自动化学报
ACTA AUTOMATICA SINICA

基于注意力机制和循环域三元损失的域自适应目标检测

周洋 韩冰 高新波 杨铮 陈玮铭

Domain Adaptive Object Detection Based on Attention Mechanism and Cycle Domain Triplet Loss

ZHOU Yang, HAN Bing, GAO Xin-Bo, YANG Zheng, CHEN Wei-Ming

在线阅读 View online: https://doi.org/10.16383/j.aas.c220938

基于注意力机制和循环域三元损失的域自适应目标检测

周 洋 [1] 韩 冰 [1] 高新波 [1, 2] 杨 铮 [1] 陈玮铭 [1]

摘 要 目前大多数深度学习算法都依赖于大量的标注数据并欠缺一定的泛化能力. 无监督域自适应算法能提取到已标注数据和未标注数据间隐式共同特征, 从而提高算法在未标注数据上的泛化性能. 目前域自适应目标检测算法主要为两阶段目标检测器设计. 针对单阶段检测器中无法直接进行实例级特征对齐导致一定数量域不变特征的缺失, 提出结合通道注意力机制的图像级域分类器加强域不变特征提取. 此外, 对于域自适应目标检测中存在类别特征的错误对齐引起的精度下降问题, 通过原型学习构建类别中心, 设计了一种基于原型的循环域三元损失 (Cycle domain triplet loss, CDTL) 函数, 从而实现原型引导的精细类别特征对齐. 以单阶段目标检测算法作为检测器, 并在多种域自适应目标检测公共数据集上进行实验. 实验结果证明该方法能有效提升原检测器在目标域的泛化能力, 达到比其他方法更高的检测精度, 并且对于单阶段目标检测网络具有一定的通用性.

关键词 无监督域自适应, 注意力机制, 循环域三元损失函数, 目标检测

引用格式 周洋, 韩冰, 高新波, 杨铮, 陈玮铭. 基于注意力机制和循环域三元损失的域自适应目标检测. 自动化学报, 2024, **50**(11): 1−16

DOI 10.16383/j.aas.c220938 **CSTR** 32138.14.j.aas.c220938

Domain Adaptive Object Detection Based on Attention Mechanism and Cycle Domain Triplet Loss

ZHOU Yang[1] HAN Bing[1] GAO Xin-Bo[1, 2] YANG Zheng[1] CHEN Wei-Ming[1]

Abstract Most current deep learning algorithms rely heavily on large amounts of annotated data and exist deficiency in generalization ability. The unsupervised domain adaptation algorithm can extract the common implicit invariant features from the labeled data and unlabeled data, so that the algorithm can achieve good generalization performance on the unlabeled data. At present, domain adaptation object detection algorithms are mainly designed as two-stage object detectors. For the one-stage object detectors, the difficulty of explicit aligning instance-level features leads to the absence of a number of domain invariant features. In this paper, an image-level domain classifier combined with channel attention mechanism is proposed to strengthen domain invariant feature extraction. In addition, to address the issue of reduced accuracy caused by inaccurate alignment of category features in domain adaptive object detection, a prototype based cycle domain triplet loss (CDTL) function was designed to construct category centers through prototype learning, thereby we can achieve precise category feature alignment guided by prototypes. One-stage object detection algorithms are used as detectors, and experiments are conducted on various domain adaptive object detection public datasets. The experimental results show that our method can effectively improve the generalization ability of the original detector on the target domain and achieves higher detection accuracy than other methods. Meanwhile, the experiment on different detector indicate our method is universal for the one-stage object detection network.

Key words Unsupervised domain adaptation, attention mechanism, cycle domain triplet loss function, object detection

Citation Zhou Yang, Han Bing, Gao Xin-Bo, Yang Zheng, Chen Wei-Ming. Domain adaptive object detection based on attention mechanism and cycle domain triplet loss. *Acta Automatica Sinica*, 2024, **50**(11): 1−16

收稿日期 2022-12-05 录用日期 2023-05-18
Manuscript received December 5, 2022; accepted May 18, 2023
国家自然科学基金 (62076190, 41831072, 62036007),陕西省重点创新产业链基金 (2022ZDLGY01-11), 西安市重点产业链技术攻关项目 (23ZDCYJSGG0022-2023), 国家空间科学数据中心青年开放课题基金 (NSSDC2302005) 资助
Supported by National Natural Science Foundation of China (62076190, 41831072, 62036007), Key Industry Innovation Chain of Shaanxi Province (2022ZDLGY01-11), Key Industry Chain Technology Research Project of Xi′an (23ZDCYJSGG0022-2023), and Youth Open Project of National Space Science Data Center (NSSDC2302005)
本文责任编委 张军平
Recommended by Associate Editor ZHANG Jun-Ping
1. 西安电子科技大学电子工程学院 西安 710071 2. 重庆邮电大学图像认知重庆市重点实验室 重庆 400065
1. School of Electronic Engineering, Xidian University, Xi′an 710071 2. Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065

随着深度学习时代的到来, 许多领域都发生着日新月异的巨大变化, 无论是智慧安防、智慧医疗亦或是目前备受关注的自动驾驶领域都得益于深度神经网络中提取到的高维语义. 目前基于神经网络[1]的深度学习方法在图像分类、目标检测、图像分割等领域取得了卓越的成绩. 但不可否认这些成功的背后都依赖于大量的标注数据, 所以目前大多数深度学习方法可以认为是数据驱动的. 通常地, 机器学习模型需要需要大量的已标注数据用于训练, 并假设训练集和测试集的数据是同分布的[2], 才能在测试阶段取得较好的效果. 但实际上如果将一个在某种特定场景 (或数据集上) 训练好的模型直接应用到另一种场景中 (或另一个数据集上), 当新场景数据与训练集数据不满足同分布假设的时候, 就会造成模型性能的大幅降低. 这是因为分布上的不一致使得直接应用于训练集外的模型发生了域迁移, 进而导致性能的退化. 这种现象在真实场景中非常常见, 例如自动驾驶场景中训练数据通常从晴朗的白天捕获而来, 而测试环境是没有标注的夜晚或者雨雪天等极端天气都会造成模型精度的骤减. 为解决以上问题, 提出了无监督域自适应方法, 旨在利用源域已有标注的数据和目标域没有标注的数据同时作为网络输入部分, 利用域自适应算法促使网络学习到域不变特征, 进而提升模型在目标域的泛化能力. 这种无监督的域自适应方法[3−4] 在早期往往通过一种距离度量来构造损失函数, 在训练过程中通过最小化这个损失函数从而拉近两个域之间的距离; 基于梯度反转层方法[5] 的提出为域自适应方向提供了一种新的思路, 与生成对抗网络[6] 中的原理类似, 通过构造一个具有梯度反转层的域分类器作为判别器, 利用对抗训练得到能够捕获域不变特性的特征提取器. 域自适应的方法目前在分类和分割任务上都取得了很好的成果并在行人重识别领域也有较好的结合[7−8], 但由于目标检测任务同时涉及到目标分类和目标框的回归使得直接应用域自适应方法存在一定困难, 所以基于域自适应方法在检测任务上的研究工作相对较少并存在一定的挑战.

目前, 大多数方法都是基于双阶段目标检测网络 Faster R-CNN (Region convolutional neural network)[9] 实现的域自适应目标检测算法. Chen 等[10] 首次将 Faster R-CNN 与域自适应算法相结合, 利用对抗特征学习的方法构建梯度反转层和域分类器实现图像级和实例级的特征对齐. Saito 等[11] 讨论了域分类器对于主干网络浅层和深层特征进行域自适应带来的不同影响, 并且使用 Focal Loss[12] 作为深层特征的域分类损失函数以解决类别不平衡问题. Shen 等[13] 进一步讨论了网络不同位置加入 Focal Loss所带来的影响. Zheng 等[14] 引入注意力机制获得权重特征图, 该特征图强调可能存在目标的区域, 并将该特征图和域分类损失加权, 使得网络更加关注于可能存在目标的区域, 同时该方法构建类别原型并计算各类原型之间的相似性, 实现类别特征的对齐. Xu 等[15] 提出一种类别正则化的策略进一步加强特征对齐, 该策略利用多标签分类器的弱定位能力去指导对抗训练. Hsu 等[16] 通过关注前景像素来实现基于中心感知的特征对齐, 从而获得更好的跨域自适应性. Chen 等[17] 在输入端使用循环对抗生成网络 (Cycle generative adversarial network, CycleGAN)[18] 将源域和目标域的图像转变成一个插值域来联结域间的鸿沟, 同时从域分类器中引入上下文特征向量来增强实例级特征的表达能力. Deng等[19] 设计了一种教师–学生蒸馏网络, 将蒸馏损失和域分类损失共同指导网络学习到域不变特征. Xu等[20] 结合图的思想, 在源域和目标域构建图结构和图一致性损失, 进而拉近两个域间的距离. Wu 等[21]提出一种基于向量分解的解耦学习方法以分离域不变表示和域特异表示, 从而促进了领域不变表示包含更多的领域无关信息.

在单阶段目标检测器上实现域自适应算法相较于双阶段检测器更为困难, 因为其缺少可以提取目标建议的区域提取网络 (Region proposal network, RPN)[9], 所以无法直接实现实例级的特征对齐. 文献 [22−24] 都是基于单阶段多检测框检测器 (Single shot multibox detector, SSD)[25] 的域自适应目标检测算法. Rodriguez 等[24] 利用伪标签自训练的思想, 先使用在源域训练好的模型在目标域推理得到伪标签, 再设计伪标签更新的策略使得模型向目标域泛化. 李威等[23] 综合源域和目标域中域不变的内容空间及域特有的属性空间表示进行多样性的图像翻译, 从而实现了一种多源域的渐进域自适应算法, 但二者[23−24] 都需要先进行源域向目标域的图像翻译, 再作为域自适应检测网络的输入进行训练, 不属于端到端的训练方式. Chen 等[22] 在图像和像素级别的对齐基础上, 构建原型特征隐式地完成实例级对齐, 但其在实例的选择上缺少目标置信度信息对实例特征进行筛选, 进而导致目标域原型存在较大的偏差. 兼具速度和精度的 YOLO (You only look once) 系列网络是广受工业界青睐的目标检测器之一, 尽管 YOLOv1 提出较早, 但 YOLO 系列检测器的发展却从未停止. 从 2015 年提出的 YOLOv1[26] 到目前最新的 YOLOv8[27], YOLO系列网络的演进更能体现出目标检测的发展. 先进的 YOLO 检测器精度和速度也已远远超过 Faster R-CNN 和 SSD 网络. Zhang 等[28] 以 YOLOv3[29] 检测器为基础实现域自适应 YOLO 目标检测算法 (Domain adaptation YOLO, DAYOLO), 但其只是简

单地将文献 [10] 中的域自适应方法迁移到 YOLOv3 上. Hnewa 等[30] 以 YOLOv4[31] 为检测器提出一种多尺度特征融合的域自适应 YOLO 目标检测网络 (Multi scale domain adaptive YOLO, MS-DAYOLO); Vidit 等[32] 以 YOLOv5[33] 作为检测器, 引入自注意力机制自适应捕获目标区域, 从而提高在目标域上的检测精度. 尽管如此, 二者都缺乏对类别特征的对齐[30, 32], 从而导致不同类别之间误对齐带来的精度下降. Li 等[34] 以 YOLOv5 作为检测器提出步进式域自适应 YOLO 目标检测算法 (Stepwise domain adaptative YOLO, S-DAYOLO), 在图像级和实例级特征对齐模块之间引入类别一致性模块, 一定程度上缓解了类别特征误对齐带来的影响.

基于此, 本文针对单阶段目标检测算法 (以 YOLO 检测器为主), 提出一种主要基于对抗特征训练的无监督域自适应单阶段目标检测算法. 首先本文设计了一种简单而有效的基于通道注意力机制的域分类器 (Channel attention domian classifier, CADC), 用于图像级特征对齐以加强图像级域不变特征的提取, 进而补充域不变信息. 该方法将 SE (Squeeze-excitation) 通道注意力机制模块[35] 与域分类器相结合, 使得网络更加关注域不变特征通道并且抑制域特异特征通道. 进一步地, 通过构造不同类别的原型特征, 设计了一种基于原型的循环域三元损失 (Cycle domain triplet loss, CDTL) 函数, 在循环域三元损失函数的指导下使不同域之间相同类别原型间的距离尽可能近, 同时使得同一个域中不同类别原型间的距离尽可能远, 进而对齐类别特征. 总的来说, 本文主要贡献如下:

1) 为了自适应地搜寻更多的具有域不变特性的特征, 提出基于通道注意力机制的图像级域分类器, 加强模型对域不变信息的学习.

2) 为了纠正特征对齐中出现的类别偏差, 设计了一种域间基于原型的循环域三元损失函数以更好地实现类别对齐, 进一步提升检测精度.

3) 通过大量实验证明本文方法的有效性, 并适用于单阶段目标检测网络, 可以为后续相关工作提供一定的参考.

1　基于注意力机制和循环域三元损失的无监督域自适应单阶段目标检测

在域自适应目标检测中往往将源域数据定义为 $\Omega_{\rm s}=\{X_{\rm s}^i, b_{\rm s}^i, y_{\rm s}^i|i=1,\cdots,N_{\rm s}\}$, 将目标域数据定义为 $\Omega_{\rm t}=\{X_{\rm t}^i|i=1,\cdots,N_{\rm t}\}$. $X_{\rm s}^i$ 和 $X_{\rm t}^i$ 分别代表在源域和目标域数据集中的第 i 幅图像, $y_{\rm s}^i\in\{1,2,\cdots,K\}$ 和 $b_{\rm s}^i$ 分别代表在源域中第 i 幅图像的类别标注和框标注, K 代表在数据集中的类别数. 本文的目标是利用已有标注的源域数据和未标注的目标域数据设计域自适应算法, 使得原检测器在目标域数据上仍具有较高的检测精度. 本文提出了基于通道注意力机制的域分类器 (CADC) 和循环域三元损失 (CDTL) 函数, 网络总体流程如图 1 所示. 图中实线代表原检测器的数据流向, 虚线代表域自适应算法的数据流向, DG代表图象级和实例级域分类器组, 包含实例级特征对齐域分类器和本文所引入的基于通道注意力机制的图像级特征对齐域分类器 CADC. 图象级和实例级域分类器组与颈部网络相对应的骨干网络特征图 F_1、F_2 和 F_3 相连接, 从而实现多尺度图像级和实例级的对齐. 同时在 F_1 的前一组卷积输出特征上引入像素级特征对齐域分类器[17], 实现浅层局部特征的对齐, 在特征图上构建循环域

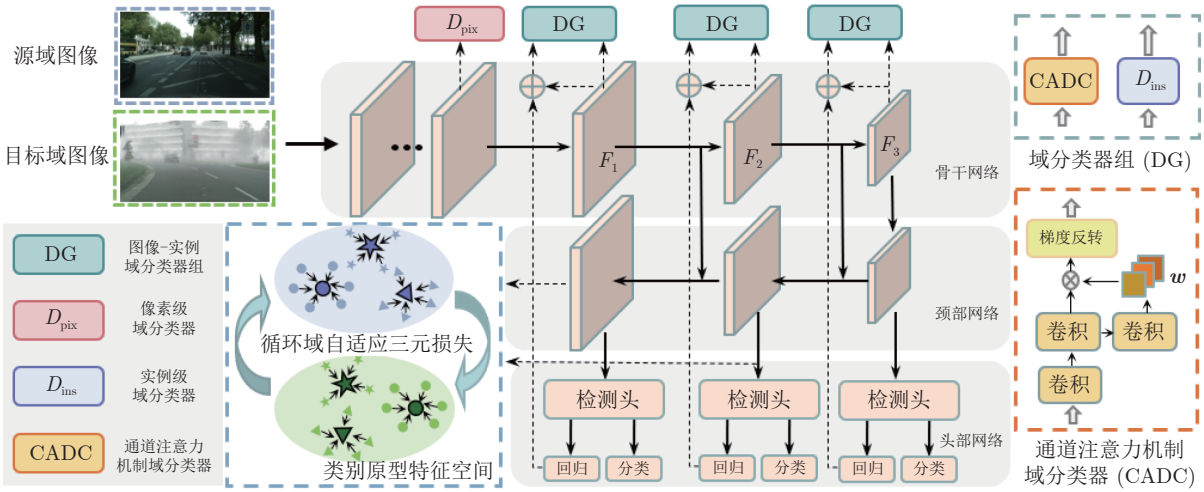


图 1　基于注意力机制和循环域三元损失的域自适应目标检测算法流程

Fig. 1　The pipeline of domain adaptive object detection based on attention mechanism and cycle domain triplet loss

三元损失函数所需的类别原型. 总的来说, 在通常的图像级对齐和实例级对齐的基础上, 本文引入的通道注意力域分类器 (CADC) 和循环域三元损失 (CDTL)函数对目标域数据和源域数据实现从图像特征到实例特征以及类别特征的分层对齐, 从而提高检测器在目标域的检测精度. 其中, CADC 可以增强域不变通道特征并同时抑制域特异通道特征, 从而使得网络能够学习到两个域之间的隐式共同特性, 进而使得图像级的特征能够较好地对齐. 而 CDTL 利用原型学习的思想, 在源域和目标域循环构建三元损失函数中的正负样本和锚示例, 最后通过最小化 CDTL实现类别对齐, 进一步提高目标检测网络在目标域数据集上的检测精度. 本文在第 1.1 节和第 1.2 节中将更为详细地对上述两种方法进行介绍.

1.1 基于通道注意力机制的域分类器

在计算机视觉中, 注意力机制可以视为一种动态选择过程, 它是根据输入的重要性自适应地加权特征来实现的. 通道注意力机制作为注意力机制的一员, 其核心思想是通过辅助网络计算每个通道对最终任务的贡献程度将其以权重的形式与网络加权, 从而使得网络偏向于对当前任务更有用的通道特征的学习. 例如 SENet[35]、ECANet[36] 和 SRM[37] 等神经网络. Wang 等[38] 提出 BatchNorm 层中较小的缩放因子所对应的通道对域自适应任务的影响较小的假设, 并对该通道权重通过剪枝正则化证明其对域自适应任务所带来的贡献较小. 受此启发, 本文直接在域自适应目标检测中的域分类器上引入通道注意力机制. 因为网络的不同通道特征对最终的域自适应具有不同的贡献程度, 并且在深度特征图中, 每个通道特征对应于输入数据的特定部分即不同的物体, 这些物体在不同域中的高维特征是相似的. 所以, 在域自适应目标检测任务中引入通道注意力机制有助于检测器学习到不同域间同类物体的共同特征. 基于此, 本文用域分类器的分类损失函数指导含有通道注意力机制的域分类器进行学习. 结合通道注意力机制的域分类器大大提升了其域判别能力, 这迫使检测器必须更加关注于两个域之间具有域不变内容特征通道的学习, 同时抑制对域自适应过程中贡献较小的通道, 以此与域判别器相抗衡, 实现特征对抗学习. 在本节中, 我们选用 SENet 中的 SE 模块对通道间的领域信息进行建模.

在以 YOLO 系列网络的颈部网络 (Neck) 部分相对应的骨干特征图 F_1, F_2 和 F_3 中引入基于通道注意力机制的图像级特征对齐域分类器. 选择这三处特征作为域分类器的输入是因为它们包含深层和丰富的多尺度语义信息, 同时其特征会跟随 Neck 部分网络不断聚合实现深层语义信息和低层纹理空间信息的融合. 本文将 SE 模块插入到域分类器中, 在训练阶段帮助网络更加关注于对域分类任务中贡献最大的通道, 进而更有效地提取到域不变特征. 因为域分类器只在训练阶段使用, 所以推理阶段该方法保持了检测器的原有结构. 具体地, 输入图像 X 通过骨干网络 (Backbone) 得到三个与 Neck 相连接的特征图 F_i ($i = 1, 2, 3$), 如式 (1) 所示.

$$F_i = Backbone(X) \tag{1}$$

其中, $Backbone$ 代表检测器的主干网络, F_i ($i = 1, 2, 3$) 为经主干网络提取到的特征层. 然后, 将其依次通过 CADC 中的卷积层 (Convolutional layer, Conv) 和 SE 模块中的平均池化 (Average pooling, AvgP)、全连接层 (Fully layer, FC) 和 Sigmoid 激活函数得到通道的域不变性权重向量 $\boldsymbol{w}_i$, 如式 (2) 所示. 最后将权重向量 $\boldsymbol{w}_i$ 与域分类器中卷积输出特征进行加权后输入到梯度反转层 (Gradient reversal layer, GRL) 中, 得到最后的域分类特征 $\boldsymbol{d}_i$ ($i = 1, 2, 3$), 如式 (3) 所示, 即

$$\boldsymbol{w}_i = Sigmoid(FC(AvgP(Conv(F_i)))) \tag{2}$$

$$\boldsymbol{d}_i = GRL(Conv(F_i)\boldsymbol{w}_i) \tag{3}$$

其中, $\boldsymbol{d}_i$ ($i = 1, 2$) 代表源域和目标域的域标签. 最后, 对这三组域分类特征分别使用交叉熵损失函数和两个 Focal 损失函数作为分类损失, 如式 (4) ~ (6) 所示, 这里 γ 为 Focal 损失函数的系数. 与 DAYOLO[28] 一样, 本文对最终的检测结果使用 ROI-Pooling 以间接获取三组实例特征 $\boldsymbol{ins}$, 并使用交叉熵损失函数作为三组实例特征的域分类损失函数, 计算过程如式 (7) 所示. 总的域分类损失 $L_{\rm DA}$ 为四者之和, 由式 (8) 计算得到.

$$L_{\rm DA1} = -\sum_i [D_i \ln(\boldsymbol{d}_1) + (1 - D_i)\ln(1 - \boldsymbol{d}_1)] \tag{4}$$

$$L_{\rm DA2} = -\sum_i [D_i(1 - \boldsymbol{d}_2)^\gamma \ln(\boldsymbol{d}_2) + (1 - D_i)\boldsymbol{d}_2^\gamma \ln(1 - \boldsymbol{d}_2)] \tag{5}$$

$$L_{\rm DA3} = -\sum_i [D_i(1 - \boldsymbol{d}_3)^\gamma \ln(\boldsymbol{d}_3) + (1 - D_i)\boldsymbol{d}_3^\gamma \ln(1 - \boldsymbol{d}_3)] \tag{6}$$

$$L_{\rm DA4} = -\sum_i [D_i \ln(\boldsymbol{ins}) + (1 - D_i)\ln(1 - \boldsymbol{ins})] \tag{7}$$

$$L_{\rm DA} = L_{\rm DA1} + L_{\rm DA2} + L_{\rm DA3} + L_{\rm DA4} \tag{8}$$

1.2　基于原型的循环域三元损失函数

在文献 [39] 中作者通过设立锚示例 $\boldsymbol{a}$、正样本示例 $\boldsymbol{p}$ 和负样本示例 $\boldsymbol{n}$ 组成以嵌入特征表示的三元组 $\{\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{n}\}$ 来构造三元损失 (Triple loss, TripleLoss) 函数, 如式 (9) 所示.

$$L = \max(d(\boldsymbol{a}, \boldsymbol{p}) - d(\boldsymbol{a}, \boldsymbol{n}) + margin, 0) \tag{9}$$

其中, 锚示例和正样本示例同类. $d(\cdot)$ 代表距离函数, 一般使用 $L2$ 范数表示. $margin$ 为超参数以控制正负样本间的距离, 同时防止模型学习到 $d(\boldsymbol{a}, \boldsymbol{p})$ 等于 $d(\boldsymbol{a}, \boldsymbol{n})$ 的特殊情况. TripleLoss 的目的是通过最小化 L 以减少锚示例和正样本示例之间的距离, 从而使得同类样本嵌入特征相互靠近、异类样本嵌入特征相互远离. 受此启发, 本文利用领域自适应中源域和目标域类别相同的固有属性, 针对实例级对齐中忽略类别信息造成不同类别特征误对齐的问题, 提出一种基于类别原型对齐的循环域三元损失函数和类别原型更新机制. 遵循原型网络[40] 的思想, 从特征图中提取类别原型 $\boldsymbol{v}_k$ 充当类别中心 k 代表类别数. 首先将源域中各类类别中心作为正样本示例 $\boldsymbol{p_i}$, $i \in \{1, 2, \cdots, K\}$, 目标域中同类样本的类别中心作为锚示例 $\boldsymbol{a_i}$, $i \in \{1, 2, \cdots, K\}$, 将源域中不同类样本的类别中心作为负样本示例 $\boldsymbol{n_i}$, $i \in \{1, 2, \cdots, K\}$, 构建三元损失函数, 进行既定迭代次数的训练. 然后交换源域和目标域, 将源域中各类类别中心作为锚示例, 而将目标域中的同类类别中心作为正样本示例, 异类类别中心作为负样本示例构建三元损失函数再次进行既定迭代次数的训练. 在整个训练过程中交替进行构成了基于原型的循环域三元损失函数, 其设计思想如图 2 所示. 图中蓝色代表源域, 橙色代表目标域, 不同类别用不同的形状表示, 其中, v_S^i 和 v_T^i 分别表示源域和目标域中第 i 个样本的类别原型.. 循环域三元损失函数可以有效缓解目标域原型构建过程中没有监督信息指导带来的误差积累, 实现更精确的类别特征对齐.

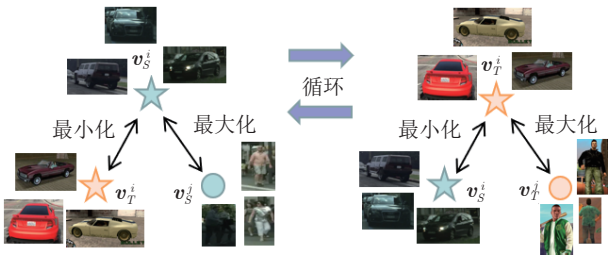


图 2　循环域三元损失函数原理

Fig. 2　Principal of cycle domain adaptive TripleLoss

具体地, 以 YOLO 检测器中骨干网络的特征图 F_1 及其对应的 YOLO 头为例. 记检测头的输出特征图为 F_{head1}, 将 F_{head1} 按照通道维度分成 3 个与锚相对应的子矩阵, 分别记为 F^q_{head1}, F^t_{head1} 和 F^v_{head1}, 然后构造类别特征矩阵 F_{cls} 和置信度特征矩阵 F_{obj}, 如式 (10) 和式 (11) 所示.

$$\begin{aligned} F_{\text{cls}} = [&F^q_{\text{head1}}(1:K), \\ &F^t_{\text{head1}}(1:K), F^v_{\text{head1}}(1:K)] \end{aligned} \tag{10}$$

$$\begin{aligned} F_{\text{obj}} = [&F^q_{\text{head1}}(1+K), \\ &F^t_{\text{head1}}(1+K), F^v_{\text{head1}}(1+K)] \end{aligned} \tag{11}$$

不同于文献 [22] 中原型的构建方式, 这里利用 YOLO 检测器独有的置信度信息矩阵 F_{obj} 对类别矩阵 F_{cls} 进行选择, 得到类别信息掩码矩阵 F_{mask}, 如式 (12) 所示.

$$F_{\text{mask}} = F_{\text{obj}} \odot F^{\text{MAX}}_{\text{cls}} \tag{12}$$

其中, $F^{\text{MAX}}_{\text{cls}}$ 通过式 (13) 得到, 即利用与每一个锚相对应的最大概率值来选择对应的类别, $\odot$ 代表哈达玛积.

$$\begin{aligned} F^{\text{MAX}}_{\text{cls}} = [&\max(F^q_{\text{head1}}(1:K)), \\ &\max(F^t_{\text{head1}}(1:K)), \max(F^v_{\text{head1}}(1:K))] \end{aligned} \tag{13}$$

值得注意的是, 此时的 F_{mask} 被类别信息所填充, 即 $F_{\text{mask}}(i, j) = l$, $l \in (1, 2, \cdots, K)$. 然后利用 F_{mask} 得到第 l 类目标在 F_1 上像素位置为 (i, j) 处的特征, 记为 F_1^l, 如式 (14) 和式 (15) 所示.

$$F_1^l = F_1 \odot F^l_{\text{mask}} \tag{14}$$

$$\begin{cases} F^l_{\text{mask}}(i, j) = 1, & F_{\text{mask}}(i, j) = l \\ F^l_{\text{mask}}(i, j) = 0, & F_{\text{mask}}(i, j) \neq l \end{cases} \tag{15}$$

最后, 通过式 (16) 可得到类别 l 对应于 F_1 的原型 $\boldsymbol{v}_l$:

$$\boldsymbol{v}_l = \frac{1}{N} \sum_{i=1}^{W} \sum_{j=1}^{H} F_1^l(i, j), \quad l \in (1, 2, \cdots, K) \tag{16}$$

其中, W 和 H 分别为 F_1^l 的宽度和高度. 由于网络训练输入受小批量 (MiniBatch) 训练数据规模的限制, 单个训练批量 (Batch) 中的有限数据所得到的原型并不能完全代表全局原型, 因此本文将单个 Batch 中得到的原型称为局部原型 $\boldsymbol{v}_l^{\text{local}}$, $l \in (1, 2, \cdots, K)$, 然后使用动量更新[41] 得到全局原型. 将动量更新参数随训练迭代次数进行自适应调整, 使得局部原型更好地拟合全局原型 $\boldsymbol{v}_l^{\text{global}}$, $l \in (1, 2, \cdots, K)$, 最终利用全局原型计算循环域三元损失函数, 其流程如算法 1 所示. 其中, Epochs 代表总训练轮数, epoch 代表加入目标域原型计算的时刻. 在训练更新过程中, 本文循环交替从源域和目标域中提取原型构建三元损失函数, 从

而减少目标域中伪标签带来的累计误差. 循环域三元损失函数计算如式 (17) 所示.

$$L_{\mathrm{CDTL}}=\begin{cases}\max(d(\boldsymbol{v}_{\mathrm{s}}^{m},\boldsymbol{v}_{\mathrm{t}}^{n})-d(\boldsymbol{v}_{\mathrm{s}}^{m},\boldsymbol{v}_{\mathrm{s}}^{n})+margin,0),\\ \qquad\qquad m\neq n,\ (iter)\bmod 3<2\\ \max(d(\boldsymbol{v}_{\mathrm{s}}^{m},\boldsymbol{v}_{\mathrm{t}}^{n})-d(\boldsymbol{v}_{\mathrm{t}}^{m},\boldsymbol{v}_{\mathrm{t}}^{n})+margin,0),\\ \qquad\qquad m\neq n,\ (iter)\bmod 3=2\end{cases}\tag{17}$$

其中, $iter$ 代表训练迭代次数, $\boldsymbol{v}_{\mathrm{s}}^{m}$ 代表源域数据 Ω_{s} 中类别 m 的全局原型特征, $\boldsymbol{v}_{\mathrm{t}}^{n}$ 代表目标域数据 Ω_{t} 中类别 n 的全局原型特征. 同理 $\boldsymbol{v}_{\mathrm{s}}^{n}$ 代表源域数据 Ω_{s} 中类别 n 的全局原型特征, $\boldsymbol{v}_{\mathrm{t}}^{m}$ 代表目标域数据 Ω_{t} 中类别 m 的全局原型特征. 本文设置从目标域中提取锚示例和从源域中提取正负示例的迭代次数为 2, 而在源域提取锚示例和在目标域提取正负示例的迭代次数为 1, 循环交替进行训练.

2 实验结果与分析

2.1 实验数据集与评价指标

2.1.1 评价指标

本文的评价指标除了使用目标检测中的平均精度 (Average precision, AP) 和平均精度均值 (Mean average precision, mAP) 进行评判外, 还提出了衡量域自适应目标检测算法域自适应能力的评价指标: 平均精度增长力 (AP growth potential, GP) 和平均精度均值增长力 (mAP growth potential, mGP). 在无监督域自适应目标检测领域, 研究人员通常仅使用源域数据进行监督训练并在目标域上进行测试的结果 (Source only, SO) 作为域自适应算法精度提升的参考. 同时将使用目标域数据进行监督训练并在目标域进行测试所得结果作为域自适应算法的目标 (Oracle). 我们可以简单地将其视为衡量域自适应算法的下限和理论上限. 但由于研究人员所选用的基础检测器不同以及训练相关参数的不同设置, 原检测器在没有域自适应算法的加持下最终的检测精度 (SO) 也会有所不同. 这使得在不同检测器和训练参数的设置下域自适应算法的比较失去了公平性. 尽管有些方法已经开始使用相较于原检测器的检测精度增量进行比较, 但又忽略了使用目标域数据进行监督训练所能达到的上限. 基于此, 平均精度增长力 (GP) 和平均精度均值增长力 (mGP) 在衡量域自适应算法带来的提升的同时, 也考虑到在基础检测器上算法精度提升的难度. 其计算如式 (18) 和式 (19) 所示.

$$GP=\frac{AP_{\mathrm{res}}-AP_{\mathrm{so}}}{AP_{\mathrm{oracle}}-AP_{\mathrm{so}}}\tag{18}$$

$$mGP=\frac{mAP_{\mathrm{res}}-mAP_{\mathrm{so}}}{mAP_{\mathrm{oracle}}-mAP_{\mathrm{so}}}\tag{19}$$

式中, AP_{res} 代表域自适应目标检测算法的检测结果.

2.1.2 实验数据集

本文实验共涉及 11 个数据集, 可将其分为 4 种实验场景: 恶劣天气场景下的领域自适应、跨摄像头的领域自适应、虚拟到现实的领域自适应以及现实到图画的领域自适应.

1) 恶劣天气场景下的领域自适应

恶劣天气场景下的领域自适应实验包括晴朗天气到浓雾天气的域自适应 CityScape→FoggyCityScape, 晴朗白天到傍晚下雨的域自适应 SunnyDay→DuskRainy 以及晴朗白天到夜晚下雨的域自适应 SunnyDay→NightRainy.

a) CityScapes: CityScapes 数据集[42] 是广泛应用于目标检测、语义分割等任务的自动驾驶场景下的数据集合. 该数据集由 Daimlerand TU Dresden 发布. 其中, 目标检测数据集共包含有 8 个类别: 汽车 (car)、卡车 (truck)、摩托车 (motor)、自行车 (bike)、火车 (train)、公共汽车 (bus)、骑手 (rider) 和人 (person). 该数据集收集于 50 个城市, 涵盖了各种各样的现实场景. 数据集包含 2 975 幅训练图像和 500 幅测试图像.

b) FoggyCityScapes: 考虑到自动驾驶场景下的复杂环境和恶劣天气的影响, 文献 [43] 使用了一个雾噪声滤波器作用于 CityScapes 数据集上, 将其渲染为雾霾场景. 与 CityScapes 一样, 该数据集包含 2 975 幅训练图像和 500 幅测试图像, 且与 CityScapes 中的数据一一对应.

c) SunnyDay, DuskRainy, NightRainy: Wu 等[21] 基于 BDD100 数据集[44] 设计了两种恶劣天气下域自适应场景, 分别为晴朗的白天到下雨的傍晚和晴朗的白天到下雨的夜晚. 其中晴朗白天数据集 (SunnyDay) 从 BDD100 数据集收集了 27 708 幅晴朗白天的图像. 下雨的傍晚数据集 (DuskRainy) 和下雨的夜晚数据集 (NightRainy) 分别包含 3 501 幅和 2 494 幅图像, 并且进行了一定程度的渲染以扩大域之间的距离. 三个数据集只包含 BDD100 中常见的 7 类交通目标, 不包含交通灯 (light)、交通牌 (sign) 以及火车 (train).

2) 跨摄像头的领域自适应

跨摄像头的域自适应实验为 KITTI→CityScapes. KITTI 数据集[45] 亦是使用范围最广的自动驾驶数据集之一, 该数据集采集了德国多个城市数小时的交通场景, 除 2D RGB 目标检测数据集外还由灰度和深度传感器采集到深度信息数据集, 在目标检测数据集中包含 7 481 幅训练图像和 7 518 幅测试图像, 共有汽车、货车、卡车、行人、骑行者、坐着的

人、有轨电车和其他 8 类. 实验中只使用汽车一个类别.

3) 虚拟到现实的领域自适应虚拟到现实的域自适应实验为 Sim10K→CityScapes.

考虑到从真实世界中收集和标注图像的困难以及高昂的成本. 2017 年, 合成数据集 Sim10K[46] 由游戏侠盗猎车手所在公司发布. 它拥有 10 K 幅图像, 但只使用一个汽车类别, 共 58 701 辆汽车实例.

4) 现实到图画的领域自适应

现实到图画的域自适应实验分别为 VOC→Clipart、VOC→Comic、VOC→Watercolor.

a) VOC: Pascal VOC[47] 是经典的真实世界目标检测数据集. 遵循文献 [22] 的实验设置, 使用 VOC-2007 和 VOC2012 的组合作为源域, 共包含 16 551 幅图像和 20 个类别.

b) Clipart1k: Clipart1k 数据集[48] 是一个与现实环境风格迥异的图画形式数据集, 包含与 VOC 相同的 20 类目标. Clipart1k 共包含 1 000 幅图像, 实验中将这 1 000 幅图像同时作为目标域的训练集和测试集.

c) Comic2k 和 Watercolor2k[48]: 二者分别为卡通和水彩风格的非现实数据集. Comic2k 和 Watercolor2k 均包含 2 000 幅图像, 其中训练集和测试集均各为 1 000 幅. 不同于 Clipart1k, Comic2k 和 Watercolor2k 只包含了 VOC 数据集中的 6 类目标.

2.1.3　实验细节

本文主要以 YOLOv3 和 YOLOv5 作为基础检测器, 验证所提出的域自适应目标检测算法的有效性. 在基于 YOLOv3 的实验中, 为确保实验的公平性, 采用 DAYOLO 中相同的实验参数配置, 即训练批量规模 (Batchsize) 为 8, 其中每个 Batch 中的 1/2 来自源域, 其余 1/2 来自目标域, 图像分辨率设置为 416× 416 像素. 另外, 本文采用与基于 YOLOv5 的域自适应方法 A-DAYOLO[32] 和 S-DAYOLO[34] 相同的实验参数设置基于 YOLOv5 的实验, 即 Batchsize 为 8、图像分辨率为 512 × 512 像素, 检测模型为 YOLOv5 系列模型中的 small 版本 (YOLOv5s). 除此之外保留所有原 YOLOv3 和 YOLOv5 的参数设置和网络结构. 以上所有实验均采用单张 RTX 3 090 显卡在 Ubuntu18.0, Pytorch1.8.1, CUDA 11.1 的环境下完成.

2.2　对比实验与分析

实验 1. CityScapes→Foggy CityScapes 恶劣天气场景下目标检测实验. 首先以 CityScape 数据集作为源域、Foggy CityScapes 作为目标域, 基于 YOLOv3 和 YOLOv5s 检测器的实验结果分别如表 1 所示. 从实验结果中可以看出, 本文方法超过目前最好的基于 YOLOv3 的域自适应算法 DAYOLO 2.3%, 达到了 38.3% 的检测精度. 同时, mGP 达到了 83.9%, 相较于 DAYOLO 提高了 22.9%. 在基于 YOLOv5 的算法中, 本文方法达到了 34.3% 的 mAP 和 83.8% 的 mGP, 精度增长力指标远高于目前已知的最优方法 S-DAYOLO. 实验证明了本文方法的有效性, 同时也说明本文方法可以适配于不用的 YOLO 系列网络. 图 3 展示了本文方法在 Foggy CityScapes 数据集的检测主观结果. 图中第 1 列为 SO 的检测结果, 第 2 列代表本文方法在 YOLOv3 上的检测结果, 第 3 列代表本文方法在 YOLOv5s 上的检测结果, 第 4 列为标签真值 (Ground truth, GT). 从主观结果可以看出, 本文方法在一定程度上弥补了在源域进行训练、在目标

表 1　不同方法在 CityScapes→Foggy CityScapes 数据集上的对比实验结果 (%)
Table 1　The results of different methods on CityScapes→Foggy CityScapes (%)

方法	检测器	person	rider	car	truck	bus	motor	bike	train	mAP	mGP
DAF[10]	Faster R-CNN	25.0	31.0	40.5	22.1	35.3	20.0	27.1	20.2	27.7	38.8
SWDA[11]	Faster R-CNN	29.9	42.3	43.5	24.5	36.2	30.0	35.3	32.6	34.3	70.0
C2F[14]	Faster R-CNN	34.0	46.9	52.1	30.8	43.2	34.7	37.4	29.9	38.6	79.1
CAFA[16]	Faster R-CNN	41.9	38.7	56.7	22.6	41.5	24.6	35.5	26.8	36.0	81.9
ICCR-VDD[21]	Faster R-CNN	33.4	44.0	51.7	33.9	52.0	34.2	36.8	34.7	40.0	—
MeGA[20]	Faster R-CNN	37.7	49.0	52.4	25.4	49.2	34.5	39.0	46.9	41.8	91.1
DAYOLO[28]	YOLOv3	29.5	27.7	46.1	9.1	28.2	12.7	24.8	4.5	36.1	61.0
本文方法 (v3)	YOLOv3	34.0	37.2	55.8	31.4	44.4	22.3	30.8	50.7	38.3	**83.9**
MS-DAYOLO[31]	YOLOv4	39.6	46.5	56.5	28.9	51.0	27.5	36.0	45.9	41.5	68.6
A-DAYOLO[32]	YOLOv5	32.8	35.7	51.3	18.8	34.5	11.8	25.6	16.2	28.3	—
S-DAYOLO[34]	YOLOv5	42.6	42.1	61.9	23.5	40.5	24.4	37.3	39.5	39.0	69.9
本文方法 (v5)	YOLOv5s	30.9	37.4	53.3	23.8	39.5	24.2	29.9	35.0	34.3	**83.8**

注: “–”表示该方法没有进行此实验; (v3) 表示检测器为 YoLov3; (v5) 表示检测器为 YoLov5s; 加粗数值表示对比实验中的最佳结果.

图 3 本文方法在 CityScapes→Foggy CityScapes 上的主观检测结果

Fig. 3 The subjective results of our method on CityScapes→Foggy CityScapes

域进行测试时存在漏检的不足, 但相对于标签真值仍存在一定的误检.

实验 2. SunnyDay→DuskRainy 恶劣天气场景下目标检测实验. 以晴朗白天数据集 SunnyDay 作为源域、下雨的傍晚数据集 DuskRainy 作为目标域进行实验. 基于 YOLOv3 和 YOLOv5s 检测器的实验结果分别如表 2 所示. 由于 DuskRainy 并没有测试集, 故这里用平均精度增量来对比实验结果. 从实验结果中可以看出, 基于 YOLOv3 的本文方法取得了 40.2% 的最高检测精度, 相对于 SO 涨幅 7.4%. 基于 YOLOv5s 的本文方法取得了 36.5% 的检测精度, 相对于 SO 涨幅 9.4%, 与目前该数据集上最佳方法 ICCR-VDD 的涨幅接近.

实验 3. SunnyDay→NightRainy 恶劣天气场景下目标检测实验. 以晴朗白天数据集 SunnyDay 作为源域、下雨的夜晚数据集 NightRainy 作为目标域进行实验. 基于 YOLOv3 和 YOLOv5s 检测器的实验结果分别如表 3 所示. 类似地, 由于 NightRainy 并未提供测试集, 所以这里仍然用平均精度增量对比不同方法的实验结果. 从实验结果中可以看出, 基于 YOLOv3 的本文方法取得了 25.3% 的最高检测精度, 相对于 SO 涨幅 5.1%. 基于 YOLOv5s 的本文方法取得了 21.5% 的检测精度, 相对于 SO 涨幅 4.7%. SunnyDay→DuskRainy 和 SunnyDay→NightRainy 的主观实验结果如图 4 所示, 图中前两行为 SunnyDay→DuskRainy 域自适应结果, 后两行为 SunnyDay→NightRainy 域自适应结果. 可以看到在本文方法的加持下原 YOLOv3 和 YOLOv5 检测器在低光照雨天的恶劣天气环境下仍然有不错的检测效果.

实验 4. KITTI→CityScapes 跨摄像头场景目标检测实验. KITTI 和 CityScapes 数据集分别是

表 2 不同方法在 SunnyDay→DuskRainy 数据集上的对比实验结果 (%)

Table 2 The results of different methods on SunnyDay→DuskRainy (%)

方法	检测器	bus	bike	car	motor	person	rider	truck	mAP	ΔmAP
DAF[10]	Faster R-CNN	43.6	27.5	52.3	16.1	28.5	21.7	44.8	33.5	5.2
SWDA[11]	Faster R-CNN	40.0	22.8	51.4	15.4	26.3	20.3	44.2	31.5	3.2
ICCR-VDD[21]	Faster R-CNN	47.9	33.2	55.1	26.1	30.5	23.8	48.1	37.8	**9.5**
本文方法 (v3)	YOLOv3	50.1	24.9	70.7	24.2	39.1	19.0	53.2	**40.2**	7.4
本文方法 (v5)	YOLOv5s	46.2	22.1	68.2	16.5	34.8	17.5	50.5	36.5	9.4

注: ΔmAP 表示 mAP 的涨幅程度.

表 3　不同方法在 SunnyDay→NightRainy 数据集上的对比实验结果 (%)
Table 3　The results of different methods on SunnyDay→NightRainy (%)

方法	检测器	bus	bike	car	motor	person	rider	truck	mAP	ΔmAP
DAF[10]	Faster R-CNN	23.8	12.0	37.7	0.2	14.9	4.0	29.0	17.4	1.1
SWDA[11]	Faster R-CNN	24.7	10.0	33.7	0.6	13.5	10.4	29.1	17.4	1.1
ICCR-VDD[21]	Faster R-CNN	34.8	15.6	38.6	10.5	18.7	17.3	30.6	23.7	**7.4**
本文方法 (v3)	YOLOv3	45.0	8.2	51.1	4.0	20.9	9.6	37.9	**25.3**	5.1
本文方法 (v5)	YOLOv5s	40.7	9.3	45.0	0.6	12.8	9.2	32.5	21.5	4.7

SO　　本文方法 (YOLOv3)　　本文方法 (YOLOv5s)　　GT

图 4　本文方法在 SunnyDay→DuskRainy 和 SunnyDay→NightRainy 上的主观检测结果
Fig. 4　The subjective results of our method on SunnyDay→DuskRainy and SunnyDay→NightRainy

由不同的摄像头捕捉而成, 具有不同的视角、尺度和环境信息. 以 KITTI 数据集作为源域、CityScapes 作为目标域, 基于 YOLOv3 和 YOLOv5s 检测器的实验如表 4 所示. 由于 KITTI 训练集图像数量远大于 CityScapes 的训练集图像数量, 且在 KITTI→CityScapes 实验中仅涉及出现最多的汽车类, 故相较于实验 1 的多目标检测而言, 该检测任务的分类分支为二分类, 所以误检较少, 更容易达到较高精度. 实验结果表明, 本文方法在 YOLOv3 和 YOLOv5s 的检测器上分别达到 61.1% 和 60.0% 的最高检测精度以及 29.4% 和 50.4% 的精度增长力.

实验 5. Sim10k→CityScapes 虚拟和现实场景的目标检测实验. Sim10k→CityScapes 上的实验具有很大的应用价值, 因为现实场景下数据的收集和标注是高成本的, 而在虚拟仿真环境下则可以很容易地获取到数据的标注信息. 通过域自适应算法使

表 4　KITTI→CityScapes 和 Sim10k→CityScapes 数据集上的对比实验结果 (%)
Table 4　The results of different methods on KITTI→CityScapes and Sim10k→CityScapes (%)

方法	KITTI→CityScapes		Sim10k→CityScapes	
	AP	GP	AP	GP
DAF[10]	38.5	21.0	39.0	22.5
SWDA[11]	37.9	19.5	42.3	30.8
C2F[14]	—	—	43.8	35.3
CAFA[16]	43.2	32.9	49.0	47.7
MeGA[20]	43.0	32.4	44.8	37.0
DAYOLO[28]	54.0	**82.2**	50.9	39.5
本文方法 (v3)	**61.1**	29.4	**60.8**	37.1
A-DAYOLO[32]	37.7	—	44.9	—
S-DAYOLO[34]	49.3	52.9	—	—
本文方法 (v5)	60.0	50.4	60.3	**56.3**

得在虚拟仿真环境下训练好的模型在真实环境也能取得不错的检测精度. 类似于 KITTI→CityScapes 实验, Sim10k→CityScapes 也仅涉及一个类别汽车, 因此其检测精度提升较小. 实验结果如表 4 所示. 从表 4 中可见, 本文方法以 YOLOv3 和 YOLOv5s 作为检测器, 分别达到 60.8% 和 60.3% 的检测精度, 同时精度增长度为 37.1% 和 56.3%. 其中精度增长度在所有方法中达到最高.

2.3 消融实验与分析

2.3.1 CADC 和 CDTL 的消融实验

在消融实验中, SO 代表仅用源域训练集进行训练并在目标域测试集上进行测试, Oracle 代表仅使用目标域训练集进行训练并在目标域进行测试, CADC 和 CDTL 分别表示单独使用通道注意力域分类器和循环域三元损失函数.

在 CityScapes→FoggyCityscapes 上的消融实验分别如表 5 和表 6 所示. 结果表明, 当只使用域通道注意力机制分类器 (CADC) 的时候, 在 YOLOv3 和 YOLOv5s 上分别提升了 8.8% 和 12.7% 的平均精度均值 (mAP), 充分证明了基于通道注意力机制的域分类器在 CityScapes→FoggyCityscapes 实验上的有效性. 当在 YOLOv3 和 YOLOv5 上仅使用循环域三元损失 (CDTL) 函数时, 平均精度也能得到 2.1% 和 4.8% 的提升. 而当二者共同作用下时, 在 YOLOv3 和 YOLOv5s 上相较于 SO 分别增加了 9.9% 和 12.9% 的平均精度, 同时在 YOLOv3 检测器上对火车类的检测甚至超过 Oracle, 达到 50.7% 的最好成绩, 进一步证明了两种方法的相辅相成.

在 SunnyDay→DuskRainy 上的消融实验结果分别如表 7 和表 8 所示. 在以 YOLOv3 作为检测器的实验中, 使用源域 SunnyDay 数据集训练模型直接应用到目标域 DuskRainy 中的检测精度为 32.8% (即表中的 SO). 另外, 对于单独加入通道注意力机制的域分类器 (即表中的 CADC), 检测精度提高到 39.6%, 而当为网络仅加入循环域三元损失进行训练时 (即表中的 CDTL), 网络检测精度提高到 35.7%. 当同时加入本文提出的通道注意力机制的域分类器和循环域三元损失函数时, 网络可以达到最高的检测精度 40.2%. 在以 YOLOv5s 作为检测器的实验中, 使用源域 SunnyDay 数据集训练模型直接应用到目标域 DuskRainy 的检测精度为 27.1% (即表中的 SO). 对于单独加入通道注意力机制的域分类器时 (即表中的 CADC), 检测精度为 35.9%, 相较于 SO 提高了 8.8%. 而当为网络仅加入循环域三元损失进行训练时 (即表中的 CDTL), 检测精度为 30.4%, 相较于 SO 提高了 3.3%. 当同时加入本文提出的通道注意力机制域分类器和循环域三元损失函数时, 网络可以达到最高的检测精度 36.5%.

在 SunnyDay→NightRainy 上的消融实验结果分别如表 9 和表 10 所示. 在以 YOLOv3 作为检测器的实验中, 使用源域 SunnyDay 数据集训练模型直接应用到目标域 NightRainy 的检测精度为 20.2% (即表中的 SO). 当加入通道注意力域分类器 (即表中的 CADC) 时, 准确率提高到 24.8%, 而当单独加入循环域三元损失 (CDTL) 函数进行训练时, 准确率提高到 21.7%. 当同时加入本文提出的通道注意力机制域分类器和循环域三元损失函数时, 达到最

表 5 CityScapes→FoggyCityscapes 数据集上基于 YOLOv3 的消融实验结果 (%)
Table 5 The results of ablation experiment on CityScapes→FoggyCityscapes based on YOLOv3 (%)

方法	person	rider	car	truck	bus	motor	bike	train	mAP
SO	29.8	35.0	44.7	20.4	32.4	14.8	28.3	21.6	28.4
CADC	**34.4**	38.0	54.7	24.4	**45.0**	21.2	**32.1**	49.1	37.2
CDTL	31.1	**38.0**	46.7	28.9	34.5	**23.4**	27.8	13.7	30.5
CADC + CDTL	34.0	37.2	**55.8**	**31.4**	44.4	22.3	30.8	**50.7**	**38.3**
Oracle	34.9	38.8	55.9	25.3	45.0	22.6	33.4	49.1	40.2

表 6 CityScapes→FoggyCityscapes 数据集上基于 YOLOv5s 的消融实验结果 (%)
Table 6 The results of ablation experiment on CityScapes→FoggyCityscapes based on YOLOv5s (%)

方法	person	rider	car	truck	bus	motor	bike	train	mAP
SO	26.9	33.1	39.9	8.9	21.1	11.3	24.8	4.9	21.4
CADC	**32.6**	37.1	52.7	**26.8**	38.1	23.0	**38.1**	32.6	34.1
CDTL	29.7	36.7	43.2	13.1	25.5	17.1	28.7	13.1	26.2
CADC + CDTL	30.9	**37.4**	**53.3**	23.8	**39.5**	**24.2**	29.9	**35.0**	**34.3**
Oracle	34.8	37.9	57.5	24.4	42.7	23.1	33.2	40.8	36.8

表 7 SunnyDay→DuskRainy 数据集上基于 YOLOv3 的消融实验结果 (%)
Table 7 The results of ablation experiment on SunnyDay→DuskRainy based on YOLOv3 (%)

方法	bus	bike	car	motor	person	rider	truck	mAP
SO	43.7	14.3	68.4	12.0	31.5	10.9	48.7	32.8
CADC	50.0	22.6	**70.8**	23.2	38.4	18.7	**53.5**	39.6
CDTL	45.4	20.1	69.2	15.2	34.8	17.2	47.8	35.7
CADC + CDTL	**50.1**	**24.9**	70.7	**24.2**	**39.1**	**19.0**	53.2	**40.2**

表 8 SunnyDay→DuskRainy 数据集上基于 YOLOv5s 的消融实验结果 (%)
Table 8 The results of ablation experiment on SunnyDay→DuskRainy based on YOLOv5s (%)

方法	bus	bike	car	motor	person	rider	truck	mAP
SO	37.2	8.4	63.8	5.5	23.7	7.9	43.4	27.1
CADC	45.6	22.1	68.2	**16.6**	34.5	15.4	50.1	35.9
CDTL	41.6	13.1	65.5	7.6	29.7	10.2	44.9	30.4
CADC + CDTL	**46.2**	**22.1**	**68.2**	16.5	**34.8**	**17.5**	**50.5**	**36.5**

表 9 SunnyDay→NightRainy 数据集上基于 YOLOv3 的消融实验结果 (%)
Table 9 The results of ablation experiment on SunnyDay→NightRainy based on YOLOv3 (%)

方法	bus	bike	car	motor	person	rider	truck	mAP
SO	39.2	5.1	44.2	0.2	14.8	6.9	30.7	20.2
CADC	44.4	8.1	50.9	0.6	20.2	**11.3**	**38.3**	24.8
CDTL	40.4	8.2	45.8	0.6	16.2	7.2	33.4	21.7
CADC + CDTL	**45.0**	**8.2**	**51.1**	**4.0**	**20.9**	9.6	37.9	**25.3**

表 10 SunnyDay→NightRainy 数据集上基于 YOLOv5s 的消融实验结果 (%)
Table 10 The results of ablation experiment on SunnyDay→NightRainy based on YOLOv5s (%)

方法	bus	bike	car	motor	person	rider	truck	mAP
SO	25.4	3.2	36.3	0.2	9.1	4.4	20.8	14.2
CADC	38.7	8.3	42.7	0.3	12.3	6.4	32.0	20.1
CDTL	34.3	6.2	44.2	0.5	11.2	8.7	30.3	19.3
CADC + CDTL	**40.7**	**9.3**	**45.0**	**0.6**	**12.8**	**9.2**	**32.5**	**21.5**

高的检测精度 25.3%. 在以 YOLOv5s 作为检测器的实验中, 使用源 SunnyDay 训练模型直接应用到目标域 NightRainy 的检测精度为 14.2% (即表中的 SO). 当单独加入通道注意力机制域分类器 (即表中的 CADC) 时, 检测精度为 20.1%, 相较于 SO 提高了 5.9%, 而当单独加入循环域三元损失 (即表中的 CDTL) 时, 检测精度为 19.3%, 相较于 SO 提高了 5.1%. 当同时加入本文提出的通道注意力机制域分类器和循环域三元损失函数时, 达到最高的检测精度 21.5%.

KITTI→CityScapes 的消融实验结果如表 11 的第 1 列数据所示, 从表 11 中可以看出, YOLOv3 在使用源域 KITTI 进行训练并在目标域 CityScapes 上进行测试的精度为 59.6% (SO), 使用目标域 CityScapes 进行训练并在目标域 CityScapes 上进行测试的精度为 64.7% (Oracle), 仅存在 5.1% 的提升空间. YOLOv5s 在使用源域 KITTI 进行训练并在目标域 CityScapes 上进行测试的精度为 54.0% (SO), 使用目标域 CityScapes进行训练并在目标域 CityScapes 上进行测试的精度为 65.9% (Oracle). 从表 11 中可以看出, CADC 和 CDTL 两种方法单独使用所带来的精度提升相当. 在 YOLOv3 上的精度均为 60.5%, 在 YOLOv5s 上的精度分别为 59.5% 和 59.0%, 当二者共同作用时使得精度得到进一步的提升, 在 YOLOv3 和 YOLOv5s 的检测器上分别达到 61.1% 和 60% 的最佳性能, 相较于 SO 提高了 1.5% 和 6%.

Sim10k→CityScapes 的消融实验如表 11 的第 2 列数据所示, 当同时加入循环域三元损失函数和域通道注意力分类器时, 在 YOLOv3 和 YOLOv5s

表 11 KITTI→CityScapes 和 Sim10k→CityScapes 数据集上的对比实验结果 (%)

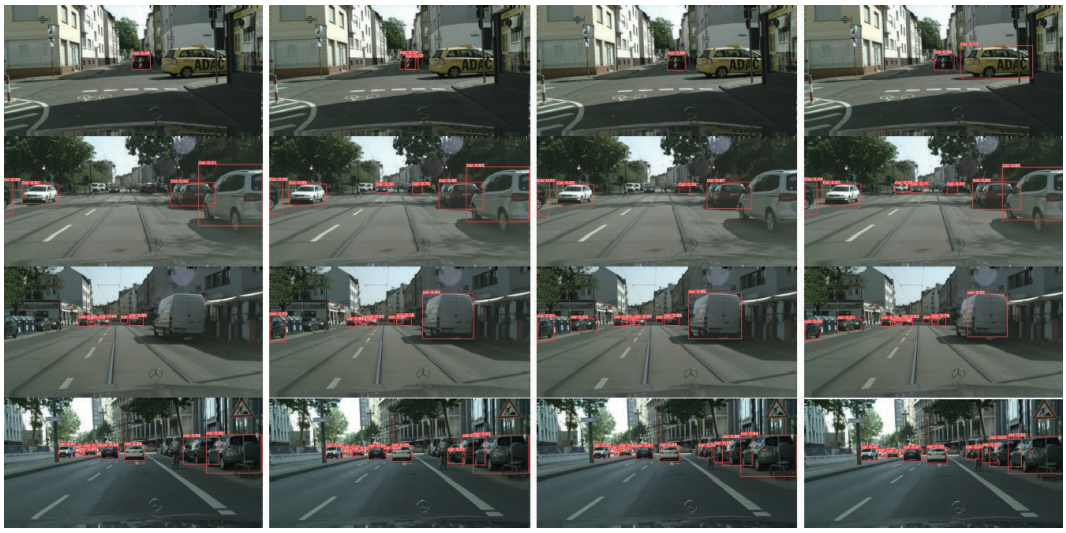
Table 11 The results of different methods on KITTI→CityScapes and Sim10k→CityScapes (%)

	方法	KITTI	Sim10k
YOLOv3	SO	59.6	58.5
	CADC	60.5	59.6
	CDTL	60.5	**60.8**
	CADC + CDTL	**61.1**	59.8
	Oracle	64.7	64.7
YOLOv5s	SO	54.0	53.1
	CADC	59.5	58.6
	CDTL	59.0	**60.3**
	CADC + CDTL	**60.0**	59.0
	Oracle	65.9	65.9

检测器上分别达到了 59.8% 和 59.0% 的检测精度. 相对于 SO 的实验结果, 分别提升了 1.3% 和 5.9%. 但在两个检测器上单独使用循环域三元损失函数的所能达到的精度均超过了与域通道注意力分类器共同作用下的精度. 我们认为这种现象的产生是因为 Sim10k 和 CityScapes 风格迥异, 即两个域之间的距离较大. 当域分类器作用时, 网络会更加注重于拉近两个域得到域不变特征, 从而疏忽了对具有判别性的实例特征的学习. 循环域三元损失 (CDTL) 从类别原型更新的角度为这类单目标检测任务提供了另一种不依赖于域分类器的解决方案.

图 5 展示了 KITTI→CityScapes 和 Sim10k→CityScapes 任务中以 YOLOv5s 作为检测器, 单独使用通道注意力机制的域分类器 (CADC) 和循环域三元损失 (CDTL) 以及共同使用两者情况下的主观检测结果. 图中每一行分别为 CityScapes 测试集中不同的场景, 前两行代表 KITTI→CityScapes 的主观实验结果, 后两行代表 Sim10k→CityScapes 的主观实验结果. 第 1 列代表 SO 的检测结果, 即不加入任何改进方法; 第 2 列表示只加入通道注意力机制的域分类器 (CADC) 得到的结果; 第 3 列表示只加入循环域三元损失 (CDTL) 得到的结果; 第 4 列表示二者共同作用的结果. 从图中可以看出, 无论是 KITTI→CityScapes 还是 Sim10k→CityScapes 上的实验, 当单独使用通道注意力机制的域分类器 (CADC) 使得网络更加注重于检测到尽可能多的目标; 而单独使用循环域三元损失 (CDTL) 所检测的目标则更为准确, 即置信度更高. 当二者共同作用下能够检测到更多具有较高置信度的目标, 弥补了漏检带来的精度降低.

此外, 为进一步验证基于通道注意力机制的域分类器和循环域三元损失函数的有效性, 本文将其直接加入以 SSD 为检测器的 I3Net 进行对应的三组域自适应目标检测实验: VOC→Clipart1k, VOC→Comic2k 和 VOC→Watercolor2k. 本文遵循 I3Net 的实验设置进行复现以确保实验对比的公平性, 实验结果分别如表 12 ~ 14 所示. 表中 CADC*表示将 SE 通道注意力机制加入到 I3Net 的域分类器中的方法. 从表 12 中 VOC→Clipart1k 的域自适应实验可以看出, 当在 I3Net 的训练过程中加入循环域三元损失函数时, 准确率提升了 1%. 当基于通道注



SO 本文方法 (YOLOv3) 本文方法 (YOLOv5s) GT

图 5 本文方法在 KITTI→CityScapes 和 Sim10k→CityScapes 上的消融实验结果

Fig. 5 The ablation experimental results of our method on KITTI→CityScapes and Sim10k→CityScapes

意力机制的域分类器和循环域三元损失函数同时加入训练时, 模型性能可以提升 2.6%. 但是在表 13 的 VOC→Comic2k 和表 14 的 VOC→Watercolor2k 实验中, 当 I3Net 训练时仅加入循环域三元损失时造成了精度的下降.

经分析, 原因如下: VOC 数据集包括 20 个类, 而 Comic2k 和 Watercolor2k数据集中仅包含 VOC 数据集中的 6 个类, 在 I3Net 的训练时, 并没有删除 VOC 中与 Comic2k 和 Watercolor2k 数据集不同类别的标注信息. 但本文提出的基于原型的循环域三元损失的方法仍然会根据源域类别构造 20 个类别原型, 这就使得网络隐式地对 Comic2k 和 Watercolor2k 数据集外的类别进行学习, 从而降低了在其测试集上的检测精度. 值得一提的是, 与通道注意力机制域分类器的结合弥补了循环域三元损失带来的精度丢失, 使得添加了通道注意力机制的 I3Net 在 VOC→Comic2k 得到了提升, 这也证明了通道注意力机制域分类器的有效性. 图 6 展示了本文方法加入到 I3Net 后的主观实验结果, 可以看到本文方法能显著提高目标检测效果. 主观和客观实验结果也表明本文方法除适用于 YOLO 目标检测网络外, 也适配于 SSD 等单阶段目标检测网络.

2.3.2　像素级对齐域分类器 D_{pixel} 的消融实验

相较于现有大多数域自适应目标检测工作中以 VGG16 作为骨干网络的 Faster R-CNN 检测器, YOLO中的骨干网络 DarkNet 具有残差连接和更深的网络结构, 因为 YOLOv3 和 YOLOv5s 中与 Neck 部分所连接的三处骨干网络 (Backbone) 特征层所在网络位置的深度已经大于 VGG16, 因此可以将其输出视为网络的深层特征. 基于此, 本文构建消融实验来验证是否在浅层网络即与 Neck 部分所连接的第一个特征层之前实现像素级对齐的必要性. 实验结果如表 15 所示, 表 15 中 D_{pixel} 代表是否在网络浅层特征处加入像素级对齐域分类器. 实验结果表明, 当加入像素级对齐域分类器时, 在基于YOLOv3和 YOLOv5s 的检测器上检测精度都能得到不同程度的提升, 从而证明了像素级对齐对于 YOLO 系列网络的必要性.

2.3.3　通道注意力域分类器 (CADC) 中损失函数选择实验

Focal 损失函数在网络不同位置的作用不尽相同, 从而对网络带来不同的影响[13]. 基于此, 本文针对 YOLO网络对不同通道注意力域分类器上使用 Focal 损失函数所带来的影响在 CityScapes→Foggy CityScapes 任务上进行实验验证, 实验中不涉及循环域三元损失函数, 结果如表 16 所示. 表 16 中, F_1, F_2, F_3 表示 YOLOv3 或者 YOLOv5s 中与 Neck 相连接的 Backbone 特征层, CE 代表交叉熵损失函数 (Cross-entropy loss function), FL 代表 Focal 损失函数 (Focal loss function). 实验结果表明, 在 YOLOv3和 YOLOv5s 上, F1 所对应的域分类器损失函数选择交叉熵损失 (CE), F2 和 F3 所

表 12　本文方法在 VOC→Clipart1k 上的实验 (%)
Table 12　The experiment on VOC→Clipart1k (%)

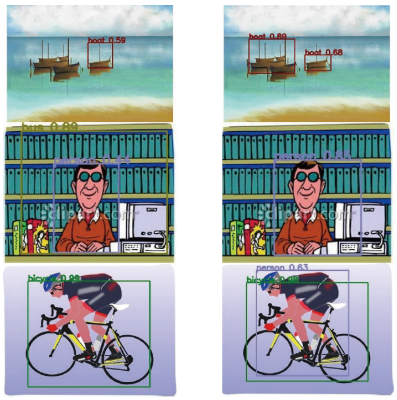
方法	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	bike	prsn	plnt	sheep	sofa	train	tv	mAP
I3Net	23.7	**66.2**	25.3	19.3	23.7	55.2	35.7	13.6	37.8	35.5	25.4	13.9	24.1	60.3	56.3	**39.8**	13.6	34.5	**56.0**	41.8	35.1
I3Net + CDTL	23.3	61.6	27.8	17.1	24.7	54.3	39.8	12.3	**41.4**	34.1	**32.2**	15.5	**27.6**	**77.9**	**57.0**	37.4	5.50	31.3	51.8	47.8	36.0
I3Net + CDTL + CADC*	**31.2**	60.4	**31.8**	**19.4**	**27.0**	**63.3**	**40.7**	**13.7**	41.1	**38.4**	27.2	**18.0**	25.5	67.8	54.9	37.2	**15.5**	**36.4**	54.8	**47.8**	**37.6**

表 13　本文方法在 VOC→Comic2k 上的实验 (%)
Table 13　The experiment on VOC→Comic2k (%)

方法	bike	bird	car	cat	dog	person	mAP
I3Net	44.9	**17.8**	31.9	10.7	23.5	46.3	29.2
I3Net + CDTL	43.7	15.1	31.5	11.7	18.6	**46.9**	27.9
I3Net + CDTL + CADC*	**47.8**	16.0	**33.8**	**15.1**	**24.4**	43.5	**30.1**

表 14　本文方法在 VOC→Watercolor2k 上的实验 (%)
Table 14　The experiment on VOC→Watercolor2k (%)

方法	bike	bird	car	cat	dog	person	mAP
I3Net	81.3	**49.6**	43.6	**38.2**	31.3	**61.7**	**51.0**
I3Net + CDTL	79.5	47.2	41.7	33.5	**35.4**	60.3	49.6
I3Net + CDTL + CADC*	**84.1**	45.3	**46.6**	32.9	31.4	61.4	50.3



I3Net I3Net + 本文方法

图 6 本文方法在 VOC→Clipart1k 上的主观结果

Fig. 6 The subjective results of our method on VOC→Clipart1k

表 15 像素级对齐对网络的影响 (%)

Table 15 The impact of pixel alignment to network (%)

方法	检测器	C→F	K→C	S→C
CDTL + CADC	YOLOv3	35.9	59.8	58.4
CDTL + CADC + D_{pixel}	YOLOv3	**37.2**	**60.5**	**59.6**
CDTL + CADC	YOLOv5	32.7	58.9	56.8
CDTL + CADC + D_{pixel}	YOLOv5	**34.1**	**59.5**	**58.6**

对应的域分类器损失函数选择 Focal 损失 (FL) 时分别达到最高的检测精度 37.2% 和 34.1%. Focal 损失函数的使用很好地缓解了域自适应目标检测过程中深层特征存在的类别不平衡的问题.

2.3.4 循环域三元损失函数的循环迭代次数 (iter) 实验

为缓解目标域伪标签误差累积造成的精度下降, 本文提出循环域三元损失函数, 即目标域和源域的原型分别作为正负样本示例和锚示例, 在既定的训练迭代次数后交换目标域和源域的原型, 将其作为锚示例和正负样本示例. 这种交替训练的策略可以使有标签的源域原型对伪标签目标域原型进行一定的纠正, 从而减小伪标签给训练带来的误导. 本文对不同循环迭代次数 iter 及所带来的影响进行实验探索, 实验结果如图 7 所示. 图中横坐标的 S 代表源域训练迭代数, T 代表目标域训练迭代数, S/T 表示以源域类别原型作为正负样本示例、目标域类别原型作为锚示例训练既定轮数后交换目标域和源域的原型分别作为锚示例和正负样本示例训练迭代既定次数. 从图 7 中可以看出, 当循环迭代次数 S/T 为 2/1 即以源域类别原型作为正负样本示例、目标域类别原型作为锚示例训练两轮, 以目标域类别原型作为正负样本示例、源域类别原型作为锚示例训练迭代 1 轮时, 在 YOLOv3 和 YOLOv5s 上能达到最佳的精度 30.5% 和 26.2%; 当源域类别原型作为正负样本示例、目标域类别原型作为锚示例训练轮数超过 2 轮时, 网络会更加倾向于源域中类别原型的学习, 而忽略了目标域中原型对域不变特征提取的帮助导致精度下降. 相反, 如果目标域类别原型作为正负样本示例、源域类别原型作为锚示例的训练轮数的增加也会导致精度的下降, 即误差的积累不能很好地得到修正. 实验表明, 无论是基于 YOLOv3 还是 YOLOv5 检测器, 当源域充当正样本示例和锚示例, 目标域充当负样本示例, 二者训练迭代次数分别为 2 和 1 时检测精度最佳.

表 16 通道注意力域分类器中损失函数的选择

Table 16 The choice of loss function in channel attention domain classifier

检测器	F_1	F_2	F_3	mAP (%)
YOLOv3/v5	CE	CE	CE	35.8/32.7
YOLOv3/v5	CE	CE	FL	36.4/33.2
YOLOv3/v5	CE	FL	FL	**37.2/34.1**
YOLOv3/v5	FL	FL	FL	37.0/33.5

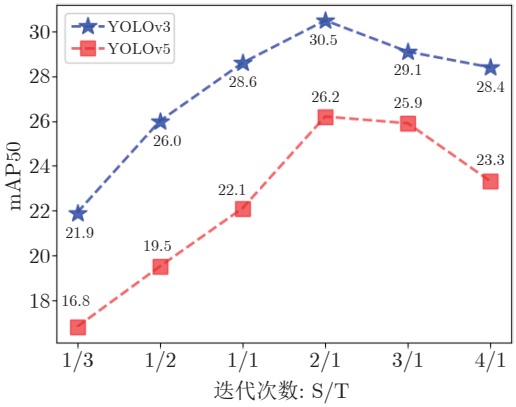


图 7 不同循环迭代训练次数在 YOLOv3 和 YOLOv5 检测器上的结果

Fig. 7 The result of different cycle iteration on YOLOv3 and YOLOv5

3 结束语

本文提出一种基于注意力机制和循环域三元损失函数的无监督域自适应单阶段目标检测算法. 首先通过在图像级域分类器中引入通道注意力机制, 使得网络更加关注于域不变特征的学习. 其次设计了一种适用于域自适应的三元损失函数引导网络实现基于类别原型的特征对齐. 分别在单阶段目标检测器 YOLOv3, YOLOv5 和 SSD 上进行实验以证明本文方法对单阶段目标检测网络的适配性. 在众多域自适应目标检测公共数据集的实验结果表明, 本文的方法在基于 YOLO 的域自适应目标检测网

络中取得最好的结果, 同时对基于 SSD 的域自适应目标检测网络也能带来精度的提升. 尽管如此, 本文所提出的循环域三元损失函数依赖于前期目标域原型伪标签的准确性, 当目标域原型伪标签误差较大时使用该方法并不能得到一个很好的检测效果. 未来可以尝试在循环域三元损失函数中使用图来表示类别中心, 从而避免原型构建过程中误差累计导致精度的丢失.

References

1 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: NIPS, 2012. 1106−1114

2 Bottou L, Bousquet O. The tradeoffs of large scale learning. In: Proceedings of the 20th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2007. 161−168

3 Shen J, Qu Y R, Zhang W N, Yu Y. Wasserstein distance guided representation learning for domain adaptation. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. 4058−4065

4 Gao Jun, Huang Li-Li, Sun Chang-Yin. A local weighted mean based domain adaptation learning framework. *Acta Automatica Sinica*, 2013, **39**(7): 1037−1052
(皋军, 黄丽莉, 孙长银. 一种基于局部加权均值的领域自适应学习框架. 自动化学报, 2013, **39**(7): 1037−1052)

5 Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 2016, **17**(1): 2096−2030

6 Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Communications of the ACM*, 2020, **63**(11): 139−144

7 Guo Ying-Chun, Feng Fang, Yan Gang, Hao Xiao-Ke. Cross-domain person re-identification on adaptive fusion network. *Acta Automatica Sinica*, 2022, **48**(11): 2744−2756
(郭迎春, 冯放, 阎刚, 郝小可. 基于自适应融合网络的跨域行人重识别方法. 自动化学报, 2022, **48**(11): 2744−2756)

8 Liang Wen-Qi, Wang Guang-Cong, Lai Jian-Huang. Asymmetric cross-domain transfer learning of person re-identification based on the many-to-many generative adversarial network. *Acta Automatica Sinica*, 2022, **48**(1): 103−120
(梁文琦, 王广聪, 赖剑煌. 基于多对多生成对抗网络的非对称跨域迁移行人再识别. 自动化学报, 2022, **48**(1): 103−120)

9 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2015. 91−99

10 Chen Y H, Li W, Sakaridis C, Dai D X, Van Gool L. Domain adaptive faster R-CNN for object detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 3339−3348

11 Saito K, Ushiku Y, Harada T, Saenko K. Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 6949−6958

12 Lin T Y, Goyal P, Girshick R, He K M, Dollar P. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **42**(2): 318−327

13 Shen Z Q, Maheshwari H, Yao W C, Savvides M. SCL: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. arXiv preprint arXiv: 1911.02559, 2019.

14 Zheng Y T, Huang D, Liu S T, Wang Y H. Cross-domain object detection through coarse-to-fine feature adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 13763−13772

15 Xu C D, Zhao X R, Jin X, Wei X S. Exploring categorical regularization for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11721−11730

16 Hsu C C, Tsai Y H, Lin Y Y, Yang M H. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 733−748

17 Chen C Q, Zheng Z B, Ding X H, Huang Y, Dou Q. Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 8866−8875

18 Zhu J Y, Park T, Isola P, Efros A A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2242−2251

19 Deng J H, Li W, Chen Y H, Duan L X. Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 4089−4099

20 Xu M H, Wang H, Ni B B, Tian Q, Zhang W J. Cross-domain detection via graph-induced prototype alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 12352−12361

21 Wu A M, Liu R, Han Y H, Zhu L C, Yang Y. Vector-decomposed disentanglement for domain-invariant object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 9322−9331

22 Chen C Q, Zheng Z B, Huang Y, Ding X H, Yu Y Z. I.3Net: Implicit instance-invariant network for adapting one-stage object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021. 12576−12585

23 Li Wei, Wang Meng. Unsupervised cross-domain object detection based on progressive multi-source transfer. *Acta Automatica Sinica*, 2022, **48**(9): 2337−2351
(李威, 王蒙. 基于渐进多源域迁移的无监督跨域目标检测. 自动化学报, 2022, **48**(9): 2337−2351)

24 Rodriguez A L, Mikolajczyk K. Domain adaptation for object detection via style consistency. In: Proceedings of the 30th British Machine Vision Conference. Cardiff, UK: BMVA Press, 2019.

25 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, et al. SSD: Single shot MultiBox detector. In: Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: Springer, 2016. 21−37

26 Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 779−788

27 Yolov8 [Online], available: https://github.com/ultralytics/yolov8, February 15, 2023

28 Zhang S Z, Tuo H Y, Hu J, Jing Z L. Domain adaptive YOLO for one-stage cross-domain detection. In: Proceedings of the 13th Asian Conference on Machine Learning. PMLR, 2021. 785−797

29 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv preprint arXiv: 1804.02767, 2018.

30 Hnewa M, Radha H. Integrated multiscale domain adaptive YOLO. *EEE Transactions on Image Processing*, 2023, **32**: 1857−1867

31 Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv: 2004.10934, 2020.

32 Vidit V, Salzmann M. Attention-based domain adaptation for single-stage detectors. *Machine Vision and Applications*, 2022, **33**(5): Article No. 65

33 YOLOv5 [Online], available: https://github.com/ultralytics/yolov5, November 28, 2022

34 Li G F, Ji Z F, Qu X D, Zhou R, Cao D P. Cross-domain object detection for autonomous driving: A stepwise domain adaptative YOLO approach. *IEEE Transactions on Intelligent*

Vehicles, 2022, **7**(3): 603−615

35 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7132−7141

36 Wang Q L, Wu B G, Zhu P F, Li P H, Zuo W M, Hu Q H. ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11531−11539

37 Lee H, Kim H E, Nam H. SRM: A style-based recalibration module for convolutional neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019. 1854−1862

38 Wang M Z, Wang W, Li B P, Zhang X, Lan L, Tan H B, et al. InterBN: Channel fusion for adversarial unsupervised domain adaptation. In: Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: ACM, 2021. 3691−3700

39 Ding S Y, Lin L, Wang G R, Chao H Y. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015, **48**(10): 2993−3003

40 Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 4080−4090

41 He K M, Fan H Q, Wu Y X, Xie S N, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 9726−9735

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

43 Sakaridis C, Dai D X, Van Gool L. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 2018, **126**(9): 973−992

44 Yu F, Chen H F, Wang X, Xian W Q, Chen Y Y, Liu F C, et al. Bdd100K: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 2633−2642

45 Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 2013, **32**(11): 1231−1237

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

47 Everingham M, Van Gool L, Williams C K I, Winn J, Zisserman A. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010, **88**(2): 303−338

48 Inoue N, Furuta R, Yamasaki T, Aizawa K. Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 5001−5009

周 洋 西安电子科技大学电子工程学院硕士研究生. 2020 获西南石油大学电子信息工程学士学位. 主要研究方向为计算机视觉和域自适应目标检测.
E-mail: yzhou_6@stu.xidian.edu.cn
(**ZHOU Yang** Master student at the School of Electronic Engineering, Xidian University. He received his bachelor degree in electronic and information engineering from Southwest Petroleum University in 2020. His research interest covers computer vision and domain adaptive detection.)

韩 冰 西安电子科技大学电子工程学院教授. 主要研究方向为智能辅助驾驶系统, 视觉感知与认知, 空间物理与人工智能交叉. 本文通信作者.
E-mail: bhan@xidian.edu.cn
(**HAN Bing** Professor at the School of Electronic Engineering, Xidian University. Her research interest covers intelligent auxiliary drive system, visual perception and cognition, and cross-disciplinary research between space physics and artificial intelligence. Corresponding author of this paper.)

高新波 西安电子科技大学教授. 主要研究方向为机器学习, 图像处理, 计算机视觉, 模式识别和多媒体内容分析. E-mail: xbgao@ieee.org
(**GAO Xin-Bo** Professor at Xidian University. His research interest covers machine learning, image processing, computer vision, pattern recognition, and multimedia content analysis.)

杨 铮 西安电子科技大学电子工程学院博士研究生. 2017 获西安电子科技大学智能科学与技术学士学位. 主要研究方向为深度学习, 目标跟踪和强化学习.
E-mail: zhengy@stu.xidian.edu.cn
(**YANG Zheng** Ph.D. candidate at the School of Electronic Engineering, Xidian University. He received his bachelor degree in intelligent science and technology from Xidian University in 2019. His research interest covers deep learning, object tracking, and reinforcement learning.)

陈玮铭 西安电子科技大学电子工程学院硕士研究生. 2019 获西安电子科技大学机械设计制造及其自动化学士学位. 主要研究方向为计算机视觉, 目标检测和遥感技术.
E-mail: wmchen@stu.xidian.edu.cn
(**CHEN Wei-Ming** Master student at the School of Electronic Engineering, Xidian University. He received his bachelor degree in mechanical design manufacture and automation from Xidian University in 2019. His research interest covers computer vision, object detection, and remote sensing.)