频域内面向目标检测的领域自适应

李岳楠[1]，徐浩宇[1]，董　浩[2]

(1. 天津大学 电气自动化与信息工程学院，天津 300072；
2. 天津津航技术物理研究所，天津 300308)

摘　要：近年来，基于深度学习的目标检测技术在机器人、自动驾驶和交通监控等领域有着广泛的应用。然而，由于训练集和测试集样本分布偏差的原因，将现成的预训练检测器应用到实际开放场景时通常会出现明显性能下降。针对该问题提出了一种频域内的领域自适应方法，利用离散余弦变换的频域能量集中特性，通过在频域内对少数重要频率系数进行处理，实现了面向目标检测的领域自适应，降低了对存储和计算资源的要求并减少了领域差异。该方法可以分为两个阶段：第一阶段使用无监督图像转换方式，将源域已标注的训练数据向目标域作转换；第二阶段采用基于对抗的领域自适应方法训练目标检测模型，对转换后的训练数据与目标域内的数据作特征适配。针对不同天气场景的目标识别实验表明：所提的频域内领域自适应方法在 4 种领域自适应对比算法中排名第一，与仅用源域数据训练的模型相比，mAP 值提升了 33.9%。

关键词：领域自适应；　目标检测；　图像转换；　频域

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.3788/IRLA20210638

Domain adaptation for object detection in the frequency domain

Li Yuenan[1], Xu Haoyu[1], Dong Hao[2]

(1. School of Electrical & Information Engineering, Tianjin University, Tianjin 300072, China;
2. Tianjin Jinhang Institute of Technical Physics, Tianjin 300308, China)

Abstract: Deep learning-based object detection technology has recently made significant progress and has a wide range of applications in robotics, autonomous driving, traffic surveillance, etc. However, due to the distribution discrepancy between the training and testing datasets, the off-the-shelf detectors pre-trained using the data in a specific domain often show apparent performance degradation when applied in wild scenarios. To address this problem, a domain adaptation method for object detection in the frequency domain is proposed. In light of the energy concentration property of the discrete cosine transform, the proposed algorithm conducts domain adaptation for object detection by processing only a few of the most significant frequency coefficients, which reduces memory and computing resource consumption and alleviates the domain shift problem. The proposed method consists of two stages. In the first stage, it translates annotated training data from the source domain to the target domain using unsupervised image-to-image translation. Adversarial domain adaptation is then applied to the object detection model to align the features of the translated data and the real data in the target domain. The experimental results of the object detection under different weather conditions show that the proposed method ranks first among the four testing algorithms. Compared with the object detection model trained

收稿日期：2022−01−20；　**修订日期：**2022−03−15

基金项目：国家自然科学基金 (61972281, 61572352)

作者简介：李岳楠，男，副教授，硕士生导师，博士，主要从事多媒体信号处理方面的研究。

with only source domain data, it can increase the mAP value by 33.9%.

Key words: domain adaptation; object detection; image translation; frequency domain

0 引 言

目标检测是计算机视觉中的一个重要任务。近年来,基于卷积神经网络 (Convolutional Neural Networks, CNN) 的工作大幅提高了目标检测的精度。目前,绝大多数目标检测算法以有监督的方式进行训练,数据标注工作需要耗费大量人力资源。此外,训练和测试样本间的差异性导致目标检测算法在新场景中的泛化能力不强。以不同天气下的检测任务为例,用晴朗天气下采集的图像训练的检测模型在雾霾天气下的检测精度通常较低。针对该问题,现有的解决方法主要分为两种:一是使用图像无监督转换的方式,将已有标注的图像 (源域) 转换到目标域,构建新的数据集进行训练;二是采用领域自适应的方式,将源域和目标域的数据映射到同一特征空间,以减小不同领域之前的差距。然而,这两种方法均存在一定的局限性。受计算资源和存储空间的限制,图像无监督转换通常仅能接受低分辨率的输入 (如 CycleGAN[1] 仅接受 256×256 和 512×512 的输入图像),对于高分辨率的输入图像,通常的做法是将原始图像降采样后输入网络,之后再升采样回原始分辨率,这种方式造成了细节内容的损失,难以获得高清晰度的输出图像且不利于后续检测任务。另一方面,领域自适应的效果也同样受到输入图像尺寸的影响。

为了减少降采样操作造成的信息丢失并节省计算资源,受到频域能量集中特性的启发,文中结合无监督图像转换和基于对抗的领域自适应两种方式,提出了一种面向目标检测的频域内的领域自适应方法。该方法分为两个阶段,第一阶段通过无监督图像转换的方式将带有标注的源域图像 (如晴天图像) 变换到与目标域 (如雾天图像) 相近的图像,并将变换后的图像所在的域定义为中间域。第二阶段通过基于对抗学习的领域自适应方法将中间域的数据与目标域 (如真实有雾图) 的数据在特征空间内作适配,两个阶段均在频域内完成。由于图像不同频带具有不同的视觉重要性,频域系数具备天然的可压缩属性。图像变换到频域后,能量集中到低频和中频频带,对少数几个频率系数处理就可以实现无监督转换和领域自适应,降低了训练和测试过程对计算资源和存储空间的要求。实验结果表明,第一阶段无监督图像转换能够生成与目标域相近的中间域图像,第二阶段基于对抗学习的领域自适应方法能够减少传统降采样操作造成的信息丢失,并显著提高目标域的检测性能。

1 相关研究工作

1.1 目标检测

近年来,绝大多数目标检测算法都采用基于卷积神经网络 CNN 的结构[2],这些工作又可以分为基于区域生成的两阶段方法和直接获得检测结果的一阶段方法。在两阶段方法中,R-CNN[3] 使用选择性搜索 (Selective Search) 得到物体的候选框,并使用支持向量机 (Support Vector Machine, SVM) 对特征进行预测。Fast R-CNN[4] 改进了特征的预测方式,使用神经网络进行检测框的分类与回归。Faster R-CNN[5] 进一步改进了 Fast R-CNN,使用区域生成网络 (Region Proposal Network, RPN) 替代耗时的选择性搜索,实现了实时目标检测算法。一阶段检测方法的代表性算法有 SSD[6]、YOLOv3[7]、RetinaNet[8] 等,这类方法能够进一步提高目标检测的实时性能。吴天舒等人[9] 结合深度可分离卷积,采用轻量化特征提取最小单元对 SSD 做轻量化处理,使其可以在移动设备上运行。遆晓光等人[10] 将视频图像向二维频域投影后,结合主动滤波和图像重构,能够检测出弱小运动目标。吴言枫等人[11] 通过提取图像中的显著性区域,并使用自适应双高斯算法分割出前景,提升了复杂天空背景下的目标检测精度。此外,还有一些方法通过改进检测器中的结构[12−13] 来提升复杂背景下以及小目标的检测精度。尽管基于卷积神经网络的检测器已经达到了较高的精度,但是现有检测模型对训练集与测试集之间分布不一致性较为敏感,在新场景的应用中泛化性能较差。

1.2 领域自适应和无监督图像转换

经典的有监督学习任务往往假设训练集和测试集分布一致,但是实际测试数据一般与理想环境下的训练数据有很大差异,迁移学习 (Transfer Learning) 是

应对这一问题的主要技术。

领域自适应 (Domain Adaptation) 是迁移学习的一种，其主要思想是将不同领域 (如不同天气的图像) 的数据映射到同一个特征空间，以减少领域之间的差距，提高模型的泛化性和鲁棒性。领域自适应一开始被用于图像分类任务，然后推广到目标检测等任务，领域自适应总体上可以分为基于人工定义约束的方式和基于对抗训练的方式。前者通过缩小两个分布之间的距离度量实现源域与目标域特征之间的对齐，常见的度量分布之间距离的方法有 KL-散度、H 散度、最大平均差距 (Maximum Mean Discrepancy, MMD) 等。Ganin 等人[14] 使用基于对抗的方法使神经网络缩减域差异，并提出了梯度反转层 (Gradient Reversal Layer, GRL)。梯度反转层应用在数据特征与域鉴别器之间，在前向传播过程中梯度保持不变，在反向传播过程中梯度方向取反，使得域鉴别器与主任务网络能够对抗地进行训练，实现了真正意义上的端到端训练，避免了生成对抗网络 (Generative Adversarial Nets, GAN) 中生成器与鉴别器交替训练的模式。近年来，一些研究通过多阶段、多尺度训练、特征融合、注意力机制、去耦合学习等方法提升了领域自适应的效果[15−18]。

无监督图像转换需要在不成对的图像样本之间学习一个映射，将一个领域的图像映射到另一个领域。无监督图像转换的方法也可以用于领域自适应。CycleGAN[1] 中提出了循环一致性损失，将图像转换到另一个领域后再使用逆映射转换回来，并要求经过循环变换的图像与输入图像一致，同时在两个领域中引入了鉴别器对相应的映射进行约束。UNIT[19] 算法中提出了共享潜空间 (Shared latent space) 思想，假设不同域的图像能够映射到同一潜空间。基于这个思想，该算法将图像在不同域之间的变换过程拆分为潜空间编码和解码两个子过程，并引入变分自编码器对潜空间向量进行约束并结合其它限制条件来提升无监督图像转换的效果。无监督图像转换尽管能够生成与目标域十分相近的图像，但在计算资源受限的条件下，图像转换网络往往只能接受低分辨率图像作为输入。此外，由于无监督图像转换本身是一个欠定问题，无法保证生成图像分布与目标域完全相同，在进行下游计算机视觉任务时仍然存在特征分布不一致的领域偏移 (Domain shift) 问题。

1.3 频域内的深度学习与领域自适应

Xu 等人[20] 首次提出在频域内训练神经网路，使用离散余弦变换 (Discrete Cosine Transform, DCT) 后的变换系数作为输入，并应用于图像分类和分割任务。

Yang 等人[21] 以一种非学习的方式对源域和目标域的图像分别进行快速傅里叶变换 (Fast Fourier Transform, FFT)，然后使用目标域图像幅值的中心 (低频) 区域替换源域图像相应的幅值并保持相位不变，之后采用快速傅里叶逆变换 (Inverse Fast Fourier Transform, IFFT) 还原出图像。该算法不需要训练，能在一定程度上实现图像间的领域变换。

2 基于频域的领域自适应方法

传统的目标检测和领域自适应方法一般在空域进行，以空域像素作为输入，在一些资源受限的场景下，例如移动设备、嵌入式系统中，由于图像数据量很大，在空域进行计算会带来巨大的计算开销。为了提高推理速度、降低通信带宽和内存开销，传统方法通常将高分辨率的空域 RGB 图像降采样为低分辨率的图像。这种方法造成的信息损失对机器视觉任务的性能有明显影响。

文中利用频域变换的能量集中特性，实现计算资源和检测性能的平衡，所提算法先将输入图像从 RGB 空间转换到 YCbCr 空间，然后使用离散余弦变换 DCT 得到图像的频域表示。在此基础上，文中提出了一种频域内面向目标检测的领域自适应方法。以不同天气下的检测任务为例，源域是晴朗天气下采集的图像，目标域是雾霾天气下采集的图像。由于晴朗天气图像和雾霾天气图像差距很大，直接在源域$\mathcal{S}$域 (晴朗天气) 图像和目标域$\mathcal{T}$域 (雾霾天气) 图像之间做领域自适应十分困难。受到 CycleGAN[1] 的启发，文中先采用无监督图像转换的方式，将源域图像转换为合成的雾霾图像 (中间域)，并记为$\mathcal{I}$域，然后使用对抗学习的方式使检测器实现在$\mathcal{I}$域和$\mathcal{T}$域之间的领域自适应，算法整体框架如图 1 所示。

文中的图像转换、领域自适应、目标检测算法均在频域内实现。

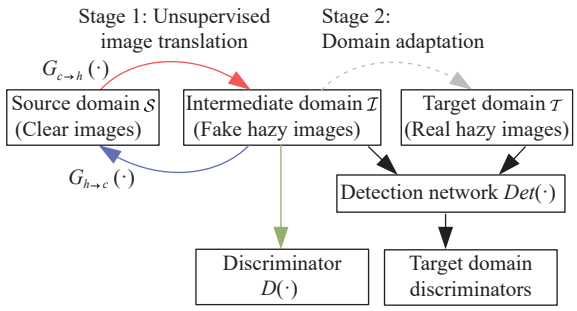


图 1 频域内两阶段的领域自适应过程

Fig.1 Two-stage domain adaptation process in the frequency domain

2.1 频域内的数据预处理

频域内的数据预处理将 RGB 空间的像素值转换为频域表示。首先，图像在 RGB 空间上进行数据增广，如随机缩放裁剪，随机翻折，对比度、亮度变换等。YCbCr 颜色空间内的每个通道划分为互不重叠的8×8大小图像块做分块 DCT 处理。假设原始图像的尺寸为$H\times W\times 3$，DCT 处理得到的系数特征尺寸为$H/8\times W/8\times 192$(YCbCr 每个通道内各 64 个系数)。为了实现数据缩减，文中对每个8×8块内的频域系数做之字型 (Zigzag) 排列，并在 YCbCr 三个颜色通道内分别选取左上角的系数，Y 空间选取 22 个系数，Cb 和 Cr 空间分别选取 21 个系数，共选取 64 个系数作为每个块的频域表示。由于不同频率下 DCT 系数的数据范围相差很大，低频分量的绝对值很大而高频分量的数量级很小。文中对输入的 DCT 系数特征作标准化处理，预先统计了数据集所有图像转换到频域后 64 个不同系数的均值和标准差，使用每个系数的均值和标准差对输入特征作标准化处理。

2.2 源域到中间域的无监督图像转换

在领域自适应的第一阶段，文中将源域$\mathcal{S}$(晴朗天气) 图像使用无监督的方式变换到目标域$\mathcal{T}$(雾霾天气)。如图 2 所示，输入的无雾图像I_c经过 DCT 处理后转换为 64 通道的 DCT 系数特征F_c，该特征经过生成网络$G_{c\to h}(\cdot)$后，变换为相同场景的有雾图像的 DCT 系数特征$F_{c\to h}$，再经过一个相同结构、不同权重的生成网络$G_{h\to c}(\cdot)$重新变换回无雾图的 DCT 系数$F_{c\to h\to c}$。经过两次转换后的 DCT 系数$F_{c\to h\to c}$与原始的 DCT 系数F_c之间计算循环一致性损失[1]，循环一致性损失定义如下：

$$L_{cyc}=\|G_{h\to c}(G_{c\to h}(F_c))-F_c\|_1 \tag{1}$$

为了适应频域的特性，增强特征提取能力，文中采用如图 2 所示的生成网络做频域内的图像转换。生成网络由 n 个$\mathcal{B}$模块级联而成。与深度残差网络 (ResNet) 类似，每个$\mathcal{B}$模块学习的是无雾特征和有雾特征之间的局部残差。在一个$\mathcal{B}$模块中，输入特征经过3×3卷积层和 ReLU 激活函数后，得到的中间结果与输入相加，再分别经过3×3卷积层、ReLU 激活函数、另一层3×3卷积层和一层通道注意力层[22]，最后与输入特征相加，得到该$\mathcal{B}$模块输出的结果。通道注意力层将通道数为c的特征分别经过$c\to\frac{c}{k}$和$\frac{c}{k}\to c$的卷积和一个 Sigmoid 函数，得到每个通道的权重系数，再将输入的特征与每个通道的权重系数相乘。由于不同频率的 DCT 系数对图像视觉效果的影响差异较大，幅度量级也各不相同，通道注意力层学习不同频率 DCT 系数的权重，有助于增强特征表示能力，能更好的刻画目标域的特性。

文中引入了一个鉴别网络用于判别频域转换的结果是否接近目标域的图像特性。鉴别网络接受转换后合成有雾图的 DCT 系数$F_{c\to h}$或者真实有雾图的 DCT 系数F_h作为输入，输出该图像属于目标域的概率。鉴别网络$D(\cdot)$前半部分由若干3×3卷积层、ReLU 激活函数，以及通道注意力层组成，后半部分由若干全连接层和 ReLU 激活函数组成，在前半部分末尾有一个展平操作，将前半部分输出的特征展平，便于与全连接层相连接。

鉴别网络$D(\cdot)$的优化目标是根据 DCT 系数来区分合成图像和目标域图像。生成网络$G(\cdot)\{G_{h\to c}(\cdot),G_{c\to h}(\cdot)\}$与鉴别网络$D(\cdot)$相对抗，目标是使合成图像和目标域图像在频域内不可区分。鉴别网络与生成网络交替训练，损失函数定义如下：

$$\min_{G(\cdot)} L_G=[D(G_{c\to h}(F_c))-1]^2 \tag{2}$$

$$\min_{D(\cdot)} L_D=[D(F_h)-1]^2+D^2(G_{c\to h}(F_c)) \tag{3}$$

DCT 处理能够将图像的尺寸缩小为原来的 1/8，并通过省略不重要的高频信息来压缩显存占用率，避免了降采样带来的信息损失，从而在同等的显存条件下能够获得分辨率更高，细节更逼真的变换效果。

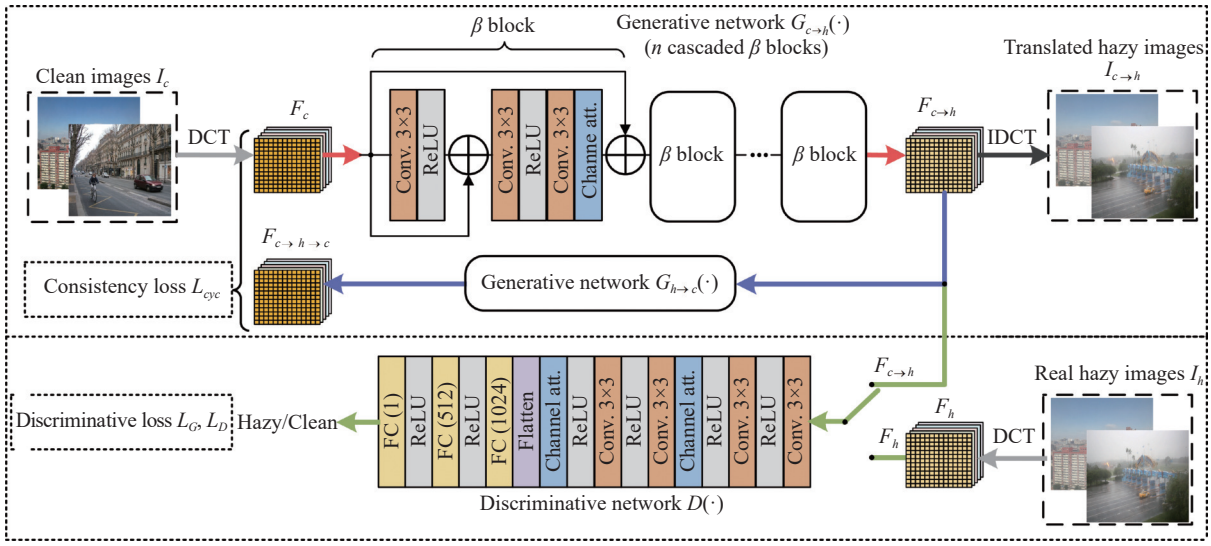


图 2 源域到中间域的无监督图像转换

Fig.2 Unsupervised image translation from source domain to intermediate domain

2.3 中间域到目标域的领域自适应

在领域自适应的第二阶段，文中在中间域 (由 2.2 节生成的合成有雾图像) 和目标域 (真实的有雾图像) 之间做领域自适应。由于$F_{c\to h}$是由无雾图像合成的，与无雾图像具有相同的场景，和无雾图像共享相同的目标检测标签 (检测框坐标和类别)，可用于有监督训练。如图 3 所示，对目标检测模型的领域自适应由两类训练样本构成，分别为带有标签信息的合成有雾图 DCT 系数$F_{c\to h}$和不带标签的真实有雾图 DCT 系数F_h。

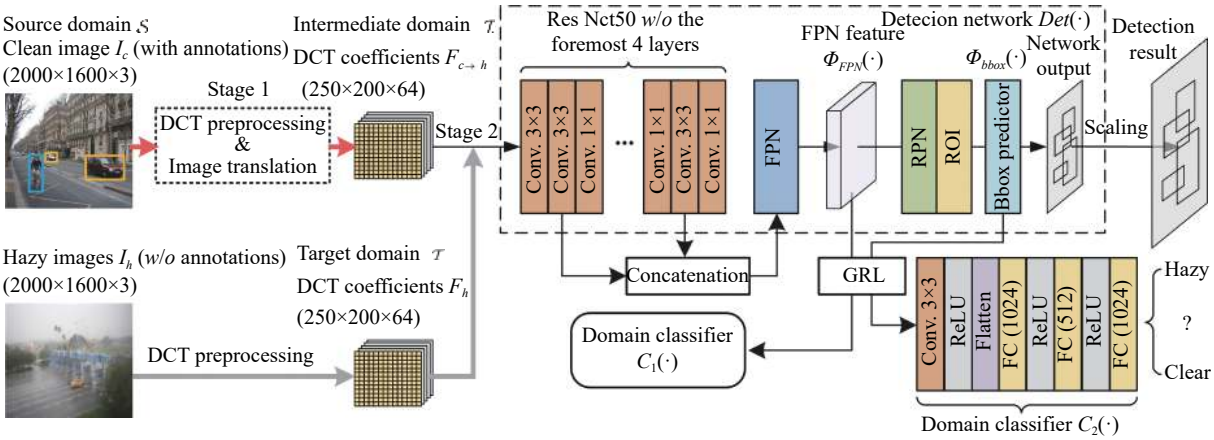


图 3 频域内面向检测的领域自适应

Fig.3 Domain adaptation for object detection in the frequency domain

为了使网络能够接受 64 通道的特征系数作为输出，文中去掉了目标检测网络骨干网络 ResNet[23] 最前面的四层，即卷积层 (卷积核为7×7，步长为 2)、批正则化 (BatchNorm)、ReLU 激活函数以及最大值池化层 (Max Pooling)。输入的 DCT 系数在经过去掉前四层的 ResNet50 后，经过特征金字塔网络 (Feature Pyramid Network, FPN) 融合特征。融合后的特征经过区域生成网络 (Region Proposal Network, RPN)、ROI 池化 (ROI Pooling) 和目标框回归操作，得到网络的预测结果，预测结果经过非极大值抑制 (Non-Maximum Suppression, NMS) 和缩放后，得到最终的检测结果。给定 DCT 系数特征 F，目标检测的损失函数定义如下：

$$L_{det}(F) = L_{rpn} + L_{cls} + L_{reg} \tag{4}$$

式中：$L_{rpn}, L_{cls}, L_{reg}$分别代表 RPN、框分类和框回归的

损失函数[5]。

为了实现领域自适应，文中分别将特征融合层和目标回归层输出的特征送入两个相同结构、不同权重的域分类器$C_1(\cdot)$和$C_2(\cdot)$，用于从目标检测网络输出特征的角度判断给定的图像是属于中间域$\mathcal{I}$还是目标域$\mathcal{T}$，如图 4 所示。域分类的结构为3×3卷积层、ReLU激活函数、展平操作以及三组全连接层和 ReLU 激活函数，最终输出 0 代表$\mathcal{I}$域，1 代表$\mathcal{T}$域，即特征属于$\mathcal{T}$域的概率。特征与域分类器之间由一层梯度反转层[14](Gradient Reversal Layer, GRL) 连接，训练过程中两个特征通过 GRL 层与域分类器相连，梯度在经过 GRL 层时会进行反转，相当于使目标检测器缩小中间域和目标域在图像上的领域差异，实现同时优化检测网络和域分类器的目的，损失函数定义如下：

$$\begin{aligned}\min_{C_1(\cdot)}\min_{C_2(\cdot)}\max_{Det(\cdot)} L_{DA} =&[C_1(\Phi_{FPN}(F_h))-1]^2+\\ &C_1^2(\Phi_{FPN}(F_{c\to h}))+\\ &[C_2(\Phi_{bbox}(F_h))-1]^2+\\ &C_2^2(\Phi_{bbox}(F_{c\to h}))\end{aligned} \tag{5}$$

式中：$Det(\cdot)$表示检测网络；$\Phi_{FPN}(\cdot)$和$\Phi_{bbox}(\cdot)$分别表示目标检测网络中的特征金字塔和目标框回归器，这两部分用于计算中间特征。

(a) 输入无雾图像
(a) Input hazy image

(b) 频域内无监督图像转换
(b) Image translation in the frequency domain

(c) 空域内无监督图像转换
(c) Image translation in the spatial domain

图 4 频域和空域空间无监督图像转换可视化对比

Fig.4 Visual comparison between the unsupervised image translation results in the frequency and spatial domains

3 实验结果和分析

3.1 数据集与实现细节

文中实验均在 Cityscapes 数据集和 Foggy Cityscapes 数据集上进行。Cityscapes 数据集是一个街道场景的图像数据集，包含 2975 张训练集图像以及 500 张验证集图像，图像的分辨率均为 2048×1024 pixel，该数据集包含物体的分割标注。为了适应目标检测任务，文中对图像分割标注中每一个连通域物体取外接矩形作为检测的标注框。

Foggy Cityscapes 是基于 Cityscapes 构建的数据集，该数据集使用 Cityscapes 提供的景深信息模拟了 3 种不同级别的雾霾天气，模拟的过程可参考原论文。该数据集包含 8895 张训练集图像以及 1500 张验证图像，即 Cityscapes 中每张图像对应 3 种不同浓度的有雾霾图像。

文中算法代码基于 PyTorch[24] 编写。在第 1 阶段无监督图像转换阶段，网络中$\mathcal{B}$模块的个数 n 为 24，Cityscapes 数据集中的无雾图像以原图尺寸 (2048×1024) 作为输入，经过 DCT 预处理后变为尺寸为 256×128×64 的系数特征。经过生成网络$G_{c\to h}(\cdot)$后，得到有雾图像的 DCT 系数特征$F_{c\to h}$。Foggy Cityscapes 数据集中的图像也作了相同的 DCT 预处理，以无监督的方式交替优化生成器$G(\cdot)\{G_{c\to h}(\cdot),G_{h\to c}(\cdot)\}$和鉴别器$D(\cdot)$。使用 Adam 优化器训练 100 代，学习率固定为2×10^{-4}。

在第 2 阶段中间域到目标域的领域自适应阶段，目标检测使用 Faster RCNN 网络。原始的 Faster RCNN 通常是将 RGB 图像缩放为短边为 600，长边不超过 1000 的图像。在文中的方法中，目标域的图像从 Foggy Cityscapes 数据集采样，使用原始图像作为输入 (2048×1024)，经过颜色和亮度增强、随机翻折数据增强后，转换到 YCbCr 空间并分块作离散余弦变换后每个块

内选取 64 个系数，最终得到150×250×64的系数特征。而中间域图像由无雾图像转换得到的有雾的 DCT 系数特征直接作为输入。预先计算了所有训练集图像的 DCT 系数的均值和方差，并对输入的 DCT 特征作标准化处理。 使用随机梯度下降 (Stochastic Gradient Descent, SGD) 算法训练，共训练 12 代，第 1 代为学习率预热 (warmup)，学习率为1×10^{-4}，第 2 代开始学习率调整为1×10^{-3}，在第 8 代和第 11 代进行学习率衰减，学习率分别变为原来的 1/10。由于输入的 DCT 特征尺寸较小，本位将锚框面积尺寸调整为$\{128^2, 64^2, 32^2, 16^2, 8^2\}$，以适应目标物体大小的变化，锚框的长宽比仍然是{1:1, 1:2, 2:1}不变。

3.2 无监督图像转换实验结果

为了可视化无监督图像转换$G_{c\to h}(\cdot)$的效果，文中对转换网络$G_{c\to h}(\cdot)$输出的系数$F_{c\to h}$作了逆离散余弦变换 (Inverse Discrete Cosine Transform, IDCT)，结果如图 4 所示，从图中可以看出，频域内的无监督图像转换能够将清晰图像进行加雾渲染生成有雾图像，转换后的图像具有目标域特性。

同时，文中也与空域中的算法 CycleGAN[1] 作了对比。为了公平起见，将文中提出的频域内的无监督转换与 CycleGAN 使用相同的骨干网络训练相同的代数，并控制模型所需运算量 GFLOPS 相同。文中算法使用原图尺寸作为输入，经过 DCT 预处理后变为 256×128×64 的 DCT 系数特征，CycleGAN 将输入图像降采样到 256×128×3，通过一层3 → 64通道的卷积层得到 256×128×64 的特征。图 4 中可视化了在 RGB 空间进行无监督图像转换的结果。 为了进行细节的对比，在图像下方可视化了局部细节放大后的结果。从图中可以看出，在相同的计算资源条件下，在频域内做图像转换能够生成细节信息更为清晰的有雾图像，对后续的目标检测任务更为有利，而 RGB 颜色空间的降采样操作造成了明显的信息损失，存在图像局部信息模糊的情况，见图 (c) 下方的局部细节放大图。

3.3 目标检测的领域自适应实验结果

文中将提出的频域内领域自适应方法与具有代表性的三种领域自适应方法[15−17] 作了比较，并以 IoU 为 0.5 报告了物体平均精确率的均值 (mean Average Precision, mAP)，结果如表 1 所示。

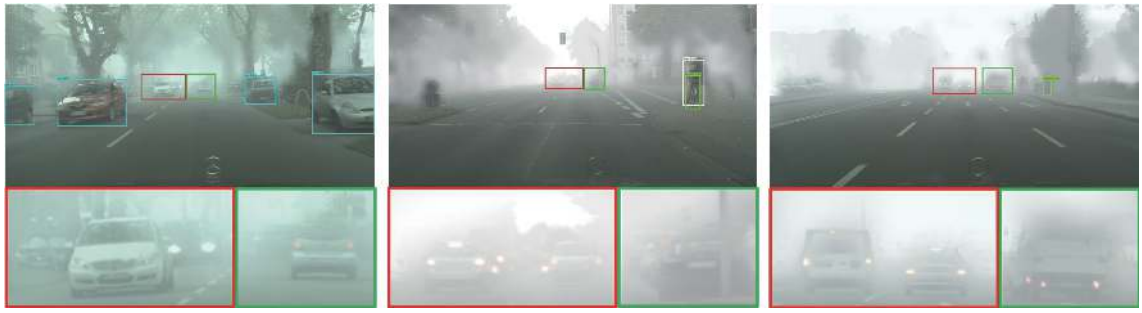
在所有对比方法中，训练集由 Cityscapes 中有检测标注的训练图像 (晴朗天气) 以及没有标注的 Foggy Cityscapes 中的训练图像 (雾霾天气) 构成。

表 1 Cityscapes→Foggy Cityscapes 不同领域自适应算法目标检测结果对比

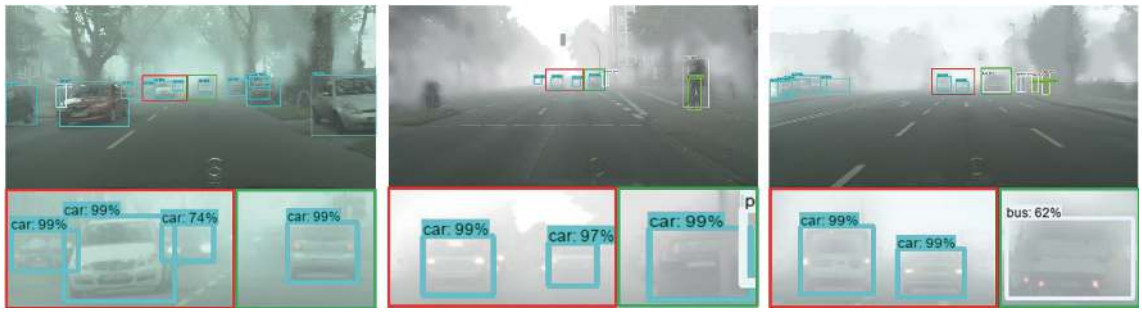
Tab.1 Object detection results of different domain adaptation algorithms on Cityscapes → Foggy Cityscapes datasets

Method	Bus	Bicycle	Car	Motor	Person	Rider	Train	Truck	mAP(@.5)
Cityscapes only	31.3%	33.8%	47.7%	20.2%	34.9%	40.5%	12.5%	17.8%	29.8%
MDA[15]	41.8%	36.5%	44.8%	30.5%	33.2%	44.2%	28.7%	28.2%	36.0%
PDA[16]	44.4%	35.9%	54.4%	29.1%	36.0%	45.5%	25.8%	24.3%	36.9%
CFF[17]	43.2%	37.4%	52.1%	34.7%	34.0%	46.9%	29.9%	30.8%	38.6%
Proposed algorithms	48.1%	42.7%	61.9%	32.1%	43.1%	49.1%	17.7%	25.4%	39.9%

测试图像均来源于 Foggy Cityscapes 提供的验证集 (雾霾天气)。表 1 中，MDA[15](Multi-level Domain Adaptation)、PDA[16](Progressive Domain Adaptation)，CFF[17](Coarse-to-Fine Feature adaptation) 是对比的 3 种领域自适应算法，数据引自参考文献 [17]。“Cityscapes Only”表示仅用源域图像图像训练，在有雾的测试集上进行测试的结果，检测结果如图 5 所示，仅用源域图像训练难以检测出雾霾中的目标，mAP 仅为 29.8%，证实了源域和目标域之间的差异。与仅用有标注的无雾图训练相比，文中提出的算法由于采用了两阶段的领域自适应方法，利用频域能量集中的特性，提高了输入特征的信息利用率，避免了降采样带来的信息损失，将 mAP 值由 29.8% 提升到 39.9%，mAP 值提高了 33.9% 左右，在 4 种对比算法中排名第一。证明了这两种策略能有效降低不同域之间差异，提高目标检测任务泛化性能。



(a) 仅 Cityscapes 训练无法检测出雾霾中的目标
(a) Detection result of "Cityscapes Only", which fails to detect objects in haze



(b) 文中所提算法的目标检测性能
(b) Detection result of the proposed algorithm

图 5 仅用 Cityscapes 训练与文中算法目标检测结果对比

Fig.5 Comparison of detection results between "Cityscapes Only"and the proposed algorithm

为了评价文中提出算法两个阶段的有效性，文中采用消融实验的方式，分别移除领域自适应阶段和无监督图像转换阶段，并评价了单个阶段目标检测的效果，结果如表 2 所示。从表中可以看出，与仅用 Cityscapes 训练相比，无监督转换方式和领域自适应方式 mAP 均有所提高，但都小于完整的两阶段算法，说明文中算法两个阶段的有效性和必要性，能够显著增强模型在无标注领域的泛化能力。

表 2 文中算法两个阶段的消融实验

Tab.2 Results of the ablation experiments corresponding to the two stages of the proposed algorithm

Algorithm	Bus	Bicycle	Car	Motor	Person	Rider	Train	Truck	mAP(@.5)
Cityscapes only	31.3%	33.8%	47.7%	20.2%	34.9%	40.5%	12.5%	17.8%	29.8%
Ours *w/o* stage 2	39.3%	38.5%	63.3%	28.0%	39.6%	42.4%	15.7%	23.6%	36.3%
Ours *w/o* stage 1	41.3%	39.0%	58.4%	28.6%	42.4%	44.7%	10.7%	23.6%	36.1%
Full model	48.1%	42.7%	61.9%	32.1%	43.1%	49.1%	17.7%	25.4%	39.9%

4 结 论

为提高目标检测的泛化性能，针对测试和训练数据分布不一致的问题，文中提出了一种频域内面向目标检测的领域自适应方法。通过频域内的无监督图像转换生成高分辨率图像，为测试集所在的域作数据扩充。算法同时采用基于对抗的领域自适应方法，进一步对齐扩充的数据和测试集数据的特征，减少了训练数据和测试数据之间的领域差异。实验结果表明，与空域的领域自适应和图像无监督转换方法相比，文中提出的方法在图像转换过程中能够生成清晰度和分辨率更高的图像。同时，利用频域的能量集中特性，能保留更多的原始图像信息，减少了由天气造成的领域差异，对交通监控等开放式目标检测的性能有着明显的提升效果。与仅用晴天图像训练的检测模型相比，领域自适应可将 mAP 值提升 33.9%。

参考文献：

[1] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 2223-2232.

[2] Fan L, Zhao H, Hu H, et al. Survey of target detection based on deep convolutional neural networks [J]. ***Optics and Precision Engineering***, 2020, 28(5): 1152-1164. (in Chinese)

[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[4] Girshick R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[5] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems, 2015, 28: 91-99.

[6] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]//European Conference on Computer Vision, 2016: 21-37.

[7] Redmon J, Farhadi A. YOLOv3: An incremental improvement [J]. ***ArXiv Preprint***, 2018, ArXiv: 1804.02767.

[8] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980-2988.

[9] Wu T, Zhang Z, Liu Y, et al. A lightweight small object detection algorithm based on improved SSD [J]. ***Infrared and Laser Engineering***, 2018, 47(7): 0703005. (in Chinese)

[10] Di X, Lin Z, Chen S. Dim moving object detection based on projection into the 2D frequency domain [J]. ***Infrared and Laser Engineering***, 2013, 42(12): 3447-3452. (in Chinese)

[11] Wu Y, Wang Y, Sun H, et al. LSS-target detection in complex sky backgrounds [J]. ***Chinese Optics***, 2019, 12(4): 853-865. (in Chinese)

[12] Gong X, Ouyang H. Improvement of tiny YOLOV3 target detection [J]. ***Optics and Precision Engineering***, 2020, 28(4): 988-995. (in Chinese)

[13] Wang C, An J, Jiang X, et al. Region proposal optimization algorithm based on convolutional neural networks [J]. ***Chinese Optics***, 2019, 12(6): 1348-1361. (in Chinese)

[14] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation[C]//International Conference on Machine Learning, PMLR, 2015: 1180-1189.

[15] Xie R, Yu F, Wang J, et al. Multi-level domain adaptive learning for cross-domain detection[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.

[16] Hsu H K, Yao C H, Tsai Y H, et al. Progressive domain adaptation for object detection[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2020: 749-757.

[17] Zheng Y, Huang D, Liu S, et al. Cross-domain object detection through coarse-to-fine feature adaptation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 13766-13775.

[18] Li H, Wan R, Wang S, et al. Unsupervised domain adaptation in the wild via disentangling representation learning [J]. ***International Journal of Computer Vision***, 2021, 129(2): 267-283.

[19] Liu M Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks[C]//Advances in Neural Information Processing Systems, 2017: 700-708.

[20] Xu K, Qin M, Sun F, et al. Learning in the frequency domain[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 1740-1749.

[21] Yang Y, Soatto S. FDA: Fourier domain adaptation for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 4085-4095.

[22] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[23] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[24] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library[C]//Advances in Neural Information Processing Systems, 2019, 32: 8026-8037.