
Article:
Liu, X., Zhou, S., Lei, T. et al. (3 more authors) (2023) First-person video domain adaptation with multi-scene cross-site datasets and attention-based methods. IEEE Transactions on Circuits and Systems for Video Technology, 33 (12). pp. 7774-7788. ISSN 1051-8215

https://doi.org/10.1109/TCSVT.2023.3281671

First-Person Video Domain Adaptation with Multi-Scene Cross-Site Datasets and Attention-Based Methods

Xianyuan Liu, Shuo Zhou, Tao Lei, Ping Jiang, Zhixiang Chen and Haiping Lu, *Senior Member, IEEE*

***Abstract*—Unsupervised Domain Adaptation (UDA) can transfer knowledge from labeled source data to unlabeled target data of the same categories. However, UDA for first-person video action recognition is an under-explored problem, with a lack of benchmark datasets and limited consideration of first-person video characteristics. Existing benchmark datasets provide videos with a single activity scene, e.g. kitchen, and similar global video statistics. However, multiple activity scenes and different global video statistics are still essential for developing robust UDA networks for real-world applications. To this end, we first introduce two first-person video domain adaptation datasets: ADL-7 and GTEA_KITCHEN-6. To the best of our knowledge, they are the first to provide multi-scene and cross-site settings for UDA problem on first-person video action recognition, promoting diversity. They provide five more domains based on the original three from existing datasets, enriching data for this area. They are also compatible with existing datasets, ensuring scalability. First-person videos have unique challenges, i.e. actions tend to occur in hand-object interaction areas. Therefore, networks paying more attention to such areas can benefit common feature learning in UDA. Attention mechanisms can endow networks with the ability to allocate resources adaptively for the important parts of the inputs and fade out the rest. Hence, we introduce channel-temporal attention modules to capture the channel-wise and temporal-wise relationships and model their inter-dependencies important to this characteristic. Moreover, we propose a Channel-Temporal Attention Network (CTAN) to integrate these modules into existing architectures. CTAN outperforms baselines on the new datasets and one existing dataset, EPIC-8.**

***Index Terms*—Action recognition, unsupervised domain adaptation, first-person vision, channel-temporal attention.**

I. Introduction

ACTION recognition is one of the most challenging problems in computer vision with wide applications including human-robot interaction [1] and video moment retrieval [2]. As first-person videos become more common with the wide usage of portable cameras, first-person video action recognition attracted much attention recently because it can offer a unique viewpoint for human daily activity analysis [3]. First-person videos are recorded from the viewpoint of the camera wearer, producing videos with non-linear and hard-to-predict head and body motion and a lack of global context [4]. First-person videos differ from third-person videos in several ways, including occlusion-free interactions with objects, a focus on hands and objects, and different motion patterns due to body and head movements [4], [5]. Many networks [6]–[8] have achieved excellent performance on third-person video benchmark datasets [6], [9], [10], benefiting from their ability to identify the human outline and analyze posture changes. However, first-person videos contain hands rather than human outlines. In addition, first-person video action recognition is valuable for human-to-robot imitation learning [11], [12]. Due to the complexity and expense of annotating new videos, the availability of labelled datasets for first-person video action recognition is still limited compared to third-person datasets.

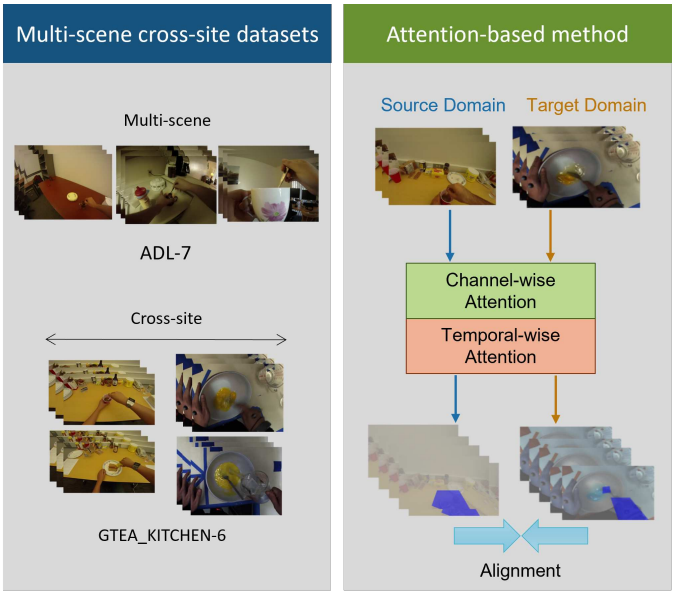


Fig. 1. We enrich datasets and promote diversity via presenting two new datasets with five more domains (left), and we improve UDA for first-person video action recognition via proposing a new channel-temporal attention network (right).

Unsupervised Domain Adaptation (UDA) for first-person video action recognition can address the sample size limitation by leveraging labeled source videos to improve performance on unlabeled target videos, e.g. via minimizing the distribution distance between source and target domains in spatial and temporal feature spaces [13]–[15]. Therefore, this paper

This work was supported in part by the China Scholarship Council (CSC) under Grant 201904910380.

X. Liu, T. Lei (corresponding author) and P. Jiang are with Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China, and with School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China and with University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: liuxianyuan16@mails.ucas.ac.cn; taoleiyan@ioe.ac.cn; jiangping@ioe.ac.cn).

S. Zhou, Z. Chen and H. Lu are with Department of Computer Science, the University of Sheffield, Sheffield S1 4DP, United Kingdom. (e-mail: shuo.zhou@sheffield.ac.uk; zhixiang.chen@sheffield.ac.uk; h.lu@sheffield.ac.uk)

focuses on UDA for first-person video action recognition. To the best of our knowledge, this topic has only been studied on two subsets of the EPIC-KITCHEN (EPIC) dataset [3], i.e. EPIC-97 [16] and EPIC-8 [17]. The shortage of benchmark datasets significantly hinders the development in this area. Meanwhile, the two existing datasets have the following limitations: 1) Their actions are in a single activity scene, e.g. cooking in a kitchen. Studying generalization performance requires exploration on more activity scenes in daily life, i.e. multi-scene problem. 2) They are for within-dataset domain adaptation (DA) challenges, leading to similar global video statistics. However, to study generalization performance, it is important to investigate across datasets with different statistics, i.e. cross-site problem. These limitations motivate us to enrich the benchmark datasets, promote diversity, and accelerate the development in this area.

Therefore, we first select and re-annotate samples from existing datasets to create two first-person video datasets with new challenges in UDA for first-person video action recognition, as shown in Fig. 1: 1) ADL-7 including three long-duration videos from the Activity Daily Living (ADL) dataset [18], on daily life actions; 2) GTEA_KITCHEN-6 including two first-person video datasets, GTEA [19] and KITCHEN [20], on cooking actions. They provide increased domain shift for UDA research. As shown in Section III, they offer two key benefits to the community. Firstly, they provide five more domains to enrich the existing three domains to significantly expand the available dataset choices: ADL-7 provides three domains with multiple daily life scenes for researchers to study against the multi-scene challenge, and GTEA_KITCHEN-6 provides two domains with a larger global video statistics difference to study the cross-site challenge. Secondly, they contain overlapping categories with EPIC-8 to ensure compatibility and scalability with all existing datasets. These benefits would help researchers in evaluating networks across more domains, exploring UDA for different scenarios, enhancing network robustness for real-world applications, and expanding these datasets in other additional ways.

There are two categories of UDA for action recognition. The first category involves giving the image-based UDA [21], [22] the ability to analyze spatio-temporal information [17], [23], [24]. Some approaches uses a two-stream structure to extract spatial and temporal information separately from videos and transfer them separately between datasets. These methods require a high space and time complexity for optical flow computation. Others follow one-stream to extract spatio-temporal information directly and transfer spatio-temporal knowledge between domains. However, most existing approaches are not end-to-end networks, requiring additional computing resources for data argumentation. Moreover, these approaches cannot learn task-specific features end-to-end, since they cannot directly generate spatio-temporal features from videos. The second category involves creating positive-negative samples to use contrastive learning techniques to transfer information between domains [25]–[27]. However, these approaches require additional processes to get positive-negative samples, e.g. different modalities or backgrounds from the same input, requiring additional computational resources and time complexity. Therefore, we consider the first category, follow one-stream and construct an end-to-end network by jointly extracting spatial and temporal information.

First-person videos have some unique characteristics, e.g. actions tend to occur in some local areas, particularly the interaction of the hands and objects. Hence, we hypothesize that networks paying more attention to such areas can leverage such characteristics to benefit common feature learning in UDA, as shown in Fig. 1. In CNN, different channels in different layers capture different characteristics. Therefore, the better weighting of channels improves feature extraction, thereby enhancing the network's performance. Attention mechanisms can improve the weighting of channels to guide the network to focus on the important components [28]. This inspires us to design attention modules that can weigh the channel-wise and temporal-wise features in the CNN layers to reveal the channel-temporal relationships for first-person videos. The term "channel-wise" denotes that algorithms are designed specifically for channels in CNN, e.g. channel-wise attention is an attention mechanism that learns and assigns the weights of attention over each channel. The term "temporal-wise" refers to the network's design specifically for the temporal dimension, e.g. temporal-wise attention learns and assigns weights over each frame. To this end, we propose a Channel-Temporal Attention (CTA) module to excite action-related spatio-temporal features in first-person videos. Moreover, the network should not only focus on these important features but also focus on the common features across domains. Therefore, we utilize an adversarial approach at the video level for alignment to minimize the discrepancy between important channels in the source and target domains. Datasets have been released at *https://github.com/XianyuanLiu/EgoAction*.

In summary, our contributions are as follows:

- We create two first-person video datasets from existing datasets to provide more benchmarking challenges for UDA: ADL-7 for multi-scene and GTEA_KITCHEN-6 for cross-site. To our knowledge, they are the only datasets besides EPIC [16], [17] for studying the first-person video UDA problems.
- We explore different image-based attention mechanisms and develop a new channel-temporal attention module to model channel-wise and temporal-wise inter-dependencies for UDA for first-person video action recognition.
- We propose a new adversarial Channel-Temporal Attention Network (CTAN) and evaluate it on our proposed and existing datasets. Our network outperforms all the UDA baselines and attention networks on average.

Our proposed attention module differs from the reported works in [28]–[37], where an individual temporal-wise attention is not taken into consideration. Our UDA network also differs from the works reported in [14], [15], [38]–[44] in three ways: 1) Our work focuses on video-based problems rather than image-based problems. 2) Our work aligns spatio-temporal features rather than just spatial features. 3) Our work uses attention-based methods to enhance feature embedding whereas the other works do not.

The remainder of this paper is organized as follows. Section II briefly introduces some related works. Next, Section III describes our datasets. Then, Section IV explains the details of the proposed network. Finally, the experiments and the implementation details are reported in Section V, with concluding remarks summarized in Section VI.

II. Related Works

This section briefly reviews the relevant fields, including datasets, attention mechanisms, and unsupervised domain adaptation approaches for images and videos.

A. Related Datasets

There are very limited benchmark datasets for UDA problem on first-person video action recognition, hindering the development in this area. To the best of our knowledge, two subsets of EPIC are the only benchmark datasets for this area, which are EPIC-8 [17] and EPIC-97 [16]. Both provide videos with daily activities captured in the kitchen [3]. EPIC-8 refers to P08, P01, P22 from EPIC as D1, D2, D3 to build three domains, providing within-dataset setting. These domains contain 8 overlapping verb categories, i.e. *put*, *take*, *open*, *close*, *mix*, *pour*, *wash* and *cut*. This dataset contains 1978, 3245 and 4871 action segments, respectively. However, all domains have the same resolutions as 640 $\times$ 480 and similar global video statistics as shown in Table I and Table II. EPIC-97 includes 97 verb categories and two domains, giving a validation set with ground-truth labels for algorithm evaluation. However, there are no ground-truth labels in the testing set of the source domain, and in the training and testing sets of the target domain. There are some problems with this validation set, such as having 84 verb categories instead of 97; roughly a quarter (22 out of 84) not overlapping between source and target; and 16 categories only containing a single video. Given these problems, we excluded it from our paper.

B. Attention Mechanisms

Attention mechanisms equip networks with the ability to focus on the informative input features, which is conducive to the full exploitation of network representational ability and the improvement of model performance [28], [32], [45], [46]. Hence, attention mechanisms have been widely used in various tasks, including image captioning [29], [47], image dehazing [28], [48], and person re-identification [30], [31]. The commonly used attention mechanisms include: Squeeze-Excitation Network (SENet) [32], Convolutional Block Attention Module (CBAM) [33], Style-based Recalibration Module (SRM) [34]. SENet enhances informative channels and fades the useless channels via using average pooling for squeezing and the sigmoid function for excitation. CBAM improves and extends SENet to channel and spatial attentions via additionally using max-pooled features and combining with a spatial attention module. SRM enhances the capacity of networks to capture global information via employing both the mean and standard deviation of the input features to excites the style information from each channel of the feature maps.

Moreover, several recent methods incorporate aforementioned attention mechanisms with temporal modeling for action recognition, e.g. Channel-wise Temporal Attention Network (CWTAN) [35], Temporal Excitation and Aggregation Network (TEA) [36] and Symbiotic Attention with Object-centric feature Alignment (SAOA) [37]. CWTAN utilizes 2D CNN and a global temporal aggregating mechanism for temporal modeling and generates attention weights for each channel in each frame. TEA utilizes 2D CNN and a local multiple temporal aggregating mechanism and generates weights for each channel. SAOA construct VerbNet and NounNet as a two-stream network to produce local/global alignment to generate attention weights for action recognition. Different from the previous attention mechanisms, this paper further exploits channel- and temporal-wise attention by individually generating attention weights for each channel and frame.

C. Related Unsupervised Domain Adaptation Approaches

Imaged-based UDA can be categorized into three approaches. The first approach is discrepancy-based, which aligns source-target distributions by minimizing a divergence that measures the distance between them, e.g. via Maximum Mean Discrepancy (MMD) [22], [41], [49] or Correlation Alignment (CORAL) [42]. The second approach is adversarial-based, constructing domain discriminators for adversarial training to reduce the discrepancy. Domain Adversarial Neural Network (DANN) [21] utilizes discriminators and Gradient Reversal Layer (GRL). Conditional Domain Adversarial Network (CDAN) [50] leverages multilinear and entropy conditioning on discriminative information. The third approach is reconstruction-based, generating reconstruction loss to conduct domain alignment, e.g. pair-wise squared reconstruction loss [43] and scale-invariant mean squared error reconstruction loss [44]. We consider the first two approaches, specifically, their extensions to video UDA, due to the high price and huge difficulty of reconstructing videos.

Video-based UDA approaches can be categorized into two parts. The first category integrates spatio-temporal video feature extractors into image-based UDA approaches [17], [24], [51], [52], including Shuffle Attend Video Domain Adaptation (SAVA) [23], Multi-Modal Self-Supervised Adversarial Domain Adaptation (MM-SADA) [17], Temporal Attentive Adversarial Adaptation Network ($\mathrm{TA^3N}$) [24]. SAVA utilizes self-supervised clip order prediction and clip attention based feature alignment for video domain adaptation. MM-SADA utilizes two-stream networks and self-supervised multi-modal UDA to learn the relationship between RGB and optical flow. $\mathrm{TA^3N}$ utilizes temporal relation module from [53] to extract spatio-temporal features and extends image-based domain adaptation to videos by adding temporal attentive alignment. The second category utilizes contrastive learning technique, e.g. Spatio-Temporal Contrastive Domain Adaptation (STCDA) [25], Cross-Modal Contrastive Domain Adaptation (CMCDA) [26] and Contrast and Mix (CoMix) [27]. STCDA and CMCDA utilize cross-modal contrastive learning networks with sampling strategies for self-supervised learning to align the feature distributions between video domains. CoMix uses

TABLE I
STATISTICS OF THE FIRST-PERSON CROSS-DOMAIN VIDEO DATASETS. GTEA_KITCHEN-6 PROVIDES TWO DATASETS WITH DIFFERENT GLOBAL VIDEO STATISTICS FOR THE CROSS-SITE PROBLEM. ADL-7 PROVIDES VARIOUS ACTIVITY SCENES FOR THE MULTI-SCENE SETTING. G: GTEA. K: KITCHEN.

Dataset	EPIC-8			GTEA_KITCHEN-6		ADL-7		
Resolution	640x480			G: 456x256 / K: 342x256		342x256		
Frame rate	60			G: 15 / K: 30		30		
Activity scene	Kitchen			Kitchen		Kitchen, office, bathroom, etc.		
Number of categories	8			6		7		
Domains	D1	D2	D3	G	K	D1	D2	D3
Number of training segments	1543	2495	3897	1166	2582	570	633	421
Number of testing segments	435	750	974	291	646	142	159	106

temporal contrastive learning over graph representations and background mixing for domain-invariance.

In contrast to the aforementioned approaches, we build our network by following a uni-modal setting and an end-to-end fashion. For a fair comparison, we degenerated and evaluated existing networks on the setting in this paper.

III. PROPOSED FIRST-PERSON VIDEO UDA DATASETS

This section mainly introduces the limitations of existing benchmark dataset, the details and benefits of our datasets: ADL-7 and GTEA_KITCHEN-6. Key statistics for these datasets are presented in Table I.

As shown in Section II-A, EPIC-8 is the only and the best benchmark dataset for UDA problem on first-person video action recognition. Despite EPIC-8 being well-established and well-organized from EPIC, it still has some limitations. On the one hand, EPIC-8 provides cooking actions in a kitchen. However, there are other actions in daily human activities not occurring in the kitchen, e.g. *put toothpaste on a toothbrush* and *put bread on a plate* both belong to the same action *put*. Still, the previous one would not occur in a kitchen. Networks trained on videos from a single activity scene may not apply to the real world, where the scenes are more complicated. On the other hand, EPIC-8 is for within-dataset DA challenges because all domains are collected from EPIC. Within-dataset would lead to similar global video statistics among all domains, e.g. resolution, illumination, contrast, etc. As shown in Table II, the difference of RGB mean and standard deviation across domains in EPIC-8 is small. Similar global video statistics would lead to a smaller domain shift.

To solve these problems, we introduce our datasets based on three of the most commonly used first-person video benchmark datasets to provide more choices to evaluate UDA for first-person video action recognition. There are two main challenges: 1) Most existing datasets are annotated with actions (verb+noun). However, these annotations have very limited overlaps and are not directly usable for domain adaptation. 2) The new datasets should keep the compatibility of annotations with EPIC to better utilize all existing datasets. To overcome these challenges, we 1) separated verbs from original action annotations because verbs have more overlapping categories; 2) matched our categories to those in EPIC-8 and extracted overlapping categories from separated verbs; 3) manually unified and organized the names of categories with similar category names, e.g. mix/stir; 4) checked all the action videos and manually annotated those lacking annotations but having actions from the eight verb categories. Detailed descriptions of each dataset are provided below.

TABLE II
COMPARISON OF RGB MEAN AND STANDARD DEVIATION (STD) AMONG DATASETS FOR UDA PROBLEM ON FIRST-PERSON VIDEO ACTION RECOGNITION. G_K-6 REFERS TO GTEA_KITCHEN-6. GAP REFERS TO THE ABSOLUTE DIFFERENCE BETWEEN DOMAINS. WE PRESENT THE AVERAGE GAP IN EPIC-8 AND ADL-7.

Dataset	Domain	RGB Mean	RGB std
EPIC-8	D1	[0.385, 0.330, 0.323]	[0.269, 0.249, 0.237]
	D2	[0.447, 0.324, 0.275]	[0.200, 0.190, 0.176]
	D3	[0.392, 0.296, 0.247]	[0.213, 0.217, 0.207]
	Gap	[0.041, 0.023, 0.051]	[0.046, 0.039, 0.041]
ADL-7	D1	[0.393, 0.365, 0.284]	[0.172, 0.175, 0.168]
	D2	[0.440, 0.362, 0.231]	[0.177, 0.186, 0.164]
	D3	[0.411, 0.321, 0.215]	[0.194, 0.211, 0.175]
	Gap	[0.031, 0.029, 0.046]	[0.015, 0.024, 0.007]
G_K-6	G	[0.555, 0.430, 0.183]	[0.132, 0.139, 0.123]
	K	[0.252, 0.243, 0.268]	[0.188, 0.186, 0.191]
	Gap	[0.303, 0.187, 0.085]	[0.056, 0.047, 0.068]

A. ADL-7

Activity Daily Living (ADL) dataset is an activity dataset of human's real daily living in first-person camera views [18], containing 10-hour action videos by 20 persons in 20 different apartments. Each video records similar actions by different persons in various scenes, which provides daily human activities besides cooking for UDA research. To make the category consistent, we make our ADL-7 including the overlapping categories in EPIC-8. We chose video P4, P6 and P11 from ADL as three domains in ADL-7 because they include the most overlapping categories, namely *put*, *take*, *open*, *close*, *mix*, *pour* and *wash*, seven categories in total. We refer to P4, P6 and P11 as D1, D2 and D3 respectively. These videos comprise one hour and 22 minutes in length. The minimum length of each action in these videos is one second, while the maximum is 46 seconds.

We enlisted the assistance of three volunteers to select and annotate videos based on the following criteria: 1) All volunteers can discern action; 2) Volunteers can see hand-object interactions in action video; 3) Volunteers can identify background changes in the same action category across domains;

Fig. 2. Example frames of three domains in ADL-7 dataset. From left to right: *put*, *take*, *open*, *close*, *mix*, *pour*, *wash*. These example frames show activity scenes are not only in a kitchen but also in a bathroom, office, and living room.

TABLE III
THE NUMBER OF ACTION SEGMENTS IN ADL-7 DATASET. D1, D2 AND D3 REFER TO P4, P6 AND P11 RESPECTIVELY IN ADL DATASET.

Domain	Split	Verb category						
		put	*take*	*open*	*close*	*mix*	*pour*	*wash*
D1	Training	37	58	36	33	86	110	210
	Testing	9	15	9	8	22	27	52
	Sum	46	73	45	41	108	137	262
D2	Training	71	158	47	39	19	202	97
	Testing	18	40	12	10	5	50	24
	Sum	89	198	59	49	24	252	121
D3	Training	34	131	40	46	87	52	31
	Testing	8	33	10	12	22	13	8
	Sum	42	164	50	58	109	65	39

4) The videos of each domain include all action categories. We have collected action videos with categories that overlap with EPIC-8 from original videos P4, P6 and P11, resulting in 250 action videos. We then removed inappropriate action videos according to the aforementioned criteria, resulting in 222 action videos. Next, we annotated these action videos based on the original verb annotation in [54] by correcting incorrect annotations and adding annotations where they are absent. We finally segmented the selected action videos into action segments according to the reorganized annotations for data augmentation. We also split action segments into training and testing sets equidistantly in each category with a ratio of 8:2. In the training process, we randomly split the training set into training and validation with a ratio of 9:1.

These processes lead to a new benchmark dataset: ADL-7. This dataset is the first in this area to extend the activity scene from the kitchen to more others, e.g. living room, bathroom, office, as shown in Fig. 2. This would benefit researchers to develop networks that are more applicable in the real world. Table III presents details about each category in ADL-7. Three domains include 2031 extracted action segments in total. The numbers of training segments in the three domains are 570, 633 and 421, while those of testing are 142, 159 and 106.

B. GTEA_KITCHEN-6

GTEA [19], [55] and KITCHEN [20] datasets are first-person video datasets recording actions in the kitchen. Their actions, like their names, are related to cooking. GTEA videos are filmed in the real kitchen, while KITCHEN videos are filmed in a temporary-built kitchen in the lab. Therefore, they have disparate video statistic. As shown in Table II, the RGB mean difference between GTEA and KITCHEN is [0.303, 0.187, 0.085], which is significantly bigger than that in EPIC-8 [0.041, 0.023, 0.051]. We use the GTEA and KITCHEN datasets as two domains to provide a cross-site setting, which is absent from EPIC datasets. Additionally, we consider category consistency.

We first collected all 115 action videos from the original GTEA dataset [19] as domain G, as each action video satisfied the requirements in the aforementioned criteria. The total length of the videos is around 35 minutes. The minimum length is one second, and the maximum is about 10 seconds. Note that all videos in GTEA are uninterrupted, continuous action. Therefore, GTEA has a shorter video length but more action videos. We then annotated action videos based on their original annotation for overlapping with ADL-7 and EPIC-8, resulting in the reduction of 55 categories to six: *put*, *take*, *open*, *close*, *mix*, *pour*. Each verb category corresponds to multiple action categories from the original dataset. We finally followed the same setting as ADL-7 to segment actions and create training/validation/testing sets.

For KITCHEN, due to the fact that few of provided action videos met the aforementioned criteria, we manually generated 339 action videos from the original KITCHEN dataset [20] as domain K after viewing each original video. The total length is about one hour and 36 minutes. The minimum and maximum length are one second and 80 seconds, respectively. We then annotated all action videos with six overlapping verb categories for our dataset. Note that the number of *mix* action segments is 3272, which is larger than the sum of other categories. In our experiment, we randomly selected a quarter of them to make *mix* have similar sample sizes with the second most category *pour* because we excluded extremely imbalanced UDA in this paper.

We finally presented the first benchmark dataset with the cross-site setting: GTEA_KITCHEN-6 dataset. As shown in Table IV, this dataset includes 1166 training segments and 291 testing segments from GTEA, 2582 training segments and 646 testing segments from KITCHEN. Fig. 3 shows resolutions of the two domains are different, where GTEA data is 456 $\times$

TABLE IV
THE NUMBER OF ACTION SEGMENTS IN GTEA_KITCHEN-6 DATASET. G REFERS TO GTEA DATASET AND K REFERS TO KITCHEN DATASET.

Domain	Split	Verb category					
		put	*take*	*open*	*close*	*mix*	*pour*
G	Training	139	310	214	117	84	302
	Testing	35	77	53	29	21	76
	Sum	174	387	267	146	105	378
K	Training	243	354	346	144	654	841
	Testing	61	88	87	36	164	210
	Sum	304	442	433	180	818	1051

(a) GTEA

(b) KITCHEN

Fig. 3. Example frames of all categories in GTEA_KITCHEN-6 dataset. First row from left to right: *put*, *take*, *open*. Second row from left to right: *close*, *mix*, *pour*.

256 and KITCHEN is 342 $\times$ 256. In addition, differences in global video statistics between domains are more significant , representing more considerable illumination and contrast difference. These would be new challenges for UDA for first-person video action recognition.

C. Benefits of Our Datasets

As shown in Table V and discussed above, our datasets, ADL-7 and GTEA_KITCHEN-6, have the following advantages over existing datasets: 1) They combine and re-annotate the original datasets (ADL, GTEA, and KITCHEN) to accommodate the multi-domain setting. 2) They provide five domains in addition to the original three from EPIC-8. This significantly expands the options available to researchers in this area for developing UDA for first-person videos. 3) They apply additional real-world activity scenes, enabling researchers to develop approaches for the complex multi-scene problem. 4) They are constructed from different benchmark datasets to support cross-site domain adaptation study, e.g. multi-source domain adaptation. In addition, they are compatible with the existing benchmark dataset, making it easy to establish new cross-domain correspondences. This enables researchers to develop UDA across all these domains. As shown in Fig. 5, their categories are imbalanced, leading to another study-worthy challenge: class-imbalanced DA. In conclusion, our datasets would enrich benchmark datasets, promote diversity, ensure compatibility and scalability, and accelerate the development of this area. In this paper, we will additionally study source-only and target-only recognition accuracy. The target-only setting means training and testing are both on the target dataset. In contrast, the source-only setting means the network trained on the source dataset is directly tested on the target dataset without UDA. These two results serve as the upper and lower bounds of UDA on these datasets.

TABLE V
COMPARISON OF DATASET SETTINGS WITH OTHER RELATED FIRST-PERSON DATASETS. G_K-6 REFERS TO GTEA_KITCHEN-6.

	First-person	Multi-domain	Multi-scene	Cross-site
ADL [18]	✓	×	×	×
GTEA [19]	✓	×	×	×
KITCHEN [20]	✓	×	×	×
EPIC-97 [16]	✓	✓	×	×
EPIC-8 [17]	✓	✓	×	×
ADL-7 (ours)	✓	✓	✓	×
G_K-6 (ours)	✓	✓	×	✓

IV. PROPOSED ATTENTION-BASED METHOD

This section outlines our proposed attention modules and integrated network for UDA for first-person video action recognition. Fig. 4 shows our proposed UDA network named as Channel-Temporal Attention Network (CTAN). In training, source and target videos are fed into a feature extractor that modifies the I3D [6] pretrained on ImageNet [56] by adding multiple channel and temporal attention (CTA) modules. Each proposed CTA module consists of a channel attention module and a temporal attention module, and is inserted into I3D to re-calibrate channel- and temporal-wise features. After feature extraction, source features are fed into an action classifier, and both source and target features are fed into a domain classifier for adversarial domain discrimination. In testing, only target videos are used as the input to the feature extractor to extract features for the action classifier to predict the action category.

A. Channel-Temporal Attention Module

In first-person video action recognition, different channels of CNN layers capture different spatio-temporal information from actions. Such information can benefit domain adaptation for action recognition. Firstly, inspired by the SENet [32] that excites informative features in input image channels, we extend the SENet to channel-wise attention (CA) module for video input, as shown in Fig. 6a. Secondly, human can usually recognize an action at a glance as long as they see small but informative temporal parts of this action. This phenomenon inspires us to extend the previous CA module to the temporal-wise attention (TA) module. This module can capture the temporal attention weights from the features. Applying the

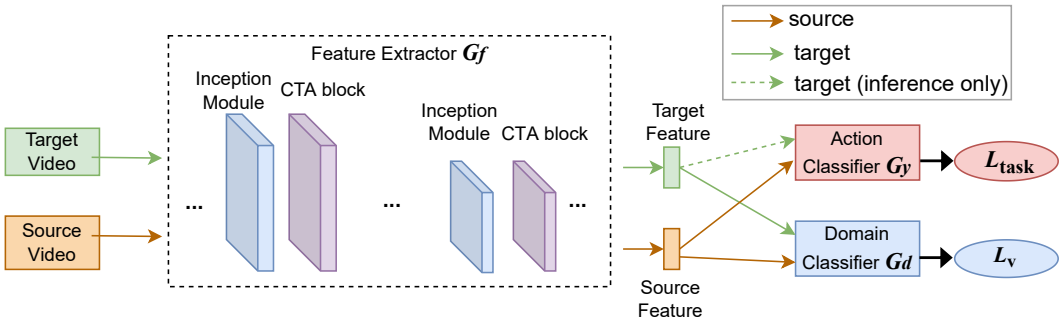


Fig. 4. Architecture of the proposed Channel-Temporal Attention Network (CTAN) for first-person video action recognition. A feature extractor, which is composed of Inception modules [6] and proposed Channel-Temporal Attention (CTA) modules, is shared by both source and target domains. The feature extractor takes labeled source videos and unlabeled target videos as the input and generates corresponding features as the output in training. Source features are fed into both action and domain classifiers, while target features are only fed to domain classifier. In testing, target videos are the only input to the feature extractor and then the action classifier.

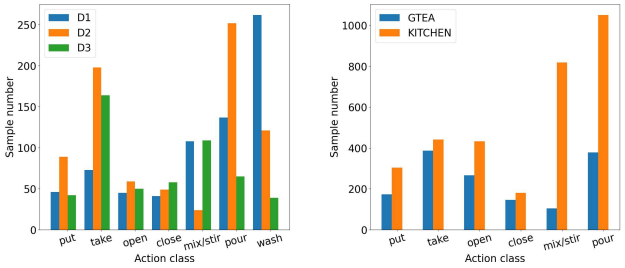


Fig. 5. The distribution of categories in our datasets.

weights to features can excite parts with important temporal information. These excited features are analogous to the informative temporal parts of actions. Finally, we integrate the CA and TA modules, as shown in Fig. 6b, into the channel-temporal attention (CTA) module described in detail below.

Given a 5D video feature $\mathbf{X} \in \mathbb{R}^{N\times T\times C\times H\times W}$. N, T and C denote batch size, temporal dimension and feature channel size. H and W correspond to height and width. First, we utilize 3D average pooling to extract channel-wise information among dimensions T, H and W, i.e.,

$$\mathbf{X}^c = \frac{1}{T\times H\times W}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{X}_{:,t,:,h,w}, \tag{1}$$

where $\mathbf{X}^c \in \mathbb{R}^{N\times 1\times C\times 1\times 1}$ is the squeezed feature.

Then, we capture the channel-reduced feature $\mathbf{X}^{cr} \in \mathbb{R}^{N\times 1\times C/r\times 1\times 1}$ for efficiency by a linear layer with parameters W^c and a reduction ratio r, as follows:

$$\mathbf{X}^{cr} = \mathrm{ReLU}(\mathbf{W}^c\mathbf{X}^c). \tag{2}$$

Another linear layer is used with parameters $\mathbf{W}^{cr}$ to restore the feature channel dimension and a sigmoid function σ is used to capture channel-attentive weights $\mathbf{A}^c \in \mathbb{R}^{N\times 1\times C\times 1\times 1}$. In order to excite the informative channels, we compute a Hadamard product $\odot$ between these weights $\mathbf{A}^c$ and the video feature $\mathbf{X}$ as

$$\mathbf{X}^{co} = \mathbf{X} + \mathbf{A}^c \odot \mathbf{X} = \mathbf{X} + \sigma(\mathbf{W}^{cr}\mathbf{X}^{cr}) \odot \mathbf{X}, \tag{3}$$

where $\mathbf{X}^{co} \in \mathbb{R}^{N\times T\times C\times H\times W}$ denotes the output of the channel attention module with the excited and enhanced channel-wise informative features. Considering that wrong channel

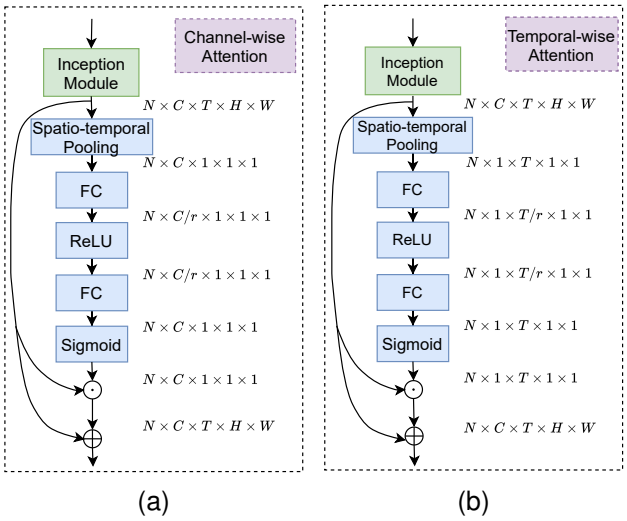


Fig. 6. Architecture of the channel-temporal attention module. (a) Channel-wise attention module. (b) Temporal-wise attention module.

attention may hurt the performance to some degree and some channel attention may suppress other information, we add a residual connection to mitigate these negative effects.

We then conduct a 3D average pooling on the video feature $\mathbf{X}^{co}$ among C, H and W to extract the squeezed temporal feature $\mathbf{X}^t \in \mathbb{R}^{N\times T\times 1\times 1\times 1}$, which is expressed by

$$\mathbf{X}^t = \frac{1}{C\times H\times W}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{X}^{co}_{:,:,c,h,w}. \tag{4}$$

We adopt one linear layer with the parameter W^t to obtain temporal-reduced feature $\mathbf{X}^{tr} \in \mathbb{R}^{N\times T/r\times 1\times 1\times 1}$, i.e.,

$$\mathbf{X}^{tr} = \mathrm{ReLU}(\mathbf{W}^t\mathbf{X}^t), \tag{5}$$

and another with parameter $\mathbf{W}^{tr}$ to restore the features. The sigmoid function σ is again adopted to obtain the temporal-attentive weights $\mathbf{A}^t \in \mathbb{R}^{N\times T\times 1\times 1\times 1}$. A residual connection is also applied to prevent temporal attention from suppressing other information and obtain the excited output $\mathbf{X}^{cto} \in \mathbb{R}^{N\times T\times C\times H\times W}$. The final excitation function can be formulated as

$$\mathbf{X}^{cto} = \mathbf{X}^{co} + \mathbf{A}^t \odot \mathbf{X}^{co} = \mathbf{X}^{co} + \sigma(\mathbf{W}^{tr}\mathbf{X}^{tr}) \odot \mathbf{X}^{co}. \tag{6}$$

Comparison with SENet. While CTAN and SENet have a similar structure of channel-wise attention, CTAN differs from SENet in the following two aspects: 1) A new dimension of channel-wise attention. The channel-wise attention of CTAN has an additional temporal dimension, T, in comparison to the structure of that of SENet. This additional dimension gives channel-wise attention of CTAN the ability of temporal modelling and video application; 2) A new type of attention. CTAN provides both temporal and channel-wise attention, whereas SENet only provides channel-wise attention. The additional temporal-wise attention can assign attention weights for each feature frame to improve the weighting of temporal dimension, hence further enhancing temporal modelling.

B. Adversarial UDA

For UDA problem, the network needs to learn common features across domains while focusing on the important features. For convenience, discrepancy-based and adversarial-based approaches such as Deep Adaptation Network (DAN) [22] and DANN [21] are easy to be adapted to our task compared with reconstruction-based UDA. In comparison, linear DAN needs a large batch size to avoid negative MMD loss, resulting in high computational demand than DANN.

In this paper, considering the limited computation resources, we use the DANN [21] in which a two-player mini-max game is constructed, but still compare its recognition performance with DAN [22] in Section V. The main idea of DANN is to add one domain classifier G_d to discriminate whether the data is from the source or target domain. G_d are trained by minimizing the discriminator loss L_d, while feature extractor G_f are trained by maximizing L_d. The aim is to outwit the discriminator to guide the feature extractor to learn common features between the source and target domain. Here, we utilize a discriminator as in DANN to align features extracted by the feature extractor across domains. The domain loss L_v is defined for each video input x_i as:

$$L_v = -\frac{1}{n} \sum_{x_i \in D_s \cup D_t} L_d(G_d(G_f(x_i)), d_i), \tag{7}$$

where D_s and D_t are source domain and target domain, respectively, n is the number of sample from both domains. d_i is the domain label of x_i. If x_i is from the source (target) domain, d_i is set as 1 (0).

C. Integration with I3D Network

Finally, we integrate the proposed modules and adversarial UDA into I3D, as illustrated in Fig. 4. Following the finding in [32] that lower layer features are typically more general, while higher layer features have greater specificity, we integrate our proposed modules into 3rd to 7th Inception modules in the I3D architecture. The domain classifier G_d and an action classifier G_y are integrated after the average pooling layer of I3D. As shown in Fig. 7, the action classifier yields the task classification loss L_{task} while the domain classifier yields the domain discriminator loss L_v. The domain classifier is composed of a binary classifier and a GRL. The binary classifier is used to discriminate between source and target

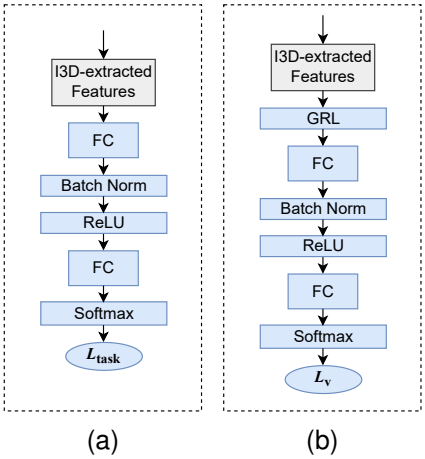


Fig. 7. Architecture of (a) the action classifier and (b) the domain classifier.

input features. GRL is used to invert the gradient during back-propagation, i.e. it multiplies the gradient by -1. The overall loss L can be expressed as follows:

$$\begin{aligned} L =& L_{task} + \lambda_v L_v \\ =& \frac{1}{n_s} \sum_{x_i \in D_s} L_y(G_y(G_f(x_i)), y_i) \\ & - \frac{\lambda_v}{n} \sum_{x_i \in D_s \cup D_t} L_d(G_d(G_f(x_i)), d_i), \end{aligned} \tag{8}$$

where λ_v is a hyper-parameter to trade-off domain adaptation with classification respectively. y_i refers to the action labels of input x_i. The whole network is trained by two cross entropy loss, L_y and L_d.

V. Experiments

We evaluate our proposed method on the three datasets: ADL-7, GTEA_KITCHEN-6 and EPIC-8 against other image-based UDA [21], [22], [50], video-based UDA [17], [24], and attention mechanisms [32]–[34].

A. Experimental Setup

Datasets. For EPIC-8, we conduct our experiment using the same settings as MM-SADA [17]. For our datasets, we first download all videos from official websites of these datasets. Then, we extract frames from videos at their respective sampling rates, which are 15 fps for GTEA and 30 fps for both ADL and KITCHEN. We finally restructure the annotations in accordance with the introduction in Section III-A and Section III-B. We follow a similar experimental protocol as EPIC-8. For data augmentation, we sample every 16-frame segment from each action video and make adjacent samples with a 4-frame overlap. All the segments are randomly divided into training and testing sets at a ratio of 8:2, as shown in Table I.

Selected hyper parameters. We utilize I3D as our feature extraction backbone and train our network end-to-end. Both the domain classifier and the action classifier are composed of two fully connected layers with a dimension of 100, a batch normalization layer, a ReLU activation function and a soft-max activation for generating predictions, as shown in

Fig. 7. We select the outputs of the final average pooling layer in I3D as the inputs for the domain classifier and the action classifier. These outputs are with a dimension of 1024. In the training stage, we use both labeled source data and unlabeled target data, however in the testing stage, we only use unlabeled target data. Inputs for both source and target are 16-frame segments sampled from the action video, with each frame scaled to 256 $\times$ 256 and then randomly cropped to 224 $\times$ 224. The optimization is performed using SGD with a momentum of 0.9 and batch size of 16. A weight decay with 5e-4 is applied to all parameters. The training stage consists of two stages. Firstly, we initialize λ_v to 0 and train the feature extractor and classifier for 10 epochs at a learning rate of 1e-2. Secondly, we follow the same strategy in [21] to increase λ_v from 0 to 1 and reduce the learning rate to train the overall network for additional 20 epochs. Code has been released at *https://github.com/XianyuanLiu/CTAN* using PyKale library [57] based on PyTorch. The experiments were conducted on an Intel Core i7-5930 3.50 GHz × 12 with 32 GB of RAM and 1 NVIDIA Titan Xp GPU with 12 GB of memory. The system uses Linux Ubuntu 18.04 with NVIDIA CUDA 10.0.

B. Comparison with Other UDA Networks

Image-based Baselines. We first extend three state-of-the-art image-based UDA networks, i.e. DAN [22], DANN [21] and CDAN [50], from image applications to video applications as our baselines. We modify three components in these networks (feature extractor, task classifier, and domain network) as outlined below. 1) replace the image feature extractor (e.g. ResNet) with a video feature extractor (I3D pretrained on ImageNet); 2) build the task classifier and domain network separately, using two fully connected layers with a dimension of 100 and a ReLU activation function; 3) add an average-pooling layer to the tail of the feature extractor to make the dimensions of the three components compatible. We follow the same experimental settings of these UDA networks for a fair comparison.

Video-based Baselines. We choose MM-SADA [17] and $\mathrm{TA^3N}$ [24] as the baselines. The default MM-SADA has a two-stream architecture accepting multi-modal input. To make the comparison fair, we develop a uni-modal MM-SADA by freezing the self-supervision alignment classifier within MM-SADA. We implement $\mathrm{TA^3N}$ in two different configurations: default and improved. For default implementation of $\mathrm{TA^3N}$, we utilize ResNet-101 as the backbone to convert the RGB frames to feature frames with a dimension of 2048. These feature frames can serve as input for $\mathrm{TA^3N}$. Considering that CTAN utilizes I3D as the backbone, we also develop an improved implementation of $\mathrm{TA^3N}$ with I3D as the backbone for a more fair comparison. We are unable to implement the uni-modal version of STCDA [25] owing to the lack of publicly available code. Considering that different platforms and configuration environments would affect the reproduction results, we evaluate CTAN and uni-modal STCDA using their accuracy differences with uni-modal MM-SADA in [25]. CMCDA [26] and CoMix [27] are excluded from the comparison with CTAN. CMCDA, unlike MM-SADA, is designed specifically for multi-modal input and developing a uni-modal version would remove much of its originality. CoMix is designed for video inputs that require intensive data argumentation, e.g. background extraction and mixing. It is inappropriate to compare CoMix to CTAN, which is developed just for raw video input. We repeat each experiment three times with different random seeds and report the average accuracy on testing target set. The best result for each task is highlighted in **bold**, and the second best is underlined.

EPIC-8. We first evaluate our network using six tasks from EPIC-8. The results are shown in Table VI. CTAN achieves the highest recognition accuracy across all tasks, whereas $\mathrm{TA^3N}$ with I3D backbone achieves the most second-best results. For image-based networks, DANN outperforms CDAN and DAN on average and improves the source-only baseline by 0.3%. CTAN outperforms DANN by 1.2% on average, proving the efficacy of proposed channel-temporal attention modules. Video-based networks with I3D backbone outperform DANN by 0.6% (MM-SADA) and 0.9% ($\mathrm{TA^3N}$), respectively, whereas CTAN still performs better than video-based baselines by 0.9% and 0.6% respectively. In addition, MM-SADA improves source-only in five out of six tasks, while our CTAN and $\mathrm{TA^3N}$ with I3D backbone improves all tasks. CTAN outperforms $\mathrm{TA^3N}$ with I3D in all tasks, with the exception of D2→D3, where they performed comparably. In addition, according to Table 3 of Ref [25], uni-modal STCDA with RGB input outperforms uni-modal MM-SADA in four out of six tasks on the EPIC-8 dataset and improves the best accuracy by 0.7% on average. Nevertheless, as shown in Table VI, our CTAN outperforms uni-modal MM-SADA in all tasks and improves the best accuracy by 0.9% on average, indicating the effectiveness of our CTAN.

ADL-7. We then evaluate CTAN on ADL-7. As shown in Table VI, in general, all networks improve the source-only baseline on average and our network improves significantly in five out of six tasks by 2.6% on average. For image-based networks, CDAN outperforms DANN and DAN in four out of six tasks, achieving the second-best performance on average among all networks. CTAN outperforms CDAN in four out of six tasks by 0.8% on average and achieves the second-best results in the remaining two off-target tasks. Both video-based networks improve the source-only baseline in four out of six tasks and $\mathrm{TA^3N}$ with I3D backbone exceeds MM-SADA by 0.9% on average. CTAN continues to outperform $\mathrm{TA^3N}$ with I3D backbone by 1.1% on average. For D1→D2 and D1→D3 tasks, only CTAN and $\mathrm{TA^3N}$ achieve better recognition accuracy than the source-only baseline. For D2→D1 and D2→D3 tasks, CTAN performs much better than DANN (by 3.9% and 5.1% respectively), MM-SADA (by 3.5% and 4.0% respectively) and source-only (by 6.0% and 4.1% respectively), despite not outperforming CDAN (D2→D1) and DAN (D2→D3).

GTEA_KITCHEN-6. We finally evaluate our proposed network using GTEA_KITCHEN-6 dataset, as shown in Table VI. Our CTAN achieves the highest average performance, 3.0% better than the source-only baseline. On the G→K task, image-based networks, excluding CDAN, improve the source-only baseline by 1.4% (DANN) and 1.0% (DAN) respectively.

TABLE VI
BEST ACCURACY (%) ON THE TARGET DOMAIN BY CTAN, COMPARED TO OTHER UDA APPROACHES ON EPIC-8 [17] AND OUR INTRODUCED ADL-7 AND GTEA_KITCHEN-6 (G_K-6) DATASETS. SO: SOURCE-ONLY, TO: TARGET-ONLY, WHICH ARE THE WORST AND BEST ACCURACY CAN BE EXPECTED FOR UDA ON THESE DATASETS. GAIN: IMPROVEMENT COMPARE TO SO. THE BEST UDA RESULT FOR EACH TASK IS IN **BOLD**, AND THE SECOND BEST IS UNDERLINED. NOTE THAT MM-SADA HERE IS THE UNI-MODAL VERSION FOR A FAIR COMPARISON WITH OTHER NETWORKS.

Method		TO	SO	DANN [21]	CDAN [50]	DAN [22]	MM-SADA [17]	TA^{3}N [24]		CTAN (Ours)
Backbone		I3D	I3D	I3D	I3D	I3D	I3D	ResNet-101	I3D	I3D
EPIC-8 [17]	D1→D2	64.7	39.4	39.4	40.7	36.3	40.1	33.1	39.6	**41.3**
	D1→D3	52.8	32.0	32.9	30.3	34.1	33.2	30.6	34.0	**35.0**
	D2→D1	60.2	35.5	36.2	36.4	36.1	36.1	32.0	36.4	**36.6**
	D2→D3	52.8	39.2	40.2	40.5	39.4	40.2	28.7	**40.6**	**40.6**
	D3→D1	60.2	38.1	37.6	38.2	37.9	38.1	28.5	38.6	**39.3**
	D3→D2	64.7	40.5	40.3	40.0	40.7	40.7	27.7	40.9	**41.3**
	Mean	59.2	37.5	37.8	37.7	37.4	38.1	30.1	38.4	**39.0**
	Gain	21.7	-	0.3	0.2	-0.1	0.6	-7.4	0.9	**1.5**
ADL-7	D1→D2	95.8	41.1	40.6	41.1	35.9	40.8	**47.4**	44.6	43.2
	D1→D3	93.5	28.6	28.1	27.3	26.6	28.1	23.5	29.0	**31.5**
	D2→D1	95.1	25.0	27.1	**34.0**	25.7	27.5	25.0	28.7	31.0
	D2→D3	93.5	24.8	23.8	26.2	**31.1**	24.9	28.4	28.9	28.9
	D3→D1	95.1	27.4	29.5	23.6	**31.2**	28.6	17.8	24.5	26.7
	D3→D2	95.8	37.5	37.5	**42.7**	36.5	38.0	32.1	37.4	38.2
	Mean	94.8	30.7	31.1	32.5	31.2	31.3	29.0	32.2	**33.3**
	Gain	64.1	-	0.4	1.8	0.5	0.6	-1.7	1.5	**2.6**
G_K-6	G→K	95.9	36.8	38.2	34.0	37.8	38.7	37.8	39.2	**41.1**
	K→G	94.5	45.9	46.5	**48.4**	43.4	46.5	26.2	47.4	47.6
	Mean	95.2	41.4	42.4	41.2	40.6	42.6	32.0	43.3	**44.4**
	Gain	53.8	-	1.0	-0.2	-0.8	1.2	-9.4	1.9	**3.0**

Each video-based network with I3D backbone improves the source-only baseline by 1.9% (MM-SADA) and 2.4% (TA^{3}N) respectively. CTAN outperforms the source-only baseline by 4.3% and image-based networks DAN by 3.3%, CDAN by 7.1% and DANN by 2.9%, as well as video-based MM-SADA by 2.4% and TA^{3}N with I3D backbone by 0.9%. On the K→G task, CDAN obtains the best performance, surpassing DANN by 1.9% and DAN by 5.0%. Although it inferior to that of CDAN, the performance of CTAN remains superior to other image- and video-based networks. On average, CTAN has the highest accuracy, surpassing DANN by 2.0%, CDAN by 3.2%, MM-SADA by 1.8% and TA^{3}N with I3D backbone by 1.1%. In comparison, CDAN performs the best on the K→G task, but the lowest (34.0%) in the G→K task.

CTAN achieves the best overall performance of these state-of-the-art image- and video-based UDA on the three datasets and has significantly improved the source-only baseline. The improvement is consistent across all pairs of domains in 13 out of 14 tasks. In contrast, other networks improve the source-only baseline in 7 (DANN), 8 (CDAN), 9 (DAN), 11 (MM-SADA) and 12 (TA^{3}N with I3D backbone) out of 14 tasks. The only off-target is the D3→D1 task in ADL-7, which is reasonable considering that D3 is the smallest and most imbalanced domain in this dataset. The Pearson correlation [58] between sample numbers from three domains demonstrates that D3 has the lowest average value (0.23), compared to D1 (0.24) and D2 (0.28). Therefore, D3 has the lowest correlation with the sample number of the other domains, making it the most imbalanced domain. It is still difficult for UDA to transfer knowledge from a small dataset with imbalanced categories to a large dataset. Similarly, the D3→D2 task is improved slightly by CTAN as shown in Table VI. On all three datasets, the default implementation of TA^{3}N with ResNet-101 as the backbone performs poorly, missing the target in the most tasks and achieving the lowest average accuracy. TA^{3}N with an I3D backbone surpasses the default implementation in 13 out of 14 tasks, demonstrating that 3D CNN is capable of providing better feature representation for spatio-temporal information. Furthermore, although CTAN achieves the best performance on two other single-scene datasets and the best average performance on ADL-7, it is not the best in some specific ADL-7 tasks. The reason is that ADL-7 captures human daily life in multiple scenes, with varying proportions of each scene in each domain, e.g. D1 contains a greater proportion of bathroom scenes than D2, whereas D2 contains more kitchen scenes. These varying proportions result in more complicated domain gaps across domains, such as considerable background and object differences, which is a greater challenge to the performance and robustness of methods compared to the single-scene problem. For example, task D1 → D2 would require a stronger focus on spatial characteristics. Therefore, TA^{3}N, which has an additional spatial domain network, outperforms our method, even though it is based on a 2D CNN backbone (ResNet-101). UDA in this multi-scene setting is a good future direction.

C. Comparison with Other Attention Modules

We then compare our proposed attention to two state-of-the-art attention modules, CBAM [33] and SRM [34], becuase CBAM and SRM have the ability to excite the channel-wise informative features, exceeding SENet in some applications. The original versions of them are designed for image applications. We first extend them to videos by adding the

TABLE VII
BEST ACCURACY (%) ON THE TARGET DOMAIN FOR OUR PROPOSED CTAN, COMPARED TO OTHER ATTENTION APPROACHES ON EPIC-8 [17], ADL-7 AND GTEA_KITCHEN-6 (G_K-6) DATASETS. GAIN: IMPROVEMENT COMPARED TO DANN [21].

Datasets	Tasks	DANN [21]	CBAM [33]	SRM [34]	CTAN (Ours)
EPIC-8 [17]	D1→D2	39.4	40.4	37.9	**41.3**
	D1→D3	32.9	35.6	**36.9**	35.0
	D2→D1	36.2	36.8	**37.6**	36.6
	D2→D3	40.2	40.4	**40.8**	40.6
	D3→D1	37.6	35.3	36.9	**39.3**
	D3→D2	40.3	40.8	38.6	**41.3**
	Mean	37.8	38.2	38.1	**39.0**
	Gain	-	0.4	0.3	**1.2**
ADL-7	D1→D2	40.6	41.0	42.5	**43.2**
	D1→D3	28.1	29.5	27.4	**31.5**
	D2→D1	27.1	28.4	28.5	**31.0**
	D2→D3	23.8	28.4	**29.1**	28.9
	D3→D1	**29.5**	27.6	26.2	26.7
	D3→D2	37.5	38.5	**39.0**	38.2
	Mean	31.1	32.2	32.1	**33.3**
	Gain	- -	1.1	1.0	**2.2**
G_K-6	G→K	38.2	38.4	38.6	**41.1**
	K→G	46.5	46.6	46.3	**47.6**
	Mean	42.4	42.5	42.5	**44.4**
	Gain	-	0.1	0.1	**2.0**

temporal dimension, T. For CBAM, 3D pooling layers and 3D convolutional layers are applied to generate spatial attention. For SRM, we replace the 2D batch normalization layer with a 3D layer and calculated the mean and standard deviation of the input features in three dimensions (temporal, height and weight). This addition not only adds the capacity for temporal modeling, but also preserves the characteristics of these attentions. We use the same settings as CTAN to evaluate them. Note that we apply residual connection to them for a fair comparison as they do not have. We evaluate them on all datasets and compare their best accuracy in Table VII.

EPIC-8. We first evaluate our network on all tasks of EPIC-8. Table VII shows all attention networks improve the baseline. CBAM improves the baseline by 0.4% on average but fails in the D3→D1 task. SRM achieves comparable performance with CBAM and have more best results. However, SRM fails in three out of these six tasks (D1→D2, D3→D1 and D3→D2). CTAN outperforms CBAM and SRM on average recognition accuracy, exceeding them by 0.8% and 0.9% respectively, and improving the baseline in all six tasks.

ADL-7. We evaluate CTAN on ADL-7. Table VII shows CBAM and SRM bring comparable improvement in baseline by 1.1% and 1.0% respectively, while SRM fails in two tasks (D1→D3 and D3→D1). All approaches fail to improve the baseline in task D3→D1. Although CTAN did not outperform CBAM (D3→D1 and D3→D2) and SRM (D2→D3 and D3→D2), it still achieved the best average performance, surpassing CBAM and SRM by 1.1% and 1.2%, respectively.

GTEA_KITCHEN-6. We finally analyze these networks using our new cross-site dataset, GTEA_KITCHEN-6. Table VII shows CBAM and SRM achieve equivalent recognition accuracy on this cross-site dataset, whereas they only benefit UDA slightly. However, CTAN outperforms other approaches by improving in both tasks, G→K and K→G (by 2.9% and 1.1%, respectively). On average, CTAN performs better than CBAM and SRM by 1.9%.

On the three datasets with within-dataset and cross-site settings, CTAN performs better than the other state-of-the-art attention modules. CBAM and SRM enhance the baseline in six and eight out of 14 tasks respectively. CTAN makes the improvement in 13 out of 14 tasks and achieves the most best results among three datasets, showing the significance of modeling temporal-wise inter-dependencies.

D. Ablation Studies and Further Analysis

Practical effectiveness of our datasets. We discuss the practical effectiveness of our datasets. On the one hand, Table II shows ADL-7 has a small difference in RGB mean and standard deviation across domains because ADL-7 videos are collected from a single dataset ADL, similar to EPIC-8. The smaller difference indicates similar global video statistics, resulting in a reduced domain gap. Nevertheless, as long as datasets contain domain gaps, datasets with a small global difference are still practically valuable for studying UDA problems. As shown in Table VI, the difference in average accuracy between source-only and target-only for EPIC-8 is 21.3%. However, it is 64.1% for ADL-7, which is significantly higher than EPIC-8, indicating that ADL-7 has a more significant domain gap than EPIC-8. The multi-scene setting brings more diverse backgrounds and objects, resulting in a larger domain gap. On the other hand, although large-scale datasets will be better, smaller datasets that are far from saturation are still valuable for advancing UDA research. Table I shows ADL-7 is smaller than EPIC-8, but as discussed above, ADL-7 has a larger domain gap. Furthermore, ADL-7 also presents a bigger challenge than EPIC-8 in Table VI. The difference of average accuracy between the best method (CTAN) and target-only for ADL-7 is 61.5%, significantly larger than that for EPIC-8 (20.2%), indicating that ADL-7 is far from being saturated. ADL-7 has a large domain gap and is not yet saturated, so it is still practically usable. Similar to ADL-7, the difference of average accuracy between source-only and target-only for GTEA_KITCHEN-6 is 53.8%, while the difference of average accuracy between the best method (CTAN) and target-only for GTEA_KITCHEN-6 is 50.8%. Both are larger than those of EPIC-8 (21.3% and 20.2% respectively). Hence, GTEA_KITCHEN-6 is also practically applicable. CTAN can narrow the gap for ADL-7 and GTEA_KITCHEN-6 by 2.9% and 2.6%, respectively, but there is still a large room for further improvement. This indicates that our datasets are challenging and far from saturated, allowing researchers to facilitate further research in this area.

Importance of our datasets. The recognition performance on a single dataset may not be sufficient for robust network exploration. Two additional datasets with distinct settings would strengthen the validity of this conclusion. Despite the fact that Table VI shows that CDAN is the second-best network on average in the within-dataset setting, we are aware that CDAN may not perform well on some cross-site domains benefited from our new cross-site dataset. Similarly, some im-

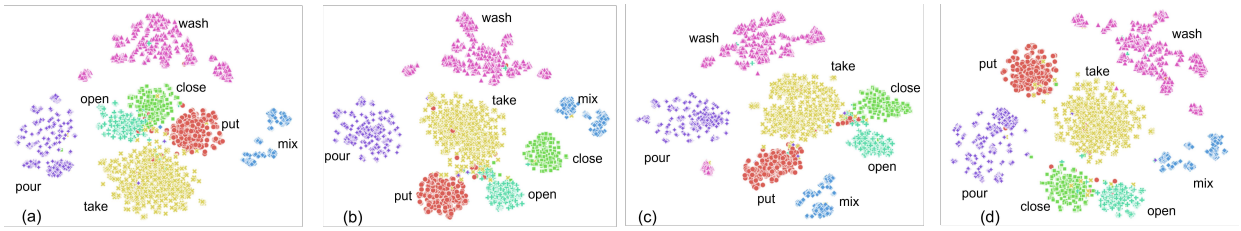


Fig. 8. t-SNE visualization on ADL-7 produced by (a) backbone without attention module, (b) with CBAM, (c) with SRM, and (d) with our proposed module.

TABLE VIII
THE EFFECT OF THE RESIDUAL CONNECTION. BEST AVERAGE ACCURACY (%) ON THE TARGET DOMAIN ACROSS ALL DOMAINS ON ADL-7 AND GTEA_KITCHEN-6 (G_K-6) DATASETS. ▲ / ▼: IMPROVEMENT / DECLINE COMPARED TO THE BASELINE [21].

		ADL-7		G_K-6	
Baseline [21]		31.1		42.4	
CBAM [33]	- w/o residual	29.6	▼-1.5	41.8	▼-0.6
	- w/ residual	32.2	▲+1.1	42.5	▲+0.1
SRM [34]	- w/o residual	30.9	▼-0.2	40.3	▼-2.1
	- w/ residual	32.1	▲+1.0	42.5	▲+0.1

TABLE IX
ABLATION STUDY OF THE ATTENTION ARRANGEMENTS ON ADL-7. BEST ACCURACY (%) BY FOUR APPROACHES OF ARRANGING THE CHANNEL AND TEMPORAL ATTENTION MODULES ARE REPORTED.

	w/o Attention	CAN	TAN	TCAN	CTAN
D1→D2	40.6	**44.8**	40.9	40.6	43.2
D1→D3	28.1	27.6	27.9	29.4	**31.5**
D2→D1	27.1	30.6	25.7	30.3	**31.0**
D2→D3	23.8	27.8	**28.9**	**28.9**	**28.9**
D3→D1	29.5	19.8	24.0	**31.9**	26.7
D3→D2	37.5	**39.9**	37.2	31.3	38.2
Mean	31.1	31.8	30.8	32.1	**33.3**
Gain	-	+0.7	-0.3	+1.0	+2.2

provements should be made before using DAN for both cross-site and multi-scene setting. Our proposed network, CTAN, achieves the best performance among all datasets, proving its robustness for multiple scenarios, e.g. single-scene, multi-scene and cross-site. Our new datasets increase the variety of available datasets, promotes diversity in this area, enabling researchers to facilitate more credible and convincing studies.

Qualitative results. Fig. 8 visualizes the distribution of the features learned without attention, with CBAM, with SRM and with our proposed module in 2D via t-SNE [59]. The visualization shows that our proposed channel-temporal attention (CTA) yields a better separation of classes, particularly in the categories *put*, *take*, *open*, and *close*. Among these categories, *open* and *close* are the two most inseparable actions, as shown in Fig. 8(a). One possible reason is the spatial information is similar in the two actions but in an opposite temporal order. Attention modules embed them in two more separable clusters as shown in Fig. 8(b)(c)(d). Fig. 8(d) shows our proposed module achieved the best performance in separating these clusters, demonstrating its effectiveness in classification.

Effect of residual connection. We investigate the effect of using residual connection. The original SRM and CBAM (without residual connection) do not perform well on the ADL-7 and GTEA_KITCHEN-6 datasets, according to Table VIII, with CBAM reducing baseline by 1.5% and SRM reducing baseline by 2.1%. On the ADL-7 and GTEA_KITCHEN-6 datasets, CBAM with residual connection performs better than the original version by 2.6% and 0.7% respectively. Similar findings are obtained by using SRM with residual connection, which enhances the original version by 1.2% and 2.2%, showing the effect of residual connection. Using residual connection can lessen the negative effect that the wrong attentions may suppress other useful information and enhance performance.

Arrangement of attentions. We also measure various module versions to investigate the channel- and temporal-wise inter-dependencies. We evaluate four different configurations of arranging the channel and temporal attention network: 1) CAN: channel-only attention; 2) TAN: temporal-only attention; 3) CTAN: sequential channel-temporal attention; 4) TCAN: sequential temporal-channel attention. As shown in Table IX, CAN outperforms TAN, showing channels benefit the network more than temporal dimensions. The reason is that channels carry more spatio-temporal information than temporal dimensions. In addition, the performance of TAN is poorer than the baseline DANN in most pairs. It means simply paying more attention to temporal information may suppress spatial information. Moreover, CTAN and TCAN outperform other networks with only one attention, showing the importance of utilizing both attentions. Finally, CTAN achieves the best accuracy among the four structures, showing that the best-arranging technique continues to improve performance.

Integration strategy of attentions. We investigate the effect of integrating attention for distinct stages. We add our channel-temporal modules into three stages of I3D: 1) early-stage: Inception blocks 3a and 3b; 2) middle-stage: Inception blocks 4c and 4d; 3) late-stage: Inception blocks 5b and 5c. As shown in Table X, 15 out of 18 tasks of stages improve the baseline, demonstrating once more the effectiveness of proposed channel-temporal attention modules. As the stage goes deeper, recognition performance on average gradually gets better because earlier layers are typically more general while later layers exhibit more significant levels of specificity. By applying attention to later layers, excited features would be more specific, which would enhance class-specific learning. Similarly, the performance difference between the early and middle stages is more significant than that between the middle and last. These findings are consistent with those in [32], showing middle stages are more class-specific than others.

Model complexity. We study the model complexity of our proposed method. Table XI shows the average accuracy and model complexity of CTAN and several UDA methods on the ADL-7 dataset. All evaluations are performed on a single

TABLE X
ABLATION STUDY OF ATTENTION INTEGRATION STRATEGIES ON ADL-7. BEST ACCURACY (%) BY INTEGRATING ATTENTION MODULES DIFFERENT STAGES ARE REPORTED.

	w/o Attention	Early	Middle	Last
D1→D2	40.6	40.7	40.8	**41.2**
D1→D3	28.1	28.6	28.9	**29.0**
D2→D1	27.1	27.9	29.0	**29.4**
D2→D3	23.8	26.6	**28.5**	28.3
D3→D1	**29.5**	24.7	26.5	26.8
D3→D2	37.5	**37.9**	37.2	37.3
Mean	31.1	31.1	31.8	**32.0**
Gain	-	0.0	+0.7	+0.9

TABLE XI
MODEL COMPLEXITY OF CTAN AND THREE COMPARING METHODS ON THE ADL-7 DATASET.

Model	DANN [21]	CDAN [50]	DAN [22]	CTAN (Ours)
FLOPs (G)	446.03	446.04	446.03	446.26
Param. (M)	12.65	13.27	12.55	13.13
Mean Acc. (%)	31.3	32.5	31.2	**33.3**

NVIDIA Titan Xp GPU. We select 16 frames from a video and resize them to 224 × 224. To evaluate speed, we utilize a batch size of 16 and ignore data loading time. Compared to DANN, CDAN and DAN, our CTAN achieves 2.0%, 0.8% and 6.1% improvement of the best average accuracy with comparable FLOPs. Due to its additional attention modules, CTAN slightly increases the parameter size by 0.48M and 0.58M when compared to DANN and DAN.

VI. CONCLUSION AND FUTURE WORK

This paper introduced two action recognition datasets with significant domain discrepancies and new challenges in UDA for first-person video action recognition, ADL-7 and GTEA_KITCHEN-6. Our new datasets enrich data and promote diversity by providing more domains and multiple settings. They also keep compatibility and scalability to better utilize existing benchmark datasets. We also proposed channel- and temporal-wise attention modules for videos to make the network focus on the important CNN channels and temporal dimensions. Finally, we proposed Channel-Temporal Attention Network, which utilizes spatio-temporal and channel attentions for videos to highlight informative video features. As a result, our network beats the image-based and video-based UDA baselines and attention networks on all three datasets.

In future work, it would be interesting to explore UDA with the effects of occlusion and camera shake in first-person videos by 1) establishing one or more new domains including occlusion and/or camera shaking and 2) studying action recognition and domain adaptation on such new data. Another future direction is to extend our proposed methods to other applications. Our attention-based method can leverage action-related information in videos, which can benefit applications including action localization [60], [61] and video object segmentation [62], [63]. Attention-based methods have improved feature embedding capabilities. Thus, applying channel-temporal attention to the feature extractor in [60]–[63] can improve feature embedding and generate better features for the proposal process. Furthermore, channel-wise attention can enhance the appearance modules [63] in generating more informative spatial features, whereas temporal-wise attention can strengthen the proposal generator [60] in capturing more action-related temporal proposal segments.

ACKNOWLEDGMENT

This work was supported in part by the China Scholarship Council (CSC) under Grant 201904910380. We thank Raivo Koot, Pawel Pukowski, and Yilin Pan for their help with data selection and annotation, as well as Yan Ge for his help with extensive comments and feedback.

REFERENCES

[1] Y. Ji, Y. Yang, F. Shen, H. T. Shen, and X. Li, "A survey of human action analysis in HRI applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2114–2128, 2019.
[2] J. Gao and C. Xu, "Learning video moment retrieval without a single annotated video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1646–1657, 2022.
[3] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "The EPIC-KITCHENS Dataset: Collection, challenges and baselines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4125–4141, 2021.
[4] A. Núñez-Marcos, G. Azkune, and I. Arganda-Carreras, "Egocentric vision-based action recognition: A survey," *Neurocomputing*, vol. 472, pp. 175–197, 2022.
[5] Y. Li, T. Nagarajan, B. Xiong, and K. Grauman, "Ego-exo: Transferring visual representations from third-person to first-person videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6943–6953.
[6] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.
[7] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6450–6459.
[8] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, 2018.
[9] Y. Ji, Y. Yang, F. Shen, H. T. Shen, and W.-S. Zheng, "Arbitrary-view human action recognition: A varying-view RGB-D action dataset," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 289–300, 2020.
[10] P. Gupta, A. Thatipelli, A. Aggarwal, S. Maheshwari, N. Trivedi, S. Das, and R. K. Sarvadevabhatla, "Quo vadis, skeleton action recognition?" *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2097–2112, 2021.
[11] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *IEEE Int. Conf. Robot. Autom.*, 2018, pp. 5628–5635.
[12] A. Nair, D. Chen, P. Agrawal, P. Isola, P. Abbeel, J. Malik, and S. Levine, "Combining self-supervised learning and imitation for vision-based rope manipulation," in *IEEE Int. Conf. Robot. Autom.*, 2017, pp. 2146–2153.
[13] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7948–7956.
[14] L. Zhang, P. Wang, W. Wei, H. Lu, C. Shen, A. van den Hengel, and Y. Zhang, "Unsupervised domain adaptation using robust class-wise matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1339–1349, 2018.
[15] X. Xu, H. He, H. Zhang, Y. Xu, and S. He, "Unsupervised domain adaptation via importance sampling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4688–4699, 2019.
[16] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100," *Int. J. Comput. Vis.*, vol. 130, no. 1, pp. 33–55, 2022.

[17] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 122–132.

[18] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2847–2854.

[19] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3281–3288.

[20] F. de la Torre, J. K. Hodgins, J. Montano, and S. Valcarcel, "Detailed human data acquisition of kitchen activities: the CMU-multimodal activity database (CMU-MMAC)," in *Proc. ACM SIGCHI Conf. Hum. Factors Comput. Syst. Workshop*, 2009.

[21] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[22] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3071–3085, 2019.

[23] J. Choi, G. Sharma, S. Schulter, and J.-B. Huang, "Shuffle and attend: Video domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 678–695.

[24] M.-H. Chen, Z. Kira, G. AlRegib, J. Yoo, R. Chen, and J. Zheng, "Temporal attentive alignment for large-scale video domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6321–6330.

[25] X. Song, S. Zhao, J. Yang, H. Yue, P. Xu, R. Hu, and H. Chai, "Spatio-temporal contrastive domain adaptation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9787–9795.

[26] D. Kim, Y.-H. Tsai, B. Zhuang, X. Yu, S. Sclaroff, K. Saenko, and M. Chandraker, "Learning cross-modal contrastive features for video domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13 618–13 627.

[27] A. Sahoo, R. Shah, R. Panda, K. Saenko, and A. Das, "Contrast and mix: Temporal contrastive video domain adaptation with background mixing," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23 386–23 400.

[28] X. Zhang, T. Wang, W. Luo, and P. Huang, "Multi-level fusion and attention-guided CNN for image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4162–4173, 2020.

[29] C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen, and X. Gao, "Task-adaptive attention for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 43–51, 2021.

[30] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3186–3195.

[31] C. Shen, G.-J. Qi, R. Jiang, Z. Jin, H. Yong, Y. Chen, and X.-S. Hua, "Sharp attention network via adaptive sampling for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3016–3027, 2018.

[32] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2020.

[33] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[34] H. Lee, H.-E. Kim, and H. Nam, "SRM: A style-based recalibration module for convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1854–1862.

[35] J. Lei, Y. Jia, B. Peng, and Q. Huang, "Channel-wise temporal attention network for video action recognition," in *IEEE Int. Conf. Multimed. Expo*, 2019, pp. 562–567.

[36] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 909–918.

[37] X. Wang, L. Zhu, Y. Wu, and Y. Yang, "Symbiotic attention for egocentric action recognition with object-centric alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020, doi: 10.1109/TPAMI.2020.3015894.

[38] S. Li, M. Yuan, J. Chen, and Z. Hu, "ADADC: Adaptive deep clustering for unsupervised domain adaptation in person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3825–3838, 2022.

[39] H. Li, N. Dong, Z. Yu, D. Tao, and G. Qi, "Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2814–2830, 2021.

[40] Q. Tian, Y. Zhu, H. Sun, S. Chen, and H. Yin, "Unsupervised domain adaptation through dynamically aligning both the feature and label spaces," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8562–8573, 2022.

[41] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2272–2281.

[42] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 01, 2016, p. 2058–2065.

[43] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 597–613.

[44] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 343–351.

[45] Y. Zhang, K. Li, K. Li, and Y. Fu, "MR image super-resolution with squeeze and excitation reasoning attention network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13 425–13 434.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[47] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10 971–10 980.

[48] S. Yin, Y. Wang, and Y.-H. Yang, "A novel image-dehazing network with a parallel attention block," *Pattern Recognit.*, vol. 102, p. 107255, 2020.

[49] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[50] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1647–1657.

[51] A. Jamal, V. P. Namboodiri, D. Deodhare, and K. Venkatesh, "Deep domain adaptation in action space." in *Proc. Br. Mach. Vis. Conf.*, vol. 2, no. 3, 2018, p. 5.

[52] B. Pan, Z. Cao, E. Adeli, and J. C. Niebles, "Adversarial cross-domain action recognition with co-attention," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 11 815–11 822.

[53] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 803–818.

[54] S. Singh, C. Arora, and C. Jawahar, "First person action recognition using deep learned descriptors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2620–2628.

[55] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 287–295.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[57] H. Lu, X. Liu, S. Zhou, R. Turner, P. Bai, R. Koot, M. Chasmai, L. Schobs, and H. Xu, "PyKale: Knowledge-aware machine learning from multiple sources in Python," in *Proc. ACM Int. Conf. Inf. Knowl. Manag.*, 2022, doi: 10.1145/3511808.3557676.

[58] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, 2009, pp. 1–4.

[59] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

[60] K. Xia, L. Wang, S. Zhou, G. Hua, and W. Tang, "Dual relation network for temporal action localization," *Pattern Recognit.*, vol. 129, p. 108725, 2022.

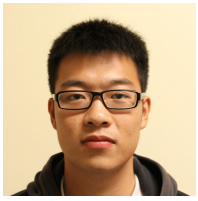
[61] K. Xia, L. Wang, S. Zhou, N. Zheng, and W. Tang, "Learning to refactor action and co-occurrence features for temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13 884–13 893.

[62] D. Zhang, J. Han, L. Yang, and D. Xu, "SPFTN: A joint learning framework for localizing and segmenting objects in weakly labeled videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 475–489, 2018.

[63] P. Huang, J. Han, N. Liu, J. Ren, and D. Zhang, "Scribble-supervised video object segmentation," *IEEE/CAA J. Automatica Sin.*, vol. 9, no. 2, pp. 339–353, 2021.

Xianyuan Liu received the B.S. degree in measuring control technology and instruments from Southeast University, Nanjing, China, in 2016. He is currently pursuing the Ph.D. degree in signal and information processing with the University of Chinese Academy of Sciences, Beijing, and in the Doctoral Program with the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China. He was a visiting researcher with the Department of Computer Science, University of Sheffield, Sheffield, UK, from 2019 to 2021.

His current research interests include action recognition, domain adaptation, and computer vision.



Shuo Zhou received the B.S. degree in information technology and education from Jiangnan University, Wuxi, China, in 2012. He received the M.Sc. and Ph.D. degrees in computer science from the University of Sheffield, Sheffield, UK, in 2017 and 2022, respectively. He is now an Academic Fellow in Machine Learning and the Deputy Head of the AI Research Engineering at the Department of Computer Science, University of Sheffield, UK.

His current research interests include interpretable machine learning and medical image analysis. He was an awardee of the Alan Turing Institute's Post-Doctoral Enrichment Awards 2022.

Tao Lei received the B.S. degree from Xihua University, Chengdu, China, in 2003. He received the M.Sc. and Ph.D. degrees in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2006 and 2013, respectively. He is now a Professor with the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China. He also serves as Vice Chairman of the Equipment and Engineering Branch of the Second Youth Innovation Promotion Association of the Chinese Academy of Sciences, Director of the Image Science and Engineering Branch of the Chinese Instrument and Control Society, and Director of Sichuan Science and Youth Federation.

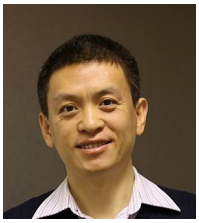
His current research interests include image processing, visual tracking, and computer vision.

Ping Jiang received the B.S. degree in computer application from Sichuan University, Chengdu, China, in 1998, and the M.Eng. degree in software engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2003. He received the Ph.D. degree in computer application technology from Sichuan University, in 2013. He is a Professor with the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China, where he serves as the Head of the Photoelectric Detection Lab.

His current research interests include image processing, computer vision, and signal processing. He has authored or co-authored more than 40 papers in peer-reviewed international journals.

Zhixiang Chen received the B.S. degree in microelectronics from Xi'an Jiaotong University, Xi'an, and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2010 and 2017, respectively. He is now a Lecturer at the Department of Computer Science, University of Sheffield, Sheffield, UK. Previously, he was a Postdoctoral Research Associate at Imperial College London and Tsinghua University.

His current research interests include computer vision and machine learning.



Haiping Lu (S'02–M'09-SM'21) received the B.Eng. and M.Eng. degrees in electrical and electronics engineering from Nanyang Technological University, Singapore, in 2001 and 2004, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Toronto, Canada, in 2008. He is now a Professor of Machine Learning at the University of Sheffield, UK, where he also serves as the Head of AI Research Engineering and Turing Academic Lead. He leads the Alan Turing Institute's interest group on meta-learning for multimodal data.

His research focuses on developing translational AI technologies for better analyzing multimodal data in healthcare and beyond. He leads the development of the PyKale library to provide more accessible machine learning from multiple data sources for interdisciplinary research, officially part of the PyTorch ecosystem. He serves as an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems and IEEE Transactions on Cognitive and Developmental Systems. He has been honoured with awards such as the Turing Network Development Award, Amazon Research Award, AAAI Outstanding PC Member Award, and IEEE CIS Outstanding PhD Dissertation Award, among others.