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依赖的数据之间存在显著差异，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。针对这一问题，本研究提出了一种领域自适应预训练方法，旨在改善黑暗环境下的行为识别性能。该方法融合了外部视觉去暗增强模型以引入关键的去暗知识，并采用跨领域自蒸馏框架来优化预训练模型，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，本方法在全监督的黑暗场景行为识别数据集上获得了 97.19%的准确率，在无源领域自适应场景数据集中，准确率提升至 49.11%，而在多源领域自适应场景数据集中，准确率达到了 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

Qinglin Xu[1,2],Yu Qiao[3], Yali Wang[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: Yali Wang E-mail：yl.wang@siat.ac.cn

Abstract: Action recognition in the dark is a challenging task in practice because it is difficult to learn robust action representations from low light environments. Furthermore, there is a domain gap between dark scenes and the data used by traditional pretrained models, which results in suboptimal results with the traditional pretrain-finetune approach, and pretraining from scratch is costly. To address this issue, a domain-adaptive pretraining method is proposed to improve action recognition 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 Top-1 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 Top-1 accuracy can reach 54.63%.

Key words: Dark Illumination; Action Recognition; Knowledge Adaptation; Domain Adaptation

Funding: This work was supported by the National Key R&D Program of

来稿日期：2023-12-25 修回日期：2024-03-04

基金项目：科技创新 2030——“新一代人工智能”重大项目(2022ZD0160505)；国家自然科学基金资助项目(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，博士研究生导师，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，博士研究生导师，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。

China(NO.2022ZD0160505), the National Natural Science Foundation of China（Grant No. 62272450)

1 引 言

随着深度学习技术的不断深化与发展，多个研究领域已在图像分类[1-3]、行为识别[4-6]、视频检索[7-8]以及图像生成[9-10]等多个领域取得了重大突破。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性的课题。近年来，借助于 3D 卷积神经网络[11-13]和 Vision Transformers(ViT)[14-16]等先进模型，行为识别技术已实现了质的飞跃。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。在黑暗环境下进行行为识别，其关键在于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，但标准的这一微调范式由于域差异而难以达成理想效果。

为了缓解黑暗场景对模型识别能力的负面影响，我们初步考虑采用视觉去暗增强模型。然而，鉴于去暗增强模型的能力有限，其生成的增强视图可能未达预期效果，甚至存在引入额外噪声的情况，因此这种方法并不能完全消除黑暗环境对模型识别能力的影响。为了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，以提升模型在黑暗场景中的识别性能，本文提出了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，在已经在大规模数据上进行过预训练的模型基础上实施进一步的预训练，旨在缩小黑暗场景与正常光照场景之间视觉表征的差异，从而增强模型在黑暗环境下进行行为识别的鲁棒性。具体来说，我们的方法使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。接着，进行跨领域知识蒸馏，目标是实现不同领域视图之间的一致性，减少在明暗场景之间的视觉表征偏差。值得注意的是，我们在实验中观察到，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。为了在知识蒸馏过程中保持模型原有的能力的同时学习目标领域的知识，我们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，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了更多的可学习表征。

本文具有以下贡献：（1）我们采用了跨领域自蒸馏学习方法，有效提升了模型在黑暗场景下进行行为识别的能力；（2）提出了渐进学习适配器使得模型能够在后预训练的过程中保持模型原有的能力，并且注入目标领域知识；（3）我们在黑暗场景下进行了广泛的行为识别实验，并且在全监督和领域自适应不同的基准测试中均取得了优于当前最先进方法的显著成果。

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，这一进展为基于 CNN 的各种视频理解任务敞开了大门。在这些方法中双流架构[12,13,18]和三维 CNN 模型[4,11,19]尤为突出，特别是，三维 CNN 因其在视频处理方面的广泛应用而受到重视，尽管它们伴随着较高的计算成本。为了应对这一问题并提高效率，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，视觉变换器（vision transformer）[24]崭露头角，已经成为图像识别领域的前沿趋势[5,14]，并广泛应用于视频理解之中。最近的许多研究利用近期涌现的图像基础模型[7,8]，并将其应用于视频领域，取得了令人瞩目的成就。然而由于预训练数据与低照度数据之间存在显著差异，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。针对这一问题，本文提出了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，而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加注重于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于设计代理任务，这些任务关注于视频的内在特征，以实现无监督学习。例如，AoT[25]利用视频帧间目标特征的连贯一致性来进行自监督学习；Fernando B 等[26]等通过预测正确与错误排序的帧序列作为自监督学习的方式，使模型能够学习视频的时序信息。随着对比学习方法的崛起，自监督预训练领域取得了令人瞩目的进展。MoCo[27]提出了一种高效的对比学习框架，采用动态记忆库来存储负样本，解决了样本特征不一致的问题。掩码自监督学习 MAE[28]的兴起，使得掩码学习在视觉领域的应用成为可能。MAE 通过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，学习图像不同区域的上下文信息。VideoMAE[29]将这一方法应用于视频自监督任务，考虑到视频数据相对于图像具有更多帧数、更丰富的运动信息和较多冗余信息，提出了一种高效的视频自监督方法。

尽管预训练数据量庞大，但其多样性及对特定领域的适应性仍然是未知的。在自然语言处理（NLP）领域，Gururangan S 等[30]探讨了跨多个领域的渐进式自监督预训练。本文则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及更高效的模型参数训练策略。

2.3 领域自适应

鉴于数据标注的高成本和数据隐私问题的日益重要性，领域自适应（domain adaptation）的研究领域获得了广泛关注，并迅速发展。在这一广泛研究领域中，本文特别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）这两个关键子领域。

无源视频领域自适应（SFVDA）主要研究由于源域数据的隐私性和数据传输所需耗费资源等问题，源域数据在迁移学习的过程中，不能够被使用情况下的域迁移问题。目前无源域适应方法中，3C-GAN [31] 和 SDDA [32]通过使用 GAN[33]生成与目标域数据分布相似带有标签的图片，然后通过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，获得域不变特征。SHOT[34]通过冻结源分类器来利用源特征分布的知识，并利用信息熵最大化和伪标签将目标域的特征匹配到源分类器。

在多源视频领域自适应(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了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。最近，MOST[37]提出了一种基于 Optimal Transport 的严格理论，用于在领域适应中利用模仿学习。在该方法中，教师分类器完美地利用源领域的知识进行处理，而学生分类器则努力模仿教师分类器在源领域中的行为。

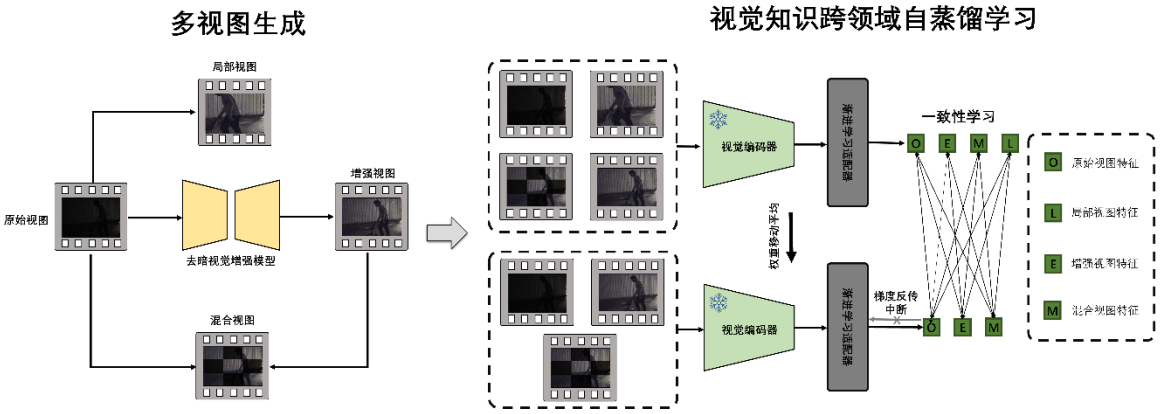

图 1 领域自适应预训练流程

Fig. 1 Domain-Adaptive Pretraining Process

与上述方法不同，本文通过充分利用外部的去暗知识，进行了领域自适应的预训练学习。同时，本文也是首次充分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

在本节中，我们将详细阐述用于黑暗场景下行为识别任务的领域自适应预训练方法。针对黑暗场景视频数据与常规预训练视频数据之间的显著差异这一核心问题，我们的方法旨在利用外部开源的去暗视觉增强模型，以将黑暗环境下的视觉知识有效整合进视觉编码器中。然而，如何高效地利用这些经过去暗处理增强后的知识，需要我们进行深入的研究与探索。直观上，我们可以尝试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，仅依赖增强视频可能导致模型忽视不同视图间的内在联系。

受到自然语言处理领域中后预训练策略的启发[30]，该策略已在自然语言处理领域取得显著成效。相关研究表明，在目标域上进行二次预训练是将预训练模型有效迁移到特定领域的有效策略。基于此，我们整合了视觉去暗增强模型和简单的图像处理技术，进而构建了四种不同类型的视图：原始视图、增强视图、混合视图以及局部视图。通过在不同的域间及同一域内实施自蒸馏的自监督学习策略，我们的方法专注于探索和学习这些多样化视图间的内在一致性。在跨域一致性的自蒸馏学习中，我们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，我们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

在本节中，我们详细介绍了用于跨领域自蒸馏学习的四种视图生成过程，如图 1 的左侧部分所示。首先，原始视图是指视频的初始表征形式，其中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 $H=W=224$。增强视图对应的是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，具体操作上，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：$\boldsymbol{X}\in\mathbb{R}^{\frac{H}{P}\times\frac{H}{P}\times(P^2\cdot C)}$，其中 H 和 W 分别指视频帧的长度和宽度，P 表示正方形补丁的边长，C 是视频帧中的通道数。$\boldsymbol{X}_o$ 和 $\boldsymbol{X}_e$ 是由原始视图和增强视图生成的补丁。然后，我们根据保持 0 和 1 的数量的差不大于 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），然后我们根据掩码 $\boldsymbol{M}$ 生成混合视图：

$$\boldsymbol{X}_t = \boldsymbol{M} \odot \boldsymbol{X}_o + (1-\boldsymbol{M}) \odot \boldsymbol{X}_e \tag{1}$$

其中$\odot$表示矩阵对应元素乘积。局部视图代表视频的局部表征，在这里，视频帧的采样策略与原始视图保持一致，但选取的结果可能不同，空间分辨率设为$H=W=96$。

3.2 渐进学习适配器

在跨领域视觉知识自蒸馏学习的过程中，我们发现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。为了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，我们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。具体而言，如图 2 所示，我们首先引入了一组可学习的适应提示向量: $\boldsymbol{P}_0 = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_M\}$，其中$M$是适应提示的数量。我们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来捕捉视觉上的中间特征。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下所示：

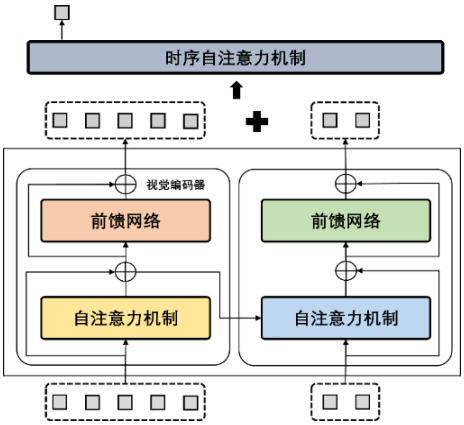

图 2 渐进学习适配器框架

Fig. 2 Progressive adaptation block framework

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

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

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

其中，η代表 layer normalization 归一化操作。$\boldsymbol{X}_{i,t}$表示第i层视觉编码器捕获视频的第t帧的帧特征。$\boldsymbol{Y}_i$表示第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。$\boldsymbol{P}_i$是经过i次交叉注意力机制后得到适应提示的参数，其初始值为$\boldsymbol{P}_0$。α代表多头注意力机制，其中包括查询、键和值三个参数。ϕ代表全连接层。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：

$$\boldsymbol{R} = \tau(\boldsymbol{Y}_n \oplus \boldsymbol{P}_n). \tag{5}$$

τ操作是为了学习不同帧之间时序信息，具体来说，首先，我们由可学习参数[24]生成的位置编码嵌入到$\boldsymbol{Y}_n$和$\boldsymbol{P}_n$的拼接中，公式（5）中的$\oplus$代表向量间的拼接操作。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：$\boldsymbol{R} = \{\boldsymbol{R}_1, \boldsymbol{R}_2, ..., \boldsymbol{R}_{T+M}\}$。在获取到不同帧的特征$\boldsymbol{R}$之后，我们之后聚合每一帧的信息，得到最终的视频特征$\boldsymbol{f}$：

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

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

我们采用了自蒸馏的方式来进行领域自适应预训练学习。这种方法通过预测同一视频不同视图之间在学生模型和教师模型的特征空间内的一致性来进行。我们的目标是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），使模型能够更全面地理解视频的潜在分布，并减少视觉表征在黑暗与光明场景之间的域差异。这种策略旨在降低模型对光照变化的敏感性，确保其在不同光照和视角变化下保持稳定性能。具体而言，我们的实验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，而学生模型则包含了四种视图。在获取了视图的特征 $\boldsymbol{f}$ 之后，我们先对其进行标准化：

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

其中 τ 表示温度参数。按照这个方法，我们可以得到标准化之后的特征：p_o、p_e、p_m 和 p_l，其分别对应于原始视图、增强视图、混合视图和局部视图。之后我们通过最小化两个视图的交叉熵损失来确保同一视频不同视图之间的一致性：

$$L=\sum_{x\in\{o,e,t\}}\sum_{y\in\{o,e,m,t\}}-p_{tx}\cdot\log(p_{sy}) \tag{8}$$

其中 p_{tx} 和 p_{sy} 分别表示教师模型和学生模型输出的标准化后的特征。与常规的知识蒸馏架构不同，我们采用自蒸馏策略，学生和教师模型拥有相同的结构和初始参数。为了避免模型由于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，我们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：

$$\theta_t\leftarrow\lambda\theta_t+\left(1-\lambda\right)\theta_s \tag{9}$$

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，我们成功构建了一种能在黑暗场景中鲁棒地提取视觉特征的编码器。在全监督设定下，模型利用黑暗场景下的训练数据进行了标准的微调学习；而在无源与多源领域自适应设定中，我们首先利用预训练模型[39]提取无监督数据的伪标签，随后以这些伪标签为进行迁移学习。整个学习流程遵循预训练-领域自适应预训练-微调的模式，有效实现了在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，我们使用 CLIP[39]中 BERT[41]来作为文本编码器，其由 12 层、512 维的 Transformer[42]和 8 个注意力头组成。对于视觉编码器，我们采用了与 DarkLight[43]相同的 R(2+1)D-34[21]。在 SFVDA 和 MSVDA 设定中，我们再次使用 CLIP[39]中的文本编码器进行文本特征提取。对于视觉编码器，我们使用 CLIP 的 VIT-B/16。在渐进式适应模块中，我们将视频级别的 token 数设置为 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。

对于第一阶段的渐进式自我预训练，我们使用 AdamW[44]优化器，学习率为 5e-5，批大小为 128。权重衰减设置为 0.04，用于正则化。在训练期间，我们冻结视觉骨架，并在

表 1　全监督黑暗场景下 ARID 实验结果
Table 1　Result for fully supervised on

模型方法	准确度（%）	
	Top1	**Top5**
Timesfomer[5]	66.57	98.31
CLIP[39]	67.15	98.58
R(2+1)D[43]	94.04	99.87
Ours + Timesfomer[5]	72.86 +6.29	97.12
Ours + CLIP[39]	74.82 +7.67	96.95
Ours + R(2+1)D[43]	97.19 +3.15	99.59

表 2　多源领域自适应下 Daily-DA 实验结果
Table 2　Result for MSVDA on Daily-DA

方法	方法类别	Daily-DA
		Daily→A11
s-DANN[45]	Adversarial-based	22.03±0.35
s-ADDA[46]		22.30±0.21
s-TA^{3}N[47]		21.76±0.16
s-ACAN[48]		23.44±0.16
c-DANN[45]		22.15±0.33
c-ADDA[46]		22.65±0.25
c- TA3N[47]		22.24±0.20
c- ACAN[48]		23.95±0.28
MDAN[35]		23.75±0.38
DCTN[49]		24.94±0.36
MDDA[50]		22.73±0.26
s-MMD[51]	Discrepancy-based	21.62±0.22
s-MCD[52]		23.80±0.28
s-CORAL[53]		21.51±0.15
c-MMD[54]		24.28±0.36
c-MCD[52]		25.68±0.28
c-CORAL[53]		23.96±0.16
LtC-MSDA[54]		24.98±0.12
MCC[55]		22.65±0.35
MOST[37]		26.28±0.46
M3SDA[56]		24.83±0.23
TAMAN[57]		29.95±0.35
ActionCLIP[58]	-	52.11±0.99
Ours		**54.63**±0.83

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，去暗增强模型我们选择使用 SCI[38],其使用简单的架构就完成了数据增强的目的。公式（7）中的τ遵循了 DINO[39]的方法，将τ的值固定在 0.1。对于公式（9）中λ，我们也使用了 DINO 中的设定，将其初始值设置为 0.996，目标值调整为 1。为了实现λ值的平滑变化，我们选用了余弦曲线调度来缓慢增加λ值。

在第二阶段，我们采用 AdamW[44]优化器，基础学习率为 5e-6 来微调预训练的参数，同时使用学习率为 5e-5 的新模块初始化可学习参数。第二阶段的权重衰减设置为 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

在本实验中，我们对首个黑暗场景下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，涵盖了 11 个不同的动作类别，为了增强实验结果的鲁棒性，数据集有 3 组划分。我们将报告三组划分 Top1 和 Top5 的平均准确率。

4.2.2 实验结果

表 1 中的结果显示，相较于以往的方法，我们提出的方法在性能上表现出显著优越性，特别是在 Top1 准确率方面，其提高幅度超过 3%。鉴于数据集仅涵盖 11 个类别，所有方法在 Top5 中均取得了良好的结果。值得注意的是，即便是先进的模型，如 Timesformer[5]和 CLIP[39]，它们在主流行为识别数据集中取得出色的成绩，但在 ARID 基准上的表现仍显不如人意，这突显了视频模型在处理重大领域差异时的局限性。然而，值得一提的是，这些方法在经过我们的领域自适应预训练后，准确率得到了显著的提升，这证明了我们提出的方法在克服领域差异方面的高效性以及即插即用性。

4.3 多源视频领域自适应

4.3.1 数据集介绍

我们选用 Daily-DA[63]数据集作为多源视频领域自适应的基准，它涵盖了正常光照和

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

方法	是否无源	Daily-DA			
		K600→A11	**MIT → A11**	**H51→A11**	**Avg**
DANN[45]	×	21.18	22.81	14.20	19.40
MK-MMD[51]	×	21.66	21.02	20.35	21.01
TA^{3}N[48]	×	19.87	21.57	14.38	18.60
SFDA[59]	√	12.57	15.96	13.08	13.87
SHOT[34]	√	12.03	15.28	13.50	13.60
SHOT++[60]	√	12.57	14.90	15.98	14.48
MA[61]	√	12.76	17.75	12.90	14.47
BAIT[62]	√	12.69	16.93	13.65	14.42
CPGA[63]	√	13.06	18.08	13.14	14.76
ATCoN[64]	√	17.21	27.23	17.92	20.79
ActionCLIP[56]	√	47.89	48.59	45.20	47.22
Ours	√	**49.61**	**50.86**	**46.87**	**49.11**

方法	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**

方法	Backbone	ARID
ActionCLIP[54]	ViT-B/16	67.15
XCLIP[65]	ViT-B/16	67.32
FrozenCLIP[66]	ViT-B/16	66.22
Ours	ViT-B/16	**74.82**

黑暗场景下的视频数据。该数据集整合了四个不同的数据集：ARID[40]（A11）、HMDB51[67]（H51）、Moments-in-Time[68]（MIT）和 Kinetics600[69]（K600）。HMDB51、Moments-in-Time 和 Kinetics 在行为识别领域得到了广泛应用，而 ARID 则是一个最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，我们将一个数据集作为目标域，而其他三个数据集则充当源域。因此，涉及四个任务：Daily→A11、Daily→H51、Daily→MIT 和 Daily→K600。值得注意的是，一旦我们从标注的源数据中获得训练有素的模型，我们将不再使用源数据，这与其他方法的做法有所不同。在呈现实验结果时。对于每种方法，我们使用不同的随机种子进行了五次独立实验，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。我们将仅展示 Daily→A11 任务，以强调我们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

表 2 列出了多源视频域适应 Daily-DA 的实验结果。通过比较可以看出，我们提出的方法在多源视频域适应任务中达到了当前的顶尖性能水平。与最先进的 TAMAN[54]方法相比，我们的方法性能提升了 24.68%。另外，我们还呈现了使用相同视觉编码器的 ActionCLIP [54]的结果，该方法通过在多源领域进行训练并使用伪标签的方式迁移到目标领域来解决域适应问题。我们的方法在这项任务上取得了+2.52%的增益，这进一步证明了我们方法在稳定解决域适应问题方面的卓越性能。

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。他包含的数据集和无源域适应是一样的。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 个不同类别，其中包括 12 项跨域动作识别任务。在本文中，我们选择展示以 ARID 为目标域数据集的 3 个跨领域任务，以突显我们提出方法在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

表 3 展示了无源视频领域自适应的实验结果。我们的方法在性能上显著超越了最先进的 ATCON[60]方法。为了进行公正的比较，我们还呈现了使用相同视觉编码器的 ActionCLIP[56]的结果，该方法是在源域数据集上进行训练，并通过利用目标域生成的伪标签来实现迁移学习。我们的方法在平均准确率上取得了+1.89%的增益，突显了其在无源视频领域自适应方面的卓越性能。

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，我们深入研究了我们提出的领域自适应预训练方法对减少黑暗场景对行为识别的影响。实验结果表明，在进行跨领域自蒸馏学习时，当我们不冻结视觉编码器对其进行训练时，模型在三个设定任务下的表现均较未进行跨领域自蒸馏学习而是直接进行

微调的基准方法差，这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。可以观察到，在锁住视觉编码器后（使用 ActionCLIP[54]中的 seqTransf 模块，并仅微调该部分），经过跨领域自蒸馏学习后，模型在三个设定下均取得了提升，其中在 MSVDA 设定下取得了显著的+1.43%提升。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 96.83%，48.41%和 53.54%的准确率，相较于仅锁住视觉编码器的策略，分别提升了 0.36%，0.70%和 1.09%。

4.5.2 与基于 CLIP 的方法比较

表 5 列出了我们的方法与基于 CLIP[39]的其他方法（ActionCLIP[54]、XCLIP[65]和 FrozenCLIP[66]）在 ARID[40]上的全监督设定的比较结果。在使用相同的 ViT-B/16 主干网络的情况下，我们的方法相较于先前的基于 CLIP[39] 的方法实现了更高的 Top1 准确度，其中我们的方法比最优的 XCLIP[61] 高出 7.50%。

4.6 结果可视化分析

如图 3 所示，我们从两个不同的视频片段中分别抽取了四帧以分析行为识别模型的性能（其中第 1 至 4 列展示喝水行为，第 5 至 8 列展示推动物体行为）。首先，第一行直接呈现了通过平均采样方法得到的视频帧序列。为了确保所有细节均可清晰观察，我们对这些原始帧进行了亮度增强处理，其结果显示于第二行。随后，第三行与第四行展示了采用领域自适应预训练方法前后的 Grad-CAM 可视化结果。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，模型能够将注意力集中在关键信息上，例如被推的物体以及手中的杯子。

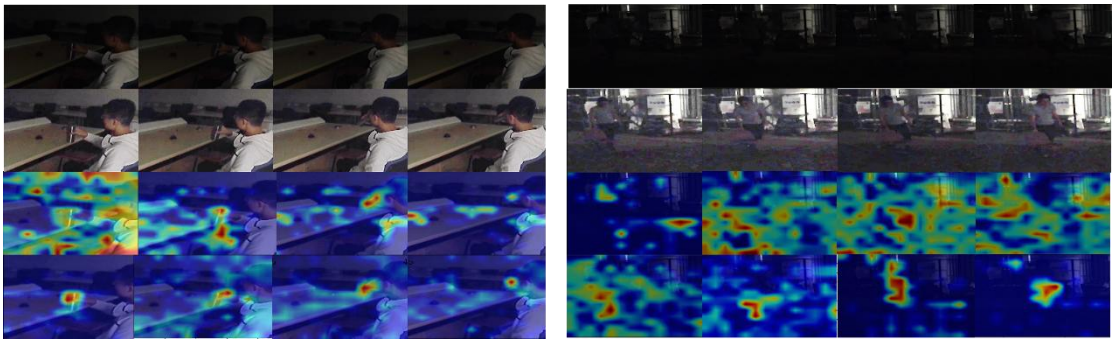

图 3 Grad-CAM 生成的视频注意力可视化
Fig. 3 The attention visualization of video generated by Grad-CAM

5 讨论与分析

在黑暗场景下进行视频理解任务是现实世界中一项具有挑战性的任务，黑暗的恶劣环境使得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的域差异，传统的预训练-微调范式并不能取得良好的效果，CLIP[54]在全监督设定下对黑暗场景行为识别 ARID 数据集的结果仅为 67.15%。融合了本文方法后，CLIP 方法的精度提升至 74.82%，取得了 7.67%的提升。同时，在选择视觉编码器为 R(2+1)D[40]的情况下，达到了目前最先进的结果 97.15%，超过之前的方法 3.19%。此外，我们的方法也适用于领域自适应，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%的结果，分别提升了 1.89%和 2.52%。可见，我们设计的跨领域自蒸馏方法能够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。本文提出了一种针对黑暗场景行为识别的有效方法，采用了预训练-后预训练-微调的迁移学习范式。尽管相较于从头开始预训练，这种策略减少了大量的计算资源和时间的投入，但与 ATCoN[60]等方法相比，仍需要额外的后预训练步骤。

6 结 论

在本论文中，面对黑暗场景下行为识别的挑战，我们提出了基于领域自适应预训练的方法。该方法通过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，使模型能够学习在不同光照条件下行为的一致性，以减少领域差异所带来的影响。通过在三个不同设定下进行的黑暗场景行为识别实验，我们验证了该方法的有效性。在未来的工作中，我们将关注更广泛的开放世界场景，设计适用于各种实际情境的预训练模型，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[J]. arXiv preprint arXiv:1409.1556, 2014.

[2] 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..

[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 L, Xiong Y, 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]//ICML. 2021, 2(3): 4.

[6] Li K, Wang Y, Gao P, et al. Uniformer: Unified transformer for efficient spatiotemporal representation learning[J]. arXiv preprint arXiv:2201.04676, 2022.

[7] Luo H, 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 W, Luo H, 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 H, Zhang H, Barber J, et al. Muse: Text-to-image generation via masked generative transformers[J]. arXiv preprint arXiv:2301.00704, 2023.

[10] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents, 2022[J]. URL https://arxiv. org/abs/2204.06125, 2022, 7.

[11] Wang X, 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 H, 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 K, Wang Y, He Y, et al. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer[J]. arXiv preprint arXiv:2211.09552, 2022.

[17] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[18] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in neural information processing systems, 2014, 27.

[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 Z, 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 X, 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 Z, Wang L, 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 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010
[25] 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.
[26] Wei D, Lim J J, 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.
[27] He K, Fan H, Wu Y, 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] Germain M, Gregor K, Murray I, et al. Made: Masked autoencoder for distribution estimation[C]//International conference on machine learning. PMLR, 2015: 881-889.
[29] Tong Z, Song Y, Wang J, et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training[J]. Advances in neural information processing systems, 2022, 35: 10078-10093.
[30] Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: Adapt language models to domains and tasks[J]. arXiv preprint arXiv:2004.10964, 2020.
[31] Kurmi V K, Subramanian V K, Namboodiri V P. 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 Q, Cao W, 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 D, Feng J. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation[C]//International conference on machine learning. PMLR, 2020: 6028-6039.
[35] Scalbert M, Vakalopoulou M, Couzinié-Devy F. Multi-source domain adaptation via supervised contrastive learning and confident consistency regularization[J]. arXiv preprint arXiv:2106.16093, 2021.
[36] Peng X, Bai Q, Xia X, 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]//Uncertainty in Artificial Intelligence. PMLR, 2021: 225-235.
[38] Ma L, Ma T, Liu R, 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 J W, Hallacy C, et al. Learning transferable visual models from natural

language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.

[40] Xu Y, Yang J, Cao H, et al. Arid: A new dataset for recognizing action in the dark[C]//Deep Learning for Human Activity Recognition: Second International Workshop, DL-HAR 2020, Held in Conjunction with IJCAI-PRICAI 2020, Kyoto, Japan, January 8, 2021, Proceedings 2. Springer Singapore, 2021: 70-84.

[41] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[42] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[43] Chen R, Chen J, Liang Z, 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[J]. 2018.

[45] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation[C]//International conference on machine learning. PMLR, 2015: 1180-1189.

[46] 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.

[47] Chen M H, 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.

[48] Xu Y, Cao H, Mao K, et al. Aligning correlation information for domain adaptation in action recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[49] Xu R, Chen Z, Zuo W, 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.

[50] Zhao S, Wang G, Zhang S, et al. Multi-source distilling domain adaptation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12975-12983.

[51] Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks[C]//International conference on machine learning. PMLR, 2015: 97-105.

[52] 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.

[53] Sun B, Feng J, Saenko K. Return of frustratingly easy domain adaptation[C]//Proceedings of the AAAI conference on artificial intelligence. 2016, 30(1).

[54] Peng X, Bai Q, Xia X, et al. Moment matching for multi-source domain adaptation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1406-1415.

[55] Wang H, Xu M, Ni B, et al. Learning to combine: Knowledge aggregation for multi-source domain adaptation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. Springer International Publishing, 2020: 727-744.

[56] Jin Y, Wang X, Long M, et al. Minimum class confusion for versatile domain

adaptation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. Springer International Publishing, 2020: 464-480.

[57] Xu Y, Yang J, Cao H, et al. Multi-source video domain adaptation with temporal attentive moment alignment network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[58] Wang M, Xing J, Liu Y. Actionclip: A new paradigm for video action recognition[J]. arXiv preprint arXiv:2109.08472, 2021.

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

[60] Liang J, Hu D, Wang Y, 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.

[61] Yang S, Wang Y, Van De Weijer J, et al. Unsupervised domain adaptation without source data by casting a bait[J]. arXiv preprint arXiv:2010.12427, 2020, 1(2): 5.

[62] Agarwal P, Paudel D P, Zaech J N, 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.

[63] Qiu Z, Zhang Y, Lin H, et al. Source-free domain adaptation via avatar prototype generation and adaptation[J]. arXiv preprint arXiv:2106.15326, 2021.

[64] Xu Y, Yang J, Cao H, et al. Source-free video domain adaptation by learning temporal consistency for action recognition[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 147-164.

[65] Ni B, Peng H, Chen M, et al. Expanding language-image pretrained models for general video recognition[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 1-18.

[66] Lin Z, Geng S, Zhang R, et al. Frozen clip models are efficient video learners[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 388-404.

[67] Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition[C]//2011 International conference on computer vision. IEEE, 2011: 2556-2563.

[68] Monfort M, Andonian A, Zhou B, 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.

[69] Carreira J, Noland E, Banki-Horvath A, et al. A short note about kinetics-600[J]. arXiv preprint arXiv:1808.01340, 2018.