软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2021,32(11):3482−3495 [doi: 10.13328/j.cnki.jos.006058]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 跨域和跨模态适应学习的无监督细粒度视频分类*

何相腾, 彭宇新

(北京大学 王选计算机研究所,北京 100080)

通讯作者: 彭宇新, E-mail: pengyuxin@pku.edu.cn

**摘 要**: 细粒度视频分类旨在识别粗粒度大类中的细粒度子类,是计算机视觉中一个极具挑战的任务.考虑到视频数据的标注成本巨大,而图像的标注成本相对较小,且细粒度图像分类已经取得了较为显著的进展,一个自然的想法是不用标注,以无监督的方式将细粒度图像分类中学习到的知识自适应地迁移到细粒度视频分类中.然而,来源不同的图像和视频之间存在着域差异和模态差异,这导致细粒度图像分类的模型不能直接应用于细粒度视频分类.为了实现无监督的细粒度视频分类,提出一种无监督辨识适应网络,能够将辨识性定位能力从细粒度图像分类迁移到细粒度视频分类.进一步,提出一种渐进式伪标签策略来迭代地引导无监督辨识适应网络学习目标域视频的数据分布.在 CUB-200-2011、Cars-196 图像数据集和 YouTube Birds、YouTube Cars 视频数据集上验证该方法跨域、跨模态的适应能力,实验结果证明了该方法在无监督细粒度视频分类上的优势.

**关键词**: 细粒度视频分类;无监督辨识适应网络;域差异;模态差异;域适应

**中图法分类号**: TP181

中文引用格式: 何相腾,彭宇新.跨域和跨模态适应学习的无监督细粒度视频分类.软件学报,2021,32(11):3482−3495. http://www.jos.org.cn/1000-9825/6058.htm

英文引用格式: He XT, Peng YX. Unsupervised fine-grained video categorization via adaptation learning across domains and modalities. Ruan Jian Xue Bao/Journal of Software, 2021,32(11):3482−3495 (in Chinese). http://www.jos.org.cn/1000-9825/6058.htm

## Unsupervised Fine-grained Video Categorization via Adaptation Learning Across Domains and Modalities

HE Xiang-Teng, PENG Yu-Xin

(Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China)

**Abstract**: Fine-grained video categorization is a highly challenging task to discriminate similar subcategories that belong to the same basic-level category. Due to the significant advances in fine-grained image categorization and expensive cost of labeling video data, it is intuitive to adapt the knowledge learned from image to video in an unsupervised manner. However, there is a clear gap to directly apply the models learned from image to recognize the fine-grained instances in video, due to domain distinction and modality distinction between image and video. Therefore, this study proposes the unsupervised discriminative adaptation network (UDAN), which transfers the ability of discrimination localization from image to video. A progressive pseudo labeling strategy is adopted to iteratively guide UDAN to approximate the distribution of the target video data. To verify the effectiveness of the proposed UDAN approach, adaptation tasks between image and video are performed, adapting the knowledge learned from CUB-200-2011/Cars-196 datasets (image) to YouTube Birds/YouTube Cars datasets (video). Experimental results illustrate the advantage of the proposed UDAN approach for unsupervised fine-grained video categorization.

**Key words**: fine-grained video categorization; unsupervised discriminative adaptation network; domain distinction; modality distinction; domain adaption

* 基金项目: 国家自然科学基金(61925201, 61771025)
Foundation item: National Natural Science Foundation of China (61925201, 61771025)
收稿时间: 2019-09-09; 修改时间: 2020-03-09; 采用时间: 2020-04-19

细粒度视觉分类(fine-grained visual categorization,简称 FGVC)是计算机视觉领域中一个重要且极具挑战的任务,其旨在对粗粒度的大类(如鸟、车等)中相似的细粒度子类(如鸟类中的小白额燕鸥、普通燕鸥和福斯特燕鸥等)进行识别.细粒度视觉分类主要有两大挑战:

(1) 类内差异大.如图 1 中的每一列所示,它们属于相同的细粒度子类,但受到不同视角和姿态等因素的影响,在外表上具有较大差异.

(2) 类间差异小.如图 1 中的每一行所示,它们属于不同的细粒度子类,但是由于它们属于同一粗粒度的大类,因此在颜色、形态等外表上差异细微,很难被区分.

这两大挑战使得细粒度视觉分类任务十分困难.现有方法一般聚焦在图像领域的细粒度分类(即细粒度图像分类),而视频领域的相关研究(即细粒度视频分类)还相对较少.但是,随着手机等移动设备上视频娱乐交友软件的快速发展,现在人们更加倾向于通过上传视频来记录他们的所见所闻以及表达他们的所感所想.视频数据的海量动态增长,使得视频的管理变得极为重要,而细粒度视频分类就是进行视频管理的重要手段之一.


Fig.1 Distinctions of domains and modalities between image and video, as well as the challenges of fine-grained visual categorization: Large variance in the same subcategory and small variance among different subcategories

图 1 图像和视频之间存在的域差异和模态差异以及细粒度视觉分类任务的挑战:"类内差异大、类间差异小"

近年来,研究者开始逐渐关注到细粒度视频分类任务.Zhu 等人[1]构建了两个细粒度视频数据集,以推动细粒度视频分类的进一步研究和应用.同时,Zhu 等人也提出了一种冗余减少注意力网络(redundancy reduction attention network,简称 RRAN)来提高细粒度视频分类的准确率.但是,RRAN 的训练依赖于大量标注的视频数据,而这些数据的标注是非常耗时耗力的,成本十分巨大.考虑到图像的标注成本相对较小,并且细粒度图像分类已经取得了较为显著的进展,一个自然的想法是不用标注,以无监督的方式将知识从细粒度图像分类迁移到细粒度视频分类.这能够有效减少视频数据的巨大标注成本,同时能够充分发挥细粒度图像分类模型的作用.

但是,从图像数据中学习到的模型很难直接应用于细粒度视频分类,主要是因为来源不同的图像和视频之间存在两种类型的差异.

(1) 域差异:如图1所示,标准的图像数据集与真实应用中用户上传的视频存在差异[2].例如:标准的图像数据具有较高的分辨率、对象一般位于图像的中心区域、背景信息也相对简单;而用户上传的视频数据则具有分辨率低、对象位置不确定、背景信息复杂等特点.这些差异均导致了两个域数据分布的不一致.

(2) 模态差异:图像只有静态的空域信息,但是除此之外,视频还有动态的时域信息,这使得细粒度视频分类更具有挑战性.

因此,将知识从图像数据迁移到视频数据包含了两层含义:(1) 从标准封闭数据集向真实应用场景的开放数据集的迁移;(2) 从空域向时域的迁移.此外,考虑到细粒度视频分类“类内差异大、类间差异小”的挑战,从图像到视频的无监督跨域和跨模态细粒度适应学习是一项极具挑战的任务.

因此,本文将知识从源域标注的图像数据迁移到目标域未标注的视频数据,旨在实现无监督的细粒度视频分类.首先,本文提出了一种无监督辨识适应网络(unsupervised discriminative adaptation network,简称 UDAN),能够将辨识性定位能力从细粒度图像分类迁移到细粒度视频分类;然后,本文提出了一种渐进式伪标签策略来迭代地引导无监督辨识适应网络学习目标域视频的数据分布.本文是细粒度视觉分类领域中,从图像到视频跨域、跨模态的无监督工作,能够有效地降低视频数据标注的巨大成本,进一步推动细粒度视频分类的研究与应用.为了验证本文 UDAN 方法的有效性,本文将辨识定位能力从 CUB-200-2011、Cars-196 图像数据集迁移到 YouTube Birds、YouTube Cars 视频数据集.实验结果验证了本文 UDAN 方法能够在无监督细粒度视频分类上取得当前最好的分类准确率.

## 1 相关工作

本节对细粒度视觉分类、域适应的相关工作进行了简单概述.其中,细粒度视觉分类是本文的目标任务,而域适应是本文的聚焦点.

### 1.1 细粒度视觉分类

细粒度视觉分类是计算机视觉领域最具挑战的任务之一,在学术界和工业界都得到了广泛关注.细粒度视觉分类在实际生活中也有着丰富的应用场景,如无人驾驶、动植物保护、癌症检测、海洋作业等,因此具有重要的研究和应用价值.

现有细粒度视觉分类一般聚焦在细粒度图像分类,而细粒度视频分类的相关研究还相对较少.本节主要从细粒度图像分类和细粒度视频分类两个方面对细粒度视觉分类进行介绍.

#### 1.1.1 细粒度图像分类

细粒度图像分类一般划分为基于定位的方法、基于编码的方法以及基于文本或属性的方法.

- 基于定位的方法

由于不同的细粒度类别之间外表相似,仅在一些局部区域存在细微的差异,因此,研究者们一般采取如下的方法流程:首先定位到图像中的辨识性区域,如鸟的头部、翅膀、尾部等,这是现有细粒度图像分类方法的关键;然后,学习并提取辨识性区域的特征以进行细粒度分类.Zhang 等人[3]分别利用对象位置信息(bounding box)和部件位置信息(part location)来训练对象检测器和部件检测器,在测试过程中,利用两个检测器来定位图像中的对象区域及其部件区域.但是对象和部件位置信息的标注极其耗时耗力,标注成本十分巨大.因此,研究者开始聚焦于如何在不使用对象和部件位置信息标注的情况下自动定位图像中的辨识性区域.

Krause 等人[4]仅使用对象位置信息来训练部件检测器,避免了部件位置信息的使用.为了进一步降低标注成本,Xiao 等人[5]首先利用选择搜索的方法(selective search)[6]对每一张图像生成多个候选图像块;然后利用对象级和部件级的注意力机制,从候选图像块中选出具有辨识性的区域.这是一种弱监督(weakly-supervised)的学习方式,既不使用对象位置信息,也不使用部件位置信息.在这之后,多个弱监督细粒度图像分类方法[7,8]相继提出,进一步推动了细粒度图像分类的研究与应用.

- 基于编码的方法

一些工作聚焦于特征表示学习,其主要方法是对卷积神经网络(convolutional neural network,简称 CNN)的特征图(feature map)进行统计编码,以获取更好的特征表示.Lin 等人[9]提出了双线性汇合方法(bilinear pooling),通过计算 CNN 特征图的格拉姆矩阵来捕获特征通道之间成对的相关关系,从而获得更好的特征表示,以提升细粒度图像分类准确率.受到双线性汇合方法的启发,Gao 等人[10]进一步提出了紧凑双线性汇合方法(compact bilinear pooling),通过 CNN 特征图低维投影的内积近似二次多项式核来降低双线性汇合方法的高维度.Cui 等人[11]进一步利用核近似获取更高阶的特征表示.Wang 等人[12]学习辨识性过滤器,并将其应用到 CNN 中使得其更加关注辨识性特征的学习.

- 基于文本或属性的方法

由于图像的文本描述信息(如这是一只白色翅膀、橙色喙的海鸥)以及图像的属性信息(如白色翅膀、橙色鸟喙等)能够提供图像中对象的细粒度辨识性信息,与图像的视觉信息互为补充,能够进一步促进图像的特征表示学习.因此,研究者开始研究基于文本或属性的方法.He 等人[13]提出联合建模文本和视觉信息的方法,挖掘二者之间的关联信息以提升细粒度图像分类的准确率.Chen 等人[14]利用属性信息来构建知识图(knowledge graph),进一步通过图卷积神经网络来学习图像的辨识性特征.

#### 1.1.2 细粒度视频分类

相比于图像,视频通常包含了更丰富的辨识性信息,因此,研究者们开始关注细粒度视频分类任务.Saito 等人[15]构建了一个细粒度视频数据集来探索运动信息在细粒度分类中的有效性.Zhu 等人[1]构建了两个大规模细粒度视频数据集,并且提出了冗余降低注意力网络来降低 CNN 模型中特征的冗余信息,从而学习得到细粒度的辨识信息.本文的目标是充分发挥在图像数据中学习到的知识,利用细粒度图像分类来实现无监督条件下的细粒度视频分类.

### 1.2 域适应

域适应(domain adaptation)任务也是计算机视觉领域的研究热点之一.随着域的变化,例如从源域数据(标准的图像分类数据集)到目标域数据(用户上传的视频),输入数据 $X$ 和输出标签 $Y$ 的分布 $P(X,Y)$会随之发生变化.影响 $P(X,Y)$变化的因素主要包括空间位置信息变化、外表多样性、图像质量等变化[16],这些因素的变化均会导致在源域数据上学习得到的模型在目标域数据上的效果很差[17].

在细粒度视觉分类领域,仅有少数工作者做了域适应的相关工作,他们一般聚焦于从源域图像数据向目标域图像数据的迁移.Gebru 等人[18]提出一种基于属性的多任务域适应算法,能够从标准的图像数据集向真实用户图像的迁移.Cui 等人[19]从大规模图像数据集中学习知识,然后适用于小规模图像数据集.这些工作都充分利用了源域图像数据中已学习到的知识,有效地提升了模型在目标域图像数据上的分析效果.

本文研究的域适应任务的源域数据是标准的图像,目标域数据是用户上传的真实视频.这涉及到模态的迁移(由图像到视频)、域的迁移、小规模数据向大规模数据的迁移.并且本文所研究的是无监督条件下的细粒度域适应任务,目标域数据的标注信息是不可以在训练过程中使用的.这是一个无监督的从图像到视频的细粒度域适应工作.

## 2 无监督细粒度视频分类

本文提出了无监督辨识适应网络,通过联合辨识最大均值差异准则(joint discriminative maximum mean discrepancy,简称 JDMMD),将在源域图像数据中学到的辨识性定位能力迁移到目标域的视频数据中.进一步,本文提出了一种渐进式伪标签策略,通过迭代的方式引导 UDAN 模型近似估计目标域视频数据的分布.

### 2.1 问题定义

本文所研究的问题是无监督条件下的细粒度视频分类,旨在将知识从标注的源域图像数据迁移到未标注的目标域视频数据中.问题的定义描述如下:

给定标注的源域图像数据 $S=\{I,Y^I\}$以及未标注的目标域视频数据 $T=\{V\}$.其中,$I$ 和 $V$ 分别表示图像数据和视频数据,$Y^I$ 表示图像数据的类别标签.由于域差异和模态差异,源域图像数据和目标域视频数据具有不同的数据分布,这导致从源域数据中学习到的知识很难直接应用于目标域数据.本文的目标是通过降低图像和视频之间跨域、跨模态的数据分布差异,进而在无监督条件下学习得到目标域视频数据的分类器.

### 2.2 无监督辨识适应网络

本文提出了无监督辨识适应网络 UDAN,以实现无监督细粒度视频分类.图 2 展示了本文 UDAN 方法的框架,采用当前先进的 CNN 网络 ResNet50[20]作为基础网络模型,并且这个基础网络模型可以替换为其他 CNN 网络.其中,提出的联合辨识最大均值差异(JDMMD)能够将辨识定位能力从图像数据迁移到视频数据,伪标签损失函数能够引导 UDAN 模型来拟合目标域视频的数据分布.

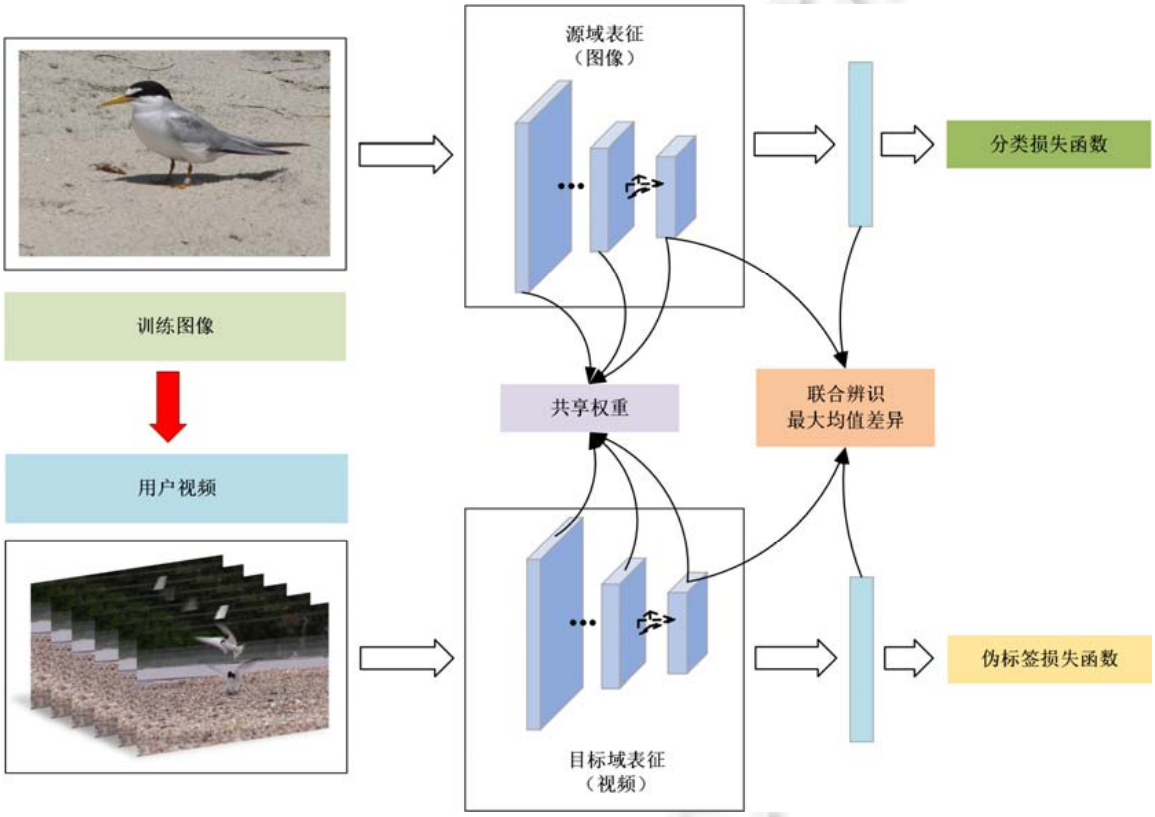


Fig.2 Unsupervised discriminative adaptation network (UDAN)

图 2 无监督辨识适应网络(UDAN)

进一步,为了实现无监督辨识适应,本文对 ResNet50 网络模型的损失函数进行了重新设计,其定义如下:

$$Loss = Loss_{cls}^{S} + Loss_{JDMMD} + Loss_{pseudo}^{T} \tag{1}$$

#### 2.2.1 源域图像数据上的分类损失

在公式(1)中,第 1 项 $Loss_{cls}^{S}$ 表示 UDAN 网络模型在标注的源域图像数据上的分类损失,它的目的是从源域数据中学习辨识性特征,其定义如下:

$$Loss_{cls}^{S} = \frac{1}{N_S}\sum_{k=1}^{N_S} ce(f(x_k^S, y_k^S)) \tag{2}$$

本文采用交叉熵损失函数 $ce(\cdot)$(cross-entropy loss function)来优化 UDAN 在源域图像数据上的学习.公式

(2)中,$N_S$ 表示源域中图像数据的数量,$x_k^S$ 和 $y_k^S$ 分别表示源域中第 $k$ 个图像数据及其类别标签.通过最小化 $Loss_{cls}^S$ 来学习源域图像数据的辨识性特征.

2.2.2 联合辨识最大均值差异损失

Yosinski 等人的研究[21]表明,CNN 网络高层的迁移能力会随着域差异的增加而降低,即越高的层其迁移能力越差.这就使得 CNN 网络的泛化能力较差,一旦迁移到另外一个域的数据上,其表现就会急剧下降.为了解决这个问题,Long 等人[22]提出了一种联合最大均值差异准则(joint maximum mean discrepancy,简称 JMMD),以使得多个网络层在跨域条件下实现联合分布对齐,其定义如下:

$$D_L(P,Q) \triangleq \| C_{Z^{S,1:|L|}}(P) - C_{Z^{T,1:|L|}}(Q) \|^2_{\otimes_{l=1}^{|L|} H^l} \tag{3}$$

本文采用 CNN 网络的最后|$L$|层输出向量(即 $P(Z^{S_1},...,Z^{S_{|L|}})$ 和 $Q(Z^{T_1},...,Z^{T_{|L|}})$)来表示源域和目标域数据的分布,它们的差异用希尔伯特空间嵌入(Hilbert space embedding)来度量.在本文 UDAN 方法中,采用 ResNet50 网路模型的最后两层,即 $L$={$pool5$,$fc$},它们的输出向量表示为 $Z$.

考虑到在细粒度视觉分类任务中,细粒度类别之间的差异一般在对象的部件上,因此,本文提出了联合辨识最大均值差异.JDMMD 能够充分利用图像中辨识性区域的特征,有效地分析细粒度类别之间的差异.所以,$P$ 和 $Q$ 重新表示为 $P(R^{S_1},...,R^{S_{|L|}})$ 和 $Q(R^{T_1},...,R^{T_{|L|}})$,其中,$R$ 表示图像中辨识性区域的特征.

为了实现 UDAN 网络模型的端到端(end-to-end)训练,本文设计了辨识性生成网络,其包含两个部分:辨识性生成层(discrimination generation layer)和感兴趣区域对齐层(RoI align layer)[23](如图 3 所示).

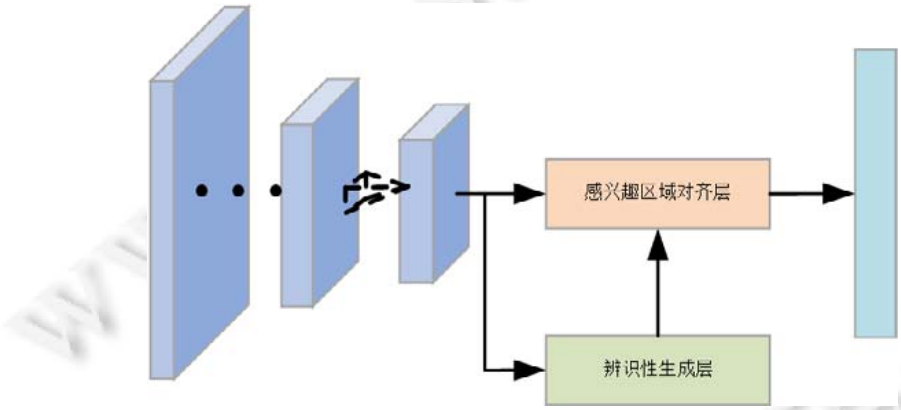


Fig.3 Architecture of the proposed discrimination generator network

图 3 本文提出的辨识性生成网络的框架

- 辨识性生成层

辨识性生成层是为了生成图像中辨识性区域的位置坐标信息,然后作为感兴趣对齐层的输入.首先,提取 ResNet50 网络模型最后一层卷积层的所有特征图,并且通过平均池化操作把它们聚合成一个特征图 $F$;然后,对聚合成的特征图 $F$ 进行上采样操作,将其尺寸变为输入图像的尺寸.特征图 $F$ 中的每一个像素值 $f_{i,j}$ 表示原始图像对应像素点($i$,$j$)的卷积响应,它表示了原始图像对应像素点的辨识性程度.所以,基于特征图 $F$ 可以获取图像中的辨识性区域.

为了获得辨识性区域的位置坐标信息,本文通过大津阈值(OTSU)[24]二值化算法和最大连通域提取算法来实现.经过这两个操作,可以获取辨识性区域的左上角坐标($x_1$,$y_1$)和右下角坐标($x_2$,$y_2$).然后,将两个坐标信息作为感兴趣区域对齐层的输入.

- 感兴趣区域对齐层

感兴趣区域对齐层是为了生成图像中辨识性区域对应的特征图,采用双线性插值算法来计算对应区域的

特征,有效避免了 RoI Pooling 的量化损失.

经过辨识性生成网络,可以获得源域数据和目标域数据对应的辨识性联合分布 $P(R^{S_1},...,R^{S_{|L|}})$ 和 $Q(R^{T_1},...,R^{T_{|L|}})$.因此,公式(1)中第 2 项 $Loss_{JDMMD}$ 对应的定义为

$$Loss_{JDMMD} \triangleq \| C_{R^{S,1:|L|}}(P) - C_{R^{T,1:|L|}}(Q) \|^2_{\otimes_{l=1}^{|L|} H^l} \tag{4}$$

与 JAN 方法[22]一样,本文采用其相似线性时间估计来适应随机梯度下降算法 SGD,以进行端到端的网络模型训练.

2.2.3 目标域视频数据上的分类损失

公式(1)中的第 3 项 $Loss^T_{pseudo}$ 表示 UDAN 网络模型在未标注的目标域视频数据上的分类损失.首先,本文对目标域视频进行抽帧;然后,利用这些视频帧来训练 UDAN 网络模型.其定义如下:

$$Loss^T_{pseudo} = \frac{1}{N_T}\sum_{k=1}^{N_T} ce(f(x_k^T, y_k^T)) \tag{5}$$

与 $Loss^S_{cls}$ 一样,本文采用交叉熵损失函数来优化本文 UDAN 方法在目标域视频数据上的学习.其中,$N_T$ 表示目标域中视频帧的数量, $x_k^T$ 和 $y_k^T$ 分别表示目标域中第 $k$ 个视频帧及其类别伪标签.伪标签的获取将在下一节中介绍.

### 2.3 渐进式伪标签策略

如果能够直接从目标域的视频数据中学习,本文 UDAN 方法将会取得更好的细粒度分类效果.在无监督条件下,目标域视频数据的标注信息不能使用,然而在源域图像数据上训练得到的 UDAN 网络模型能够较为准确地对目标域的部分视频帧进行分类.因此,本文提出了一种渐进式伪标签策略,能够有效地给目标域的部分视频帧打标签,从而利用这些带有伪标签信息的视频帧数据对 UDAN 网络模型进行微调(fine-tune).对于目标域的视频帧,需要满足下列条件才能获取伪标签:

$$\max(softmax(x_K^T)) > \tau \tag{6}$$

为了使得所选择的目标域视频帧具有较高的置信度,本文将 $\tau$ 设置为 0.99.通过利用带有伪标签的视频帧数据进行 UDAN 的训练,UDAN 能够直接学习到目标域视频数据的特定知识.进一步,再次利用训练后的 UDAN 执行伪标签操作,会获得更多的视频帧数据以作训练之用.通过这样一种渐进式的迭代过程,UDAN 能够进一步提升细粒度分类能力.

## 3 实 验

在实验部分,为了验证本文 UDAN 方法的有效性,本文将 CUB-200-2011 数据集[25]、Cars-196 数据集[26]作为源域图像数据集,YouTube Birds 数据集[1]、YouTube Cars 数据集[1]作为目标域视频数据集,旨在将知识从图像迁移到视频.

### 3.1 数据集介绍

CUB-200-2011、Cars-196、YouTube Birds 和 YouTube Cars 这 4 个细粒度数据集的划分见表 1,分别包含了训练集和测试集,其具体介绍如下.

(1) CUB-200-2011 数据集[25]是使用最为广泛的细粒度图像数据集,共包含 11 788 张图像,涵盖了 200 个鸟类的细粒度子类别,如小白额燕鸥、普通燕鸥和福斯特燕鸥等.其数据划分如下:训练集包含 5 994 张图像,测试集包含 5 794 张图像.每一张图像都有详细的标注信息,包括图像级的类别标签、图像中对象的位置标注信息、15 个部件的位置标注信息以及 312 个二值属性标注信息.在本文实验中,仅使用了图像级的类别标签信息.

(2) Cars-196 数据集[26]共包含 16 185 张图像,涵盖了 196 个车类的细粒度子类别,如现代伊兰特 2007、丰田红杉 SUV 2012 等.其数据集划分如下:训练集包含 8 144 张图像,测试集包含 8 041 张图像.每一张

图像有两种标注信息:1 个图像级的类别标签和 1 个图像中对象级的位置标注信息.

(3) YouTube Birds 数据集[1]是新近构建的大规模细粒度鸟类视频数据集,共包含 18 350 个视频.与 CUB-200-2011 数据集相同,涵盖了 200 个鸟类的细粒度子类别,而且二者的子类别种类完全相同.视频数据来源于 YouTube 视频网站用户上传的真实视频,每个视频时长不超过 5 分钟.数据集的划分如下:训练集包含 12 666 个视频,测试集包含 5 684 个视频.每个视频仅有一个视频级的类别标签信息.为了验证无监督细粒度视频分类的可行性和有效性,在本文的实验中没有使用视频的标注信息.

(4) YouTube Cars 数据集[1]同样是新近构建的大规模细粒度车类视频数据集,共包含 15 220 个视频.与 Cars-196 数据集相同,涵盖了 196 个车类的细粒度子类别,而且二者的子类别种类完全相同.数据集的划分如下:训练集包含 10 259 个视频,测试集包含 4 961 个视频.每个视频仅有 1 个视频级的类别标签信息.为了验证无监督细粒度视频分类的可行性和有效性,本文的实验中没有使用视频的标注信息.

**Table 1** Data partitions on four fine-grained datasets

**表 1** 4 个细粒度数据集的数据划分

| 数据集 | 训练集 | 测试集 |
|---|---|---|
| CUB-200-2011 | 5 994 张图像 | 5 794 张图像 |
| Cars-196 | 8 144 张图像 | 8 041 张图像 |
| YouTube Birds | 12 666 个视频 | 5 684 个视频 |
| YouTube Cars | 10 259 个视频 | 4 961 个视频 |

### 3.2 评价任务和指标

#### 3.2.1 评价任务

为了评价本文 UDAN 方法的有效性,本文设计了两种适应任务,分别是图像到视频帧的适应任务以及图像到视频的适应任务.以 CUB-200-2011 和 YouTube Birds 两个数据集为例,见表 2.

(1) 图像到视频帧的适应任务($I\rightarrow F$):本文将 CUB-200-2011 数据集的训练图片作为源域数据,将 YouTube Birds 数据集的测试视频作为目标域数据.对于 YouTube Birds 数据集的视频,本文抽取中间帧作为目标域测试视频帧.

(2) 图像到视频的适应任务($I\rightarrow V$):本文将 CUB-200-2011 数据集的训练图片作为源域数据,将 YouTube Birds 数据集的测试视频作为目标域数据.

需要注意的是:在训练过程中使用的是标注的 CUB-200-2011 数据集的训练图片和未标注的 YouTube Birds 数据集的训练视频,在测试过程中使用的是 YouTube Birds 数据集的测试视频.

**Table 2** Two types of adaptation tasks

**表 2** 两种适应任务

| 适应任务 | 源域数据 | 目标域数据 |
|---|---|---|
| 图像到视频帧($I\rightarrow F$) | CUB-200-2011 的训练集图像 | YouTube Birds 的测试集视频中间帧 |
| 图像到视频($I\rightarrow V$) | CUB-200-2011 的训练集图像 | YouTube Birds 的测试集视频 |

#### 3.2.2 评价指标

在本文的两种适应任务实验中,采用准确率(accuracy)作为评价指标来验证本文 UDAN 方法的有效性.准确率的定义如下:

$$Accuracy=\frac{R_a}{R} \tag{7}$$

其中,$R$ 表示测试集中视频或者视频帧的总数,$R_a$ 表示正确分类的视频或者视频帧的数目.

### 3.3 实现细节

为了方便后续研究者与本文 UDAN 方法进行公平对比,本节从基础 CNN 模型、视频处理和训练细节这 3

个方面对本文 UDAN 方法的实现细节进行介绍.

(1) 基础 CNN 模型:本文采用 ResNet50 网络模型[20]作为基础 CNN 模型.为了获得更好的细粒度视频分类准确率,本文在 ResNet50 网络模型的基础上作了一些改动.具体地,将输入图片的裁剪尺寸(crop size)设置为 448×448,在最后一层卷积层后面加一个平均池化层(average pooling layer),其内核尺寸为 14,步长为 1.

(2) 视频处理:对于 YouTube Birds 和 YouTube Cars 这 2 个视频数据集,本文实验部分仅对其 RGB 视频帧进行分析处理.具体地,在图像到视频帧和图像到视频的适应任务中,对于每个训练视频,本文等间隔地抽取 5 帧作为训练视频帧数据.在测试过程中,两个适应任务的设置有所不同:对于图像到视频帧适应任务,本文抽取测试视频的中间帧作为测试数据;对于图像到视频适应任务,本文对每个测试视频等间隔地抽取 25 帧作为测试数据.

(3) 训练细节:为了获得最好的细粒度视频分类效果,在训练过程中,本文采用了两次迭代来进行伪标签的生成.对于伪标签生成的阈值 $\tau$,本文设置为 0.99 和 0.9.在本文 UDAN 网络模型的训练过程中,采用梯度下降算法 SGD 进行优化,设置批尺寸(batch size)大小为 8,权值衰减系数(weight decay)为 0.0005,冲量系数(momentum)为 0.9.本文设置初始学习率(learning rate)为 1e-5,每训练 6 个 epoch 学习率以 0.5 的系数减小.

### 3.4 源域数据与目标域数据之间的差异

本节展示了源域数据与目标域数据之间的差异,以 CUB-200-2011 和 YouTube Birds 两个数据集为例.与文献[18]一样,首先利用源域数据对模型进行训练,然后将训练好的模型迁移到目标域进行测试.为了充分验证源域与目标域之间的差异,本文执行了两种适应任务:图像到视频帧适应任务($I\rightarrow F$)和图像到视频适应任务($I\rightarrow V$).结果见表 3(其中两个数据集分别用 $S$ 和 $T$ 表示),其中,源域和目标域分别表示为 $S$ 和 $T$.

**Table 3** Results of adaptation between CUB-200-2011 and YouTube-Birds datasets

**表 3** 从 CUB-200-2011 数据集到 YouTube Birds 数据集的适应结果

| 数据集 | | 准确率(%) | |
|---|---|---|---|
| 训练集 | 测试集 | $I\rightarrow I/F$ | $I\rightarrow V$ |
| $S$ | $S$ | 85.2 | – |
| $S$ | $T$ | 34.2 | 40.7 |
| $S+T$ | $T$ | 44.4 | 60.6 |

首先,本文利用标注的 CUB-200-2011 数据集的训练图像训练 ResNet50 网络模型;然后,在 CUB-200-2011 数据集的测试图像上验证 ResNet50 网络模型的细粒度分类效果.如表 3 所示,ResNet50 网络模型能够取得不错的细粒度分类效果,即 85.2%的准确率.

然后,本文验证了 $I\rightarrow F$ 和 $I\rightarrow V$ 两种适应任务下,ResNet50 网络模型的适应能力.

(1) $I\rightarrow F$:直接利用在 CUB-200-2011 训练集图像数据上学习到的 ResNet50 网络模型进行 YouTube Birds 测试集中视频中间帧的测试,细粒度分类准确率出现了断崖式的下降,从 85.2%下降到 34.2%.

(2) $I\rightarrow V$:同样,直接利用在 CUB-200-2011 训练集图像数据上学习到的 ResNet50 网络模型进行 YouTube Birds 测试集中视频的测试,细粒度分类准确率同样下降严重,只取得了 40.7%的准确率.相比 $I\rightarrow F$ 任务取得了较高的细粒度分类准确率,这是因为视频相比单一视频中间帧包含更多、更丰富、更有用的信息.

最后,本文验证了额外使用 YouTube Birds 训练集数据对于 $I\rightarrow F$ 和 $I\rightarrow V$ 两种适应任务的影响.

(1) $I\rightarrow F$:同时利用 CUB-200-2011 训练集图像数据和 YouTube Birds 训练集的视频帧来训练 ResNet50 网络模型,对 YouTube Birds 测试集中的视频中间帧进行测试.细粒度分类准确率为 44.4%,相比仅使用 CUB-200-2011 训练集图像数据准确率提升了 10.2%.这充分体现了源域(即 CUB-200-2011 数据集)与目标域数据(即 YouTube Birds 数据集)之间的巨大差异.

(2) $I\rightarrow V$:同上述一样的训练方式,对 YouTube Birds 测试集中的视频进行测试.相比仅使用 CUB-200-2011 训练集图像数据准确率提升了 19.9%.

从上述分析可以看出,源域与目标域数据之间存在巨大的差异,$I\rightarrow F$ 和 $I\rightarrow F$ 这两种适应任务是非常具有挑战性的.此外,即使使用了 YouTube Birds 训练集的数据,细粒度分类效果依旧不理想,这说明了对用户上传的视频数据进行细粒度分类同样是一个非常具有挑战性的任务.

### 3.5 无监督细粒度视频分类

本节通过无监督细粒度视频分类任务来验证本文 UDAN 方法的有效性,将知识从标注的图像数据迁移到未标注的视频数据.本文将目标域的视频数据划分为两种模态:视频帧(每个视频的中间帧)及视频.同样地,执行两种适应任务:$I\rightarrow F$ 和 $I\rightarrow V$.本节从以下两个方面对实验结果进行分析比较:(1) 与现有方法对比;(2) 剥离实验.

#### 3.5.1 与现有方法对比

本节将本文 UDAN 方法与现有先进(state-of-the-art)方法进行了对比,并给出了详细的分析.为了公平对比,本文 UDAN 方法与对比方法在 $I\rightarrow F$ 和 $I\rightarrow V$ 这两种适应任务中均采用相同的实验设置.表 4、表 5 分别展示了从 CUB-200-2011 到 YouTube Birds 和从 Cars-196 到 YouTube Car 两组数据集上的结果.

**Table 4** Results on two types of adaptation tasks: Image-to-frame adaptation ($I\rightarrow F$), and image-to-video adaptation ($I\rightarrow V$) on CUB-200-2011 and YouTube Birds datasets

**表 4** CUB-200-2011 和 YouTube Birds 两个数据集上 $I\rightarrow F$ 和 $I\rightarrow V$ 两种适应任务上的结果

| 对比方法 | 准确率(%) | |
|---|---|---|
| | $I\rightarrow F$ | $I\rightarrow V$ |
| 本文 UDAN 方法 | 42.5 | 58.3 |
| JAN[22] | 36.5 | 46.4 |
| ResNet50[20] | 34.2 | 40.7 |
| ICAN[27] | 32.9 | 42.3 |
| MCD[28] | 30.1 | 43.9 |
| I3D[29] | – | 40.7 |

**Table 5** Results on two types of adaptation tasks: image-to-frame adaptation ($I\rightarrow F$), and image-to-video adaptation ($I\rightarrow V$) on Cars-196 and YouTube Cars datasets

**表 5** Cars-196 和 YouTube Cars 两个数据集上 $I\rightarrow F$ 和 $I\rightarrow V$ 两种适应任务上的结果

| 对比方法 | 准确率(%) | |
|---|---|---|
| | $I\rightarrow F$ | $I\rightarrow V$ |
| 本文 UDAN 方法 | 15.3 | 44.6 |
| JAN[22] | 10.4 | 15.5 |
| ResNet50[20] | 14.3 | 30.4 |
| ICAN[27] | 10.9 | 28.0 |
| I3D[29] | – | 40.9 |

表 4 展示了从 CUB-200-2011 到 YouTube Birds 上两种适应任务的结果,验证了本文 UDAN 方法的有效性,表明其在两种适应任务上都取得了最好的细粒度分类效果,与现有最好的方法相比分别提升了 6%和 11.9%.由于 $I\rightarrow F$ 和 $I\rightarrow V$ 这两种适应任务上的趋势一致,本节以 $I\rightarrow V$ 适应任务为例,从以下 3 个方面给出具体的分析比较.

(1) 与基础网络模型的对比.在实验中,本文 UDAN 方法采用 ResNet50 网络模型[20]作为基础 CNN 模型,因此首先与 ResNet50 网络模型的结果进行对比.直接利用 ResNet50 网络模型,在两种适应任务上,细粒度分类效果均比较差.在 $I\rightarrow V$ 适应任务上只有 40.7%的准确率.而本文 UDAN 方法可以将细粒度分类准确率提升 17.6%,这表明其具有缩短域差异和模态差异的能力.图 4 展示了在 CUB-200-2011 和 YouTube Birds 两个数据集上从 ResNet50 网络模型到本文 UDAN 方法,目标域视频的数据分布变化.从图 4 可以看出,本文的 UDAN 方法能够有效地将细粒度类别的数据区分开,而 ResNet50 网络模型下相同类别的数据相对分散而不够紧凑.进一步地,从表 3 中的结果可以看出,即使同时使用了标注的

源域图像训练数据和目标域的视频数据,细粒度视频分类准确率仅有 60.7%,只比本文 UDAN 的无监督方法高 2.4%.这充分验证了本文 UDAN 方法在无监督细粒度视频分类上的有效性.

(2) 与现有适应方法的对比.本文的 UDAN 方法与现有的适应方法进行了对比,如联合适应网络(joint adaptation network,简称 JAN)[22]、增量式的协同对抗网络(incremental collaborative and adverarial network,简称 ICAN)[27]和最大分类差异(maximum classifier discrepancy,简称 MCD)[28].JAN 基于联合最大均值差异,对多个网络层进行跨域的分布对齐.这是本文 UDAN 方法的基础结构.本文并没有采用从头训练的方式,而是利用在 CUB-200-2011 训练集上预训练的 RestNet50 网络模型对 JAN 方法进行初始化,然后在此基础上进行训练.这有效提升了模型的细粒度分类效果.相比 JAN 方法,本文的 UDAN 方法取得了较大的提升,在 $I \rightarrow V$ 适应任务上,将细粒度分类准确率从 46.4%提升到 58.3%.ICAN 方法在 CNN 特征提取中采用了多个域分类器(domain classifier)以学习与域相关和不相关的特征.相比 ICAN 方法,本文 UDAN 方法的细粒度分类准确率取得了 16.0%的提升.这主要是因为本文的 UDAN 方法中提出的联合辨识最大均值差异(JDMMD)准则能够有效地将辨识定位能力从图像数据迁移到视频数据.

(3) 与现有的有监督的视频分类方法的对比.本文也与膨胀三维卷积(inflated 3D ConvNet,简称 I3D)方法进行了对比.I3D 将二维的卷积层膨胀为三维的卷积层,首先利用二维的卷积层进行初始化,之后再利用视频数据进行训练.从表 4 的结果可以看到,即使使用了标注的目标域的视频训练数据,I3D 方法的细粒度视频分类准确率依然比本文 UDAN 方法低,这验证了本文 UDAN 方法能够有效地将知识从标注的图像数据迁移到未标注的视频数据.

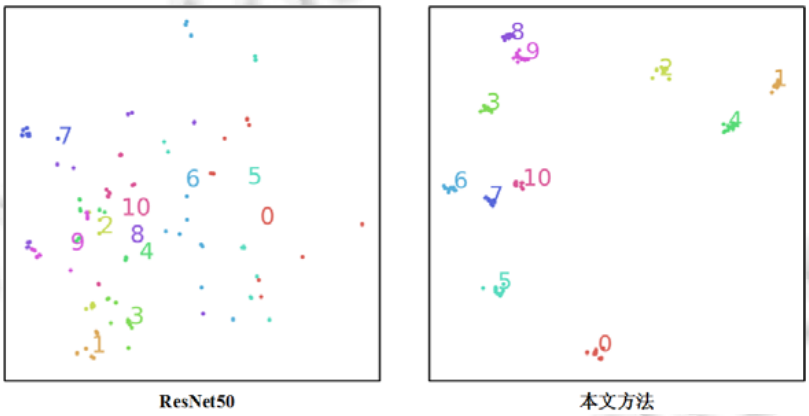


Fig.4 Variation of distribution of target video data, from ResNet50 model to our UDAN approach on CUB-200-2011 and YouTube Birds datasets

图 4 在 CUB-200-2011 和 YouTube Birds 数据集上从 ResNet50 模型到本文 UDAN 方法,目标视频的数据分布变化

从上述 3 个方面的分析对比,本文 UDAN 方法的有效性得到了验证.从 Cars-196 到 YouTube Car 上两种适应任务的结果如表 5 所示.本文的 UDAN 方法与基础网络模型、现有适应方法、现有有监督视频分类方法的对比趋势,与从 CUB-200-2011 到 YouTube Birds 上两种适应任务的结果一致.这进一步验证了本文 UDAN 方法在适应任务上的有效性.在下一节中,通过剥离实验进一步验证其有效性.

### 3.5.2 剥离实验

在本节中,通过剥离实验验证本文 UDAN 方法每个组成部分的效果,以及在渐进式伪标签策略中迭代次数对于细粒度分类效果的影响.

- 基线实验结果分析

为了验证本文 UDAN 方法每个组成部分的效果,在基线方法 JAN[22]上依次加入新提出的联合辨识最大均值差异准则(JDMMD)和渐进式伪标签策略(PL),以 CUB-200-2011 和 YouTube Birds 两个数据集上的实验为例,结果如图 5 所示.

(1) 在基线方法 JAN(表示为 baseline)的基础上加入联合辨识最大均值差异准则(表示为+JDMMD),可以在 $I\rightarrow F$ 和 $I\rightarrow V$ 这两种适应任务中分别提升 2.2%和 8.2%的细粒度分类准确率.这是因为联合辨识最大均值差异准则能够在域适应任务中有效地高亮辨识性区域的特征,同时削弱背景信息的负面影响.

(2) 进一步,加入渐进式伪标签策略(表示为+JDMMD+PL)可以在 $I\rightarrow F$ 和 $I\rightarrow V$ 这两种适应任务中分别再提升 3.8%和 3.7%的细粒度分类准确率.这是因为渐进式伪标签策略可以在未标注的目标域视频数据中选择置信度高的视频数据并打标签,进而利用这些伪标签视频数据进行训练,从而使得 UDAN 模型能够直接从目标域视频数据中获取信息.

- 渐进式伪标签策略的迭代次数影响

在渐进式伪标签策略中,不同的迭代次数对于本文 UDAN 方法的细粒度分类准确率有一定的影响.

以 CUB-200-2011 和 YouTube Birds 这两个数据集上的实验为例,图 6 中的结果验证了渐进式伪标签策略在 $I\rightarrow F$ 和 $I\rightarrow V$ 这两种适应任务中的有效性,以及渐进式地迭代能够进一步提高细粒度分类准确率.但是,细粒度分类的准确率不会始终随着迭代次数的增加而提升,当迭代一定次数后,准确率开始提升得比较平缓甚至出现下降的情况.在本文实验中,为了在 $I\rightarrow F$ 和 $I\rightarrow V$ 这两种适应任务均取得最好的细粒度分类准确率,迭代次数设置为 2.

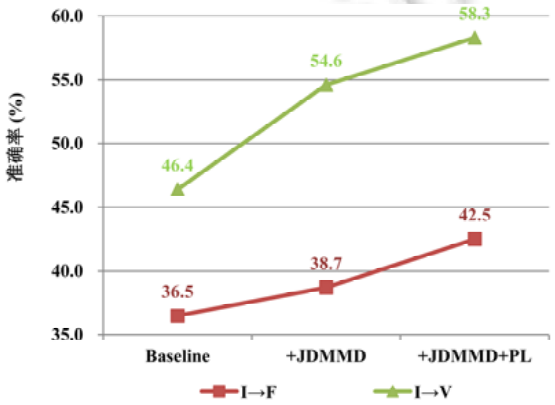


Fig.5 Effect of each component in our UDAN approach

图 5 本文 UDAN 方法中每个组成部分的影响

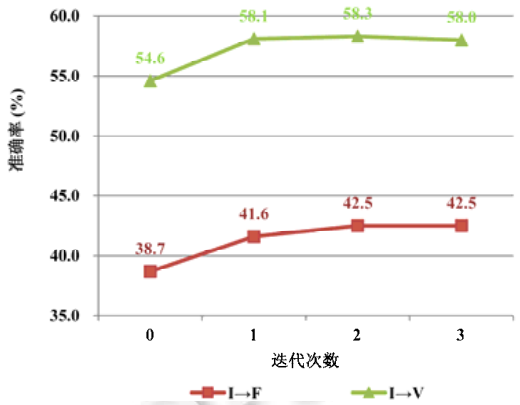


Fig.6 Effect of iteration number in progressive pseudo labeling strategy

图 6 渐进式伪标签策略中迭代次数的影响

## 4 结 论

本文提出了无监督辨识适应网络,能够将知识从源域标注的图像数据迁移到目标域未标注的视频数据.首先,本文提出了联合辨识最大均值差异,将从源域图像数据学习到的辨识性定位能力迁移应用于目标域视频数据;进一步,本文提出了一种渐进式伪标签策略来迭代地引导无监督辨识适应网络来近似目标域视频的数据分布.在实验部分,本文将知识从 CUB-200-2011/Cars-196 图像数据集迁移到 YouTube Birds/YouTube Cars 视频数据集,实验结果证明了本文方法在无监督细粒度视频分类上的优势.

下一步工作主要从以下两个方面展开:(1) 发现更多更精细的辨识性区域,以进一步降低域差异和模态差异;(2) 探索使用少量目标域标注的视频数据对于细粒度视频分类的影响.

**References**:


[1] Zhu C, Tan X, Zhou F, Liu X, Yue KY, Ding ER, Ma Y. Fine-grained video categorization with redundancy reduction attention. In: Proc. of the European Conf. on Computer Vision (ECCV). Berlin: Springer-Verlag, 2018. 139−155.

[2] Torralba A, Efros AA. Unbiased look at dataset bias. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2011. 1521−1528.

[3] Zhang NN, Donahue J, Girshick R, Darrell T. Part-based *r*-CNNs for fine-grained category detection. In: Proc. of the Int'l Conf. on Machine Learning (ICML). New York: ACM, 2014. 834−849.

[4] Krause J, Jin HL, Yang JC, Li FF. Fine-grained recognition without part annotations. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015. 5546−5555.

[5] Xiao TJ, Xu YC, Yang KY, Zhang JX, Peng YX, Zhang Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015. 842−850.

[6] Uijlings JRR, van de Sande KEA, Gevers T, Smeulders AWM. Selective search for object recognition. Int'l Journal of Computer Vision (IJCV), 2013,104(2):154−171.

[7] Fu JL, Zheng HL, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017. 4438−4446.

[8] He XT, Peng YX, Zhao JJ. Stackdrl: Stacked deep reinforcement learning for fine-grained visual categorization. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence (IJCAI). San Francisco: Morgan Kaufmann Publishers, 2018. 741−747.

[9] Lin TY, Chowdhury AR, Maji S. Bilinear CNN models for fine-grained visual recognition. In: Proc. of the Int'l Conf. of Computer Vision (ICCV). Piscataway: IEEE, 2015. 1449−1457.

[10] Gao Y, Beijbom O, Zhang N, Darrell T. Compact bilinear pooling. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016. 317−326.

[11] Cui Y, Zhou F, Wang J, Liu X, Lin YQ, Belongie S. Kernel pooling for convolutional neural networks. In: Proc. of the IEEE Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017. 3049−3058.

[12] Wang YM, Morariu VI, Davis LS. Learning a discriminative filter bank within a CNN for fine-grained recognition. In: Proc. of the IEEE Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2018. 4148−4157.

[13] He XT, Peng YX. Fine-grained image classification via combining vision and language. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017. 5994−6002.

[14] Chen TS, Lin L, Chen RQ, Wu Y, Luo XN. Knowledge-embedded representation learning for fine-grained image recognition. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence (IJCAI). San Francisco: Morgan Kaufmann Publishers, 2018. 627−634.

[15] Saito T, Kanezaki A, Harada T. Ibc127: Video dataset for fine-grained bird classification. In: Proc. of the IEEE Int'l Conf. on Multimedia and Expo (ICME). Piscataway: IEEE, 2016. 1−6.

[16] Kalogeiton V, Ferrari V, Schmid C. Analysing domain shift factors between videos and images for object detection. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI), 2016,38(11):2327−2334.

[17] Ben-David S, Blitzer J, Crammer K, Pereira F. Analysis of representations for domain adaptation. In: Proc. of the Neural Information Processing Systems (NeurIPS). Cambridge: MIT Press, 2007. 137−144.

[18] Gebru T, Hoffman J, Li FF. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In: Proc. of the IEEE Int'l Conf. onComputer Vision (ICCV). Piscataway: IEEE, 2017. 1358−1367.

[19] Cui Y, Song Y, Sun C, Howard A, Belongie S. Large scale fine-grained categorization and domain-specific transfer learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2018. 4109−4118.

[20] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016. 770−778.

[21] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Proc. of the Neural Information Processing Systems (NeurIPS). Cambridge: MIT Press, 2014. 3320−3328.

[22] Long MS, Zhu H, Wang JM, Jordan MI. Deep transfer learning with joint adaptation networks. In: Proc. of the Int'l Conf. on Machine Learning (ICML). New York: ACM, 2017. 2208−2217.

[23] He KM, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV). Piscataway: IEEE, 2017. 2980−2988.

[24] Otsu N. A threshold selection method from gray-level histograms. IEEE Trans. on Systems, Man, and Cybernetics (TCYB), 1979, 9(1):62−66.

[25] Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD birds-200-2011 dataset. Technical Report, CNS-TR-2011-001, Pasadena: California Institute of Technology, 2011.

[26] Krause J, Stark M, Deng J, Li FF. 3D object representations for fine-grained categorization. In: Proc. of the Int'l Conf. of Computer Vision Workshop (ICCVW). Piscataway: IEEE, 2013. 554−561.

[27] Zhang WC, Ouyang WL, Li W, Xu D. Collaborative and adversarial network for unsupervised domain adaptation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2018. 3801−3809.

[28] Saito K, Watanabe K, Ushiku Y, Harada T. Maximum classifier discrepancy for unsupervised domain adaptation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2018. 3723−3732.

[29] Carreira J, Zisserman A. Quo Vadis, action recognition? A new model and the kinetics dataset. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017. 4724−4733.

**何相腾**(1991－),男,博士,主要研究领域为细粒度图像分类,细粒度跨媒体检索,多模态内容理解.

**彭宇新**(1974－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为跨媒体分析与推理,图像视频识别与理解,计算机视觉,人工智能.