Graph Interaction Networks for Relation Transfer in Human Activity Videos

Yansong Tang, Yi Wei, Xumin Yu, Jiwen Lu, *Senior Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*— Recent years have witnessed rapid progress in employing graph convolutional networks (GCNs) for various video analysis tasks where graph-based data abound. However, exploring the transferable knowledge between different graphs, which is a direction with wide and potential applications, has been rarely studied. To address this issue, we propose a graph interaction networks (GINs) model for transferring relation knowledge across two graphs. Different from conventional domain adaptation or knowledge distillation approaches, our GINs focus on a "self-learned" weight matrix, which is a higher-level representation of the input data. And each element of the weight matrix represents the pair-wise relation among different nodes within the graph. Moreover, we guide the networks to transfer the knowledge across the weight matrices by designing a task-specific loss function, so that the relation information is well preserved during transfer. We conduct experiments on two different scenarios for video analysis, including a new proposed setting for unsupervised skeleton-based action recognition across different datasets, and supervised group activity recognition with multi-modal inputs. Extensive experiments on six widely used datasets illustrate that our GINs achieve very competitive performance in comparison with the state-of-the-arts.**

***Index Terms*— Graph convolutional network, skeleton-based action recognition, group activity recognition, transfer learning.**

I. Introduction

THERE are substantial graph-based data existing in various video analysis tasks. For example, skeleton-based sequence for action recognition (Fig. 1(b)), sport video with multiple people for group activity recognition (Fig. 1(c)) and many others [1]–[9]. During the past decades, great efforts have been devoted to modelling the dependency of different nodes in graphs (*e.g.* probabilistic graphical models [10], [11]). More recently, inspired by the success of deep convolutional neural networks (DCNNs) on grid-based data, a series of works have been proposed to employ convolution operators on the graph-based topology with deep hierarchical architectures [12]–[14]. With this developed graph convolutional network (GCN), rapid progress has been achieved on a variety of tasks like action recognition [15], [16], scene graph generation [17], social relationship understanding [2], *etc*.

Based on the rich graph-based data and the success of GCN on a single graphical model, we move a step towards the interaction of two graphs, which is a new direction that has been rarely explored. More intuitively, suppose we are given two graphs within some correlated structure, *e.g.* unpaired graphs sampled from different distributions or paired graphs from different modalities, can we transfer the relation[1] between them? This question is important as it extends more applications for graphical models, such as (1) unpaired relation transfer across different datasets in Fig. 1(b), and (2) paired relation transfer across different modalities in Fig. 1(c).[2] To address this issue, a straightforward way is to adopt the existing approaches on domain adaptation [20]–[23] or knowledge distillation [24], [25]. However, these methods are designed for grid-based data, and only transfer knowledge at the intermediate feature level or the softmax score level, which may lose the relation information in graph-based data after transferring.

To tackle this, we propose a graph interaction networks (GINs) architecture for transferring relation knowledge between different graphs in this paper. Specifically, we first generate a "self-learned" weight matrix for each graph, which is sent into graph convolutional layer with the node features. And output features of this layer can be utilized for the original task. During the optimization period, we enforce the two weight matrices, which contain the relation information of the corresponding graphs, to be close with each other. In this way, we can leverage the relation knowledge in a graph with more advantages (*e.g.,* with supervisory signals or higher-quality data) to guide the learning process of another graph and achieve better performance. In order to demonstrate the effectiveness of our method, we conducted experiments on a new proposed unsupervised setting for skeleton-based action recognition and the conventional supervised setting for group activity recognition. Experimental results on four datasets for skeleton-based action recognition and two datasets for group activity recognition have shown the advantages of our proposed method compared with the state-of-the-arts.

Manuscript received August 14, 2019; revised December 2, 2019 and January 31, 2020; accepted February 3, 2020. Date of publication February 11, 2020; date of current version September 3, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001, and in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, Grant U1713214, and Grant 61672306. This article was recommended by Associate Editor W. Li. *(Corresponding author: Jiwen Lu.)*

The authors are with the State Key Laboratory of Intelligent Technologies and Systems, Beijing Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China, and also with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: tys15@mails.tsinghua.edu.cn; y-wei19@mails.tsinghua.edu.cn; yuxumin16@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; jzhou@tsinghua.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2020.2973301

[1] In this paper, "relation" refers to the pairwise property between any two nodes of a graph.

[2] We provide detailed descriptions of these tasks in the caption of Fig. 1.

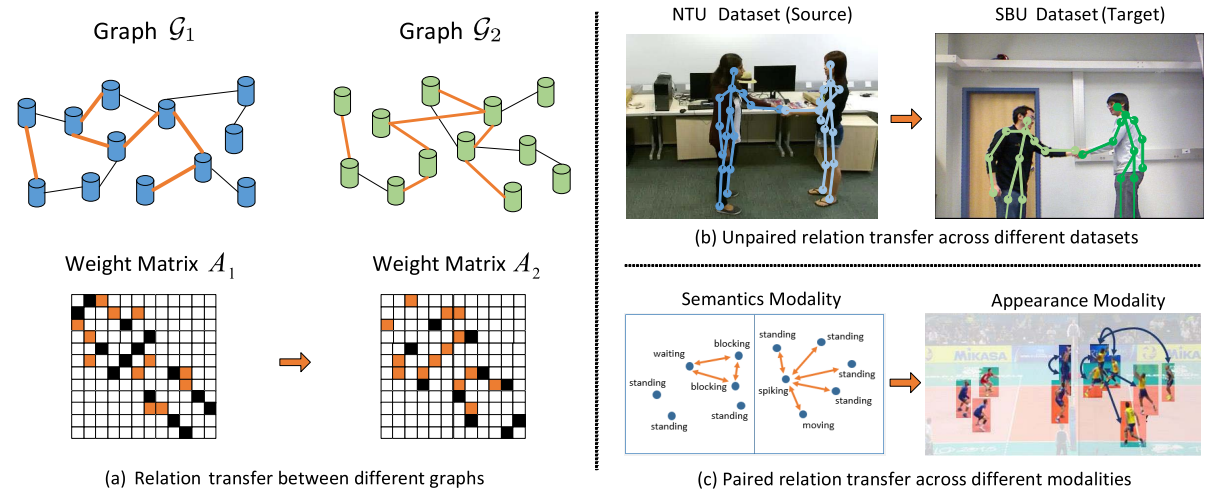


Fig. 1. **(a) Illustration of our basic idea to transfer the relation knowledge across graphs.** Here, relation refers to the pairwise property between any two nodes of a graph [18], [19]. To be specific, in the two presented graphs $\mathcal{G}_k (k = 1, 2)$, the element $A_{k,ij}$ in weight matrix A_k represents the relation between the i-th and j-th nodes in $\mathcal{G}_k$. Since there would be many different relations in a graph, for brevity, we only use the orange links, black links and disconnection to denote strong relation, ordinary relation and weak relation. In order to transfer the relation knowledge from $\mathcal{G}_1$ to $\mathcal{G}_2$, we enforce the two weight matrices A_1, A_2 to be close with each other during training. **(b) Unpaired relation transfer across different datasets.** In the figure, the graph nodes denote different joints and the links represent the bones of the body. Suppose we have unpaired skeleton-based action samples from two different datasets, where action labels are available on one dataset (source) but unavailable on the other dataset (target), the goal is to transfer the relation knowledge learned in the source dataset to the target dataset for skeleton-based action recognition. **(c) Paired relation transfer across different modalities.** In the figure, the graph nodes denote different people and the links represent strong relations. Given paired samples corresponding to the same video on different modalities, *e.g.,* a set of individual action labels (semantics modality) and tracklets of different people (appearance modality), we aim to transfer the relation knowledge from the semantic modality to the appearance modality for group activity recognition.

Our main contributions are summarized as below:

1) In contrast to conventional works on graph convolutional network which focus on a single graph, we investigate the problem of knowledge transfer across different graph-based data. This is a rarely studied direction with wide and potential applications.
2) Unlike most existing knowledge transfer methods which process the intermediate vanilla features, our GINs transfer knowledge between the weight matrices of the corresponding graphs, through which the relation information can be well preserved.
3) Different from conventional fully supervised settings, we explore a new unsupervised domain adaptation setting for skeleton-based action recognition. It is more close to real-world applications, and experimental results have revealed the great challenge of this new setting.
4) We have conducted extensive experiments on four datasets for skeleton-based action recognition and two datasets for group activity recognition. The comparison with the state-of-the-arts and ablation study have shown the effectiveness of our proposed method.

II. Related Work

A. Graph Convolutional Network

Motivated by the success of deep convolutional neural network on grid-based data, various approaches have been proposed to employ data-driven methods on graph-based structures [12]–[14], [26], [27]. Given a set of node features of a graph, which lies in the non-Euclidean space, these works aim to perform convolution operations to capture the local information of the neighboring nodes by a new architecture called Graph Convolutional Network (GCN). Recently, GCNs have also been applied into various tasks for computer vision [3], [4], [28]. For example, Yang *et al.* [17] proposed a graph R-CNN model to capture contextual information between objects and relations for scene graph generation. Wang *et al.* [2] developed a Graph Reasoning Model (GRM) to explore the persons-object interaction for social relationship understanding. There are also several approaches similar to our work in adopting GCN for human action recognition, where the graph model was built based on skeleton-based sequences [16] or RGB videos [15] respectively. Different from these works, we attempt to transfer the relation knowledge between two GCNs in different datasets or with multi-modal inputs.

B. Skeleton-Based Action Recognition

Thanks to the rapid development of 3D sensors and algorithms for pose estimation, skeleton-based action recognition has attracted more and more attention in the research field recently [29]–[31]. For a survey we refer to [32]–[34]. Here we review several works relevant to this paper, which utilized the technique of graph convolutional network. Since the human body in skeleton-based data lies in a graph-based structure, where the hinged joints and rigid bones can be modeled as the nodes and edges, Tang *et al.* [35] employed GCN to reason the dependency of different joints in spatial domain. Yan *et al.* [16] considered a skeleton-based sequence as a spatial-temporal graph, and designed an ST-GCN architecture to perform action recognition. More recently, a variety of works have been proposed to improve the process graph construction [36]–[39].

Unlike these works which follow the fully-supervised learning paradigm, we explore a new unsupervised domain adaptation setting and investigate whether the knowledge between different skeleton-based videos can be transferred.

C. Domain Adaptation

Domain adaptation, which aims to transfer the knowledge from the source domain to the target domain, has wide applications in computer vision [40]–[42]. The main challenge for domain adaptation is to reduce the shift between the data from different distributions. To address this, early shallow methods focused on exploring domain-invariant representations based on hand-crafted features [43]–[45], while recent works leveraged the power of deep neural networks to learn more transferable representations. As for the deep models, one direction is to include adaptation layers in DNN to close the distributions of the intermediate feature across different domains [21]. Another direction [20]–[23], [46], which is inspired by Generative Adversarial Networks (GANs), contains two subnetworks as domain discriminator and feature extractor. The domain discriminator aims to distinguish features from different domains, while the feature extractor aims to confuse the domain discriminator. Thus the extracted features of two domains cannot be discriminated [20]. Along the latter direction, we move a new step towards transferring the relation knowledge between graph-based data across different domains (*i.e.* skeleton-based sequences across different datasets). More recently, Ding *et al.* [47] presented a graph adaptive knowledge transfer (GAKT) model for unsupervised domain adaptation. Significantly different from this work, which performed graph-based *label* propagation for target samples, we aim to transfer the relation information in the graphed-based data across source domain and target domain. Yang *et al.* [48] explored the latent graphs between pairs of data units from large-scale unlabeled data and transferred the graphs to the downstream task. This has been shown to be effective on several NLP tasks and pixel-level image classification. Inspired by this idea, we explore the transferable knowledge between graphs for skeleton-based action recognition in domain adaptation scenario, and further extend it to another task for group activity recognition with multi-modal inputs.

D. Group Activity Recognition

Group activity recognition is an important branch of human behaviour understanding, which presents significant research value for some real-world applications like sport video analysis, traffic surveillance and many others. The main challenge of this problem is to model the relationship of different people, and numbers of works have been proposed to address this issue based on hand-crafted features [49]–[51] and deep learning models [52]–[59] respectively. Recently, Tang *et al.* [59] leveraged the attention knowledge in the semantics modality (the words of individual actions and group activity), to guide the learning process of appearance modality. In this paper, we move a further step which transfers the relation knowledge across the two modalities and shows the superior performance.

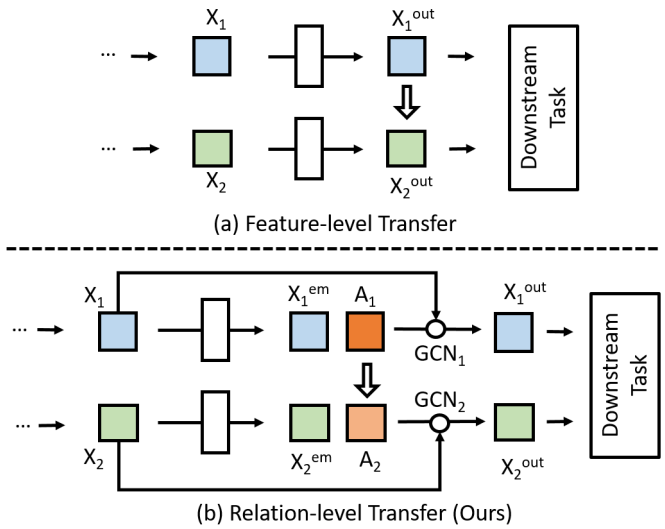


Fig. 2. Comparison of our GINs with conventional methods for knowledge transfer. (a) Conventional methods usually perform transfer at intermediate feature level. (b) Our transfers the intrinsic relation across different graphs, which can better preserve the relation information after transfer.

E. Knowledge Distillation

As a pioneering work, Hinton *et al.* [25] proposed the concept of "knowledge distillation", which aims to utilize the knowledge in a network with better performance (Teacher) to guide another network (Student) during optimization. Towards this direction, a series of approaches have been proposed to enforce "Student" to mimic "Teacher" based on their softmax outputs [25] or the intermediate features [24], [60], [61], [61], [62] of the two networks. More recently, video analysis has also benefited from knowledge distillation. For human activity analysis, several works have been proposed to utilize privileged information (*e.g.,* RGBD videos) during training, and only single modality (*e.g.,* only RGB videos) during testing [59], [63], [64]. Different from these works, which performed mimicking at vanilla features or the attention scores [59] of the two networks, we attempt to transfer the knowledge across different graphs through the weight matrices containing relation information of the input data.

III. Approach

A. Graph Interaction Networks

We present our main idea in Fig. 2. Conventional approaches for knowledge transfer usually focus on the intermediate vanilla features, which might lose relation information of the graph-based data. In comparison, our GINs aim to transfer the relation knowledge across different graphs, during which the relation knowledge is well preserved.

As illustrated in Fig. 2(b), we aim to transfer the relation knowledge from the upper graph to the bottom graph. For the upper graph $\mathcal{G}_1(X_1, A_1)$, $X_1 = \{x_{1,1}, x_{1,2}, \ldots, x_{1,N}\}$ denote the node features, while A_1 represents a weight matrix which captures the relationship between different nodes [12], [14], [16], [27]. Mathematically, the element A_{ij} encodes the connection weight between the i-th node and the j-th node, which can be formulated as $A_{1,ij} = f(x_{1,i}, x_{1,j})$. In practice,

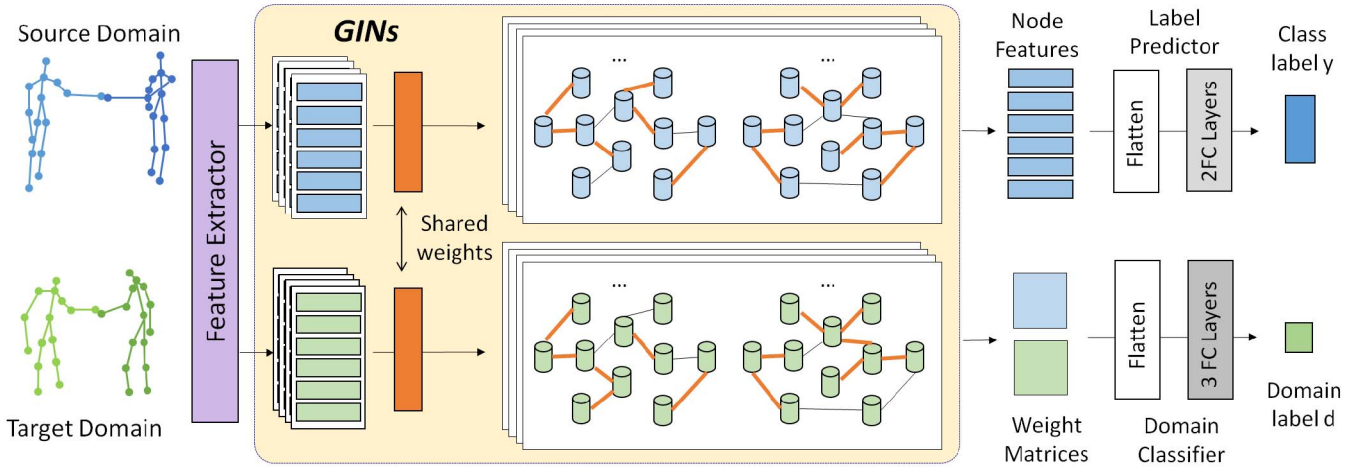


Fig. 3. The framework of our method for unsupervised skeleton-based action recognition during training. The input sequences come from two datasets with different distributions. First, we send the data from source domain and target domain into the same base network to extract joint-level features. Then we feed those features as the input to our GINs, aiming at getting the relationship among the N joints for each person and the output node features for the downstream task. On one hand, we use the output feature from the source domain to train the label predictor under the supervision of source labels. On the other hand, we feed the weight matrices from two domains into a domain classifier. By reversing the loss of the domain classifier during back-propagation, our GINs are capable to reduce the domain shift between the two weight matrices, which contain the relationship among N joints from different domains.

there are different versions for f (we compare them in the section IV.C) and we apply the following method empirically. In order to obtain symmetric weight matrix A_1, we first embed X_1 into a set of latent vectors $X_1^{em} = \{x_{1,1}^{em}, x_{1,2}^{em}, \ldots, x_{1,N}^{em}\}$ by a fully-connected layer. Then we calculate the dot product of X_1^{em} and its transposition $(X_1^{em})^T$ as follow:

$$A_1 = X_1^{em} \cdot (X_1^{em})^T. \tag{1}$$

Along with the weight matrix A_1, we sent the node features X_1 into a graph convolutional layer (denoted as GCN in Fig. 2(b)) for non-grid structure representation learning:

$$X_1^{out} = A_1 X_1 W_1, \tag{2}$$

where W_1 is a learnable weight matrix. For the bottom graph $\mathcal{G}_2(X_2, A_2)$, $X_2 = \{x_{2,1}, x_{2,2}, \ldots, x_{2,N}\}$, we perform the same process on X_2 and obtain A_2 and X_2^{out} correspondingly. In summary, this core part of our GINs can be formulated as a module with multi-input and multi-output as:

$$X_1^{out}, X_2^{out}, A_1, A_2 = GINs(X_1, X_2). \tag{3}$$

The output features X_1^{out}, X_2^{out} are utilized for the downstream task. The weight matrices A_1, A_2, embodied with the relation information, are adopted for the relation transfer between two graphs.

1) Objective Function: Our GINs targets at two objectives. The first is to learn discriminative representations for each graph. The second is to transfer the knowledge from $\mathcal{G}_1$ to $\mathcal{G}_2$. Towards these goals, we minimize the following loss function during the optimization process:

$$J = J_O^{task} + \lambda J_T^{task}(A_1, A_2). \tag{4}$$

Here λ is a trade-off parameter, "$task$" can be instantiated to a specific task, such as SAR (skeleton-based action recognition) or GAR (group activity recognition). The total loss J contains two terms. The first term J_O^{task} is the original loss of the task (*e.g.,* recognition loss). And the second term $J_T^{task}(A_1, A_2)$ is a transfer loss, which enforces the adjacent matrices A_1, A_2 from two domains to be close with each other.

B. Unpaired Relation Transfer

First, we show that our GINs can be employed for unpaired relation transfer across different datasets, which consists of graph-based data (*e.g.* skeleton-based videos).

1) Preliminaries: We consider an unsupervised domain adaptation setting for skeleton-based action recognition. Suppose we have a source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ of n_s labeled skeleton-based sequences and a target domain $\mathcal{D}_t = \{(\mathbf{x}_j^t)\}_{j=1}^{n_t}$ of n_t unlabeled examples. The source domain and target domain are sampled from different joint distributions $P(\mathbf{X}_s, \mathbf{Y}_s)$ and $Q(\mathbf{X}_t, \mathbf{Y}_t)$. The goal of this section is to develop a model, which generalizes well on the target domain, only with the supervision of source domain labels during the training stage. Since the key challenge of this problem is the shift between two domains, numbers of works have explored to reduce the distribution discrepancy. Motivated by the generative adversarial networks (GANs), a series of approaches have been proposed based on the framework of two-player minimax game [20]–[22]. The first player is a domain discriminator, which aims to distinguish the source domain from the target domain. The second player, which is a feature extractor trained simultaneously, targets at confusing the domain discriminator. Inspired by this idea, we make further exploration on graph-based data. Rather than transferring the knowledge at the feature level, we pay more attention to the intrinsic relation in the graphs.

2) GINs for Skeleton-Based Action Recognition: Skeleton-based action recognition has been explored widely in the past few years. Motivated by the fact that the joints and bones of a skeleton-based body can be considered as the nodes and edges of a graph respectively, numbers of works have explored

the graph-based model for this task and achieved promising progress. Different from most previous works, which focus on the supervised learning setting on one dataset, we explore a new setting in unsupervised learning across two different datasets. The first dataset is a source domain with labels, and the second dataset is a target domain without labels. This is a scenario which is more close to real-world applications.

In this problem, the input is a skeleton-based video with a size of $L \times N \times 3$, where L, N and 3 denote the number of video frames, the number of joints and the number of dimensions of 3D positions, respectively. First, for all samples in the source domain and target domain, we employ a model to extract a set of node features as $\{X_l\}_{l=1}^{L} = \{X_{s,l}\}_{l=1}^{L} \cup \{X_{t,l}\}_{l=1}^{L}$,[3] where $X_{s,l}, X_{t,l} \in R^{N\times d}$ denote the representation at the lth frame of source domain and target domain, respectively. Then, we sent them into our GINs module to model the relationship of different nodes as follow:

$$X_{s,l}^{out}, X_{t,l}^{out}, A_{s,l}, A_{t,l} = GINs(X_{s,l}, X_{t,l}). \tag{5}$$

Here $X_{s,l}^{out}, X_{t,l}^{out}$ refer to the output features of the GINs, and $A_{s,l}, A_{t,l}$ are the output weight matrices. The subscripts l, s and t represent the $l-$th frames, source domain and target domain, respectively.

During the training stage, in order to utilize the supervision labels of source domain, we flatten the $\{X_{s,l}^{out}\}_{l=1}^{L}$ into a vector, and send into a label predictor G_y with two fully connected layers to output the final prediction. We perform max-pooling[4] on $\{A_{s,l}\}_{l=1}^{L}, \{A_{t,l}\}_{l=1}^{L}$, and send the outputs A_s, A_t into the domain classifier G_d, which contains three fully connected layers. The loss function can be written as follow according to the framework of Equation (4):

$$\begin{aligned} J &= J_O^{SAR} + \lambda J_T^{SAR}(A_s, A_t) \\ &= \frac{1}{n_s}\sum_{x_i\in\mathcal{D}_s} J_{CE}(G_y(GINs_X(x_i)), y_i) \\ &\quad -\frac{\lambda_{SAR}}{n_s+n_t}\sum_{x_i\in\mathcal{D}_s\cup\mathcal{D}_t} J_d(G_d(GINs_A(x_i)), d_i), \end{aligned} \tag{6}$$

where $GINs_A(x_i)$ and $GINs_X(x_i)$ denote the output of weight matrix and node feature of GINs respectively. d_i is a domain label, which equals 1 if x_i is from the source domain, and 0 otherwise. The first term of J is the classification loss for the main classification task, where we calculate the cross-entropy between the softmax output and the ground-truth action label in the source domain. And the second term is the transfer loss, which is designed for aligning the weight matrices from the two domains. The second loss is subtracted to confuse the domain discriminator during the training phase and help GINs to learn the domain-invariant features. We follow the DANN [20] and use the gradient reversal layer. During the testing stage, we feed the feature X_s^{out} into the G_y layer and predict the final label.

[3]We present the details of feature extraction in Section IV.B.

[4]We study the influence of max-pooling operation on weight matrices in Section IV.C

C. Paired Relation Transfer

In this subsection, we further employ our GINs for paired relation transfer across different modalities. As an application example, we study the problem of group activity recognition, which aims to discern the activity label of a group of people. As suggested in previous works, we can adopt a set of tracklets of different people as pre-processed inputs. More recently, Tang *et al.* [59] has explored another auxiliary input modality, *i.e.* the words of individual actions, during training phase.[5] They designed a Teacher Network for these semantics-based modalities and a Student Network for the appearance-based modality. Following this setting, we show that our GINs are capable to transfer the relation knowledge from the Teacher Network to the Student Network.

1) GINs for Group Activity Recognition: On one hand, we denote the features of the Student Network as $X_a = \{x_{a,l}\}_{l=1}^{L}$, where $x_{a,l} \in R^{N\times d}$ stands for the appearance features of N people in the lth frame (See Section IV.B for details of feature extraction). On the other hand, we represent the semantics features (*i.e.* several words of individual actions) as $X_w \in R^{N\times d}$. Thus, the $\{X_a, X_w\}$ is a pair-wise input of different modalities associated with the same video. Since a group of people can be considered as a graph, where each node represents a single person, and edge denotes the relationship between different people, we can model the pair-wise input as two graphs. At the lth frame, we feed the inputs into the GINs module as:

$$X_{a,l}^{out}, X_w^{out}, A_{a,l}, A_w = GINs(X_{a,l}, X_w). \tag{7}$$

Then we apply two attention pooling layers. For the Teacher Network, the input of the attention pooling layers is $X_w^{out} = \{x_n^{teacher}\}_{n=1}^{N}$. The attention pooling layer is derived as follow:

$$s_n = ReLU(W_1 x_n^{teacher} + b_1), \tag{8}$$

$$\alpha_n = exp(s_n)/\sum_{j=1}^{N} exp(s_j), \tag{9}$$

$$w^{teacher} = \sum_{n=1}^{N} \alpha_n \cdot x_n^{teacher}, \tag{10}$$

where W_1 and b_1 denote the weighted matrix and biased term. s_n and α_n refer to the score and the normalized attention for the n-th person in semantics domain. The obtained $w^{teacher}$ is sent into an fc layer for group activity recognition. For the Student Network, the input of the attention pooling layer is $X_a^{out} = \{\{x_{l,n}^{student}\}_{l=1}^{L}\}_{n=1}^{N}$. Similar to the Teacher Network, the attention pooling layer is designed as below:

$$s_{l,n} = ReLU(W_2 x_{l,n}^{student} + b_2), \tag{11}$$

$$\beta_{l,n} = exp(s_{l,n})/\sum_{j=1}^{N} exp(s_{l,j}), \tag{12}$$

$$w_l^{student} = \sum_{n=1}^{N} \beta_{l,n} \cdot x_{l,n}^{student}. \tag{13}$$

[5]We should not use these labels in the testing stage as they are not available.

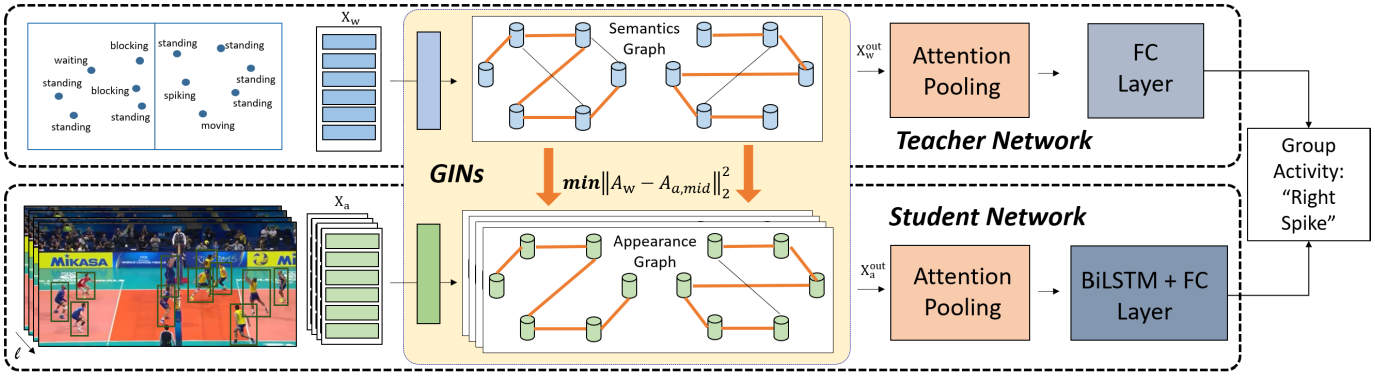


Fig. 4. Illustration of our approach for group activity recognition based on the Teacher-Student paradigm. During training, the inputs are a set of tracklets and individual action words, which can be treated as two types of graph-based data in different modalities (we denote them as appearance graph and semantics graph). We feed their extracted features into our GINs simultaneously, which aims to distill the relation knowledge in the semantics graph to guide the learning process of the appearance graph. We aggregate the output node features of the graph by attention pooling and send them into a classifier to obtain the label of group activity. During testing, as the words (labels) of individual actions are not available, we can not use them nor their weight matrix. Therefore, we only employ the student network in testing period. Note that the network parameters of the Teacher Network and the Student Network are not shared.

The output feature $\{w_l^{student}\}_{l=1}^{L}$ are fed into a BiLSTM layer and another fc layer to obtain the final result.

2) Relation Transfer by Graph Interaction: For semantics and appearance modalities, there are two graph modules and they consistently model the relationship of different people, thus it is reasonable to consider these two modules jointly. As the performance of the Teacher Network is better than the Student Network,[6] we attempt to employ the relation knowledge of the Teacher Network to guide the Student Network. In practice, we first train the Teacher Network with the provided labels of training samples. And with our proposed GINs, the Student Network is enforced to mimic the relation knowledge in the Teacher Network during the optimization course with the following loss function:

$$\begin{aligned} J &= J_O^{GAR} + \lambda J_T^{GAR}(A_w, A_a) \\ &= J_{CE}(P_a, y) + \lambda_{GAR}||A_w - A_{a,mid}||_2^2. \end{aligned}$$

Here λ_{GAR} is the hyper-parameters to balance the effects of two different terms to make a good trade-off. The first term is a cross-entropy loss for activity recognition. P_a and y represent the softmax output and ground-truth label of the Student Network. The second term aims to enforce the Student's relation to preserve the Teacher's semantics relation. We adopt the MSE loss to measure the distance between the weight matrices of the Teacher Network and Student Network. In the process of training, we optimize all the parameters of our Student Network by the backpropagation through time (BPTT) algorithm [65]. It is worth attention that the Teacher Network is only allowed to guide the Student Network during the training phase, as the ground-truth label $Y = \{y_n\}_{n=1}^{N}$ is not available at the testing stage.

3) Utilizing Location Information: Previous works have shown the privileges of location information for action recognition. To further boost the recognition performance, we make full use of the location information, which is associated with the provided tracklets. Hence, no extra annotations are required. Inspired by [15], which embeds the position information for graph construction, we define another weight matrix A^{loc}. First we calculate the central location of the mth person as $c_m = (\gamma\, x_{m,mid}/W_I, \gamma\, y_{m,mid}/H_I)$. Here, we use $x_{m,mid}$ and $y_{m,mid}$ to represent the central positions coordinates of the input tracklets. W_I and H_I are width and height of the video frame. and γ is a scaled parameter, which is set to 10 empirically. Instead of being learnt by networks, each element a_{mn} in A^{loc} is defined based on the spatial coordinates as: $a_{mn} = exp(-||c_m - c_n||_2^2/2)$. In this way, a_{mn} will turn to be a large value if two people are closed to each other, and vice versa. Parallel with the original GINs module, we feed the node features into other GCN models:

$$X_{a,l}^{loc,out} = A_{a,l}^{loc} X_{a,l} W_a, \quad X_w^{loc,out} = A_{w,mid}^{loc} X_w W_w. \tag{14}$$

The output features $X_{a,l}^{loc,out}$, $X_w^{loc,out}$ are concatenate with $X_{a,l}^{out}$, X_w^{out} for the downstream task. We do not perform transfer on A_{loc}, as it is fixed based on the location information. Besides, the location feature of a single person is a 4-dimension vector $f_{loc} = [x_1, y_1, x_2, y_2]$, which is normalized into $[-1,1]$ based on the width and height of the corresponding frame. The location feature is utilized twice, concatenated at the node feature level and the layers before the final classification for better performance.

D. Discussion

In this subsection, we provide some deeper discussion on our proposed method.

1) Discussion on the Insight of GINs: Recent works have shown the great importance of relation information for various video analysis tasks, especially for skeleton-based action recogniton [16], [35]–[39] and group activity recognition [57], [58]. In this paper, we further explore the transferable relation knowledge across different graphs. Therefore, $\mathcal{G}_2$ (Section III.A) can better mine relation information from $\mathcal{G}_1$ which enables it to learn more powerful knowledge. Motivated by the fact that not all the features need to be transferred, our

[6]Because the inputs of Teacher Network are the ground-truth label of individual actions, while the Student Network takes the tracklets as inputs and requires a more complex feature learning process.

GINs make alignment based on the weight matrices of different graphs at the higher level. Specifically, for skeleton-based action recognition, conducting fully alignment on the two distributions of node features is unreasonable, because some information (*e.g.* viewpoint, height) is not strictly required to be similar in the two domains. However, the relations of different joints are irrelevant to the two domains and should maintain consistently. And for group activity recognition, different people have different contributions to the final results. Therefore, it is unnecessary to transfer all the people's features and more reasonable to transfer their relations.

2) Discussion on Relations to be Transferred: (1) For skeleton-based action recognition, there might already be some kind of relation graphs such as the physical structure of human body, but it is not sufficient for the recognition task. For example, the relation between the two hands is important for recognizing the action "clapping hand", but the two hands are disconnected physically. Therefore, it is important to learn other "action-aware relation" (*i.e.,* the relation which is important for recognizing the action class). Actually, there have been a series of works utilizing the automatical learning scheme to explore this kind of relation for skeleton-based action recognition [16], [36], [37], [66]. And in this paper, we study the unsupervised domain adaptation setting and aim to transfer this relation learned in the source dataset to the target dataset. In the source dataset, the relation is learned with the supervision of action label. Thus it is more privileged for recognizing the action than that in the target dataset where the action label is unavailable. (2) For group activity recognition, similarly, we aim to explore the "activity-aware relation" of different people for the recognition task. For example, discovering the strong relation between the "spiking" and "setting" players would benefit recognizing the group activity "spike". In fact, this kind of relation has also been explored by several methods in an automatically learning scheme [58], [67]. In this work, we target at transferring this relation learned in the semantics domain to that of appearance domain. Since the data in the semantics domain has higher-quality than that in the appearance domain, the semantics relation is better for recognizing the group activity. We further demonstrate this viewpoint in Fig. 9.

3) Discussion on the Loss Function: In this paper, we propose a general objective function for GINs as $J = J_O^{task} + \lambda J_T^{task}(A_1, A_2)$, where J_O^{task} is the classification loss and $J_T^{task}(A_1, A_2)$ is a transfer loss, which enforces the adjacent matrices A_1, A_2 from two domains to be close with each other. Specifically, we design two formats of J_T^{task} for the two tasks according to their different goals. For unpaired relation transfer (the skeleton-based action recognition task), the core is to reduce the *distribution* discrepancy between the source domain and target domain, and it is hard to find paired samples to perform mimicking. Therefore, we employ the adversarial loss for J_T^{SAR}. For paired relation transfer (the group activity recognition task), the goal is to enforce each *sample* in the appearance domain to mimic its paired *sample* in the semantics domain. Hence, we adopt the MSE loss for J_T^{GAR}.

IV. Experiments

A. Datasets and Experiment Settings

For *skeleton-based action recognition*, we adopted NTU RGB+D dataset (NTU) [68], SBU Kinect Interaction dataset (SBU) [69], Online RGBD Action dataset (ORGBD) [70] and MSRDaily Activity3D dataset (MSRDA3D) [71]. We conducted experiments under two new-proposed settings as NTU $\rightarrow$ SBU and ORGBD $\rightarrow$ MSRDA3D.[7]

1) NTU $\rightarrow$ SBU: The NTU RGB+D dataset is a large-scale dataset for skeleton-based action recognition, which contains 56880 skeleton-based videos of 60 action categories. In comparison, SBU is a smaller dataset, which comprises 282 skeleton-based sequences of 8 classes. In order to perform unsupervised domain adaptation, we made the following two pre-process steps on the NTU dataset. (1) We selected the sequences of the 8 categories[8] corresponding to those in SBU. Finally, 7513 videos were selected out. (2) As SBU and NTU were collected by Kinect v1 and Kinect v2, the number of joints for each person are 15 and 25 respectively. Hence, we only used the 15 corresponding joints[9] in the NTU dataset.

2) ORGBD $\rightarrow$ MSRDA3D: Both the Online RGBD Action dataset and the MSRDaily Activity3D dataset are captured by the Kinect device, and the number of joints for each person is 20. For the unsupervised domain adaptation task, we adopted the 5 action categories which exist in both datasets.[10] As results, we obtained 240 and 100 videos from the ORGBD dataset and the MSRDA3D dataset respectively.

For *group activity recognition*, we evaluated Volleyball dataset [72] and Collective Activity (CA) dataset [73].

Volleyball Dataset [72]: The Volleyball dataset is currently the largest benchmark for group activity recognition. It consists of 4830 clips which are trimmed from 55 long sport videos and each clip contains 10 frames. The annotations include the tracklets of players, 9 individual action labels (waiting, setting, digging, falling, spiking, blocking, jumping, moving and standing) and 8 group activity categories (right set, right spike, right pass, right winpoint, left winpoint, left pass, left spike and left set). We follow the protocol adopted in [72] to separate the training/testing sets. Our experimental results are based on the evaluation metrics of Multi-class Classification Accuracy (MCA) and Mean Per Class Accuracy (MPCA).

3) Collective Activity (CA) Dataset [73]: The Collective Activity dataset is a widely used dataset for group activity recognition. It contains 44 video clips, which are labeled with 6 individual action classes (NA, crossing, walking, waiting, talking and queueing) and 5 group activity labels (crossing, walking, waiting, talking and queueing). Similar to the Volleyball dataset, there are 10 frames in each short clip. We adopt the training and testing splits as suggested in [49]. Noticing

[7] A $\rightarrow$ B denotes that we used A as source domain dataset and B as target domain dataset.

[8] We selected the action ID $\{59, 60, 51, 52, 58, 55, 56, 50\}$ in the NTU dataset, which correspond to $\{1 \sim 8\}$ in the SBU dataset.

[9] We chose the joints with ID $\{4, 3, 2, 9, 10, 12, 5, 6, 8, 17, 18, 19, 13, 14, 15\}$ in order of the SBU dataset.

[10] We chose the action ID $\{0, 1, 2, 4, 5\}$ in the ORGBD dataset, which corresponds to $\{1, 2, 6, 4, 3\}$ in the MSRDA3D dataset.

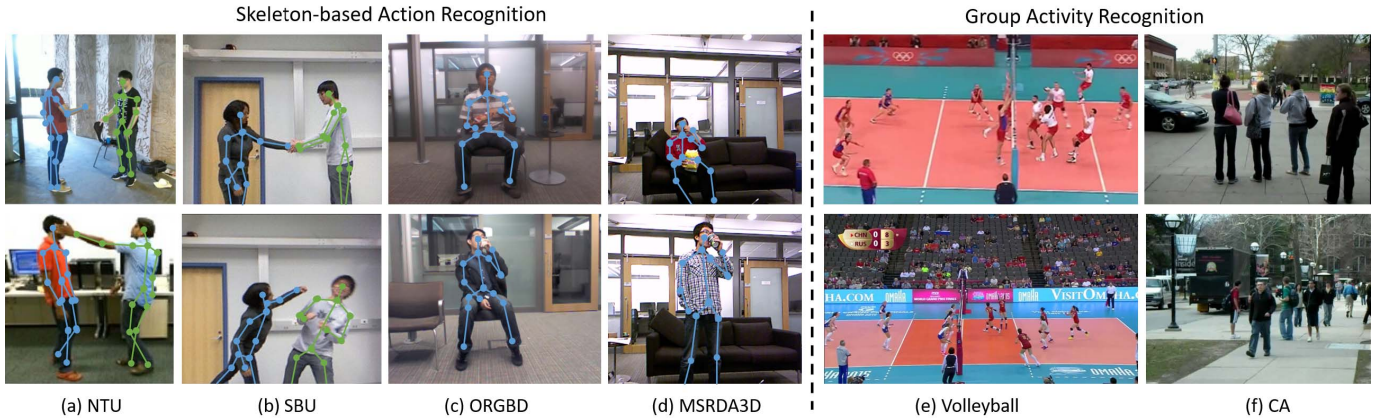


Fig. 5. Examplar frames from different datasets. We conducted experiments on (a) NTU RGB+D dataset (NTU), (b) SBU Kinect Interaction dataset (SBU), (c) Online RGBD Action dataset (ORGBD) and (d) MSRDaily Activity3D dataset (MSRDA3D) for skeleton-based action recognition. We evaluated (e) Volleyball dataset and (f) Collective Activity (CA) dataset for group activity recognition.

the clarification in [73] which originally presented the dataset, the "walking" activity is actually an individual action but not a collective activity, we adopted the experimental setup in [74], to combine the "walking" and "crossing" categories into a new class of "moving". We report the performance of Mean Per Class Accuracy (MPCA), based on which we can better compare with the previous results under this setting.

B. Implementation Details

1) Skeleton-Based Action Recognition: In our experiments, we followed the standard protocol for unsupervised domain adaptation [20], [21]. For each person, we first subtracted the coordinates of the torso (NTU $\rightarrow$ SBU) or spine (ORGBD $\rightarrow$ MSRDA3D) from the other joints. Then we normalized each input video to a fixed length L with bilinear interpolation operation. L was set to 16 in our experiments. Similar to [75], we produced motion data by making the difference between two adjacent frames at the coordinates dimension and sent the two-stream data into the Feature Extractor network. For each stream, the input data passed through two convolution layers to learn the point-level representation, and other two convolution layers to learn the co-occurrence information between joints [75]. The output features correspond to the node features X described in Section III.A, which were then sent into our GINs module. We set the trade-off parameter $\lambda_{SAR} = \lambda^*$ [20] as follow:

$$\lambda^* = \lambda_p = \frac{2}{1 + exp(-\gamma \cdot p)} - 1, \tag{15}$$

where p linearly changed from 0 to 1 and γ was set to be 10 in our experiments. This schedule gradually increased λ^* from 0 to 1 with the training process. We set the initial learning rate for the base network to 0.0001 and discriminator network to 0.00005. For each batch, we randomly selected 128 unpaired videos (64 from source domain and 64 from target domain).

2) Group Activity Recognition: The input of Teacher Network is a set of words (labels) of individual actions. We encoded them into one-hot vectors and passed them through an fc layer with the size of 32. We also utilized position coordinates and transformed them into 32 dimensions using another fc layer. Then we concatenated these two types of features as the input X_w for GINs. We set the initial learning rate to 0.003 for the Teacher Network. The input of the Student Network is a set of tracklets. We adopted the same scheme in [59] to extract RGB features based on VGG16 and LSTM networks. For better model the dynamic information, we also combined the optical flows [59], [76] and location coordinates. We encoded RGB features, optical flow features and position features to 1024 dimensions, 1024 dimensions and 256 dimensions respectively. Then we concatenated these encoded features as the initial appearance representation X_a. We utilized the same adjacency matrix in the Teacher Network to guide those in the Student Network. Since the Volleyball dataset is much larger than the CA dataset, we stacked three graph convolution layers for the Student Network for the previous one, and utilized one graph convolution layer for the other to avoid overfitting. The hidden size of the bidirectional LSTM layer is 128. During the Teacher guided training process, the Student Network was optimized with the initial learning rate of 0.00003 for the Volleyball dataset and 0.0001 for the CA dataset. As for ratio of different parts of losses, we set $\lambda_{GAR} = 1$. For both the Teacher Network and Student Network, we adopted Adam optimization method. The batchsize was set to be 16 for the Volleyball dataset and 8 for the CA dataset.

C. Evaluation on Skeleton-Based Action Recognition

In this task, we compare our method with the following baselines and state-of-the-art methods:

- Source Only: The "Source Only" model is trained without utilization of data from target domain (there is no domain classifier subnetwork in this model).
- Geo Transfer: In order to see whether the geometric transfer can reduce the domain shift between two datasets, we evaluated "Geo Transfer" method by rotating the people based on the joints of "left-shoulder", "right-shoulder" to align the x axis, and "torso", "neck" (NTU $\rightarrow$ SBU) or "spine", "shoulder center"(ORGBD $\rightarrow$ MSRDA3D) to align the y axis. This makes people in all videos facing to the camera frontally.

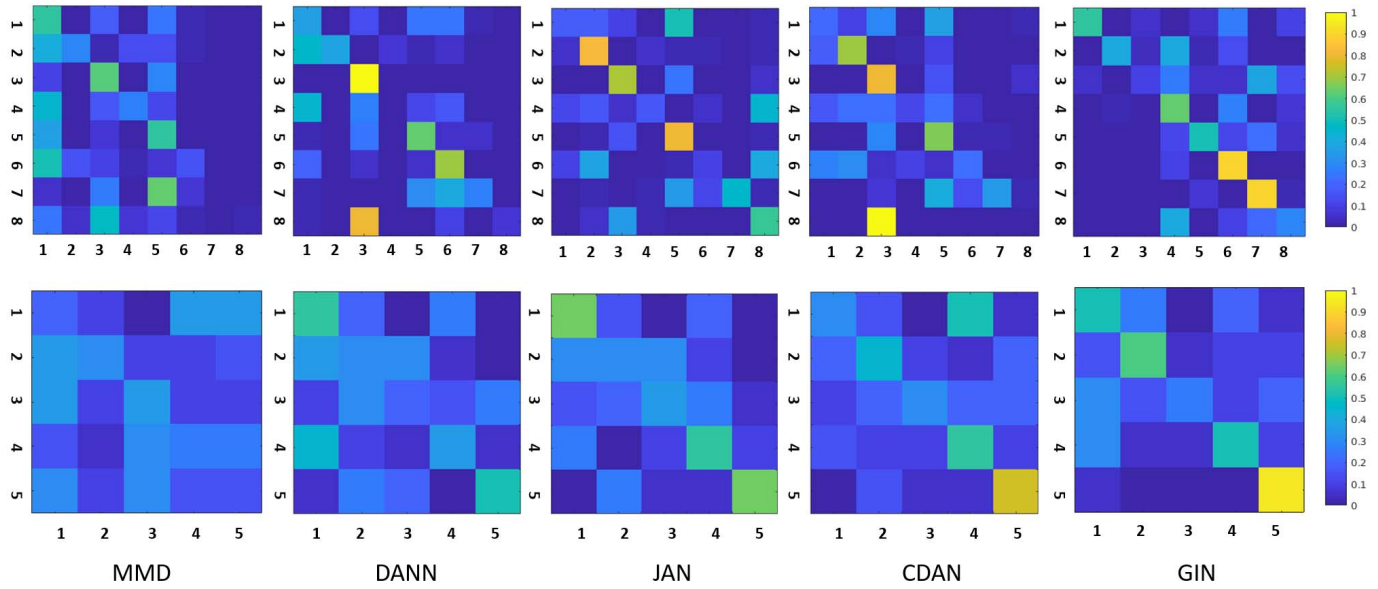


Fig. 6. The confusion matrices of different methods on the SBU (top row) and ORGBD (bottom row) datasets. The ground truth labels are shown on the vertical axis, while the predicted labels are displayed on the horizontal axis. Labels of the NTU → SBU: 1: punching, 2: exchanging something, 3: hugging, 4: handshaking, 5: pushing, 6: kicking, 7: walking apart, 8: walking towards. Labels of the ORGBD → MSRDA3D: 1: drinking, 2: eating, 3: using laptop, 4: making phone call, 5: reading book.

TABLE I
COMPARISON OF THE SKELETON-BASED ACTION RECOGNITION ACCURACY (%) UNDER THE UNSUPERVISED DOMAIN ADAPTATION SETTING. SOURCE DOMAIN: NTU DATASET, TARGET DOMAIN: SBU DATASET

Method	Accuracy	Year
Source Only	38.4	-
Geo Transfer	35.8	-
MMD [77]	31.4	2015
DANN [20]	46.3	2017
JAN [21]	47.6	2017
CDAN [22]	39.9	2018
GAKT [47]	31.8	2018
BSP [78]	32.4	2019
GINs	**50.7**	

- MMD [77], CDAN [22], DANN [20], JAN [21], GAKT [47] and BSP [78]: These are recent state-of-the-arts for unsupervised domain adaptation, we employed the value reported in the corresponding original papers for the parameter λ_{SAR}.

1) Experimental Results on NTU → SBU: As shown in Table I, we first find that "Geo Transfer" method causes a decline in performance based on "Source only". This is mainly because all the skeleton sequences in "NTU → SBU" setting contain two people, and the geometric transfer method cannot deal well with the interaction action (In contrast, this simple spatial transformation can improve the performance on individual actions. Please see the results in "ORGBD → MSRDA3D"). Hence, other transfer learning methods should be explored. Moreover, we find our GINs model achieves 50.7% accuracy, which outperforms other compared state-of-the-art methods. We further show some confusion matrices in Fig. 6, where our model has clear advantages on recognizing the actions of "kicking" and "walking apart".

TABLE II
ABLATION STUDY OF THE SKELETON-BASED ACTION RECOGNITION ACCURACY (%) UNDER THE UNSUPERVISED DOMAIN ADAPTATION SETTING. SOURCE DOMAIN: NTU DATASET, TARGET DOMAIN: SBU DATASET

Method	Construct Graph?	Transfer Level	Acc.
DANN [20]	No	Vanilla Features	46.3
Node-Trans^1	Yes	(X_1,X_2)	48.2
Node-Trans^2	Yes	(X_1^{em},X_2^{em})	45.9
Node-Trans^3	Yes	(X_1^{out},X_2^{out})	48.0
GINs	Yes	(A_1,A_2)	**50.7**

2) Ablation Study: We provide some ablation study results in Table II, where DANN [20] corresponds to the method in Fig. 2(a) that transfers the vanilla features. For Node-Trans^1, Node-Trans^2 and Node-Trans^3, they construct graphs as Fig. 2(b), but perform transfer at node feature level ($X_1 \rightarrow X_2$), ($X_1^{em} \rightarrow X_2^{em}$) and ($X_1^{out} \rightarrow X_2^{out}$). Actually, performing transfer at ($A_1 \rightarrow A_2$) is a more flexible constraint on the two graphs, as it does not strictly align the node level features across two graphs. As a result, our GINs achieve 4.4%, 2.5%, 4.8% and 2.7% improvements on DANN, Node-Trans^1, Node-Trans^2 and Node-Trans^3 respectively. Convincingly, these results demonstrate the effectiveness of our proposed scheme for relation transfer.

3) Analysis on the Weight Matrix: As the weight matrix is an important factor in graph convolutional network, we study different ways to construct the weight matrix A aside from our original formulation in Equation (1). As shown in Table III, we explore 4 methods to build the weight matrix A based on the i-th and j-th embedded node features x_i^{em} and x_j^{em}. The "Relation Module" is proposed in [19], and our proposed method is equal to the "Product", which

TABLE III

ANALYSIS ON DIFFERENT FORMULATIONS OF WEIGHT MATRIX A FOR SKELETON-BASED ACTION RECOGNITION. SOURCE DOMAIN: NTU DATASET, TARGET DOMAIN: SBU DATASET

Method	A_{ij}	Accuracy
Sum	$\|\|x_i^{em} + x_j^{em}\|\|_2^2$	38.7
Relation Module [19]	$g_\theta(concat[x_i^{em}; x_j^{em}])$	48.2
Gaussian Distance	$exp(-\|\|x_i^{em} - x_j^{em}\|\|_2^2/2)$	46.1
Product	$< x_i^{em}, x_j^{em} >$	**50.7**

TABLE IV

STUDY ON THE MAX-POOLING OPERATION. WE REPORT THE SKELETON-BASED ACTION RECOGNITION ACCURACY (%) UNDER THE NTU→SBU SETTING

Method	GINs (with max-pooling)	GINs (without max-pooling)
Accuracy	50.7	45.4

TABLE V

ANALYSIS ON DIFFERENT λ_{SAR} FOR SKELETON-BASED ACTION RECOGNITION. SOURCE DOMAIN: NTU DATASET, TARGET DOMAIN, SBU DATASET

λ_{SAR}	$0.01\lambda^*$	$0.1\lambda^*$	λ^*	$10\lambda^*$
Accuracy	50.3	49.5	**50.7**	44.1

calculates the inner-product of x_i^{em} and x_j^{em} for A_{ij}. As a result, we observe that the "Product" method achieves the best performance of 50.7%, which is 12.0%, 2.5% and 4.6% higher than the "Sum", "Relation Module" and "Gaussian Distance" respectively.

4) Analysis on the Max-Pooling Operation on Weight Matrices: We study the effect of max-pooling method on weight matrices at the setting of NTU→SBU in Table IV. The "GINs (without max-pooling)" denotes passing the matrices at each time step to the domain classifier, which causes 5.3% decrease based on the "GINs (with max-pooling)". This is because not every frame in the source video need to be strictly aligned with that in the target video, and the max-pooling is a proper operation to aggregate the temporal information of a video.

5) Analysis on the Hyper-Parameter λ_{SAR}: We also study the effect on different λ_{SAR} in Table V. When $\lambda_{SAR} \leq \lambda^*$, the accuracy varies relatively slightly and achieves the maximum value of 50.7% at λ^*. When λ_{SAR} increases to $10\lambda^*$, the accuracy drops to 44.1% which indicates that λ_{SAR} cannot be too large.

6) Analysis on the Influence of Different Amounts of Training Samples: We conduct experiments under the NTU→SBU setting to study the influence of the amount of source domain data. We use the training samples in the source domain under the ratio of [25%, 50%, 75%, 100%]. We evaluate our GINs and the baseline method DANN [20], and present the compared results in the following Fig. 7. As it shows, both two methods can achieve better results with more training data, and our proposed GINs are more effective. Moreover, our GINs yield a relatively promising result of 41.4% with only 50% training data from source domain. It outperforms the accuracy

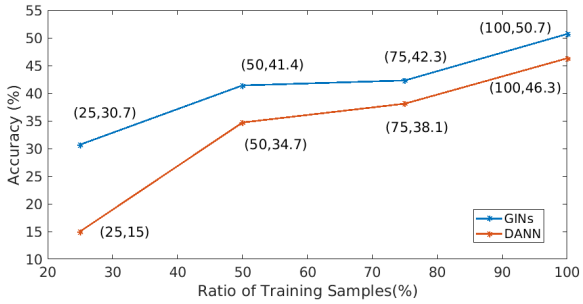


Fig. 7. Study on different ratios of training samples in the source domain (NTU dataset) for skeleton-based recognition under the unsupervised domain adaptation setting (NTU→SBU).

TABLE VI

COMPARISON OF THE SKELETON-BASED ACTION RECOGNITION ACCURACY (%) UNDER THE UNSUPERVISED DOMAIN ADAPTATION SETTING. SOURCE DOMAIN: ORGBD DATASET, TARGET DOMAIN: MSRDA3D DATASET

Method	Raw Data	Geo Transfer	Year
Source Only	18.7	48.3	-
MMD [77]	25.0	25.5	2015
DANN [20]	35.2	39.3	2017
JAN [21]	34.6	49.2	2017
CDAN [22]	31.3	48.7	2018
GAKT [47]	34.9	48.4	2018
BSP [78]	31.2	41.3	2019
GINs	**40.2**	**51.5**	

of 38.4% obtained by the "source only" method using 100% training data.

7) Experimental Results on ORGBD → MSRDA3D:

8) Discussion: Although great progress have been achieved in recent years for supervised skeleton-based action recognition [37], [66], [75], the overall experimental results are much lower under the unsupervised setting. This indicates the great challenges for this problem, which leaves plenty of room for the future work to achieve further improvements. Moreover, we have also conducted experiments on SBU → NTU, which adopted SBU as a source dataset and NTU as a target dataset. However, the results on all methods are inferior (less than 25%).[11] This is because the size of SBU dataset is rather small, thus the samples of the source domain are insufficient to optimize the model well. In the future, it is desirable to collect dataset with larger scale for this interesting problem.

We then conduct experiments on the ORGBD → MSRDA3D setting, where the skeleton sequences only contain a single person. As shown in Table VI, We test 7 methods and "Source Only" with "Raw Data" and "Geo Tranfer". "Raw Data" refers to the original data from the dataset, while the "Geo Transfer" refers to the data after the spatial transformation. As a result, we can see "Geo Transfer" can improve the result for all the methods in our experiments, which indicates the effectiveness of the spatial transformation for action performed by a single person. We notice that

[11] [MMD, DANN, JAN, CDAN, GAKT, BSP, GINs] achieve an accuracy(%) of [18.4, 19.3, 19.3, 23.1, 21.2, 19.8, 24.1] respectively.

TABLE VII

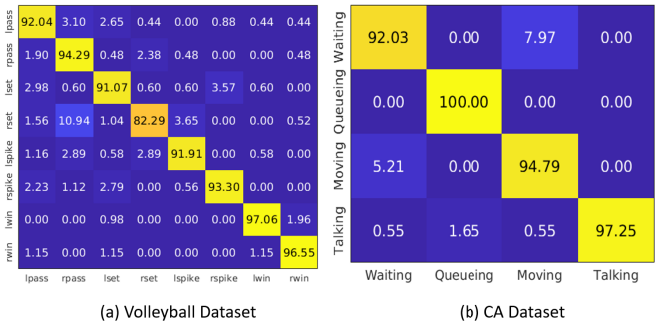
COMPARISON WITH [79]. SPTS+GCN$_{FW}$ IS EXPLORED IN THE "FUTURE WORK" SECTION [79]. AM REFERS TO THE ATTENTION MODEL, GCN$_{pos}$, GCN$_{fc}$ AND GCN$_{pr}$ DENOTE THE GRAPH CONVOLUTIONAL NETWORKS WITH DIFFERENT WEIGHT MATRICES. SEE THE TEXT FOR MORE DETAILS

Method	SPTS [79]	SPTS+GCN [79]	SPTS+GCN$_{FW}$ [79]	GINs(Ours)
Student Network	AM	GCN$_{pos}$ + AM	GCN$_{pos}$+GCN$_{fc}$ + AM	GCN$_{pos}$+GCN$_{pr}$ + AM
Teacher Network	AM	GCN$_{pos}$ + AM	GCN$_{pos}$+GCN$_{fc}$ + AM	GCN$_{pos}$+GCN$_{pr}$ + AM
Transfer Level	AM	AM	GCN$_{fc}$	GCN$_{pr}$
MCA(%)	90.7	91.2	91.1	**91.7**

TABLE VIII

COMPARISON OF DIFFERENT EXPERIMENTAL SETTINGS ON THE VOLLEYBALL DATASET. DURING TRAINING, BOTH TEACHER AND TEACHER* UTILIZE THE GROUND-TRUTH OF INDIVIDUAL ACTIONS. DURING TESTING, TEACHER EMPLOYED THE GROUND-TRUTH LABEL, WHILE TEACHER* USED PREDICTED LABELS

Method	MCA	MPCA
Teacher	93.3	91.8
Teacher*	74.0	73.4
Student-only (baseline)	90.5	90.8
Student-soft [25]	90.2	90.4
Student-feature	90.1	90.5
Student-attention	91.1	91.7
Student-full (GINs)	**91.7**	**92.3**



(a) Volleyball Dataset (b) CA Dataset

Fig. 8. The confusion matrices of our method on the Volleyball dataset (left) and CA dataset (right).

MMD and DANN methods are weaker than "Source Only" when using "Geo Transfer" pre-process. This is because the "Geo Transfer" operation has changed the structure of the skeleton-based video, which might bring more difficulties for some specific methods like MMD and DANN. Our proposed GINs achieve the best performance of 40.2% and 51.5% and exceeds the state-of-the-arts. We further display the comparison of the confusion matrices using "Geo Transfer" in Fig. 6, where the GINs show its superiority for classifying the action "reading book".

Similar to the previous subsection, we also conducted experiments on MSRDA3D → ORGBD, where MSRDA3D is a smaller dataset with 100 videos and ORGBD with 240 samples is relatively larger. However, the results are still unideal as SBU → NTU. This phenomenon further suggests that the amount of training data in the source domain is essential for this problem and larger dataset is encouraged to be collected as a future work.

D. Evaluation on Group Activity Recognition

1) Experimental Results on the Volleyball Dataset: Table IX presents the experimental results of our GINs model compared with recent approaches for group activity recognition. We observe that our GINs achieve 91.7% and 92.3% accuracy on MCA and MPCA metrics respectively, which outperform the state-of-the-arts. We further show the confusion matrix of our method in Fig. 8(a). As it shows, our model can recognize most action categories well except the activity "rset" ("right-set"), which are more likely to be confused with the activity "rpass" ("right-pass").

Table VII further compares our GINs with the models proposed in [79], which also apply the Teacher-Student framework for the group activity recognition task. We highlight the differences and our advantages as follows: (1) The main contributions of [79] are SPTS and SPTS+GCN. SPTS contains two attention models (AMs) in the Teacher and Student Networks, and the transfer is performed at the attention level. SPTS+GCN builds two graph convolutional networks (GCNs$_{pos}$) upon the Teacher Network and Student Network based on the position information of different people. But still, the transfer is performed at the attention level. In comparison, we perform transfer at the relation level. In this way, the relation information in the Teacher Network can be better preserved. (2) In the "Future Work" section, [79] mentioned the direction to transfer the knowledge across different graphs and make a preliminary study in the "Supplementary Material" part (denoted as "SPTS+GCN$_{FW}$" in Table VII). Specifically, it performs transfer based on other graph convolutional networks (GCNs$_{fc}$), where the n-th row of weight matrix A is simply obtained by the fully-connected layer of the node features. However, this strategy breaks the symmetric characteristic of the graph and causes a slight decrease over SPTS+GCN. To address this, we calculate the inner-product of the embedded node features for GCNs$_{pr}$, which guarantees the symmetric characteristic of the weight matrix during graph construction. (3) Experimental results in Table VII have also demonstrated the effectiveness of our proposed GINs compared with the models in [79]. Moreover, unlike [79] which focuses on the application of group activity analysis, we further show that our GINs can generalize well on other tasks like skeleton-based action recognition.

2) Ablation Study: We further present different settings of our method in Table VIII. On one hand, the Teacher Network,

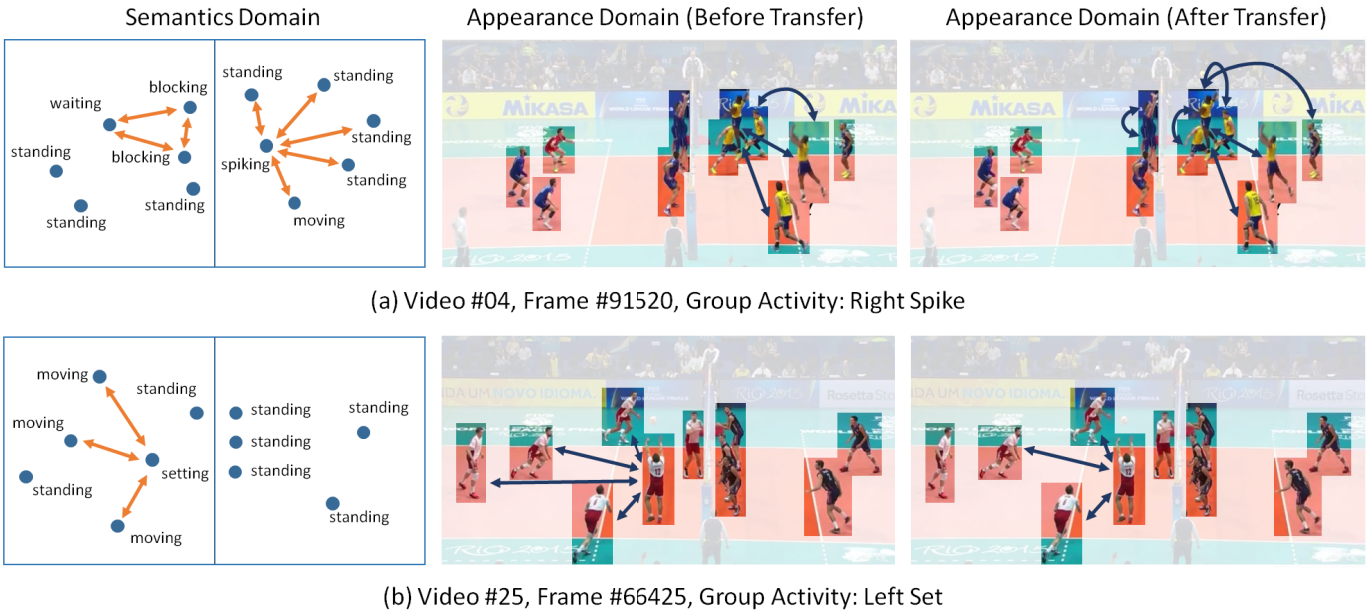


Fig. 9. The visualization results. We present the learned relation knowledge of Teacher Network and Student Network (before and after relation transfer). The relation scores between two players arrange from 0 to 1, as it is the output of a sigmoid function. For brevity, we only draw the relation with the score higher than 0.8.

TABLE IX

COMPARISON OF THE GROUP ACTIVITY RECOGNITION ACCURACY (%) ON THE VOLLEYBALL DATASET

Method	MCA	MPCA	Year
HDTM [52]	86.8	85.8	2016
CERN-2 [53]	83.3	83.6	2017
SSU [54]	90.6	–	2017
SRNN [56]	83.5	–	2018
stagNet [57]	89.3	84.4	2018
RCRG [58]	89.5	–	2018
SPTS [79]	90.7	90.0	2018
SPTS+GCN [79]	91.2	91.4	2019
GINs	**91.7**	**92.3**	

TABLE X

ANALYSIS ON DIFFERENT λ_{GAR} FOR GROUP ACTIVITY RECOGNITION ON THE VOLLEYBALL DATASET

λ_{GAR}	0.01	0.1	1	10
MCA	91.2	91.5	**91.7**	91.3
MPCA	91.9	92.2	**92.3**	92.1

TABLE XI

ANALYSIS ON DIFFERENT FORMULATIONS OF WEIGHT MATRIX A FOR GROUP ACTIVITY RECOGNITION ON THE VOLLEYBALL DATASET

Method	A_{ij}	MCA	MPCA
Sum	$\|\|x_i^{em} + x_j^{em}\|\|_2^2$	90.8	91.1
Relation Module [19]	$g_\theta(concat[x_i^{em}; x_j^{em}])$	91.1	91.7
Gaussian Distance	$exp(-\|\|x_i^{em} - x_j^{em}\|\|_2^2/2)$	91.5	**92.3**
Product	$< x_i^{em}, x_j^{em} >$	**91.7**	**92.3**

which takes ground-truth of individual action during training and testing phase, achieves very promising performance. However, the individual action labels are often not available at test time in a real-world scenario. And unfortunately, the Teacher* network, which takes the predicted individual action labels as inputs during testing, achieves poor results of 74.0% (MCA) and 73.4% (MPCA) respectively. This is because the predicted words (label) of individual action might sometimes be inaccurate, and this will heavily harm the performance of the Teacher* Network, which is sensitive to the inputs. On the other hand, we also conducted experiments based on other transfer methods. Student-feature indicates that we used the feature from the last layer of the Teacher Network to supervise that of the Student Network accordingly. We also evaluated Knowledge Distillation [25] method (Student-soft), which is the pioneering work in knowledge distillation area. For fair comparison with [59] which did not employ GCN models on both Student Network and Teacher Network, we performed mimicking based on the attention scores (Student-attention) but not weight matrices of our GINs. As shown in Table VIII, our GINs obtain 1.2% (MCA) and 1.5% (MPCA) improvements based on the baseline model. Moreover, it consistently outperforms other transfer methods, which has further demonstrated the effectiveness of our proposed scheme for relation transfer.

3) Analysis on the Hyper-Parameter λ_{GAR}: We present of effects on different λ_{GAR} in Table X. Both MCA and MPCA reach the peak when $\lambda_{GAR} = 1$, which indicates that the importance of the transfer loss J_T^{GAR} is equal to that of the classification loss J_O^{GAR}.

4) Analysis on the Weight Matrix: Similar to the skeleton-based action recognition, we further explore different formulations of the weight matrix A for the task of group activity recognition. From the results displayed in Table XI, we observe that the "Product" method yields the best performance (91.7% MCA and 92.3% MPCA), which is slightly higher than the "Distance" method (91.5% MCA and 92.3%

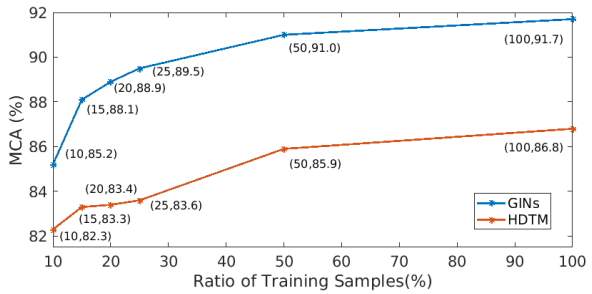


Fig. 10. Study on different ratios of training samples in the Volleyball dataset for group activity recognition.

TABLE XII
COMPARISON OF THE GROUP ACTIVITY RECOGNITION ACCURACY (%) ON THE CA DATASET

Method	MPCA	Year
Cardinality kernel [50]	88.3	2015
CERN-2 [53]	88.3	2017
RMIC [74]	89.4	2017
SBGAR [55]	89.9	2017
MTCAR [80]	90.8	2012
SPTS [79]	95.7	2018
SPTS+GCN [79]	95.8	2019
GINs	**96.0**	

MPCA). Hence, in the following experiments, we also adopt the "Product" method.

5) Analysis on the Influence of Different Amounts of Training Samples: We evaluate our GINs and the baseline method HDTM [52]. For both two methods, we extracted the features based on the same model, which is pre-trained on the individual labels. We then use these features under the ratio of [10%, 15%, 20%, 25%, 50%, 100%] during training. As shown in the following Fig. 10, more training data can bring more benefits for both two methods. Specifically, when the ratio of training samples increases from 10% to 15%, there is a significant gain of GINs, which is also larger than the improvement of HDTM. And when the ratio rises to 50%, our GINs achieve a promising accuracy of 91.0%, which turns to be a relatively saturated result.

6) Visualization: We further present several visualization results on the Volleyball dataset. As shown in Fig. 8(a), after the supervision of Teacher Network, the Student Network discovers the strong relationships between two "blocking" people on the left. Furthermore, on the right side, the "spiking" the player has been linked to more players as he is the key person to identify the "right-spike" group activity. In Fig. 8(b), the Teacher Network only assigns the high relation score to between the "setting" and "moving" players. So the connection between the "setting" and "standing" people has been cut after relation transferring.

7) Experimental Results on the CA Dataset: We finally conduct experiments on the CA dataset and present the experimental result in Table XII. Our GINs obtain the result of 96.0%, which exceeds the other state-of-the-arts. We discover that on the CA dataset the improvements of GINs based on the SPA [59] and SPACI [79] are slight, while on the Volleyball dataset our GINs outperform them clearly. We analyze the reasons as follow: First, the Volleyball dataset is greatly larger than the CA dataset in scale. Since our GINs are trained in a data-driven scheme, it requires more training data to achieves a better result. Second, there are about 12 people in each video in the Volleyball dataset, while the number of people in each sequence of the CA dataset is about 4. Moreover, the relationship of different people in the Volleyball dataset is much more complex than that of the CA dataset. Hence, our GINs can further show its advantage on the previous dataset as it is designed for relation modeling. We further show the confusion matrix in Fig. 8(b), which illustrates that our proposed method can well distinguish the activity of "queueing" and "talking".

V. CONCLUSION

In this paper, we have developed a graph interaction networks (GINs) for transferring relation knowledge in human activity videos. With the proposed method, we have explored two different tasks, including unsupervised skeleton-based action recognition across datasets, and supervised group activity recognition with multi-modal inputs. Both quantitative and qualitative experimental results have demonstrated the effectiveness of our GINs. In the future, it is an interesting direction to employ our method for other tasks related to the graphical model, such as social relationship understanding and human body parsing. Moreover, it is desirable to collect new datasets with larger scale for the task of skeleton-based action recognition under the unsupervised domain adaptation setting.

ACKNOWLEDGMENT

The authors would like to thank X. Wang for his general help on reimplementing the baselines of unsupervised domain adaptation task, and Y. Rao for valuable discussion.

REFERENCES

[1] S. Park, B. X. Nie, and S.-C. Zhu, "Attribute and-or grammar for joint parsing of human pose, parts and attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1555–1569, Jul. 2018.
[2] Z. Wang, T. Chen, J. S. J. Ren, W. Yu, H. Cheng, and L. Lin, "Deep reasoning with knowledge graph for social relationship understanding," in *Proc. IJCAI*, 2018, pp. 1021–1028.
[3] X. Qi, L. Renjie, J. Jiaya, F. Sanja, and U. Raquel, "3D graph neural networks for RGBD semantic segmentation," in *Proc. ICCV*, 2017, pp. 5209–5218.
[4] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proc. CVPR*, 2018, pp. 6857–6866.
[5] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. CVPR*, 2017, pp. 3097–3106.
[6] S. Zhang, Y. Sui, S. Zhao, and L. Zhang, "Graph-regularized structured support vector machine for object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1249–1262, Jun. 2017.
[7] S. Cho and H. Byun, "A space-time graph optimization approach based on maximum cliques for action detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 661–672, Apr. 2016.
[8] W. Chen, L. Cao, X. Chen, and K. Huang, "An equalized global graph model-based approach for multicamera object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 11, pp. 2367–2381, Nov. 2017.
[9] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2527–2542, Dec. 2017.

[10] D. Koller and N. Friedman, *Probabilistic Graphical Models—Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[11] T. Wu and S.-C. Zhu, "A numerical study of the bottom-up and top-down inference processes in and-or graphs," *Int. J. Comput. Vis.*, vol. 93, no. 2, pp. 226–252, Jun. 2011.

[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017, pp. 1–14.

[13] J. Bruna, W. Zaremba, A. Szlam, and Y. Lecun, "Spectral networks and locally connected networks on graphs," in *Proc. ICLR*, 2014, pp. 1–14.

[14] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. NeurIPS*, 2016, pp. 3844–3852.

[15] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. ECCV*, 2018, pp. 413–431.

[16] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI*, 2018, pp. 7444–7452.

[17] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. ECCV*, 2018, pp. 690–706.

[18] P. W. Battaglia *et al.*, "Relational inductive biases, deep learning, and graph networks," *CoRR*, vol. abs/1806.01261, pp. 1–40, Oct. 2018.

[19] A. Santoro *et al.*, "A simple neural network module for relational reasoning," in *Proc. NeurIPS*, 2017, pp. 4967–4976.

[20] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 59:1–59:35, 2016.

[21] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. ICML*, 2017, pp. 2208–2217.

[22] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. NeurIPS*, 2018, pp. 1647–1657.

[23] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. CVPR*, 2018, pp. 3723–3732.

[24] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. CVPR*, 2017, pp. 7130–7138.

[25] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPSW*, 2014, pp. 1–9.

[26] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[27] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[28] X. Chen, L. Li, L. Fei-Fei, and A. Gupta, "Iterative visual reasoning beyond convolutions," in *Proc. CVPR*, 2018, pp. 7239–7248.

[29] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 807–811, Mar. 2018.

[30] C. Cao, C. Lan, Y. Zhang, W. Zeng, H. Lu, and Y. Zhang, "Skeleton-based action recognition with gated convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 11, pp. 3247–3257, Nov. 2019.

[31] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio–temporal visual attention on skeleton image sequences," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 29, no. 8, pp. 2405–2415, Aug. 2019.

[32] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Comput. Vis. Image Understand.*, vol. 158, pp. 85–105, May 2017.

[33] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L.-Y. Duan, and A. K. Chichung, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[34] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Comput. Vis. Image Understand.*, vol. 171, pp. 118–139, Jun. 2018.

[35] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. CVPR*, 2018, pp. 5323–5332.

[36] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. CVPR*, 2019, pp. 3595–3603.

[37] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. CVPR*, 2019, pp. 12026–12035.

[38] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. CVPR*, 2019, pp. 7912–7921.

[39] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. CVPR*, 2019, pp. 1227–1236.

[40] L. Zhang *et al.*, "Unsupervised domain adaptation using robust class-wise matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1339–1349, May 2019.

[41] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, "Hybrid CNN and dictionary-based models for scene recognition and domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1263–1274, Jun. 2017.

[42] J. Li, Y. Wu, and K. Lu, "Structured domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1700–1713, Aug. 2017.

[43] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf, "Correcting sample selection bias by unlabeled data," in *Proc. NeurIPS*, 2006, pp. 601–608.

[44] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[45] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *Proc. ICML*, 2013, pp. 222–230.

[46] X. Wang, L. Li, W. Ye, M. Long, and J. Wang, "Transferable attention for domain adaptation," in *Proc. AAAI*, 2019, pp. 5345–5352.

[47] Z. Ding, S. Li, M. Shao, and Y. Fu, "Graph adaptive knowledge transfer for unsupervised domain adaptation," in *Proc. ECCV*, 2018, pp. 36–52.

[48] Z. Yang *et al.*, "GLoMo: Unsupervised learning of transferable relational graphs," in *NeurIPS*, 2018, pp. 8950–8961.

[49] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1549–1562, Aug. 2012.

[50] H. Hajimirsadeghi, W. Yan, A. Vahdat, and G. Mori, "Visual recognition by counting instances: A multi-instance cardinality potential kernel," in *Proc. CVPR*, 2015, pp. 2596–2605.

[51] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. Zhu, "Joint inference of groups, events and human roles in aerial videos," in *Proc. CVPR*, 2015, pp. 4576–4584.

[52] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proc. CVPR*, 2016, pp. 1971–1980.

[53] T. Shu, S. Todorovic, and S. Zhu, "CERN: Confidence-energy recurrent network for group activity recognition," in *Proc. CVPR*, 2017, pp. 4255–4263.

[54] T. M. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in *Proc. CVPR*, 2017, pp. 3425–3434.

[55] X. Li and M. C. Chuah, "SBGAR: Semantics based group activity recognition," in *Proc. ICCV*, 2017, pp. 2895–2904.

[56] S. Biswas and J. Gall, "Structural recurrent neural network (SRNN) for group activity analysis," in *Proc. WACV*, 2018, pp. 1625–1632.

[57] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. V. Gool, "stagNet: An attentive semantic RNN for group activity recognition," in *Proc. ECCV*, 2018, pp. 104–120.

[58] M. S. Ibrahim and G. Mori, "Hierarchical relational networks for group activity recognition and retrieval," in *Proc. ECCV*, 2018, pp. 742–758.

[59] Y. Tang, Z. Wang, P. Li, J. Lu, M. Yang, and J. Zhou, "Mining semantics-preserving attention for group activity recognition," in *Proc. ACM MM*, 2018, pp. 1283–1291.

[60] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," in *Proc. ICLR*, 2014, pp. 1–13.

[61] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. ICLR*, 2017, pp. 1–13.

[62] T. Chen, I. J. Goodfellow, and J. Shlens, "Net2Net: Accelerating learning via knowledge transfer," in *Proc. ICLR*, 2015, pp. 1–12.

[63] Z. Luo, J. Hsieh, L. Jiang, J. C. Niebles, and L. Fei-Fei, "Graph distillation for action detection with privileged modalities," in *Proc. ECCV*, 2018, pp. 174–192.

[64] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Proc. ECCV*, 2018, pp. 106–121.

[65] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990.

[66] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proc. ECCV*, 2018, pp. 106–121.
[67] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in *Proc. CVPR*, 2016, pp. 4772–4781.
[68] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. CVPR*, 2016, pp. 1010–1019.
[69] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. CVPRW*, 2012, pp. 28–35.
[70] G. Yu, Z. Liu, and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *Proc. ACCV*, 2014, pp. 50–65.
[71] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. CVPR*, 2012, pp. 1290–1297.
[72] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "Hierarchical deep temporal models for group activity recognition," *CoRR*, vol. abs/1607.02643, pp. 1–14, Jul. 2016.
[73] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *Proc. ICCVW*, 2009, pp. 1282–1289.
[74] M. Wang, B. Ni, and X. Yang, "Recurrent modeling of interaction context for collective activity recognition," in *Proc. CVPR*, 2017, pp. 7408–7416.
[75] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *Proc. IJCAI*, 2018, pp. 786–792.
[76] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. CVPR*, 2017, pp. 1647–1655.
[77] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. ICML*, 2015, pp. 97–105.
[78] X. Chen, S. Wang, and M. L. J. Wang, "Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation," in *Proc. ICML*, 2019, pp. 4013–4022.
[79] Y. Tang, J. Lu, Z. Wang, M. Yang, and J. Zhou, "Learning semantics-preserving attention and contextual interaction for group activity recognition," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4997–5012, Oct. 2019.
[80] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *Proc. ECCV*, 2012, pp. 215–230.

Yansong Tang received the B.S. degree from the Department of Automation, Tsinghua University, China, in 2015, where he is currently pursuing the Ph.D. degree with the Department of Automation. His current research interest lies in human behavior understanding for computer vision. He has authored six scientific articles in this area, where four articles are published in top journals and conferences, including IEEE TIP, CVPR, and ACM MM. He serves as a regular reviewer member for a number of journals and conferences, e.g., TPAMI, TIP, TCSVT, CVPR, ICCV, and AAAI. He has obtained the National Scholarship of Tsinghua University in 2018.

Yi Wei received the B.S. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Automation. His research interest concentrates on computer vision, especially 3D vision.

Xumin Yu is currently pursuing the B.S. degree with the Department of Electronic Engineering, Tsinghua University, Beijing, China. His research interests are in computer vision, especially action recognition and action detection.

Jiwen Lu (Senior Member, IEEE) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He has authored or coauthored over 240 scientific articles in these areas, where more than 70 of them are the IEEE TRANSACTIONS articles (including 14 T-PAMI articles) and more than 50 of them are CVPR/ICCV/ECCV articles. He was/is a member of the Multimedia Signal Processing Technical Committee and the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society, and a member of the Multimedia Systems and Applications Technical Committee and the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He was a recipient of the National 1000 Young Talents Program of China in 2015, and the National Science Fund of China for Excellent Young Scholars in 2018, respectively. He is also the Program Co-Chair of IEEE ICME'2020 and AVSS'2020. He serves the Co-Editor-of-Chief of the *Pattern Recognition Letters*, an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE, and *Pattern Recognition*.

Jie Zhou (Senior Member, IEEE) received the B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. From then to 1997, he served as a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. In recent years, he has authored more than 100 articles in peer-reviewed journals and conferences. Among them, more than 30 articles have been published in top journals and conferences such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and CVPR. He received the National Outstanding Youth Foundation of China Award. He is an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and two other journals.