Multi-Source Video Domain Adaptation With
Temporal Attentive Moment Alignment Network

Yucong Xu, Member, IEEE, Jianfei Yang, Member, IEEE, Haozhi Cao, Keyu Wu,
Min Wu, Senior Member, IEEE, Zhengguo Li, Fellow, IEEE, and Zhenghua Chen, Senior Member, IEEE

Abstract—Multi-Source Domain Adaptation (MSDA) is a more practical domain adaptation scenario in real-world scenarios, which relaxes the assumption in conventional Unsupervised Domain Adaptation (UDA) that source data are sampled from a single domain and match a uniform data distribution. The MSDA is more challenging due to the existence of different domain shifts between distinct domain pairs. When considering videos, the negative transfer would be provoked by spatial-temporal features and can be formulated into a more challenging Multi-Source Video Domain Adaptation (MSVDA) problem. In this paper, we address the MSVDA problem by proposing a novel Temporal Attentive Moment Alignment Network (TAMAN) which aims for effective feature transfer by dynamically aligning both spatial and temporal feature moments. The TAMAN further constructs robust global temporal features by attending to dominant domain-invariant local temporal features with high local classification confidence and low disparity between global and local feature discrepancies. To facilitate future research on the MSVDA problem, we introduce comprehensive benchmarks, covering extensive MSVDA scenarios. Empirical results demonstrate a superior performance of the proposed TAMAN across multiple MSVDA benchmarks.

Index Terms—Multi-source, video domain adaptation, action recognition, moment alignment, dataset.

I. INTRODUCTION

VIDEO-BASED tasks (e.g., action recognition [1], [2], [3]) have been studied widely due to their applications. Among existing methods, neural networks have made remarkable advances in these tasks thanks to the emergence of large-scale labeled datasets for training and testing. However, sufficient labeled training videos may not be readily available in real-world scenarios owing to the high cost of video data annotation. Subsequently, various Unsupervised Domain Adaptation (UDA) and Video Unsupervised Domain Adaptation (VUDA) methods have been introduced to transfer knowledge from a labeled source domain to an unlabeled target domain by reducing discrepancies between source and target domain.

Though existing UDA and VUDA methods [4], [5], [6], [7] enable the transfer of knowledge across domains, they normally assume that the source data are sampled from a single domain and match a uniform data distribution. Such an assumption may not hold in real-world applications. In practice, with the availability of a variety of large-scale labeled public datasets, source data are more likely to be collected from multiple datasets. This scenario is defined as Multi-Source Domain Adaptation (MSDA) which relaxes the constraint of identical source data distribution by assuming that source data are sampled from multiple domains corresponding to different data distributions. The MSDA problem is more difficult owing to the existence of different levels of domain shifts among source domains and between different source-target domain pairs, which adversely affects the alignment of target data, resulting in negative transfer.

Besides spatial features that are used in image representations, temporal features are also the key component in video representations. The presence of the additional features engenders a novel Multi-Source Video Domain Adaptation (MSVDA) problem, which aims to transfer networks trained with data from the multiple source domains to the target domain. The MSVDA empowers models trained in a collection of large-scale video datasets (e.g., UCF101 [8] and HMDB51 [9]) to be employed directly to smaller-scale datasets (e.g., ARID [10]) without label supervision. For the MSVDA, negative transfer would be triggered if we directly reduce the divergence between multiple domain pairs regardless of inconsistent domain shifts caused by distinct spatial and temporal feature distributions, as shown in Fig. 1.

The key towards solving the MSVDA problem lies in leveraging the additional temporal features explicitly to mitigate negative transfer in the MSVDA. To tackle the negative transfer in the MSVDA, we argue that the temporal features should be utilized from two perspectives: firstly, effective global temporal features should be constructed with attention to dominant local temporal features with higher local classification confidence, alleviating the probability of provoking the negative transfer caused by temporal features; secondly,
the temporal features should contribute towards the overall feature alignment process together with the spatial features, alleviating the possible misalignment of the spatial features. Empirical results as shown in Sec. V justify such argument.

To this end, we propose a novel Temporal Attentive Moment Alignment Network (TAMAN) to address the challenges in the MSVDA uniformly. The TAMAN first constructs robust global temporal features with high transferability by attentively combining local temporal features which represent the different characteristics of the overall motion. The attention strategies depend on both the local temporal feature classification confidence, as well as the disparity between the global and local feature discrepancies. Meanwhile, the TAMAN aligns spatial-temporal features jointly by aligning the moments of both spatial and temporal features across all domain pairs, mitigating the possible negative transfer caused by the misalignment of spatial features.

To promote MSVDA relevant research, we further propose two sets of comprehensive benchmarks, utilizing both widely used public datasets in action recognition and a more recent dataset built with dark videos. The proposed benchmarks are: (i) Daily-DA, constructed with the ARID [10], HMDB51 [9], Kinetics [11], and Moments-in-Time [12] datasets; and (ii) Sports-DA, constructed with the UCF101 [8], Sports-1M [13], and Kinetics datasets. The proposed benchmarks cover extensive MSVDA scenarios with distinct domain shifts across included domains.

In summary, our contributions are threefold.

- Firstly, we formulate a novel practical and challenging Multi-Source Video Domain Adaptation (MSVDA) problem. To the best of our knowledge, this is an initial work that investigates multi-source domain adaptation in the video classification field, especially action recognition.
- Secondly, we analyze the challenges of the MSVDA problem and propose a novel TAMAN to address the challenges. The TAMAN learns robust global temporal features with local temporal attention strategies, while utilizing moments of both the spatial and temporal features for feature alignment across domain pairs jointly.
- Finally, we introduce two sets of MSVDA benchmarks and exhibit the capability of TAMAN, which achieves superior performances across all the proposed MSVDA benchmarks.

II. RELATED WORK

A. Unsupervised Domain Adaptation (UDA)

Current UDA methods aim to distill shared knowledge across domains with the labeled source domain and unlabeled target domain, thus improving the transferability of models. In general, these methods could be divided into three categories: a) reconstruction-based methods [14], [15], [16], [17], where domain-invariant features are obtained by encoders trained under data-reconstruction schemas, typically formulated as encoder-decoder networks; b) adversarial-based methods [5], [7], [18], [19], [20], which are inspired by the success of GAN [21], and are designed with additional domain discriminators that are trained jointly with feature generators in an adversarial manner [22], minimizing adversarial losses [23]; and c) discrepancy-based methods [4], [24], [25], [26], [27], which alleviate domain shifts across source-target domain pairs by employing various metric learning schemas, such as MMD [24], CORAL [28] and KL-divergence [29]. Discrepancy-based methods do not require additional network structures (e.g., domain classifiers), thus are more stable and easy to train. More recently, with the wide applications of videos in various fields, there has been increasing research for Video Unsupervised Domain Adaptation (VUDA). The success of obtaining domain-invariant features with the above UDA methods extends to VUDA, with multiple VUDA methods proposed for tasks such as action recognition [6], [30], [31], [32] and action segmentation [33].

B. Multi-Source Domain Adaptation (MSDA)

Though UDA and VUDA methods have made outstanding progress, current approaches generally assume that the training source data are sampled from a single domain and follow a uniform data distribution. A more general and practical scenario that relaxes this assumption is denoted as Multi-Source Domain Adaptation (MSDA) [34], [35], which enables models to transfer knowledge from multiple sources. Earlier MSDA methods rely on either hand-crafted feature representations [36], [37] or pre-trained classifiers [38], [39]. These works demonstrate the applications of MSDA in fields such as image classification [34], [36] and multimedia classification [37], [40].

With the advances in deep neural networks, various end-to-end MSDA methods have been proposed in conjunction with deep learning. Among these, MDAN [41] aligns the target domain to source domains globally with adversarial learning, applying a domain discriminator for each source-target domain pair and a single task classifier. Meanwhile, DCTN [42] improves on MDAN by deploying a separate task classifier for each source domain, with the final result being a weighted combination of the output predictions. Further, MDDA [43] introduces a source distillation mechanism for fine-tuning both the feature extractor and the task classifier while CMSS [44]...
introduces a dynamic curriculum that updates the error rate of domain discriminators constantly. Meanwhile, M3SDA [45] utilizes a moment matching component for transferring knowledge. More recently, LiC-MSDA [46] propose to assist model prediction on the target domain by learning from how source domains interact with each other leveraging on knowledge graphs, along with a Relation Alignment Loss to constrain the global and local relations of feature representations. Further, MCC [47] measures the pairwise class confusion on the target domain, while MOST [48] propose a rigorous OT-based theory to leverage imitation learning for MSDA.

Despite the notable progress made in the MSDA with its various applications, the current MSDA approaches are mostly built for image-based MSDA, with both the source and target domains being image data. However, the MSVDA, which focuses on video-based knowledge transfer, has not been dealt with. The MSVDA is more challenging due to the possibility that negative transfer could be triggered by temporal features, which do not exist in images. We propose to tackle the MSVDA with a novel method that constructs robust global temporal features with local temporal attention strategies while utilizing moments of both spatial and temporal features for effective feature alignment.

III. PROPOSED METHOD

In the scenario of Multi-Source Video Domain Adaptation (MSVDA), we are given a collection of $M$ source domains denoted as $S = \{S_1, S_2, \ldots, S_M\}$, with domain $S_m = (\{Vi_{S_m} \mid \forall i \in S_m\})_{i=1}^{n_m}$ containing $n_m$ i.i.d. labeled videos associated with $K$ classes and characterized by a probability distribution of $p_{S_m}$.

A single target domain $T = \{Vi_T\}_{i=1}^{n_T}$, with $n_T$ i.i.d. unlabeled videos characterized by a probability distribution of $p_{T}$ is accessed. We assume that the unlabeled target domain videos share the same label space with the source domain videos. To tackle the MSVDA problem, our goal is to build a robust network capable of learning transferable features across the multiple video source domains and the video target domain, while minimizing the target classification risk. In contrast to the conventional VUDA, the MSVDA is more challenging owing to the existence of both domain shifts between the different source-target domain pairs and among the different source domains.

Moreover, while there are existing MSDA approaches that tackle the negative effect brought by the extra domain shifts, these approaches are mostly built for image-based MSDA problems. Negative transfer in these problems would only be provoked by domain shifts caused by spatial features. However, videos contain temporal features which represent the temporal correlation information, thus negative transfer could be further triggered by different domain shifts w.r.t. temporal features. Since only spatial features are relevant for image-based MSDA problems, temporal features have not been considered in existing MSDA works. Therefore, a novel Temporal Attentive Moment Alignment Network (TAMAN) is proposed to transfer from multiple video source domains while alleviating the negative transfer with a full usage of effective temporal features built in an attentive manner. We start with a brief review of discrepancy-based MSDA approaches utilizing moment alignment, proceeded by a detailed description of the proposed TAMAN.

A. Discrepancy-Based MSDA With Moment Alignment

The goal of conventional UDA and its variants is to align the data distributions of source and target domains. Discrepancy-based approaches are widely used thanks to their ability in alleviating domain shift through metric learning schemas without additional network components. Given the feature extractor $G_f$, the source classifier $C_S$, with $x_S$ and $x_T$ being collections of $n_S$ source domain samples and $n_T$ target domain samples, the overall objective of discrepancy-based DA methods are generally formulated as:

$$L = \frac{1}{n_S} \sum_{x \in S} L_S(C_S(G_f(x)), y) + \lambda_d \cdot d(G_f(x_S), G_f(x_T)),$$

(1)

where $L_S$ stands for the source classification loss with $y$ being the ground truth label of input $x$ from the source domain, while $\lambda_d$ is the trade-off parameter for the cross-domain discrepancy $d$. While various forms of discrepancies have been proposed, a major line of which is moment-based. Minimizing such discrepancies could therefore be viewed as moment alignment schemas. Typical examples include MMD [24], which matches the first moments of distributions, and CORAL [28], which matches the second moments of distributions.

While these moment alignment methods could align data distributions under conventional UDA settings, their performances degrade substantially when applying directly to the MSDA tasks, due to the negative transfer caused by the domain shifts between the different source-target domain pairs and within the different source domains.

As proven in [45], the upper bound of the target classification risk relates closely with the pairwise cross-moment discrepancy between the target domain and each source domain, denoted as $d_{CM}(S, T)$, which can be formulated as:

$$d_{CM}(S, T) = \sum_{j=1}^{M} \lambda_j \sum_{k} d_{CM}^k(S_j, T),$$

(2)

where $d_{CM}^k(S_j, T)$ is the $k$-th moment discrepancy between the $j$-th source domain and the target domain. Since $d_{CM}^k(\cdot, \cdot)$ is a metric, it follows the triangle inequality, formulated as:

$$d_{CM}^k(S_j, S_j) \leq d_{CM}^k(S_j, T) + d_{CM}^k(S_j, T).$$

(3)

This implies that the cross-moment discrepancy between domains $S_j, T$ is lower bounded by the pairwise discrepancies between source domains. Combining Equations 2 and 3, a moment distance is introduced as:

$$d_{CM}(S, T) = \sum_{k} \left( \frac{1}{M} \sum_{i=1}^{M} \|E(x_{S_i}^k) - E(x_T^k)\|_2^2 + \left( \frac{M}{2} \right)^{-1} \sum_{i,j=1 \atop i \neq j}^{M} \|E(x_{S_i}^k) - E(x_{S_j}^k)\|_2^2 \right).$$

(4)

The moment distance $d_{CM}$ enables effective multi-source domain adaptation through minimizing both the source-target
domain discrepancies and discrepancies among the different source domains. The overall objective function of MSDA is finally formulated as suggested in [45]:

$$L_{ms} = \sum_{j=1}^{M} \frac{1}{|S_j|} \sum_{x \in S_j} L_{S_j}(C_{S_j}(G_f(x)), y) + \lambda_d d_{CM}(S, T),$$

(5)

where $L_{S_j}$ stands for the source classification loss for the $j$-th source classifier with $y$ being the ground truth of input $x$ from the $S_j$, while $\lambda_d$ is the trade-off parameter.

B. Temporal Attentive Moment Alignment Network

To achieve the MSVDA, an intuitive approach would be to apply the model matching to video data directly by integrating videos into Eq. 5, i.e., $x \rightarrow V$. Meanwhile, the original image feature extractors (e.g., 2D-CNN) could simply be substituted with video feature extractors (e.g., 3D-CNN).

Despite the simplicity of the moment alignment, empirical results suggest that such a method is insufficient to deal with the negative transfer in the MSDVA well, leading to inferior adaptation results. We can expect such inferior results, since the video representations obtained through common feature extractors (i.e., convolutional neural network (CNN)-based extractors) focus primarily on spatial features. In contrast, temporal features are normally obtained vaguely with a simple pooling process across the temporal dimension, leading to tremendous distribution shifts. Without explicit considering the temporal features, domain adaptation through moment alignment may only be performed on the spatial features. The negative transfer provoked by the temporal features could not be addressed by simply reducing the moment distance.

Based on the above observation, we introduce a novel Temporal Attentive Moment Alignment Network (TAMAN) to perform multi-source adaptation by exploiting both spatial and temporal features, with its structure shown in Fig. 2.

To utilize the temporal features for moment alignment, a key prior is to obtain the temporal features explicitly. Compared to conventional CNN-based extractors (e.g., 3D-ResNet [49]) whose temporal features are obtained implicitly via a temporal pooling, the Temporal Relation Network (TRN) [50] is adopted. This is thanks to its ability to extract temporal features through reasoning over the correlations between spatial representations, which coincides with the human approach on recognizing actions and is therefore more effective than other spatial-temporal networks [51]. The TRN has also been adopted in other VUDA tasks and approaches, such as PVDA [52] and SFVDA [53], bringing state-of-the-art results. With the frame-level spatial features obtained from the shared spatial feature extractor $G_{sp}$, the $v$-th input video from domain $S_m$ with $h$ frames is expressed as $V_{S_m} = \{f^{(1)}_{S_m}, f^{(2)}_{S_m}, \ldots, f^{(h)}_{S_m}\}$. Here $f^{(i)}_{S_m}$ is the $i$-th frame-level spatial feature of the $v$-th video from domain $S_m$. For clarity, the subscript $v$ is omitted in subsequent equations. The TRN constructs the global temporal features of $V_{S_m}$ denoted by $t_{S_m}$ by aggregating multiple clip-level local temporal features, each of which is built from $r$ temporally-ordered frames with $r \in [2, h]$. A clip-level local temporal feature is defined as:

$$t_{S_m}^{(r)} = \sum_{z} g_{lt}^{(r)}(V_{S_m}^{(r)});$$

(6)

Here $(V_{S_m}^{(r)})_z = \{f^{(a)}_{S_m}, f^{(b)}_{S_m}, \ldots\}_z$ represents the $z$-th clip that contains $r$ temporally-ordered frames, with frame indices $a$ and $b$. Note that $b > a$ but $a$ and $b$ can be nonconsecutive. The local temporal feature $t_{S_m}^{(r)}$ is computed by integrating the collection of temporal-ordered frame-level spatial features through an integration function $g_{lt}^{(r)}$, implemented as a Multi-layer Perceptron (MLP).

The global temporal features could be obtained by simple aggregation strategies applied to all the local temporal features (e.g., an average operation). However, the contribution of each local temporal feature is empirically not equal, which...
formulation of the connection for more stable optimization, and a motivates us to develop a local attention mechanism with two attention strategies for dominant domain-invariant local temporal features. Firstly, inspired by findings in [30], we enable TAMAN to focus on more transferable local temporal features. To this end, TAMAN learns the global temporal features that correspond to a higher local class prediction confidence. Specifically, the prediction of each local temporal feature is first obtained by applying the classifier of domain $S_m$ to feature $l_t*S_m$, denoted as $\hat{y}_{lt,S_m} = C_{S_m}(l_t*S_m)$. It indicates the probability of the local temporal feature classified as each video class. Suppose there are a total of $K$ video classes, the confidence of prediction $\hat{y}_{lt,S_m}$ is defined as the additive inverse of its entropy computed over all the probabilities as:

$$C(\hat{y}_{lt,S_m}) = \sum_{c=1}^{K} \hat{y}_{lt,S_m,c} \log(\hat{y}_{lt,S_m,c}),$$

(7)

where $\hat{y}_{lt,S_m,c}$ corresponds to the prediction of the $c$-th class. The local confidence weight corresponding to the local temporal feature $l_t*S_m$ is generated by adding a residual connection for more stable optimization, and a tanh function for constraining the weights within the range of $[0,1]$. The formulation of the local confidence weight is thus:

$$w_{C,m} = \tanh(1 + C(\hat{y}_{lt,S_m})).$$

(8)

Secondly, inspired by temporal action localization and action detection tasks, it is believed that most actions would be observed in a local temporal range [54], [55], [56], therefore effective global temporal features should be constructed by focusing on the dominant local temporal features while discarding the ineffective clips that may lead to domain shifts. Due to the fact that target videos are unlabeled, it is impossible to obtain the prediction accuracies of each local temporal feature. Instead, the dominance weight is derived from the disparity between the global and local temporal feature discrepancies. Formally, the raw global temporal features from the source domain $S_m$ and the target domain $T$, denoted as $\hat{t}_{S_m}$ and $\hat{t}_{T}$, are obtained by a simple additive aggregation of the clip-level local temporal features, i.e., $\hat{t}_{S_m} = \sum_r l_t^{(r)}_{S_m}$ and $\hat{t}_{T} = \sum_r l_t^{(r)}_{T}$. The feature discrepancy is defined based on the cross-moment discrepancy in Eq. 4, where the moment-based local temporal discrepancy $d_{lt}^{(r)}$ is formulated as:

$$d_{lt}^{(r)}(S,T) = \sum_{k=1}^{M} \frac{1}{M} \sum_{i=1}^{M} \|E((l_t^{(r)})^k) - E((l_t^{(r)})^k)\|_2 + \left(\frac{M}{2}\right)^{-1} \sum_{i,j\in[1,M]} \|E((l_t^{(r)})^k) - E((l_t^{(r)})^j)\|_2).$$

(9)

The global temporal discrepancy $d_l$ is defined similarly. The dominance weight is subsequently generated by the disparity between $d_l$ and $d_{lt}^{(r)}$, computed as $d_{lt}^{(r)} = |d_l - d_{lt}^{(r)}|$. The dominance weight $w_{dom}$ is therefore formulated as:

$$w_{dom}^{(r)} = e^{d_{lt}^{(r)}} / \sum_c e^{d_{lt}^{(r)}}.$$

(10)

Finally, the global temporal feature is an attentive aggregation of all local temporal features, with the local attention weight $w_{C,m}^{(r)}$ being the multiplication of the local confidence weight and the dominance weight, i.e., $w_{C,m}^{(r)} = w_{C,m}^{(r)}w_{dom}^{(r)}$. It is further normalized such that $\sum_r w_{C,m}^{(r)} = 1$. The global temporal feature for the source data in domain $S_m$, denoted as $\hat{t}_{S_m}$ is therefore formulated as:

$$\hat{t}_{S_m} = \sum_r w_{C,m}^{(r)}l_t^{(r)}_{S_m}.$$ 

(11)

With the global temporal features extracted, the TAMAN aims to perform feature alignment for the spatial and temporal features jointly. This is achieved by minimizing the moment-based feature discrepancies $d_f$ and $d_t$ concurrently. Both $d_f$ and $d_t$ are defined equivalently with Eq. 9. Overall, the objective function for the TAMAN is expressed as:

$$L_{ams} = \sum_{j=1}^{M} \frac{1}{n_{S_j}} \sum_{u} L_{S_j}(C_{S_j}(t_{S_j}), y_{u,S_j}) + \lambda_d d_f(S,T) + \lambda_d d_t(S,T),$$

(12)

where $L_{S_j}$ stands for the classification loss for the $j$-th source classifier with $y_{u,S_j}$ being the ground truth of the $u$-th input source video from domain $S_j$, while $\lambda_d$ is the trade-off parameters for the moment-based spatial and temporal feature discrepancies respectively.

During the testing phase, the target data are first propagated through the spatial and temporal feature extractors, and then the $M$ classifiers trained by the source data. To obtain the final classification prediction, the outputs from all the classifiers $P_j = C_{S_j}(t_{T}), j \in [1,M]$ are combined. The most intuitive method is to average all the outputs. Yet, since the domain shift between different source-target domain pairs are different, their target accuracies also vary. To address this issue, we propose a weighted ensemble schema to combine the outputs effectively. The idea behind the prediction weight $w_p$ is that the final prediction should focus on the classifier whose output is of higher certainty. Given that the sum of the weights, i.e., $\sum_{j=1}^{M} w_p j$ should be 1, the prediction weight is defined as:

$$w_{p,j} = \sigma(\sum_c P_{j,c} \log(P_{j,c})).$$

(13)

Here $K$ is the number of video classes, $P_{j,c}$ corresponds to the prediction of the $c$-th class from the $j$-th classifier, while $\sigma$ is the softmax function performed across the $M$ classifiers, i.e., $\sigma(x_j) = \exp(x_j) / \sum_{j=1}^{M} \exp(x_j)$. The final prediction is therefore the weighted sum of predictions from each classifier guided by the prediction weight $w_p$.

In summary, TAMAN addresses MSVDA by constructing highly transferable global temporal features by attentively combining local temporal features, focusing on those that are more transferable with higher local class prediction confidence.
and those that are more dominant with lower disparity between the global and local discrepancies, while discarding the ineffective clips. TAMAN further aligns the spatial and temporal features jointly by minimizing separate moment-based discrepancies across all domain pairs. To highlight the novelty of TAMAN, we compare our TAMAN with previous MSDA and VUDA methods. Specifically, we compare with M3SDA [45], LtC-MSDA [46], TA3N [30], and ACAN [32]. The methods are compared from two perspectives: the tasks they tackle and the techniques leveraged, as shown in Table I.

Table I: Detailed Comparison of TAMAN With Related But Different MSDA and VUDA Methods

| Method       | Publication | Task                                                                 | Techniques                                                                 |
|--------------|-------------|----------------------------------------------------------------------|-----------------------------------------------------------------------------|
| M3SDA [45]   | ICCV-19     | Multi-Source Domain Adaptation (MSDA): labeled source data collected from multiple domains corresponding to different data distribution, target data are unlabeled, image-based. | MSDA leverages a moment matching component operating on the image features which aligns the multiple source domains with the target domain while aligning the source domains with each other simultaneously. |
| LtC-MSDA [46] | ECCV-22     | Multi-Source Domain Adaptation (MSDA): labeled source data collected from multiple domains corresponding to different data distribution, target data are unlabeled, image-based. | (a) LtC-MSDA assists model prediction on the target domain by learning from the interactions across source domains leveraging on knowledge graphs; (b) LtC-MSDA further constrains the global and local relations of image feature representations through applying a Relation Alignment Loss on the category level.         |
| TA3N [30]    | ICCV-19     | Video Unsupervised Domain Adaptation (VUDA): labeled source video data collected from single domain, target video data are unlabeled, video-based.                      | (a) TA3N aligns videos across the source and target domains by applying adversarial-based domain adaptation with a single video domain discriminators across both spatial and local temporal features; (b) TA3N attends to the local temporal features with high domain discriminability. |
| ACAN [32]    | TNNLS-22    | Video Unsupervised Domain Adaptation (VUDA): labeled source video data collected from single domain, target video data are unlabeled, video-based. | (a) ACAN adopts an adversarial-based framework with a single video domain discriminator and aligns the correlation features of source and target videos in the form of long-range spatiotemporal dependencies; (b) ACAN further improves the correlation alignment by aligning the joint distribution of correlation information across the source and target video domains through minimizing a Pixel Correlation Discrepancy. |
| TAMAN (Ours) | -           | Multi-Source Video Domain Adaptation (MSVDA): labeled source video data collected from multiple domains corresponding to different video data distribution, target video data are unlabeled, video-based. | (a) TAMAN constructs robust global temporal features with high transferability by attracting different local temporal features; (b) TAMAN adopts a local attention mechanism to focus on local temporal features that are more transferable with higher local class prediction confidence and are more dominant with lower disparity between the global and local feature discrepancies; (c) TAMAN aligns both the spatial and temporal features jointly by aligning the moments of spatial and temporal features across all domain pairs. |

IV. MSVDA Benchmarks

There are very limited cross-domain benchmark datasets for VUDA and its variant tasks. For the few cross-domain datasets available such as UCF-HMDBfull [30] for VUDA and HMDB-ARD | partial | [52] for Partial Video Domain Adaptation (PVDA), the source domains are always constrained to be a single domain. To facilitate MSVDA research, we propose two sets of comprehensive benchmarks, namely the Daily-DA and the Sports-DA datasets. Both datasets cover extensive MSVDA scenarios and provide adequate baselines with distinct domain shifts to facilitate future MSVDA research.

A. Daily-DA Dataset

The Daily-DA dataset comprises of videos with common daily actions. It is constructed from four action datasets: ARID (A) [10], HMDB51 (H) [9], Moments-in-Time (M) [11], and Kinetics (K) [12] (using the Kinetics-600 version). Among them, the HMDB51, Moments-in-Time, and Kinetics datasets are widely used for action recognition benchmarking collected from various public video platforms (e.g. YouTube, Flickr). The ARID is a more recent dataset, comprised with videos shot under adverse illumination conditions. Statistically, videos in the ARID are characterized by their low RGB mean value and standard deviation, which results in larger domain gaps between the ARID and other video domains. A total of 8 overlapping classes are collected, which are listed in Tab. II, resulting in a total of 18,949 videos. There are 2,776 training videos and 1,289 testing videos from ARID; 560 training videos and 240 testing videos from HMDB51; 4,000 training videos and 400 testing videos from Moments-in-Time; and 8,959 training videos and 725 testing videos from Kinetics. When performing MSVDA, one dataset is selected as the target domain, with the other three datasets as the source domains. We therefore construct four MSVDA tasks: Daily→A, Daily→H, Daily→M, and Daily→K. The training and testing splits are separated following the official splits for each dataset. Fig. 3 shows the comparison of sampled frames from sampled classes in the Daily-DA dataset.

B. Sports-DA Dataset

The Sports-DA dataset comprises of videos with common sport actions, and is built from three large-scale action datasets: UCF101 (U) [8], Sports-1M (S) [13], and Kinetics (K) (also using the Kinetics-600 version). Compared to the Daily-DA, this dataset is much larger in terms of both the number of classes and videos. A total of 23 overlapping classes are collected which are listed in Tab. III, resulting in a

Table II: List of Overlapping Classes for Daily-DA Dataset

| ARID Class | HMDB51 Class | Moments-in-Time Class | Kinetics Class |
|------------|--------------|-----------------------|----------------|
| Drink      | drink        | drinking              | drinking shots |
| Jump       | jump         | jumping               | jumping bicycle |
| Pick       | pick         | picking               | picking fruit  |
| Push       | push         | pushing               | pushing bike   |
| Run         | run          | running               | running on treadmill |
| Walk       | walk         | walking               | walking the dog |
| Wave       | wave         | waving                | waving hand    |
Fig. 3. Sampled frames of videos from sampled classes in Daily-DA.

Fig. 4. Sampled frames of videos from sampled classes in Sports-DA.

total of 40,718 videos, making the Sports-DA dataset one of the largest cross-domain video datasets introduced. There are 2,145 training videos and 851 testing videos from UCF101; 14,754 training videos and 1,900 testing videos from Sports-1M; and 19,104 training videos and 1,961 testing videos from Kinetics. As videos in both the original Sports-1M and Kinetics dataset are provided as YouTube links, we ensure that the collected videos are still valid. Invalid links are all omitted during collection. The Sports-DA dataset is designed to validate the effectiveness of MSVDA approaches on large-scale video data. Similar to the Daily-DA dataset, one dataset is selected as the target domain, with the other two datasets as the source domains when performing MSVDA, resulting in three MSVDA tasks: Sports → U, Sports → S, and Sports → K. We follow the official split for separating the training and testing sets. Fig. 4 shows the comparison of sampled frames from sampled classes in the Sports-DA dataset.

To further demonstrate the challenge of both benchmarks, we provide the baseline results under the vanilla VUDA [32] setting (i.e., single source), utilizing the discrepancy based MMD approach [23] and the adversarial based DANN approach [24] as demonstrated in Tab. IV and Tab. V. We also present the results without applying any domain adaptation approach (i.e., source-only results).

V. EXPERIMENTS

In this section, we evaluate our proposed TAMAN by conducting cross-domain action recognition on MSVDA benchmarks proposed in Sec. IV. We present superior results on both proposed benchmarks. Ablation studies and empirical analysis of TAMAN are also presented to justify our design.

A. Experimental Settings

Cross-domain action recognition tasks are performed on both the Daily-DA and Sports-DA datasets, with a total of 7 cross-domain settings as presented in Sec. IV. Following standard UDA evaluation protocols [57], source videos are labeled while target videos are strictly unlabeled. For fair comparison, both the proposed TAMAN and all compared methods employ the Temporal Relation Network (TRN) [50] as the feature extractor backbone, which is pretrained on the ImageNet [58]. All experiments are implemented with PyTorch [59] library. To obtain video features, we instantiate the TRN with the ResNet-101 [60] as the backbone for video feature extraction for both source and target domain videos, with the model pretrained on the ImageNet [58]. The TRN has been widely adopted in previous video domain adaptation tasks, such as VUDA [30] and PVDA [52] and brought state-of-the-art results on these tasks through its capability in obtaining explicit temporal features, as presented in Sec. III-B. The source and target feature extractors share parameters. New layers are trained from scratch, and their learning rates are set to 0.001. The pretrained layers which outputs the frame-level spatial features $f$ are frozen.

The stochastic gradient descent (SGD) algorithm [61] is used for optimization, with the weight decay set to 0.0001 and the momentum to 0.9. The batch size is set to 64 per GPU per domain. Our initial learning rate is set to 0.001 and is divided by 10 for three times during training. We train our networks with a total of 100 epochs for the Daily-DA dataset and 40 epochs for the Sports-DA dataset. The trade-off weight for the moment-based spatial and temporal feature discrepancies $\lambda_{df}$ and $\lambda_{dt}$ are set to 0.005 and 0.01. All experiments are conducted using two NVIDIA RTX 2080 Ti GPUs.

B. Overall Results and Comparisons

We first compare the TAMAN with various UDA/VUDA and MSDA approaches, which include: (i) adversarial-based methods: DANN [23], ADDA [18], TA3N [30],
the improvements lessen to around 7% for other tasks on the source-only results by more than 27% for Daily-

large. As presented in Tab. VI, the TAMAN improves on the significant when the domain shift between source domains is

ments of TAMAN towards the source-only results are less
domain gaps.

Further, it could be observed that for all prior UDA/VUDA and

strategies are employed: (i) single-best (‘s-’), where the adaptation is performed for each source-target pair with the best result selected; and (ii) source-combined (‘c-’), where all source domains are combined to form a domain. The results are presented in Tab. VI. Following [45], we report the mean and standard deviation (std) of the top-1 accuracy with 5 runs under identical network settings. For comparison, we also report the results of the backbone TRN trained without any adaptation approaches.

Results in Tab. VI demonstrate the effectiveness of TAMAN, achieving the best results on all MSVDA tasks and outperforming all prior approaches by noticeable gains. Notably, the TAMAN outperforms all the image-based MSDA approaches (e.g., MDAN, DCTN, and MDDA) consistently by an average of more than 10% relative improvements in mean accuracy. This empirically justifies the effectiveness of constructing the temporal attentive robust global temporal features which are more transferable while incorporating both the spatial and temporal features for feature moment alignment. Further, it could be observed that for all prior UDA/VUDA and MSDA approaches, the adaptation results are inferior to that of the backbone TRN trained with supervised source data only and tested on the target data (source-only results).

In particular, the negative effect is more severe for the Daily-

DA, with an average of 12 out of 17 approaches evaluated suffering from the negative transfer. This owes to the fact that the Daily-DA dataset contains data collected from ARID

ACAN [32], MDAN [41], DCTN [42] and MDDA [43]; and (ii) discrepancy-based methods: MMD [24], MCD [26], CORAL [28], LiC-MSDA [46], MCC [47], MOST [48], and M3SDA [45]. For UDA/VUDA approaches (i.e., DANN, ADDA, TA\textsuperscript{2}N, ACAN, MMD, MCD, and CORAL), two strategies are employed: (i) single-best (‘s-’), where the adaptation is performed for each source-target pair with the best result selected; and (ii) source-combined (‘c-’), where all source domains are combined to form a domain. The results are presented in Tab. VI. Following [45], we report the mean and standard deviation (std) of the top-1 accuracy with 5 runs under identical network settings. For comparison, we also report the results of the backbone TRN trained with supervised source data only and tested on the target data (source-only results).

C. Ablation Studies and Feature Visualization

To further validate the efficacy of TAMAN and justify its design, we perform detailed ablation studies and feature visualization. The ablation studies are conducted from four perspectives: (i) local attention weight and its components; (ii) different strategies for obtaining dominance weight; (iii) different prediction ensemble schemas; and (iv) significance of leveraging both spatial and temporal feature moment discrepancies. Ablation studies are conducted with the Daily→A and Daily→H tasks.

1) Local Attention Weight: We evaluate TAMAN against 3 variants to justify the design of the local attention weight:

(a) TAMAN w/o local confidence, where the local confidence weights are set to be equal for all local temporal features;
(b) TAMAN w/o dominance, where the local attention weight does not incorporate dominance weights; and
(c) TAMAN w/o local attention, where the global temporal features are built by additive aggregation of local temporal features. Results presented in Tab. VII clearly demonstrate the necessity of both the local confidence weight and the dominance weight, both of which help construct the robust global temporal features for alignment. By employing either weight, TAMAN learns more transferable temporal features given the better result compared to all approaches evaluated in Tab. VI. It is also noted that though TAMAN w/o local attention falls behind TAMAN by a notable gap, it still performs better than most image-based MSDA approaches, justifying the need for joint alignment of both the spatial and temporal features.

2) Obtaining Dominance Weights: We propose the dominance weight which is obtained from the disparity between the global and local feature discrepancies in Sec. III-B. Alternatively, the dominance weight could be obtained directly by comparing the local temporal feature discrepancies. Therefore, we justify the current strategy for obtaining the dominance weight by evaluating TAMAN against TAMAN w/o

| Source-only | ARID | HMDB51 | MIT | Kinetics |
|-------------|------|--------|-----|----------|
| ARID        | N/A  | 28.75  | 22.25| 27.17    |
| HMDB51      | 16.84| N/A    | 26.50| 38.90    |
| MIT         | 24.36| 44.17  | N/A  | 64.69    |
| Kinetics    | 24.52| 38.75  | 25.75| N/A      |

| Source-only | ARID | HMDB51 | MIT | Kinetics |
|-------------|------|--------|-----|----------|
| ARID        | N/A  | 28.75  | 22.25| 27.17    |
| HMDB51      | 16.84| N/A    | 26.50| 38.90    |
| MIT         | 24.36| 44.17  | N/A  | 64.69    |
| Kinetics    | 24.52| 38.75  | 25.75| N/A      |

| Source-only | UCF101 | Sports-1M | Kinetics |
|-------------|---------|------------|----------|
| ARID        | N/A     | 46.32      | 62.13    |
| Sports-1M   | 80.02   | N/A        | 68.86    |
| Kinetics    | 85.90   | 55.16      | N/A      |

| Source-only | UCF101 | Sports-1M | Kinetics |
|-------------|---------|------------|----------|
| ARID        | N/A     | 47.21      | 64.17    |
| Sports-1M   | 73.55   | N/A        | 69.22    |
| Kinetics    | ARID    | HMDB51     | N/A      |

| Source-only | UCF101 | Sports-1M | Kinetics |
|-------------|---------|------------|----------|
| ARID        | N/A     | 46.74      | 61.77    |
| Sports-1M   | 81.20   | N/A        | 68.35    |
| Kinetics    | 86.60   | 55.05      | N/A      |
 dominance and two other variants: (a) TAMAN w min. \(d_{tr}^{dom} \) dominance, where the global temporal features are set to focus on the local temporal feature with the minimum cross-domain moment discrepancy; and (b) TAMAN w max. \(d_{tr}^{dom} \) dominance, whose global temporal features attend to the local temporal feature with maximum cross-domain moment discrepancy. As shown in Tab. VIII, the results justify the design of the dominance weight through the disparity of discrepancies. While the other two strategies are computed with ease, their inferior results to TAMAN w/o dominance show that the sub-optimal dominance weight may negatively affect the global temporal features.

3) Prediction Ensemble Schemas: TAMAN utilizes a weighted ensemble schema based on prediction certainty. To justify such an approach, we compare TAMAN with the following variants: (a) TAMAN by avg, whose final prediction is ensembled by directly averaging across outputs from each classifier; and (b) TAMAN by src. only accuracy, whose prediction is ensembled following [45], with the weights of each prediction output derived by the source only accuracy between each source-target domain pair. As demonstrated in Tab. IX, the performance improvement of the ensemble strategy in TAMAN is marginal, indicating that the domain-variant feature learning plays a more vital role in MSVDA. It is noted that though the ensemble method in [45] is more effective than simple averaging, it requires the evaluation of source-only results with each individual source-target domain pair, resulting in more computation and less efficiency.

4) Spatial and Temporal Feature Moment Discrepancies: The TAMAN leverages moment-based discrepancies for both spatial and temporal features. To better understand the significance of leveraging both spatial and temporal features for computing moment-based discrepancies, we compare the proposed TAMAN with variants TAMANdf and TAMANdt, where (a) TAMANdf only optimizes the moment-based spatial discrepancy, while (b) TAMANdt only optimizes the moment-based temporal discrepancy. Note that both the TAMANdf and TAMANdt optimize the classification loss for the source classifiers. As compared in Tab. X, optimizing both the moment-based spatial discrepancy and the moment-based temporal discrepancy complements each other, justifying the need to optimize both discrepancies for ultimate performance.
knowledge transfer. It should also be noted that while both the TAMAN$_{df}$ and TAMAN$_{df}$ perform inferior against TAMAN, the performance of TAMAN$_{df}$ falls behind by a larger gap. This proves the importance of aligning temporal features for MSVDA. Meanwhile, with effective temporal features constructed, TAMAN$_{df}$ still outperforms most image-based MSDA approaches, justifying the significance of effective temporal feature construction in tackling MSVDA.

5) Feature Visualization: We further plot the t-SNE embeddings [62] of the features learned by the TRN, c-DANN, M3SDA, and TAMAN for the Sports→U task with class information in the target domain as shown in Fig. 5. It can be observed that the features learned by TAMAN are much more clustered. This justifies the effectiveness of the features extracted by TAMAN with local attention, which possesses higher discriminability. On the contrary, the features learned by c-DANN are even less clustered compared with the TRN backbone. This suggests negative transfer where features are misaligned across the multiple source domains and are of lower discriminability. The above observations imply the superiority of our TAMAN in tackling the MSVDA.

VI. Conclusion and Future Work

In this work, we propose a novel method for tackling Multi-Source Video Domain Adaptation (MSVDA). In contrast to prior works where only spatial features are aligned, TAMAN deals with MSVDA by dynamically aligning both the spatial and temporal feature moments. TAMAN also attends to dominant domain-invariant local temporal features with high local classification confidence and low disparity between global and local feature discrepancies. We further pioneer in introducing novel MSVDA benchmarks to facilitate future MSVDA research. Our proposed TAMAN tackles MSVDA well, supported by extensive experiments and ablation studies across the proposed MSVDA benchmarks.

While the TAMAN has proved to be effective for the MSVDA, there is still room for improvements. While the TAMAN constructs global temporal features adaptively, it transfers knowledge from all source domains equally. Intuitively, with different domain shifts between different source domains and the target domain, knowledge transferred from the different source domains should be combined adaptively, focusing on the more relevant domain. Such an idea is relevant to adaptive transfer learning [63] where relevant samples are selected for better adaptation. Meanwhile, With the increasing emphasis on data privacy, methods that require source data access could raise serious privacy issues. Source-Free Video Domain Adaptation [64], [65] have been proposed to cope with privacy concerns which can be further extended to MSVDA.

REFERENCES

[1] T. V. Nguyen, Z. Song, and S. Yan, “STAP: Spatial-temporal attention-aware pooling for action recognition,” IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 1, pp. 77–86, Jan. 2015.
[2] 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, Jan. 2021.
[3] L. Liu, L. Shao, X. Li, and K. Lu, “Learning spatio-temporal representations for action recognition: A genetic programming approach,” IEEE Trans. Cybern., vol. 46, no. 1, pp. 158–170, Jan. 2016.
[4] Y. Zhang, T. Liu, M. Long, and M. Jordan, “Bridging theory and algorithm for domain adaptation,” in Proc. 36th Int. Conf. Mach. Learn., 2019, pp. 7404–7413.
[5] 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, Dec. 2020.
[6] Z. Gao, Y. Zhao, H. Zhang, D. Chen, A.-A. Liu, and S. Chen, “A novel multiple-view adversarial learning network for unsupervised domain adaptation action recognition,” IEEE Trans. Cybern., vol. 52, no. 12, pp. 13197–13211, Dec. 2022.
[7] Y. Kim and S. Hong, “Adaptive graph adversarial networks for partial domain adaptation,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 1, pp. 172–182, Jan. 2022.
[8] K. Soonroo, A. R. Zamir, and M. Shah, “UCF101: A dataset of 101 human actions classes from videos in the wild,” 2012, arXiv:1212.0402.
[9] 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.
[10] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, “ARID: A new dataset for recognizing action in the dark,” in Proc. Int. Workshop Deep Learn. Hum. Activity Recognit. Singapore: Springer, 2021, pp. 70–84.
[11] W. Kay et al., “The kinetics human action video dataset,” 2017, arXiv:1705.06950.
[12] M. Monfort et al., “Moments in time dataset: One million videos for event understanding,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 502–508, Feb. 2020.
[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1725–1732.
[14] 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. Cham, Switzerland: Springer, 2016, pp. 597–613.
[15] I.-H. Huo, D. Liu, D. T. Lee, and S.-F. Chang, “Robust visual domain adaptation with low-rank reconstruction,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 2168–2175.
[16] R. Aljundi and T. Tuytelaars, “Lightweight unsupervised domain adaptation by convolutional filter reconstruction,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 508–515.
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 was a receiver of the Nanyang President’s Graduate Scholarship. His research focuses on video understanding and analysis based on deep learning and transfer learning. He was the Co-Organizer of the UG2+ Challenge for Computational Photography and Visual Recognition, held in conjunction with CVPR 2021 and CVPR 2022. He is currently a Research Scientist with the Institute for Infocomm Research, A*STAR, Singapore, and a Lecturer with NTU.

Jianfei Yang (Member, IEEE) received the B.Eng. degree from the School of Data and Computer Science, Sun Yat-sen University, in 2016, and the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2021. He used to work as a Senior Research Engineer at the University of California, Berkeley. He is currently a Presidential Post-Doctoral Research Fellow and an Independent Principal Investigator with NTU. His research focuses on Artificial Intelligence of Things (AIoT), such as wireless sensing and computer vision based on deep learning and transfer learning. He won many International AI challenges in computer vision and interdisciplinary research fields. He received the Best Ph.D. Thesis Award from NTU.

Haozhi Cao received the B.Eng. degree from the School of Electrical Engineering and Automation, Wuhan University, 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. 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.

Keyu Wu received the B.Eng. degree in bioengineering from the National University of Singapore, Singapore, in 2013, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2020. She is currently a Scientist with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. She has won the CVPR UG2+ Challenge in 2021. Her current research interests include reinforcement learning, transfer learning, self-supervised learning, and related applications.

Min Wu (Senior Member, IEEE) received the B.S. degree in computer science from the University of Science and Technology of China (USTC) in 2006, and the Ph.D. degree in computer science from Nanyang Technological University (NTU), Singapore, in 2011. He is currently a Senior Scientist with the Machine Intelligence Department, Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. He received the best paper awards in InCoB 2016 and DASFAA 2015, and the finalist Academic Paper Award in IEEE PHM 2020. He also won the CVPR UG2+ Challenge in 2021 and the IJCAI competition on repeated buyers prediction in 2015. His current research interests include machine learning, data mining, and bioinformatics.

Zhenggua Li (Fellow, IEEE) received the B.Sci. degree in applied mathematics and the M.Eng. degree in automatic control from Northeastern University, Shenyang, China, in 1992 and 1995, respectively, and the Ph.D. degree in automatic control from Nanyang Technological University, Singapore, in 2001. He is currently with the Agency for Science, Technology and Research, Singapore. He has coauthored one monograph, more than 200 journal/conference papers including more than 60 IEEE TRANSACTIONS, and eleven granted USA patents, including normative technologies on scalable extension of H.264/AVC and HEVC. His research interests include video processing and delivery, computational photography, switched and impulsive control, sensor fusion, and physics-driven deep learning. He has been actively involved in the development of H.264/AVC and HEVC since 2002. He had three informative proposals adopted by the H.264/AVC and three normative proposals adopted by the HEVC.

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 Laboratory Head with 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 data-efficient and model-efficient learning with related applications in smart city and smart manufacturing. He has won several competitive awards, such as the First Place Winner for CVPR 2021 UG2+ Challenge, 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 CPHM 2020. He serves as an Associate Editor for Neurocomputing (Elsevier) and IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT. He is currently the Vice Chair of IEEE Sensors Council Singapore Chapter.