第 14 卷 第 1 期
2025 年 1 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 14 No. 1
Jan. 2025

引文格式：
许清林, 乔宇, 王亚立. 基于领域自适应预训练的黑暗场景下行为识别研究 [J]. 集成技术, 2025, 14(1): 25-38.
Xu QL, Qiao Y, Wang YL. Domain-adaptive pretraining for action recognition in the dark scenes [J]. Journal of Integration Technology, 2025, 14(1): 25-38.

基于领域自适应预训练的黑暗场景下行为识别研究

许清林[1,2] 乔 宇[3] 王亚立[1,3*]

[1]（中国科学院深圳先进技术研究院 深圳 518055）
[2]（中国科学院大学 北京 100049）
[3]（上海人工智能实验室 上海 200232）

摘 要 黑暗场景与传统预训练模型所使用的数据之间的域差距导致传统的预训练-微调策略难以达到理想效果，而从头开始的预训练则代价高昂。针对此问题，该研究提出一种领域自适应预训练方法，旨在改善黑暗场景下的行为识别性能。该方法通过融合外部视觉去暗增强模型，引入关键的去暗知识，并利用跨领域自蒸馏框架优化预训练模型，可有效减小明暗场景间视觉表征的域差异。在一系列黑暗场景行为识别实验中，该方法在全监督的黑暗场景行为识别数据集中的准确率达 97.19%；在无源领域自适应场景数据集中的准确率提升至 49.11%；而在多源领域自适应场景数据集中的准确率达 54.63%。

关键词 黑暗场景；行为识别；迁移学习；领域自适应
中图分类号 TP183 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20231225001

Domain-Adaptive Pretraining for Action Recognition in the Dark Scenes

XU Qinglin[1,2] QIAO Yu[3] WANG Yali[1,3*]

[1](*Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)
[3](*Shanghai Artificial Intelligence Laboratory*, *Shanghai* 200232, *China*)
[*]*Corresponding Author*: *yl.wang@siat.ac.cn*

Abstract The domain gap between dark scenes and the data used by traditional pretrained models leads to suboptimal performance with the conventional pretrain-finetune approach, and pretraining from scratch is costly. To address this issue, a domain-adaptive pretraining method is proposed to improve action recognition

收稿日期：2023-12-25 **修回日期**：2024-03-04
基金项目：国家重点研发计划项目（2022ZD0160505）；国家自然科学基金项目（62272450）
作者简介：许清林，硕士研究生，研究方向为计算机视觉、行为识别和多模态学习等；乔宇，博士，研究员，研究方向为计算机视觉、深度学习、行为识别、场景识别、人脸识别和目标检测等；王亚立（通讯作者），博士，研究员，研究方向为计算机视觉、深度学习和行为识别等，E-mail：yl.wang@siat.ac.cn。

performance in the dark environments. The method integrates an external vision enhancement model for de-darkening to introduce critical knowledge for dark scene processing. It also employs a cross-domain self-distillation framework to reduce the domain gap of visual representations between illuminated and dark scenes. Through extensive experiments in various dark environment action recognition settings, the proposed approach can achieve a Top1 accuracy of 97.19% on the dark dataset of fully supervised action recognition. In the source-free domain adaptation on the Daily-DA dataset, the accuracy can be improved to 49.11%. In the multi-source domain adaptation scenario on the Daily-DA dataset, the Top1 accuracy can reach 54.63%.

Keywords dark scenes; action recognition; transfer learning; domain adaptation

Funding This work is supported by National Key Research and Development Program of China (2022ZD0160505), National Natural Science Foundation of China (62272450)

1 引 言

随着深度学习技术的不断深化和发展，计算机视觉已在图像分类[1-3]、行为识别[4-6]、视频检索[7-8]和图像生成[9-10]等多个领域取得重大突破。在众多研究挑战中，行为识别任务因其复杂性而被公认为是一项极具挑战的课题。近年来，借助 3D 卷积神经网络[11-13]和视觉变换器(vision Transformers，ViT)[14-16]等先进模型，行为识别技术已实现质的飞跃。尽管如此，在光照条件较差的场景下获得鲁棒的动作表示仍是一个棘手的难题，尤其是在弱光环境下进行准确的行为识别。此外，如何将在光照充足的环境中训练好的模型迁移到低照度条件下，尤其是在缺少有监督数据的情况下进行领域自适应，亦成为现实世界亟待解决的问题。在黑暗场景下进行行为识别的关键是如何减小预训练数据与目标域数据之间的领域差异带来的影响。常规做法是使用预训练模型，并在目标域数据集上进行微调，但标准的微调范式因域差异而难以达到理想效果。

为缓解黑暗场景对模型识别能力的负面影响，本文初步考虑采用视觉去暗增强模型。然而，去暗增强模型的能力有限，生成的增强视图可能无法达到预期效果，甚至存在引入额外噪声的情况，因此不能完全消除黑暗场景对模型识别能力的影响。为有效解决由领域差异造成的问题，并最大化利用视觉增强模型引入的黑暗域知识，提升模型在黑暗场景中的识别性能，本文提出一种领域自适应预训练策略。该策略采用跨领域自蒸馏学习方法，在已被大规模数据预训练过的模型的基础上，实施进一步预训练，旨在缩小黑暗场景与正常光照场景之间的视觉表征差异，增强模型在黑暗场景下进行行为识别的鲁棒性。具体而言，本文方法先利用视觉去暗增强模型处理原始视图，从而生成增强视图，并结合原始视图和增强视图在补丁级别上获取混合视图，然后进行跨领域知识蒸馏，旨在实现不同领域视图之间的一致性，减少明暗场景之间的视觉表征偏差。此外，本实验观察到，直接进行后续预训练可能会破坏模型原有的泛化能力，导致性能下降。为在知识蒸馏过程中保持模型原有能力的同时学习目标领域的知识，本实验在自蒸馏过程中冻结了视觉骨干网络，并设计了一种渐进学习适配器，以协助模型在进一步的预训练中学习新的领域知识。这种渐进学习适配器通过收集每个骨干层的多层次特征，为跨领域自蒸馏学习提供更

多的可学习表征。

本文的主要内容如下：(1)利用跨领域自蒸馏学习方法，有效提升模型在黑暗场景下进行行为识别的能力；(2)提出渐进学习适配器，使模型能在后预训练的过程中保持模型原有的能力，并注入目标领域知识；(3)在黑暗场景下进行广泛的行为识别实验，并在全监督和领域自适应的基准中均取得了显著的结果。

2　国内外研究现状

2.1　视频理解

卷积神经网络(convolutional neural networks，CNN)[17]显著推进了图像识别领域的发展，为基于 CNN 的各种视频理解任务开辟了新的发展路径。在这些视频理解方法中，双流架构[12-13,18]和三维 CNN 模型[4,11,19]逐渐发展成为两大主流方向。特别地，三维 CNN 在视频处理方面应用广泛，并因此受到重视，尽管它们伴随着较高的计算成本。为应对此问题，并提高效率，研究人员开始探索一些技术，如空间和时间卷积分解[20-21]，以及在二维 CNN 架构中引入时间模块[22-23]。近年来，ViT[24]崭露头角，已成为图像识别领域的主要趋势[5,14]，并广泛应用于视频理解。最近的许多研究将近期涌现的图像基础模型[7-8]应用于视频领域，取得了令人瞩目的成效。然而，由于预训练数据与低照度数据之间存在显著差异，因此，这些强大的视频模型在低照度条件下的性能往往不尽如人意。针对此问题，本文提出一种有效的迁移学习策略，旨在提升视频骨干模型在低照度条件下进行行为识别的适应能力。

2.2　自监督学习

在有监督的表征学习领域，模型往往专注于输入数据与其对应标签之间的关联性，忽视了视频数据的内在结构。相比之下，自监督学习策略则更注重挖掘视频数据的内部结构。初期的视频自监督学习方法主要依赖设计代理任务，这些任务关注视频的内在特征，可实现无监督学习。例如：AoT 利用视频帧间目标特征的连贯一致性进行自监督学习[25]；Fernando 等[26]通过判断帧序列的正确性实施自监督学习，其中正确排序的帧序列被视为正样本，而乱序的帧序列则作为负样本，从而使模型学习视频的时序结构。随着对比学习方法的崛起，自监督预训练领域取得了显著进展。He 等[27]提出一种高效的对比学习框架，利用动态记忆库存储负样本，解决了样本特征不一致的问题。MAE[28]的兴起使得掩码学习在视觉领域的应用成为可能。MAE 将图像分割成补丁，并使用编解码器架构重建被掩盖的区域，以学习图像不同区域的上下文信息。VideoMAE[29]首次将掩码重构的方法应用于视频自监督学习任务，与图像不同，视频包含更多帧，提供了更丰富的动态信息和更多的信息冗余。因此，VideoMAE 提出一个针对视频的高效掩码重构自监督方法，以最大限度地利用视频数据的这些独特属性。然而，对于黑暗场景等特定领域，由于数据分布差异较大，直接使用预训练模型进行微调通常难以取得理想效果，而从头开始针对性地预训练新模型，又会消耗大量计算资源和时间成本。因此，在自然语言处理(NLP)领域，Gururangan 等[30]探讨了跨多个领域的渐进式自监督预训练。本文专注于计算机视觉领域，探索更有针对性的预训练任务设计和更高效的模型参数训练策略。

2.3　领域自适应

由于数据标注的成本较高和数据隐私问题日益严峻，因此，领域自适应(domain adaptation)被广泛关注，并迅速发展。在领域自适应研究中，本文聚焦于无源视频领域自适应(source-free video domain adaptation，SFVDA)和多源视频领域自适应(multi-source video domain adaptation，

MSVDA）两个关键子领域。

SFVDA 主要研究源域数据的隐私性和数据传输所需耗费的资源等问题，以及源域数据在迁移学习的过程中，不能被使用时的域迁移问题。在目前的无源域适应方法中，3C-GAN[31]和 SDDA[32]通过使用 GAN[33]生成与目标域数据分布相似的带有标签的图片，然后基于对抗的域自适应方法将新的目标风格数据与原始目标数据对齐，以获得域不变特征。SHOT[34]通过冻结源分类器利用源特征分布的知识，并通过信息熵最大化和伪标签将目标域的特征匹配到源分类器。

在 MSVDA 中，MDAN[35]为多源域适应下的分类和回归问题提供了平均情况下的泛化边界，并通过对抗学习实现目标域与源域的全局对齐。M3SDA[36]提供了多种复杂的对抗训练策略，并引入了一个模型，用于匹配源特征和目标特征分布矩阵。Most[37]引入了基于 optimal transport 的方法，有效促进了领域自适应过程中模仿学习的效果，其中教师分类器完美地利用源域的知识进行处理，而学生分类器则努力模仿教师分类器在源域中的行为。

与上述方法不同，本文利用外部的去暗知识进行领域自适应的预训练学习。同时，本文探索了多模态预训练模型在领域自适应中的潜在能力。

3 领域自适应预训练

针对黑暗场景视频数据与常规预训练视频数据之间的差异，本文方法旨在利用外部开源的去暗视觉增强模型，将黑暗环境下的视觉知识有效整合到视觉编码器中。然而，如何高效利用这些经过去暗处理增强后的知识，需要进行深入的研究与探索。直观上，可以直接使用经过去暗处理的视频进行识别，但增强模型的局限性和引入的噪声是不容忽视的问题。此外，仅依赖去暗之后的视频可能导致模型忽视不同视图间的内在联系。

在自然语言处理领域，后预训练策略已证明其可以显著提升模型性能[30]。研究指出，通过在目标域上再次实施预训练，可有效地将通用预训练模型适应到特定应用领域，从而进一步提高模型的适用性和效能[30]。基于此，本文整合了视觉去暗增强模型和简单的图像处理技术，构建了 4 种不同类型的视图：原始视图、增强视图、混合视图和局部视图。本文方法通过在跨域和同域场景中实施自蒸馏的自监督学习策略，致力于学习不同视图之间的内在一致性。在跨域一致性的自蒸馏学习中，本研究团队的目标是确保不同光照条件下的视觉表征一致性，以减少低光照场景对视觉识别的影响。对于域内一致性学习，本文专注于掌握同一动作在运动和空间变化上的一致性。

3.1 多视图生成

本节详细介绍了用于跨领域自蒸馏学习的 4 种视图生成过程，如图 1 所示。原始视图指视频的初始表征形式，视频帧以固定频率均匀采样，每帧保留完整的空间信息，空间分辨率为 $H=W=224$。其中 H 和 W 分别为视频帧的长度和宽度。增强视图指在原始视图上使用去暗视觉增强模型[38]后得到的视频帧。混合视图是原始视图和增强视图的混合。首先，仿照 ViT 的处理方式，原始视图和增强视图被划分为不重叠的补丁 $\boldsymbol{X}\in\mathbb{R}^{\frac{H}{P}\times\frac{H}{P}\times\left(P^2\cdot C\right)}$。其中，$P$ 为正方形补丁的边长；C 为视频帧中的通道数。其次，在保持 0 和 1 的数量差不大于 1 的设定下，生成一个随机掩码 $\boldsymbol{M}\in\{0,1\}^{\frac{H}{P}\times\frac{H}{P}}$（由 0 和 1 组成的矩阵）。最后，根据掩码 $\boldsymbol{M}$ 生成混合视图：

$$\boldsymbol{X}_{\mathrm{t}}=\boldsymbol{M}\odot\boldsymbol{X}_{\mathrm{o}}+\left(1-\boldsymbol{M}\right)\odot\boldsymbol{X}_{\mathrm{e}} \tag{1}$$

其中，$\odot$ 为矩阵对应元素乘积；$\boldsymbol{X}_{\mathrm{o}}$ 和 $\boldsymbol{X}_{\mathrm{e}}$ 分别为原始视图和增强视图生成的补丁。局部视图是视频的局部表征，本文视频帧的采样策略与原始

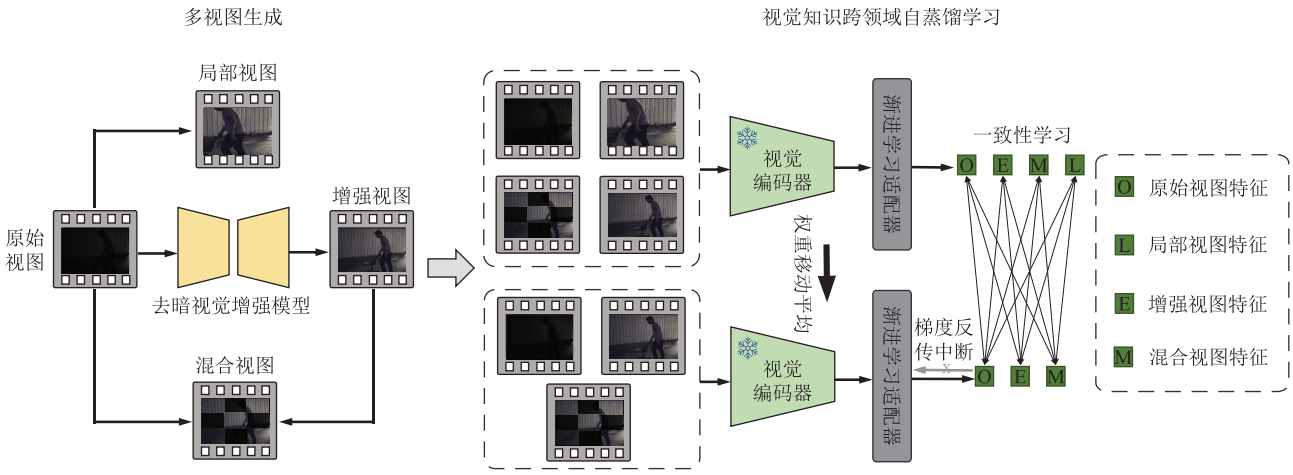

图 1　领域自适应预训练流程

Fig. 1　Process of domain-adaptive pretraining

视图一致，但选取结果可能不同，空间分辨率为 $H=W=96$。

3.2　渐进学习适配器

研究表明，在跨领域视觉知识自蒸馏学习的过程中，直接对预训练模型进行后续预训练会损害其原有的泛化能力，从而降低识别效果。因此，为在知识蒸馏过程中既保留模型原有的能力，又获取新的领域知识，本文提出了渐进学习适配器模块。在跨领域视觉知识自蒸馏学习过程中，视觉主干被冻结，仅对新引入的渐进学习适配器模块进行训练。适配器能从每个视觉骨干层中收集各层次的特征，为知识蒸馏提供更广泛的可训练特征集，从而促进模型学习目标领域的知识。如图 2 所示，本文首先引入了一组可学习的适应提示向量 $\boldsymbol{P}_0=\{p_1,p_2,\cdots,p_N\}$，其中，$N$ 为适应提示的数量；其次，从视觉编码器的每一层收集特征，并通过交叉自注意力机制利用适应提示捕捉视觉上的中间特征；最后，将适应提示作为查询，而从视觉主干提取的中间特征则作为键和值。交叉自注意力机制的操作如式(2)～式(4)所示：

$$\boldsymbol{Y}_i=\eta\left[\left(\boldsymbol{X}_{i,1},\boldsymbol{X}_{i,2},\cdots,\boldsymbol{X}_{i,T}\right)\right] \tag{2}$$

$$\tilde{\boldsymbol{P}}_i=\boldsymbol{P}_{i-1}+\alpha_i\left(\boldsymbol{P}_{i-1},\boldsymbol{Y}_i,\boldsymbol{Y}_i\right) \tag{3}$$

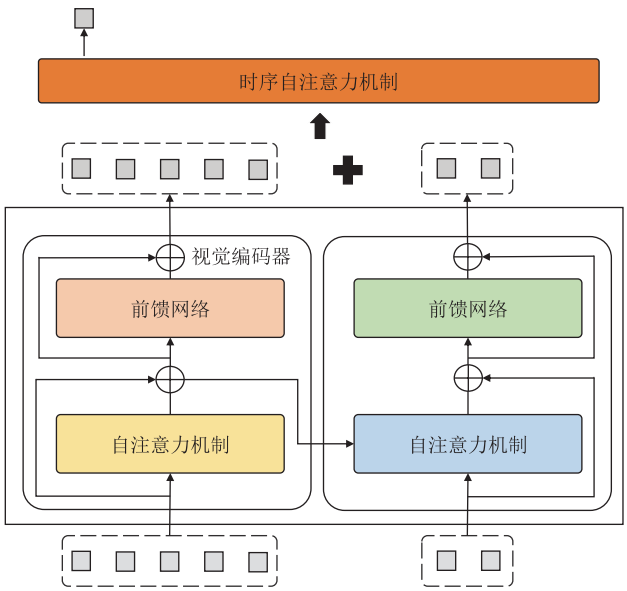

图 2　渐进学习适配器框架

Fig. 2　Framework of progressive learning adaptation

$$\boldsymbol{P}_i=\tilde{\boldsymbol{P}}_i+\phi_i\left(\tilde{\boldsymbol{P}}_i\right) \tag{4}$$

其中，$\boldsymbol{Y}_i$ 为第 i 层视觉编码器的视频特征经过归一化处理后得到的结果；η 为 layer normalization 操作；$\boldsymbol{X}_{i,t}$ 为第 i 层视觉编码器捕获视频的第 t 帧的视觉特征，T 为视频的总帧数；$\boldsymbol{P}_i$ 为经过 i 次交叉注意力机制后得到适应提示的参数，其初始值为 $\boldsymbol{P}_0$，$\tilde{\boldsymbol{P}}_i$ 为自注意力机制的中间计算结果；α_i 为第 i 层的多头注意力机制，包括查询、键和值

3 个参数；ϕ 为全连接层。之后，渐进学习适配器通过结合适应提示和视觉骨干特征共同处理相关信息，并理解不同帧之间的上下文信息，表示如下：

$$\boldsymbol{R}=a\left(\boldsymbol{Y}_K \oplus \boldsymbol{P}_K\right) \tag{5}$$

其中，K 为视觉编码器的总层数；$\boldsymbol{Y}_K$ 和 $\boldsymbol{P}_K$ 分别为视觉编码器最后一层输出的视觉特征和适应提示向量经过最后一次交叉注意力机制后得到适应提示的参数；$\oplus$ 为向量间的拼接操作；a 为学习不同帧之间的时序信息方法。具体来说，首先，由可学习参数[24]生成的位置编码嵌入到 $\boldsymbol{Y}_n$ 和 $\boldsymbol{P}_n$ 的拼接中，接着，这一拼接向量被送入多头自注意力机制中，以获得经过时序上下文加工后的视觉特征：$\boldsymbol{R}=\left\{\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_{T+M}\right\}$。在获取到不同帧的特征 $\boldsymbol{R}$ 之后，聚合每一帧的信息，得到最终的视频特征 $\boldsymbol{f}$：

$$\boldsymbol{f}=\mu\left(\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_{T+M}\right) \tag{6}$$

其中，μ 表示对所有的特征向量进行平均求和。

3.3 训练网络

本文采用自蒸馏方式进行领域自适应预训练学习，通过最小化同一视频在学生模型和教师模型特征空间中不同视图的表征差异来实现，旨在通过学习跨域一致性(匹配不同光照环境下的视图)和域内一致性(匹配局部视角和全局视角)，使模型能更全面地理解视频的潜在分布，减少视觉表征在黑暗与光照场景之间的域差异，降低模型对光照变化的敏感性，确保其在不同光照和视角下保持性能稳定。在跨领域知识蒸馏方面，将教师模型输入设定为原始视图、增强视图和混合视图，而学生模型则包含 4 种视图。在获取了各视图的特征 $\boldsymbol{f}$ 后，先对其进行标准化：

$$p[i]=\frac{\exp\left(\boldsymbol{f}[i]/\tau\right)}{\sum_{i=1}^{n}\exp\left(\boldsymbol{f}[i]/\tau\right)} \tag{7}$$

其中，τ 为温度参数；n 为特征向量的维度。由此可得标准化后的特征：p_{o}、p_{e}、p_{m} 和 p_{l}，分别对应原始视图、增强视图、混合视图和局部视图。之后通过最小化两个视图的交叉熵损失，确保同一视频不同视图之间的一致性，表示如下：

$$L=\sum_{x\in\{\mathrm{o,e,l}\}}\sum_{z\in\{\mathrm{o,e,m,l}\}}-p_{\mathrm{t}x}\cdot\log\left(p_{\mathrm{s}z}\right) \tag{8}$$

其中，L 表示损失；$p_{\mathrm{t}x}$ 和 $p_{\mathrm{s}z}$ 分别为教师模型和学生模型输出的标准化后的特征；x 和 z 分别为视图选择的输入类型。与常规的知识蒸馏架构不同，本文采用自蒸馏策略，学生和教师模型拥有相同的结构和初始参数。为避免模型因教师和学生模型输出相同结果而发生崩溃，在领域自适应预训练过程中，本文冻结了教师模型的参数，并通过权重移动平均(EMA)的方式更新参数：

$$\theta_{\mathrm{t}}\leftarrow\lambda\theta_{\mathrm{t}}+\left(1-\lambda\right)\theta_{\mathrm{s}} \tag{9}$$

其中，θ_{t} 和 θ_{s} 是教师和学生的模型参数；$\lambda\in[0,1]$。而学生模型采取梯度更新的方式进行学习。完成领域自适应预训练后，一种能在黑暗场景中鲁棒地提取视觉特征的编码器被成功构建。在全监督设定下，视觉编码器利用黑暗场景下的训练数据进行标准的微调学习；而在无源和多源领域自适应设定中，本文先利用预训练模型[39]提取无监督数据的伪标签，再利用这些伪标签进行迁移学习。整个学习流程遵循预训练-领域自适应预训练-微调的模式，有效实现了黑暗场景下的跨领域行为识别。

4 结 果

4.1 模型框架和实现细节

在 ARID[40]的全监督基准上，本文将 CLIP[39]中的 BERT[41]作为文本编码器，其由 12 层、512 维的 Transformer[42]和 8 个注意力头组成。本文采用了与 DarkLight[43]相同的 R(2+1)D-34[21]视觉编码器。在 SFVDA 和 MSVDA 的设定中，本文再次使用 CLIP[39]中的文本编码器进行文本特征提取。除延续使用 CLIP 文本编码器

外，还首次将其 ViT-B/16 视觉编码器应用于视频领域适应任务，深入探索了多模态预训练模型在视频领域的迁移能力。在渐进式适应模块中，本文将视频级别的 token 数设为 2，并在 R(2+1)D-34 中使用 2 层时间自注意力机制，而在 CLIP ViT-B/16 中使用 6 层。

在第一阶段的渐进式自我预训练中，本文使用 AdamW[44]优化器进行正则化，学习率为 5×10^{-5}，批大小为 128，权重衰减设为 0.04。在训练期间，本文冻结视觉骨架，并在 CLIP ViT-B/16 中稀疏采样 8 帧视频，在 R(2+1)D-34 中稀疏采样 64 帧视频。其中，去暗增强模型使用 SCI[38]，其仅使用简单的架构即可完成数据增强的目的。式(7)中的 τ 遵循了 DINO[39]的方法，将 τ 的值固定在 0.1。式(9)中的 λ 使用了 DINO 中的设定，初始值设为 0.996，目标值调整为 1。为实现 λ 值的平滑变化，本文利用余弦曲线调度缓慢增加 λ 值。

在微调阶段，本文采用 AdamW[44]优化器(基础学习率为 5×10^{-6})微调预训练的参数，同时使用学习率为 5×10^{-5} 的新模块初始化可学习参数。微调阶段的权重衰减设为 0.2，帧采样设置与第一阶段相同，批大小为 32。所有实验均在使用 8 个 A6000 GPU 的 PyTorch 上进行。

4.2　全监督黑暗场景下行为识别

4.2.1 数据集介绍

实验中，本研究团队对首个黑暗场景下的行为识别基准数据集进行实验，即 ARID[40]。该数据集包含 3 784 个视频片段，涵盖 11 个不同的动作类别，为增强实验结果的鲁棒性，数据集被划分为 3 组。本文将报告在 3 组数据集划分计算得到的 Top1 和 Top5 平均准确率。

4.2.2 实验结果

由表 1 可知，与以往的方法相比，本文方法的性能较优，特别是在 Top1 准确率方面，其提高幅度超过 3%。由于 ARID[40]数据集仅包含 11 个类别，所有方法在 Top5 预测中都表现出了较好的效果，因此本文不在实验结果中展示和比较 Top5 的指标。此外，先进的模型，如 Timesformer[5]和 CLIP[39]，虽然在主流行为识别数据集中取得了出色的成绩，但在 ARID 基准上的表现仍不尽如人意，这突显了视频模型在处理重大领域差异时的局限性。然而，这些方法在经过本文的领域自适应预训练后，准确率得到了显著提升，表明本文方法在克服领域差异方面的高效性和即插即用性。

表 1　全监督设定下 ARID 结果

Table 1　Result for fully supervised setting on ARID

模型方法	准确率 (%)	
	Top1	Top5
Timesfomer[5]	66.57	98.31
CLIP[39]	67.15	98.58
R(2＋1)D[43]	94.04	99.87
领域自适应预训练＋Timesfomer[5]	72.86＋6.29	97.12
领域自适应预训练＋CLIP[39]	74.82＋7.67	96.95
领域自适应预训练＋R(2＋1)D[43]	97.19＋3.15	99.59

4.3　多源视频领域自适应

4.3.1 数据集介绍

Daily-DA[45]数据集涵盖了正常光照和黑暗场景下的视频数据，本文将其作为多源视频领域自适应的基准。

该数据集整合了 4 个不同的数据集：ARID[40](A11)、HMDB51[46](H51)、Moments in Time[47](MIT)和 Kinetics-600[48](K600)。HMDB51、Moments in Time 和 Kinetics-600 在行为识别领域得到了广泛应用，而 ARID 则是一个新的黑暗场景下行为识别数据集，由低光照条件下录制的视频组成。在多源视频域适应的情况下，本文将一个数据集作为目标域，而其他 3 个数据集则充当源域。因此，涉及 4 个任务：Daily→A11、Daily→H51、Daily→MIT 和 Daily→K600。此外，一旦从标注的源数据中获得训练有素的模型，则本方法将不再使用源数

据，这与其他方法的做法不同。在呈现实验结果时，对于每种方法，本文使用不同的随机种子进行了 5 次独立实验，并报告了这些实验的平均值和标准差，以展示结果的一致性和可靠性。本文仅展示 Daily→A11 任务，以强调本文模型将正常光照情况下训练得到的模型无监督迁移到黑暗场景下的卓越能力。

4.3.2 实验结果

多源视频领域适应 Daily-DA 的实验结果如表 2 所示。通过比较可知，本文方法在多源视频领域适应任务中的性能较好。与最先进的 TAMAN[45] 方法相比，本文方法的性能提升了 24.68%。另外，本文还呈现了使用相同视觉编码器的 ActionCLIP[54] 的结果，该方法使用相同的方式将伪标签作为监督迁移到目标领域。本文方法在这项任务上取得了 2.52% 的增益，表明本文方法在稳定解决域适应问题方面的性能较好。

表 2 Daily-DA 多源视频领域自适应结果

Table 2 Results for multi-source video domain adaptation on Daily-DA

方法	方法类别	Top1 准确率(%)
		Daily→A11
s-DANN[49]	基于对抗的方法	22.03±0.35
s-ADDA[50]		22.30±0.21
s-TA$_3$N[51]		21.76±0.16
s-ACAN[52]		23.44±0.16
c-DANN[49]		22.15±0.33
c-ADDA[50]		22.65±0.25
c-TA$_3$N[51]		22.24±0.20
c-ACAN[52]		23.95±0.28
MDAN[35]		23.75±0.38
DCTN[53]		24.94±0.36
MDDA[54]		22.73±0.26
s-MMD[55]	基于差异的方法	21.62±0.22
s-MCD[56]		23.80±0.28
s-CORAL[57]		21.51±0.15
c-MMD[55]		24.28±0.36
c-MCD[56]		25.68±0.28
c-CORAL[57]		23.96±0.16
LtC-MSDA[36]		24.98±0.12
MCC[58]		22.65±0.35
MOST[37]		26.28±0.46
M3SDA[59]		24.83±0.23
TAMAN[45]		29.95±0.35
ActionCLIP[60]		52.11±0.99
本文方法		54.63±0.83

4.4 无源视频领域自适应

4.4.1 数据集介绍

无源视频领域自适应 Daily-DA[60] 是一个具有挑战性的数据集，涵盖了正常光照和黑暗场景下的视频数据。Daily-DA 包含的数据集和多源域适应是一样的。无源视频领域自适应 Daily-DA 数据集包含了 18 949 个视频，涵盖了 8 个类别，包括 12 项跨域动作识别任务。本文选择展示以 ARID 为目标域数据集的 3 个跨领域任务，以突显本文方法将正常光照情况下训练得到的模型无监督迁移到黑暗场景下的性能。

4.4.2 实验结果

无源视频领域自适应的实验结果如表 3 所示。与较先进的 ATCoN[66] 方法相比，本文方法的性能较优。为公正地比较，本文还呈现了使用相同视觉编码器的 ActionCLIP[60] 的结果，该方法在源域数据集上进行训练，并通过目标域生成的伪标签实现迁移学习。本文方法在平均准确率上取得了 1.89% 的增益，表明其在无源视频领域自适应方面的性能较优。

4.5 消融实验

4.5.1 领域自适应预训练的作用

如表 4 所示，本文深入研究了所提出的领域自适应预训练方法在减小黑暗场景对行为识别影响方面的效果。实验结果表明，在进行跨领域自蒸馏学习时，当不冻结视觉编码器对视觉编码器进行训练时，与未进行跨领域自蒸馏学习而是直接进行微调的基准方法相比，模型在 3 个设定任务下的结果均较差，表明在目标领域对预训练模型进行后预训练时，对预训练模型的主干网络进行学习可能破坏模型原本的知识。由表 4 可

表 3　Daily-DA 无源视频领域适应的结果

Table 3　Results for source-free video domain adaptation on Daily-DA

方法	是否无源	Top1 准确率(%)			
		K600→A11	MIT→A11	H51→A11	平均值
DANN[49]	×	21.18	22.81	14.20	19.40
MK-MMD[55]	×	21.66	21.02	20.35	21.01
TA_3N[51]	×	19.87	21.57	14.38	18.60
SFDA[61]	√	12.57	15.96	13.08	13.87
SHOT[34]	√	12.03	15.28	13.50	13.60
SHOT＋＋[62]	√	12.57	14.90	15.98	14.48
MA[63]	√	12.76	17.75	12.90	14.47
BAIT[64]	√	12.69	16.93	13.65	14.42
CPGA[65]	√	13.06	18.08	13.14	14.76
ATCoN[66]	√	17.21	27.23	17.92	20.79
ActionCLIP[60]	√	47.89	48.59	45.20	47.22
本文方法	√	49.61	50.86	46.87	49.11

知，在锁住视觉编码器后(使用 ActionCLIP[60]中的 seqTransf 模块，并仅微调该部分)，经过跨领域自蒸馏学习后，模型在 3 个设定下均取得了提升，其中，在 MSVDA 设定下提升了 1.43%。引入渐进式学习适配器后，模型的性能进一步提升，在全监督、SFVDA 和 MSVDA 的设定下分别达到了 97.19%、49.11% 和 54.63% 的准确率，与仅锁住视觉编码器的策略相比，分别提升了 0.36%、0.70% 和 1.09%。

表 4　领域自适应预训练学习的作用

Table 4　The effect of domain-adaptive pretraining

方法	Top1 准确率(%)		
	ARID	SFVDA	MSVDA
基准方法	96.30	47.22	52.11
训练编码器	70.21	32.15	41.23
锁住编码器	96.83	48.41	53.54
渐进学习适配器	97.19	49.11	54.63

4.5.2 与基于 CLIP 的方法比较

本文方法与基于 CLIP[39]的其他方法(ActionCLIP[60]、XCLIP[67]和 FrozenCLIP[68])在 ARID[40]上的全监督设定的比较结果如表 5 所示。在使用相同的 ViT-B/16 主干网络的情况下，与先前的基于 CLIP 的方法相比，本文方法的 Top1 准确率更高，比 XCLIP 高 7.50%。

表 5　与基于 CLIP 的方法的对比

Table 5　Comparison with CLIP-base methods

方法	Top1 准确率(%)	
	主干网络	ARID
ActionCLIP[60]	ViT-B/16	67.15
XCLIP[67]	ViT-B/16	67.32
FrozenCLIP[68]	ViT-B/16	66.22
本文方法	ViT-B/16	74.82

4.6　结果可视化分析

如图 3 所示，本文从两个不同的视频片段中分别抽取了 4 帧，以分析行为识别模型的性能(其中：喝水行为见第 1～4 列，推动物体行为见第 5～8 列)。通过平均采样方法得到的视频帧序列如第 1 行所示。为确保所有细节均可清晰观察，本文对这些原始帧进行了亮度增强处理，如第 2 行所示。采用领域自适应预训练方法前后的 Grad-CAM 可视化结果如第 3 行和第 4 行所示。可以观察到，在没有进行领域自适应预训练的情况下，模型很难在黑暗中聚焦到关键物体。然而，通过领域自适应预训练后，模型能将注意力集中在关键信息上，如被推的物体和手中的杯子。

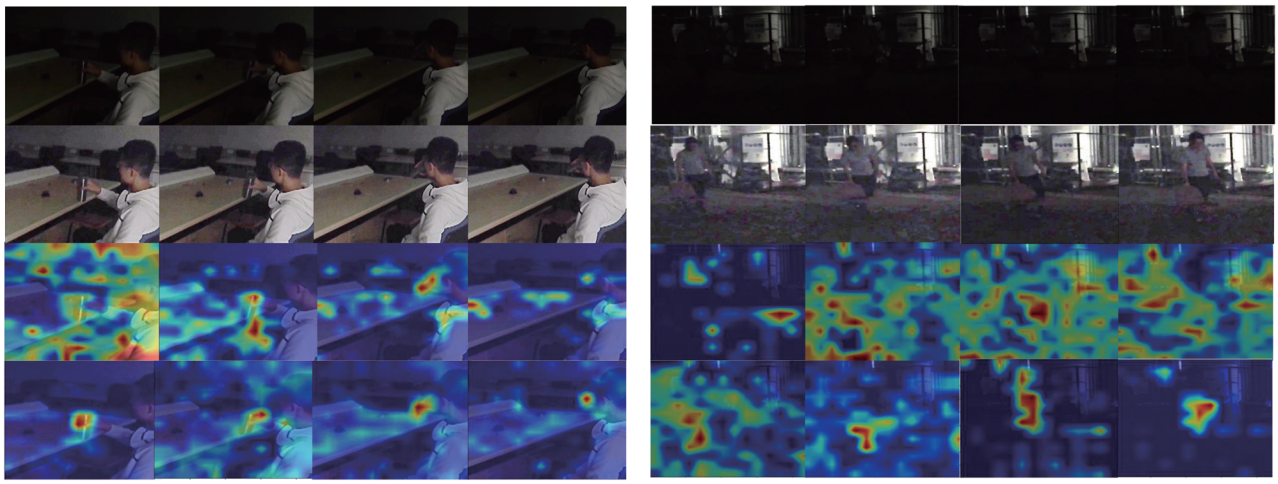

图 3 Grad-CAM 生成的视频注意力可视化

Fig. 3 The attention visualization of video generated by Grad-CAM

5 讨论与分析

目前，在黑暗场景下进行视频理解是一项具有挑战性的任务，黑暗环境导致视频数据中很多重要的细节容易被模型忽视。同时，由于视觉模型预训练数据与目标域黑暗场景数据存在域差异，因此，传统的预训练-微调范式并不能取得良好的效果，CLIP[60]在全监督设定下对黑暗场景行为识别 ARID 数据集的 Top1 准确率仅为 67.15%。融合了本文方法后，CLIP 方法的 Top1 准确率提升至 74.82%，提升了 7.67%。同时，本方法在选择视觉编码器为 R(2+1)D[40]的情况下，达到了 97.19% 的 Top1 准确率，与之前的方法相比，本文方法实现了 3.15% 的显著提升。此外，本文方法还适用于领域自适应，融合了本文方法后，在无源域适应和多源域适应设定下的 Daily-DA 数据集上，取得了 49.11% 和 54.63% 的 Top1 准确率，分别提升了 1.89% 和 2.52%。由此可知，本文的跨领域自蒸馏方法能有效减少不同光照条件下特征的差异，提升黑暗场景下模型视频理解的能力。本文提出一种针对黑暗场景行为识别的有效方法，采用预训练-后预训练-微调的迁移学习范式，通过在目标领域对通用预训练模型进行再次预训练，可有效减少目标领域与预训练数据之间的域差异。与从头开始预训练相比，后预训练策略的计算资源和时间投入虽然大量减少，但与 ATCoN[66]等方法相比，仍需要额外的后预训练步骤。

6 结 论

面对黑暗场景下行为识别的挑战，本文提出了基于领域自适应预训练的方法。该方法从外部的去暗增强模型获取去暗知识，进而获得正常光照场景下的视图。通过跨领域自蒸馏学习策略，使模型可以学习提取不同光照环境下行为的一致性表征，以减少领域差异带来的影响。通过在 3 个设定下进行黑暗场景行为识别实验，本文验证了基于领域自适应预训练方法的有效性。在未来的工作中，本研究团队将关注更广泛的开放世界场景，设计适用于各种实际情景的预训练模型，并提出统一有效的迁移策略。

参 考 文 献

[1] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [Z/OL].

arXiv Preprint, arXiv: 1409.1556, 2014.

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

[3] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[4] Wang LM, Xiong YJ, Wang Z, et al. Temporal segment networks for action recognition in videos [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(11): 2740-2755.

[5] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? [C] // Proceedings of the 38th International Conference on Machine Learning, 2021: 1-3.

[6] Li KC, Wang YL, Gao P, et al. UniFormer: unified Transformer for efficient spatiotemporal representation learning [Z/OL]. arXiv Preprint, arXiv: 2201.04676, 2022.

[7] Luo HS, Ji L, Zhong M, et al. CLIP4Clip: an empirical study of CLIP for end to end video clip retrieval and captioning [J]. Neurocomputing, 2022, 508: 293-304.

[8] Wu WH, Luo HP, Fang B, et al. Cap4Video: what can auxiliary captions do for text-video retrieval? [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 10704-10713.

[9] Chang HW, Zhang H, Barber J, et al. Muse: text-to-image generation via masked generative Transformers [Z/OL]. arXiv Preprint, arXiv: 2301.00704, 2023.

[10] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents [Z/OL]. arXiv Preprint, arXiv: 2204.06125, 2022.

[11] Wang XL, Girshick R, Gupta A, et al. Non-local neural networks [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7794-7803.

[12] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1933-1941.

[13] Feichtenhofer C, Fan HQ, Malik J, et al. SlowFast networks for video recognition [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6202-6211.

[14] Liu Z, Ning J, Cao Y, et al. Video Swin Transformer [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 3202-3211.

[15] Arnab A, Dehghani M, Heigold G, et al. ViViT: a video vision Transformer [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 6836-6846.

[16] Li KC, Wang YL, He YN, et al. UniFormerV2: spatiotemporal learning by arming image ViTs with video UniFormer [Z/OL]. arXiv Preprint, arXiv: 2211.09552, 2022.

[17] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks [C] // Proceedings of the Advances in Neural Information Processing Systems, 2012, 25.

[18] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [C] // Proceedings of the Advances in Neural Information Processing Systems, 2014: 568-576.

[19] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks [C] // Proceedings of the IEEE International Conference on Computer Vision, 2015: 4489-4497.

[20] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6450-6459.

[21] Qiu ZF, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks [C] // Proceedings of the IEEE International Conference on Computer Vision, 2017: 5533-5541.

[22] Li Y, Ji B, Shi XT, et al. TEA: temporal excitation and aggregation for action recognition [C] //

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 909-918.

[23] Liu ZY, Wang LM, Wu W, et al. TAM: temporal adaptive module for video recognition [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 13708-13718.

[24] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [Z/OL]. arXiv Preprint, arXiv: 2010.11929, 2010.

[25] Wei DL, Lim JJ, Zisserman A, et al. Learning and using the arrow of time [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8052-8060.

[26] Fernando B, Bilen H, Gavves E, et al. Self-supervised video representation learning with odd-one-out networks [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3636-3645.

[27] He KM, Fan HQ, Wu YX, et al. Momentum contrast for unsupervised visual representation learning [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 9729-9738.

[28] He KM, Chen XL, Xie SN, et al. Masked autoencoders are scalable vision learners [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16000-16009.

[29] Tong Z, Song YB, Wang J, et al. VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training [C] // Proceedings of the Advances in Neural Information Processing Systems, 2022: 10078-10093.

[30] Gururangan S, Marasović A, Swayamdipta S, et al. Don’t stop pretraining: adapt language models to domains and tasks [Z/OL]. arXiv Preprint, arXiv: 2004.10964, 2020.

[31] Kurmi VK, Subramanian VK, Namboodiri VP. Domain impression: a source data free domain adaptation method [C] // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 615-625.

[32] Li R, Jiao QF, Cao WM, et al. Model adaptation: unsupervised domain adaptation without source data [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 9641-9650.

[33] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139-144.

[34] Liang J, Hu DP, Feng JS. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation [C] // Proceedings of the 37th International Conference on Machine Learning, 2020: 6028-6039.

[35] Scalbert M, Vakalopoulou M, Couzinié-Devy F. Multi-source domain adaptation via supervised contrastive learning and confident consistency regularization [Z/OL]. arXiv Preprint, arXiv: 2106.16093, 2021.

[36] Peng XC, Bai QX, Xia XD, et al. Moment matching for multi-source domain adaptation [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1406-1415.

[37] Nguyen T, Le T, Zhao H, et al. Most: multi-source domain adaptation via optimal transport for student-teacher learning [C] // Proceedings of the Uncertainty in Artificial Intelligence, 2021: 225-235.

[38] Ma L, Ma TY, Liu RS, et al. Toward fast, flexible, and robust low-light image enhancement [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5637-5646.

[39] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] // Proceedings of the 38th International Conference on Machine Learning, 2021: 8748-8763.

[40] Xu YC, Yang JF, Cao HZ, et al. ARID: a new dataset for recognizing action in the dark [C] // Proceedings of the Deep Learning for Human Activity Recognition, 2021: 70-84.

[41] Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [Z/OL]. arXiv Preprint, arXiv: 1810.04805, 2018.

[42] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Proceedings of the 31st Conference on Neural Information Processing Systems, 2017: 30.

[43] Chen R, Chen JJ, Liang ZX, et al. DarkLight networks for action recognition in the dark [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 846-852.

[44] Loshchilov I, Hutter F. Fixing weight decay regularization in adam [Z/OL]. arXiv Preprint, arXiv: 1711.05101, 2018.

[45] Xu YC, Yang JF, Cao HZ, et al. Multi-source video domain adaptation with temporal attentive moment alignment network [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 3860-3871.

[46] Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition [C] // Proceedings of the 2011 International Conference on Computer Vision, 2011: 2556-2563.

[47] Monfort M, Andonian A, Zhou BL, et al. Moments in time dataset: one million videos for event understanding [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(2): 502-508.

[48] Carreira J, Noland E, Banki-Horvath A, et al. A short note about Kinetics-600 [Z/OL]. arXiv Preprint, arXiv: 1808.01340, 2018.

[49] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation [C] // Proceedings of the 32nd International Conference on Machine Learning, 2015: 1180-1189.

[50] Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7167-7176.

[51] Chen MH, Kira Z, AlRegib G, et al. Temporal attentive alignment for large-scale video domain adaptation [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6321-6330.

[52] Xu YC, Cao HZ, Mao KZ, et al. Aligning correlation information for domain adaptation in action recognition [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(5): 6767-6778.

[53] Xu RJ, Chen ZL, Zuo WM, et al. Deep cocktail network: multi-source unsupervised domain adaptation with category shift [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3964-3973.

[54] Zhao SC, Wang GZ, Zhang SH, et al. Multi-source distilling domain adaptation [C] // Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020: 12975-12983.

[55] Long MS, Cao Y, Wang JM, et al. Learning transferable features with deep adaptation networks [C] // Proceedings of the 32nd International Conference on Machine Learning, 2015: 97-105.

[56] Saito K, Watanabe K, Ushiku Y, et al. Maximum classifier discrepancy for unsupervised domain adaptation [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3723-3732.

[57] Sun BC, Feng JS, Saenko K. Return of frustratingly easy domain adaptation [C] // Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016: 2058-2065.

[58] Wang H, Xu MH, Ni BB, et al. Learning to combine: knowledge aggregation for multi-source domain adaptation [C] // Proceedings of the European Conference on Computer Vision, 2020: 727-744.

[59] Jin Y, Wang XM, Long MS, et al. Minimum class confusion for versatile domain adaptation [C] // Proceedings of the European Conference on Computer Vision, 2020: 464-480.

[60] Wang MM, Xing JZ, Liu Y. ActionCLIP: a new paradigm for video action recognition [Z/OL]. arXiv Preprint, arXiv: 2109.08472, 2021.

[61] Kim Y, Cho D, Han K, et al. Domain adaptation without source data [J]. IEEE Transactions on Artificial Intelligence, 2021, 2(6): 508-518.

[62] Liang J, Hu DP, Wang YB, et al. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 8602-8617.

[63] Yang SQ, Wang YX, van de Weijer J, et al. Unsupervised domain adaptation without source data by casting a BAIT [Z/OL]. arXiv Preprint, arXiv: 2010.12427, 2020.

[64] Agarwal P, Paudel DP, Zaech JN, et al. Unsupervised robust domain adaptation without source data [C] // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 2009-2018.

[65] Qiu Z, Zhang YF, Lin HB, et al. Source-free domain adaptation via avatar prototype generation and adaptation [Z/OL]. arXiv Preprint, arXiv: 2106.15326, 2021.

[66] Xu YC, Yang JF, Cao HZ, et al. Source-free video domain adaptation by learning temporal consistency for action recognition [C] // Proceedings of the European Conference on Computer Vision, 2022: 147-164.

[67] Ni BL, Peng HW, Chen MH, et al. Expanding language-image pretrained models for general video recognition [C] // Proceedings of the European Conference on Computer Vision, 2022: 1-18.

[68] Lin ZY, Geng SJ, Zhang RR, et al. Frozen CLIP models are efficient video learners [C] // Proceedings of the European Conference on Computer Vision, 2022: 388-404.