Source-free Video Domain Adaptation by Learning Temporal Consistency for Action Recognition

摘要

基于视频的无监督域适应（VUDA）：这些方法需要在适应过程中不断访问源数据。然而，在许多实际应用中，源视频域中的主题和场景应该与目标视频域中的主题和场景无关。并且，源视频数据属于隐私，难以获取。

在本文中，我们提出了一种新的关注**时间一致性网络(ATCoN)**，通过学习时间一致性来解决SFVDA（Source-Free Video-based Domain Adaptation）问题，该网络由跨局部时间特征执行的两个新的一致性目标(即**特征一致性和源预测一致性**)保证。

注释：**时间一致性**是指在不同领域或时间范围内调整时间特征和预测，以提高视频模型的性能。

介绍

每个局部时间特征都不相同，局部时间特征可能不包含相似的语义信息。因此，整体时间特征可能包含模糊的语义信息，并且不会具有区别性。相反，我们假设对于源视频，提取的局部时间特征不仅具有区别性，而且彼此之间一致，具有相似的特征分布模式，这意味着相似的语义信息。这种假说被称为**跨时间假说**。如果目标数据与源数据分布一致，我们假设目标数据学习了类源表示，因此目标数据表示应该满足跨时间假设。我们的方法被设计成局部时间特征在特征表示上是一致的，这将导致相应的整体时间特征是有效的和有区别的。

由于只有具有源分类器的源模型可用于自适应，因此目标数据对源数据分布的相关性与目标数据在源分类器上的预测高度相关。

因此，为了更好地使目标时间特征适应源分类器，相应的局部时间特征与源数据分布的相关性也应该保持一致。这种一致性可以解释为局部时间特征相对于固定源分类器的源预测一致性。此外，为了提高视频特征的可分辨性，需要将局部时间特征精心组合，构建整体时间特征。

ATCoN通过关注局部时间特征，进一步使目标数据适应源数据分布，其与源数据分布的相关性具有更高的置信度，表示为更高的源预测置信度。

总之，我们的贡献有三方面。首先，我们提出了无源视频域自适应(SFVDA)问题。据我们所知，这是第一个研究基于视频任务的无源传输的研究，旨在解决VUDA中的**数据隐私问题**。

ATCoN的目的是通过学习由特征一致性和源预测一致性组成的时间一致性，获得满足跨时间假设的有效的、有判别性的整体时间特征。ATCoN通过关注具有高源预测置信度的局部时间特征，进一步将目标数据与源数据分布对齐，而无需访问源数据。该方案只提供训练良好的源视频模型和未标记的目标领域数据进行自适应。

结论

制定了具有挑战性但现实的无源视频域适应(SFVDA)问题，该问题解决了视频中的数据隐私问题。我们提出了一种新的ATCoN来有效地解决SFVDA。最后证明了优越性

相关工作

尽管图像的SFDA研究取得了进展，但SFVDA尚未得到解决。

本方法

在无源视频域自适应(SFVDA)场景中，我们只得到一个由空间特征提取器G_S、$_{sp}$、时间特征提取器$G_{S,t}$和分类器H_S，以及一个未标记的目标域$D_T = \{V_{iT}\}^{nT}{}_{i=1}$，具有$n_T$以$p_T$的概率分布为特征的 i.i.d视频。源模型通过训练其参数θ_S、s_p、θ_S、t、θ_H生成，标记的源域$D_S = \{(V_{iS}, y_{iS})\}^{nS}{}_{i=1}$，包含$n_S$个视频。我们假设标记的源域视频和未标记的目标域视频共享相同的C类，但在将源模型调整为D_T时，D_S是不可访问的。

相比之下，SFVDA采用的是时间关系网络(Temporal Relation Network, TRN)[50]，因为它能够通过对空间表示之间的相关性进行推理，获得更精确的时间特征，这与人类识别动作的方式相对应。

$$lt_{iS}^{(r)} = \sum_m g_S^{(r)}((V_{iS}^{(r)})_m), \tag{1}$$

$(V^{(r)}{}_{iS})_m = \{f^{(a)}{}_{iS}, f^{(b)}{}_{iS}, \ldots\}_m$是具有r个时序帧的第m个clip。a和b是帧序号。 积分函数$g^{(r)}{}_S$

最终的整体时间特征t_{iS}是应用于所有局部时间特征的简单平均聚合,

$$\mathbf{t}_{iS} = \frac{1}{k-1} \sum_r lt_{iS}^{(r)}$$

通过在t_{iS}上应用源分类器H_S进一步计算源预测（p_S）。**源模型以标准交叉熵损失作为目标函数进行训练**，其公式为:

$$\mathcal{L}_{S,ce} = -\frac{1}{n_S} \sum_{i=1}^{n_S} y_{iS} \log \sigma(H_S(\mathbf{t}_{iS})), \tag{2}$$

其中σ是softmax函数，其C -th元素定义为$\sigma_c(x)$ =

$$exp(x_c) / \sum_{c=1}^{C} exp(x_c)$$

为了使源模型更具判别性和可转移性，从而更好地对目标数据进行对齐，我们进一步采用了标签平滑技术[35]，使提取的特征分布在均匀分离的紧密聚类中。进一步表示为:

$$\mathcal{L}'_{S,ce} = -\frac{1}{n_S} \sum_{i=1}^{n_S} y'_{iS} \log \sigma(H_S(\mathbf{t}_{iS})),$$

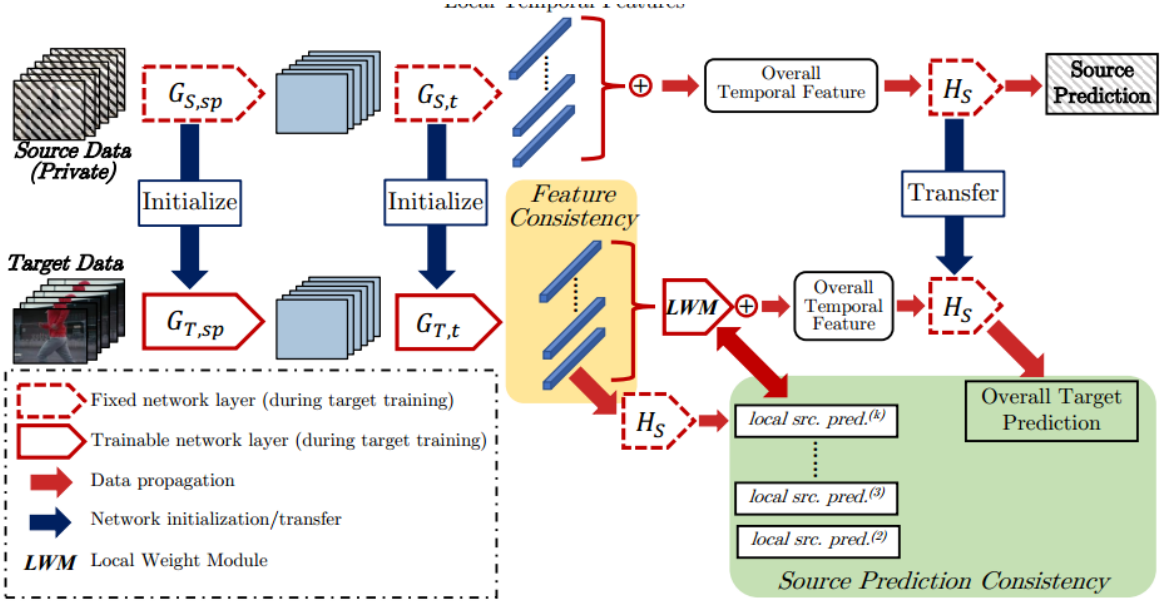
$y'{}_{iS}$是平滑的标签

$$y'_{iS} = (1 - \epsilon) y_{iS} + \epsilon / C$$

ϵ为平滑参数，值设为0.1。

在没有目标标签或源数据的情况下，以自监督的方式提取有效的**总体时间特征**，这些特征具有判别性并符合跨时间假设;另一方面，通过关注**局部时间特征**来对齐源数据分布，对其与源数据分布的相关性具有更高的置信度。

拟议ATCoN的结构



目标时空特征提取器$G_{T、sp}$ $G_{T、t}$采用与$G_{S、sp}$ $G_{S、t}$、$G_{T、sp}$、G_T相同的网络架构，$G_{T,sp}$和$G_{T,t}$分别由$G_{S、sp}$、$G_{S、t}$进行初始化。local src.pred.(k)在第 k 帧处的局部源预测。整体时间特征是通过学习局部时间特征的时间一致性以及直接在局部时间特征上应用源分类器H_S产生的各自的局部源预测来获得的。同时，为了对目标局部时间特征进行集中聚合，进一步设计了局部权重模块(LWM)。

如果局部时间特征一致，则$lt^{(r1)}{}_T$与$lt^{(r2)}{}_T$之间的互相关矩阵应该接近单位矩阵。互相关矩阵表示为:

$$\mathcal{C}^{r1r2} = \left(\hat{lt}_T^{(r1)}\right)^T \hat{lt}_T^{(r2)}, \tag{4}$$

式中，ˆlt 为归一化局部时间特征，计算为:

$$\hat{lt} = \frac{lt - \mathbb{E}(lt)}{\sqrt{Var(lt) + \varepsilon}}, \tag{5}$$

ε为数值稳定性的小偏置值。

互相关矩阵C^{r1r2}是一个大小为d × d的方阵，其中d为局部时间特征的维数。由于理想情况下C^{r1r2}应该接近单位矩阵，特征一致性损失应该最大化各自局部时间特征的相似性，同时减少组件之间的冗余。

因此，对于C^{r1r2}的特征一致性损失表示为:

$$\mathcal{L}_{fc}^{r1r2} = \sum_i (1 - \mathcal{C}_{ii}^{r1r2})^2 + \lambda \sum_i \sum_{j \neq i} (\mathcal{C}_{ij}^{r1r2})^2, \tag{6}$$

其中i, j∈[0,d−1]为局部时间特征维数的指标，λ为权衡常数。

最终的特征一致性损失计算为所有相互关联矩阵的平均特征一致性损失，每个矩阵对应于一对局部时间特征。**最终的特征一致性损失**表示为:

$$\mathcal{L}_{fc} = \frac{1}{N_{fc}} \left(\sum_{r1} \sum_{r2 \neq r1} \mathcal{L}_{fc}^{r1r2} \right), \tag{7}$$

其中N_{fc} $=P^{k-1}{}_2$为局部时间特征对的总数。

此外，**由于同一输入视频的局部时间特征应该通过最小化L_{fc}来保持一致，因此它们与源数据分布的相关性也应该保持一致。**由于源分类器包含源数据分布，因此这种相关性可以通过源分类器对局部时间特征的预测来近似。换句话说，目标局部时间特征对源数据分布相关性的一致性相当于目标局部时间特征对源预测的一致性。同时，对各自的局部时间特征进行聚合，得到目标整体时间特征。它应该包含与局部时间特征相似的运动信息。因此，源预测的一致性预测可以扩展到整体时间特征。

局部源预测：

$$p_{lt,T}^{(r)} = H_S(lt_T^{(r)})$$

平均局部源预测

$$\bar{p}_{lt,T} = \frac{1}{k-1}\sum_{r=2}^{k} p_{lt,T}^{(r)}$$

为了实现源预测的一致性，我们的目标是最小化每个局部源预测与平均局部源预测之间的差异:

$$\mathcal{L}_{pc}^{local} = \frac{1}{k-1}\left(\sum_{r=2}^{k} KL(\log\sigma(p_{lt,T}^{(r)}) \| \log\sigma(\bar{p}_{lt,T}))\right), \quad (8)$$

KL(p‖q) 代表Kullback–Leibler (KL) 差异。

通过将H_S应用于目标总体时间特征t_T来计算总体目标预测$p_{t,T}$，这是一个简单的平均聚合，应用于局部时间特征$lt^{(2)}_T$，…，$lt^{(k)}_T$。为了将$p_{t,T}$纳入源预测一致性，我们的目标是最小化$p_{t,T}$与$\bar{p}_{lt,T}$之间的绝对差值，定义为:

$$\mathcal{L}_{pc}^{overall} = \sum_{c=1}^{C} |\log\sigma_c(p_{t,T}) - \log\sigma_c(\bar{p}_{lt,T})|. \quad (9)$$

最终的源预测一致性是通过联合最小化每个局部源预测与平均局部源预测之间的预测偏差，以及整体目标预测与平均局部源预测之间的预测偏差来实现的，表示为:

$$\mathcal{L}_{pc} = \alpha_{local}\mathcal{L}_{pc}^{local} + \alpha_{overall}\mathcal{L}_{pc}^{overall}$$

其中α局部和α整体是权衡常数。因此，**通过对源预测一致性损失和特征一致性损失进行联合优化来实现学习时间一致性**，表示为:

$$\mathcal{L}_{tc} = \beta_{fc}\mathcal{L}_{fc} + \beta_{pc}\mathcal{L}_{pc}$$

其中β_{fc}和β_{pc}为权衡超参数。

Local Weight Module (LWM).

观察到总体时间特征t_T是通过简单地对所有局部时间特征进行平均来构建的。这是不合理的，因为每个局部时间特征的重要性通常是不均衡的。因此，我**们提出了局部权重模块(LWM)来为局部时间特征分配局部权重**，以进行后续的关注聚合。

$p^{(r)}_{lt,T}$的置信度如下：

$$\mathbb{C}(p_{lt,T}^{(r)}) = \sum_{c=1}^{\mathcal{C}} \sigma_c(p_{lt,T,c}^{(r)})\log\sigma_c(p_{lt,T,c}^{(r)}). \quad (10)$$

最后通过加入残差连接得到局部时间特征$lt^{(r)}_T$所对应的局部相关权值，以实现更稳定的优化，表示为:

$$w_{lt_T^{(r)}} = 1 + \mathbb{C}(p^{(r)}_{lt,T}).$$

对局部相关权值进行加权，得到加权整体时间特征t'_T，即相应加权局部时间特征的平均聚合，计算为:

$$\mathbf{t}'_T = \frac{1}{k-1} \sum_r w_{lt_T^{(r)}} \, lt_T^{(r)}.$$

同时，对局部源预测$p^{(r)}_{lt,T}$进一步应用局部相关权值，**通过关联加权的局部源预测学习源预测一致性**:

$$p^{(r)\prime}_{lt,T} = w_{lt_T^{(r)}} p^{(r)}_{lt,T}.$$

我们从两个方面进一步完善ATCoN:

（1）信息最大化

因此，我们对加权整体时间特征应用信息最大化(IM)损失，如下:

$$\begin{aligned}\mathcal{L}_{IM} = &-\mathbb{E}_{V_T \in \mathbf{D}_T} \sum_{c=1}^{C} \sigma_c(H_S(\mathbf{t}'_T(V_T))) \log \sigma_c(H_S(\mathbf{t}'_T(V_T))) \\ &+ \sum_{c=1}^{C} KL\left(\mathbb{E}_{V_T \in \mathbf{D}_T}[\sigma_c(H_S(\mathbf{t}'_T(V_T)))] \,\|\, \frac{1}{C}\right),\end{aligned} \tag{11}$$

式中$t'_T(V_T)$为目标视频V_T对应的加权总体时间特征，σ_c为softmax中的第c个元素。E表示期望值运算.

（2）自监督伪标签生成

为了进一步改善缺少目标标签的ATCoN的分类对齐，我们遵循[20]，并以自监督的方式为目标视频生成伪标签。具体来说，伪标签是通过总体时间特征上重复的k-means聚类过程生成的。其中，c类的初始质心:

$$\mathbf{c}_c^{(0)} = \frac{\sum_{V_T \in \mathbf{D}_T} \sigma_c(H_S(\mathbf{t}'_T(V_T)))\, \mathbf{t}'_T(V_T)}{\sum_{V_T \in \mathbf{D}_T} \sigma_c(H_S(\mathbf{t}'_T(V_T)))}. \tag{12}$$

随后，目标数据V_T的初始伪标签由其最近的质心得到，定义为:

$$\hat{y}_{V_T} = \arg\min_c \cos(\mathbf{t}'_T(V_T), \mathbf{c}_c^{(0)}),$$

其中cos(·，·)为余弦距离函数。在初始伪标签的基础上，进一步更新初始质心，更可靠地表征目标域的类别分布，公式为:

$$\mathbf{c}_c^{(1)} = \frac{\sum_{V_T \in \mathbf{D}_T} \mathbb{I}(\hat{y}_{V_T}=c)\, \mathbf{t}'_T(V_T)}{\sum_{V_T \in \mathbf{D}_T} \mathbb{I}(\hat{y}_{V_T}=c)}, \tag{13}$$

其中I(·)为指示函数。伪标签最终在更新后的质心后更新为:

$$\hat{y}_{V_T} = \arg\min_c \cos(\mathbf{t}'_T(V_T), \mathbf{c}_c^{(1)}).$$

通过伪标签的交叉熵损失进一步训练ATCoN为:

$$\mathcal{L}_{T,ce} = -\frac{1}{n_T}\sum_{i=1}^{n_T} \hat{y}_{V_T} \log \sigma(H_S(\mathbf{t}'_T(V_T))), \quad (14)$$

其中nT为目标视频的总数。

总体目标:综上所述，**给定一个训练好的源模型，ATCoN的总体优化**目标表示为:$L = \beta_{tc}L_{tc} + \beta_{IM}L_{IM} + \beta_{ce}L_{T,ce}$，其中$\beta_{tc}$、$\beta_{IM}$和$\beta_{ce}$为权衡超参数。

注释：标签平滑技术的原理是通过在分类模型中引入标签平滑损失函数，来减少模型对于训练数据中标签的过度依赖，从而提高模型的泛化性能。标签平滑损失函数通常会将**真实标签的概率分布与均匀分布进行加权平均，并与模型预测的概率分布进行比较**，从而计算出损失值。

互相关矩阵是一种在计算机视觉中常用的矩阵，用于计算两个图像之间的相似度。互相关矩阵可以通过将两个图像的像素值进行卷积运算得到，其中每个元素表示两个图像在相应位置上的像素值之间的相似度。

归一化处理是一种常用的数据预处理技术，用于将数据映射到一个固定的范围内，以便更好地进行分析和比较。通常情况下，**归一化处理会将原始数据进行线性变换，使得数据的取值范围缩小到一个固定的区间内（例如[0,1]或[-1,1]）**，并保持数据的分布形态不变。归一化处理可以**使得不同特征之间具有可比性**，从而更好地进行特征选择、降维和分类等任务。

KL 差异的计算结果越小，表示两个分布越相似；计算结果越大，表示两个分布差异越大。在计算机视觉中，KL 差异常用于衡量图像或特征之间的相似度。公式如下：KL(P||Q) = ∑_iP(i)*log(P(i)/Q(i))

信息最大化的核心思想是选择那些具有**最大信息增益的样本**进行标注或训练，从而使得模型在学习过程中获得更多的信息，提高模型的泛化性能。信息最大化的损失函数会将未标注样本的信息增益作为权重，用于计算模型的损失值。因此，**具有最大信息增益的样本会被赋予更高的权重，对模型的训练起到更大的作用。**

信息最大化损失函数是一种用于训练生成模型的损失函数其目标是让生成模型生成的样本具有最大的信息熵，即最大化样本的不确定性，从而使生成的样本更加多样化和丰富。

k-means聚类：通过质心收敛进行聚类。

5 分钟带你弄懂 K-means 聚类 - 知乎 (zhihu.com)

在计算机视觉中，**数据传播**（Data Propagation）通常指的是将数据从网络的一层传递到下一层的过程。

在计算机视觉中，**Fixed network layer（固定网络层）**通常指的是在神经网络中的一层，其权重参数在训练过程中被固定不变，也就是说，**这一层的参数不会被更新。**

网络迁移（Network Transfer）是指将已经训练好的模型应用于新问题的过程。在计算机视觉中，通常会使用预训练好的模型来解决新问题，这个过程就是网络迁移。

余弦距离函数的取值范围是[-1,1]，当两个向量完全相同时，余弦距离函数的取值为1；当两个向量正交时，余弦距离函数的取值为0；当两个向量方向完全相反时，余弦距离函数的取值为-1。因此，余弦距离函数可以用于衡量两个向量之间的相似度或者差异度。需要注意的是，余弦距离函数只考虑了两个向量之间的夹角，而没有考虑它们的模长。因此，**在使用余弦距离函数时，需要保证两个向量的模长相等或者已经进行了归一化处理**，以避免模长对相似度的影响。

实验

我们通过三个跨域动作识别基准，包括UCF-HMDB$_{full}$[2]、Daily-DA[43]和Sports-DA[43]来评估我们提出的ATCoN。这些基准测试涵盖了广泛的跨域场景。

实验设置

在三个基准中，**UCF-HMDB$_{full}$**是使用最广泛的跨域视频数据集之一，该数据集包含来自两个公共数据集的视频:UCF101 (U101)[33]和HMDB51 (H51)[16]，共12个动作类3,209个视频，有2个跨域动作识别任务。

同时，**Daily-DA**是一个更具挑战性的数据集，它包含了普通视频和低照度视频。它由四个数据集构成:ARID (A11)[42]、HMDB51 (H51)、Moments-in-Time (MIT)[24]和Kinetics (K600)[14]。HMDB51、Moments-in-Time和Kinetics被广泛用于动作识别基准测试，而ARID是一个较新的暗数据集，由在恶劣光照条件下拍摄的视频组成。Daily-DA共包含8个类18,949个视频，共包含12个跨域动作识别任务。

Sports-DA是一个大规模的跨域视频数据集，由UCF101 (U101)、Sports-1M (S1M)[13]和Kinetics (K600)构建而成。共计40718个视频和23个动作种类。

为了公平比较，所有方法都采用**TRN[50]**作为视频特征提取的主干，源模型在**ImageNet上进行预训练**[5]。在基**于假设迁移和标记迁移的无监督域自适应方法**之后，插入批处理归一化[12]和一个额外的全连接层，同时对最后一个全连接层应用权值归一化[32]。所有实验均使用PyTorch[27]库实现。

总体结果和比较

我们报告了目标域上的前1精度，在每种方法具有相同设置的5次运行中平均。

我们将ATCoN与最先进的SFDA方法以及几种具有竞争力的**UDA/VUDA方法**进行了比较。这些包括:SFDA[15]、SHOT[20]、SHOT++[21]、MA[18]、BAIT[47]和CPGA[28]，它们是为**无源改编而设计**的;以及针对UDA/VUDA场景设计的DANN[6]、MK-MMD[22]和TA3N。我们还报告了仅源模型(TRN)的结果，该模型是通过将源数据训练的模型直接应用于目标数据而获得的。

Table 1. Results for SFVDA on UCF-HMDB$_{full}$ and Sports-DA.

Methods	Source-free	UCF-HMDB$_{full}$			Sports-DA						
		U101→H51	H51→U101	Avg.	K600→U101	K600→S1M	S1M→U101	S1M→K600	U101→K600	U101→S1M	Avg.
TRN	-	72.78	72.15	72.47	86.41	66.95	85.31	71.05	49.29	43.32	67.06
DANN	✗	74.44	75.13	74.79	86.60	66.79	89.32	70.53	61.77	48.73	70.62
MK-MMD	✗	74.72	79.69	77.21	86.49	66.18	87.37	71.43	64.17	**49.24**	70.81
TA^{3}N	✗	78.14	84.83	81.49	88.24	**70.56**	83.32	75.54	57.51	46.37	70.26
SFDA	✓	69.86	74.98	72.42	86.10	60.02	85.37	68.04	55.75	43.58	66.48
SHOT	✓	74.44	74.43	74.44	91.19	64.95	88.84	72.02	53.93	43.58	69.09
SHOT++	✓	71.11	68.13	69.62	90.01	63.11	88.01	70.34	44.75	40.95	66.20
MA	✓	74.45	67.36	70.91	91.04	65.95	87.84	71.88	60.75	39.41	69.48
BAIT	✓	75.33	76.36	75.85	92.27	66.61	88.33	72.85	57.25	44.67	70.33
CPGA	✓	75.82	68.16	71.99	89.42	66.26	86.49	72.55	55.22	44.53	69.08
ATCoN	✓	**79.72**	**85.29**	**82.51**	**93.62**	69.70	**90.64**	**75.99**	**65.24**	47.90	**73.85**

Table 2. Results for SFVDA on Daily-DA.

Methods	Source-free	Daily-DA											
		K600→A11	K600→H51	K600→MIT	MIT→A11	MIT→H51	MIT→K600	H51→A11	H51→MIT	H51→K600	A11→H51	A11→MIT	A11→K600	Avg.
TRN	-	20.87	36.66	29.00	22.11	43.75	53.10	13.81	22.00	37.10	17.20	14.75	24.38	27.89
DANN	✗	21.18	37.50	21.75	22.81	43.33	**58.76**	14.20	29.50	38.24	20.11	**19.75**	27.03	29.51
MK-MMD	✗	**21.66**	36.25	24.00	21.02	**50.42**	58.48	**20.35**	25.75	33.79	18.75	18.00	26.07	29.55
TA^{3}N	✗	19.87	37.67	31.53	21.57	43.01	55.47	14.38	25.71	38.39	14.92	15.56	23.42	28.49
SFDA	✓	12.57	44.95	27.50	15.96	35.19	49.23	13.08	24.25	24.86	16.29	13.25	25.22	25.19
SHOT	✓	12.03	44.58	29.50	15.28	36.67	51.04	13.58	24.25	21.24	17.08	14.00	24.35	25.30
SHOT++	✓	12.57	40.83	28.75	14.90	41.67	46.34	15.98	22.25	33.10	15.42	12.50	21.76	24.42
MA	✓	12.76	45.82	30.00	17.75	37.36	53.54	12.90	25.00	22.19	16.67	15.25	24.29	26.13
BAIT	✓	12.69	45.73	30.00	16.93	39.64	53.00	13.65	25.50	21.17	15.70	14.50	25.52	26.17
CPGA	✓	13.06	46.02	30.75	18.08	39.21	55.09	13.14	26.25	25.54	19.19	16.50	26.72	26.46
ATCoN	✓	17.21	**48.25**	**32.50**	**27.23**	47.35	57.66	17.92	**30.75**	**48.55**	**26.67**	17.25	**31.05**	**33.53**

我们提出的ATCoN甚至超过了在13个跨域任务下使用可访问源数据训练的VUDA方法的性能，而我们的方法的平均精度始终高于所有VUDA方法在三个基准上的评估。这进一步验证了ATCoN在构造有效时间特征方面的能力。

注释：**Top-1准确率（Top-1 Accuracy）**是指在分类任务中，模型预测的结果中最可能的类别与真实类别相同的比例。也就是说，如果我们将所有样本的预测结果按照置信度从高到低排序，那么Top-1准确率就是排在第一位的预测结果与真实类别相同的比例。

消融实验和特征可视化

消融研究从两个方面考察ATCoN:**一是时间一致性的组成部分;其次是LWM生成的局部相关权值的应用。**所有消融研究都是利用UCF-HMDBfull数据集进行的，具有2个跨域动作识别任务，而TRN作为特征提取器的主干。

我们针对4种变体评估ATCoN，以验证所提出的时间一致性损失Ltc的设计:ATCoN- fc，其中仅学习特征一致性;ATCoN-PC†和ATCoN-PC，其中只学习源预测一致性，不包括ATCoN-PC†的整体目标预测;最后是ATCoN-TC，其中只学习具有特征一致性和源预测一致性的时间一致性损失。变体都不使用IM损失和伪标记。

Methods	U101→H51	H51→U101
Source-only (TRN)	72.78	72.15
ATCoN	**79.72**	**85.29**
ATCoN-FC	77.78	83.36
ATCoN-$PC^{\dagger}$	76.67	82.83
ATCoN-PC	77.50	83.01
ATCoN-TC	78.89	84.59

(a) Components of temporal consistency

与学习时间一致性对基线模型性能的改善相比，同时使用IM损失和伪标记的性能增益是微不足道的。通过比较经验证明，ATCoN成功的关键在于对时间一致性的学习。

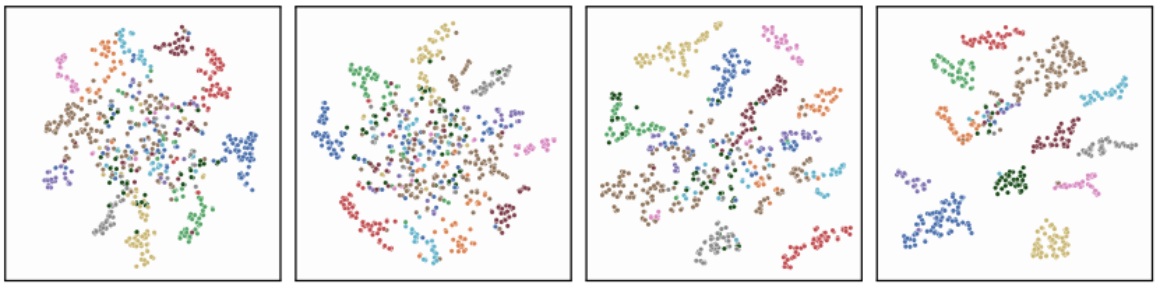
我们将ATCoN与3种变体进行比较:ATCoN- NA，其中没有插入LWM，因此根本无法获得w_{lt};ATCoN-A@F，其中w_{lt}仅用于获取整体时间特征t'_T;和ATCoN-A@P，其中w_{lt}仅用于获得加权局部源预测$p^{(r)}_{lt,T'}$。在上述三种变体的训练过程中，同时采用了IM损失和伪标签生成。

Methods	U101→H51	H51→U101
Source-only (TRN)	72.78	72.15
ATCoN	**79.72**	**85.29**
ATCoN-*NA*	78.33	83.89
ATCoN-*A@F*	79.17	84.93
ATCoN-*A@P*	78.61	84.41

(b) Application of *local relevance weight*

表3(b)所示，无论在何处应用本地相关性权重，都会带来一致的改进，这证明了使用这种权重的必要性。但需要注意的是，与学习时间一致性带来的改善相比，这种改善是相对微不足道的。

我们进一步绘制了ATCoN、CPGA和SHOT在H51→U101任务中学习到的整体时间特征的t-SNE嵌入图，并在目标域中包含了类信息。



(a) Source-Only (b) CPGA (c) SHOT (d) ATCoN

我们可以清楚地看到，ATCoN学习到的特征比其他网络学习到的特征聚类程度要高得多。这验证了ATCoN学习到的特征具有更高的判别性，从而获得了更好的SFVDA性能。