法之一，本节介绍的无监督域自适应目标检测算法也基本都基于Faster R-CNN框架. 本节将目标检测无监督域自适应的方法分为两类进行介绍：基于生成对抗思想和基于其他方法的无监督域自适应方法, 典型的方法概括在表7中.

4.1　基于生成对抗的无监督域自适应方法

与图像分类任务一样，目标检测任务中的基于生成对抗思想的无监督域自适应方法也可分为两类，即在特征空间和输入空间进行域自适应的方法.

4.1.1　目标检测特征空间

在目标检测特征空间进行无监督域自适应的算法在经典目标检测构架中加入域鉴别器进行对抗学习. Chen等人[32]较早地在Faster R-CNN构架中加入域鉴别器进行目标检测域自适应. 由于目标检测需要对图中的一个或多个物体进行识别和定位，因此域偏移可能同时发生在图像层面(如图像风格、光照等)和实例层面(如对象外观、大小等)，Chen等人[32]针对图像和实例分别设计了域鉴别器[29], 从两个层面分别进行域自适应.

分类任务的域自适应常常将图像作为一个整体对齐不同域的特征分布，而目标检测任务需同时识别和定位多个物体，因此目标检测的自适应需要更关注包含目标的局部区域，一些方法[111,112]尝试从这些局部区域提取更多的有用信息. He等人[111]提出包含较低层次特征的分层对齐模块和较高层次聚合特征的对齐模块的一个多层对抗域自适应框架. 主体网络中含三个域鉴别器用于三个不同层次特征的对抗学习，同时使用梯度反转层和尺度缩减层以减小特征尺度；检测器部分将检测结果聚合进行语义对齐，将聚合后的特征输入域鉴别器并加入权梯度反转层以惩罚难混淆的样本. Zhu等人[112]将域自适应问题分为两步进行, 即包含目标的局部区域的寻找和对齐，方法的关键在于挖掘与目标检测直接相关的区域并对齐. 随后，Xu等人[113]认为在目标检测中除了需要关注图片中有目标的重点区域, 还需要关注较难进行域自适应的目标. 因此, 提出一个分类一致性正则化项, 同时在图像级和实例级进行目标检测，将检测的一致与否作为判断目标检测难易的标准.

由于目标检测中不同域中的目标大小和分布情况可能不同, 强行匹配会改变语义, 适得其反; 而纹理和颜色等局部特征的强匹配是有意义的，因为它们不会改变语义层次. 基于此，Saito等人[114]提出了一种基于强局部对齐和弱全局对齐的目标检测域自适应方法，在低层次特征和高层次特征两个层面进行域自适应，即在纹理、颜色等低层次特征进行强对齐(交叉熵损失)，在涉及语义信息的高层次特征进行弱对齐(对难以分类的样本进行部分对齐). 目标检测任务中，同一张图片中包含的目标类别往往各不相同且不同类别的目标样本分布各异. 之前的一些工作[32,111,112,114]没有考虑到这些信息，而是把图片中不同种类目标的分布当作整体进行域自适应，因此Zheng等人[115]提出一种两阶段的域自适应方法，从粗粒度到细粒度进行特征对齐. 在粗粒度自适应阶段，采用注意力机制提取前景目标并将其视作一个整体在多层网络中对齐其边缘分布. 在细粒度阶段，对于前景目标中的所有不同类别，按类别分别对齐来自不同域的前景目标. 为了减小域间的错误类别对齐，VS等人[116]使用多个鉴别器来对齐不同域中语义相同的类别. VS等人提出有记忆的注意力机制以区分不同类别的目标，从而将不同种类的目标送到相应的鉴别器中. 而Xu等人[117]首先建立不同候选框之间的图关系以获得更精确的实例特征表示，然后通过增强类间的紧凑度和类间的可分离度来分别对齐不同目标类别. 此外，Xu等人还通过赋予出现频次较少的类别较高权重来提升这类目标的自适应效果.

表 7 无监督域自适应方法在目标检测中的应用

Table 7 Object detection with UDA

分类		年份	方法	鉴别器输入	生成方向[1)]	数据集[2)]
生成对抗	特征空间	2018	Chen等人[32]	特征	–	SIM→C, C→FC, K→C
		2019	He等人[111]	特征	–	SIM→C, C→FC, K→C
		2019	Zhu等人[112]	特征	–	SIM→C, C→FC, K→C
		2019	Saito等人[114]	特征	–	C→FC, G→C 等
		2020	Xu等人[113]	特征	–	C→FC 等
		2020	Zheng等人[115]	特征	–	SIM→C, C→FC, K→C
		2020	Xu等人[117]	特征	–	SIM→C, C→FC, K→C
		2021	MeGA-CDA[116]	特征	–	C→FC, SIM→C, K⇌C
	输入空间	2019	Kim等人[48]	图片	S→其他风格	C→FC 等
其他		2019	Roychowdhury等人[118]	–	–	WIDER→CS6, BDD[3)]
		2021	ST3D[119]	–	–	Waymo→K, Waymo→Lyft, Waymo→nuScenes, nuScenes→K[3)]

1) 此生成方向指在图像分类输入空间上使用的风格迁移网络中的生成器的生成方向, S代表源域, T代表目标域.

2) 数据集名称均为表5中数据集的缩写形式.

3) 人脸识别数据集WIDER[120]、CS6[121], 行人检测数据集BDD[122], 无人驾驶数据集Waymo[123]、nuScenes[124]、Lyft[125], 由于这些数据集在本综述中仅出现一次, 因此未在表2中列出.

4.1.2 目标检测输入空间

在目标检测输入空间进行域自适应的算法除了使用域鉴别器和目标预测任务网络进行对抗学习, 还加入了额外的GAN进行数据增强. Kim等人[48]使用CycleGAN进行数据增强, 生成源域风格到目标域风格之间的多个风格的图像, 且这些图像与源域共享标签, 这样则获得了更多新的有标签的训练数据. 特别地, 为了学到多个域的不变特征, 这里的域鉴别器进行多分类.

4.2 基于其他方法的无监督域自适应方法

一些方法使用自监督训练的方法进行目标检测的域自适应, 通过在源域上训练的模型为目标域生成伪标签以使得目标域以有监督的方式进行训练. Roychowdhury等人[118]利用在源域上训练好的检测器和跟踪器为目标域中的图片生成目标检测任务生成伪标签. 其中, 使用跟踪器是为了避免检测器漏掉目标域视频中的某些目标. 此外, 本方法还使用了知识蒸馏的思想. 随后, 研究人员开始探究3D目标检测的域自适应. Yang等人[119]首先提出随机对象缩放策略, 在源域上对三维检测器进行预训练, 以减轻源域偏差的负面影响. 然后, 在目标域上交替使用两种策略改进检测器, 即伪标签更新和利用课程数据增强进行模型训练.

相比于图像分类任务, 目标检测域自适应的研究方法较少, 仍处于起步阶段. 从目前的发展态势来看, 目标检测任务中使用的域自适应方法与图像分类任务中使用的域自适应方法的基本思想相似. 在未来的研究中, 可继续借鉴其他发展较为成熟的计算机视觉任务中使用的无监督域自适应方法, 结合目标检测任务自身进行域自适应时遇到的问题, 提升目标检测模型的迁移性.

5 语义分割

随着图像分类无监督域自适应方法的发展成熟, 人们开始研究像素级视觉任务的域自适应问题. 本节对语义分割中的无监督域自适应方法做了分类介绍和总结, 归纳在表8中. 由于重构思想和伪标签被广泛应用于语义分割无监督域自适应方法中, 表中还总结了各方法是否使用重构思想和伪标签. 在本节中, 源域数据大都来自于虚拟数据集, 含有语义标签, 而目标域数据来自于真实数据集, 不含语义标签.

5.1 基于生成对抗的无监督域自适应方法

生成对抗思想也被广泛应用于语义分割任务的无

表 8　无监督域自适应在语义分割中的应用

Table 8　Semantic segmentation with UDA

分类		年份	方法	鉴别器输入			生成方向[1)]	重构[2)]	伪标签	特点	数据集[3)]
				输入	特征	输出					
生成对抗	特征空间	2016	FCNWild[27]	–	全局、类别	–	–	–	–	–	G, S→C, C→B
		2017	Chen等人[74]	–	全局、类别	–	–	–	√	–	S→C, C→N
		2018	Huang等人[126]	–	全局(逐层)	–	–	–	–	–	G, S→C
		2018	ROAD[129]	–	空间	–	–	–	–	知识蒸馏	G→C
		2019	SSF-DAN[127]	–	√	–	–	–	√	–	G, S→C, C→N
		2019	SIBAN[131]	–	√	–	–	–	–	信息瓶颈	G, S→C
		2020	CSCL[128]	–	全局、类别	–	–	–	√	考虑负迁移, 类对齐, 强化学习	G, S→C, C→N
		2020	Pan等人[130]	–	√	–	–	–	√	熵值排序, 课程学习	G, S, Y→C
	输出空间	2018	AdaptSegNet[132]	–	√	√	–	–	–	–	G, S→C, C→N
		2019	Tsai等人[133]	–	–	图片块	–	–	–	–	G, S→C, C→O
		2019	ADVENT[134]	–	–	√	–	–	√	熵最小化	G, S→C
		2019	DADA[135]	–	–	√	–	–	–	深度信息	S→C, M
		2019	Luo等人[136]	–	–	类别	–	–	–	协同训练	G, S→C
		2020	FADA[137]	–	–	类别	–	–	–	–	G, S→C, C→N
		2020	SIM[138]	–	–	√	–	–	√	前背景分离, 类对齐	C→C, S→C
		2021	SFDA[139]	–	√	图片块	–	–	√	源域数据不可知	G, S→C, C→N
	输入空间	2018	FCAN[49]	–	√	–	S⇌T	–	–	–	G→C, B
		2018	DCAN[140]	–	–	–	S→T	–	–	双重对齐	G, S→C
		2018	Zhu等人[141]	√	–	–	S⇌T	–	–	损失函数	G, S→C
		2018	Sankaranarayanan等人[143]	√	–	–	S⇌T	√	–	–	G, S→C
		2018	CYCADA[146]	√	√	–	S⇌T	√	–	–	G→C, S
		2018	I2I Adapt[148]	√	√	–	S⇌T	√	–	–	G→C
		2019	DISE[144]	√	√	√	S⇌T	√	–	特征提取	G, S→C
		2019	CrDoCo[147]	√	√	–	S⇌T	√	–	深度, 光流	G, S→C, C→N
		2019	Li等人[149]	√	√	–	S⇌T	√	√	双向	G, S→C
		2019	TGCF-DA[151]	√	–	–	S→T	–	√	自集成, 教师-学生网络	G, S→C
		2019	MADAN[153]	√	√	–	S⇌T	√	–	多源	G, S→C, B
		2019	GIO-Ada[154]	√	–	√	S→T	–	–	深度辅助语义	vK→K, S→C
		2020	RPT[142]	–	√	–	S⇌T	–	–	语义约束	G, S→C
		2020	Yang等人[145]	√	–	√	T→S	√	√	语义重构图片	G, S→C
		2020	Kim等人[152]	–	√	√	→S	–	√	多纹理, 自监督学习	G, S→C
其他方法	课程学习	2017	Curriculum[155]	–	–	–	–	–	–	–	G, S→C
		2018	CMAda[156]	–	–	–	–	–	–	有雾场景	
		2019	Sakaridis等人[157]	–	–	–	–	–	–	光照变化	
		2019	PyCDA[158]	–	–	–	–	–	√	自监督学习	G, S→C
	自监督	2018	CBST[159]	–	–	–	–	–	√	类平衡	G, S→C, C→N
		2018	Saleh等人[160]	–	–	–	–	–	√	前背景分离	
		2019	CRST[161]	–	–	–	–	–	√	置信度正则化	G, S→C
		2020	MLSL[163]	–	–	–	–	–	√	空间独立, 语义一致	G, S→C
		2020	CCM[164]	–	–	–	–	–	√	考虑负迁移	G, S→C
		2020	IAST[165]	–	–	–	–	–	√	对抗学习, 实例	G, S→C
		2021	MetaCorrection[162]	–	–	–	–	–	√	元学习、噪声建模	G, S→C
		2021	Ma等人[166]	–	–	–	–	–	√	光度矫正	G, S→C
		2021	MSDA[167]	–	–	–	–	–	√	多源域	G, S, Y, G+S, G+Y, S+Y, G+S+Y→C
	熵最小化	2019	Chen等[170]	–	–	–	–	–	–	最大平方损失	G, S→C, C→N
		2020	CrCDA[171]	–	–	–	–	–	–	上下文信息	G, S→C

1) 此生成方向指在图像分类输入空间上使用的风格迁移网络中的生成器的生成方向, S代表源域, T代表目标域.

2) 这里的重构不包括在输入空间使用风格迁移网络重构图像.

3) 数据集名称均为表5中数据集的缩写形式.

监督域自适应中. 这类方法一般包含两个网络, 分别用于预测输入图像的语义分割结果和作为域鉴别器判别输入数据来自源域还是目标域. 两个网络通过对抗学习使语义分割网络能够提取域不变特征(使从两个域提取的特征具有相似的分布), 最终使得域鉴别器无法分辨出输入数据来自于何域. 按照鉴别器输入数据的不同, 可分为在特征空间对抗和在输出空间上对抗两类. 还有一些方法通过一个额外的图片风格迁移网络实现数据增强, 称为在语义分割输入空间上进行域自适应的方法. 本小节对在语义分割特征空间、输出空间和输入空间进行生成对抗学习的方法分别进行介绍.

5.1.1 语义分割特征空间

在语义分割的特征空间进行生成对抗学习的方法将语义分割网络从不同域提取的特征输入域鉴别器, 通过域鉴别器和语义分割网络的对抗学习进行域自适应. 语义分割任务的域偏移主要来自两方面, 不同域之间整体外观的差别和不同域之间不同语义类别分布的差异(甚至语义类别不同), 两者共同导致了不同域特征空间的边缘分布不一致[27]. 基于此, 一些方法[27,74]分别从全局特征和不同语义类别特征两个层面进行对抗学习. Hoffman等人[27]除进行全局特征对齐外, 还将源域上不同语义类别的分布作为先验迁移到目标域上, 对目标域上对应的语义类别进行域自适应. 而Chen等人[74]在目标域使用网格级的伪标签代替像素级的伪标签进行对抗学习以实现类内对齐, 同时提出NTHU数据集, 并利用其中包含的时间信息确定场景中的静态物体作为对抗学习的先验信息. 随后, Huang等人[126]认为仅对齐语义分割网络中的某一层输出特征是不够的, 因此其通过对齐语义分割网络中的多个中间网络层的特征引入了更多约束, 同时上层网络输出的特征经过对齐被输入到下层网络中, 也降低了下层网络输出特征的对齐难度.

由于Chen等人[74]在目标域中使用全部的伪标签进行对抗学习, 这导致网络的训练无可避免地会受错误伪标签的影响. 因此, Du等人[127]仅使用置信度较高的像素区域作为伪标签, 并分别对不同的语义类别使用不同的域鉴别器, 此外还使用渐进置信策略以解决语义级对抗学习中不同域之间语义标签不一致的问题. 类似地, Dong等人[128]考虑到了负迁移的存在, 指出不是所有的信息都是可迁移的, 强制性迁移所有知识只会造成更差的结果. 针对负迁移, Dong等人提出一个基于批判性迁移的对抗框架以突出可转移的知识而忽略不可转移的知识, 对目标样本生成基于置信度的伪标签, 并利用强化学习使正迁移收益最大化.

除在GAN中加入自监督训练的方法外, 一些工作还引入了知识蒸馏[39]、熵[38]、课程学习[35]等的概念. Chen等人[129]使用知识蒸馏方法以避免语义分割模型在合成数据集上过拟合. 此外, 提出一种空间可知的域自适应方法, 通过将训练图片拆分成不同块, 利用图片不同区域块中蕴含的内在几何关联, 对图片块分别进行对抗学习. Pan等人[130]不仅考虑了不同域之间的差异, 还考虑了目标域中不同图片之间的差异, 利用基于熵的排序函数, 将目标域分为简单的和困难的两部分进行课程学习. 除了考虑如何进行特征对齐, 对抗训练能否达到稳定也逐渐被关注. 与语义分割网络相比, 域鉴别器在对抗学习中更占优势, 因此语义分割网络和域鉴别器的对抗学习很难达到稳定. 鉴于此, Luo等人[131]利用信息瓶颈(Information bottleneck)方法压缩复杂的特征, 仅提取与语义分割任务相关的特征, 以使对抗学习更容易达到平衡. 此外, 为平衡不同语义类别, 保证出现频次较少的类也能顺利迁移, Luo等人还提出了基于重要性的信息瓶颈方法用于对抗训练.

在本小节中, 语义分割任务通过特征对齐来实现域自适应. 然而, 由于语义分割任务需要考虑图像形状、纹理、上下文等的复杂信息, 仅利用特征对齐进行域自适应往往不能取得满意的效果. 且语义分割任务在特征空间进行对齐存在两点挑战. 首先, 对齐不同域特征的边缘分布难以约束语义的一致性, 举例来说, 目标域上对应于汽车的特征可能映射为源域上对应的自行车的特征. 其次, 高层的特征对齐无法建模低层外观差异. 因此, 由于语义分割任务依赖图片风格、纹理、结构等信息, 仅在特征层面进行域自适应的效果并不理想[49]. 由于仅仅利用特征对齐难以在语义分割任务上获得较好的结果, 很多方法开始探究在语义分割输出空间和输入空间上进行域自适应.

5.1.2 语义分割输出空间

注意到语义分割的输出空间(语义分割结果)包含了丰富的结构信息, 例如, 即使来自于不同域的图像外观非常不同, 它们的语义分割结果也可能很相似. 因

此, 可以在语义分割输出空间进行域自适应, 使源域和目标域语义预测的标签分布尽可能相似[132,133]. 与在语义分割特征空间上进行对抗生成学习相同, 在语义分割输出空间上进行对抗生成学习的方法一般也包含两个网络, 一个输出语义分割结果的语义分割网络和一个用于鉴别语义分割结果来自源域图片还是目标域图片的域鉴别器. 通过对抗学习, 语义分割网络的目的是输出在目标域和源域上的分布尽量相同的语义分割结果, 以混淆域鉴别器. 本小节提及的方法都在语义分割的输出空间上进行了对抗学习, 同时这些方法大多也同时在语义分割的特征空间上进行了对抗学习, 具体在何空间上进行域自适应可由表8中的鉴别器输入区分.

Tsai等人[132,133]较早地在语义输出空间上使用对抗思想, 以约束模型在目标域和源域上输出分布相同的语义分割结果. 特别地, 文献[132]还对语义分割网络中的两个层进行特征层面的域自适应, 文献[133]对语义分割结果分块(Patch-wise), 先对源域块的语义分割结果聚类, 然后利用对抗学习, 使目标域预测的语义块的聚类结果和源域的聚类结果尽可能相近, 以此实现域自适应. 之后, Vu等人[134,135]尝试将其他损失和信息融入语义分割域自适应框架中. 文献[134]结合对抗损失和熵损失以最小化目标域预测结果的熵值进行无监督域自适应, 文献[135]利用源域上的场景深度信息作为额外的监督信号, 这是因为域偏移不会影响场景深度信息在空间结构上的一致性, 方法通过同时预测源域和目标域上的场景深度和语义进行对抗学习.

与Chen等人[74]相似, 考虑到语义分割中不仅需要对全局特征的边缘分布对齐, 还需要对不同类别标签的条件分布分别进行对齐, Luo等人[136]结合对抗学习和协同训练[108]提出一个类别级对抗网络, 使用源域和目标域的数据联合训练两个语义分割网络, 以检测不同域的语义分割类别是否得以对齐. 和Luo等人[136]通过协同训练间接对齐语义类别不同, Wang等人[137]直接将类别信息输入域鉴别器, 提出了一种细粒度的类别级特征对齐的对抗学习策略, 同时保留了跨域语义分割的结构信息. 由于背景与实例相比, 域之间的差异较小, Wang等人[138]在进行域自适应时对实例和背景采取不同的策略. 方法分别对每个背景类进行对齐操作, 而由于不同实例之间差异较大, 因此对于每个不同的实例类别, 只要求目标域中的实例与源域中最相似的实例保持一致以考虑实例的个体差异的影响, 防止过度对齐. 考虑到有些情况下源域数据较难获得, Liu等人[139]仅使用在源域上训练的模型和目标域数据进行无监督域自适应. 该方法分为知识迁移和模型自适应两部分, 以从源模型中恢复和保存源域信息, 并从目标领域中提取有价值的信息进行自监督训练.

5.1.3　语义分割输入空间

除了在特征空间和输出空间使用域鉴别器进行对抗学习, 一些工作通过在输入空间上使用额外的GAN在源域和目标域图片间进行风格转换, 实现数据增强, 以提升语义分割的域自适应效果. 由于这些方法在进行域自适应任务前, 先在输入空间上进行图像风格迁移, 因此我们将这些方法归为在语义分割输入空间上进行对抗学习的方法. 要注意的是, 本小节中介绍的方法也在特征或输出空间上进行了域自适应. 通过风格变换, 可以得到有着目标域风格的带语义分割标签的数据, 从而起到了数据增强的效果.

Zhang等人[49]较早将图片风格迁移用于语义分割域自适应, 除了利用GAN在源域和目标域之间进行风格转换, 还通过额外的对抗网络学习域不变特征. 同时其他一些工作, 通过改善特征对齐方法[140]或损失函数[141]提升域自适应的效果. 由于图片每个信道的特征图的统计数据中都蕴含着与图片风格相关的信息, 因此可以分别利用每个信道的信息进行图片风格转换. 基于此, Wu等人[140]在GAN和语义分割网络中分别采用了基于信道的特征对齐方法, 利用信道中包含的图片风格和结构信息, 提升域自适应结果. 研究发现, 当网络在源域上表现很好时, 模型在源域过拟合, 导致网络迁移能力不好; 而在源域上表现很差时, 网络又不能很好地完成语义分割任务. 因此, Zhu等人[141]认为, 当网络在源域上的语义分割结果即将接近最好却不曾达到时, 可以避免网络过拟合又能获得较好的语义分割结果. 这是通过在对抗学习中使用一种鲁棒损失函数, 用保守损失(Conservative loss)代替交叉熵损失, 对源域上语义分割结果很好或很差的情况都进行惩罚, 而鼓励表现折中的情形实现的. 后来, Zhang等人[142]在文献[49]的基础上, 提出了多层次的语义预测一致性约束, 包含图片块级、聚类级和上下文语义预测一致性. 首先约束每个图片块上的语义预测相近, 然后对图片块聚类, 使每个类上的语义信息相近, 最后通过

上下文的逻辑信息约束图像的语义趋于一致.

在一些方法[143~145]中, 生成器不仅被用来进行图片风格转换, 而且被用来重构图片, 然后将原图片和重构后的图片送入鉴别器进行对抗学习. Sankaranarayanan等人[143]通过编码器提取源/目标域图像的特征, 然后使用生成网络从特征空间上重建源/目标域的图像, 鉴别器则用来判别输入的图像是源/目标域图像还是生成器生成的图像, 以此进行生成对抗训练. 与Bousmalis等人[106]相似, Chang等人[144]将图像的结构和纹理信息解耦, 即分别提取源域和目标域的域不变特征和各自有代表性的特征, 以此进行组合, 进行图片重建和风格转换. 与Sankaranarayanan等人[143]和Chang等人[144]利用图片重建图片不同, Yang等人[145]利用语义分割网络输出的分割结果重建图片. 具体来说, 先将目标域图片转换成有着源域风格的图片, 然后将目标域和有着源域风格的目标域图片输入语义分割网络预测语义分割结果, 最后利用不同域的语义分割结果重建图片.

为了确保图片风格迁移后语义信息仍被保留, 一些方法[146~149]使用CycleGAN[150]中的循环一致损失作为额外约束. Hoffman等人[146]同时在输入空间和特征空间上进行域自适应, 先使用CycleGAN进行图片风格转换, 将所有图片都用于训练, 然后提取特征进行对抗训练. 考虑到尽管目标域和源域在图像风格迁移前后的风格可能存在差异, 但同一张图片迁移前后的语义输出结果也应当一致. Chen等人[147]引入跨域一致性损失, 此框架不仅可以很好地用于语义分割, 还能在深度估计和光流估计中获得很好的性能. Murez等人[148]结合了多种方法[14,27,93,103,150]中的损失函数, 以使得从不同域中提取的特征是不可区分的, 并利用这些特征同时在两个域上重建图像. 上述一些方法[140,146]先使用GAN进行图像风格迁移, 然后再进行对抗训练, 因此这些方法中的风格迁移模型在语义分割过程中不再参与训练, 两个过程是割裂开来的. 这就导致如果预先训练好的风格迁移网络性能不佳, 语义分割网络的分割结果也会受影响. 基于此, Murez等人[148]还通过语义分割模型提升风格迁移网络的性能. 而Li等人[149]提出一个双向学习的网络框架, 交替训练图像风格迁移网络和语义分割网络, 形成闭环, 两个网络相互提升. 特别地, Li等人[149]还使用了自监督训练的思想, 用有伪标签的目标域数据微调在源域上训练出的语义分割模型, 从而也提升了图像风格迁移模型的效果.

一些方法[151,152]将知识蒸馏、自监督训练的方法加入到生成对抗学习中. Choi等人[151]使用知识蒸馏的思想, 通过一个教师-学生网络提升语义分割网络在目标域上的性能. Kim等人[152]利用图片风格迁移算法改变合成图像的纹理, 得到风格多样的合成图片. 通过使用风格多样的合成数据训练, 可以防止语义分割模型过拟合到一个特定的(合成)纹理, 最后通过自监督训练得到目标域的伪标签对模型进行微调. 考虑到现有的方法一般利用单个源域进行域自适应, 而现实生活中域是复杂多变的, Zhao等人[153]利用多个源域进行训练. 对于计算机引擎来说, 同时生成场景语义和深度标签也比较容易, 且场景深度信息中包含的几何约束受域偏移的影响较小. 因此, Chen等人[154]提出了一种利用场景深度信息辅助语义分割的域自适应方法, 在输入和输出空间上同时利用场景深度信息进行对抗学习以减少域偏移: 在输入空间上, 同时利用语义、场景深度和图片信息进行图片风格迁移, 将合成图像转换为真实图像的风格; 在输出空间上, 同时输出语义分割和深度估计结果, 利用合成和真实图片生成的语义分割和深度估计结果进行对抗性训练, 保持语义与深度估计的相关性.

5.2 基于其他方法的无监督域自适应方法

除了使用生成对抗思想进行语义分割无监督域自适应外, 其他一些方法, 如课程学习、自监督学习、熵最小化等也被应用于语义分割无监督域自适应中, 本小节对这些方法进行了概括总结.

5.2.1 课程域自适应

Zhang等人[155]提出了一种基于课程学习的域自适应方法并用于语义分割. 先学习目标域中两个“简单”的任务: 所有图像的语义标签分布和图片中特征显著的局部区域(对于城市街道场景来说, 建筑、街道、汽车等的大小和空间关系特征较为显著)的语义标签, 然后在学好的“简单”任务的基础上, 学习比较“困难”的像素级语义分割任务. 课程学习除可被用于合成与真实数据集之间的域自适应, 也可用于进行不同环境间的域自适应. Sakaridis等人[156]将课程学习用于有雾场景的语义分割, 先学习有标签的合成薄雾场景数据集

中的语义分割，然后将学到的知识应用于无标签的真实浓雾场景下的语义分割中. 随后，Sakaridis等人[157]尝试在有标签的光照良好的场景下训练语义分割模型, 并迁移到无标签的夜晚场景中. 通过先学习较为简单的光照条件良好的情况下的图片，然后在此基础上，再学习更困难的夜晚场景下的语义分割. 考虑到课程学习和自监督训练有很多相似点，Lian等人[158]结合课程学习和自监督训练进行域自适应. 首先, 构建了一个包含目标域图像、图像块和像素的标签分布的金字塔型课程学习. 然后, 使用自监督训练预测此金字塔模型的语义输出结果. 下一小节具体介绍了自监督训练与无监督域自适应的相结合的方法. 此外, 一些方法将课程学习的思想用于对抗模型中, 如上文提及的文献[130].

5.2.2　自监督训练

自监督训练也被用于语义分割中进行域自适应. Zou等人[159]提出了一种基于自监督训练的无监督域自适应框架, 在目标域上迭代生成伪标签, 并使用这些伪标签对模型进行再训练来实现域自适应. 同时, 为了避免比重大的类别标签逐渐在伪标签的生成上占据绝对优势，Zou等人还提出了一个类平衡自监督训练框架, 引入空间先验来优化生成的标签. 考虑到图片背景在不同的域中相似，而实例则在不同域之间有较大的差异，Saleh等人[160]对实例和背景采取不同的策略进行域自适应. 在使用源域数据训练网络时，使用一般的语义分割网络处理背景，而对前景进行实例分割，然后使用在源域上预训练好的模型输出目标域的伪标签. 由于伪标签中常常存在噪声, 使得某些类会被赋予置信度较高的错误标签不断迭代. 因此, Zou等人[161]提出了一种置信度正则化的自监督训练框架. 与文献[159]中使用独热编码作为伪标签不同，Zou等人[161]把伪标签看作是连续的概率值, 进行交替联合优化. Guo等人[162]引入噪声转移矩阵来建模目标域内伪标签的噪声分布以缓解伪标签中噪声的影响，并使用元学习策略缓解伪标签中的噪声问题，提高模型对目标数据的泛化能力.

考虑到上述方法不能利用全局上下文信息，Iqbal等人[163]先将图像分成小块, 再对语义分割结果相同的不同小块施加相同的约束，从而生成空间独立、语义一致的伪标签. 此外，此方法在特征空间上设计了一种全局语义一致的伪弱标签，使得网络能够学习到占比较小的语义标签. Li等人[164]首先利用源域图片训练一个初始语义分割模型，并使用该模型生成目标域图片的伪标签，然后利用源域的准确标注和目标域的伪标签，从水平和垂直两个方向构造图像的语义布局矩阵，进行矩阵匹配，找出与目标域语义布局相似的源域图像. 最后，选择置信度较高的伪标签进一步进行迭代训练. 为了有效地提高伪标签的质量，Mei等人[165]设计了一个带实例自适应选择器的伪标签生成策略，同时提出了伪标签区域平滑和非伪标签区域锐化策略. 与在输入空间上使用风格迁移网络的方法[49,140]不同，Ma等人[166]提出光度矫正模块，在LAB色彩空间上对齐不同域间各个通道的色彩和亮度的分布, 同时提出两个类别级的正则化方法, 使源域类别间距离比类别内距离更大，并约束目标域中矫正后的目标图像的预测与其对应的矫正前的伪标签一致. 同样地，He等人[167]也首先在LAB色彩空间上进行图片风格变换, 对齐源域和目标域. 特别地, He等人还同时利用多个源域信息与目标域协同训练以提升模型的迁移精度.

5.2.3　熵最小化

熵最小化[38]是半监督学习[168,169]中最流行的方法之一. 然而, 当将熵最小化方法用于语义分割的无监督域自适应时，方法往往会更关注目标域中容易迁移的样本, 使得较难迁移的样本训练不充分. 因此, Chen等人[170]提出了最大平方损失以解决这个问题, 同时方法还引入了图像级权重因子用于缓解目标域中语义类别不平衡的问题. Huang等人[171]通过匹配不同域图片间的局部区域，利用其上下文关联，使源域与目标域的局部语义趋于一致来提升域自适应结果，以最大最小化熵值的对抗形式训练模型.

本节对语义分割中的无监督域自适应方法做了分类介绍. 最初的相关工作仅在特征空间上使用域鉴别器，但由于像素级任务预测量较大，且预测输出高度结构化，仅在特征空间上进行域自适应较为困难[49,132]. 鉴于此，很多学者开始探索其他空间上的域自适应, 包括语义分割的输出空间和输入空间. 在语义分割输出空间上进行域自适应利用了语义分割结果通常在全局空间分布相似性较大，上下文关联紧密的特点[132]. 而在语义分割输入空间进行域自适应的一般想法是对源域和目标域的图像进行风格转换，从而起到

数据增强的作用，减小域偏移，这一般通过GAN实现[49]. 相比于其他视觉任务，语义分割域自适应的研究较多，可以考虑将其中的一些经典方法扩展到其他任务中.

6 深度估计

与语义分割任务相似，深度估计任务的无监督域自适应主要也依赖于生成对抗的思想，本节对深度估计中的无监督域自适应方法做了分类总结和介绍，整理在表9中. 但要注意到的是，与语义分割相比，针对深度估计无监督域自适应的相关研究还比较少.

6.1 基于生成对抗的无监督域自适应方法

与其他视觉任务相似，生成对抗模型也被广泛用于深度估计的无监督域自适应中. 方法包含两个存在对抗关系的网络，分别用于估计场景深度和鉴别数据来源. 与其他视觉任务相似，额外的用于图片风格迁移的GAN也被用于深度估计的无监督域自适应中，在深度估计的输入空间进行训练数据增强，以提升域自适应的结果. 本小节根据输入鉴别器数据来自于深度估计的特征空间还是输出空间，以及是否进行数据增强，将方法分为在特征空间、输出空间及输入空间进行域自适应三类介绍.

6.1.1 深度估计特征空间

在深度估计特征空间使用域鉴别器的方法通过判断编码器提取的中间特征来自于源域还是目标域与深度估计网络进行对抗学习，从而实现域自适应. Ren等人[33]提出一个联合估计场景深度、表面法线和实例轮廓的多任务学习框架. 由于多任务的标签很难获取，因此先在合成数据集上训练，再通过对抗学习将合成和真实数据集的特征进行对齐，从而减小域偏移的影响. Bozorgtabar等人[172]利用合成数据集和单目图片序列共同进行单目深度估计，在特征空间上使用生成器和鉴别器，利用生成器将真实数据的特征转换到合成数据空间上，利用鉴别器判别. 此外，方法通过利用合成数据集中的几何信息，恢复了单目深度估计的绝对尺度，生成尺度一致的场景深度和相机相对位移. Vankadari等人[173]预先利用无监督单目深度估计的方法[34]训练一个深度估计网络，然后分别提取日间和夜晚图片信息的特征后输入鉴别器，最后将提取夜晚特征的编码器和用于日间场景深度估计的解码器结合，用以估计夜晚条件下的场景深度，实现深度估计模型在日间向夜晚场景中的迁移.

6.1.2 深度估计输出空间

本小节总结了在深度估计的输出空间上使用鉴别器进行对抗学习的方法，即通过一个域鉴别器判断输

表 9 无监督域自适应方法在深度估计中的应用

Table 9 Depth estimation with UDA

分类		年份	方法	鉴别器输入			生成方向[1)]	重构[2)]	特点	数据集[3)]
				输入	特征	输出				
生成对抗	特征空间	2018	Ren等人[33]	–	√	–	–	–	深度、表面法线、实例轮廓	SU, Sc→P
		2019	SynDeMo[172]	–	√	–	–	–	单目深度估计	vK→K
		2020	ADFA[173]	–	√	–	–	–	夜间场景	O
	输出空间	2018	AdaDepth[174]	–	√	√	–	–	–	PB→ND, vK→K
		2019	UM-Adapt[175]	–	√	√	–	–	多任务学习	PB→ND, G→K, C
	输入空间	2018	Atapour等人[50]	√	√	√	S⇌T	–	CycleGAN	K→M3
		2018	T^2Net[176]	√	√	–	S→T	√	–	SU→ND, vK→K
		2019	GASDA[178]	√	–	–	S⇌T	–	双目深度估计	vK→K
		2020	S^2Net[179]	√	–	–	S→T	–	单目深度估计、语义一致	vK→K

1) 此生成方向指在图像分类输入空间上使用的风格迁移网络中的生成器的生成方向, S代表源域, T代表目标域.

2) 这里的重构不包括在输入空间使用风格迁移网络重构图像.

3) 数据集名称均为表5中数据集的缩写形式.

入是来自何域的深度估计图. Kundu等人[174]使用两个域鉴别器分别从特征层面和输出层面进行对抗学习, 将从包含图片和相应的场景深度真值的合成数据集中学习到的模型参数迁移到无真值的真实数据集中. 特别地, 为缓解对抗学习中模式崩塌的问题, 提出内容一致性约束, 确保输出的场景深度图与输入图像的空间结构信息保持一致. 随后, Kundu等人[175]提出两种新的正则化策略: 基于轮廓的内容正则化和利用跨任务蒸馏模块探索任务间一致性的正则化, 以同时估计场景深度、表面法线和语义分割结果. 此外, 此方法中没有使用传统的域鉴别器, 而是将跨任务的蒸馏损失作为目标函数以最小化域偏移.

6.1.3 深度估计输入空间

在深度估计输入空间进行域自适应的方法首先利用额外的GAN进行域间的图片风格迁移, 再引入对抗思想进行深度估计域自适应. 在输入空间预先进行图片风格迁移的方法先将含有场景深度真值的源域图片转变成有着目标域风格的图片, 进行数据增强, 再引入对抗思想进行深度估计域自适应训练, 因此这类工作一般包含一个风格迁移网络和一个深度估计网络.

Atapour等人[50]利用CycleGAN进行图片风格迁移从而实现数据增强, 同时使用源域场景深度真值和预测的场景深度图进行对抗学习, 训练源域图片的深度估计模型. 然后利用CycleGAN中的生成器将目标域图片转换成源域的风格, 以便利用训练的深度估计模型预测场景深度值. Zheng等人[176]提出广义的风格迁移网络: 若网络输入为真实图片, 网络重建图片; 若输入为合成图片, 网络将合成图片的风格转变为真实图片的风格. 因此, 这个广义的风格迁移网络能够更好地提取真实图片的特征, 实现特征对齐. 与其他计算机视觉任务不同, 深度估计任务可以在没有真值信号的情况下, 利用双目图片对之间或单目视频序列之间的几何约束以无监督的方式预测场景深度值[34,177]. 由于这种几何约束受域偏移的影响较小, Zhao等人[178]同时利用合成数据中的场景深度真值标记和真实双目图像中的对极几何约束进行域自适应. Cheng等人[179]将网络分为两步训练, 首先使用源域数据进行风格迁移和深度估计网络训练, 然后固定风格迁移网络, 通过源域图片的有监督信号和目标域图片的自监督信号训练深度估计网络. 此外, 考虑到图片风格转换前后的语义分割标签应该不变, 因此, 对风格转变前后的图片上语义分割结果进行一致性约束.

由于深度估计任务能通过图片间的几何约束减小域偏移的影响, 因此一些语义估计的域自适应方法[147,154]将这种场景深度信息加入到语义分割的无监督域自适应中. 而语义分割任务中的标签信息蕴含空间结构信息, 因此未来可以考虑在深度估计任务中加入语义信息来提升深度估计的域自适应结果.

7 未来研究方向

近年来, 随着深度神经网络的发展, 基于深度学习的无监督域自适应方法也得到了飞速发展. 无监督域自适应方法在计算机视觉任务中的发展大大提升了深度学习模型的迁移性, 使得模型在没有数据标签的新环境中也能表现良好, 为未来深度学习模型在更多领域的落地和使用奠定了基础. 从目前无监督域自适应的发展趋势来看, 基于生成对抗思想的方法是计算机视觉领域域自适应的主流方法. 从不同任务的发展情况来看, 不同任务间的域自适应方法发展不平衡. 虽然不同的计算机视觉任务各有特点, 但无监督域自适应的基本思想是相通的, 因此不同任务之间可以互相借鉴, 促进各个任务的发展. 目前来看, 无监督域自适应在计算机视觉任务中主要存在以下几个潜在的研究方向.

(1) 常识性信息的加入　通过加入例如道路是直的、车辆应在路边停靠、天空距离很远等常识性信息来提升计算机视觉任务的效果, 可以通过加入知识图谱等方法来提升网络对于上下文语境信息的理解.

(2) 单任务到多任务　进行多任务学习, 例如同时学习语义、场景深度、光流等像素级任务. 虽然不同计算机视觉任务各有特点, 但在它们之间也存在一定的相关性, 因此可以通过探索不同任务间的内在关联, 利用这种内在关联增强无监督域自适应的约束. 几个不同但相关的任务之间信息共享, 互相协作以提升网络性能.

(3) 二维到三维　从二维计算机视觉任务扩展到三维, 例如从2D目标检测扩展到3D目标检测, 从2D语义分割扩展到3D语义分割, 从光流估计扩展到场景流的域自适应等. 三维任务相比于二维任务来说, 难度更大, 所需的信息量更多, 往往需要进行多模态信息的融合, 因此需要探索如何利用多源信息进行域自适应,

以及多源数据之间如何互补以提升网络性能.

(4) 更富挑战性的场景迁移 目前主要的域自适应工作着重于探索如何在不同风格的数据集之间迁移或从合成数据集迁移到真实数据集. 未来可以更加注重于探索如何将域自适应方法应用到更富挑战性的场景中, 如光照条件变化(白天与黑夜), 天气条件变化(有雨、有雾、有霾等), 季节变化等, 这些场景对于算法的实际应用和应用意义重大.

(5) 轻量级网络 目前的深度估计域自适应网络结构复杂, 以基于生成对抗模型的深度估计域自适应方法为例, 由图片风格迁移网络和视觉任务网络组成, 训练过程需要耗费大量时间与计算资源. 因此, 亟需研究一种轻量级的方法, 使算法更具实用价值.

(6) 实际应用 无监督域自适应算法减轻了域偏移对于深度学习模型的影响, 使得深度学习模型能在不同环境中都能取得良好的效果. 因此可以将基于无监督域自适应算法的深度学习模型应用到实际场景中, 并与决策和控制方法融合, 赋予智能体自主感知、决策和控制的能力, 在自主无人系统中发挥作用.

8 结束语

基于深度学习的无监督域自适应方法在图像分类、目标检测、语义分割和深度估计等计算机视觉任务上取得了突出成果, 为在缺少真值标签的域上实现相关计算机视觉任务提供了可能性. 本文从是否使用生成对抗模型为依据进行分类, 概括总结了近年来基于深度学习的无监督域自适应方法在典型计算机视觉任务上的研究进展. 无监督域自适应的研究为理论算法的实际落地和使用提供了可能性, 其与深度学习的融合为该领域注入了新的活力, 是未来计算机视觉与深度学习领域的一大值得研究的方向.

参考文献

1 Yuan Y, Tang X, Zhou W, et al. Data driven discovery of cyber physical systems. Nature Commun, 2019, 10: 1–9

2 Yuan Y, Ma G, Cheng C, et al. A general end-to-end diagnosis framework for manufacturing systems. Natl Sci Rev, 2020, 7: 418–429

3 Tang Y, Zhao C, Wang J, et al. Perception and decision-making of autonomous systems in the era of learning: An overview. arXiv preprint, arXiv: 2001.02319, 2020

4 Li H, Duan H B. Verification of monocular and binocular pose estimation algorithms in vision-based uavs autonomous aerial refueling system. Sci China Tech Sci, 2016, 59: 1730–1738

5 Zhu Z S, Su A, Liu H B, et al. Vision navigation for aircrafts based on 3D reconstruction from real-time image sequences. Sci China Tech Sci, 2015, 58: 1196–1208

6 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 2012 Advances in Neural Information Processing Systems (NIPS). Lake Tahoe, USA: MIT Press, 2012. 1097–1105

7 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770–778

8 Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of the 2015 Advances in Neural Information Processing Systems (NIPS). Montreal, Canada: MIT Press, 2015. 91–99

9 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 3431–3440

10 Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell, 2017, 40: 834–848

11 Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. In: Proceedings of the 2014 Advances in Neural Information Processing Systems (NIPS). Montreal, Canada: MIT Press, 2014. 2366–2374

12 Zhao C Q, Sun Q Y, Zhang C Z, et al. Monocular depth estimation based on deep learning: An overview. Sci China Tech Sci, 2020, 63: 1612–1627

13 Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 3213–3223

14 Ghifary M, Kleijn W B, Zhang M, et al. Deep reconstruction-classification networks for unsupervised domain adaptation. In: Proceedings of the

2016 European Conference on Computer Vision (ECCV). Amsterdam, Netherlands: Springer, 2016. 597–613
15 Zhang C, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization. arXiv preprint, arXiv: 1611.03530, 2016
16 Zhang C, Wang J, Yen G G, et al. When autonomous systems meet accuracy and transferability through AI: A survey. Patterns (New York, N.Y.), 2020, 1: 100050
17 Torralba A, Efros A A. Unbiased look at dataset bias. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, USA: IEEE, 2011. 1521–1528
18 Tommasi T, Patricia N, Caputo B, et al. A deeper look at dataset bias. In: Proceedings of 37th German Conference on Pattern Recognition (GCPR). Springer, 2017. 37–55
19 Patel V M, Gopalan R, Li R, et al. Visual domain adaptation: A survey of recent advances. IEEE Signal Process Mag, 2015, 32: 53–69
20 Csurka G. Domain adaptation for visual applications: A comprehensive survey. arXiv preprint, arXiv: 1702.05374, 2017
21 Wilson G, Cook D J. A survey of unsupervised deep domain adaptation. ACM Trans Intell Syst Technol, 2020, 11: 1–46
22 Wang M, Deng W. Deep visual domain adaptation: A survey. Neurocomputing, 2018, 312: 135–153
23 Pan S J, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng, 2009, 22: 1345–1359
24 Weiss K, Khoshgoftaar T M, Wang D D. A survey of transfer learning. J Big Data, 2016, 3: 9
25 Saenko K, Kulis B, Fritz M, et al. Adapting visual category models to new domains. In: Proceedings of the 2010 European Conference on Computer Vision (ECCV). Crete, Greece: Springer, 2010. 213–226
26 Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks. J Mach Learn Res, 2016, 17: 2096–2030
27 Hoffman J, Wang D, Yu F, et al. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint, arXiv: 1612.02649, 2016
28 Tan C, Sun F, Kong T, et al. A Survey on deep transfer learning. In: Proceedings of International conference on Artifical Neural Networks. Springer, 2018. 270–279
29 Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: Proceedings of the 2015 International Conference on Machine Learning (ICML). Lille, France: ACM, 2015. 1180–1189
30 Tzeng E, Hoffman J, Zhang N, et al. Deep domain confusion: Maximizing for domain invariance. arXiv preprint, arXiv: 1412.3474, 2014
31 Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks. In: Proceedings of the 2015 International Conference on Machine Learning (ICML). Lille, France: ACM, 2015. 97–105
32 Chen Y, Li W, Sakaridis C, et al. Domain adaptive faster R-CNN for object detection in the wild. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 3339–3348
33 Ren Z, Jae Lee Y. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 762–771
34 Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 1851–1858
35 Bengio Y, Louradour J, Collobert R, et al. Curriculum learning. In: Proceedings of the 2009 International Conference on Machine Learning (ICML). Montreal, Canada: ACM, 2009. 41–48
36 Doersch C, Gupta A, Efros A A. Unsupervised visual representation learning by context prediction. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1422–1430
37 Zhu X, Goldberg A B. Introduction to semi-supervised learning. Synth Lectures Art Intell Mach Learn, 2009, 3: 1–130
38 Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. In: Proceedings of the 2005 Advances in Neural Information Processing Systems (NIPS). Vancouver, Canada: MIT Press, 2005. 529–536
39 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint, arXiv: 1503.02531, 2015
40 Long M, Zhu H, Wang J, et al. Unsupervised domain adaptation with residual transfer networks. In: Proceedings of the 2016 Advances in Neural Information Processing Systems (NIPS). Barcelona, Spain: MIT Press, 2016. 136–144
41 Sun B, Feng J, Saenko K. Return of frustratingly easy domain adaptation. In: Proceedings of the 2016 AAAI Conference on Artificial Intelligence (AAAI). Phoenix, Arizona, USA: AAAI, 2016
42 Sun B, Saenko K. Deep coral: Correlation alignment for deep domain adaptation. In: Proceedings of the 2016 European Conference on Computer Vision (ECCV). Amsterdam, Netherlands: Springer, 2016. 443–450

43 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of the 2014 Advances in Neural Information Processing Systems (NIPS). Montreal, Canada: MIT Press, 2014. 2672–2680

44 Tzeng E, Hoffman J, Darrell T, et al. Simultaneous deep transfer across domains and tasks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 4068–4076

45 Liu M Y, Tuzel O. Coupled generative adversarial networks. In: Proceedings of the 2016 Advances in Neural Information Processing Systems (NIPS). Barcelona, Spain: MIT Press, 2016. 469–477

46 Saito K, Ushiku Y, Harada T, et al. Adversarial dropout regularization. arXiv preprint, arXiv: 1711.01575, 2017

47 Saito K, Watanabe K, Ushiku Y, et al. Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 3723–3732

48 Kim T, Jeong M, Kim S, et al. Diversify and match: A domain adaptive representation learning paradigm for object detection. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 12456–12465

49 Zhang Y, Qiu Z, Yao T, et al. Fully convolutional adaptation networks for semantic segmentation. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 6810–6818

50 Atapour-Abarghouei A, Breckon T P. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 2800–2810

51 Andrychowicz M, Denil M, Colmenarejo S G, et al. Learning to learn by gradient descent by gradient descent. In: Proceedings of the 2016 Advances in Neural Information Processing Systems (NIPS). Barcelona, Spain: MIT Press, 2016. 3988–3996

52 Wang K, Gou C, Duan Y, et al. Generative adversarial networks: Introduction and outlook. IEEE/CAA J Autom Sin, 2017, 4: 588–598

53 LeCun Y. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998

54 Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The KITTI dataset. Int J Robotics Res, 2013, 32: 1231–1237

55 Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgbd images. In: Proceedings of the 2012 European Conference on Computer Vision (ECCV). Florence, Italy: Springer, 2012. 746–760

56 Zhang Y, Song S, Yumer E, et al. Physically-based rendering for indoor scene understanding using convolutional neural networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 5287–5295

57 Song S, Yu F, Zeng A, et al. Semantic scene completion from a single depth image. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 1746–1754

58 McCormac J, Handa A, Leutenegger S, et al. SceneNet RGB-D: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2678–2687

59 Ros G, Sellart L, Materzynska J, et al. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 3234–3243

60 Gaidon A, Wang Q, Cabon Y, et al. Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 4340–4349

61 Richter S R, Vineet V, Roth S, et al. Playing for data: Ground truth from computer games. In: Proceedings of the 2016 European Conference on Computer Vision (ECCV). Amsterdam, Netherlands: Springer, 2016. 102–118

62 Sakaridis C, Dai D, Van Gool L. Semantic foggy scene understanding with synthetic data. Int J Comput Vis, 2018, 126: 973–992

63 Wrenninge M, Unger J. Synscapes: A photorealistic synthetic dataset for street scene parsing. arXiv preprint, arXiv: 1810.08705, 2018

64 LeCun Y, Matan O, Boser B, et al. Handwritten zip code recognition with multilayer networks. In: Proceedings of the 1990 International Conference on Pattern Recognition. Atlantic City, NJ, USA: IEEE, 1990. 35–40

65 Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning, In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011. 2011

66 Moiseev B, Konev A, Chigorin A, et al. Evaluation of traffic sign recognition methods trained on synthetically generated data. In: International Conference on Advanced Concepts for Intelligent Vision Systems. Catania, Italy: Springer, 2013. 576–583

67 Stallkamp J, Schlipsing M, Salmen J, et al. The german traffic sign recognition benchmark: A multiclass classification competition. In: Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN). San Jose, CA, USA: IEEE, 2011. 1453–1460

68 Gong B, Shi Y, Sha F, et al. Geodesic flow kernel for unsupervised domain adaptation. In: Proceedings of the 2012 IEEE Conference on

Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA: IEEE, 2012. 2066–2073
69 Venkateswara H, Eusebio J, Chakraborty S, et al. Deep hashing network for unsupervised domain adaptation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 5018–5027
70 Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images, In: Handbook of Systemic Autoimmune Diseases. 2009
71 Coates A, Ng A, Lee H. An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Florida, USA: JMLR, 2011. 215–223
72 Peng X, Usman B, Kaushik N, et al. Visda: The visual domain adaptation challenge. arXiv preprint, arXiv: 1710.06924, 2017
73 Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge. Int J Comput Vis, 2010, 88: 303–338
74 Chen Y H, Chen W Y, Chen Y T, et al. No more discrimination: Cross city adaptation of road scene segmenters. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 1992–2001
75 Maddern W, Pascoe G, Linegar C, et al. 1 year, 1000 km: The oxford robotcar dataset. Internat J Robot Res, 2017, 36: 3–15
76 Neuhold G, Ollmann T, Rota Bulo S, et al. The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 4990–4999
77 Zhou B, Khosla A, Lapedriza A, et al. Places: An image database for deep scene understanding. arXiv preprint, arXiv: 1610.02055, 2016
78 Saxena A, Min Sun A, Ng A Y. Make3d: Learning 3d scene structure from a single still image. IEEE Trans Pattern Anal Mach Intell, 2008, 31: 824–840
79 Griffin G, Holub A, Perona P. Caltech-256 object category dataset, 2007
80 Borgwardt K M, Gretton A, Rasch M J, et al. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics, 2006, 22: e49–e57
81 Long M, Zhu H, Wang J, et al. Deep transfer learning with joint adaptation networks. In: Proceedings of the 2017 International Conference on Machine Learning (ICML). Sydney, Australia: ACM, 2017. 2208–2217
82 Rozantsev A, Salzmann M, Fua P. Beyond sharing weights for deep domain adaptation. IEEE Trans Pattern Anal Mach Intell, 2018, 41: 801–814
83 Wang Y, Li W, Dai D, et al. Deep domain adaptation by geodesic distance minimization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2651–2657
84 Morerio P, Murino V. Correlation alignment by riemannian metric for domain adaptation. arXiv preprint, arXiv: 1705.08180, 2017
85 Courty N, Flamary R, Habrard A, et al. Joint distribution optimal transportation for domain adaptation. In: Proceedings of the 2017 Advances in Neural Information Processing Systems (NIPS). Long Beach, CA, USA, 2017. 3730–3739
86 Redko I, Habrard A, Sebban M. Theoretical analysis of domain adaptation with optimal transport. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Skopje, Macedonia: Springer, 2017. 737–753
87 Bhushan Damodaran B, Kellenberger B, Flamary R, et al. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 447–463
88 Vallender S. Calculation of the wasserstein distance between probability distributions on the line. Theor Probab Appl, 1974, 18: 784–786
89 Xu R, Liu P, Wang L, et al. Reliable weighted optimal transport for unsupervised domain adaptation. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 4394–4403
90 Das D, Lee C G. Unsupervised domain adaptation using regularized hyper-graph matching. In: 2018 25th IEEE International Conference on Image Processing (ICIP). Athens, Greece: IEEE, 2018. 3758–3762
91 Das D, Lee C S G. Sample-to-sample correspondence for unsupervised domain adaptation. Eng Appl Artificial Intelligence, 2018, 73: 80–91
92 Kang G, Jiang L, Yang Y, et al. Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 4893–4902
93 Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 7167–7176
94 Sohn K, Shang W, Yu X, et al. Unsupervised domain adaptation for distance metric learning. In: Proceedings of the 2018 International Conference on Learning Representations (ICLR). Vancouver, Canada: 2018
95 Cui S, Wang S, Zhuo J, et al. Gradually vanishing bridge for adversarial domain adaptation. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 12455–12464

96 Shen J, Qu Y, Zhang W, et al. Wasserstein distance guided representation learning for domain adaptation. arXiv preprint, arXiv: 1707.01217, 2017

97 Long M, Cao Z, Wang J, et al. Conditional adversarial domain adaptation. In: Proceedings of the 2018 Advances in Neural Information Processing Systems (NIPS). Montreal, Canada: MIT Press, 2018. 1640–1650

98 Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint, arXiv: 1411.1784, 2014

99 Xia H, Ding Z. Structure preserving generative cross-domain learning. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 4364–4373

100 Wei G, Lan C, Zeng W, et al. MetaAlign: Coordinating domain alignment and classification for unsupervised domain adaptation. In: Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. 16643–16653

101 Li S, Zhang J, Ma W, et al. Dynamic domain adaptation for efficient inference. In: Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. 7832–7841

102 Bousmalis K, Silberman N, Dohan D, et al. Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 3722–3731

103 Sankaranarayanan S, Balaji Y, Castillo C D, et al. Generate to adapt: Aligning domains using generative adversarial networks. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 8503–8512

104 Russo P, Carlucci F M, Tommasi T, et al. From source to target and back: Symmetric bidirectional adaptive gan. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 8099–8108

105 Ye S, Wu K, Zhou M, et al. Light-weight calibrator: A separable component for unsupervised domain adaptation. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 13736–13745

106 Bousmalis K, Trigeorgis G, Silberman N, et al. Domain separation networks. In: Proceedings of the 2016 Advances in Neural Information Processing Systems (NIPS). Barcelona, Spain: MIT Press, 2016. 343–351

107 Wang K, Zhang D, Li Y, et al. Cost-effective active learning for deep image classification. IEEE Trans Circuits Syst Video Technol, 2016, 27: 2591–2600

108 Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers. IEEE Trans Knowl Data Eng, 2005, 17: 1529–1541

109 Saito K, Ushiku Y, Harada T. Asymmetric tritraining for unsupervised domain adaptation. arXiv preprint, arXiv: 1702.08400, 2017

110 Yue X, Zheng Z, Zhang S, et al. Prototypical crossdomain self-supervised learning for few-shot unsupervised domain adaptation. In: Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. 13834–13844

111 He Z, Zhang L. Multi-adversarial faster-rcnn for unrestricted object detection. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 6668–6677

112 Zhu X, Pang J, Yang C, et al. Adapting object detectors via selective cross-domain alignment. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 687–696

113 Xu C D, Zhao X R, Jin X, et al. Exploring categorical regularization for domain adaptive object detection. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 11724–11733

114 Saito K, Ushiku Y, Harada T, et al. Strongweak distribution alignment for adaptive object detection. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 6956–6965

115 Zheng Y, Huang D, Liu S, et al. Crossdomain object detection through coarse-to-fine feature adaptation. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 13766–13775

116 VS V, Oza P, Sindagi V A, et al. MeGA-CDA: Memory guided attention for category-aware unsupervised domain adaptive object detection. In: Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. 4516–4526

117 Xu M, Wang H, Ni B, et al. Cross-domain detection via graph-induced prototype alignment. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 12355–12364

118 RoyChowdhury A, Chakrabarty P, Singh A, et al. Automatic adaptation of object detectors to new domains using selftraining. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 780–790

119 Yang J, Shi S, Wang Z, et al. St3d: Selftraining for unsupervised domain adaptation on 3d objectdetection. In: Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. 10368–10378

120 Yang S, Luo P, Loy C C, et al. Wider face: A face detection benchmark. In: Proceedings of the 2016 IEEE Conference on Computer Vision and

Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 5525–5533

121 Kalka N D, Maze B, Duncan J A, et al. Ijb-s: Iarpa janus surveillance video benchmark. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, 2018. 1–9

122 Yu F, Xian W, Chen Y, et al. Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint, arXiv: 1805.04687, 2018

123 Sun P, Kretzschmar H, Dotiwalla X, et al. Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 2446–2454

124 Caesar H, Bankiti V, Lang A H, et al. Nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 11621–11631

125 Kesten R, Usman M, Houston J, et al. Lyft level 5 perception dataset 2020. 2019

126 Huang H, Huang Q, Krahenbuhl P. Domain transfer through deep activation matching. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 590–605

127 Du L, Tan J, Yang H, et al. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 982–991

128 Dong J, Cong Y, Sun G, et al. CSCL: Critical semantic-consistent learning for unsupervised domain adaptation. In: Proceedings of the 2020 European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 745–762

129 Chen Y, Li W, Van Gool L. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 7892–7901

130 Pan F, Shin I, Rameau F, et al. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 3764–3773

131 Luo Y, Liu P, Guan T, et al. Significanceaware information bottleneck for domain adaptive semantic segmentation. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 6778–6787

132 Tsai Y H, Hung W C, Schulter S, et al. Learning to adapt structured output space for semantic segmentation. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 7472–7481

133 Tsai Y H, Sohn K, Schulter S, et al. Domain adaptation for structured output via discriminative patch representations. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 1456–1465

134 Vu T H, Jain H, Bucher M, et al. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 2517–2526

135 Vu T H, Jain H, Bucher M, et al. Dada: Depth-aware domain adaptation in semantic segmentation. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 7364–7373

136 Luo Y, Zheng L, Guan T, et al. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 2507–2516

137 Wang H, Shen T, Zhang W, et al. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In: Proceedings of the 2020 European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 642–659

138 Wang Z, Yu M, Wei Y, et al. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 12635–12644

139 Liu Y, Zhang W, Wang J. Source-free domain adaptation for semantic segmentation. In: Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. 1215–1224

140 Wu Z, Han X, Lin Y L, et al. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 518–534

141 Zhu X, Zhou H, Yang C, et al. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 568–583

142 Zhang Y, Qiu Z, Yao T, et al. Transferring and regularizing prediction for semantic segmentation. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 9621–9630

143 Sankaranarayanan S, Balaji Y, Jain A, et al. Learning from synthetic data: Addressing domain shift for semantic segmentation. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 3752–3761

144 Chang W L, Wang H P, Peng W H, et al. All about structure: Adapting structural information across domains for boosting semantic segmentation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 1900–1909

145 Yang J, An W, Wang S, et al. Label-driven reconstruction for domain adaptation in semantic segmentation. In: Proceedings of the 2020 European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 480–498

146 Hoffman J, Tzeng E, Park T, et al. Cycada: Cycleconsistent adversarial domain adaptation. In: Proceedings of the 2018 International Conference on Machine Learning (ICML). Stockholm, Sweden: ACM, 2018. 1989–1998

147 Chen Y C, Lin Y Y, Yang M H, et al. Crdoco: Pixel-level domain transfer with cross-domain consistency. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 1791–1800

148 Murez Z, Kolouri S, Kriegman D, et al. Image to image translation for domain adaptation. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 4500–4509

149 Li Y, Yuan L, Vasconcelos N. Bidirectional learning for domain adaptation of semantic segmentation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 6936–6945

150 Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2223–2232

151 Choi J, Kim T, Kim C. Self-ensembling with ganbased data augmentation for domain adaptation in semantic segmentation. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 6830–6840

152 Kim M, Byun H. Learning texture invariant representation for domain adaptation of semantic segmentation. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 12975–12984

153 Zhao S, Li B, Yue X, et al. Multi-source domain adaptation for semantic segmentation. In: Proceedings of the 2019 Advances in Neural Information Processing Systems (NIPS). Vancouver, Canada: MIT Press, 2019. 7287–7300

154 Chen Y, Li W, Chen X, et al. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 1841–1850

155 Zhang Y, David P, Gong B. Curriculum domain adaptation for semantic segmentation of urban scenes. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2020–2030

156 Sakaridis C, Dai D, Hecker S, et al. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 687–704

157 Sakaridis C, Dai D, Gool L V. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 7374–7383

158 Lian Q, Lv F, Duan L, et al. Constructing self-motivated pyramid curriculums for crossdomain semantic segmentation: A non-adversarial approach. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 6758–6767

159 Zou Y, Yu Z, Vijaya Kumar B, et al. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 289–305

160 Saleh F S, Aliakbarian M S, Salzmann M, et al. Effective use of synthetic data for urban scene semantic segmentation. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 86–103

161 Zou Y, Yu Z, Liu X, et al. Confidence regularized self-training. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 5982–5991

162 Guo X, Yang C, Li B, et al. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In: Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. 3927–3936

163 Iqbal J, Ali M. MLSL: Multi-level self-supervised learning for domain adaptation with spatially independent and semantically consistent labeling. In: Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Snowmass Village, CO, USA:

IEEE, 2020. 1864–1873

164 Li G, Kang G, Liu W, et al. Content-Consistent Matching for Domain Adaptive Semantic Segmentation. Springer, 2020

165 Mei K, Zhu C, Zou J, et al. Instance adaptive self-training for unsupervised domain adaptation. In: Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 415–430

166 Ma H, Lin X, Wu Z, et al. Coarse-to-fine domain adaptive semantic segmentation with photometric alignment and category-center regularization. In: Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. 4051–4060

167 He J, Jia X, Chen S, et al. Multi-source domain adaptation with collaborative learning for semantic segmentation. In: Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021. 11008–11017

168 Lee D H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML 2103 Workshop on Challenges in Representation Learning. Atlanta, USA: ACM, 2013

169 Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. In: Proceedings of Advances in Neural Information Processing Systems (NIPS). Vancouver, British Columbia, Canada, 2004. 529–536

170 Chen M, Xue H, Cai D. Domain adaptation for semantic segmentation with maximum squares loss. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 2090–2099

171 Huang J, Lu S, Guan D, et al. Contextualrelation consistent domain adaptation for semantic segmentation. In: Proceedings of the 2020 European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 705-722

172 Bozorgtabar B, Rad M S, Mahapatra D, et al. SynDeMo: Synergistic deep feature alignment for joint learning of depth and ego-motion. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 4210–4219

173 Vankadari M B, Garg S, Majumdar A, et al. Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In: Proceedings of the 2020 European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 443–459

174 Nath Kundu J, Krishna Uppala P, Pahuja A, et al. Adadepth: Unsupervised content congruent adaptation for depth estimation. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 2656–2665

175 Kundu J N, Lakkakula N, Babu R V. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In: Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 1436–1445

176 Zheng C, Cham T J, Cai J. T2net: Synthetic-torealistic translation for solving single-image depth estimation tasks. In: European Conference on Computer Vision (ECCV). 2018. 767–783

177 Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 270–279

178 Zhao S, Fu H, Gong M, et al. Geometry-aware symmetric domain adaptation for monocular depth estimation. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 9788–9798

179 Cheng B, Saggu I S, Shah R, et al. S^3Net: Semantic-aware self-supervised depth estimation with monocular videos and synthetic data. In: Proceedings of the 2020 European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 52–69

A survey on unsupervised domain adaptation in computer vision tasks

SUN QiYu, ZHAO ChaoQiang, TANG Yang & QIAN Feng

School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

As one of the typical applications of Industrial Internet, Internet of Vehicles (IoV) develops rapidly in recent years. It relies on the interconnection of information, where the transferability of the accurate perception is of great importance. Though deep learning has accelerated the development of computer vision tasks, traditional deep learning-based methods still have strong reliance on manually annotated training data and are poor at generalizing knowledge to new environments. For computer vision tasks, since it is difficult to collect training data with ground truth, it is agent to improve the generalization ability of deep learning models and alleviate their dependence on manually annotated labels. Unsupervised domain adaptation (UDA) methods apply deep learning models to extract and align features from data in different domains, which ensures the satisfactory generalization performance of deep learning-based computer vision algorithms. This paper focuses on the challenges and applications of UDA in some typical computer vision tasks. Firstly, the definition, significances, application difficulties, basic methods and relevant datasets of deep learning-based UDA methods are introduced. Then, the mainstream UDA methods in typical computer vision tasks are introduced separately. Finally, the prospective technical development trends and a summary are given.

unsupervised domain adaptation, computer vision, deep learning, transfer learning, autonomous system

doi: 10.1360/SST-2021-0150