



## 一种基于对抗学习的实时跟踪模型设计

李志鹏<sup>1</sup>, 张 睿<sup>2</sup>

(1. 复旦大学 软件学院, 上海 201203; 2. 复旦大学 计算机实验教学中心, 上海 201203)

**摘 要:** 目标跟踪指在视频帧中找到感兴趣目标的运动位置, 广泛应用于环境感知、安防监控和无人驾驶等领域。为进行高效的目标跟踪, 建立一种基于对抗学习和特征压缩的相关滤波器目标跟踪模型。为了同时兼顾精度与速度, 在模型中引入特征提取优化、特征压缩和特征聚合等步骤。在提取图像特征前, 采用对抗学习方法解决特征提取模型中训练数据与任务数据分布不匹配的问题。在特征压缩阶段, 应用双通道自编码器结构和特征聚合来增强模型对图像风格的泛化能力。实验结果表明, 与非实时跟踪算法相比, 该模型在精度损失不超过3%的情况下能获得明显的速度提升, 其跟踪速度高达103FPS。

**关键词:** 目标跟踪; 对抗学习; 自编码器; 相关滤波器; 表示学习

开放科学(资源服务)标志码(OSID):



中文引用格式: 李志鹏, 张睿. 一种基于对抗学习的实时跟踪模型设计[J]. 计算机工程, 2021, 47(6): 262-270.

英文引用格式: LI Zhipeng, ZHANG Rui. Design of a real-time tracking model based on adversarial learning[J]. Computer Engineering, 2021, 47(6): 262-270.

## Design of a Real-Time Tracking Model Based on Adversarial Learning

LI Zhipeng<sup>1</sup>, ZHANG Rui<sup>2</sup>

(1. School of Software, Fudan University, Shanghai 201203, China;

2. Computer Teaching and Experiment Center, Fudan University, Shanghai 201203, China)

**[Abstract]** Target tracking is the process of finding the position of the moving target in the video frame, and is widely used in the fields of environmental perception, security monitoring and unmanned driving. To improve the efficiency of target tracking, this paper proposes a target tracking model using correlation filter based on adversarial learning and feature compression. The model includes the optimization of feature extraction, feature compression and feature aggregation, which improve both the accuracy and the speed of the model. Before feature extraction, adversarial learning is used to solve the problem of the mismatch of training data and task data distribution in the feature extraction model. In the stage of feature compression, a two-way autoencoder structure and feature aggregation are used to enhance the generalization of the image style. Experimental results show that compared with the non-real-time tracking algorithm, the model significantly improves the speed, with the tracking rate reaching 103FPS. At the same time, the loss of the model accuracy is within 3%.

**[Key words]** target tracking; adversarial learning; autoencoder; correlation filter; representation learning

DOI: 10.19678/j.issn.1000-3428.0057920

### 0 概述

目标跟踪是指在视频的每一帧中找到感兴趣目标的运动位置, 其被广泛应用于人机交互、视频监控和自动驾驶等领域, 也是计算机视觉的一个重要分支。在设计目标跟踪算法时, 需要同时考虑复杂环境中的跟踪精度以及跟踪速度, 以满足实时性需求。随着深度学习研究的深入, 卷积神经网络(Convolutional Neural

Network, CNN)大幅提升了跟踪算法的性能和稳定性<sup>[1]</sup>, 但是计算复杂度的提升使得单一CNN算法难以满足实时性需求<sup>[2]</sup>。基于相关滤波器的目标跟踪算法一般具有较高的计算效率, 现有实时跟踪算法大多使用基于图像卷积特征的相关滤波器<sup>[3]</sup>, 但是此类算法保持高跟踪速度的同时难以保证高跟踪精度。随着实际应用中精度需求的不断提升, 相关滤波器输入维数的增长也会限制算法的跟踪速度<sup>[4]</sup>。因此, 设计同时具

基金项目: 教育部专项基金“大数据驱动的临床辅助诊断系统研究”(2018A11005)。

作者简介: 李志鹏(1993—), 男, 硕士研究生, 主研方向为图像处理、机器学习、嵌入式开发; 张睿, 高级工程师。

收稿日期: 2020-03-31 修回日期: 2020-05-11 E-mail: 17212010018@fudan.edu.cn

有高精度和高速度的目标跟踪模型具有重要意义。

在现有的跟踪算法中,研究人员应用迁移学习的思想<sup>[5]</sup>,使用VGG-NET<sup>[6]</sup>等预训练卷积网络提取图像特征,大幅提高了算法性能。但是直接应用VGG-NET并非最优方案,原因是对于特定跟踪场景,待处理的图像分布通常与VGG-NET训练数据的分布不同,这类分布差异问题会影响图像特征的可靠性。为了解决这一问题,本文采用对抗学习方法<sup>[7]</sup>,在不需要跟踪场景图像标签的情况下解决分布差异问题,使模型在特定任务中表现更稳定。另外,由于特征维数是限制模型计算效率的主要因素,本文利用自编码器将图像特征压缩到低维空间<sup>[8]</sup>。在实际应用中,模型要处理包含多类目标和环境风格的图像,因此,本文设计一种双通道自编码器结构并在训练时优化判别损失函数,以提高模型的泛化能力。在特征压缩后使用相关滤波器进行目标跟踪。具体地,本文提出一种应用深度压缩特征的相关滤波器实时跟踪模型(TDFC),使用对抗学习和特征压缩提高跟踪算法的精度和速度<sup>[9]</sup>。应用对抗学习方法优化图像特征提取过程,使得跟踪场景数据与特征提取模型VGG-NET的预训练ImageNet<sup>[10]</sup>数据分布一致,得到针对任务场景的优化图像卷积特征。在此基础上,提出一种基于自编码器的双通道结构模型,将图像特征压缩到低维空间,并通过类别标签信息优化模型的训练过程。

## 1 研究背景

现有的目标跟踪算法主要分为两类:

1)第一类是基于相关滤波器的算法,此类算法一般具有较高的计算效率。文献[11]提出的KCF算法利用快速傅里叶变换和循环矩阵降低算法的时间复杂度。SAMF<sup>[12]</sup>在KCF的基础上,结合HOG特征和CN特征,并对目标尺度变化进行检测。文献[13]与SAMF类似,除了聚合多种特征外,还提出一种三维滤波器结构,实现对目标尺度的自适应。STC<sup>[14]</sup>在贝叶斯框架下对跟踪目标及其上下文的时空关系建模,得到目标和周围区域的统计相关性。但是上述相关滤波器跟踪算法会在跟踪精度方面存在局限性。为了解决这一问题,HCF<sup>[15]</sup>算法在KCF的基础上,将HOG特征替换为分层卷积特征,在不同层训练相关滤波器以提高跟踪精度。DeepSRDCF<sup>[16]</sup>在使用卷积特征的同时还加入惩罚项以改善边界的影响。这些改进算法通过使用图像卷积特征使跟踪精度得到提升,但是图像卷积特征的通道数远多于原始图像的通道数,这会导致计算复杂度提升以及跟踪速度降低。

2)第二类是基于神经网络的算法。文献[17]提出的SO-DLT算法使用CNN作为获取特征和分类结果的模型,先在非跟踪数据上进行离线预训练,然后

基于跟踪数据调整参数,以解决跟踪过程中数据不足的问题。文献[18]在SO-DLT思想的基础上,在多域学习模型中使用带标注的视频数据。文献[19]用小型卷积网络将多层卷积特征稀疏化,得到用于跟踪的判别式特征。文献[20]使用循环神经网络对目标物体建模,提升模型鉴别相似物体的能力。文献[21]通过强化学习来学习物体的连续性动作,从而检测目标变化。虽然上述基于神经网络的算法能达到很高的跟踪精度,但随着网络复杂度的提高,跟踪速度不可避免地会有所下降,难以满足实时性的需求。

## 2 模型和算法

本文提出的模型主要包含特征提取、特征压缩、目标跟踪等步骤。在特征提取步骤中,首先使用对抗学习方法调整跟踪场景图像的分布,然后使用预训练的VGG-NET提取图像的卷积特征。在特征压缩步骤中,采用基于自编码器的双通道网络结构,结合类别信息优化训练过程以降低图像特征维度,再对双通道特征执行聚合操作得到压缩特征。最后,将压缩特征作为相关滤波器的输入以实现目标跟踪。

### 2.1 特征提取优化

在使用VGG-NET提取图像特征时,预训练数据与跟踪场景任务域数据存在分布不一致的问题,该问题会影响图像特征的有效性。因此,本文采用对抗学习方法优化特征提取过程,从而解决该问题。

#### 2.1.1 优化方法的理论支持

为了解决VGG-NET预训练数据(ImageNet图像)与跟踪场景任务域数据分布不一致的问题,需要对齐跟踪场景图像和ImageNet图像的分布,即降低VGG-NET在跟踪场景数据上的迁移学习误差。根据文献[22]中关于迁移学习模型误差分析的理论,对于 $\forall h \in H$ ,迁移学习模型的期望误差 $E_T(h)$ 有如下性质:

$$E_T(h) \leq E_S(h) + \frac{1}{2} d_{H\Delta H}(S, T) + \lambda \quad (1)$$

其中, $H$ 表示假设函数族(亦可理解为模型族), $h$ 是 $H$ 的一个实例, $E_S(h)$ 表示迁移学习模型在源域(即ImageNet图像服从的分布)数据集上的误差项, $E_T(h)$ 表示迁移学习模型在目标域(即跟踪场景图像服从的分布)数据集上的误差项, $d_{H\Delta H}$ 是衡量一对分类器之间差异的项, $\lambda$ 表示一对分类器间的共有误差。在式(1)右侧的各项中, $E_S(h)$ 为模型在源域数据上的期望误差,是实验中得到的常量, $d_{H\Delta H}$ 是变量, $\lambda$ 是常数。因此,为了降低VGG-NET模型在跟踪场景图像上的期望误差(即式(1)左侧的 $E_T(h)$ )的大小,只需对式(1)右侧的 $d_{H\Delta H}$ 项进行分析, $d_{H\Delta H}$ 定义如下:

$$d_{H\Delta H} = 2 \sup_{(h, h') \in H^2} \left| E \left[ I \left( h(x_s) \neq h'(x_s) \right) \right] - E \left[ I \left( h(x_T) \neq h'(x_T) \right) \right] \right| \quad (2)$$

其中,  $\sup$  表示最小上界,  $x_s$  表示源域数据,  $x_T$  表示目标域数据,  $E$  为求期望运算,  $I$  为布尔函数, 其参数为真时输出 1, 否则输出 0,  $x$  为输入样本,  $S$  表示源域 (即 ImageNet 图像服从的分布),  $T$  表示目标域 (即跟踪场景图像服从的分布),  $h$  和  $h'$  表示不同的假设函数实例, 等价于结构相同但参数不同的分类器模型,  $h(\cdot)$  表示模型输出。

在实际应用中, 经过训练的分类器  $h$  和  $h'$  在源域数据上具有较高的分类精度, 即对于  $\forall x \sim S$ ,  $h$  和  $h'$  关于输入  $x$  的预测值接近标签值,  $h$  和  $h'$  关于源域数据的输出总体趋于一致, 故式(2)中的  $E \left[ I \left( h(x_s) \neq h'(x_s) \right) \right]$  是接近于 0 的较小数值, 可从式(2)中移除。因此, 式(2)可进一步简化为:

$$d_{H\Delta H} = 2 \sup_{(h, h') \in H^2} E \left[ I \left( h(x_T) \neq h'(x_T) \right) \right] \quad (3)$$

式(3)右边可理解为结构相同但参数不同的分类器  $h$  和  $h'$  对目标域数据  $\forall x \sim T$  的预估值分布差导的期望的最小上界。通过上述过程, 可将式(1)中误差项  $E_T(h)$  的优化问题转化为式(3)中  $d_{H\Delta H}$  的优化问题。

将式(3)中的最小上界改写为  $\min \max$  形式, 得到式(4):

$$d_{H\Delta H} = \min_{(h, h')} \max_{(h, h')} E \left[ I \left( h(x_T) \neq h'(x_T) \right) \right] \quad (4)$$

引入本文设计的对抗学习模型结构后, 将式(4)改写为:

$$d_{H\Delta H} = \min_G \max_{(D_1, D_2)} E \left[ I \left( D_1 \circ G(x_T) \neq D_2 \circ G(x_T) \right) \right] \quad (5)$$

其中,  $D_1$  和  $D_2$  表示结构相同但参数不同的判别器,  $G$  表示生成器。式(4)中的  $x$  被修正为原始样本经生成器处理后的输出  $G(x)$ ,  $h$  被具体化为判别器  $D_1$ ,  $h'$  被具体化为判别器  $D_2$ 。

按照式(5)给出的目标进行优化, 可以使  $d_{H\Delta H}$  趋近最低值, 结合前文对式(1)右侧各项的分析, 这一优化操作可以限制迁移学习模型在目标域数据上期望误差  $E_T(h)$  的上界, 从而有利于模型在目标域上的误差趋近最低值。更重要的是, 训练后的生成器  $G$  可以将跟踪场景图像映射到 ImageNet 图像服从的分布上。

## 2.1.2 优化模型设计

本节根据 2.1.1 节中的理论推导设计优化方案。

训练数据包含 ImageNet 的部分标签数据 (记作  $\{X_s, Y_s\}$ ) 以及从跟踪场景采集的无标签图像 (记作  $\{X_T\}$ )。

按照 2.1.1 节的理论分析, 与特征提取优化相关的对抗学习模型应当按照式(5)进行设计和优化, 式(5)将式(4)中的模型输出项 ( $h(x)$  和  $h'(x)$ ) 具体化为生成器和判别器的联合输出项 ( $D_1 \circ G(x)$  和  $D_2 \circ G(x)$ ), 其中,  $G$ 、 $D_1$ 、 $D_2$  三项为对抗学习模型包含的主要结构,  $G$  是生成器,  $D_1$  和  $D_2$  是判别器, 由于式(4)中的  $h$  和  $h'$  来自同一假设函数族  $H$ , 因此  $D_1$  和  $D_2$  是一对结构相同但参数不同的判别器。基于对抗学习的网络结构如图 1 所示, 生成器  $G$  的输入是从  $\{X_s, Y_s\}$  和  $\{X_T\}$  中采样的图像, 分别记作  $(x_s, y_s)$  和  $x_T$ 。判别器  $D_1$  和  $D_2$  以生成器  $G$  的输出为输入, 是两个结构相同但参数不同的图像分类器, 由卷积层和全连接层组成,  $P_1$  和  $P_2$  分别是判别器  $D_1$  和  $D_2$  输出的预测向量, 用于计算损失函数。

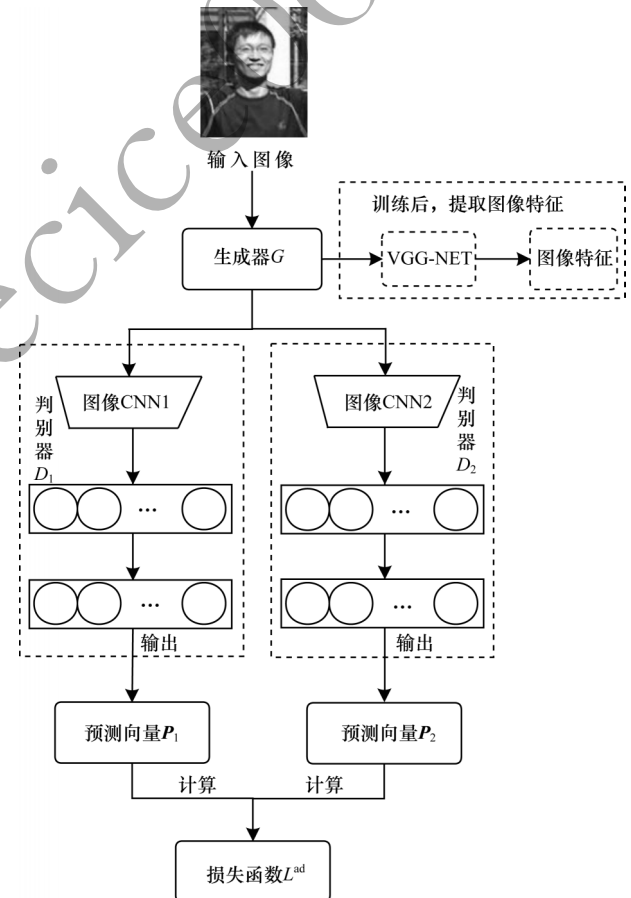


图1 基于对抗学习的网络结构

Fig.1 Network structure based on adversarial learning



网络优化目标是降低 VGG-NET 在任务场景数据上的误差,提高图像特征的有效性。为达到此目标,需要利用两个判别器间的差异信息,式(5)中的布尔函数项  $I(D_1 \circ G(x) \neq D_2 \circ G(x))$  指定了判别器差异的衡量方式,即判别器  $D_1$  和  $D_2$  的输出向量差异。在保证整体模型在 ImageNet 数据集上具有高分类精度的前提下,最大化判别器  $D_1$  和  $D_2$  的预测差异(式(5)中的  $\max$ );对于生成器  $G$ ,其输出应使判别器  $D_1$  和  $D_2$  的输出一致(式(5)中的  $\min$ )。对生成器参数和判别器参数进行交替优化,即可优化式(5)中的  $d_{H\Delta H}$  项,使得生成器  $G$  能够将任务场景图像映射到 ImageNet 数据服从的分布上。

### 2.1.3 优化步骤

具体优化步骤如下:

**步骤1** 以不同的参数初始化判别器  $D_1$  和  $D_2$ 。根据 2.1.2 节的讨论,首先需要保证整体模型在 ImageNet 数据集上具有高分类精度,因此,第一步的优化目标是 minimize 一个交叉熵损失函数,如式(6)所示:

$$L_1^{\text{ad}}(\mathbf{x}_s, \mathbf{y}_s) = -\frac{1}{n} \sum_{i=1}^n (y_{s,i} \log_a \mathbf{P}_{1,i} + y_{s,i} \log_a \mathbf{P}_{2,i}) \quad (6)$$

其中,  $(\mathbf{x}_s, \mathbf{y}_s)$  表示带标签的 ImageNet 样本,  $n$  为类别数量,  $\mathbf{P}_1$  和  $\mathbf{P}_2$  分别是  $D_1$  和  $D_2$  的输出向量,  $i$  是向量索引。

**步骤2** 固定生成器  $G$ , 调整判别器  $D_1$  和  $D_2$  的参数。在这一步中,结合式(5)及相关讨论,优化目标是对于相同的输入数据,  $D_1$  和  $D_2$  输出的预测向量,即  $\mathbf{P}_1$  和  $\mathbf{P}_2$  之间的差异尽可能大。为了衡量这一差异,定义包含距离度量的损失函数,分别如式(7)和式(8)所示:

$$L_2^{\text{ad}}(\mathbf{x}_s, \mathbf{y}_s) = -\frac{1}{n} \sum_{i=1}^n (y_{s,i} \log_a \mathbf{P}_{1,i} + y_{s,i} \log_a \mathbf{P}_{2,i}) \quad (7)$$

$$L_3^{\text{ad}}(\mathbf{x}_T) = -\frac{1}{n} \sum_{i=1}^n (\mathbf{P}_{1,i} - \mathbf{P}_{2,i})_1 \quad (8)$$

其中,  $|\cdot|_1$  表示 L1 范数。式(7)为分类损失项,适用于带标签的 ImageNet 数据,由于模型优化依赖的式(5)的推导前提是模型实例  $h$  和  $h'$  在源域数据上具有高分类精度,因此需要式(7)来提升模型对 ImageNet 样本的分类精度。式(8)为距离度量项,适用于无标签的任务场景图像。式(7)和式(8)共同构成本步骤的优化目标。

**步骤3** 固定判别器  $D_1$  和  $D_2$ , 调整生成器  $G$  的参数。在这一步中,优化目标是输入图像经  $G$  处理后,  $D_1$  和  $D_2$  的输出尽可能一致,损失函数如式(9)所示:

$$L_4^{\text{ad}}(\mathbf{x}_T) = \frac{1}{n} \sum_{i=1}^n (\mathbf{P}_{1,i} - \mathbf{P}_{2,i})_1 \quad (9)$$

参考文献[23]中对抗生成网络的训练过程,特征提取优化模型具体的训练方法如下:

1) 在 ImageNet 数据集上训练生成器  $G$  和判别器  $D_1, D_2$  (即上文步骤1)。

2) 调整判别器  $D_1$  和  $D_2$ , 最大化预测差异 (即上文步骤2)。

3) 调整生成器  $G$ , 最小化预测差异 (即上文步骤3)。

4) 重复上述第2步和第3步,直至达到最大迭代次数。

随着损失函数的收敛,经过调整的生成器  $G$  可以将任务场景图像映射到 ImageNet 数据服从的分布上,从而解决数据分布不匹配的问题。

经过本文设计的优化方案,对于给定的跟踪场景图像,使用生成器  $G$  对其处理后,将  $G$  的输出作为 VGG-NET 的输入,即可得到更有效的图像卷积特征。

### 2.1.4 特征提取模型实现细节

本文生成器  $G$  采用全卷积结构,判别器  $D_1, D_2$  是包含一组卷积层和两个全连接层的图像分类器,每一个卷积层中卷积核空间上的大小为  $3 \times 3$ ,全连接层的神经元数量分别为 2 048 和 1 024。取 VGG-NET 的第二层 feature map 作为图像特征,因此,输出维度是  $224 \times 224 \times 64$ 。模型训练时批量大小为 16,采用 ADAM<sup>[24]</sup> 优化算法来优化损失函数,学习率设置为  $10^{-4}$ ,其他超参数取文献[24]中的默认值。

## 2.2 双通道自编码器

自编码器是一种无监督特征压缩方法,通过优化重构成本,能在保留重要信息的同时去除数据中的冗余。为了实现高压压缩率,自编码器包含多个隐藏层,第  $i$  个编码器层的计算方式为:  $R^{w \cdot h \cdot c_i} \rightarrow R^{w \cdot h \cdot c_{i+1}}$ ,从而将图像卷积特征的通道数逐层降低。本文方法采用双通道自编码器结构和特征聚合操作来提高模型的泛化能力。

双通道自编码器的模型结构如图2所示。模型以端到端的方式进行训练,输入是 VGG-NET 输出的图像卷积特征。编码器1和解码器1组成一个卷积自编码器,每一个卷积层中卷积核空间上的大小是  $3 \times 3$ ,激活函数是 relu 函数。编码器2和解码器2组成一个去噪自编码器,样本输入到去噪自编码器之前,按照一定概率对其加入噪声,噪声分三种:第一种是随机将图像特征中的值置为0或高斯噪声,这类似于信息传输中的随机扰乱;第二种是随机选取特征中的一个通道,将该通道上的值置为0;第三种是随机调换特征中两个区域位置下所有通道的值,由于卷积层平移等变的性质,这相当于在原始图像中执行区域互换,可以修正位置因素对特征的影响。在上述过程中,每个样本至多被加入一种噪声。

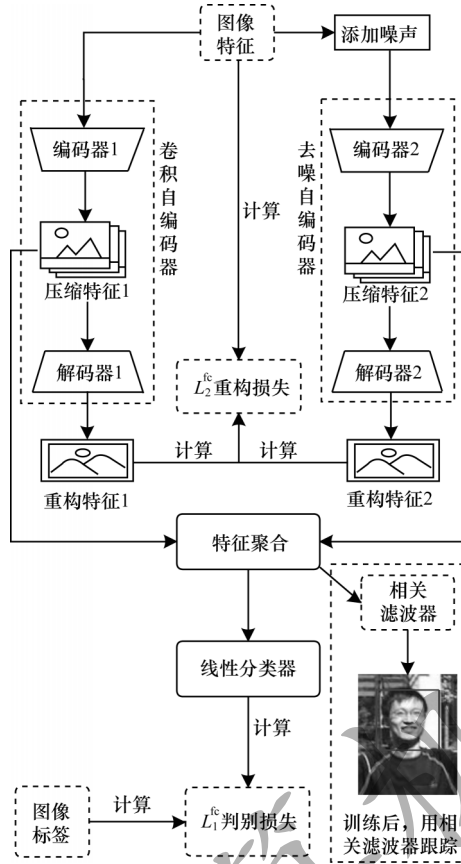


图2 双通道自编码器网络结构

Fig.2 Network structure of dual channel autoencoder

深度自编码器在应用时容易出现过拟合问题,本文方法采取两个措施来防止过拟合:

1)第一个措施源自压缩特征应保留足够判别式信息的思想,即压缩后的特征可通过线性分类器预测其原有的类别标签。具体的实现方式如图2所示,“压缩特征1”和“压缩特征2”经“特征聚合”(聚合函数的具体形式在下文介绍)操作后输入到一个“线性分类器”中,该“线性分类器”预测输入的类别标签,然后根据预测结果计算判别损失,损失函数如式(10)所示:

$$L_1^c(c_s, y_s) = \frac{1}{2} \|w^T c_s - y_s\|_2^2 \quad (10)$$

其中,  $c_s$  表示带标签 ImageNet 样本  $x_s$  的压缩特征,  $y_s$  是  $x_s$  的标签,  $w$  是线性分类器的参数矩阵,  $\|\cdot\|_2^2$  表示平方 L2 范数。

2)第二个措施是引入多级重构误差函数,除了考虑完整自编码器中输入和输出间的重构误差,还考虑自编码器子结构的重构误差,损失函数如式(11)所示:

$$L_2^c(x) = \frac{1}{4} \sum_{i=1}^2 \sum_{j=1}^{l_i} [(x - AE_j^{(i)}(x))^2] \quad (11)$$

其中,  $x$  表示 ImageNet 样本特征或跟踪场景样本特征,  $i$  是双通道分支索引,  $AE^{(i)}$  指代图2结构中的两个自编码器之一,  $l_i$  表示相应自编码器中编码器结构的层数,

由于自编码器结构对称的特点,解码器的层数也是  $l_i$ ,  $AE_j$  表示自编码器的子结构,即  $AE_1$  表示保留编码器的第一个隐藏层、解码器的最后一个隐藏层而组成的子结构,  $AE_2$  表示保留编码器的前两个隐藏层、解码器的最后两个隐藏层而组成的子结构,  $AE_j$  表示保留编码器的前  $j$  个隐藏层、解码器的最后  $j$  个隐藏层而组成的子结构,  $AE_l$  则表示完整的自编码器结构。

结合式(10)、式(11)可以得到特征压缩过程的整体优化目标:

$$L^c = L_1^c + \alpha L_2^c \quad (12)$$

其中,超参数  $\alpha$  衡量损失函数各部分的重要程度。

经过以上两个措施,能够有效防止模型出现过拟合问题。此外,训练过程还应用数据增强方法进一步防止过拟合问题,包括水平翻转、竖直翻转、色调变化、对比度变化等操作。

在图2中,双通道自编码器输出的降维特征还需要经过特征聚合操作,以提高模型的泛化能力。聚合函数中的参数能够与模型其他部分一同进行端到端的训练,其具体形式如下:

1)线性聚合,这种聚合方式假设每一个自编码器起到同等作用,如式(13)所示:

$$c = \frac{1}{2} \sum_{i=1}^2 m^{(i)} f^{(i)} \quad (13)$$

其中,  $c$  是聚合后的压缩特征,  $i$  是双通道分支索引,  $f$  是对应分支的降维特征,  $m$  是变换矩阵。

2)加权线性聚合,简单的线性聚合不能反映每个特征的重要性差异,因此,加权线性聚合方式为每一种特征赋予一定的权重,如式(14)所示:

$$c = \frac{1}{2} \sum_{i=1}^2 w^{(i)} (m^{(i)} f^{(i)}) \quad (14)$$

其中,  $w$  是权重向量。

3)加权非线性聚合,这种聚合方式引入非线性变换,提高聚合层的表达能力,从而建模更复杂的统计相关性,如式(15)所示:

$$c = \frac{1}{2} \sum_{i=1}^2 w^{(i)} \sigma(m^{(i)} f^{(i)}) \quad (15)$$

其中,  $\sigma$  表示 sigmoid 函数。

对于上述三种聚合方式,线性聚合的参数量最少,因此,其聚合能力最弱;加权线性聚合和加权非线性聚合对模型总体性能的影响差别不大,这是因为先前的特征提取优化和特征压缩都是非线性过程,所以加权非线性聚合中  $\sigma$  函数的作用不是特别明显。

本文卷积自编码器的层数为4层,去噪自编码器为6层。特征样本输入到去噪自编码器前,被添加噪声的几率为30%。特征融合方式选择线性加权

聚合。训练时的批量大小为16,采用ADAM优化算子,学习率设置为 $2 \times 10^{-6}$ ,算子超参数取文献[24]中的默认值。

基于本节设计的模型结构和优化方法,可以实现图像特征的高效压缩,提高算法的计算效率。

### 2.3 相关滤波器

傅里叶域循环矩阵的性质使得经过训练的相关滤波器能以较低的计算开销完成目标跟踪任务。在本文模型中,为了对场景中的目标进行跟踪,需要将2.2节得到的压缩特征输入相关滤波器,以得到跟踪结果。相关滤波器的参数可表示为:

$$\mathbf{w} = \mathbf{F}^{-1} \left( \frac{\mathbf{c}' \odot \mathbf{r}'}{\mathbf{c}' \odot \mathbf{c}'_{*} + \lambda} \right) \quad (16)$$

其中, $\mathbf{w}$ 是相关滤波器参数, $\mathbf{F}^{-1}$ 是逆傅里叶变换, $\mathbf{c}$ 为图像特征, $\mathbf{r}$ 为响应窗口, $\mathbf{c}'$ 和 $\mathbf{r}'$ 为相应向量经过傅里叶变换的结果, $\mathbf{c}'_{*}$ 为共轭向量, $\lambda$ 为常数。

按照式(16)更新相关滤波器后,给定待跟踪的图像特征,计算得到的响应窗口如式(17)所示:

$$\mathbf{r} = \mathbf{F}^{-1} (\mathbf{w}' \odot \mathbf{c}'_{\text{new},*}) \quad (17)$$

其中, $\mathbf{c}_{\text{new}}$ 是待跟踪图像特征, $\mathbf{c}'_{\text{new},*}$ 是共轭向量, $\mathbf{w}'$ 是 $\mathbf{w}$ 经傅里叶变换的结果。

通过响应窗口 $\mathbf{r}$ 即可得到目标跟踪结果。相关滤波器是本文算法在功能实现时的重要一环,相关滤波器的原理和其他细节本文不再赘述。

## 3 实验结果与分析

### 3.1 训练集和评估方式

本文模型的训练数据来自ImageNet<sup>[10]</sup>和OTB-100数据集<sup>[25]</sup>。为了评估算法的性能,在实验中统计算法的跟踪精度和跟踪速度(FPS)信息。实验中算法的跟踪精度基于“Location error threshold”和“Overlap threshold”计算而得。其中,基于“Location error threshold”的跟踪精度指算法估计的目标位置与人工标注中心点间的距离小于给定阈值的视频帧所占的百分比;基于“Overlap threshold”的跟踪精度指算法估计的目标范围与人工标注框重叠比例大于给定阈值的视频帧所占的百分比。

实验中所使用的软硬件平台设置:硬件环境为Intel i7-7700K CPU @ 4.20 GHz, 16 GB 内存, NVIDIA GTX1080Ti GPU;软件环境为Python3.6, keras, tensorflow, Matlab。

### 3.2 模型性能分析

在OTB-100数据集上验证本文跟踪算法及其他算法的效果,相关量化指标结果如表1和表2所示,表中记录的跟踪精度是基于“Location error threshold”为20像素所计算,默认的特征聚合方式为加权线性聚合。

表1 本文算法在不同配置下的性能比较

Table 1 Performance comparison of this algorithm under different configurations

算法-配置	跟踪精度/%	跟踪速度(FPS)
TDFC	89.1	103.1
TDFC-nolinear	88.5	103.0
TDFC-mean	87.2	103.1
TDFC-noAugmentation	85.9	101.9
TDFC-noLabel	82.1	102.7
TDFC-noEnsemble	79.9	103.2
TDFC-noLevel	77.5	102.5
TDFC-noGenerator	75.9	103.6

表2 不同跟踪算法的性能比较

Table 2 Performance comparison of different tracking algorithms

算法	跟踪精度/%	跟踪速度(FPS)
TDFC	89.1	103.1
SCT <sup>[26]</sup>	84.9	41.0
SiamFC <sup>[27]</sup>	79.8	79.0
DSST <sup>[13]</sup>	75.6	32.0
KCF <sup>[11]</sup>	74.5	115.5
ADNET <sup>[21]</sup>	91.7	2.8
SANET <sup>[20]</sup>	90.6	0.9
FCNT <sup>[19]</sup>	86.2	2.9
DeepSRDCF <sup>[16]</sup>	85.3	1.8

表1统计本文算法在不同配置下的跟踪精度和速度,TDFC是本文算法的最优配置,TDFC-nolinear将聚合函数改为加权非线性聚合,TDFC-mean将聚合函数改为线性聚合,TDFC-noAugmentation中移除了2.2节提到的数据增强处理,TDFC-noLabel中训练双通道自编码器时忽略了判别式损失 $L_1^{\text{fc}}$ ,仅考虑重构损失 $L_2^{\text{fc}}$ ,TDFC-noEnsemble中移除了双通道自编码器结构和特征聚合操作,仅使用两个分支中的去噪自编码器,TDFC-noLevel中训练双通道自编码器时将多级重构损失 $L_2^{\text{fc}}$ 简化为完整自编码器结构输入层和输出层间的重构损失,TDFC-noGenerator在提取图像卷积特征前移除了基于对抗学习的优化步骤。表2统计本文算法以及其他跟踪算法的跟踪性能,对比算法包括SCT<sup>[26]</sup>、SiamFC<sup>[27]</sup>、DSST<sup>[13]</sup>、KCF<sup>[11]</sup>、ADNET<sup>[21]</sup>、SANET<sup>[20]</sup>、FCNT<sup>[19]</sup>、DeepSRDCF<sup>[16]</sup>,以上算法均按照原参考文献设计实现。其中,SCT、SiamFC、DSST、KCF是实时跟踪算法,ADNET、SANET、FCNT、DeepSRDCF是非实时跟踪算法,从表2可以看出,与现有跟踪算法相比,本文算法的跟踪精度损失不超过3%。图3所示为本文算法在不同配置下的性能统计信息,其中,图3(a)的横轴统计量为“Location error threshold”,图3(b)的横轴统计量为“Overlap threshold”。



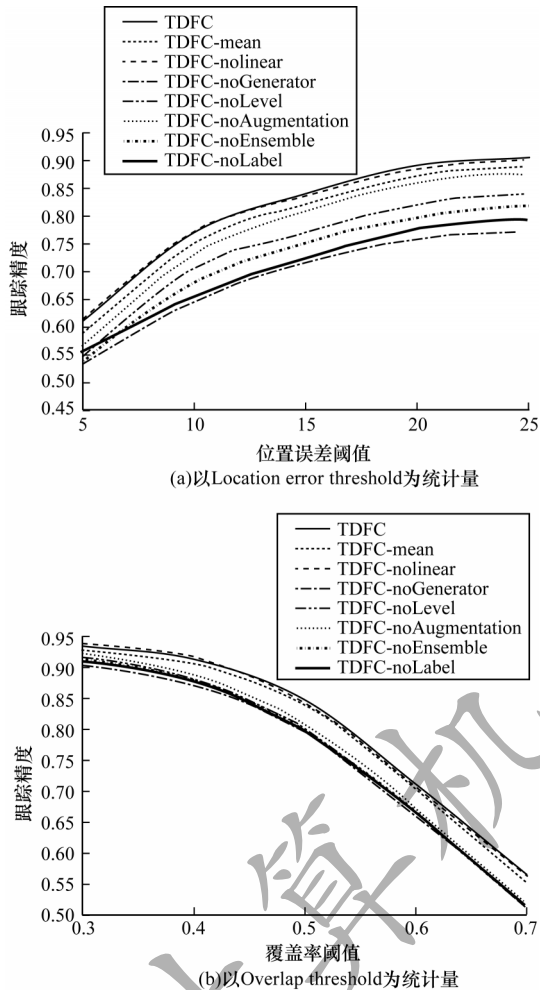


图3 本文算法在不同配置下的跟踪精度统计

Fig.3 The tracking accuracy statistics of the algorithm under different configurations

对比上述结果可以看出:聚合函数的选择对跟踪效果有一定影响,使用非线性聚合的TDFC-nonlinear由于 $\sigma$ 函数引起的梯度消失,模型更难以训练,其精度比使用加权线性聚合的TDFC略低;与使用线性聚合的TDFC-mean相比,TDFC有1.9%的跟踪精度提升,线性聚合的参数数量相对更少,不能充分反映模型间重要程度的差异;与TDFC-noAugmentation的对比说明应用数据增强方法给模型带来了3.2%的跟踪精度提升;与TDFC-noLabel相比,TDFC有7%的精度提升,说明通过保留判别式信息,能够有效防止自编码器出现过拟合,提高算法的性能;与TDFC-noEnsemble相比,TDFC的跟踪精度有9.2%的提升,表明双通道自编码器结构和特征聚合提升了模型的泛化性,得到的压缩特征有助于提高跟踪精度;与TDFC-noLevel相比,TDFC有11.6%的精度提升,说明2.2节定义的多级重构损失函数有助于改进自编码器模型的训练,提高最终的跟踪精度;与TDFC-noGenerator相比,TDFC有13.2%的精度提升,说明2.1节提出的迁移学习分布不匹配问题确实存在,基于对抗学习的优化方法能够解决此问题,有助于提高跟踪算法的精度。

图4所示为特征提取优化模型的收敛情况,可以看出,在距离损失逐渐降低的同时,判别器 $D_1$ 和 $D_2$ 对训练数据的平均分类精度逐渐升高,当迭代(epoch)次数达到45时,模型趋于收敛。以上数据和分析结果表明本文设计中的各步骤具有有效性。

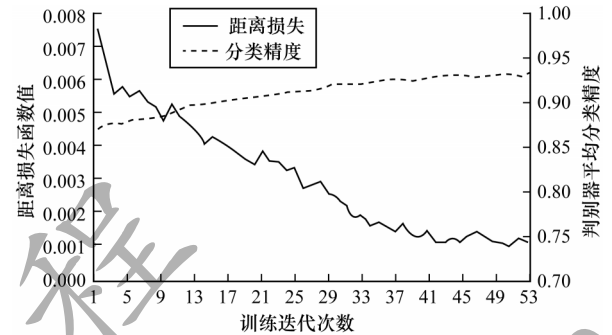


图4 距离损失函数的收敛情况

Fig.4 Convergence of distance loss function

为了进一步验证2.1节中对抗学习方法的有效性,用t-SNE<sup>[28]</sup>工具对部分ImageNet数据和跟踪场景数据进行降维处理及可视化操作,数据集可视化情况如图5所示。

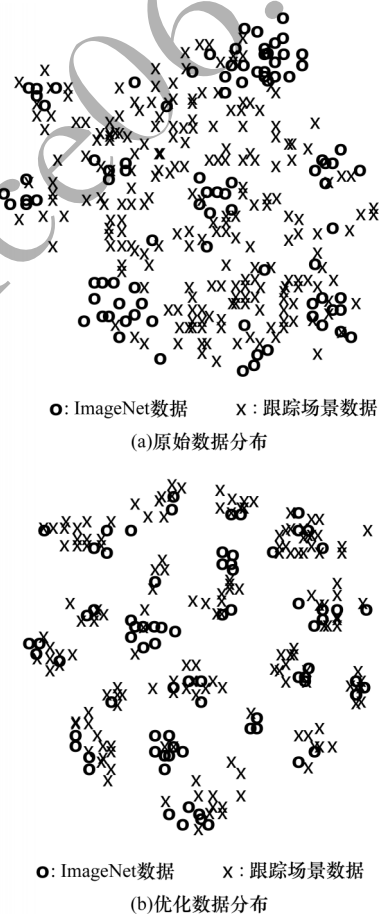
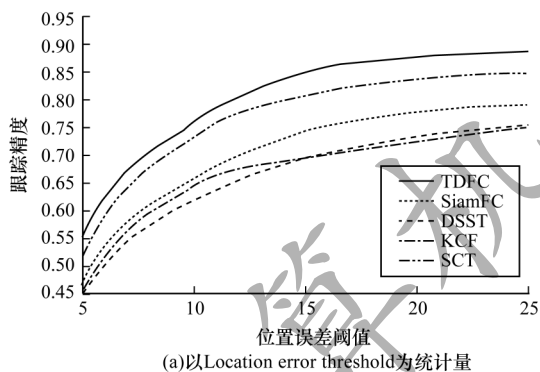


图5 ImageNet数据和跟踪场景数据分布的可视化效果

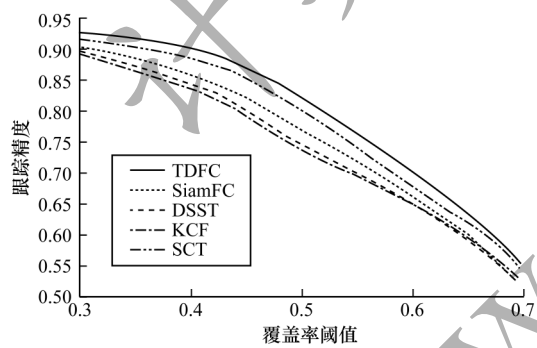
Fig.5 Visualization of distribution of ImageNet data and tracking scene data

图5(a)展示了未经2.1节对抗学习方法优化的图像数据分布,其中,ImageNet数据展现出一定的可分性,跟踪场景数据的分布特征并不明显,与ImageNet数据显示出不一致的分布情况;图5(b)展示了经过2.1节对抗学习方法优化后的数据分布情况,可以发现,优化后数据的分布趋于一致,虽然跟踪场景数据与ImageNet数据分布未完全匹配,但跟踪场景数据呈现出与ImageNet数据聚类中心保持一致的趋势,并呈现出一定的可分性,进一步验证了对抗学习方法对解决数据分布不匹配问题的有效性。

选择最优算法配置TDFC(本文模型)并与其他目标跟踪算法进行对比实验。图6所示为各算法的性能统计信息,其中,图6(a)的横轴统计量为“Location error threshold”,图6(b)的横轴统计量为“Overlap threshold”。



(a)以Location error threshold为统计量



(b)以Overlap threshold为统计量

图6 不同跟踪算法的跟踪精度统计

Fig.6 Tracking accuracy statistics of different tracking algorithms

通过上述跟踪算法的性能比较结果可以看出,本文设计的对抗学习方法和高效特征压缩使得TDFC的跟踪精度和速度均高于SCT、SiamFC和DSST算法。与KCF相比,TDFC有14.6%的跟踪精度提升,虽然跟踪速度低于KCF,但103.1FPS的跟踪速度足以满足实时跟踪的需求,符合精度与速度兼具的特点。综上,与其他跟踪算法相比,本文算法能以更高的精度对目标进行实时跟踪。

图7所示为本文算法以及其他跟踪算法在OTB-100数据集上的部分跟踪效果。

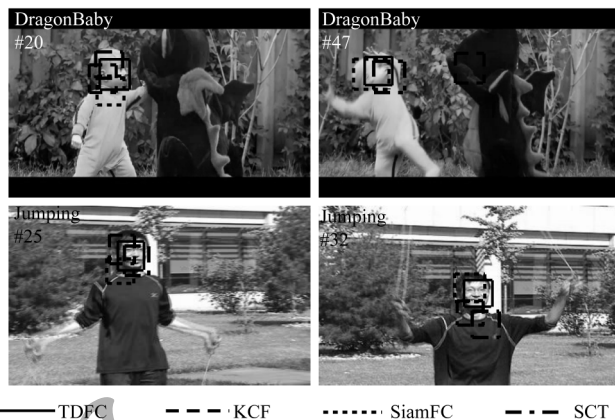


图7 不同算法的跟踪效果对比

Fig.7 Tracking effect comparison of different algorithms

#### 4 结束语

本文提出一种基于对抗学习和特征压缩的高精度实时跟踪算法。使用对抗学习方法优化图像特征提取过程,设计双通道自编码器结构压缩图像特征并结合类别信息优化训练过程。实验结果表明,相比现有的实时跟踪算法,该算法具有明显的精度优势,且在有限的精度损失下能够取得较大的速度提升。下一步将降低本文算法在跟踪时的计算复杂度,并将其扩展到DCF、SO-DLT等其他跟踪框架中。

#### 参考文献

- [1] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA; IEEE Press, 2016: 4293-4302.
- [2] TAO R, GAVVES E, SMEULDERS A W M. Siamese instance search for tracking [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA; IEEE Press, 2016: 1420-1429.
- [3] DANELLJAN M, HAGER G, SHAHBAZ KHAN F, et al. Convolutional features for correlation filter based visual tracking [C]//Proceedings of IEEE International Conference on Computer Vision Workshops. Washington D. C. , USA; IEEE Press, 2015: 58-66.
- [4] DANELLJAN M, BHAT G, SHAHBAZ KHAN F, et al. ECO: efficient convolution operators for tracking [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA; IEEE Press, 2017: 6638-6646.
- [5] WANG L, OUYANG W, WANG X, et al. STCT: sequentially training convolutional networks for visual tracking [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA; IEEE Press, 2016: 1373-1381.
- [6] CHATFIELD K, SIMONYAN K, VEDALDI A, et al. Return of the devil in the details: delving deep into convolutional nets [EB/OL]. [2020-02-05]. <http://de.arxiv.org/pdf/1405.3531>.



- [ 7 ] SAITO K, WATANABE K, USHIKU Y, et al. Maximum classifier discrepancy for unsupervised domain adaptation [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2018: 3723-3732.
- [ 8 ] YAN X, YANG J, SOHN K, et al. Attribute2Image: conditional image generation from visual attributes[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 776-791.
- [ 9 ] DU B, XIONG W, WU J, et al. Stacked convolutional denoising auto-encoders for feature representation[J]. IEEE Transactions on Cybernetics, 2016, 47(4): 1017-1027.
- [ 10 ] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
- [ 11 ] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(3): 583-596.
- [ 12 ] LI Y, ZHU J. A scale adaptive kernel correlation filter tracker with feature integration[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2014: 254-265.
- [ 13 ] DANELLJAN M, HÄGER G, KHAN F, et al. Accurate scale estimation for robust visual tracking[C]//Proceedings of British Machine Vision Conference. Washington D. C. , USA: IEEE Press, 2014: 15-22.
- [ 14 ] ZHANG K, ZHANG L, LIU Q, et al. Fast visual tracking via dense spatio-temporal context learning[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2014: 127-141.
- [ 15 ] MA C, HUANG J B, YANG X, et al. Hierarchical convolutional features for visual tracking[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C. , USA: IEEE Press, 2015: 3074-3082.
- [ 16 ] DANELLJAN M, HÄGER G, SHAHBAZ KHAN F, et al. Learning spatially regularized correlation filters for visual tracking[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C. , USA: IEEE Press, 2015: 4310-4318.
- [ 17 ] WANG N, LI S, GUPTA A, et al. Transferring rich feature hierarchies for robust visual tracking[EB/OL]. [2020-02-05]. <https://arxiv.org/pdf/1501.04587.pdf>.
- [ 18 ] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2016: 4293-4302.
- [ 19 ] WANG L, OUYANG W, WANG X, et al. Visual tracking with fully convolutional networks[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C. , USA: IEEE Press, 2015: 3119-3127.
- [ 20 ] FAN H, LING H. SANet: structure-aware network for visual tracking[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington D. C. , USA: IEEE Press, 2017: 42-49.
- [ 21 ] YUN S, CHOI J, YOO Y, et al. Action-decision networks for visual tracking with deep reinforcement learning[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2017: 2711-2720.
- [ 22 ] BEN-DAVID S, BLITZER J, CRAMMER K, et al. A theory of learning from different domains[J]. Machine Learning, 2010, 79(1/2): 151-175.
- [ 23 ] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[EB/OL]. [2020-02-05]. <https://papers.nips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>.
- [ 24 ] KINGMA D P, BA J. Adam: a method for stochastic optimization [EB/OL]. [2020-02-05]. <http://de.arxiv.org/pdf/1412.6980>.
- [ 25 ] WU Y, LIM J, YANG M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.
- [ 26 ] CHOI J, JIN CHANG H, JEONG J, et al. Visual tracking using attention-modulated disintegration and integration [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2016: 4321-4330.
- [ 27 ] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking [C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 850-865.
- [ 28 ] MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.

编辑 吴云芳