Aligning Correlation Information for Domain Adaptation in Action Recognition

Yuecong Xu, *Member, IEEE*, Haozhi Cao, Kezhi Mao, *Member, IEEE*, Zhenghua Chen, *Senior Member, IEEE*, Lihua Xie, *Fellow, IEEE*, and Jianfei Yang, *Member, IEEE*

***Abstract*— Domain adaptation (DA) approaches address domain shift and enable networks to be applied to different scenarios. Although various image DA approaches have been proposed in recent years, there is limited research toward video DA. This is partly due to the complexity in adapting the different modalities of features in videos, which includes the correlation features extracted as long-range dependencies of pixels across spatiotemporal dimensions. The correlation features are highly associated with action classes and proven their effectiveness in accurate video feature extraction through the supervised action recognition task. Yet correlation features of the same action would differ across domains due to domain shift. Therefore, we propose a novel adversarial correlation adaptation network (ACAN) to align action videos by aligning pixel correlations. ACAN aims to minimize the distribution of correlation information, termed as *pixel correlation discrepancy* (PCD). Additionally, video DA research is also limited by the lack of cross-domain video datasets with larger domain shifts. We, therefore, introduce a novel HMDB-ARID dataset with a larger domain shift caused by a larger statistical difference between domains. This dataset is built in an effort to leverage current datasets for dark video classification. Empirical results demonstrate the state-of-the-art performance of our proposed ACAN for both existing and the new video DA datasets.**

***Index Terms*— Action recognition, adversarial, correlation, dark videos, domain adaptation (DA).**

I. Introduction

ACTION recognition has long been studied thanks to its applications in various fields. Despite achieving promising results, most research assumes that the distribution of the test data is in line with that of the train data. Meanwhile, due to the high cost of annotating videos, it is desirable if networks trained in one domain could be directly applied to another. However, significant decrease in performances is observed when networks are applied to cross-domain scenarios. To alleviate the impact of domain shift, studies have been conducted on unsupervised domain adaptation (UDA), which aims to leverage data from the labeled source domain to boost performance on the unlabeled target domain [1], [2]. Previously, UDA has been mostly explored on image-based tasks, such as image recognition [3], [4], [5], object detection [6], [7], [8], and person re-identification [9], [10].

Comparatively, there is limited research toward applying DA methods to videos for tasks such as action recognition. This is mainly due to the fact that videos contain data with more modalities, which complicates the adaptation process. Earlier works use the same adaptation strategies as that for image DA while utilizing 3-D convolutional neural networks (3D-CNNs) instead of 2-D CNNs (2D-CNNs) for feature extraction. However, these works produce inferior results due to the fact that the simple strategy of substituting feature extractor ignores the different characteristic between spatial and temporal features. Current improvements in DA methods for video tasks focus on improving alignment along the temporal direction. Such improvements are in line with the additional temporal information provided in videos compare to images. They are achieved mainly through applying attention mechanisms to features of video segments sampled across the temporal direction [11], [12]. Alternatively, auxiliary tasks such as clip order prediction [13] are utilized to extract robust temporal representation [14].

Intuitively, the correlation features in videos in the form of long-range spatiotemporal pixel dependencies are highly associated with an action. In supervised action recognition, such correlation features have been recently exploited to aid the extraction of accurate video features. One significant example is the nonlocal neural network [15], inspired by the nonlocal mean operation for image denoising [16], [17]. The spatiotemporal features are constructed by extracting correlation features, obtained by performing self-attention [18], [19], [20]. The correlation features have brought significant increase in network performance compared to utilizing temporal features only [15], [21], [22], [23], [24]. This is thanks to the fact that

Manuscript received 1 August 2021; revised 18 March 2022 and 16 August 2022; accepted 4 October 2022. This work was supported in part by the Agency for Science, Technology and Research (A*STAR), Singapore, through its Career Development Award under Grant C210112046; and in part by the Nanyang Technological University (NTU) Presidential Postdoctoral Fellowship, "Adaptive multimodal learning for robust sensing and recognition in smart cities" project fund, Nanyang Technological University, Singapore. *(Yuecong Xu and Haozhi Cao contributed equally to this work.) (Corresponding author: Jianfei Yang.)*

Yuecong Xu is with the Institute for Infocomm Research (I^{2}R), Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: xuyu0014@e.ntu.edu.sg).

Haozhi Cao, Kezhi Mao, Lihua Xie, and Jianfei Yang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: haozhi001@e.ntu.edu.sg; ekzmao@ntu.edu.sg; yang0478@e.ntu.edu.sg).

Zhenghua Chen is with the Institute for Infocomm Research (I^{2}R) and the Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: chen0832@e.ntu.edu.sg).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3212909.

Digital Object Identifier 10.1109/TNNLS.2022.3212909

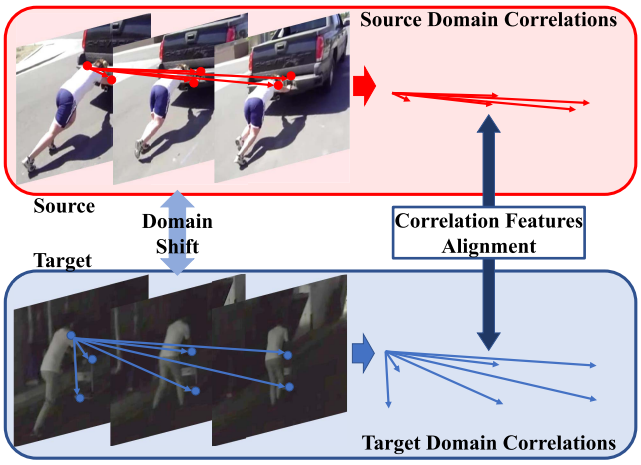

Fig. 1. Illustration of our proposed correlation features alignment. Correlation features are extracted as long-range dependencies of pixels across spatiotemporal dimensions. For the same action in the source and target domains, their corresponding correlation features are distinct due to the different postures of the actors. While correlation features are highly associated with the action, alignment of video features should include the alignment of correlation features. Here, we show two samples with the action "Push" from HMDB51 (top) and ARID (bottom).

temporal features only correlate to local pixel dependencies, while long-range dependencies are captured by correlation features. However, correlation features of the same action could be very different, as depicted in Fig. 1. The same action "Push" sampled from two different datasets results in distinct correlation information. Given the close relation between correlation features and the action, it is therefore reasonable to not only align spatial and temporal features alone but also to align correlation features. We therefore propose an adversarial correlation adaptation network (ACAN) that aligns correlation features in an adversarial manner.

For an action within a domain, its correlation features, and the embedded correlation information, would be similar, thanks to the similar appearance and postures of the actors. Yet outliers may be presented in each domain, which may impact the transferability of the network. To cope with such impact, we propose that the joint distribution of correlation information should be aligned. We believe that such a joint distribution of correlation information could be computed as the covariance of the correlation information [25], implemented as its corresponding Gram matrix [26], [27]. Therefore, aligning the correlation features of two domains is interpreted as minimizing the difference between the Gram matrices of the correlation information. While direct minimization of the Gram matrix difference could come at a price of decreasing network discriminability and high-computational cost, we propose to minimize the *pixel correlation discrepancy* (PCD).

Besides the complexity of the process of video data, the lack of research in DA methods for action recognition and other video-based tasks are also partly due to the lack of sufficient and meaningful cross-domain video datasets. Apart from current video DA datasets, we proposed a new HMDB-ARID dataset from HMDB51 [28] and a recent dark video dataset, ARID [29]. The different illumination conditions of videos in HMDB51 and ARID causes larger domain shift, making the HMDB-ARID dataset more challenging.

Our main contributions are summarized as follows.

1) We proposed a novel ACAN network for domain adaptation (DA) in action recognition by aligning correlation features in the form of long-range spatiotemporal dependencies across domains, which has not been explored by prior works.
2) We further improve the effectiveness of correlation alignment by aligning the joint distribution of correlation information of different domains through minimizing PCD.
3) We introduce a more challenging video DA dataset: the HMDB-ARID dataset. To our knowledge, this is the first video DA dataset that includes videos shot under different illumination, which possess larger domain shift than current video DA datasets.
4) We perform extensive experiments, whose results demonstrate the effectiveness of our proposed method, achieving the state-of-the-art performance across multiple current and novel video DA datasets.

The rest of this article is organized as follows: related works of unsupervised domain-adaptation in video-based tasks, such as action recognition are discussed in Section II. In Section III, we introduce our proposed ACAN with the process of minimizing PCD thoroughly. Further, in Section IV, we introduce our proposed HMDB-ARID dataset in detail. After that, we present and analyze the experimental results of our proposed ACAN on previous and our novel video DA datasets, with a thorough ablation study on the design of ACAN in Section V. Finally, we conclude the article and propose our future work in Section VI.

II. Related Works

A. Action Recognition

Action recognition has shown great progress with the use of CNNs for extracting accurate video features and representations. There exist mainly two branches of work. One of which utilizes the two-stream structure [30], [31], [32], [33], [34], [35], [36], extracting video features through CNNs from both optical flow and RGB inputs. The other path utilizes the 3D-CNN structure [37], [38], [39], [40], [41], [42], [43] to extract video features by extracting spatial and temporal features jointly with only RGB inputs. This path has made further progress by introducing separable CNN [44], [45], improving the efficiency of video feature extraction.

More recently, correlation features in the form of long-range spatiotemporal dependencies have been exploited for further improvements in action recognition. One significant example of which is inspired by the nonlocal means for image filtering task [16], termed the nonlocal block [15], and is introduced with the nonlocal neural network for capturing correlation between spatiotemporal pixels. Works as in [21], [46], and [47] also improve video feature extraction using the same idea, but utilizing different methods such as attention [21], [46] or relation modules [47]. Despite the great progress made in

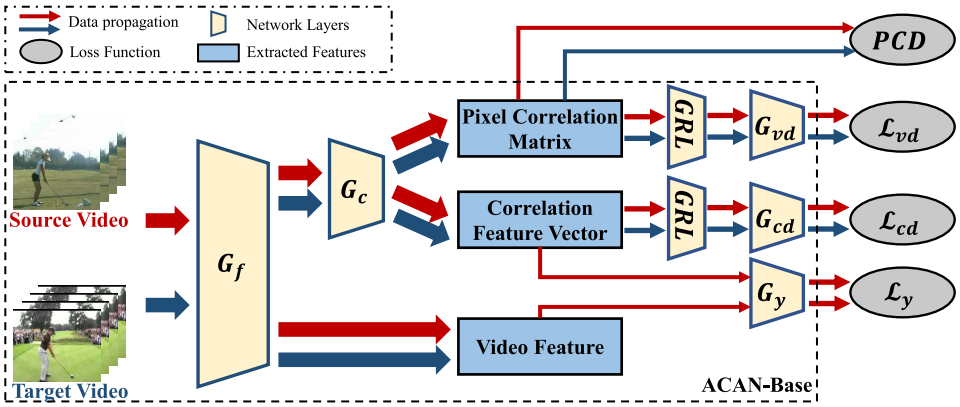

Fig. 2. Overview of the structure of ACAN. We first generate video features with a shared 3D-CNN encoder for both source and target domain videos with the same spatial and temporal dimensions. Source and target correlation feature vectors are obtained through high-level video features, extracted from a deeper layer of the encoder. An adversarial domain loss is applied to both the video features and the correlation feature vectors for aligning the video features and correlation feature vectors. Further, aligning the joint correlation information distribution requires the alignment of the Gram matrices constructed from the pixel correlation matrices (PCM). To achieve this, we further introduce the PCD. Figure best viewed in color and zoomed in.

action recognition, most models rely on the target supervised data for fine-tuning on the target dataset, and thus could not be applied to different domains or scenarios without sufficient labels or annotations. To this end, unsupervised DA helps improve the transferability of models so that they could be applied without access to target labels during training.

B. Unsupervised Domain Adaptation

In recent years, there has been a rise of research interest in DA, which aims to distill shared knowledge across domains and improve the transferability of models. In our work, we focus on UDA, when labeled target data are not available. With the success of generative adversarial network (GAN) [48], [49], researchers have proposed to construct adversarial loss [3] for DA. Various adversarial-based DA methods [3], [4], [50], [51], [52] have been proposed for a wide range of image-based tasks, such as image recognition [4], [5], [53], [54], object detection [6], [7], [55], semantic segmentation [56], [57], [58], [59], and person re-identification [9], [10], [60], [61].

Despite the progress in UDA for image-based tasks, there have been few works on UDA for video-based tasks (VUDA), such as action recognition [11], [12], [14], [62] and action segmentation [63]. Compared to direct integration of UDA approaches to videos through a simple change of feature extractor, most of these works adapt temporal features more effectively. However, temporal features only correlate to local pixel dependencies. Meanwhile, none of them have explored the alignment of correlation features that correlate to long-range pixel dependencies, which are highly associated with actions and have proven its effectiveness in supervised tasks, yet may be very different across different domains. We therefore propose to align correlation information for better video feature alignment.

III. Method

In video UDA, we are given a source domain with N_s labeled videos $\mathcal{D}_s = \{(V_s^i, y_s^i)\}_{i=1}^{N_s}$, and a target domain with N_t unlabeled videos $\mathcal{D}_t = \{V_t^j\}_{j=1}^{N_t}$. The source and target domains are characterized by two underlying probability distributions p_s and p_t, respectively. The goal of video UDA is to construct a network capable of learning transferable features and minimizing a target classification risk.

Current video DA approaches still rely on aligning only spatial and/or temporal features which correlate local pixel dependencies and fail to align correlation features which correlate long-range pixel dependencies. To cope with this challenge, we propose an adversarial correlation alignment network (ACAN) to align cross-domain correlation features in an adversarial manner. We further introduce the PCD, motivated by the theoretical results in style transfer. We begin this section by presenting the base architecture of ACAN, denoted as ACAN-base, followed by an illustration on the minimization of PCD.

A. Base Architecture

Fig. 2 presents the base architecture of our proposed ACAN, illustrated as ACAN-Base. During training, given a source and target video pair (V_s^i, V_t^j), the source and target video features f_s^i, f_t^j are obtained through a shared 3D-CNN encoder $G_f(.;\theta_f)$. To ensure that both the shared encoder is applicable on both the source and target data, the input source and target videos share the same spatial and temporal dimensions. This is achieved by sampling sequentially the same number of frames from both source and target videos, while each frame is resized and cropped directly. Meanwhile, the high-level source and target video feature f_{hs}^i, f_{ht}^j are extracted from a deeper layer of $G_f(.;\theta_f)$ (e.g., conv4 layer). The high-level video features are processed by a shared correlation extraction module G_c where the correlation features of the input videos are extracted. The results are the source and target pixel correlation matrices $\mathbf{M}_s^i, \mathbf{M}_t^j$ as well as the source and target correlation feature vectors f_{cs}^i, f_{ct}^j. $G_c(.;\theta_c)$ is built based on the nonlocal operation [15], which extracts the correlation features as long-range dependencies between spatiotemporal pixels. To preserve both local and long-range spatiotemporal

pixel dependencies, the source correlation feature vector and video feature f_{cs}^i, f_s^i are concatenated to form the overall feature representation of source video V_s^i, which would be input to a classifier G_y for action predictions. The action class prediction loss $\mathcal{L}_y$ is computed with respect to the predictions from G_y, formulated as

$$\mathcal{L}_y = \frac{1}{N_s}\sum_{i=1}^{N_s} L_y\big(G_y\big(f_{\text{cs}}^i \oplus f_s^i\big), y_i\big) \tag{1}$$

where L_y is the cross entropy loss function and $\oplus$ denotes the concatenation operation.

To accommodate the domain shift between source and target domains, adversarial-based UDA approaches are proved to perform well on image data [3], [50], [51], [52] and language data [64]. We also leverage such technique for VUDA, which aims to align the global distributions with additional domain discriminators that are trained with the feature generators in a min-max fashion. Domain discriminators are designed to discriminate the video features while the feature generators are trained to deceive the domain discriminators. Here the feature generators are referred to as the combination of G_f and G_c. We adopted separate domain discriminators for the source/target video features $f_*^\star$ $(* \in (s,t), \star \in (i,j))$ and the source/target correlation features $f_{c*}^\star$. The two domain discriminators are denoted as the video domain discriminator $G_{vd}(.;\theta_{vd})$ and the correlation domain discriminator $G_{cd}(.;\theta_{cd})$. During the adversarial training process, the parameters θ_{vd} and θ_{cd} are learned by minimizing the video domain loss $\mathcal{L}_{vd}$ and the correlation domain loss $\mathcal{L}_{cd}$, respectively, which are formulated as

$$\mathcal{L}_{vd} = \frac{1}{N_s}\sum_{i=1}^{N_s} L_b\big(G_{vd}\big(f_s^i\big), d_i\big) + \frac{1}{N_t}\sum_{j=1}^{N_t} L_b\Big(G_{vd}\Big(f_t^j\Big), d_j\Big) \tag{2}$$

$$\mathcal{L}_{cd} = \frac{1}{N_s}\sum_{i=1}^{N_s} L_b\big(G_{cd}\big(f_{\text{cs}}^i\big), d_i\big) + \frac{1}{N_t}\sum_{j=1}^{N_t} L_b\Big(G_{vd}\Big(f_{ct}^j\Big), d_j\Big) \tag{3}$$

where L_b is the binary cross-entropy loss of the domain discriminators, while d_i and d_j are the domain label for the source and target domains, respectively. Meanwhile, the parameters of the feature extractors θ_f and θ_c are learned to maximize the domain losses simultaneously. To achieve uniform minimization of the action class prediction loss and the maximization of the domain losses, a gradient reverse layer (GRL) [3] is inserted before each domain discriminator as in Fig. 2.

The overall loss function to be optimized can therefore be formulated as

$$\mathcal{L} = \mathcal{L}_y - (\lambda_v \mathcal{L}_{vd} + \lambda_r \mathcal{L}_{cd}) \tag{4}$$

where λ_v and λ_r are the trade-off weights for the video domain loss and correlation domain loss, respectively.

B. Minimizing Pixel Correlation Discrepancy

In the ACAN-Base network, the same DA approach is applied to both video and correlation features. However, it remains a question *whether such an approach is the most effective way for aligning correlation features across different domains?* Aligning correlation features can be further achieved through aligning the joint distribution of correlation information. The joint distribution could be computed as the covariance of correlation information, implemented as its corresponding Gram matrix. The key to the above question therefore lies in the expression of the correlation information. As illustrated in Fig. 2, correlation features are extracted from G_c, whose structure is shown in Fig. 3. For the ith input video, we define the pixel correlation matrix (PCM) $\mathbf{M}_*^i$ as

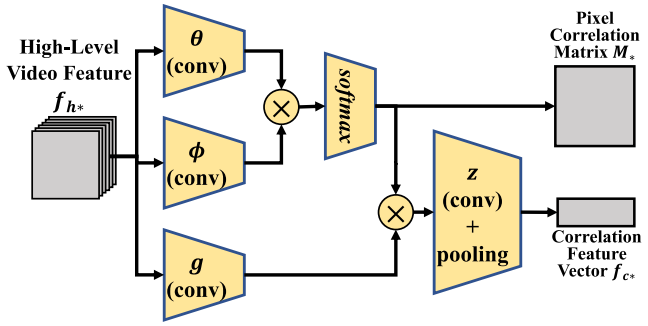

Fig. 3. Structure of the correlation extraction module G_c. G_c extract correlation features (pixel correlation matrix $\mathbf{M}_*$ and correlation feature vector f_{c*}) through the high-level video feature f_{h*}. It is built upon the nonlocal operation. $\mathbf{M}_*$ is obtained through multiplication of f_{h*} projected on latent spaces, and represents the correlation between each spatiotemporal pixel feature. f_{c*} is further obtained by multiplying the $\mathbf{M}_*$ f_{h*} projected on the latent space, followed by pooling operation over spatiotemporal dimensions. Projection functions are implemented with convolution layers of $1 \times 1 \times 1$ kernel.

$$\mathbf{M}_*^i = \varphi\Big(\theta\big(f_{h*}^i\big)^T \phi\big(f_{h*}^i\big)\Big) \tag{5}$$

where φ is the softmax operation. Both $\theta(\cdot)$ and $\phi(\cdot)$ are linear functions projecting the high-level video features to latent spaces. In practice, they are implemented as convolution layers with a kernel size of $1 \times 1 \times 1$. The value $\mathbf{M}_{*,pq}^i$ at the (p,q) position of PCM represents the correlation between the video feature at spatiotemporal point p, $f_{h*,p}^i$, and the video feature at spatiotemporal point q, $f_{h*,q}^i$. We argue that PCM could be viewed as the correlation information of the video. Therefore, the joint correlation information distribution is constructed as the Gram matrix of the PCM, denoted as $\mathcal{G}^i \in \mathbb{R}^{N_M \times N_M}$, where N_M is the number of spatiotemporal points in the feature map $\theta(f_{h*}^i)$. $\mathcal{G}^i$ is computed by

$$\mathcal{G}^i = {\mathbf{M}_*^i}^T \mathbf{M}_*^i. \tag{6}$$

The alignment of correlation features thus requires the minimization of the distance between the Gram matrices $\mathcal{G}$, termed as the video covariance loss $\mathcal{L}_{vs}$, formulated by

$$\mathcal{L}_{vs} = ||\mathbf{E}(\mathcal{G}_s) - \mathbf{E}(\mathcal{G}_t)||^2 \tag{7}$$

where the subscripts s and t denotes the Gram matrices for source and target videos respectively. However, such computation is inefficient, requiring a cost of $O({N_M}^2)$. Furthermore, improving network transferability through minimizing $\mathcal{L}_{vs}$ comes at the price of decreasing network discriminability.

To minimize $\mathcal{L}_{vs}$ more efficiently while causing less impact on the network's discriminability, we simplify according to the theory in [65].

Theorem 1: Given the Gram matrices $\mathcal{G}_s, \mathcal{G}_y$ constructed from source and target features $\mathbf{M}_s$, $\mathbf{M}_t$, the minimization of distance between the Gram matrices $\mathcal{L}_{vs}$ can be seen as a distribution alignment process from $\mathbf{M}_t$ to $\mathbf{M}_s$.

As proven in [65], the above theorem indicates that minimizing $\mathcal{L}_{vs}$ could be reformulated as minimizing the distribution discrepancy of $\mathbf{M}_t$ and $\mathbf{M}_s$. Set the underlying distributions of $\mathbf{M}_s$ be p_{Ms} and that of $\mathbf{M}_t$ be p_{Mt}. Here we propose the PCD, denote as $d_M(p_{Ms}, p_{Mt})$. Computing and minimizing this discrepancy are achieved by representing the distributions p_{Ms} and p_{Mt} as elements on the reproducing kernel Hilbert space (RKHS). As such, the distribution discrepancy could be defined as distance of distribution embedded elements on the RKHS.

Further, to align the distributions of p_{Ms} and p_{Mt} in a more fine-grained manner, it is important to align the distributions taking the relations between relevant classes into consideration. That is to align p_{Ms} and p_{Mt} within the same action classes in source and target domains, instead of aligning it only in by the global distributions. The overall PCD is therefore formulated as

$$d_M(p_{Ms}, p_{Mt}) \triangleq \mathbf{E}_{cl} \left|\left| \mathbf{E}_{p_{Ms}(cl)}[\zeta(\mathbf{M}_s)] - \mathbf{E}_{p_{Mt}(cl)}[\zeta(\mathbf{M}_t)] \right|\right|^2_{\mathcal{H}} \tag{8}$$

where $\mathbf{E}_{p_{M*}(cl)}$ is the mean embedding of distribution p_{M*} for action class cl on the RKHS $\mathcal{H}$. The feature map ζ is closely related to the RKHS characteristic kernel k by $k(\mathbf{M}_s, \mathbf{M}_t) = \langle \zeta(\mathbf{M}_s), \zeta(\mathbf{M}_t) \rangle$. The use of mean embedding for each class enables our PCD to align distributions of correlation information within each action class instead of only focusing on the global correlation information distribution. In practice, we may further assume that each video belongs to a certain action class with a class-related weight w_{cl}. We therefore could estimate PCD in (8) as

$$d_M(p_{Ms}, p_{Mt}) = \frac{1}{Cl} \sum_{cl=1}^{Cl} \left|\left| \sum_{i=1}^{N_s} w^i_{\text{scl}} \zeta\left(\mathbf{M}^i_s\right) - \sum_{j=1}^{N_t} w^j_{\text{tcl}} \zeta\left(\mathbf{M}^j_t\right) \right|\right|^2_{\mathcal{H}} \tag{9}$$

where Cl is the number of action classes. When computing the weight of a source video for a certain action class, given that the labels are provided, the weight w^i_{scl} is computed by

$$w^i_{\text{scl}} = \frac{y^i_s}{\sum_{k=1}^{N_s} y^k_s} \tag{10}$$

whereas for the target videos, since the labels are not available, we cannot compute the weight w^j_{tcl} directly. Instead, we utilize the output from the action classifier G_y which characterizes the probability of assigning a given video to an action class. This is denoted as the pseudo-label for a target video and is computed by

$$y^j_t = G_y\left(f^j_t \oplus f^j_{ct}\right). \tag{11}$$

The resulting pseudo-labels of the target videos could be used as in (10) for computing the weight of a target video for an action class. Finally, since the feature map ζ cannot be computed directly in most cases, we expand (9) while utilizing the characteristic kernel k. The PCD could therefore be reformulated as

$$\begin{aligned} d_M(p_{Ms}, p_{Mt}) = \frac{1}{C} \sum_{c=1}^{C} \Bigg(& \sum_{i=1}^{N_s} \sum_{i'=1}^{N_s} w^i_{sc} w^{i'}_{sc} k\left(\mathbf{M}^i_s, \mathbf{M}^{i'}_s\right) \\ & + \sum_{j=1}^{N_t} \sum_{j'=1}^{N_t} w^i_{tc} w^{j'}_{tc} k\left(\mathbf{M}^j_t, \mathbf{M}^{j'}_t\right) \\ & - 2 \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} w^i_{sc} w^j_{tc} k\left(\mathbf{M}^i_s, \mathbf{M}^j_t\right) \Bigg) \end{aligned} \tag{12}$$

where the kernel k would typically be of Gaussian form, hence $k(\mathbf{M}^i_s, \mathbf{M}^j_t) = -\exp((\| \mathbf{M}^i_s - \mathbf{M}^j_t \|^2 / 2\sigma^2))$. The overall optimization objective is thus formulated as

$$\mathcal{L} = \mathcal{L}_y - (\lambda_v \mathcal{L}_{vd} + \lambda_r \mathcal{L}_{cd}) + \lambda_d d_M \tag{13}$$

where λ_d is the trade-off weight for the PCD. Minimizing our proposed PCD is superior in effective alignment of cross-domain correlation features thanks to its relatively solid theoretical motivation. While aligning video features could also be achieved by minimizing feature discrepancies directly through methods such as MMD [66], CORAL [67], these discrepancies cannot measure the correlation difference between the source and the target domains as in PCD which matters to video DA. Therefore, applying MMD or CORAL for video feature alignment produces inferior performances than our proposed approach as illustrated in Section V. For inference, we follow the steps as indicated in Algorithm 1 and obtain the action recognition predictions for the unlabeled target domain videos. Note that the video indices i, j are omitted for simplicity.

Algorithm 1 Inference ACAN for Target Domain Videos

Input: Target data $V_t \in \mathcal{D}_t$, trained feature generators G_f, G_c, and trained classifier G_f
Output: Predicted action class y_t
Obtain target video feature $f_t = G_f(V_t)$
Obtain target correlation feature vector $f_{ct} = G_c(V_t)$
Concatenate f_t and f_{ct} to form the overall feature representation of V_t by $f_t \oplus f_{ct}$
$y_t = G_y(f_t \oplus f_{ct})$

IV. HMDB-ARID Dataset

There are very limited cross-domain benchmark datasets for video DA tasks, therefore hindering the research for video DA. Previous cross-domain datasets introduced for video DA [62], [68], [69] are of very small scale, with not more than six classes, and typically less than 1000 videos. The lack of classes and data over these cross-domain datasets introduces limited

TABLE I
COMPARISON OF RGB MEAN AND STANDARD DEVIATION (STD) OVER COMMON ACTION RECOGNITION DATASETS AND THE ARID DATASET

Dataset	RGB Mean	RGB Std
HMDB51	[0.424,0.364,0.319]	[0.268,0.255,0.260]
UCF101	[0.409,0.397,0.358]	[0.266,0.265,0.270]
Kinetics	[0.432,0.395,0.377]	[0.228,0.222,0.217]
ARID	[0.079,0.074,0.073]	[0.101,0.098,0.090]

domain discrepancy, and therefore the performances of DA approaches are saturated. More recently, larger cross-domain video datasets, such as UCF-HMDB$_{\text{full}}$ have been introduced with larger domain discrepancies.

Though larger cross-domain datasets have been introduced, both domains included in these datasets are still based on current well-established action recognition datasets. These action recognition datasets may include different classes with different videos, yet most of them are collected on public video platforms. This would lead to similar video statistics among these datasets, as compared in Table I. Similar video statistics suggest high probability of similar scenarios exist among current action recognition datasets, thus the domain shift between these datasets may not be significant. Consequently, the difficulty of adapting the same model across the different domains with similar video statistics or similar scenarios may be trivial. Video DA approaches that perform well in these cross-domain video datasets may not be well applicable in real-world applications where the gap between domains may be much larger than current cross-domain datasets. We argue that video DA approaches would be more useful for bridging with video domains with large distribution shifts, such as dark videos (adverse illumination) or hazy videos (adverse contrast).

To explore how to leverage current datasets to boost performance on videos shot in adverse environments, we propose a novel cross-domain dataset. It incorporates both the current action recognition dataset and a more recent dark dataset, ARID [29], whose videos are shot under adverse illumination conditions. Compared with current action recognition datasets, videos in ARID are characterized by low brightness and low contrast. Statistically, videos in ARID possess much lower RGB mean value and standard deviation (std), as presented in Table I. The larger statistical differences between ARID and current action recognition datasets, such as HMDB51 [28], would strongly suggest a larger domain shift between the different datasets.

The ARID dataset includes a total of 11 human action classes. These includes *drink, jump, pick, pour, push, run, sit, stand, turn, walk, and wave*. When proposing the cross-domain HMDB-ARID dataset, we include all 11 action classes in ARID and HMDB51. For both datasets, we follow the official split method to separate the train and validation sets. The HMDB-ARID dataset thus includes 770 training videos and 330 validation videos from HMDB51, and 2288 training videos and 823 validation videos from ARID. Fig. 4 shows the comparison of sampled frames from HMDB-ARID dataset. Compared to previous video DA datasets, besides containing larger domain shift, our dataset also contains a larger number of total videos for both training and validation, as illustrated in Table II.

V. EXPERIMENTS

In this section, we evaluate our proposed ACAN performing cross-domain action recognition on two video DA datasets: UCF-HMDB$_{\text{full}}$ and our new HMDB-ARID. We present the state-of-the-art results on both datasets. We also present detailed ablation studies and qualitative analysis of our proposed ACAN to verify our design.

A. Experimental Settings and Details

We perform action recognition tasks on both the UCF-HMDB$_{\text{full}}$ dataset and our new HMDB-ARID dataset. The UCF-HMDB$_{\text{full}}$ dataset [11] is introduced as an expansion of the original UCF-HMDB$_{\text{small}}$ dataset [68], with more classes and larger domain discrepancy. The UCF-HMDB$_{\text{full}}$ contains a total of 3209 videos with 12 action classes, all from the original UCF101 [70] and HMDB51 [28] datasets. It includes two settings: UCF→HMDB and HMDB→UCF, where the direction of the arrow symbol is set from the source domain toward the target domain. We use the same splits as provided in the original paper [11]. The novel HMDB-ARID dataset is as introduced in Section IV, and also consist of two settings: HMDB→ARID and ARID→HMDB. For all four settings, we report the top-1 accuracy on the target dataset, averaged on 5 runs with identical settings for each approach.

Our experiments are implemented using the PyTorch [71] library. To obtain video features, we instantiate two 3D-CNNs, I3D [39] and MFNet [72], as G_f for both source and target domain videos. Both I3D and MFNet are utilized thanks to its performance on current action recognition benchmarks (namely UCF101 [70], HMDB51 [28], and Kinetics [73]). MFNet is also utilized due to its lightweight structure, which enables it to achieve comparable results to that of I3D while requiring a fraction of the parameters and computation power needed.

The source and target feature extractors share parameters. Following the implementation in [72] and [39], the inputs for both I3D and MFNet as the source or target feature extractors are frame sequences of 16 frames sampled sequentially from the original input source or target video. Each frame is of the same resolution obtained by resizing such that the shorter edge is of 240 pixels and cropping the original frame to resolution 224×224. The correlation extraction module takes the high-level video feature from the output of layer4 in I3D and the output of conv4 layer in MFNet as inputs, which are feature maps of size 14×14. The stochastic gradient descent algorithm [74] is used for optimization, with the weight decay set to 0.0001 and the momentum to 0.9 for both I3D and MFNet. During training, the batch size is set to 8 samples per GPU. Empirically, our initial learning rate is set to 0.005 and is divided by 10 after 20 and 35 epochs. λ_v is set to 0.5 while

TABLE II
COMPARISON OF CURRENT AND OUR NOVEL VIDEO DA DATASETS

Statistics	UCF-HMDB$_{small}$	UCF-Olympic	UCF-HMDB$_{full}$	HMDB-ARID
Video Length (seconds)	1-21	1-39	1-33	1-30
Video Classes #	5	6	12	11
Training Video #	UCF:482/HMDB:350	UCF:601/Olympic:250	UCF:1438/HMDB:840	HMDB:770/ARID:2288
Validation Video #	UCF:189/HMDB:150	UCF:240/Olympic:54	UCF:571/HMDB:360	HMDB:330/ARID:823

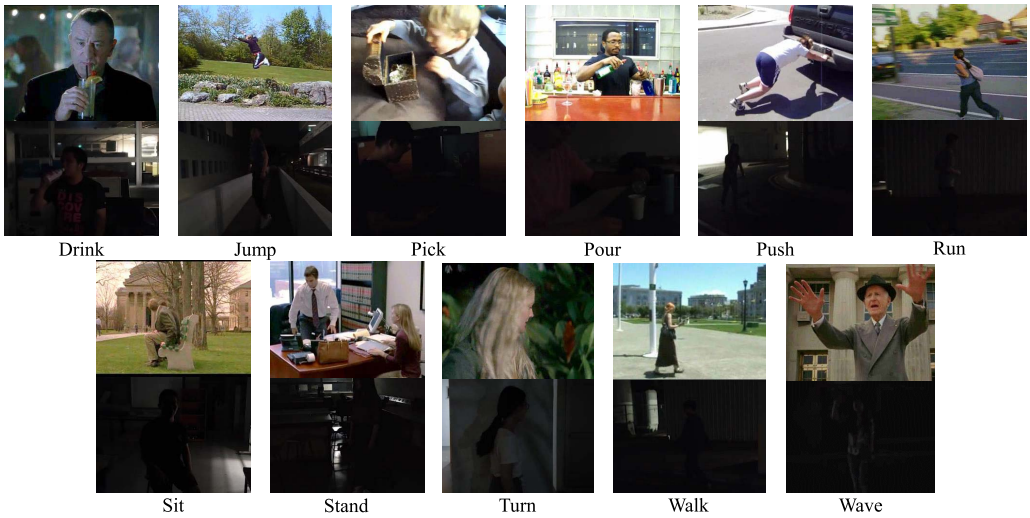

Fig. 4. Sampled frames for each action class from the videos in HMDB-ARID. Note that the sampled frames from HMDB51 are shown in the upper row, whereas the sampled frame from ARID are shown in the lower row. Best viewed zoomed in.

λ_r and λ_d are both set to 1.0 through empirical results. All experiments are conducted using two NVIDIA GP100 GPUs.

B. Overall Results

There are limited studies focusing on applying DA approaches to the action recognition task. Here we first compare previous methods utilizing the UCF-HMDB$_{\text{full}}$ benchmark. These include TA^{3}N [11], TCoN [12], and SAVA [14]. Due to the different encoders used for the different methods, we report both: 1) the "Source only" results, where the network is trained with supervised source data only and validated on the target data, and is the lower bound performance for the adaptation process; and 2) the "Target only" results, where the network is directly trained and validated with supervised target data and is the upper bound performance for the adaptation process. The comparison of performance should focus on the networks' improvement with respect to the performance with the "Source only" setting. The comparison should also focus on the distance between the network's performance and the performance with the "Target only" setting. For the performance of TA^{3}N, we follow the works in [14] and obtain the results by running the publicly available code. Table III shows the comparison of performances between our proposed ACAN and the methods as mentioned on UCF-HMDB$_{\text{full}}$.

The performance results in Table III shows that our proposed ACAN achieves the best result under the HMDB→UCF setting and very competitive performance under the UCF→HMDB setting when using either MFNet or I3D as the encoder. More specifically, our ACAN with the MFNet encoder achieves 85.8% top-1 accuracy for UCF→HMDB setting, indicating that the improvement brought by ACAN toward the lower bound of the UCF→HMDB setting is 7.2%. This is significantly higher than that brought by SAVA (1.9%) and TA^{3}N (2.2%). The large improvement brought by ACAN enables our network to perform better on UCF→HMDB setting despite the lower bound of MFNet is lower than that of I3D [39]. Under this setting, our ACAN is also closer to the upper bound of the encoder, with a gap of 10.2%. Comparatively, the gap to the upper bound performance is 15.5% for TA^{3}N and 12.8% for SAVA. Similarly, our ACAN with I3D encoder also performs better than both TA^{3}N and SAVA. Comparatively, ACAN with

TABLE III
RESULTS ON THE TWO SETTINGS FOR UCF-HMDB$_{\text{FULL}}$

Method	Encoder	UCF→ HMDB	HMDB→ UCF
Source Only	TRN-Res101	73.1%	73.9%
TA^{3}N	TRN-Res101	75.3%	79.3%
TCoN	TRN-Res101	**87.2%**	89.1%
Target Only	TRN-Res101	90.8%	95.6%
Source Only	I3D	80.3%	88.8%
SAVA	I3D	82.2%	91.2%
ACAN(Ours)	I3D	85.4%	**93.8%**
Target Only	I3D	95.0%	96.8%
Source Only	MFNet	78.6%	88.4%
ACAN(Ours)	MFNet	**85.8%**	93.2%
Target Only	MFNet	96.0%	97.1%

TABLE IV
RESULTS ON THE TWO SETTINGS FOR HMDB-ARID

Method	Encoder	HMDB→ ARID	ARID→ HMDB
Source Only	TRN-Res101	17.8%	15.7%
TA^{3}N	TRN-Res101	22.4%	19.8%
Target Only	TRN-Res101	52.8%	50.9%
Source Only	MFNet	48.3%	37.9%
DANN	MFNet	50.7%	40.6%
MK-MMD	MFNet	50.2%	40.1%
MCD	MFNet	47.6%	36.8%
CORAL	MFNet	51.3%	41.7%
ACAN(Ours)	MFNet	**58.0%**	**46.4%**
Target Only	MFNet	76.1%	67.6%

I3D encoder outperforms SAVA by 3.2% while sharing the I3D as the common video feature encoder with SAVA. This further demonstrates the superiority of ACAN over current video DA methods.

The superiority of ACAN further strengthens under the HMDB→UCF setting. Under these settings when utilizing MFNet as the video feature encoder, our proposed ACAN gains a 4.8% improvement toward the lower bound performance, which is greater than that brought by SAVA (2.4%). When utilizing I3D as the video feature encoder as in SAVA, our proposed ACAN gains an exceptional 5.0% improvement toward the lower bound performance. The larger increase built upon the strong I3D encoder enables our ACAN to achieve the best result under this setting with 93.8% top-1 accuracy. The gap toward the upper bound performance is also the smallest for ACAN using the I3D encoder, with 3.0% compared to 16.3% for TA^{3}N, 6.5% for TCoN, and 5.6% for SAVA.

We further compare performances of several methods on our novel HMDB-ARID dataset, with both HMDB→ARID and ARID→HMDB settings, as shown in Table IV. Note that both settings are more challenging, given that the gap between the lower bound performance (trained with supervised source data) and the upper bound performance (trained with supervised target data) is larger compared to the settings for UCF-HMDB$_{\text{full}}$. In addition to comparing with the TA^{3}N with TRN-Res101 [47] encoder, we also compare with performances with other typical DA approaches, e.g., DANN [3], MK-MMD [66], MCD [75], and CORAL [67], all with MFNet as the encoder.

The performance results in Table IV indicate that our proposed ACAN achieves the best results in either setting related to our novel HMDB-ARID dataset. Our ACAN achieves a top-1 accuracy of 58.0% for the HMDB→ARID setting and 46.4% for the ARID→HMDB setting. Our ACAN also brings the most significant improvement with respect to the lower bound performance, with 9.8% and 8.5% for the two settings respectively. Comparatively, TA^{3}N which does not utilize correlation alignment only brings 4.6% and 4.1% increase with respect to the lower bound performance. This shows that previous methods that fail to align correlations would not be able to effectively handle the larger domain shift caused by a more significant difference in video statistics. Note that the gap to the upper bound performance obtained by training with supervised target data is still relatively large, suggesting further improvements could be made on this novel HMDB-ARID dataset.

TABLE V
ABLATION EXPERIMENTS ON INCLUDING CORRELATION FEATURES, ON UCF→HMDB AND HMDB→ARID SETTINGS

Method	UCF→HMDB	HMDB→ARID
Source only w/o. correlation	76.1%	48.1%
Source only w. correlation	78.6%	48.3%
Adv. DA w/o. correlation	80.2%	50.7%
Adv. DA w. correlation	84.2%	52.6%

C. Ablation Studies

We further justify our proposed design of ACAN through thorough ablation studies. Specifically, we first examine the performance of our ACAN in four scenarios and justify the need for introducing correlation features in the extraction process, the use of two separate domain losses, and the introduction of PCD. We also introduce an alternative form of the joint correlation information distribution difference minimization to compare and justify our current design of PCD. All ablation studies are conducted under the UCF→HMDB and HMDB→ARID settings, with the batch size and other training parameters as mentioned in Section V-A. The MFNet [72] is instantiated as the encoder for all ablation studies.

1) Necessity of Correlation Feature Alignment: We first justify the need for correlation features for alignment, which is achieved by: 1) comparing the "Source only" results with and without the introduction of correlation features and 2) comparing the use of adversarial DA approaches with and without correlation features. Results in Table V justifies the use of correlation features, where such strategy consistently improves the performance of the network under both "Source only" training and when DANN method is used for DA. It could also be observed that the use of correlation features brings more improvement when the DANN method is applied. Such observation is consistent with our argument of improving video feature alignment using correlation alignment.

2) Effectiveness of Domain Loss $\mathcal{L}_d$: We then justify our design of the domain loss $\mathcal{L}_d$, which is the weighted sum of $\mathcal{L}_{vd}$ and $\mathcal{L}_{cd}$. We compare with the variants of ACAN where either $\mathcal{L}_{vd}$ or $\mathcal{L}_{cd}$ alone is used as the domain loss, denoted as ACAN$-\mathcal{L}_{cd}$ and ACAN$-\mathcal{L}_{vd}$. We also tested on the case where the domain loss is not applied (hence aligning correlation features by minimizing PCD alone), denoted as MFNet+PCD. As indicated in Table VI, both losses contribute to the effective alignment of video features. The removal of either loss brings a decrease in network performance for both dataset settings. Further decrease is observed when no domain loss is applied. Meanwhile, the domain discriminators corresponding to either domain loss bring only a negligible growth in computation cost. Hence it is worthwhile to include two separate domain discriminators, with two domain losses for the overall domain loss $\mathcal{L}_d$.

3) Effectiveness of PCD: PCD is introduced for improving the effectiveness of correlation alignment by matching the joint correlation information distribution of video domains. We examine the effect of PCD through comparing with the ACAN variant without PCD, which is ACAN-Base as shown

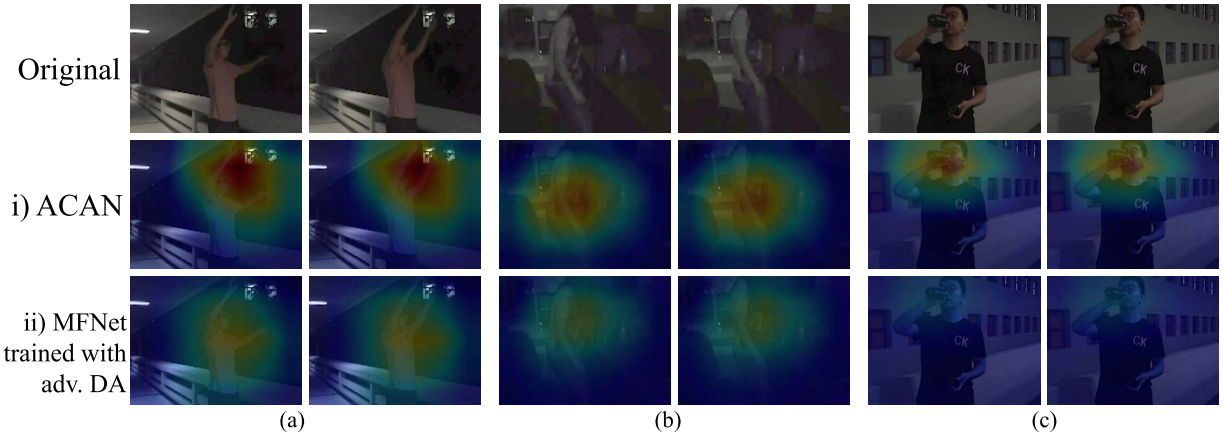

Fig. 5. Class activation maps (CAMs) on ARID, utilizing i) ACAN and ii) MFNet trained with adversarial DA approach. CAMs are obtained from three actions: (a) "Wave," (b) "Stand," and (c) "Drink." We also show the original frames at the top row from which the CAMs are computed. Original frames are tuned brighter for visualization.

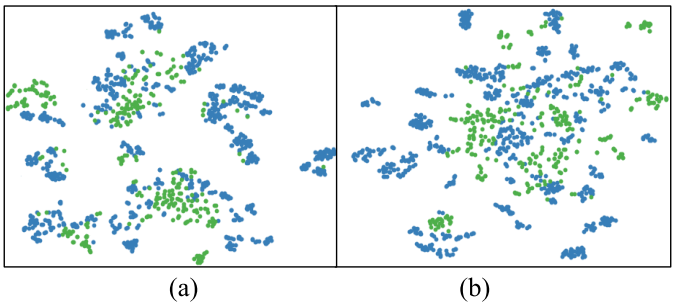

Fig. 6. Comparison of t-SNE visualization of video features of both source and target domains under HMDB→ARID. Video features are obtained from (a) ACAN and (b) MFNet trained with the adversarial DA approach. Green dots represent the data from the source domain while the blue dots represent the data from the target domain.

TABLE VI

ABLATION EXPERIMENTS ON THE DOMAIN LOSS $\mathcal{L}_d$ ON UCF→HMDB AND HMDB→ARID SETTINGS

Method	UCF→HMDB	HMDB→ARID
ACAN	85.8%	58.0%
ACAN−$\mathcal{L}_{cd}$	84.9%	56.9%
ACAN−$\mathcal{L}_{vd}$	84.5%	56.7%
MFNet + PCD	83.8%	56.1%

in Fig. 2. The results in Table VII demonstrates the effectiveness of PCD, whose absence results in a noticeable 1.6% accuracy decrease for UCF→HMDB setting, and a significant 5.4% accuracy decrease for HMDB→ARID setting. Though the introduced PCD improves the effectiveness of correlation alignment greatly, minimizing PCD involves kernel estimation which increase computation cost. Inspired by the hypothesis presented in [76], minimizing the joint distribution difference, and hence the distance between distributions p_{Ms} and p_{Mt}, could also be achieved through matching the norm of p_{Ms} and p_{Mt} toward a shared restrictive scalar R. The computation of distribution distance with this method is simpler given that no kernel estimation is required. In this case, the equation for the overall loss (13) is reformulated as

$$\begin{aligned}\mathcal{L} = \mathcal{L}_y - (\lambda_v \mathcal{L}_{vd} + \lambda_r \mathcal{L}_{cd}) \\ + \lambda_{\text{dist}} \left(L_{\text{dist}} \left(\frac{1}{N_s} \sum_{i=1}^{N_s} n\left(\mathbf{M}_s^i\right) R \right) \right. \\ \left. + L_{\text{dist}} \left(\frac{1}{N_t} \sum_{j=1}^{N_t} n\left(\mathbf{M}_t^j\right), R \right) \right). \end{aligned} \tag{14}$$

Here L_{dist} is the distance loss between the norm of PCMs and the restrictive scalar R, and is implemented as L_2-distance, while $n(\cdot)$ denotes the norm function. R is set to 25 during the experiments. We denote the variant of ACAN with loss function in (14) as ACAN (l2-norm) and compare with the original ACAN. The results in Table VII shows that the variant formulated by (14) could still bring noticeable improvement compared to the ACAN-Base where the distributions of p_{Ms} and p_{Mt} are not aligned. However, compared to PCD, the improvement is relatively minor, which further justifies the effectiveness of the current design of PCD.

TABLE VII

ABLATION ON PCD AND ALTERNATIVE WAY OF MINIMIZING JOINT CORRELATION INFORMATION DISTRIBUTION DIFFERENCE, ON UCF→HMDB AND HMDB→ARID SETTINGS

Method	UCF→HMDB	HMDB→ARID
ACAN	85.8%	58.0%
ACAN-Base	84.2%	52.6%
ACAN (l2-norm)	85.0%	54.2%

D. Qualitative Analysis

To better understand the effect of ACAN, we perform qualitative analysis on trained networks. We first present the class activation maps (CAMs) [77] of the target ARID videos with ACAN and with MFNet (encoder) trained with adversarial DA approach in Fig. 5. The dark videos in ARID make it difficult for accurate video features to be

extracted. Therefore, if correlation alignment is not utilized, the network may fail to focus on the actual action in the target domain. Instead, it may only briefly focus on the whole actor [see Fig. 5(ii-a)], or on unrelated background [see Fig. 5(ii-b)]. With the involvement of correlation features and its alignment, ACAN is able to focus on the waving hand for the "Wave" action, or the person standing for the "Stand" action, thus showing much stronger performance on the HMDB→ARID setting. Further, we visualize the distribution of the source and target domains under the HMDB→ARID setting with t-SNE [78], as shown in Fig. 6. It could be observed that our proposed ACAN can group both the data from the source domain (green dots) and data from the target domain (blue dots) into denser clusters. Our ACAN could also match the target domain data with source domain data more accurately.

VI. Conclusion and Future Work

In this work, we propose a novel DA method for action recognition across different domains. The new ACAN aligns correlation features in an adversarial manner while minimizing joint correlation information distribution differences by minimizing PCD. We further introduce a novel video DA dataset, HMDB-ARID, with a larger domain shift, and is the first video DA dataset that includes videos shot in adverse conditions. Our method obtains the state-of-the-art results on both the UCF-HMDB$_{\text{full}}$ and HMDB-ARID datasets. We further justify our design via thorough ablation studies and validate the effectiveness of ACAN with qualitative results.

Although state-of-the-art performances have been achieved by the proposed ACAN, we observe that the gap to the upper bound performance obtained by training with supervised target data is still relatively large as depicted in Table IV, suggesting further improvements could be made on the novel HMDB-ARID dataset. Additionally, cross-domain video datasets that involve a variety of large domain shift scenarios, such as blurry or hazy videos may be explored. Video DA approaches that cope with these different large domain shift scenarios would also be further investigated.

References

[1] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015.

[2] S. Zhao et al., "A review of single-source deep unsupervised visual domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 473–493, Feb. 2022.

[3] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[4] X. Ma, T. Zhang, and C. Xu, "Deep multi-modality adversarial networks for unsupervised domain adaptation," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2419–2431, Sep. 2019.

[5] Q. Kang, S. Yao, M. Zhou, K. Zhang, and A. Abusorrah, "Effective visual domain adaptation via generative adversarial distribution matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3919–3929, Sep. 2021.

[6] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3339–3348.

[7] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11457–11466.

[8] S. Song et al., "Deep domain adaptation based multi-spectral salient object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 128–140, 2022.

[9] F. Yang et al., "Part-aware progressive unsupervised domain adaptation for person re-identification," *IEEE Trans. Multimedia*, vol. 23, pp. 1681–1695, 2021.

[10] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," 2020, *arXiv:2001.01526*.

[11] 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. (ICCV)*, Oct. 2019, pp. 6321–6330.

[12] B. Pan, Z. Cao, E. Adeli, and J. C. Niebles, "Adversarial cross-domain action recognition with co-attention," in *Proc. AAAI*, 2020, pp. 11815–11822.

[13] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang, "Self-supervised spatiotemporal learning via video clip order prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10334–10343.

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

[15] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[16] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jul. 2005, pp. 60–65.

[17] H. Li and C. Y. Suen, "A novel non-local means image denoising method based on grey theory," *Pattern Recognit.*, vol. 49, pp. 237–248, Jan. 2016.

[18] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[19] H. Chen, D. Jiang, and H. Sahli, "Transformer encoder with multi-modal multi-head attention for continuous affect recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 4171–4183, 2021.

[20] Y. Zhang, Y. Gong, H. Zhu, X. Bai, and W. Tang, "Multi-head enhanced self-attention network for novelty detection," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107486.

[21] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A^2-Nets: Double attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 352–361.

[22] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-relation networks for video classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1430–1439.

[23] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, and F. Xu, "Compact generalized non-local network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6510–6519.

[24] N. Lu et al., "MASTER: Multi-aspect non-local network for scene text recognition," *Pattern Recognit.*, vol. 117, Sep. 2021, Art. no. 107980.

[25] J. A. Rice, *Mathematical Statistics and Data Analysis*. Boston, MA, USA: Cengage Learning, 2006.

[26] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[27] M. Ramona, G. Richard, and B. David, "Multiclass feature selection with kernel Gram-matrix-based criteria," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 10, pp. 1611–1623, Oct. 2012.

[28] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[29] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "ARID: A new dataset for recognizing action in the dark," 2020, *arXiv:2006.03876*.

[30] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[31] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4768–4777.

[32] A. Tran and L.-F. Cheong, "Two-stream flow-guided convolutional attention networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3110–3119.

[33] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3-D convNet fusion for action recognition in videos with arbitrary size and length," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 634–644, Mar. 2018.

[34] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 363–378.

[35] L. Wang et al., "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[36] Z. Tu et al., "Multi-stream CNN: Learning representations based on human-related regions for action recognition," *Pattern Recognit.*, vol. 79, pp. 32–43, Jul. 2018.

[37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[38] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "ConvNet architecture search for spatiotemporal feature learning," 2017, *arXiv:1708.05038*.

[39] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[40] K. Liu, W. Liu, C. Gan, M. Tan, and H. Ma, "T-C3D: Temporal convolutional 3D network for real-time action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7138–7145.

[41] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[42] H. Yang et al., "Asymmetric 3D convolutional neural networks for action recognition," *Pattern Recognit., J. Pattern Recognit. Soc.*, vol. 85, pp. 1–12, Jan. 2019.

[43] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio–temporal attention networks for action recognition and detection," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2990–3001, Nov. 2020.

[44] 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.*, Jun. 2018, pp. 6450–6459.

[45] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 305–321.

[46] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. P. Graf, "Attend and interact: Higher-order object interactions for video understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6790–6800.

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

[48] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[49] F. Liu, L. Jiao, and X. Tang, "Task-oriented GAN for PolSAR image classification and clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2707–2719, Sep. 2019.

[50] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.

[51] J. Hoffman et al., "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[52] H. Zou, Y. Zhou, J. Yang, H. Liu, H. P. Das, and C. J. Spanos, "Consensus adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5997–6004.

[53] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4068–4076.

[54] J. Zhang, W. Li, and P. Ogunbona, "Joint geometrical and statistical alignment for visual domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1859–1867.

[55] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 687–696.

[56] Y. Zou, Z. Yu, B. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 289–305.

[57] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. P. Perez, "DADA: Depth-aware domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7364–7373.

[58] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "CrDoCo: Pixel-level domain transfer with cross-domain consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1791–1800.

[59] D. Guan, J. Huang, S. Lu, and A. Xiao, "Scale variance minimization for unsupervised domain adaptation in image segmentation," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107764.

[60] R. Panda, A. Bhuiyan, V. Murino, and A. K. Roy-Chowdhury, "Adaptation of person re-identification models for on-boarding new camera(s)," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106991.

[61] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.

[62] A. Jamal, V. P. Namboodiri, D. Deodhare, and K. Venkatesh, "Deep domain adaptation in action space," in *Proc. BMVC*, 2018, p. 264.

[63] M.-H. Chen, B. Li, Y. Bao, G. AlRegib, and Z. Kira, "Action segmentation with joint self-supervised temporal domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9454–9463.

[64] L. Fu, T. H. Nguyen, B. Min, and R. Grishman, "Domain adaptation for relation extraction with domain adversarial neural network," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, vol. 2, 2017, pp. 425–429.

[65] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," 2017, *arXiv:1701.01036*.

[66] 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.

[67] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 443–450.

[68] W. Sultani and I. Saleemi, "Human action recognition across datasets by foreground-weighted histogram decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 764–771.

[69] T. Xu, F. Zhu, E. K. Wong, and Y. Fang, "Dual many-to-one-encoder-based transfer learning for cross-dataset human action recognition," *Image Vis. Comput.*, vol. 55, pp. 127–137, Nov. 2016.

[70] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[71] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[72] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Multi-fiber networks for video recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 352–367.

[73] W. Kay et al., "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[74] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Berlin, Germany: Springer-Verlag, 2010, pp. 177–186. [Online]. Available: https://link.springer.com/book/10.1007/978-3-7908-2604-3

[75] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3723–3732.

[76] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1426–1435.

[77] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[78] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, Nov. 2008.

Yuecong Xu (Member, IEEE) received the B.Eng. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore, in 2017, and the Ph.D. degree from NTU in 2021.

He is currently a Research Scientist with the Institute for Infocomm Research, A*STAR, Singapore, and a Lecturer at NTU. His research focuses on video understanding and analysis based on deep learning and transfer learning.

Dr. Xu was a recipient of the Nanyang President's Graduate Scholarship. He was the Co-Organizer of the UG2+ Challenge for Computational Photography and Visual Recognition, held in conjunction with CVPR 2021 and CVPR 2022.

Haozhi Cao received the B.Eng. degree from the School of Electrical Engineering and Automation, Wuhan University, Wuhan, China, in 2019, and the M.Eng. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore, in 2021, where he is currently pursuing the Ph.D. degree with the School of Electrical and Electronic Engineering.

He is also working as a Research Associate with the Centre for Advanced Robotics Technology (CARTIN), NTU. His research interests include deep learning with applications in video understanding, transfer learning, and multi-modal learning.

Kezhi Mao (Member, IEEE) received the B.Eng. degree from Jinan University, Jinan, China, in 1989, the M.Eng. degree from Northeastern University, Shenyang, China, in 1992, and the Ph.D. degree from The University of Sheffield, Sheffield, U.K., in 1998.

He was a Lecturer with Northeastern University, Shenyang, China, from March 1992 to May 1995; a Research Associate with The University of Sheffield from April 1998 to September 1998; and a Research Fellow with Nanyang Technological University, Singapore, from September 1998 to May 2001, where he was also an Assistant Professor with the School of Electrical and Electronic Engineering from June 2001 to September 2005 and has been an Associate Professor since October 2005. His areas of interests include computational intelligence, pattern recognition, text mining, knowledge extraction, cognitive science, big data, and text analytic.

Zhenghua Chen (Senior Member, IEEE) received the B.Eng. degree in mechatronics engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2011, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2017.

He is currently a Scientist and the Lab Head of the Institute for Infocomm Research, and an Early Career Investigator with the Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore. His research interests include smart sensing, data analytics, machine learning, transfer learning, and related applications.

Dr. Chen has won several competitive awards, such as the First Place Winner of CVPR 2021 UG2+ Challenge, the A*STAR Career Development Award, the First Runner-Up Award for Grand Challenge at IEEE VCIP 2020, and the Finalist Academic Paper Award at IEEE ICPHM 2020. He is currently the Vice Chair of IEEE Sensors Council Singapore Chapter. He serves as an Associate Editor for *Neurocomputing* (Elsevier) and IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT.

Lihua Xie (Fellow, IEEE) received the B.E. and M.E. degrees in electrical engineering from the Nanjing University of Science and Technology, Nanjing, China, in 1983 and 1986, respectively, and the Ph.D. degree in electrical engineering from The University of Newcastle, Callaghan, NSW, Australia, in 1992.

Since 1992, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he is currently a Professor and served as the Head of Division of Control and Instrumentation from July 2011 to June 2014. He held teaching appointments with the Department of Automatic Control, Nanjing University of Science and Technology, from 1986 to 1989, and a Changjiang Visiting Professorship with the South China University of Technology, Guangzhou, China, from 2006 to 2011. His research interests include robust control and estimation, networked control systems, multi-agent control, and unmanned systems.

Dr. Xie is a fellow of the Academy of Engineering Singapore, a fellow of IFAC, and a fellow of Chinese Automation Association. He has served as an Editor of IET Book Series in Control and an Associate Editor for a number of journals, including IEEE TRANSACTIONS ON AUTOMATIC CONTROL, *Automatica*, IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS.

Jianfei Yang (Member, IEEE) received the B.Eng. degree from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, in 2016, and the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2021, where he received the Best Ph.D. Thesis Award.

He used to work as a Senior Research Engineer with the University of California at Berkeley, Berkeley, CA, USA. He is currently a Presidential Post-Doctoral Research Fellow and an Independent Principal Investigator at NTU. His research focuses on Artificial Intelligence of Things (AIoT), such as wireless sensing and computer vision based on deep learning and transfer learning.

Dr. Yang won many international AI challenges in computer vision and interdisciplinary research fields.