Multi-View Multi-Human Association With Deep Assignment Network

Ruize Han, Yun Wang, Haomin Yan, Wei Feng, *Member, IEEE*, and Song Wang, *Senior Member, IEEE*

***Abstract*— Identifying the same persons across different views plays an important role in many vision applications. In this paper, we study this important problem, denoted as Multi-view Multi-Human Association (MvMHA), on multi-view images that are taken by different cameras at the same time. Different from previous works on human association across two views, this paper is focused on more general and challenging scenarios of more than two views, and none of these views are fixed or priorly known. In addition, each involved person may be present in all the views or only a subset of views, which are also not priorly known. We develop a new end-to-end deep-network based framework to address this problem. First, we use an appearance-based deep network to extract the feature of each detected subject on each image. We then compute pairwise-similarity scores between all the detected subjects and construct a comprehensive affinity matrix. Finally, we propose a Deep Assignment Network (DAN) to transform the affinity matrix into an assignment matrix, which provides a binary assignment result for MvMHA. We build both a synthetic dataset and a real image dataset to verify the effectiveness of the proposed method. We also test the trained network on other three public datasets, resulting in very good cross-domain performance.**

***Index Terms*— Human association, multi-view association, wearable cameras, maximum multi-clique problem.**

I. Introduction

MULTIPLE cameras can simultaneously take images/videos of the same scene from different views, which provide complementary information for many important vision tasks, such as video surveillance. One important problem in this setting is to identify the same persons across the images taken by different-view cameras, which we refer to as Multi-view Multi-Human Association (MvMHA) in this paper. Previous works usually use a network of fixed cameras, whose views can be estimated and calibrated in advance for multi-view correspondence. However, the fixed cameras suffer from the problem of limited coverage and pre-specified view angles. In this paper, we focus on MvMHA across images/videos taken by multiple wearable (moving) cameras, such as phone cameras, GoPro, Google Glass, etc. [1]–[4], as shown in Fig. 1. The proposed MvMHA problem has many practical applications in real world. A typical example is video surveillance – in an outdoor scenario without pre-installed cameras, we can associate and analyze the videos taken by the cameras worn by several law enforcement officials for collaborative tracking, human activity recognition, important/abnormal person detection, etc. This can provide multi-perspective all-round information for video surveillance. *For all the above analysis, the first step is to associate the humans across the multiple views (wearable cameras).*

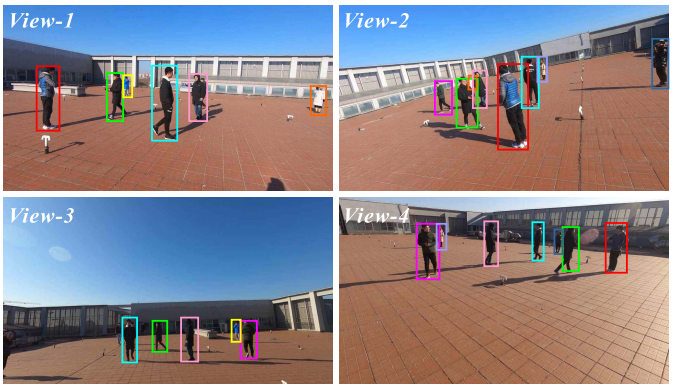

Fig. 1. An illustration of the proposed MvMHA problem. Four views are shown here and subjects bounded by the same colour box in different views represent the same person.

As in prior works [2], [5], we assume the images/videos taken by multiple wearable cameras are temporally synchronized, i.e., the corresponding images/frames across different views are taken at (roughly) the same time. This can be achieved by synchronizing clocks in these cameras. While MvMHA can be treated as a person re-identification (re-id) problem from a general perspective – for each subject detected in one view, re-identifying him/her in the other views, it brings

Manuscript received May 13, 2021; revised October 29, 2021 and December 10, 2021; accepted December 10, 2021. Date of publication January 26, 2022; date of current version February 15, 2022. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U1803264 and Grant 62072334 and in part by the Natural Science Foundation of Tianjin under Grant 18JCYBJC15200 and in part by the Tianjin Research Innovation Project for Postgraduate Students under Grant 2021YJSB174. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Raymond Fu. *(Ruize Han and Yun Wang contributed equally to this work.) (Corresponding author: Song Wang.)*

Ruize Han, Yun Wang, Haomin Yan, and Wei Feng are with the School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China, also with the Key Research Center for Surface Monitoring and Analysis of Cultural Relics (SMARC), State Administration of Cultural Heritage, Tianjin 300350, China, and also with the Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin 300350, China (e-mail: han_ruize@tju.edu.cn; 2019216099@tju.edu.cn; yan_hm@tju.edu.cn; wfeng@tju.edu.cn).

Song Wang is with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208 USA (e-mail: songwang@cec.sc.edu).

Digital Object Identifier 10.1109/TIP.2021.3139178

in new challenges such as unknown and significant view differences, illumination differences, mutual occlusions, and background clutters [5], as shown in Fig. 1. More importantly, the existing re-id methods only measure and threshold the similarity between two subjects for pairwise matching, without a global consideration of multiple human association across $C > 2$ views.

Previous works on multi-view human association are usually focused on the simplest case of two views. Many of them either adopt the person re-id like approach by detecting and matching the appearance/motion features of individual subjects in pairwise way [2], [5], or use pairwise similarity to build an affinity matrix between multiple detected subjects across two views. For the latter, a pairwise assignment method, e.g., Hungarian algorithm [6], Deep Hungarian Network (DHN) [7], or Deep Graph Matching Network [4], [8], [9], can be applied to achieve the human association results between the two views. Such methods may not handle well the more general and challenging scenarios of $C > 2$ views without considering the cyclic consistency among all the views. Several recent works try to solve $C > 2$ multi-view association by joint optimization, e.g., Constrained Binary Integer Programming (BIP) [10]–[12], Spectral Clustering [13], and Alternating Direction Method of Multipliers (ADMM) [14]. However, these methods are not end-to-end ones and their performance is highly dependent on the tuning of many hyper-parameters.

In this paper, we model MvMHA for $C > 2$ views as a constrained optimization problem and develop a new end-to-end framework to address it. The proposed framework consists of two components: affinity matrix construction and multi-view multi-clique assignment. For the former, we use an affinity network to compute the similarity between any two subjects detected in any two different views. For the latter, we propose a new Deep Assignment Network (DAN) by modeling the multi-view constraint conditions as the unsupervised training losses. In particular, we adopt four specific constraints for the MvMHA problem, and then represent the constraints as *differentiable loss functions* with theoretical exploration. With that, we propose to combine the image feature extraction, affinity matrix calculation together with the assignment optimization *in an end-to-end framework for joint training*. We build a synthetic dataset and a real dataset collected by the wearable cameras for the training and testing of the whole framework. To better evaluate the proposed method, we further test our method on three public datasets. Extensive experiments on all the datasets verify the effectiveness of the proposed method. The main contributions of this paper are:

- We propose a constrained optimization model for MvMHA and an end-to-end framework to solve the optimization problem. To the best of our knowledge, this is *the first work to model such multi-view assignment problem* using a differentiable network with constraint losses.
- We propose a Deep Assignment Network (DAN) to model the *constrained multi-view multi-clique assignment problem*, which is implemented by two popular backbone networks. On both of them, we verify the effectiveness of the proposed unsupervised constraint loss.
- We build two datasets for the training and testing of MvMHA. Extensive experiments on our datasets and other three public datasets verify the effectiveness of our method. The datasets and code are released to the public at https://github.com/RuizeHan/DAN4Ass.

II. Related Work

Person re-identification (re-id), has been widely studied, resulting in many traditional and deep-learning algorithms, especially on extracting or learning discriminative visual representations that are robust to the change of views, image resolution, illumination and/or background. For deep-learning based re-id, in [15] a data augmentation approach is developed to smooth the camera style disparities and enhance the learning capability of features. In [16], a two-stream network is used to aggregate local appearance similarities for person re-id. In [17], GANs are used to generate high-quality cross-id composed images to augment the insufficient training data. More works can be found in a recent survey for re-id [18]. In our work, we will first use a re-id network to estimate the pairwise similarity between subjects, with which we further develop an assignment network to achieve MvMHA across $C > 2$ views.

Cross-view person identification (CVPI) across two wearable cameras was first introduced in [19], in which a deep-learning network is designed to learn view-invariant motion features from the two videos. These motion features are then combined with the appearance features to decide whether the subjects shown in these two videos are the same person or not. Later, Liang *et al.* [2] introduce a new metric of confidence for each joint in 3D pose estimation and show that such 3D pose features can be combined with motion and appearance features for improving the CVPI performance. It also studies the effectiveness of several different feature-combination strategies. As in person re-id, CVPI aims to match a pair of subjects, each from a different video, while the proposed MvMHA in this paper aims to identify multiple subjects across $C > 2$ views.

Multi-view multi-object association tries to match the objects detected from different views using a multi-camera system and most of its existing works focus on pairwise correspondences. Hungarian algorithm [6] is widely used for finding such an optimal pairwise correspondence by solving the underlying allocation problems in polynomial time. Deep Hungarian Network (DHN) [7] models the Hungarian algorithm in an differentiable deep network layer. Deep Graph Matching Network [4], [8], [9] uses an end-to-end model for the graph matching underlying the pairwise correspondence. In general, without an integrative consideration of the multi-view property, such pairwise correspondence can not handle well the cases of $C > 2$ views.

There are several related works that solve the multi-view association by joint optimization over all the views. A self-supervised feature descriptor is proposed for multi-view person association [20]. Various methods have been developed for modeling the relations of multiple objects across multiple views, such as a Constrained Binary Integer Programming (BIP) in [10]–[12], the random-walk algorithm

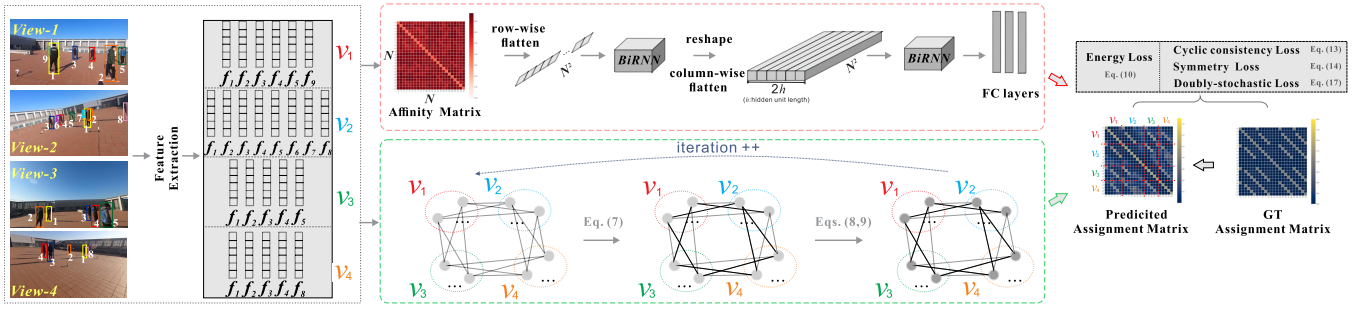

Fig. 2. An illustration of the proposed framework. Given bounding-box detections on C synchronous images taken by different cameras, an off-the-shelf person re-identification network is used to extract features $\boldsymbol{f}_v^i$ for the subject v in view i, followed by a deep assignment network (DAN) to compute the assignment matrix as the MvMHA result. We use either RNN (top) or GNN (bottom) as the network architecture for the proposed DAN. Specifically, in the RNN based model, the input affinity matrix is first flattened into a vector with length N^2 flowing the row-wise order and fed to the BiRNN module. The output of the first BiRNN is then reshaped as $N \times N \times 2h$ and each channel is flattened to a vector with the length of N^2 by following the column-wise order and next fed to the second BiRNN. Finally, an FC layer is used to generate the predicted assignment matrix. In the GNN based model, the extracted features are used as the node features of the GNN, which iteratively performs the affinity matrix update (note the change of edge thickness) using Eq. (7), and the graph node/edge feature update (note the change of node gray-levels) using Eqs. (8) and (9), respectively. In the end, we apply the proposed loss functions for end-to-end training of the whole framework.

in [21], the loop constraints in [22], the multi-graph matching in [23]–[25], the cycle-consistency constraint in [26], [27], and quadratic assignment problem (QAP) in [28], [29]. Such iterative algorithms for MvMHA may involve several hyper-parameters to be tuned and cannot be implemented as an end-to-end trainable models. Various constraints have been used in these methods, which, however, calculate the affinity matrix independently and then implement a combinatorial optimization method to solve the problem. We hope to explore a new data-driven and learning-guided method in this direction. The proposed work is the *first attempt to jointly model the feature extraction, affinity matrix calculation and the MvMHA optimization into an end-to-end framework*. It is different from 1) previous deep models focusing on the case of *pairwise assignment* problem, e.g., DHN, and 2) the *constraint optimization* problem for multi-view cases solved by the classical algorithms, e.g., BIP method.

III. The Proposed Method

A. Problem Formulation

1) Pairwise Association: We first briefly review the pairwise association across two views. Suppose there are two cameras in the scene and they produce a pair of synchronous images. The pairwise affinity scores between these two views, i.e., i and j, can be represented by $\mathbf{A}_{ij} \in \mathbb{R}^{n_i \times n_j}$, where n_i and n_j denote the number of detected subjects (bounding boxes) in views i and j, respectively. The association between these two views can be estimated using a pairwise assignment matrix $\mathbf{P}_{ij} = \{0,1\}^{n_i \times n_j}$. Deriving the assignment matrix from the affinity matrix is a linear assignment problem. It can be solved by Hungarian algorithm to maximize the matrix inner product $\langle \mathbf{P}_{ij}, \mathbf{A}_{ij} \rangle$, which is denoted as a similarity metric between the two matrices, for predicting the assignment matrix $\hat{\mathbf{P}}_{ij}$.

2) Multi-View Association: In our problem, suppose there are $C > 2$ cameras in the scene, leading to C synchronous images from different views as shown in Fig. 2. We aim to predict the assignment matrix $\mathbf{P}$ for all views

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \dots & \mathbf{P}_{1C} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \dots & \mathbf{P}_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{C1} & \mathbf{P}_{C2} & \dots & \mathbf{P}_{CC} \end{pmatrix}, \tag{1}$$

where $\mathbf{P}_{ij}, i, j = 1, 2, \dots, C$ are the pairwise assignment matrices, and $\mathbf{P} \in \{0,1\}^{N \times N}$ is the assignment matrix among all $N = \sum_{i=1}^{C} n_i$ detected subjects (bounding boxes) in all C views. For MvMHA, we take $\mathbf{A} \in [0,1]^{N \times N}$ as input and output the optimal assignment matrix $\mathbf{P}$, where we denote $\mathbf{A}$ to be an $N \times N$ matrix by concatenating the pairwise affinity matrices $\mathbf{A}_{ij}$ in the same order as their corresponding $\mathbf{P}_{ij}$, $i, j = 1, 2, \dots, C$ in $\mathbf{P}$. We propose to maximize the following energy function for MvMHA

$$\begin{aligned} \hat{\mathbf{P}} &= \arg\max_{\mathbf{P}} \ \langle \mathbf{P}, \mathbf{A} \rangle, \\ &\text{s.t. } \mathbf{P} \in \mathcal{S} \end{aligned} \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and the energy function encodes the correspondence of the assignment matrix and the affinity matrix. Different from simply stacking the pair-wise information together, the multi-view association is subject to several constraints $\mathcal{S}$, which encodes the structural compatibilities for the desired association matrix $\mathbf{P}$. After calculating the pair-wise information $\mathbf{A}_{ij}$, these constraints reflect the *global structural properties among all the views*. For example, we require that all the concatenations among the subjects in different views to form a closed loop, as shown in Fig. 3. This way, we consider to use auxiliary loss to effectively penalize the red-line concatenations. In the following, we discuss these constraints for the global structural properties in detail.

Constraint 1 (Closed-Loop Constraint): The same person appearing in different views should be connected as a closed loop, as shown in Fig. 3. Following the previous work [30], we assume $\mathcal{U}$ as the set of unique people in the scene (under all C views). For each view pair i and j, we have $\mathbf{P}_{ij} = \mathbf{P}_i^u \, \mathbf{P}_j^u$,

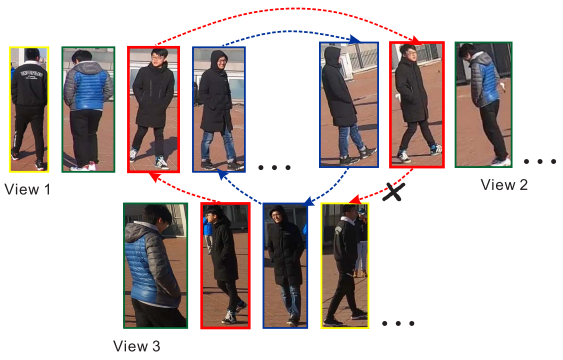

Fig. 3. An illustration of the closed-loop constraint, which requires that all the concatenations among the subjects in different views form a closed loop. Specifically, the blue concatenation lines follow this constraint while the red ones do not. This way, the loss inferred from the closed-loop constraint can effectively penalize the red-line concatenations.

where $\mathbf{P}_i^u \in \mathbb{R}^{n_i \times |\mathcal{U}|}$ denotes the assignment matrix between the n_i subjects in view i and $\mathcal{U}$. We concatenate all the $\mathbf{P}_i^u$ row by row and get $\mathbf{P}_u \in \mathbb{R}^{N \times |\mathcal{U}|}$, the cyclic-consistency constraint requires that

$$\mathbf{P} = \mathbf{P}_u \mathbf{P}_u^T, \tag{3}$$

which implies that all the concatenations among the subjects in different views form $|\mathcal{U}|$ closed loops. Note, the number of elements in each loop is no more than C.

Constraint 2 (Symmetric Constraint): MvMHA result is commutative and we impose symmetric constraints to the matrix $\mathbf{P}$

$$\begin{aligned} \mathbf{P}_{ij} &= \mathbf{P}_{ji}^{\mathrm{T}}, \quad 1 \le i,\ j \le C,\ i \ne j, \\ \mathbf{P}_{ii} &= \mathbf{I}_{n_i \times n_i}, \quad 1 \le i \le C, \end{aligned} \tag{4}$$

where $\mathbf{I}$ is the identity matrix with the dimension of the number of subjects n_i in view i.

Constraint 3 (Doubly-Stochastic Constraint): One person can appear in one view at most once. As a result, each row/column of matrix $\mathbf{P}_{ij}$ can have at most one element value of 1, and the other elements must be 0. This can be formulated as

$$\mathbf{0} \le \mathbf{P}_{ij}\mathbf{1} \le \mathbf{1}, \quad \mathbf{0} \le \mathbf{P}_{ij}^T\mathbf{1} \le \mathbf{1}, \tag{5}$$

where $\mathbf{P}_{ij} \in \mathbb{R}^{n_i \times n_j}$ and $\mathbf{0}, \mathbf{1}$ denote an all-0 or all-1 vector with dimension matched to its multiplied matrix or the corresponding result vector.

Constraint 4 (Element Constraint): Each of the assignment matrix $\mathbf{P}$ takes the value of either 1 or 0, which reflects the matching between two subjects or not, i.e.,

$$p_{mn} \in \{0, 1\}, \quad 1 \le m,\ n \le N, \tag{6}$$

where p_{mn} denotes the mth-row nth-column element of the matrix $\mathbf{P}$.

B. Deep Assignment Network

In the following, we elaborate on a deep assignment network (DAN) to solve this constrained optimization problem, which takes the appearance features of the subjects as input and generates the assignment matrix $\mathbf{P}$ as the output. The appearance features of the subjects are computed by the affinity network, which will be introduced in Section III-D in detail. In the above formulation, we may have $C > 2$ views and the number of detected subjects in each view, i.e., n_i, and the total number of the detected subjects in all the views, i.e., N, are not fixed values. Therefore, the proposed deep assignment network (DAN) has to be able to handle the input features from uncertain number of subjects. For this purpose, we consider two kinds of networks – recurrent neural networks (RNN) and graph neural networks (GNN) – for the proposed DAN, respectively, both of which have been successfully used in previous work for pairwise-view subject matching [7], [8]. As illustrated in Fig. 2, we use either RNN (top) or GNN (bottom) as the network architecture for DAN.

RNN Based Framework: Inspired by [31], we use bi-directional RNN network. We compute the similarity between the features generated by the affinity network and construct the affinity matrix $\mathbf{A}$. As shown in the top stream of Fig. 2, the input affinity matrix $\mathbf{A}$ is first flattened to a vector by following the row-wise order and fed to the first BiRNN. The output of the first BiRNN is then reshaped and flattened to a vector by following the column-wise order and fed to the second BiRNN. Note that, the weights of the two BiRNN are not shared. We can see that there are two 'flatten' procedures in the the RNN module. This is inspired by the original Hungarian algorithm that alternately performs the row-wise and column-wise subtracting operations. Specifically, given an affinity matrix using the Hungarian algorithm for assigning task, first, all elements in each row subtracts the row minimum, and then, we similarly subtract the column minimum from each column. The operation stops until an optimal assignment exists among the zeros in the matrix. Here the 'flatten' procedures are used for simulating the above alternate row-wise and column-wise operations. Later, three fully-connected (FC) layers are applied, followed by a sigmoid function to achieve the final assignment matrix.

GNN Based Framework: We first construct a non-fully-connected undirected graph taking all the subjects as graph nodes. Specifically, there is no edge between two nodes representing the subjects appearing in the same views, since two subjects in the same view can not be the same person. All other node pairs are connected by an edge. The feature of each subject generated by the affinity network is used as the node feature and the edge feature is initialized by the distance between the involved node features, i.e., the absolute value of difference between two node feature vectors. The adjacency matrix is initialized by the affinity matrix $\mathbf{A}$. Similar to the previous GNN [4], [32], we iteratively update the graph, including the node, edge and affinity matrix. The final output of the affinity matrix is taken as the desired assignment matrix. The proposed GNN based framework contains three main phases – affinity matrix update, graph node update and graph edge update.

1) Affinity Matrix Update: Given a graph $\mathcal{V}$, the GNN updates the adjacency matrix $\mathbf{A}$ to infer the current affinity relation among different nodes, according to the node and edge

features, by

$$a_{\upsilon,w}^{(k)} = \sigma(\mathrm{F_A}(\mathbf{x}_{\upsilon}^{(k-1)}, \mathbf{x}_{w}^{(k-1)}, \mathbf{x}_{\upsilon,w}^{(k-1)})), \quad \upsilon, w \in \mathcal{V}, \tag{7}$$

where $\mathbf{x}_{\upsilon}$ denotes the feature vector of node υ and $\mathbf{x}_{\upsilon,w}$ denotes the feature vector for edge $e = (\upsilon, w)$. The affinity matrix $\mathbf{A}^{(k)} = [a_{\upsilon,w}^{(k)}] \in \mathbb{R}^{N\times N}$ encodes current (k-th iteration) connection relation predictions. $\mathrm{F_A}$ is a connectivity readout network that maps an edge representation into the affinity weight and σ is an activation function.

2) Graph Node Update: We update node representations $\mathbf{x}_{\upsilon}$ via considering all the incoming node and edge information weighted by the corresponding connectivity

$$\mathbf{x}_{\upsilon}^{(k)} = \sigma(\sum_{w} a_{\upsilon,w}^{(k-1)} \mathrm{F_V}(\mathbf{x}_{\upsilon}^{(k-1)}, \mathbf{x}_{w}^{(k-1)}, \mathbf{x}_{\upsilon,w}^{(k-1)})), \tag{8}$$

where $\mathrm{F_V}$ represents a node update network. We set $\mathbf{x}_{\upsilon}^{(0)}$ as the initial node representation, which is from the feature extraction by the pre-trained re-id network.

3) Graph Edge Update: We compute the distance between two node features and get the corresponding edge feature representations

$$\mathbf{x}_{\upsilon,w}^{(k)} = \mathrm{abs}(\mathbf{x}_{\upsilon}^{(k)} - \mathbf{x}_{w}^{(k)}), \quad \upsilon, w \in \mathcal{V}, \tag{9}$$

where abs denotes the operation to take the absolute value.

We iteratively update the adjacency matrix and node and edge representations for K iterations then obtain the final output of the affinity matrix, i.e., $\mathbf{A}^{(K)}$, which is taken as the desired assignment matrix. The adjacency matrix update functions $\mathrm{F_A}$ in Eq. (7) is implemented by a four-layer convolution network. The node update function $\mathrm{F_V}$ in Eq. (8) is implemented by the fully-connected-layer and gated recurrent unit (GRU) network. Besides, the activity function σ in Eq. (7) and (8) are sigmoid function.

Supervised Loss: We next discuss the loss for the proposed DAN. First, we define the supervised loss using the ground-truth assignment matrix as

$$\mathcal{L}_{\mathrm{E}} = \begin{cases} -\alpha(1-\hat{p}_{mn})^{\gamma}\log(\hat{p}_{mn}) & \text{if } \tilde{p}_{mn} = 1, \\ -(1-\alpha)(\hat{p}_{mn})^{\gamma}\log(1-\hat{p}_{mn}) & \text{if } \tilde{p}_{mn} = 0, \end{cases} \tag{10}$$

where $\tilde{p}_{mn}$ and $\hat{p}_{mn}$ represent the element of the ground-truth assignment matrix and the corresponding output of DAN, respectively. Here we adopt the focal loss [33] because the the number of positive and negative training samples are imbalanced. α and γ are two pre-defined parameters.

C. Unsupervised Constraint Loss

The above RNN/GNN based framework just considers the transformation between the affinity matrix $\mathbf{A}$ and the assignment matrix $\mathbf{P}$, i.e., the energy function defined in Eq. (2). However, the structural compatibilities $\mathcal{S}$ of $\mathbf{P}$ are not considered. Next we discuss the constraints defined in Eqs. (3-6). Note that, in the above section III-A, we have *theoretically* defined the constraints in a *rigorous formulation*. Differently, in this section, we represent the above constraints as *differentiable loss functions* as in Eqs. (13, 14, 17), which can be used for the end-to-end network training.

1) Closed-Loop Loss: From Eq. (3), the closed-loop constraint requires the assignment matrix $\mathbf{P} \in \mathbb{R}^{N\times N}$ can be factorized as $\mathbf{P}_u\mathbf{P}_u^T$ with $\mathbf{P}_u \in \mathbb{R}^{N\times|\mathcal{U}|}$. Following the theory of matrices [34], the above constraint $\mathbf{P} = \mathbf{P}_u\mathbf{P}_u^T$ needs $\mathbf{P}$ to satisfy that

$$\mathbf{P} \succeq 0, \quad \mathrm{rank}(\mathbf{P}) \leq |\mathcal{U}|, \tag{11}$$

i.e., $\mathbf{P}$ is a *positive-semidefinite* and *low-rank* matrix, where $|\mathcal{U}|$ denotes the number of unique people in the scene and N is the total number of detections (bounding boxes) in C views.

Note that, the constraint in Eq. (11) is implicit and non-differentiable, and the number of persons in the scene, i.e., $|\mathcal{U}|$, is unknown in advance. Next, we discuss how to transform this constraint to a differentiable loss function. Specifically, we use the nuclear norm $\|\mathbf{P}\|_*$ (sum of singular values) to approximate the above constraints of the matrix.

Inference: Given the real symmetric matrix $\mathbf{P}$, let $\mathbf{e} = (e_1, e_2, \ldots, e_N)$ denote the eigenvalues of $\mathbf{P}$, and the singular value vector of it can be represented as $\mathbf{e}^+$, i.e., the absolute value of $\mathbf{e}$. It is not hard to get

$$\|\mathbf{P}\|_* = \|\mathbf{e}^+\|_1, \quad \mathrm{rank}(\mathbf{P}) = \|\mathbf{e}^+\|_0, \tag{12}$$

which represent the sum of singular values and the number of non-zero singular values, respectively. We make following inference: 1) With $\|\mathbf{P}\|_* = \sum_{i=1}^{N}|e_i| \geq \sum_{i=1}^{N} e_i$, minimizing the nuclear norm $\|\mathbf{P}\|_*$ drives $|e_i|$ to approximate e_i, iff. $e_i \geq 0$, i.e., $\mathbf{P}$ is positive semidefinite. 2) From Eq. (12), we found that the low-rank constraint is equivalent to minimize the l_0 norm, which is a notoriously difficult problem. Therefore, we use the optimal convex approximation of l_0 norm, i.e., the l_1 norm to replace it. This way, decreasing $\|\mathbf{P}\|_*$ also compels the low-rank constraint. As inferred above, the positive-semidefinite and low-rank properties are defined as minimizing the following closed-loop loss

$$\mathcal{L}_{\mathrm{C}} = \|\mathbf{P}\|_*. \tag{13}$$

2) Symmetric Loss: For Constraint 2 in Eq. (4), we can use the matrix norm as a loss to reflect the matrix symmetry, i.e.,

$$\mathcal{L}_{\mathrm{S}} = \|\mathbf{P} - \mathbf{P}^T\|_2. \tag{14}$$

3) Doubly-Stochastic Loss: For Constraint 3 in Eq. (5), it restrains the row sum and column sum of pairwise assignment matrix $\mathbf{P}_{ij}$ to be within [0, 1]. Take the row sum of each matrix, referred to as r, for example, we can use the following loss function to achieve the above constraint

$$\mathrm{L}(r) = \begin{cases} 0 & 0 \leq r \leq 1, \\ l & \text{otherwise}, \end{cases} \tag{15}$$

where $l > 0$ is a pre-set parameter. However, such a function is non-differentiable. Therefore, we approximate it by a differentiable function

$$\tilde{\mathrm{L}}(r) = \begin{cases} \frac{1}{2}r^2 & r < 0, \\ 0 & 0 \leq r \leq 1, \\ \frac{1}{2}(r-1)^2 & r > 1. \end{cases} \tag{16}$$

This way, Constraint 3 can be described by a loss

$$\mathcal{L}_{\mathrm{D}} = \sum_{i,j=1}^{C} \tilde{\mathrm{L}}(r_{ij}) + \sum_{i,j=1}^{C} \tilde{\mathrm{L}}(c_{ij}), \tag{17}$$

where C is the number of views, r_{ij} and c_{ij} denote the row/column sum of $\mathbf{P}_{ij}$.

4) Element constraint: requires the binarization of each element in $\mathbf{P}$. This is not differentiable in the neural network inference. For the convenience of optimization, we relax it to a real value in the range [0, 1] as

$$0 \leq p_{mn} \leq 1, \tag{18}$$

and make use of the sigmoid function to ensure that the output takes values in the range of [0, 1].

Discussion. The above loss functions, in part, are *the necessary conditions (but not the sufficient conditions)* of the corresponding constraints. We relax the constraints to make some of them to be modeled as differentiable components. One example is the low-rank constraint in Eq. (11), from which we can derive the differentiable loss in Eq. (13). This way, the descent of such loss, as the necessary condition, can compel the results to satisfy the corresponding properties. We clarify that such loss function is the relaxed condition but not the strict guarantee (namely sufficient conditions) for the constraint to be always satisfied. Even so, the proposed unsupervised losses are effective for constraining the structure of the assignment matrix. We will show the effectiveness of all the proposed losses in the experiments. Finally, we define the total loss as

$$\mathcal{L} = \mathcal{L}_{\mathrm{E}} + \mathcal{L}_{\mathrm{C}} + \mathcal{L}_{\mathrm{S}} + \mathcal{L}_{\mathrm{D}}, \tag{19}$$

where we adaptively tune the loss weights as in [35] during training. This way, the multiple loss weights in our problem can be adaptively tuned thereby avoiding simple linear weighting, which also alleviates the model sensibility caused by the manual tuning of parameters.

D. The Framework

1) End-to-End Training: Given the input set of subjects detected on C synchronous images taken by different cameras, we finally put together an end-to-end framework for MvMHA as shown in Fig. 2. Following the setting in previous work [14], we first use an off-the-shelf person re-identification network to extract features of each detected subject. We adopt the CamStyle [15] trained on the Market-1501 dataset [36] as the pre-trained person re-id network to extract features of each subject detected in the form of a bounding box. Specifically, we feed the cropped bounding boxes of each detected subject in each view to the CamStyle network, and obtain the descriptor for each bounding box from the 'pool5' layer. We then calculate the Euclidean distance between the feature vectors and apply a sigmoid function to map the distances to values in $[0, 1]$ as the appearance affinity score between the corresponding pair of detected subjects. This way, we obtain the global affinity matrix $\mathbf{A}$. After that the deep assignment network (DAN) described above is used to get the assignment matrix $\mathbf{P}$ over all the views as the MvMHA result. Besides the respective training of the affinity network and DAN, since all the modules are differentiable, we can jointly train both of them in an end-to-end manner.

2) Implementation Details: In our experiments, for the output of the proposed network, we use a threshold, i.e., 0.5, to convert the obtained assignment matrix $\mathbf{P}$ to a binary matrix as the MvMHA result. In Eq. (15), the parameter l is set to 2 to reduce the impact of simple samples and α in Eq. (10) is set empirically according to the ratio of positive and negative samples. We implement our method based on the PyTorch framework and use the optimizer of stochastic gradient descent to optimize the network parameters. Our model is implemented on an NVIDIA GTX-2080Ti GPU. For the proposed RNN based network, the training process takes 20 epochs with a learning rate of 1×10^{-5}. For the GNN based network, the learning rate is set as 1×10^{-3} and the training process takes 70 epochs.

IV. Experiments

A. Dataset and Evaluation Metrics

1) Synthetic Matrix Dataset (SMD): We build a synthetic dataset to simulate the problem of MvMHA by constructing the corresponding assignment matrix and affinity matrix. We first generate the synthetic assignment matrix $\mathbf{P}$ which satisfies the constraints of the MvMHA problem. Then, we construct the corresponding affinity matrix $\mathbf{A}$ by adding Gaussian-distribution noise $\mathcal{N}(0, \sigma^2)$ to each $\mathbf{P}$, where the variance σ is set to 0.5. The noise is added to each upper triangular element of the assignment matrix independently and then copied to the corresponding lower triangular elements to ensure the symmetry of the constructed affinity matrix. We set the number of views C to be in the range of 3 to 6, and the number of detected bounding boxes in each view to be in the range of 2 to 20. For evaluation, we generate totally $2,400$ pairs of $\mathbf{P}$ and $\mathbf{A}$ with 600 pairs for each value of C. For training, we generate additional 700 samples with no overlap of our testing dataset.

2) Synthetic MvMHA Image Dataset (MvMHA-S): Adopting the famous 3D modeling engine Unity and the open source toolkit PersonX [37], we build a synthetic image dataset to simulate the multi-view multi-human scenes. We chose a city scenario similar to the real-world environment as the background of the dataset. In each image of the dataset, ten different subjects are randomly placed in the scene, and four moving cameras with overlapped area coverage are randomly placed on the periphery. Under the simulation environment, we can accurately obtain the label and bounding box of each subject without manual annotation. Specifically, the same subject across all views in an image group is labeled with the same ID for the MvMHA task. We generate totally 4,000 images with 31,006 human bounding boxes.

3) Real-World MvMHA Image Dataset (MvMHA-R): We also build a new real image dataset, referred to as MvMHA-R dataset, for the MvMHA task. This dataset is collected from the videos [4] using four GoPro wearable cameras to cover an area present with multiple people from significantly different directions, e.g., near 90 degree view-angle difference.

This way, we obtain four synchronous videos, from which we extract the synchronous $1,728$ frames with a maximum number of 10 persons in each frame. We use the manually annotated bounding boxes and the cross-view human ID labels for each subject on all $1,728 \times 4 = 6,912$ images for training and testing, which contains totally 44,131 human bounding boxes.

4) Public Datasets: Besides the self-collected dataset, we have also apply our method to three public datasets including CVMHT [10], APIDIS [38] and Campus [39], which have 3, 4 and 4 views, with a maximum number of 12, 13 and 16 persons, respectively. These three datasets contain 720, 1,000 and 1,222 test images per view, respectively. Among them, CVMHT has recorded multiple actors walking around using several wearable cameras. APIDIS[1] and Campus[2] recorded the basketball game and daily campus scene, respectively, both of which are collected from the real life. Note that, these datasets only contain the testing datasets without the training data. So, we directly apply the network trained on MvMHA-R training set to these three public datasets for testing to verify the cross-domain effectiveness of our method. In testing stage, we use the bounding boxes provided by these datasets.

5) Evaluation Metrics: We use precision ($\mathcal{P}$), recall ($\mathcal{R}$), and F_1 score ($\mathcal{F}$) for evaluation. Precision and recall denote the ratio of the true-positive pairwise matches against all predicted-positive and all real-positive matches, respectively. F_1 score is computed as $\mathcal{F} = \frac{2 \times \mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}}$.

B. Baselines

• **Re-id** is a simple baseline that directly uses the person re-id network to compute the affinity matrix and take the maximum (in each row) as 1 and others as 0 to obtain the assignment result. For fair comparison, we use the re-id method [15] trained on the Market-1501 dataset [36] - the same network structure is also used for feature extraction in our method.

• **Hungarian** [6] is widely used for finding optimal pairwise correspondence by solving the underlying allocation problems. It can be used to solve the proposed MvMHA problem by matching every two views separately. Specifically, for fair comparison, we adopt the same affinity matrix as in our framework and then solve the cross-view subject assignment problem between each pair of views using the Hungarian algorithm.

• **DHN** [31] follows the same setting as Hungarian except that a deep RNN network is used to implement the Hungarian algorithm in a deep learning manner. We train the DHN on our training dataset for fair comparison.

• **Clustering** [40] denotes the spectral clustering algorithm, which takes the input of the same affinity matrix as in the proposed method. The spectral clustering algorithm makes positive correlation within the same cluster and the negative connection across different clusters. The assignment matrix $\mathbf{P}$ is constructed according to the clustering results.

[1]https://sites.uclouvain.be/ispgroup/index.php/Softwares/APIDIS
[2]http://web.cs.ucla.edu/ yuanluxu/research/mv_track.html

TABLE I
COMPARATIVE RESULTS ON SYNTHETIC MATRIX DATASET WITH DIFFERENT NUMBER OF VIEWS IN TERMS OF $\mathcal{F}$ SCORES (%)

Method	C=3	C=4	C=5	C=6
Hungarian	39.95	35.76	32.86	29.82
DHN	69.37	68.43	65.01	65.91
Clustering	48.87	59.53	63.58	70.38
Ours-R	**77.53**	**76.20**	**72.16**	**72.95**

TABLE II
COMPARATIVE RESULTS ON MVMHA-S AND MVMHA-R DATASETS

Method	MvMHA-S			MvMHA-R		
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Re-id	61.21	66.79	63.88	53.82	69.84	60.79
+ Hungarian	66.03	70.17	68.04	57.07	76.00	65.19
+ DHN	68.98	67.75	68.36	74.49	81.19	77.69
+ Clu.	45.63	70.90	55.52	56.48	64.98	60.44
+ Clu. w $\lvert\mathcal{U}\rvert$	67.13	58.43	62.48	66.68	62.04	64.27
GMN	43.10	45.65	44.34	44.06	58.79	50.36
PCA-GMN	71.07	75.29	73.12	66.68	88.91	76.20
Mv Match.	84.54	27.34	41.32	58.39	32.69	41.92
Ours-R	**86.73**	**93.77**	**90.12**	86.77	86.54	86.65
Ours-G	79.36	79.57	79.48	**89.81**	**86.70**	**88.23**

• **Deep Graph Matching.** We also include two end-to-end methods for comparison, in which both feature extraction and later assignment are achieved by combined network training. **GMN** [9] and **PCA-GMN** [4], [8] both build graph models, on which deep graph matching is applied to learn the pairwise correspondence for cross-view subject association. Both methods run 20 epochs in the training process.

• **Multi-view Matching** [14]. We select a related multi-view multi-human association method for comparison, which follows [29] and models such task as a cycle-consistency-aware multi-way matching optimization problem to cluster the detected humans in multiple views. For this method, we use the default parameter settings in [14].

C. Results

1) Association Results on SMD: We evaluate the classic assignment algorithms and the proposed RNN based DAN on the synthetic matrix dataset (SMD), which take the affinity matrix as input and output the assignment matrix. As shown in Table I, the proposed method shows better F_1 score than other three compared methods on all the subsets with different number of views. This verifies the effectiveness of the proposed deep assignment network.

2) Comparison on MvMHA Dataset: Table II shows the performance of our method and the above baseline methods on our collected MvMHA-S and MvMHA-R datasets. As shown in the first five rows, for fair comparison, we evaluate the methods for assignment problem, i.e., Re-id, Hungarian, DHN, Clustering, by giving the same pre-computed affinity matrix as in our method. For clustering methods, we can also set the number of clusters to be the true number of humans $|\mathcal{U}|$ as a prior, which we refer to as 'Clustering w $|\mathcal{U}|$' in Table II.

TABLE III
ABLATION STUDY OF DAN WITH DIFFERENT LOSSES ON MVMHA-R

Method	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
w only $\mathcal{L}_{\mathrm{E}}$	74.49	81.19	77.69
+ $\mathcal{L}_{\mathrm{C}}$	78.72	85.16	81.81
+ $\mathcal{L}_{\mathrm{C}}$+ $\mathcal{L}_{\mathrm{S}}$	80.10	84.10	82.30
+ $\mathcal{L}_{\mathrm{C}}$ + $\mathcal{L}_{\mathrm{S}}$ + $\mathcal{L}_{\mathrm{D}}$ (Ours-R)	**86.77**	**86.54**	**86.65**
w only $\mathcal{L}_{\mathrm{E}}$	80.61	89.22	84.70
+ $\mathcal{L}_{\mathrm{C}}$	85.99	**89.42**	87.67
+ $\mathcal{L}_{\mathrm{D}}$	84.02	88.61	86.26
+ $\mathcal{L}_{\mathrm{C}}$ + $\mathcal{L}_{\mathrm{D}}$ (Ours-G)	**89.81**	86.70	**88.23**

This way, we can see both our methods with RNN (Ours-R) and GNN (Ours-G) as backbone achieve the better F_1 scores against the comparative methods on both datasets. We can also see that, the performances of other compared end-to-end frameworks, i.e., GMN and PCA-GMN, do not perform as well as the proposed framework. Similar results can be found in the comparison with the multi-view matching algorithm, i.e., 'Mv Match.', which only extracts the human feature using a pre-trained network and does not integrate the feature extraction and human matching into a whole framework as our method. This may lead to limited representation ability and discrimination of the features.

D. Ablation Study

In this section, we conduct ablation analysis to evaluate the effect of different constraints in the proposed method on MvMHA-R dataset. As shown in the top half of Table III, we investigate the effectiveness of the constraint losses on the proposed method using RNN. We can see that the F_1 score is improved by adding the proposed constraints one by one. We can also see that the doubly-stochastic loss ($\mathcal{L}_{\mathrm{D}}$) and the cyclic-consistency loss ($\mathcal{L}_{\mathrm{C}}$) contribute a lot to the final performance. One possible reason is that, besides the doubly-stochastic loss constrains the local pairwise matching matrix for each view pair, the cyclic-consistency loss aggregates the multi-view information in a global way, which makes the proposed method different from the existing pairwise associations. We can get the similar results for the proposed method using GNN, as shown in the bottom half of Table III. Note that, the adjacent matrix maintains symmetry when updating GNN and we do not apply the symmetric loss ($\mathcal{L}_{\mathrm{S}}$) in the proposed method using GNN.

E. Cross-Domain Evaluation

We also test the proposed method on other three public multi-view multi-human analysis datasets, i.e., CVMHT [10], APIDIS [38] and Campus [39]. Table IV shows the testing results on these three public datasets. Since these datasets do not have training data, all the methods involving network training, including our proposed method, are learned using our collected MvMHA dataset (its training set). We can see that Hungarian, DHN and Clustering methods all show very poor performance, mainly because they perform pairwise association by ignoring the cyclic consistency across $C > 2$ views. Compared to the proposed end-to-end framework, other training-based methods chosen for comparison produce lower performance, partly because they are more sensitive to the domain difference between the training and testing datasets. The better performance achieved by the proposed method justifies its robustness to the domain difference. We find that the proposed GNN based framework (Ours-G) does not perform as well as the RNN (Ours-R) on the cross-domain evaluation. It can be explained that the GNN takes the features as the input, which is more sensitive to the distribution difference between the training and testing dataset. Similar results can be found in other GNN based approaches, such as GMN and PCA-GMN. On the contrary, the RNN based framework takes the affinity matrix as input, which depends less on the training dataset and may show greater robustness on cross-domain testing.

F. Qualitative Evaluation

We show sample visual results of our method for MvMHA in Fig. 4. In Fig. 4(a), we use green dashed lines to show a matching result generated by an appearance-based re-id method. We can see that it mistakenly matches two different people in view 3 with labels *B* and *C* to the same subject *A* in view 1. In Fig. 4(b), the blue dashed lines indicate a matching result of Hungarian algorithm, which generates an inaccurate association between views 1 and 3, without considering the cyclic-consistency constraint. The dashed red lines in both figures show that the proposed method corrects such errors by considering the cyclic consistency. The complete association results by the proposed MvMHA framework is shown by the color of the bounding boxes – associated subjects have the same color across different views.

We also show sample visual results on the synthetic dataset MvMHA-S in Fig. 5. In Fig. 5(a), we use the red dashed lines to show a matching result generated by our method. We can see that our method match all the same subjects in different views correctly. This shows that the proposed method with the closed-loop constraint considers the cyclic consistency of the same subject among all views. We also show the results generated by pairwise matching approach DHN with the yellow lines. We can see that it mistakenly matches two different people in view 4 and generates an inaccurate association between views 2 & 4 and 3 & 4 without considering the cyclic-consistency constraint.

G. Failure Cases

Figure 5(b) shows a failure case of the proposed method, where the associated subjects produced by the proposed Ours-R are bounded by the same color boxes across four views. The subject in dashed box *A* in view 3 is incorrectly associated to the subjects in red bounding boxes in the other three views. The subject in dashed box *B* in view 3 is not associated to any subjects in other three views and this is also incorrect. The former is caused by the limited appearance feature of the subject and the latter failure is caused by the mutual occlusions of the involved subjects in some views.

TABLE IV
TESTING RESULTS ON THREE OTHER PUBLIC DATASETS

Method	CVMHT			APIDIS			Campus		
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Re-id	62.24	80.09	70.04	42.16	62.53	50.36	42.36	43.41	42.88
Re-id + Hungarian	65.68	77.45	71.08	43.78	46.21	44.96	31.94	31.94	31.94
Re-id + DHN	71.45	87.51	78.45	48.34	61.19	54.01	46.18	44.99	45.58
Re-id + Clustering	72.49	76.17	73.78	40.99	61.40	49.16	20.66	47.96	28.88
Re-id + Clustering w s	78.30	71.69	74.69	55.49	53.82	54.65	37.43	42.22	39.68
GMN	47.03	64.09	53.90	34.64	35.20	34.92	18.48	18.48	18.48
PCA-GMN	34.74	46.40	39.53	19.22	19.53	19.37	11.31	11.31	11.31
Ours-G	62.66	72.82	67.33	34.44	50.98	41.10	32.70	33.43	33.06
Ours-R	86.13	80.54	**83.05**	72.93	58.23	**64.76**	62.04	45.52	**52.51**

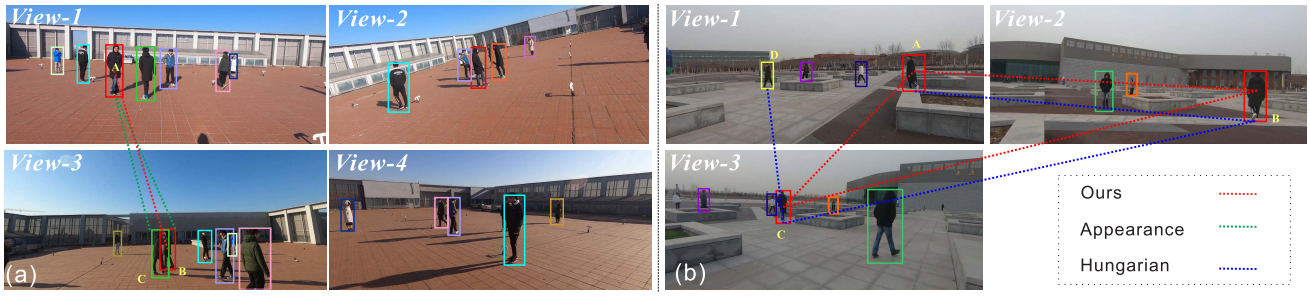

Fig. 4. An illustration of qualitative results in (a) MvMHA-R dataset and (b) CVMHT dataset. The associated subjects identified by the proposed MvMHA framework are bounded by identical-color boxes in different views.

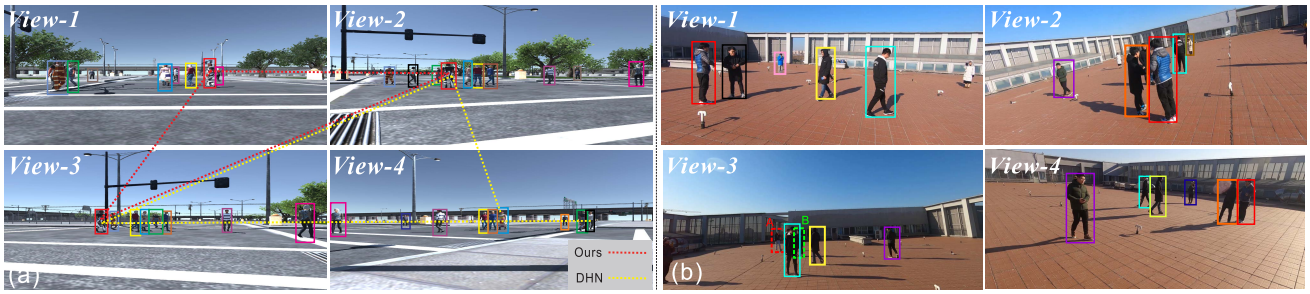

Fig. 5. An illustration of qualitative results in (a) MvMHA-S dataset, where the associated subjects identified by the proposed MvMHA framework are bounded by identical-color boxes in different views, and (b) CVMHT-R dataset with failures, where the subject A in View-3 is incorrectly associated to the subjects in red bounding boxes in the other three views and the subject B in View-3 is not associated to the corresponding subjects in other three views.

In the future, we plan to integrate more types of features, e.g., the spatial-aware feature to make our method more robust.

H. Applications

The proposed method has many applications in multi-view video analysis. We provide two examples as below.

1) Multi-View Human Tracking: We first extend the proposed MvMHA to the task of multi-view multi-human tracking. Specifically, we select a multi-view tracking dataset [41] containing 46 sequences in 12 video groups, with each group containing three to four views. As shown in Table V, we first apply two state-of-the-art multi-object tracking (MOT) algorithms, i.e., Tracktor++ [42] and TraDeS [43], to track the subjects in each view. We also implement a *baseline* method with the ResNet50 pre-trained on ImageNet for extracting the human feature with a post-processing strategy in Deep Sort [44] to achieve the multi-object tracking. We use the classical ID-based MOT metrics [45], i.e., ID precision (IDP), recall (IDR), and ID F_1 measure (IDF_1) for tracking performance evaluation. We can see that the baseline method performs more poorly than the single-view MOT method. We integrate our MvMHA results into the multi-view tracking to collaboratively track the subjects in multiple views. We find that the proposed MvMHA can clearly improve the simple baseline method and outperforms the comparative MOT methods. We also evaluate the cross-view association performance using the metrics in this paper, as shown in the right of Table V. For the single-view MOT algorithms, we provide the ground-truth cross-view human association at the first frame and propagate the association results in the subsequent frames by the single-view temporal tracking. The baseline method directly adopts the similarity among the extracted features followed by a Hungary algorithm to get the association results. We can see that the baseline with the use of MvMHA produces superior performance in the association task.

2) Multi-View Human Action Recognition: Multi-view human action recognition is a relatively new task. We adopt the proposed method to handle a multi-view multi-human activity recognition task with several modifications. Specifically, we use the dataset in [4], which was originally constructed

TABLE V
COMPARATIVE RESULTS ON MULTI-VIEW HUMAN TRACKING

Method	Tracking			Association		
	IDP	IDR	IDF_1	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Tracktor++ [42]	64.6	61.1	62.8	45.6	23.6	31.1
TraDeS [43]	67.1	65.7	66.4	53.4	30.8	39.0
Baseline	56.1	48.6	52.1	14.6	3.2	5.2
Baseline w MvMHA	76.1	65.3	70.3	74.4	43.2	54.7

TABLE VI
COMPARATIVE RESULTS ON MULTI-VIEW HUMAN ACTION RECOGNITION

Method	View	Top-1 Acc.	Top-2 Acc.
Baseline (ARG)	View 1	83.21	89.63
	View 2	80.96	87.78
	View 3	77.54	86.50
	View 4	77.99	82.49
	All View Average	79.94	86.60
w MvMHA	All View	83.93	87.28
w GT association	All View	87.79	91.28

for the interaction detection in multi-human scene. Different from the setting in [4] that only focuses on the interaction activity by considering two views, we aim to recognize the action of each subject in the scene using all four views in this dataset. We adopt a state-of-the-art group activity recognition method, i.e., ARG [46], as our baseline method for recognizing the human actions. As shown in Table VI, we first directly implement the ARG algorithm on the video from each view and evaluate the action recognition accuracy (using Top-1 and Top-2 metrics), respectively. We also calculate the average accuracy on all views as shown in the fifth row. Then, we adopt a simple strategy to achieve the multi-view human action recognition with the proposed MvMHA result. In particular, we first identify the same person appearing in different views. Then, for each subject with the (frame-level) action recognition result in each view, we integrate the human action label by selecting the result with the highest confidence. We can see that the human action recognition performance can be improved by adopting such simple strategy. We further investigate the oracle results in the multi-view setting, as shown in the last row. We can see that, with the ground-truth association results, the human action recognition performance can be further improved. This can be explained that the mutual occlusions usually hinder the action recognition with only one view and this issue can be alleviated by combining multi-view information using the proposed MvMHA method.

V. CONCLUSION

In this paper, we studied the Multi-view Multi-Human Association (MvMHA) problem by developing a new end-to-end deep-network based framework. The framework is composed of an appearance-based affinity network to obtain the affinity matrix, and a new Deep Assignment Network (DAN) to transform the affinity matrix into the assignment matrix. The proposed DAN considers multiple constraints for MvMHA across $C > 2$ views and these constraints are then converted to respective losses for network training. We built both a synthetic and a real image datasets for network training and performance evaluation. Experimental results on both datasets verified the effectiveness of the proposed method and each of its components. Testing on other three public datasets demonstrated the cross-domain robustness of our method.

REFERENCES

[1] R. Han, W. Feng, Y. Zhang, J. Zhao, and S. Wang, "Multiple human association and tracking from egocentric and complementary top views," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 2, 2021, doi: 10.1109/TPAMI.2021.3070562.

[2] G. Liang, X. Lan, K. Zheng, S. Wang, and N. Zheng, "Cross-view person identification by matching human poses estimated with confidence on each body joint," in *Proc. AAAI*, Apr. 2018, pp. 7089–7097.

[3] R. Han, J. Zhao, W. Feng, Y. Gan, L. Wan, and S. Wang, "Complementary-view co-interest person detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2746–2754.

[4] J. Zhao, R. Han, Y. Gan, L. Wan, W. Feng, and S. Wang, "Human identification and interaction detection in cross-view multi-person videos with wearable cameras," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2608–2616.

[5] K. Zheng *et al.*, "Learning view-invariant features for person identification in temporally synchronized videos taken by wearable cameras," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2858–2866.

[6] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.

[7] Y. Xu, A. Sep, Y. Ban, R. Horaud, L. Leal-Taixe, and X. Alameda-Pineda, "How to train your deep multi-object tracker," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6787–6796.

[8] R. Wang, J. Yan, and X. Yang, "Learning combinatorial embedding networks for deep graph matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3056–3065.

[9] A. Zanfir and C. Sminchisescu, "Deep learning of graph matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2684–2693.

[10] R. Han *et al.*, "Complementary-view multiple human tracking," in *Proc. AAAI*, Apr. 2020, pp. 10917–10924.

[11] A. R. Zamir, A. Dehghan, and M. Shah, "GMCP-tracker: Global multi-object tracking using generalized minimum clique graphs," in *Proc. ECCV*, 2012, pp. 343–356.

[12] A. Dehghan, S. M. Assari, and M. Shah, "GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4091–4099.

[13] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6036–6046.

[14] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou, "Fast and robust multi-person 3D pose estimation from multiple views," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7792–7801.

[15] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7792–7801.

[16] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3219–3228.

[17] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.

[18] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 26, 2021, doi: 10.1109/TPAMI.2021.3054775.

[19] K. Zheng, H. Guo, X. Fan, H. Yu, and S. Wang, "Identifying same persons from temporally synchronized videos taken by multiple wearable cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 105–113.

[20] M. Vo, E. Yumer, K. Sunkavalli, S. Hadap, Y. Sheikh, and S. G. Narasimhan, "Self-supervised multi-view person association and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2794–2808, Aug. 2021.

[21] M. Cho, J. Lee, and K. M. Lee, "Reweighted random walks for graph matching," in *Proc. ECCV*, 2010, pp. 492–505.

[22] C. Zach, M. Klopschitz, and M. Pollefeys, "Disambiguating visual relations using loop constraints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1426–1433.

[23] J. Yan, Y. Li, W. Liu, H. Zha, X. Yang, and S. M. Chu, "Graduated consistency-regularized optimization for multi-graph matching," in *Proc. ECCV*, 2014, pp. 407–422.

[24] J. Yan, J. Wang, H. Zha, X. Yang, and S. Chu, "Consistency-driven alternating optimization for multigraph matching: A unified approach," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 994–1009, Mar. 2015.

[25] J. Yan, M. Cho, H. Zha, X. Yang, and S. Chu, "Multi-graph matching via affinity optimization with graduated consistency regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1228–1242, Jun. 2015.

[26] Q.-X. Huang and L. Guibas, "Consistent shape maps via semidefinite programming," *Comput. Graph. Forum*, vol. 32, no. 5, pp. 177–186, 2013.

[27] T. Zhou, Y. J. Lee, S. X. Yu, and A. A. Efros, "FlowWeb: Joint image set alignment by weaving consistent, pixel-wise correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1191–1200.

[28] X. Zhou, M. Zhu, and K. Daniilidis, "Multi-image matching via fast alternating minimization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4032–4040.

[29] Q. Wang, X. Zhou, and K. Daniilidis, "Multi-image semantic matching by mining consistent features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 685–694.

[30] D. Pachauri, R. Kondor, and V. Singhz, "Solving the multi-way matching problem by permutation synchronization," in *Proc. NIPS*, 2013, pp. 1860–1868.

[31] Y. Xu, Y. Ban, X. Alameda-Pineda, and R. Horaud, "DeepMOT: A differentiable framework for training multiple object trackers," in *Proc. CVPR*, 2019, pp. 1–11.

[32] S. Qi, W. Wang, B. Jia, J. Shen, and S. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. ICCV*, 2018, pp. 401–417.

[33] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017, pp. 2980–2988.

[34] F. R. Gantmakher, *The Theory Matrices*. Providence, RI, USA: American Mathematical Society, 1959.

[35] M. Wang, Y. Lin, G. Lin, K. Yang, and X.-M. Wu, "M2GRL: A multi-task multi-view graph representation learning framework for web-scale recommender systems," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 2349–2358.

[36] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[37] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 608–617.

[38] F. Chen and C. De Vleeschouwer, "Personalized production of basketball videos from multi-sensored data under limited display resolution," *Comput. Vis. Image Understand.*, vol. 114, no. 6, pp. 667–680, 2010.

[39] Y. Xu, X. Liu, Y. Liu, and S.-C. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4256–4265.

[40] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[41] Y. Gan, R. Han, L. Yin, W. Feng, and S. Wang, "Self-supervised multi-view multi-human association and tracking," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 282–290.

[42] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 941–951.

[43] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, and J. Yuan, "Track to detect and segment: An online multi-object tracker," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12352–12361.

[44] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.

[45] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. CVPR*, Oct. 2016, pp. 17–35.

[46] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9964–9997.

Ruize Han received the B.S. degree in mathematics and applied mathematics from the Hebei University of Technology, China, in 2016, and the M.E. degree in computer technology from Tianjin University, China, in 2019, where he is currently pursuing the Ph.D. degree with the College of Intelligence and Computing. His major research interest is visual intelligence, specifically including multi-camera video collaborative analysis and visual object tracking. He is also interested in solving preventive conservation problems of cultural heritages via artificial intelligence.

Yun Wang received the B.E. degree from the School of Computer Science and Technology, Northeastern University at Qinhuangdao, China, in 2019, and the M.E. degree in computer technology from Tianjin University, China, in 2022. Her research interest focuses on multi-camera video collaborative analysis, especially for multi-view human association.

Haomin Yan received the B.E. degree from the School of Electrical and Information Engineering, Tianjin University, China, in 2020, where he is currently pursuing the postgraduate degree with the College of Intelligence and Computing. His research interest focuses on human action analysis, especially for the weakly supervised individual action detection and social group activity detection.

Wei Feng (Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong in 2008. From 2008 to 2010, he was a Research Fellow at the Chinese University of Hong Kong and the City University of Hong Kong. He is currently a Professor with the School of Computer Science and Technology, College of Computing and Intelligence, Tianjin University, China. His major research interests are active robotic vision and visual intelligence, specifically including active camera relocalization and lighting recurrence, general Markov Random Fields modeling, energy minimization, active 3D scene perception, SLAM, and video understanding. Recently, he focuses on solving preventive conservation problems of cultural heritages via computer vision and machine learning. He is an Associate Editor of *Neurocomputing* and *Journal of Ambient Intelligence and Humanized Computing*.

Song Wang (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana Champaign (UIUC), Champaign, IL, USA, in 2002. He was a Research Assistant with the Image Formation and Processing Group, Beckman Institute, UIUC, from 1998 to 2002. In 2002, he joined the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA, where he is currently a Professor. His current research interests include computer vision, image processing, and machine learning. He is currently serving as the Publicity/Web Portal Chair of the Technical Committee of Pattern Analysis and Machine Intelligence of the IEEE Computer Society, an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTION ON MULTIMEDIA, and *Pattern Recognition Letters*. He is a member of the IEEE Computer Society.