基于渐进多源域迁移的无监督跨域目标检测

李 威 [1, 2] 王 蒙 [1, 2]

摘 要 针对目标检测任务中获取人工标注训练样本的困难, 提出一种在像素级与特征级渐进完成域自适应的无监督跨域目标检测方法. 现有的像素级域自适应方法中, 存在翻译图像风格单一、内容结构不一致的问题. 因此, 将输入图像分解为域不变的内容空间及域特有的属性空间, 综合不同空间表示进行多样性的图像翻译, 同时保留图像的空间语义结构以实现标注信息的迁移. 此外, 对特征级域自适应而言, 为缓解单源域引起的源域偏向问题, 将得到的带有标注的多样性翻译图像作为多源域训练集, 设计基于多领域的对抗判别模块, 从而获取多个领域不变的特征表示. 最后, 采用自训练方案迭代生成目标域训练集伪标签, 以进一步提升模型在目标域上的检测效果. 在 Cityscapes & Foggy Cityscapes 与 VOC07 & Clipart1k 数据集上的实验结果表明, 相比现有的无监督跨域检测算法, 该检测框架具更优越的迁移检测性能.

关键词 迁移学习, 域自适应, 目标检测, 多源域, 自训练

引用格式 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

DOI 10.16383/j.aas.c190532

Unsupervised Cross-domain Object Detection Based on Progressive Multi-source Transfer

LI Wei[1, 2] WANG Meng[1, 2]

Abstract To address the difficulty of collecting manually labeled training samples for object detection tasks, this paper proposes an unsupervised cross-domain object detection method that gradually adapts the model at pixel level and feature level. The existing pixel-level domain adaptive methods generate translated images with a single style and inconsistent content structure. To solve this problem, this paper embeds the input images into domain-invariant content space and domain-specific attribute space, then cooperates different space representations to synthesize diverse translated images that preserve the spatial semantic information to enable label transfer. In addition, for feature-level domain adaptation, to alleviate the source-bias problem caused by single source domain, we treat the generated diverse labeled images as source domain data and design a multi-domain discriminator to get multi-domain-invariant representations. Finally, To further enhance the detection performance on the target domain, we propose a self-training framework to alternatively generate pseudo labels on target training data. The exploratory experiment results from the Cityscapes & Foggy Cityscapes dataset and VOC07 & Clipart1k dataset demonstrate that compared with the current unsupervised cross-domain detection methods, the proposed detection framework achieves better transferability.

Key words Transfer learning, domain adaptation, object detection, multi-source domain, self training

Citation Li Wei, Wang Meng. Unsupervised cross-domain object detection based on progressive multi-source transfer. *Acta Automatica Sinica*, 2022, **48**(9): 2337−2351

目标检测作为一类计算机视觉的基础任务, 能对图像前景对象进行定位及分类, 在智能驾驶、安防监控等领域有着广泛的应用[1−2]. 近年来, 伴随着深度卷积神经网络[3] 的发展, 目标检测在检测精度和时效性上均取得了一系列重大突破. 基于深度学习的目标检测方法, 目前主要分为 2 类: 1) 两阶段检测器, 如区域卷积网络 (Region convolution neural network, R-CNN)[4]、快速区域卷积网络 (Fast R-CNN)[5]、超快速区域卷积网络 (Faster R-CNN)[6] 等, 这类检测器首先通过区域提取网络得到感兴趣的区域, 再进一步对这些区域进行分类和回归; 2) 单阶段检测器, 如一见即得检测器[7]、单发多框检测器 (Single shot multi-box detector, SSD)[8] 等. 这类检测器中, 直接对不同特征层上的预设边框进行分类和回归, 从而提升了检测速度. 虽然这些检测方法均取得了不错的效果, 但在许多

收稿日期 2019-10-25 录用日期 2020-03-11
Manuscript received October 25, 2019; accepted March 11, 2020
国家自然科学基金 (61563025) 和云南省科技计划项目 (2016FB-109) 资助
Supported by National Natural Science Foundation of China (61563025) and Yunnan Science and Technology Department of Science and Technology Project (2016FB109)
本文责任编委 刘青山
Recommended by Associate Editor LIU Qing-Shan
1. 昆明理工大学信息工程与自动化学院 昆明 650500 2. 昆明理工大学云南省人工智能重点实验室 昆明 650500
1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500

实际场景中却不能得到有效应用. 一方面, 训练基于深层网络的检测器需要海量的标注数据, 而从数据的收集到标注, 都是一件耗时费力的事. 此外, 大部分人工数据标注缺乏统一的标准, 会不可避免地引入人为偏差. 另一方面, 现有的目标检测方法一般假设训练数据与测试数据服从独立同分布, 而在实际应用中却难以满足, 从而导致在某数据集上训练好的检测模型难以泛化到其他场景. 例如, 用天气良好时采集的图片训练得到的检测模型, 在有雾的情况下检测性能会急剧下降. 如图 1 所示, 上边为天气良好情况下收集的图片, 下边为有雾天气下的数据, 这 2 个数据集在风格、光照以及颜色等方面存在差异. 针对上述问题, 本文主要研究无监督跨域目标检测算法. 其中, 源域数据集 (如图 1 中上行图片) 有分类标注与边界框标注, 而目标域没有标注信息 (如图 1 中下行图片). 将大量易得的标注数据的知识迁移到其他不易得且缺乏标注的数据域中, 以提升检测器在不同场景下的适应能力, 是本文的主要研究目的.

图 1 Cityscapes[9] (上) 与 Foggy Cityscapes[10] (下) 示例图
Fig. 1 Examples from Cityscapes[9] (up) and Foggy Cityscapes[10] (bottom)

针对目标域标注数据稀缺、领域分布异构等问题, 目前主要有两类方法. 一类是弱监督的目标检测方法[11−12]. 给定只有分类标注的数据集, 通过区域提取网络得到感兴趣的区域, 然后再设计分类器并用分类标注进行训练. 相对于强监督的目标检测方法, 这种方法的检测效果较差. 另外一类, 可概括为无监督域自适应方法[13], 通过源域到目标域的域自适应, 将源域中的标注信息迁移到目标域, 从而提升目标域数据集上的检测精度. 为实现源域与目标域的语义对齐, 采用了最小化源域与目标域之间度量距离的方法, 如相关对齐[14] 和最大均值差异[15] 等. 这种基于度量的方法取得了一定的效果, 但在深度卷积网络中, 由于数据被映射到高维空间, 效果有时反而更差[16]. 尽管无监督域自适应方法在图像分类和分割等任务中均取得了不错的效果, 但在目标检测方面的研究仍然不足. 已有为数不多的研究[17−27], 主要采用像素级对齐[17−18] 或特征级对齐[18−25] 来实现源域知识到目标域的迁移. 其中, 像素级对齐主要采用图像翻译的方法来实现, 如采用循环对抗生成网络 (Cycle generative adversarial network, CycleGAN)[28] 等, 通过生成含有源域数据的内容信息与目标域数据的风格信息的图片, 从而将源域中的标注信息迁移到生成图像. 特征级对齐在特征层加入判别器, 通过构造对抗生成网络 (Generative adversarial networks, GAN)[29] 使判别器无法将源域特征从目标域特征中分辨出来, 进而拉近两个领域之间的特征分布. 例如, Inoue 等[17] 提出一种渐进弱监督跨域目标检测方法, 先采用 CycleGAN[28] 生成含有源域数据空间语义信息和目标域风格特征的图片, 并将源域中的标注信息迁移到生成图像上; 然后使用在源域数据上训练好的检测模型在这些生成图片上进行微调; 最后, 使用在目标域上预测生成的伪标签进一步训练, 并得到在目标域上的检测模型. 类似的, 加噪标签[26] 直接使用在源域数据上训练的检测器在目标域上预测生成伪标注, 然后使用一个分类模块对伪标签进行修正并与源域数据联合训练, 以得到一个更具鲁棒性的检测器. Chen等[19] 在 Faster R-CNN[6] 的基础上, 通过实例级与图像级的域自适应, 实现了检测模型的泛化. 在此基础上, 文献 [20−25]通过不同特征层的对齐, 实现了不同领域之间深层特征与浅层特征的适配. 以上工作主要面向单源域到单目标域的检测迁移问题, 为了进一步有效利用众多不同领域之间的相关知识, 一些研究者将目光转向了更具挑战性的多源域到单目标域的迁移问题. Wang 等[27] 提出了一个基于注意力机制的域自适应检测框架, 实现了从多个源域到单目标域的检测任务. 其困难在于需要收集大量不同的源域数据集. 此外, Kim 等[18] 探索了如何生成多样性的翻译图片来实现多源域适配, 但其图像转换过程尚未利用目标域特有的属性特征, 以使得生成图像与目标域特征分布更加相似.

上述无监督域自适应方法的提出, 证明了基于迁移的目标检测模型的有效性, 但仍存在以下 3 方面问题: 1) 在像素级对齐时, 采用 CycleGAN[28] 等图像翻译方法生成的样本, 多样性不够, 不能保持语义结构的连续性; 或是人为设置源域样本的多样性, 而没有充分利用目标域的属性特征; 2) 特征级对齐方面, 大多只考虑单源域到单目标域的迁移, 没有考虑多源域到单目标域迁移的情景. 特征对齐网络在训练过程中, 其判别性主要取决于有标注信息的源域数据, 迁移性则取决于源域特征与目标域特征之间的相似性. 在单源域自适应方法中, 由于单一风格的源域图像通常只包含部分信息, 因此检测模型的判别性容易偏向于仅有的单一源域表示, 从而影响目标域上的性能; 3) 部分方法仅针对某一特定检测模型, 例如 Chen 等[19] 提出的实例级域自

适应方法在单阶段的检测模型中难以实现. 为尝试解决这些困难, 本文提出了一个渐进对齐的无监督跨域目标检测框架, 主要工作如下: 1) 对图片特征进行分解, 分别得到域不变的结构内容特征与域特有的风格属性特征, 以使得生成样本更好地保持原数据的空间结构信息. 并且, 通过源域与目标域之间两类特征的结合, 能够生成多样性的数据样本, 这些不同风格属性的生成图片丰富了源域样本的多样性; 2) 设计了一个基于对抗网络的多域分类器, 并将生成的具有不同属性特征的样本加入到源域数据集中, 使检测器能在多个源域数据集上训练, 并且目标域特征分布可以由多个与其风格近似的源域数据来拟合, 从而获取多领域不变的特征表示; 3) 采用自训练框架进一步提升目标域上的检测性能. 源域和目标域通过像素级对齐和多源域特征对齐后, 检测模型在目标域上可以预测生成质量较高的伪标签, 从而避免了直接使用源域数据训练的模型预测生成伪标注质量差的问题. 实验表明,采用这种渐进域自适应的训练方式, 显著地提升了检测模型的迁移性能.

1　基于渐进多源域迁移的跨域目标检测方法

1.1　问题描述

在本文研究的无监督跨域目标检测任务中, 源域数据集有分类标注与边界框标注, 而目标域没有标注信息. 定义源域数据集为 $X_S = \{x_S^i | i = 1, \cdots, n_S\}$, 标注集为 $Y_S = \{y_S^{ic},\ y_S^{ib} | y_S^{ic} \in C,\ i = 1, \cdots, n_S\}$, 目标域数据为 $X_T = \{x_T^j | j = 1, \cdots, n_T\}$. 其中 n_S 和 n_T 分别表示源域与目标域的数据大小, y_S^{ic} 和 y_S^{ib} 分别为第 i 张图片的类别标注集合与边框标注集合, C 为源域数据的类别集合. 并且, 目标域数据的类别集合是源域类别集合的子集. 本文研究的目的是利用源域中丰富的数据与标注信息, 通过迁移学习的方法, 将源域中的知识迁移到目标域中, 以提升目标域测试集上的检测性能.

1.2　基本检测模型

考虑到实际应用中检测的时效性要求, 本文采用单阶段检测器 SSD[8] 作为基本检测模型. 在 SSD 模型中, 首先通过基础网络 VGG16[30] 提取特征, 然后加入尺寸不同的特征层, 并分别在 6 个不同尺度的特征层上获得检测边框集合与对应的分类置信度, 再对所得边框进行非极大值抑制, 从而得到最终检测结果. 训练过程中, SSD 的目标损失函数为:

$$L_{det} = \frac{1}{N}(L_{conf}(x, p) + \alpha L_{loc}(x, l, g)) \tag{1}$$

式中, N 为与标注边框相匹配的默认框个数. 若 $N = 0$ 则误差为 0, 定位误差 L_{loc} 为平滑 L_1 损失函数, L_{loc} 为分类置信度损失, α 为平衡定位与分类损失的权重参数, x 为当前预测框的类别匹配信息 ($x_{i,j}^p = \{0, 1\}$ 代表当前第 i 个预测框匹配类别 p 的第 j 个目标框真值), l 为预测边框, g 为真实标注边框.

1.3　像素级域自适应网络

像素级域自适应网络, 主要通过源域与目标域之间的图像翻译来实现. 本文借鉴了基于非纠缠表示的图像翻译[31] 模型中的特征分解思想. 图像翻译网络框架如图 2 (a) 所示: 其组成有内容编码器 $\{E_S^c, E_T^c\}$, 及其对应的内容判别器 $\{D_{adv}^c\}$、属性编码器 $\{E_S^a, E_T^a\}$、生成器 $\{G_S, G_T\}$ 和判别器 $\{D_S, D_T\}$. 以输入源域数据集 $X_S = \{x_S^i | i = 1, \cdots, n_S\}$ 为例, 内容编码器 E_S^c 将输入数据映射到一个共享的域不变内容空间, 属性编码器 E_S^a 将其映射到一个域特有的属性空间 (Domain-specific attribute space, DSAS), 生成器 G_S 将 $E_S^a(X_S)$ 与 $E_T^c(X_T)$ 作为输入得到从目标域到源域空间的翻译图片, 而判别器 D_S 则用于判断是生成图片还是原图片. 此外, D_{adv}^c 用于判断内容空间是来自于源域还是目标域, 从而生成源域与目标域共享的域不变内容特征.

1.3.1　特征表示分解

在使用深度卷积神经网络进行特征提取的过程中, 将输入数据特征分解为两个部分: 域不变内容空间和域特有属性空间. 其中, 域特有属性空间主要用于模拟在给定相同语义内容情况下领域特有的特征, 诸如图片的纹理、风格等, 而域不变内容空间则用于提取不同领域之间的共同信息, 主要包含共有的空间语义信息. 给定源域数据集 X_S, 内容编码器 E_S^c 得到跨域共享的特征表示 $z_S^c = E_S^c(X_S)$, 而属性编码器得到私有的特征表示 $z_S^a = E_S^a(X_S)$. 同理, 给定目标域数据集 X_T, 得到 $z_T^c = E_T^c(X_T)$ 和 $z_T^a = E_T^a(X_T)$. 为获得两个图像域之间的共享特征, z_S^c 与 z_T^c 将同时输入一个判别器 D_{adv}^c, 通过对抗生成的训练方式, 使得来自两个图像域的内容特征分布近似, 损失函数为:

$$\begin{aligned} L_{adv}^c(E_S^c, E_T^c, D_{adv}^c) = {} & \mathrm{E}_{x_S \sim p(X_S)} \left[\frac{1}{2} \ln D_{adv}^c(z_S^c) + \right. \\ & \left. \frac{1}{2} \ln(1 - D_{adv}^c(z_S^c)) \right] + \mathrm{E}_{x_T \sim p(X_T)} \cdot \\ & \left[\frac{1}{2} \ln D_{adv}^c(z_T^c) + \frac{1}{2} \ln(1 - D_{adv}^c(z_T^c)) \right] \end{aligned} \tag{2}$$

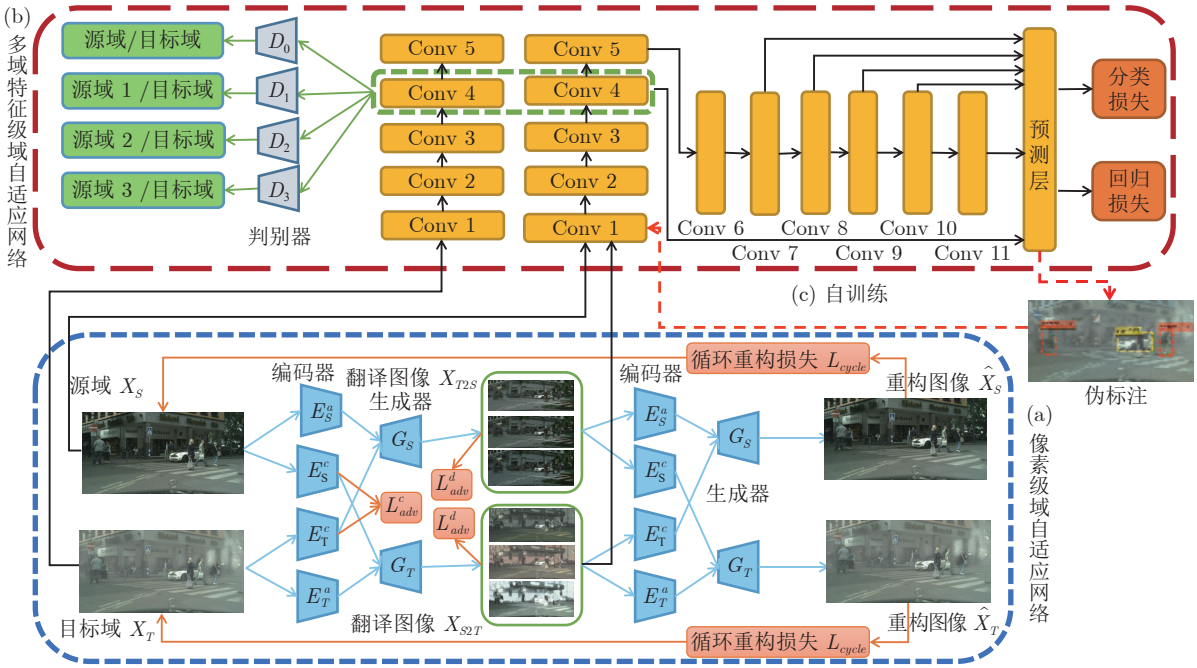


图 2 无监督跨域目标检测方法结构图

Fig. 2 Diagram for unsupervised cross-domain object detection

在特征分解过程中, $z^a \in \mathbf{R}^8$. 在测试过程中对领域特有的属性特征表示 z^a 进行随机采样, 令 z^a 近似于高斯分布, 如图 3 所示. 主要通过 Kullback-Leibler (KL) 散度来实现:

$$L_{KL} = \mathrm{E}[D_{KL}((z^a)||\mathrm{N}(0,1))] \tag{3}$$

式中, $D_{KL}(p||q) = -\int p(z)\ln(p(z)/q(z))\mathrm{d}z$.

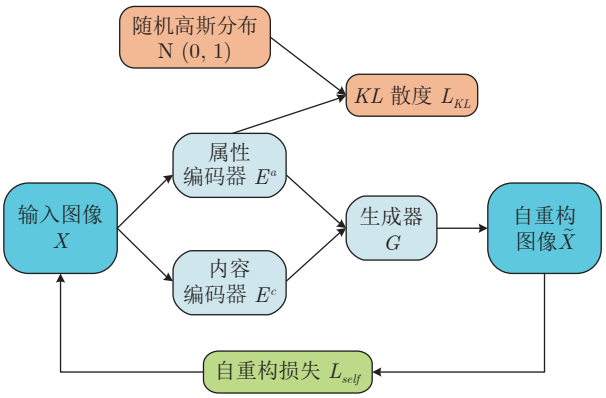


图 3 损失函数

Fig. 3 Loss function

特征分解网络结构及其参数设置如图 4 所示. 内容编码器 E^c 由 3 个卷积层和 4 残差层组成, 并使用了实例归一化[32]. 为了将源域与目标域映射到一个共享的空间, 最后一个卷积层将共享参数. 属性编码器 E^a 由 6 个卷积层组成, 内容判别器网络 D^c_{adv} 由 4 个卷积网络组成.

1.3.2 多样性图像翻译

对于生成器 G_S, 将 z^a_S 和 z^c_T 作为输入, 以生成 $T \to S$ 的翻译图片 $X_{T2S} = G_S(z^c_T, z^a_S)$, 并使用判别器 D_S 来判断图片是否为翻译图片. 同理, 通过 G_T 得到 $S \to T$ 的翻译图片 $X_{S2T} = G_T(z^c_S, z^a_T)$, 并使用判别器 D_T 来分辨图片是否为翻译图片. 这个过程中, 损失函数 L^d_{adv} 如下:

$$\begin{aligned} L^d_{adv}(G_S, G_T, E^c_S, E^a_S, E^c_T, E^a_T) = {} & \mathrm{E}_{\substack{x_S\sim p(X_S)\\ x_T\sim p(X_T)}} \cdot \\ & \left[\frac{1}{2}\ln D_S(x_S) + \frac{1}{2}\ln(1 - D_S(x_{T2S}))\right] + \\ & \mathrm{E}_{\substack{x_S\sim p(X_S)\\ x_T\sim p(X_T)}} \left[\frac{1}{2}\ln D_T(x_T) + \frac{1}{2}\ln(1 - D_T(x_{S2T}))\right] \end{aligned} \tag{4}$$

此外, 为了保留图像的结构内容信息, 将翻译图像映射回原图片空间时, 引入循环一致性约束, 并使用循环重构损失引导模型在 2 个图像域相互转换过程中保留图像固有特征. 具体而言, 将翻译图片 X_{T2S} 分解为 $E^c_S(X_{T2S})$ 和 $E^a_S(X_{T2S})$, X_{S2T} 分解为 $E^c_T(X_{S2T})$ 和 $E^a_T(X_{S2T})$, 并分别作为 G_S 和 G_T 的输入, 以将翻译图片映射回原图片空间, 并分别生成图片 $\hat{X}_S = G_s(z^c_{S2T}, z^a_{T2S})$ 和 $\hat{X}_T = G_T(z^c_{T2S}, z^a_{S2T})$. 使用 L_1 范式作为重构误差为:

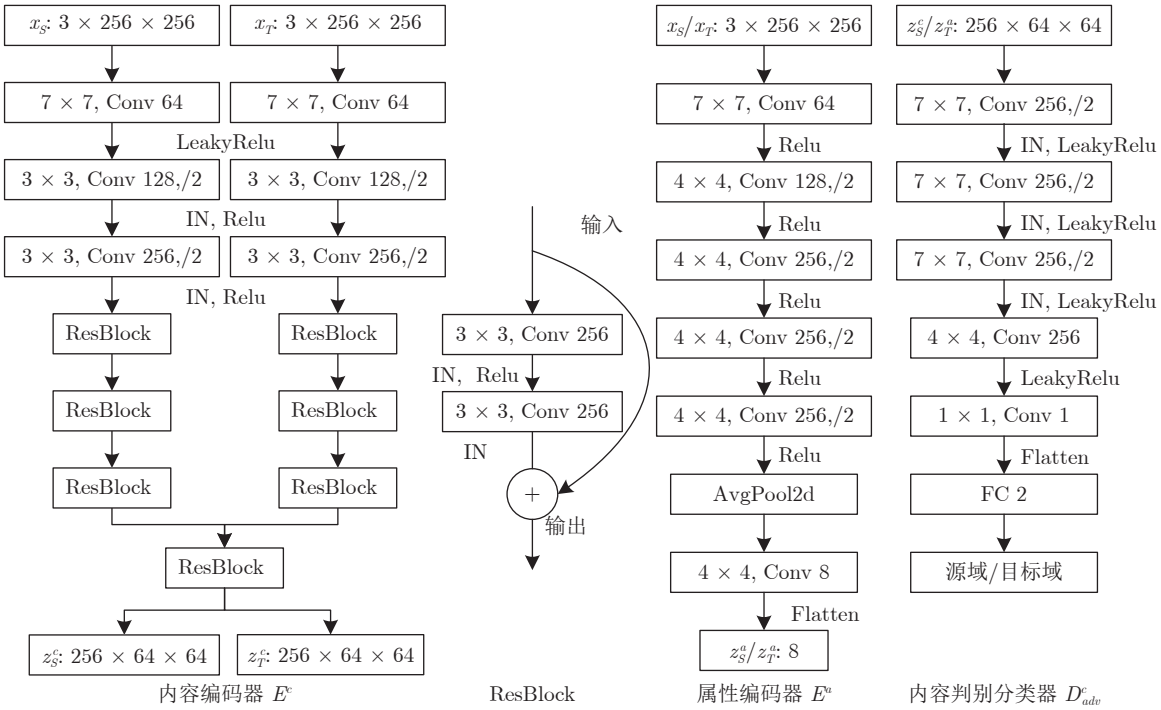


图 4　分解表示所采用模块网络结构

Fig. 4　Modular network structures used in the disentangled representation framework

$$L_{cycle}(G_S, G_T, E_S^c, E_S^a, E_T^c, E_T^a) = \mathrm{E}_{x_S\sim p(X_S)}\left[\|G_S(z_{S2T}^c, Z_{T2S}^a) - x_S\|_1\right] + \mathrm{E}_{x_T\sim p(X_T)}\left[\|G_T(z_{T2S}^c, Z_{S2T}^a) - x_T\|_1\right] \tag{5}$$

在此基础上, 进一步引入自重构误差. 在源域与目标域得到内容特征$\{z_S^c,\ z_T^c\}$与属性特征$\{z_S^a,\ z_T^a\}$后, 再分别将其映射回原来的图像空间, 得到$G_S(z_S^c,\ z_S^a)$和$G_T(z_T^c,\ z_T^a)$, 用于保持图像的一致性, 如图 3 所示, 并将生成的图像记为$\widetilde{X}$. 若使用L_1误差, 则自重构误差L_{self}为:

$$L_{self}(G_S, G_T, E_S^c, E_S^a, E_T^c, E_T^a) = \mathrm{E}_{x_S\sim p(X_S)}[\|G_S(z_S^c, z_S^a) - x_S\|_1] + \mathrm{E}_{x_T\sim p(X_T)}[\|G_T(z_T^c, z_T^a) - x_T\|_1] \tag{6}$$

在多样性图片翻译过程中, 生成器$\{G_S,\ G_T\}$与判别器$\{D_S,\ D_T\}$的网络结构如图 5 所示, 其使用了实例归一化以增强图像风格迁移效果. 整个图像翻译网络框架如图 2(a) 所示, 其训练过程为:

$$\min\ \lambda_{adv}^c L_{adv}^c + \lambda_{adv}^d L_{adv}^d + \lambda_{cycle}L_{cycle} + \lambda_{self}L_{self} + \lambda_{KL}L_{KL} \tag{7}$$

式中, λ_{adv}^c、λ_{adv}^d、λ_{cycle}、λ_{self}和λ_{KL}为超参数, 用于平衡各项目标函数.

在训练过程中, 可以将源域中的图像内容与目标域中的任意属性特征相结合, 生成从源域空间映射到目标域空间的各种不同风格的多样性图片$X_{S2T} = \{x_{S2T}^{m,i}|m = 1,\ 2, \cdots, M;\ i = 1, 2, \cdots, n_S\}$, M表示生成不同的风格数, n_S表示每一个风格属性的数据集大小, 其与源域数据集大小相同. 由于这些生成图片保持了源域图片的空间语义结构信息, 因此可以将源域数据中的标注信息迁移到翻译图片X_{S2T}中. 相应地, 其标注为$Y_{S2T} = \{Y_{S2T}^{m,i}|m = 1, 2, \cdots, M;\ i = 1, 2, \cdots, n_S\}$.

1.4　多域特征级域自适应网络

特征级域自适应的主要目的是使得源域与目标域在特征表示分布上尽可能相似, 典型的方法是通过对抗生成网络来实现. 文献 [33] 将源域特征与目标域特征作为判别器D的输入, 通过在判别器前面加入梯度反向层, 使得判别器无法分辨出特征层来自哪一个样本域, 进而得到域不变的特征表示. 文献 [18, 34] 指出, 在单源域到单目标域的迁移任务中, 容易得到次优解. 由于风格单一的源域图像只包含部分信息, 因此得到的特征表示具有偏向性. 而使用多个风格不同的源域数据, 可以得到不同方面的特征信息, 从而使得多域不变的特征表示具有更强的泛化性能.

通过前述的图像翻译, 得到多样性数据集$X_{S2T} = \{X_{S2T}^{m,i}|m = 1, 2, \cdots, M;\ i = 1, 2, \cdots, n_S\}$, 及其标注

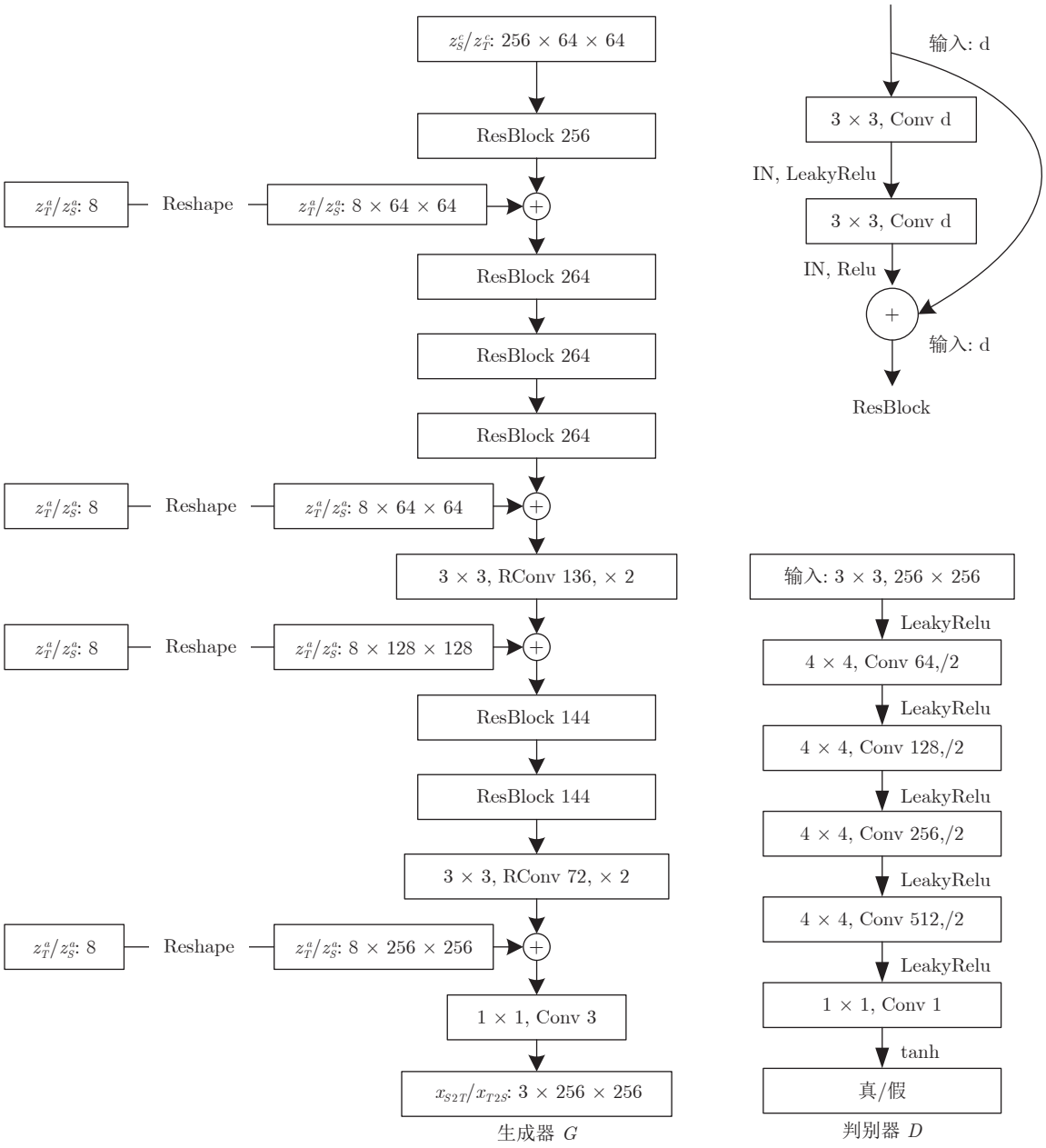


图 5 图像翻译中采用的生成器与判别器网络结构

Fig. 5 Network structures of the generator and the discriminator used in image-to-image translation

$Y_{S2T} = \{Y_{S2T}^{m,i} | m = 1, 2, \cdots, M;\ i = 1, 2, \cdots, n_S\}$, 并将其加入到源域数据集, 则源域数据集共有 $M+1$ 个, 目标域共有 1 个. 以 SSD 作为基本检测器, 将 X_S、X_{S2T} 和 X_T 输入 SSD 的基本特征提取网络, 分别得到源域与目标域的特征表示 $F_S(X_S)$、$F_S(X_{S2T})$ 和 $F_T(X_T)$. 每一个源域特征, 将分别与目标域特征作为二分类领域分类器 $\{D_m | m = 0, 1, 2, 3\}$ 的输入, 进而得到多域不变特征表示, 如图 6 所示. 其中, 每一个源域特征表示标记为 1, 目标域特征表示标记为 0, 则多域对抗判别器损失函数为:

$$\begin{aligned} L_d^{multi}(X_S, X_{S2T}, X_T) = \\ & \mathrm{E}_{x_s \sim X_S} \ln(1 - D_0(F_S(x_S))) + \\ & \mathrm{E}_{x_T \sim X_T} \ln D_0(F_T(x_T)) + \\ & \sum_{m=1}^{M} (\mathrm{E}_{x_{S2T} \sim X_{S2T}} \ln(1 - D_m(F_S(x_{S2T}^m))) + \\ & \mathrm{E}_{x_T \sim X_T} \ln D_m(F_T(x_T))) \end{aligned} \tag{8}$$

将目标域特征作为生成特征, 则对抗损失函数为:

$$L_g^{multi}(X_T)=\sum_{m=0}^{M}\mathrm{E}_{x_T\sim X_T}\ln(1-D_m(F_T(x_T))) \quad (9)$$

此时, 多源域的检测模型目标损失函数为:

$$L_{det}^{multi}(F,Y)=L_{det}(F(x_S),y_S)+\sum_{m=1}^{M}L_{det}(F(x_{S2T}^m),y_{S2T}^m) \quad (10)$$

联合训练多源域分类器与检测模型, 训练过程如下:

$$\max_D\min_F L_{det}^{multi}(F)+\lambda\cdot L_g^{multi}(D(F)) \quad (11)$$

式中, 超参数 λ 用于控制对抗损失的重要性.

在训练过程中, 判别器 $\{D_m|m=0,1,2,3\}$ 的网络结构均由三个卷积层与三个全连接层组成, 并使用了批归一化[35]. 三个卷积层通道数分别为 512、256 和 128, 步长均为 2. 三个全连接层维度分别为 512、256 和 1, 均使用 LeakyRelu 激活函数. 在训练过程中, 将 SSD 的 Conv4_3_relu 特征层作为域分类器 D 的输入, 此时卷积特征层为 $512\times38\times38$, 经过 3 个卷积层后大小变为 $128\times10\times10$, 之后再将特征层转变为一维向量作为全连接层的输入.

1.5　自训练

自训练是半监督学习的一种常用方法, 旨在使用预训练模型在没有标注的图片上自动生成伪标注, 并使用伪标注进行全监督训练. 在无监督跨域检测任务中, 源域数据与目标源数据分布不一致, 在源域数据上训练好的模型很难泛化到目标域, 使得在目标域训练集上的预测结果存在大量漏检与误检. 而使用这些带有 “噪音”的伪标签进行迭代自训练时, 会进一步强化这些错误的信息, 并导致更多错误标签的生成. 为了有效地解决这个问题, 本文采取渐进自训练方法, 使用像素级对齐和多源域特征对齐后的检测模型在目标域训练集上进行预测, 从而提升伪标签的质量. 具体而言, 设数据集的类别集合为 C, 则在目标域上生成的伪标签为 $\tilde{Y}_T=\{\tilde{y}_T^{jc},\tilde{y}_T^{jb}|\tilde{y}_T^{jc}\in C,\ j=1,2,\cdots,n_T\}$. 其中 $\tilde{y}_T^{jc}$ 与 $\tilde{y}_T^{jb}$ 分别为第 j 张图片的分类标注集合与边框标注集

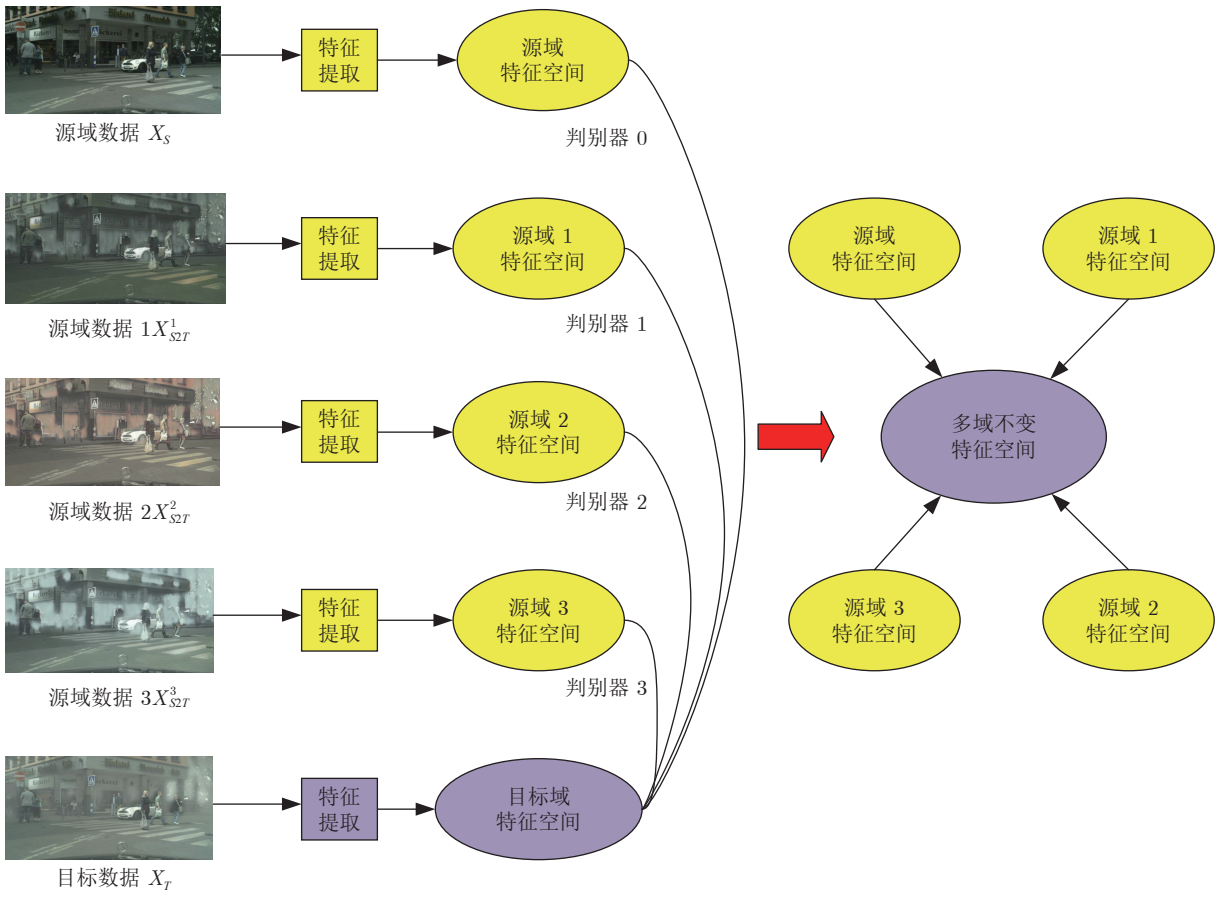


图 6　多域不变特征表示

Fig. 6　Multi-domain-invariant representation

合, n_T 为目标域数据大小. 使用训练好的检测模型对目标域数据进行预测, 设定阈值 θ, 当预测边框的分类置信得分大于阈值 θ 时, 则将对应的边框与类别加入到伪标签中, 并在目标域训练集上的得到最终的伪标签 $\tilde{Y}_T$. 使用目标域训练集及其生成的伪标签进行训练, 过程如下:

$$\min_{w} L_{st}(w, \tilde{Y}_T) = L_{det}(x_T, \tilde{y}_T; w)) \tag{12}$$

式中, w 为检测模型训练参数. 以上自训练过程可以多次迭代进行, 以渐进提升伪标签的质量.

1.6 提出方法整体框架

根据上述各模块描述, 提出方法整体框架如图 2 所示. 图 2(a) 为像素级域自适应网络框架, 其通过基于特征分解的图像翻译, 将源域图像 X_S 转换为 X_{S2T}, 并将源域的标注信息迁移到生成的图片中. 图 2(b) 为多域特征自适应网络框架. 将图 2(a) 中生成的翻译图像 X_{S2T} 加入到源域中, 实现多源域特征对齐的对抗训练. 图 2(c) 为自训练操作, 用图 2(b) 中训练好的模型对目标域数据进行预测生成伪标签, 并进一步做微调训练, 得到最终的检测模型.

2 实验结果

2.1 实验数据与评价指标

为了证明提出检测模型的有效性, 分别在 2 组迁移集上进行实验, 包括 Cityscapes[9] → Foggy Cityscapes[10] 和 VOC07[36] → Clipart1k[17], 并使用检测平均精度 (mAP) 作为评价指标. 两组迁移集具体情况如下:

1) 移集 1: Cityscapes → Foggy Cityscapes. Cityscapes 作为源域数据集, Foggy Cityscapes 作为目标域数据集. 其中, Cityscapes 共有 2 975 张训练图片, Foggy Cityscapes 是在 Cityscapes 数据集中加入合成雾制作而成, 其训练数据大小为 2 975, 有 500 张测试图片. 源域与目标域数据均有 8 个检测类别, 图片分辨率均为 1 024 × 2 048, 在训练过程中, 将图像尺寸设置为 300 × 300.

2) 迁移集 2: VOC07 → Clipart1k. VOC07 作为源域数据集, Clipart1k 作为目标域数据集. 其中, VOC07 中的训练集和验证集均作为源域训练数据集, 共有 5 011 张图片; Clipart1k 共 1 000 张图片, 训练集与测试集分别为 500 张. 源域与目标域数据均有 20 个检测类别, 在训练过程中, 将图像尺寸设置为 300 × 300.

2.2 实验设置

本文提出了一种渐进对齐的无监督跨域目标检测方法. 其训练主要分为基本检测模型、像素级特征对齐、多源域特征对齐和自训练 4 个步骤:

1) 基本检测模型: 使用源域数据, 参照 SSD[8] 的参数设置, 得到一个基本的检测模型.

2) 在像素级对齐网络中, 实现多样性的图像翻译. 输入图像大小为 256 × 256, 训练批次大小为 1, 所有网络模型的权重使用均值为 0、方差为 0.02 的高斯分布进行随机初始化. 分别设置参数 λ^c_{adv} =1, λ^d_{adv} =1, λ_{cycle} =10, λ_{self} =10, λ_{KL} =0.01. 采用 Adam[37] 优化算法, 一阶矩估计的指数衰减率 β_1 设定为 0.5, 二阶矩估计的指数衰减率 β_2 设定为 0.999. 共训练 180 个周期, 内容判别器 D^c_{adv} 初始学习率为 4×10^{-5}, 其他网络结构的初始学习率为 1.0×10^{-4}, 在训练 90 个周期后, 学习率均减小为原来的 0.1 倍. 然后, 将基本检测模型作为预训练模型, 并将生成的多样性图像作为输入, 参照 SSD 的训练参数, 得到一个检测模型.

3) 在多源域特征对齐网络中, 使用 SSD 作为基本的检测器, 由于显存的限制训练批次大小设置为 6. 在训练过程中, 检测网络使用像素级对齐网络中训练好的模型作为预训练模型, 初始学习率为 0.001, 训练周期为 30 000, 每到 10 000 次迭代周期时学习率变为原来的 0.1 倍, 其他参数设置均与 SSD 中相同. 领域分类器加在 VGG16 网络中 Conv4_3_relu 层, 平衡参数 $\lambda = 1$, 其网络权重使用均值为 0、方差为 0.02 的高斯分布进行随机初始化. 领域分类器的学习率为 1.0×10^{-4}, 采用 Adam[37] 优化算法, 一阶矩估计的指数衰减率 β_1 设定为 0.9, 二阶矩估计的指数衰减率 β_2 设定为 0.99.

4) 在自训练过程中, 使用多源特征对齐网络训练好的模型作为初始模型, 学习率为 1.0×10^{-5}, 训练批次样本数为 16, 共训练10 000 批次, 其他设置与 SSD 相同. 自训练过程共迭代 3 次, 每一轮迭代过程都以上一轮的最终模型预测生成伪标注, 并作为预训练模型进行微调训练. 以上所有实验均在 Ubuntu18.04 操作系统上完成, 并使用 pytorch1.0、python3.6 和显卡 GeForce RTX 2070 进行模型训练.

2.3 实验结果分析

通过上述的实验方案, 分别得到了迁移集 1 和迁移集 2 中对目标域的检测结果, 如表 1 所示. 其中, 基线方法为只使用源域数据训练得到的检测模型. 在全监督方法中, 将基线方法得到的模型作为预训练模型, 再使用带有标注信息的目标域训练数据进行训练, 该方法在目标域测试集上得到的结果可作为最终检测性能的上限. 由表 1 可以看出, 本文方法的每一步操作均提升了性能. 具体而言, 在

表 1　不同目标检测方法 mAP 性能对比 (%)
Table 1　Comparison of different detection methods on performance of mAP (%)

方法	迁移集 1	迁移集 2
基线方法 (SSD300)	17.4	23.2
DAAN[33]	25.9	28.4
CycleGAN[17]	27.9	30.0
DT[17]	23.3	25.6
像素级对齐	29.5	31.8
多源特征对齐	32.7	36.2
自训练	20.1	23.9
多源特征对齐 + 自训练	32.9	38.6
全监督	33.0	48.4

Cityscapes → Foggy Cityscapes 的迁移实验中, 通过生成多样性 ($M = 3$) 翻译图像, 实现了像素级对齐, 将检测结果提升 12.1%. 进一步地实施多源域特征对齐, 检测结果由初始的 17.4% 提升到 32.7%; 单独采用自训练方法, 检测结果提升了 2.7%. 最后, 通过综合多源特征对齐与自训练方法, 检测结果提升到了 32.9%, 只比全监督检测结果低 0.1%. 在 VOC07 → Clipart1k 实验中, 通过结构化多样性图像翻译, 生成 $M = 3$ 种不同风格的图片. 在像素级对齐实验中, 相比基线模型检测平均精度提升了 8.6%; 在多源特征对齐试验中, 检测结果由 23.2% 提升到 36.2%; 通过自训练, 检测结果提升了 0.7%; 综合本文所提出的所有模块, 最终检测结果提升了 15.4%. 同时, 本文也与其他方法进行了对比, 主要包括域自适应对抗网络 (Domain-adaption adversarial network, DAAN)[33]、CycleGAN 以及域迁移 (Domain transform, DT)[17]. 其中 DAAN 主要通过对抗生成网络实现了源域与目标域特征级对齐, 在训练时, 将领域分类器加在 SSD 网络中的 Conv4_3_relu 层. CycleGAN得到从源域到目标域上的翻译图片, 将源域中的标注信息迁移到翻译图片, 并使用在源域数据上训练的检测模型在翻译图片上做微调训练. DT 中的方法与本文的更为接近, 其在 CycleGAN 的基础上, 进一步的使用训练好的模型在目标域数据上生成伪标签并进行微调, 以得到最终的检测模型. 不同于本文设定阈值得到伪标注, DT 将在目标域训练集上分类得分最高的预测边框作为伪标签. 由表 1 可知, 本文方法优于以上各种方法. 以 Cityscapes → Foggy Cityscapes 的迁移实验为例, 相比 DAAN, 本文最终结果提升了 7%. CycleGAN 与本文中的像素级自适应的思想类似. 不同的是, 本文基于特征分解的图像翻译, 其生成的样本具有多样性, 从而使得翻译图片包含了目标域中更多不同方面的信息. 由表 1 可以看出, 相较于 CycleGAN[17] 方法, 本文提出的像素级自适应网络的检测性能提升了 1.6% (27.9% 比 29.5%). DT 在进一步使用自训练方法后, 性能反而降低了 4.6% (27.9% 比 23.3%), 其原因在于 DT 通过取首位排名分类得分对应的预测边框作为图像的伪标注, 存在大量分类得分较低的错误标注, 并遗漏了许多可能为正样本的标注. 而本文中采用的基于阈值选取伪标注的方法, 则可以避免大量的错误标注与遗漏标注, 从而更好地提升检测性能.

此外, 由图 7 和图 8 可以看出, 本文方法在大多数类别上取得了最好的检测效果, 实现了类别级的检测迁移性能提升. 图 9 和图 10 则分别给出了分类置信度阈值为 0.5 时迁移集 1 和迁移集 2 中目标域上不同方法的检测结果. 可以看出, 其他方法中均存在不同程度的错检和漏检情况, 而本文方法得到的检测结果明显更好.

2.3.1　基于 Faster R-CNN 检测框架的实现与比较

本文的实验主要基于 SSD 检测框架完成, 为了证明本文方法具有更广的适用性, 以 Faster R-CNN 为基本检测模型, 并在 Cityscapes → Foggy Cityscapes 迁移集上进行验证. 具体而言, 在 Faster R-CNN 检测器中, 以 VGG16 作为基本的特征提取网络, 输入图像较短边大小设置为 600. 在训练基本检测模型性过程中, 依照 Faster R-CNN[6] 中的参数设置. 在像素级域自适应网络中, 使用基本检测模型为预训练模型, 学习率设置为 0.001, 迭代训练 10 个周期. 在特征级域适应方法中, 学习率设置为 0.001, 迭代训练 10 个周期. 其他参数设置均与 Faster R-CNN[6] 中相同. 平衡参数 $\lambda = 1$, 领域分类器加在 VGG16 网路中 Conv5_3_relu层, 学习率为 0.0001, 采用 Adam 优化算法, 一阶矩估计的指数衰减率 β_1 设定为 0.9, 二阶矩估计的指数衰减率 β_2 设定为 0.99. 在自训练过程中, 只进行一次迭代训练. 取阈值 $\theta = 0.5$, 使用多源特征对齐网络训练好的模型作为预训练模型, 学习率为 0.0001, 单批次样本数为 1, 共迭代训练 20 000 次, 其他设置与 Faster R-CNN[6] 相同. 此外, 不同于以上采用分步渐进训练的方法, 同时设计以 VGG16 作为预训练模型, 将像素级与特征级域自适应网络进行联合训练. 其中, 初始学习率为 0.01, 训练次数为 60 000, 在迭代次数为 40 000 时, 学习率变为原来的 0.1 倍. 其他其他设置与 Faster R-CNN[6] 相同. 整个训练过程中训练批次大小设置为 1. 实验结果如表 2 所示, 本文在对像素级对齐与特征级对齐网络逐步训

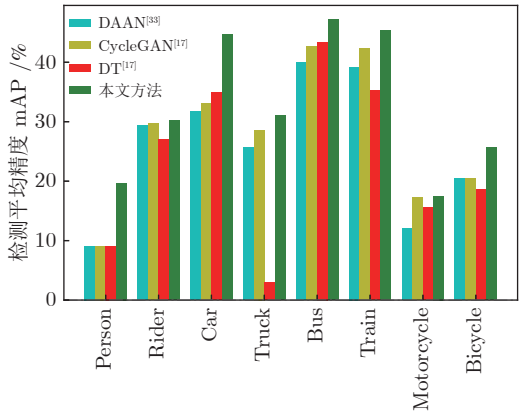


图 7 在 Cityscapes → Foggy Cityscapes 实验中不同方法在所有 8 个类别上的 mAP 表现

Fig. 7 Percategory mAP performance of different approaches over all the 8 categories on the experiment Cityscapes → Foggy Cityscapes

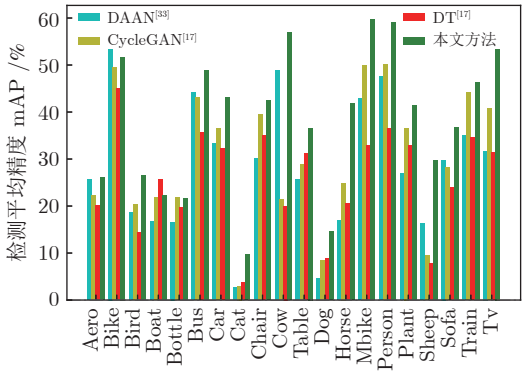


图 8 在 VOC07 → Clipart1k 实验中不同方法在所有 20 个类别上的 mAP 表现

Fig. 8 Percategory mAP performance of different approaches over all the 20 categories on the experiment VOC07 → Clipart1k

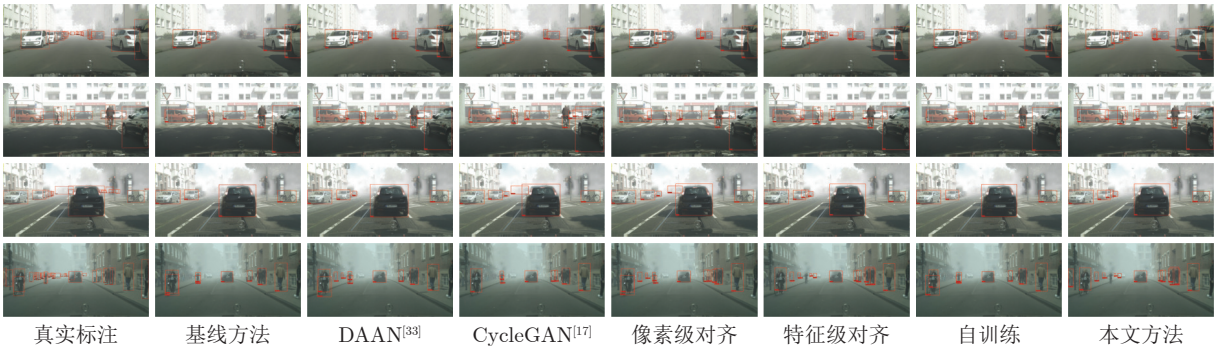


图 9 多种方法在 Cityscapes → Foggy Cityscapes 实验中检测结果对比

Fig. 9 Comparison of different detection methods in the Cityscapes → Foggy Cityscapes experiment

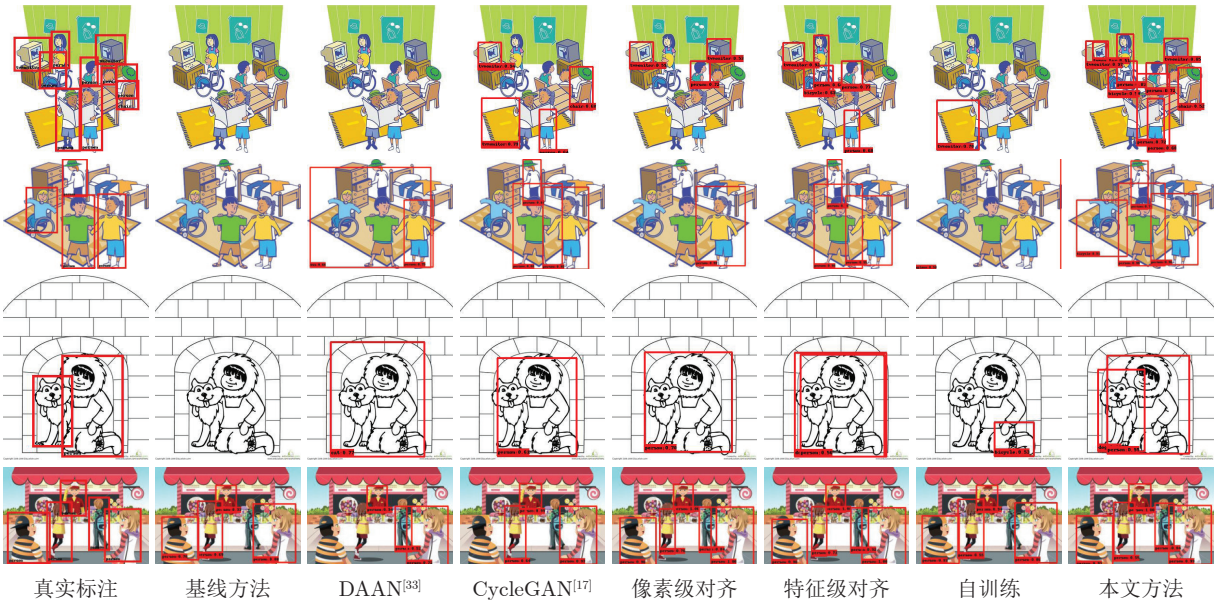


图 10 不同方法在 VOC07 → Clipart1k 实验中检测结果对比

Fig. 10 Comparison of different detection methods in the VOC07 → Clipart1k experiment

表 2　在 Cityscapes → Foggy Cityscapes 实验中基于 Faster R-CNN 的不同跨域检测方法性能对比 (%)
Table 2　Comparison of different cross-domain detection methods based on Faster R-CNN detector in Cityscapes → Foggy Cityscapes (%)

方法	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	mAP
基线方法 (Faster R-CNN)	24.7	31.7	33.0	11.5	24.4	9.5	15.9	28.9	22.5
域自适应 Faster R-CNN[19]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
DT[17]	25.4	39.3	42.4	24.9	40.4	23.1	25.9	30.4	31.5
选择性跨域对齐[21]	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
多对抗超快速区域卷积网络[23]	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
强弱分布对齐[20]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
域自适应表示学习[18]	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
一致性教师客体关系[22]	30.6	41.4	44.0	21.9	38.6	40.6	28.3	35.6	35.1
多层域自适应[24]	33.2	44.2	44.8	28.2	41.8	28.7	30.5	36.5	36.0
加噪标签[26]	35.1	42.1	49.2	30.1	45.3	26.9	26.8	36.0	36.5
像素级 + 多域特征对齐 (联合训练)	32.3	42.5	49.1	26.5	44.6	32.8	31.5	35.6	36.9
像素级对齐	33.1	43.0	49.4	28.0	43.3	35.2	**35.4**	36.3	38.0
多域特征对齐	33.0	43.8	48.5	26.7	45.2	**44.6**	30.8	**37.0**	38.7
像素级 + 多域特征对齐 + 自训练	**35.7**	**45.6**	**51.7**	**31.0**	**47.0**	41.4	30.3	36.7	**39.9**
全监督	35.4	47.1	52.4	29.6	42.7	46.3	33.8	38.4	40.7

练时, 在目标域上的检测性能为 38.7%. 然后再进一步自训练, 平均准确率达到了 39.9%, 这比原始的 Faster R-CNN 模型提升了 17.4%, 而相较于全监督方法, 只差 0.8%. 同时, 相比对像素级对齐与特征级对齐网络进行联合训练, 分步渐进训练的方法取得了更好的效果, 检测平均精度要高出 1.8% (38.7% 比 36.9%).

为了验证本文方法的有效性, 本文与当前最新进的 9 种无监督跨域目标检测方法进行了对比. 其中, 域自适应 Faster R-CNN[19] 采用图像级与实例级特征对齐的方法, 实现源域与目标域的对齐; DT[17] 先使用 CycleGAN 得到从源域到目标域的翻译图像, 以实现像素级的域自适应. 然后再使用自训练方式, 以进一步减小源域与目标域之间在高层语义特征的域差异; 选择性跨域对齐[21] 为了缓解全局特征对齐的局限性, 通过聚类的方式得到不同的提取区域, 以实现更细节的局部对齐; 多对抗超快速区域卷积网络[23]、强弱分布对齐[20] 和多层域自适应[24] 通过对不同特征层的对齐, 以实现源域与目标域浅层特征与深层特征的适配. 域自适应表示学习[18] 使用 CycelGAN 生成多样性的图像, 然后再实现了多领域不变的特征表示; 一致性教师客体关系[22] 则使用了一致性教师训练的方法实现的方法实现高效的跨域检测; 加噪标签[26] 则采用在目标域上生成伪标注并进一步对伪标注进行修正的方式来提升在目标域上的检测性能. 由表 2 可以看出, 本文方法取得了更好的跨域检测性能, 即便在不使用自训练方法的情况下, 在特征级对齐网络中得到的检测结果也比当前最好的方法加噪标签高出 2.2%. 具体来说, 域自适应 Faster R-CNN、选择性跨域对齐、多对抗超快速区域卷积网络、强弱分布对齐、多层域自适应和 MTOR 主要使用了不同策略的特征级对齐方法, 相比于本文采用的多域对抗的方法, 本文得到了更好的检测性能. DT 和 DMRL 均使用 CycleGAN 生成从源域到目标域的翻译图像, 即便在只使用像素级对齐网络的情况下, 本文的检测结果也更优. 加噪标签则主要是使用自训练的策略, 与本文得到的结果最为接近. 加噪标签通过在源域上训练好的模型在目标域上预测生成带有噪声的伪标注, 然后使用分类网络对这些伪标注进行修正, 并进一步用于自训练. 这种自训练策略, 值得本文借鉴. 最后, 通过对每一个类别检测结果的对比, 可以看到本文提出的方法不仅实现了平均检测精度的最优, 而且也实现了类级别的跨域检测性能提升, 域检测性能提升和性能提升.

2.3.2　不同数据集上的性能比较

在 Cityscapes → Foggy Cityscapes 实验中, 源域与目标域训练数据数量相同, 且 Foggy Cityscapes 主要由 Cityscapes 加入雾生成, 二者之间有着完全相同的空间结构信息. 此时, 源域与目标域数据差异相对较小. 在 VOC07 → Clipart1k 实验中, 源域有 5 011 张图片, 目标域只有 500 张训练图片, 而且源域与目标域空间信息不尽相同. 因此, 这组数据中源域与目标域差异相对较大. 图 11 分析

了本文提出方法的每一成分对结果的影响. 可以看出, 在迁移集 1 上的迁移效果较好, 这也与迁移集 1 中源域与目标域差异更小的看法相符. 其中, 像素级对齐在迁移集 1 上效果提升更明显, 而加入多源域特征对齐后, 在迁移集 2 上有更大的提升 (3.2% 比 4.4%). 在单独使用自训练的情况下, 在迁移集 1 上的检测提升性能更好 (2.7% 比 0.7%), 而在进一步采取像素级对齐与特征级对齐后, 自训练方法在迁移集 2 上效果更明显 (0.2% 比 2.4%). 这是因为, 相对而言, 迁移集 1 中的域差异比迁移集 2 中的更小, 则在只使用源域数据训练得到的检测模型在迁移集 1 中可以生成更好的伪标注. 在采取像素级对齐与特征级对齐后, 检测模型在迁移集 1 中的结果已经相当接近全监督下的检测结果, 再使用自训练则容易发生过拟合. 而在迁移集 2 则可以得到质量更好的初始伪标注, 从而更有利于检测性能的提升. 由上可见, 不同的方法在不同的数据集上有不同的效果, 但综合不同的方法可以弥补各自方法的不足, 进而实现更好的迁移检测性能.

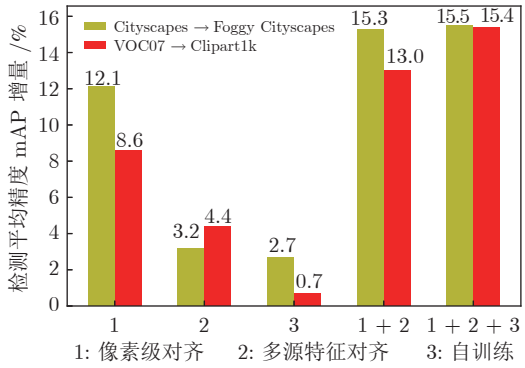


图 11 每一成分对 mAP 的提升

Fig. 11 The mAP gain of each component

2.3.3 源域数量的影响

通过基于结构分解的多样性图像翻译, 可以得到不同风格属性的翻译图片, 并将其作为源域数据集. 在得到多样性翻译图像时, 有两种策略: 1) 将源域图像与随机的属性特征相结合, 如随机噪声; 2) 将源域图像与指定的属性图片相结合, 这里主要指目标域中的图片属性. 图 12 展示了由 Cityscapes → Foggy Cityscapes 生成的 3 种不同风格的图片. 其中, 第一列为输入的不同内容属性图片, 最上面一行为 3 种不同的目标域风格属性, 其在颜色、色调、纹理、风格等方面存在差异. 通过将每一张源域内容图片与目标域的风格属性相结合, 从而可以为每一张内容图片生成多种带有目标域不同风格的翻译图像. 他们分别保留了源域图片的空间内容特征, 却带有不同的风格属性. 这样生成的多样性图片包含了目标域不同方面的信息, 通过特征提取可以得到多样性的特征表达, 然后再使用多域特征对齐网络, 得到多个领域不变的特征表示, 从而具有更好的鲁棒性与泛化性能. 此外, 通过将源域图片与随机属性如高斯噪声相结合, 也能生成随机的多样性翻译图像. 由图 12 可以看出, 使用目标域属特征生成的翻译图像在表观特征上与目标域更为相似. 不同的是, 使用 CycleGAN 只能得到单一属性的翻译图像.

源域数据的多样性直接影响到最终的检测结果, 表 3 给出了源域数量 M 对实验结果的影响. 当 $M = 0$ 时, 为基本的检测模型. 可以直观地看出, 在像素级对齐和多源域特征对齐实验中, 随着源域数据多样性 M 的增加, 在目标域上的检测结果不断提升. 在多样性图像翻译过程中, 可以将源域图片的内容特征与任意的目标域风格属性特征相结合, 因此可以得到多种不同风格的翻译图片. 受限于显卡内存, 本文只取了 $M = 3$, 在实际应用中可以取更大的 M 值, 并在理论上得到比本文报告中更好的检测迁移效果. 同时, 不同的属性特征也会影响到最终的检测性能. 表 4 给出了不同属性特征对目标域

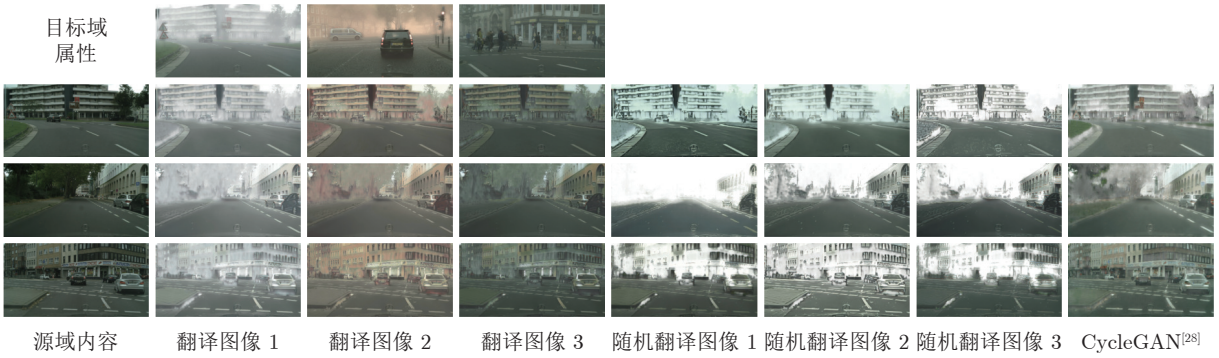


图 12 图像翻译结果示例图

Fig. 12 Sample results of translated images

最终检测结果的影响. 由表 4 可以看出, 在基于 SSD 或 Faster R-CNN 的跨域检测框架下, 通过使用目标域指定属性得到的检测结果都明显优于随机属性得到的检测结果.

表 3　在 Cityscapes → Foggy Cityscapes 实验中源域数量 M 对检测性能的影响 (%)

Table 3　Impact of the number of source domains M on the detection performance in Cityscapes → Foggy Cityscapes (%)

M	0	1	2	3
像素级对齐 mAP	17.4	27.3	28.9	29.5
多源域特征对齐 mAP	17.4	29.6	30.3	32.7

表 4　Cityscapes → Foggy Cityscapes 实验中属性特征对检测性能的影响 (%)

Table 4　Impact of attribute features on the detection performance in Cityscapes → Foggy Cityscapes (%)

方法	框架	mAP
像素级对齐 (随机属性)	SSD	29.0
像素级对齐	SSD	29.5
多源域特征对齐 (随机属性)	SSD	31.6
多源域特征对齐	SSD	32.7
像素级对齐 (随机属性)	Faster R-CNN	34.7
像素级对齐	Faster R-CNN	38.0
多源域特征对齐 (随机属性)	Faster R-CNN	36.5
多源域特征对齐	Faster R-CNN	38.7

2.3.4　参数 λ 敏感性分析

在多源域特征对齐的训练过程中, 式 (11) 中参数 λ 的设置对检测损失与对抗损失的平衡起到关键作用. 表 5 给出了 VOC07 → Clipart1k 实验中, 不同 λ 取值得到的检测结果. 从表中可以看出, 在多源域特征对齐网络中, 参数 λ 的取值过大或过小都不利于最终的检测结果. 当参数 λ 过小时, 多源域判别器的梯度反向传播值相对较小, 因此不能很好地训练判别器以得到多个领域不变的特征表示; 当参数 λ 过大时, 多源域判别器会反向传播不正确的梯度值, 将不利于检测性能的提升.

表 5　在 VOC07 → Clipart1k 实验中参数 λ 的敏感性分析 (%)

Table 5　Sensitivity analysis of λ in VOC07 → Clipart1k (%)

λ	0.1	0.5	1.0	1.5	2.0
mAP	34.0	36.1	36.3	36.1	35.7

2.3.5　阈值 θ 的敏感性分析

在自训练过程中, 根据在目标域训练集上的预测边框分类得分来选取伪标注. 当阈值 θ 取值较高时, 尽管得到的伪标注更为可信, 但会遗漏大量的有用标注. 当阈值 θ 值较小时, 预测分类得分较低的边框包含其中, 从而造成大量的错误标注. 因此, 阈值 θ 设定直接影响到生成的伪标注的质量. 表 6 给出了 VOC07 → Clipart1k 试验中, 不同 θ 取值得到的检测结果. 可以看到, 在第一轮自训练过程中, 当 $\theta = 0.2$ 时取得了最好的检测效果. 由于目标域训练数据比较少 (只有 500 张图片), 当阈值 θ 较大时, 大量的图片上无法生成伪标注. 此外, 本文分析了多轮自训练的策略. 通过设置不同的阈值 θ, 在每轮自训练后, 选取效果最好的 θ. 由于第 1 轮自训练后模型的性能渐进提升, 在下一轮自训练时, 将只选取更大的 θ, 以生成更为可靠的伪标注. 如表 6 所示, 总共进行了 3 轮自训练. 在第 2 轮自训练时, 在阈值 $\theta = 0.6$ 或 $\theta = 0.7$ 时取得了最好的效果. 而在第 3 轮自训练时, 已无法再提升模型的检测性能. 通过这种多轮次与渐进提升阈值 θ 的自训练策略, 可以有效提升在目标域上的检测性能.

表 6　在 VOC07 → Clipart1k 实验中阈值 θ 的敏感性分析 (%)

Table 6　Sensitivity analysis of θ in VOC07 → Clipart1k (%)

θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
第 1 轮 mAP	37.8	**38.3**	38.1	37.8	37.7	37.2	36.2	35.7	35.0
第 2 轮 mAP	–	–	38.0	38.3	38.4	**38.6**	**38.6**	38.3	38.4
第 3 轮 mAP	–	–	–	–	–	–	**38.6**	38.3	38.4

3　结束语

本文提出了一种基于渐进对齐的无监督跨域目标检测算法. 首先, 针对现有图像翻译中生成图像风格单一、语义结构信息不一致的问题, 通过图像特征分解实现图像的结构化翻译, 将源域的内容特征与目标域的任意属性特征结合, 生成了从源域到目标域映射的多样性图片, 并将源域的标注信息迁移到生成数据, 实现了像素级域自适应; 其次, 为了避免单源域迁移中特征对齐时出现的源域偏向性问题, 设计多领域自适应网络, 得到多领域不变的特征表示, 实现了多样性特征级域自适应; 最后, 通过自训练在目标域上生成伪标签, 进一步提升了模型在目标域上的检测性能. 多个数据集上的实验结果表明, 本文提出的算法取得了令人满意的效果. 与此同时, 由于本文在实现迁移的过程中给予了每个源域样本同等的权重考虑, 而没有考虑不同样本对目标域的迁移效果, 这个问题可作为开展下一步研

究工作的方向.

References

1 Liu L, Ouyang W L, Wang X A, Paul W. F, Jie C, Liu X W, et al. Deep learning for generic object detection: A survey. arXiv preprint, 2018, arXiv: 1809.02165

2 Zhang Hui, Wang Kun-Feng, Wang Fei-Yue. Advances andperspectives on applications of deep learning in visual objectdetection. *Acta Automatica Sinica*, 2017, **43**(8): 1289−1305
(张慧, 王坤峰, 王飞跃. 深度学习在目标视觉检测中的应用进展与展望. 自动化学报, 2017, **43**(8): 1289−1305)

3 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 (NIPS), Lake Tahoe, USA: IEEE, 2012. 1097−1105

4 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, USA: IEEE, 2014. 580−587

5 Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile: 2015. 1440−1448

6 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 (NIPS), Montreal, Canada: IEEE, 2015. 91−99

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

8 Liu W, Anguelov D, Erhan D, Szegedy C, Reed E, Fu C Y, et al. SSD: Single shot multi-box detector. In: Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, Netherlands: Springer International Publishing, 2016. 21−37

9 Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler N, Benenson R, 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 Computer Society, 2016. 3213−3223

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

11 Li D, Huang J B, Li Y L, Wang S J, Yang M H. Weakly supervised object localization with progressive domain adaptation. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Amsterdam, Netherland: 2016. 3512−3520

12 Bilen H, Vedaldi A. Weakly supervised deep detection networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Amsterdam, Netherland: 2016. 2846− 2854

13 Zhang Xue-Song, Zhuang Yan, Yan Fei, Wang Wei. Status and development of transfer learning basedcategory-level object recognition and detection. *Acta Automatica Sinica*, 2019, **45**(7): 1224−1243
(张雪松, 庄严, 闫飞, 王伟. 基于迁移学习的类别级物体识别与检测研究与进展. 自动化学报, 2019, **45**(7): 1224−1243)

14 Sun B C, Feng J S, Saenko K. Return of frustratingly easy domain adaptation. In: Proceedings of the 2016 Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA: AAAI Press, 2016. 2058−2065

15 Long M S, Cao Y, Wang J M, Jordan M. Learning transferable features with deep adaptation networks. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: 2015. 97−105

16 Peng X C, Usman B, Saito K, Kaushik N, Hoffman J, Saenko K. Syn2Real: A new benchmark for synthetic-to-real visual domain adaptation. arXiv preprint, 2018, arXiv: 1806.09755

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

18 Kim T, Jeong M K, Kim S, Choi S, Kim C. 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, USA: 2019. 12456−12465

19 Chen Y H, Li W, Sakaridis C, Dai D X, Gool L V. 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, USA: 2018. 3339−3348

20 Saito K, Ushiku Y, Harada T, Saenko K. Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA: 2019. 6956−6965

21 Zhu X G, Pang J M, Yang C Y, Shi J P, Lin D H. Adapting object detectors via selective cross-domain alignment. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA: 2019. 687−696

22 Cai Q, Pan Y W, Ngo C W, Tian X M, Duan L Y, Yao T. Exploring object relation in mean teacher for cross-domain detection. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA: 2019. 11457−11466

23 He Z W, Zhang L. Multi-adversarial faster-RCNN for unrestricted object detection. In: Proceedings of the 2019 IEEE International Conference on Computer Vision Workshops, Seoul, South Korea: 2019. 6667−6676

24 Xie R C, Yu F, Wang J C, Wang Y Z, Zhang L. Multi-level domain adaptive learning for cross-domain detection. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshops. Seoul, South Korea: 2019. 3213−3219

25 Wang T, Zhang Y P, Yuan L, Feng J S. Few-shot adaptive faster R-CNN. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 7173−7182

26 Khodabandeh M, Vahdat A, Ranjbar M, Macready W G. A robust learning approach to domain adaptive object detection. In: Proceedings of the 2019 IEEE International Conference on Computer Vision Workshops. Seoul, South Korea: 2019. 480−490

27 Wang X D, Cai Z W, Gao D S, Vasconcelos N. Towards universal object detection by domain attention. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA: 2019. 7289−7298

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

29 Goodfellow I J, Pouget-Abadie J, Mehdi M, Bing X, David W F, Sherjil O, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada: 2014. 2672−2680

30 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint, 2014, arXiv:

1409.1556

31 Lee H Y, Tseng H Y, Huang J B, Singh M, Yang M H. Diverse image-to-image translation via disentangled representations. In: Proceedings of the Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: 2018. 36–52

32 Ulyanov D, Vedaldi D, Lempitsky V. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: 2017. 4105–4113

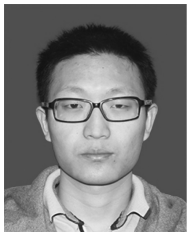
33 Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. In: Proceedings of the 2017 Domain Adaptation in Computer Vision Applications. Cham, Switzerland: 2017. 189–209

34 Zhao H, Zhang S H, Wu G H, Moura J, Costeira J P, Gordon G J. Adversarial multiple source domain adaptation. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Montreal, Canada: 2018. 8568–8579

35 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: 2015. 448–456

36 Everingham M, Gool L J, Williams C, Winn J, Zisserman A. Semantic the pascal visual object classes (VOC) Challenge. *International Journal of Computer Vision*, 2010, **88**(2): 303–338

37 Diederik P K, Jimmy B. Adam: A method for stochastic optimization. arXiv preprint, 2014, arXiv: 1412.6980



李 威 昆明理工大学信息工程与自动化学院硕士研究生. 主要研究方向为图像处理, 计算机视觉以及模式识别. E-mail: leesoon2049@gmail.com

(**LI Wei** Master student at the School of Information Engineering and Automation, Kunming University of Science and Technology. His research interest covers image processing, computer vision and pattern recognition.)

王 蒙 博士, 昆明理工大学信息工程与自动化学院副教授. 主要研究方向为图像处理, 计算机视觉以及模式识别. 本文通信作者.
E-mail: wmeng06@126.com

(**WANG Meng** Ph.D., associate professor at the School of Information Engineering and Automation, Kunming University of Science and Technology. His research interest covers image processing, computer vision and pattern recognition. Corresponding author of this paper.)