

Cross-Graph Domain Adaptation for Skeleton-based Human Action Recognition

Haitao Tian

School of Electrical Engineering
and Computer Science
University of Ottawa
Ottawa, ON, Canada
htian026@uottawa.ca

James Dickens

School of Electrical Engineering
and Computer Science
University of Ottawa
Ottawa, ON, Canada
jdick088@uottawa.ca

Pierre Payeur

School of Electrical Engineering
and Computer Science
University of Ottawa
Ottawa, ON, Canada
ppayeur@uottawa.ca

Abstract— Recent research on human action recognition is largely facilitated by skeletal data, a compact graph representation composed of key joints of the human skeleton that is efficiently extracted by body tracking systems and that offers the merit of being robust to environmental variations. However, the skeleton resolution and joint connectivity of the extracted skeletons may vary with sensor devices, which results in different skeleton graph representations on collected data. This paper investigates a cross skeleton graph domain adaptation approach where a skeleton action recognition model is trained upon a source skeletal data domain but is expected to adapt onto a target domain configured with a different skeleton graph. It proposes an adversarial learning framework where a generation space is developed on which the model learns valid skeletal action knowledge from the source graph domain. Interaction with an embedded discrimination space is employed to extract heterogeneous graph features from the target domain. Optimization of the generation space and the discrimination space is realized alternatively under adversarial learning which guarantees action-aware and domain-agnostic skeletal knowledge, thus forming a joint human action recognition model effectively functioning on both graph domains. In experiments, the paper evaluates the proposed method by incorporating graph convolutional networks into two skeleton action recognition benchmarks, NTU-RGB+D and Northwestern-UCLA, where comparisons are conducted to demonstrate the effectiveness of the proposed approach. Code will be available at <https://github.com/tht106/CrossGraphDA>.

Keywords: *domain adaptation; skeleton data; human action recognition; unsupervised learning*

I. INTRODUCTION

Human skeletal data has been widely utilized to address a variety of human activity analysis tasks, such as action recognition [1], interactive gaming [2], and video surveillance [3], where human movements are interpreted by the trajectories of human skeleton joints captured by motion capture (MoCap) systems in a pre-configured environment. However, skeletal data representations may vary significantly in graph configurations, involving different skeleton resolution (i.e., number of joints) and joint connectivity (i.e., relative position of joints being tracked), depending on the MoCap systems used for data collection. For instance, the different generations of Microsoft Kinect rely on different

skeleton body tracking algorithms and sensing modalities [4], resulting in a high degree of skeletal representation discrepancy in joint permutations, numbers, and positions [5], as depicted in Fig. 1.

The discrepancy in skeleton graph leads to domain shift which remains a challenge for research on skeleton-based human action recognition. First, in many practical scenarios, the development of skeleton-based human action recognition models leverages copious training data available from a label-rich source environment equipped with a well-calibrated MoCap system, to train an optimal model transferrable to a target application domain in the real world. However, when the target domain involves utilizing a different MoCap system whose body tracking skeleton graph configuration differs from that of the source domain, the model transfer tends to achieve suboptimal performance and the model's scalability is precluded in practical scenarios. Second, a common practice to deal with model transfer is to retrain the model with sufficient data samples collected with the MoCap system in the target application domain. However, data collection and annotation are extremely labor-intensive and time-consuming. Moreover, the resulting retrained model becomes dependent to a single domain, leading to poor generalization performance.

This paper proposes an unsupervised domain adaptation framework, which allows a joint skeletal human action recognition model to learn from heterogeneous graph configurations. The unsupervised learning method treats the skeletal action knowledge and domain graph features separately, made possible by the development of a generation space and a discrimination space. In the generation space, advantage is taken of sufficient well-labeled data from the source domain to supervise a graph convolutional network (GCN) to learn *action aware semantics*. In the discrimination space, a small number of unlabeled skeletal data samples facilitate an adversarial learning paradigm in which the GCN integrates *domain-agnostic graph knowledge*. In general, the interaction between the generation space and the discrimination space transfers learnt knowledge from a data-rich source domain and a data-scarce target domain, resulting in a joint model for both the source and the target domains.

The contribution of the unsupervised domain adaptation approach is twofold. First, it provides the first investigation on skeletal graph discrepancy in human action recognition.

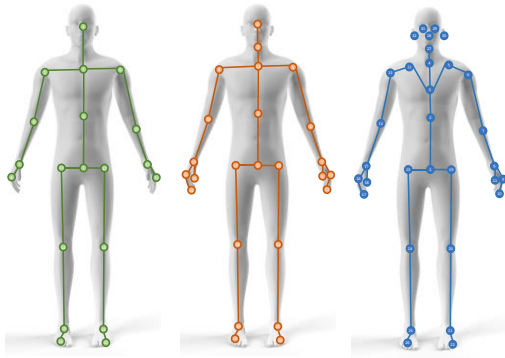


Figure 1. Skeletal joints detected by different generations of Microsoft Kinect. From left to right: Microsoft Kinect V1, Microsoft Kinect V2, and Microsoft Azure Kinect.

Second, it introduces a novel paradigm that allows to elaborate a single human action recognition model that functions effectively on different skeleton graphs, which will improve the scalability of existing skeletal action recognition models in varying real-world environments.

II. RELATED WORK

A. Human Action Recognition

Human action recognition [1] addresses the perception by computer vision of a variety of human actions. Recent research on human action recognition is largely facilitated by the encoding of human skeletons, a compact spatial-temporal sequence of 3D features extracted by processing sensors output while imaging human subjects and offering the merit of being robust to variations in the environment.

With the goal of further boosting state-of-the-art performance, GCNs have dominated recent research [6] [7] [8] on skeletal data-based human action recognition by involving skeletal data into end-to-end supervised learning. AS-GCN [6] introduces directed graph neural networks to model the temporal dependencies among skeletal joints. ST-GCN [7] employs spatial-temporal graph convolutional networks to capture both spatial and temporal information in skeletal data, achieving competitive results in action recognition. CTR-GCN [8] proposes a relations modeling framework for a GCN with the goal of interpreting relationships between joints.

Domain shift is observed in computer vision as a result of data distribution discrepancy emerging from domain changes [9][10]. In the context of skeletal data-based human action recognition, such a data representation shift leads to heterogeneous skeleton graph representations between the source domain and the target domain, in which a GCN model is prone to overfit to the source skeleton graph representation and suffers from the adaptation to a target domain. Consequently, the scalability of GCNs is significantly limited for practical human action recognition tasks.

B. Domain Adaptation

Domain adaptation (DA) [11] [12] has been successfully leveraged to approach the generalization issue of machine learning solutions in computer vision, by alleviating discrepancies between distinct domain distributions to improve the performance of generalization in unsupervised manners. For instance, DANN [13] introduces an effective domain adaptation procedure which aligns feature representations across domains by using a domain classifier to enforce domain invariance. Vu et al. [14] utilize entropy minimization to encourage unambiguous cluster assignments to maintain the target domain's clusters separable in image semantic segmentation.

However, research on domain adaptation in skeletal data-based human action recognition is quite limited. Mitsuzumi et al. [15] propose a data-efficient domain adaptation approach to learning cross-subject skeletal action recognition by utilizing phase randomization-based data augmentation. Skeleton-CutMix [16] proposes a simple and effective skeleton augmentation framework for supervised domain adaptation which hallucinates new skeleton representations by using pairs of skeletons from the source and target domains. But the existing research generally overlooks the cross-skeleton graph domain adaptation issue, and there is not an effective method proposed to address it.

C. Adversarial Learning

Recently, adversarial learning has been combined with domain adaptation motivated by the fact that generative adversarial networks (GAN) [17] excel at extracting essential semantic features that resemble a target distribution. UNIT [18] focuses on unsupervised image-to-image translation for domain adaptation. It employs a shared latent space and adversarial training to achieve domain-invariant representations. Hoffman *et al.* [19] embed a discrimination network (discriminator) into FCN-based segmentation networks as an adaptation component. By imposing the DA operation, the segmentation network not only learns discriminative representations but also invariant encodings from different domains. Fukushi *et al.* [20] propose a few-shot generative model based upon cross-domain regularization and entropy regularization, which is effective for transferring the diversity of the motions contained in the source to a target domain.

III. METHODOLOGY

This section details the methodology for the proposed unsupervised domain adaptation in skeleton-based human action recognition. It begins with mathematical preliminaries about supervised skeletal action recognition, upon which it details the framework for unsupervised cross-graph domain adaptation.

A. Preliminaries

1) **Skeletal action recognition.** Human action movements can be composed of a sequence of skeletons, each of which consisting of a number of key body joints, and interpreting the trajectories of human movement over time. In the context of graph convolution, skeletons can be naturally treated as a

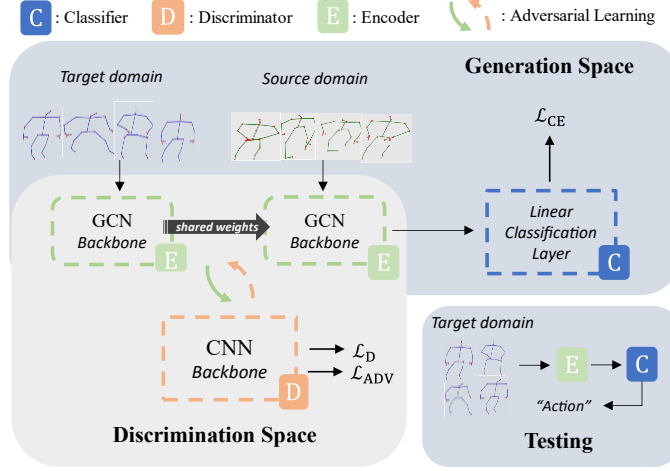


Figure 2. A conceptual overview of the proposed adaptation network for human action recognition.

graph-structured data form in which skeletal joints compose graph nodes, connected by skeletal bones, namely, node edges. A skeletal action recognition model considers a skeletal sequence input, $(\mathbf{X}, \mathbf{A}, \mathbf{y})$, where the $\mathbf{X} \in \mathbb{R}^{T \times V \times 3}$ is the sequence composed of T frames of skeletons, each of which is configured by the graph $\mathbf{A} \in \{0, 1\}^{V \times V}$ denoting the node connectivity of V skeletal joints; $\mathbf{y} \in \mathbb{R}^L$ is the corresponding action label of \mathbf{X} from L action categories. In graph convolution, the GCN model aggregates action semantic representations by involving the skeletal sequence \mathbf{X} into end-to-end supervised learning with the sequence-wise annotation \mathbf{y} . It employs an encoder \mathbf{E} to perform graph convolution on the input which generates high-dimension feature encodings, $\mathbf{E}(\mathbf{X}, \mathbf{A}; \boldsymbol{\theta}_E)$, where $\boldsymbol{\theta}_E$ is the learnable network weights of \mathbf{E} . A classifier \mathbf{C} , parameterized by $\boldsymbol{\theta}_C$, then accepts the feature encodings and outputs action classifications on a soft-max layer. The optimization of the GCN, i.e., $\mathbf{C}(\mathbf{E}(\cdot))$ can be solved by typical neural network optimization schemes, such as stochastic gradient descent (SGD), with a cross-entropy loss:

$$L_{CE} = -\mathbf{y} \cdot \log[\mathbf{C}(\mathbf{E}(\mathbf{X}, \mathbf{A}; \boldsymbol{\theta}_E); \boldsymbol{\theta}_C)] \quad (1)$$

2) **The cross-graph shift.** Despite the significant effectiveness of graph convolution on skeletal data, GCNs still demonstrate intrinsic limitations when learning from heterogenous skeleton graphs. Specifically, considering a labelled source domain, $\mathcal{D}_S = \{\mathbf{X}^s, \mathbf{A}^s, \mathbf{y}^s\}$, where $\mathbf{A}^s \in \mathbb{R}^{V_s \times V_s}$ and $\mathbf{y}^s \in \mathbb{R}^L$, and an unlabeled target domain, $\mathcal{D}_T = \{\mathbf{X}^t, \mathbf{A}^t\}$, where $\mathbf{A}^t \in \mathbb{R}^{V_t \times V_t} \neq \mathbf{A}^s$, the graph convolution of the encoder \mathbf{E} on the source domain can be implemented by operating conventional 2D convolution on input \mathbf{X}^s , weighted by the adjacent matrix \mathbf{A}^s , which can be formulated as: $\mathbf{E}(\mathbf{X}^s, \mathbf{A}^s; \boldsymbol{\theta}_E) = \mathbf{E}_{\boldsymbol{\theta}_E}(\mathbf{X}^s) \cdot \mathbf{A}^s$, where $\mathbf{E}_{\boldsymbol{\theta}_E}$ is the optimal encoder parameterized by $\boldsymbol{\theta}_E$ obtained in (1), and \cdot denotes the tensor product. Clearly, the knowledge distilled by $\mathbf{E}_{\boldsymbol{\theta}_E}$ on

the source domain resides on the graph \mathbf{A}^s . In this way, given a target domain sequence \mathbf{X}^t associated with the graph \mathbf{A}^t , the cross graph discrepancy between \mathbf{A}^s and \mathbf{A}^t prevents the knowledge of $\mathbf{E}_{\boldsymbol{\theta}_E}$ from properly transferring to the target domain by formulating $\mathbf{E}_{\boldsymbol{\theta}_E}(\mathbf{X}^t) \cdot \mathbf{A}^t$.

B. Cross-graph Domain Adaptation

The goal of cross-graph domain adaptation is to learn a joint GCN model capable of functioning effectively on two skeleton domains, associated with the graphs, \mathbf{A}^s and \mathbf{A}^t , respectively. However, given the absence of label statistics in the target domain, it is impossible to fine-tune an effective model with supervision in the target domain. To this end, an adversarial learning framework is proposed which incorporates a generation space to learn *action aware semantics* with a traditional GCN model while embedding a discrimination space to enforce the generation space to exploit *graph invariant representations* at the same time. The adversarial learning strategy is formulated in three successive stages, as detailed below. Fig. 2 depicts the overall concept.

1) **In the generation space**, the framework takes the GCN model, i.e., $\mathbf{C}(\mathbf{E}(\cdot))$, as the base structure to learn action aware semantics. Optimization of the encoder parameters, $\boldsymbol{\theta}_E$ is realized by supervised learning with the labeled source domain data by solving a cross-entropy loss:

$$L_{CE}(\mathbf{C}, \mathbf{E}) = -\frac{1}{|\hat{\mathcal{X}}_S|} \sum_{(\mathbf{X}^s, \mathbf{y}^s) \in \hat{\mathcal{X}}_S} [\mathbf{y}^s \log \mathbf{C}(\mathbf{E}(\mathbf{X}^s, \mathbf{A}^s; \boldsymbol{\theta}_E); \boldsymbol{\theta}_C)] \quad (2)$$

where $\hat{\mathcal{X}}_S$ denotes a batch of data samples from the source domain \mathcal{D}_S .

2) **In the discrimination space**, an auxiliary network, discriminator \mathbf{D} , is employed to classify the graph feature encodings from different domains. Specifically, the discrimination space involves batch data from the two domains, $\hat{\mathcal{X}}_S$ and $\hat{\mathcal{X}}_T$, as a dual input. The discriminator \mathbf{D} , which is embedded on the end of the encoder of the base

network, accepts the feature encodings, $\mathbf{E}_{\theta_E}(\hat{\mathcal{X}}_s, \mathbf{A}^s)$ and $\mathbf{E}_{\theta_E}(\hat{\mathcal{X}}_t, \mathbf{A}^t)$, that are aggregated on the generation space (where \mathbf{E}_{θ_E} denotes the encoder \mathbf{E} parameterized by θ_E according to (2)), and performs binary classification upon their original domain labels. The development of \mathbf{D} effectively circumvents the absence of label information in the target domain while taking the advantage of unlabeled skeleton samples for domain adaptation. The optimization on \mathbf{D} is solved by a binary cross-entropy loss:

$$\mathcal{L}_D(\mathbf{D}) = \log[1 - \mathbf{D}(\mathbf{E}_{\theta_E}(\hat{\mathcal{X}}_s, \mathbf{A}^s); \theta_D)] + \log[\mathbf{D}(\mathbf{E}_{\theta_E}(\hat{\mathcal{X}}_t, \mathbf{A}^t); \theta_D)] \quad (3)$$

where θ_D is the learnable weights of the discriminator. Note that the network weights of the encoder θ_E are fixed at this stage and only contribute to the forward propagation. The optimization based on (3) aims to endow the discriminator \mathbf{D} with the ability to discriminate the heterogeneous graph representations upon different domains. The structure of \mathbf{D} will be detailed in Figure 3 and Section IV. B.

3) **Adversarial optimization.** Lastly, the encoder \mathbf{E} is optimized by the loss function in (4) with supervision of the discriminator \mathbf{D}_{θ_D} , parameterized by θ_D , to produce graph-agnostic skeletal representations.

$$\mathcal{L}_{\text{ADV}}(\mathbf{E}) = \log[\mathbf{D}_{\theta_D}(\mathbf{E}(\hat{\mathcal{X}}_t, \mathbf{A}^t); \theta_D)] \quad (4)$$

In (4), the network weights of the discriminator θ_D are fixed and only contribute to the forward propagation.

The network components, classifier \mathbf{C} , encoder \mathbf{E} , and discriminator \mathbf{D} , are updated alternatively by minmax adversarial optimization formulated in (5):

$$\mathbf{C}_{\theta_C}, \mathbf{D}_{\theta_D}, \mathbf{E}_{\theta_E} = \arg \min_{\mathbf{C}, \mathbf{D}} \max_{\mathbf{E}} \mathcal{L}(\mathbf{C}, \mathbf{D}, \mathbf{E}) \quad (5)$$

$$\mathcal{L}(\mathbf{C}, \mathbf{D}, \mathbf{E}) = L_{\text{CE}}(\mathbf{C}, \mathbf{E}) + \mathcal{L}_D(\mathbf{D}) + \lambda \mathcal{L}_{\text{ADV}}(\mathbf{E}) \quad (6)$$

where λ denotes the hyperparameter adapting the weights of adversarial optimization.

After adaptation, the learnt network $\mathbf{C}_{\theta_C}(\mathbf{E}_{\theta_E}(\cdot))$ is utilized to infer testing samples from the target domain by formulating: $\mathbf{C}_{\theta_C}(\mathbf{E}_{\theta_E}(\mathcal{X}^t, \mathbf{A}^t))$.

IV. EXPERIMENTS

This section experimentally evaluates the effectiveness of the proposed unsupervised adaptation framework by investigating cross-graph skeletal action recognition tasks that leverage two public datasets.

A. Datasets

NTU RGB+D [21] is a large-scale human skeleton action dataset composed of 56,880 samples covering 60 human daily actions. The skeletal data set is recorded in indoor scenes with three *Microsoft Kinect V2* cameras mounted in different setups. Each camera provides a skeleton sequence encoded over 25 joints. The samples from the ‘‘cross subject’’ training set [21] are used for adaptation training.

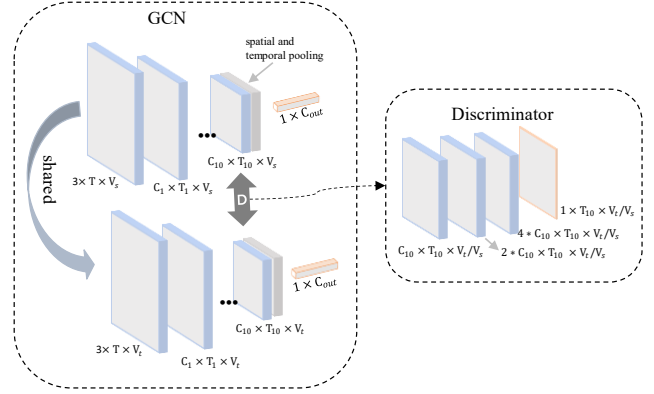


Figure 3. Illustration of network architectures for the proposed adaptation framework. First, the input layer of the GCN backbone accepts domain-specific skeletal data with the size of $3 \times T \times V_s$ and $3 \times T \times V_t$, followed by ten graph convolution blocks configured with the channel number $\{C_1, \dots, C_{10}\}$. \mathbf{D} accepts high-dimensional feature encodings generated in the tenth block with the size of $C_{10} \times T_{10} \times V_s$ and $C_{10} \times T_{10} \times V_t$ and forward propagates through three normal convolutional layers with stride 2 and kernel size 4. Each of the first two layers is followed by a leaky ReLU. The channel number of \mathbf{D} is $\{2C_{10}, 4C_{10}, 1\}$.

Northwestern-UCLA [22] is a relatively small skeletal dataset composed of 1494 short sequences of skeletons covering 10 human actions. This dataset is recorded by one *Microsoft Kinect V1* which provides skeleton representations encoded over 20 joints. The discrepancy on the number of joints and variations in camera setups with respect to **NTU RGB+D** acquisition configuration provides an effective evaluation scenario for domain adaptation. The skeleton coordinates of both datasets are normalized into $[-1, 1]$ during training and testing.

B. Implementation

The proposed network is deployed using PyTorch on a NVIDIA 3090 GPU. It uses CTR GCN [8] as the GCN backbone in the generation space, where encoder \mathbf{E} is composed of ten blocks of graph convolution (refer to [8] for more details). Discriminator \mathbf{D} is embedded on the final block of encoder \mathbf{E} , which is composed of a four-layer CNN as detailed in Fig. 3. The GCN backbone, CTR GCN, is optimized by using SGD, with momentum as 0.9, weight decay $4e-10$, and learning rate 0.1. The discriminator \mathbf{D} is updated by the Adam optimizer with the learning rate as $1e-3$ and betas $\{0.9, 0.99\}$. The hyperparameter λ in (6) is set as 0.08. The performance of the proposed method is evaluated under the metric of Top-1 Accuracy [7].

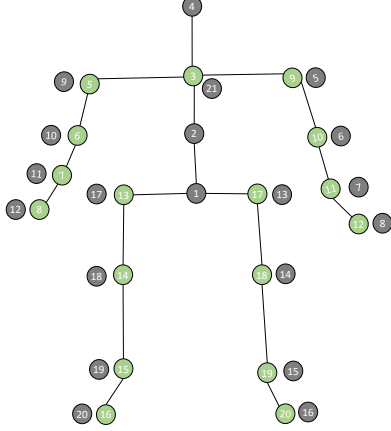


Figure 4. Joint remapping from Microsoft Kinect V2 to Microsoft Kinect V1. The joints connected by lines compose the remapped 20-joint skeleton under the same skeleton configuration (joint number and permutation) used by Microsoft Kinect V1. Each green joint on the skeleton is remapped from a nearby gray joint whose coordinates are indexed by the number indicated in the gray circle. For instance, the third joint on the remapped skeleton has the location information of the twenty-first joint of the source skeleton graph. Joints 1, 2 and 4 are the same under both mappings. Furthermore, five joints, 3-neck, 22-tip of the left hand, 23-left thumb, 24-tip of the right hand, and 25-right thumb, of the source skeleton graph are dropped after remapping.

C. Results

Two cross-graph skeletal action recognition tasks are considered to validate the performance of the proposed framework.

1) Test case on remapped skeleton representation

In the first cross-graph domain adaptation experiment, the original training dataset of **NTU RGB+D** (recorded with a *Microsoft Kinect V2*) is considered as the source domain. The skeletons from the **NTU RGB+D** are then remapped from 25 to 20 joints as per the matching rules detailed in Fig. 4, to mimic a target domain that is recorded by *Microsoft Kinect V1*. Experimental results on this cross-graph domain adaptation task are summarized in Table 1. First, a source-only model is trained with exclusive full supervision using the original training dataset composed of 25-joint skeletons with annotations to serve as a comparative. The source-only model reaches 83.3% on the test dataset supported by 25-joint encodings but achieves only 62.9% on the skeletons remapped to 20 joints. It demonstrates that a reduction on the skeleton resolution results in suboptimal performance for the GCN model. To overcome the issue, the model is adapted with the proposed adversarial learning scheme. In the second row of Table 1 it is observed that a significant performance improvement is achieved on the downsized 20-joint skeleton encodings (from 62.9% to 80.1%). At the same time, the adapted model exhibits stable performance when processing

Table 1. Comparative performance of the proposed method under the metric of Top-1 Accuracy.

NTU RGB+D 25 joints remapped to 20 joints		
Model	25 joints	20 joints
Source only	83.3%	62.9%
With adaptation	83.7%	80.1%

skeletons with 25 joints, as it was trained on. These results demonstrate that the proposed method successfully learns a joint model across skeleton graphs with different encodings and inherent resolution.

2) Test case on different skeleton encodings

The second experiment studies a more practical but challenging cross-graph action recognition scenario. Specifically, it considers **NTU RGB+D** with 25-joint skeleton encodings as the source domain and **Northwestern UCLA** with 20-joint encodings as the target domain. The skeleton sequences under the common 7 actions (i.e., "drop", "pickup", "throw", "sitting down", "standing up", "wear jacket", and "take off jacket") between the two domains are used for cross-graph domain adaptation training and testing. The overall experimental results over these 7 actions are detailed in Table 2. First, the "Source-only" model (trained on data (4672 samples) from **NTU RGB+D**) reaches 71.6% average Top-1 Accuracy on the 7 common actions when tested on 370 skeleton samples from **Northwestern UCLA** with fewer joints. After applying domain adaptation with 799 unlabeled training samples from the target domain, the proposed method gains an average performance boost of 14.7% on the target test dataset. Meanwhile, the proposed model with adaptation also outperforms the source-only model in a vast majority of action categories, with only very minor discrepancies in the weaker class of "sitting down". To explain the upper limit at 86.3% in Top-1 Accuracy, it is conjectured that the variations between **NTU RGB+D** and **Northwestern UCLA** in data collection settings, such as camera setups and environmental configurations, are significant enough and contribute complex data discrepancy in skeletal representations, presenting a significant challenge to any methods for skeleton-based human action recognition.

D. Ablation Study

1) Domain adaptation with fewer target domain samples

This work performs unsupervised learning for cross skeleton graph-based human action recognition by only using unlabeled target domain samples. Nevertheless, the development of the proposed domain adaptation framework still imposes data collection in the target domain. Further experimental investigation examines the principle that drives adaptation effectiveness with lower amounts of target domain samples to circumvent constraints on data usage. Specifically, it conducts an experiment where the GCN model is learnt with the same domain adaptation implementation detailed in the task "**NTU RGB+D** to **Northwestern UCLA**" but altering

Table 2. Experimental results of the adaptation task “NTU RGB+D to Northwestern UCLA”.

NTU-RGB+D to NW-UCLA	Drop	Pickup	Throw	Sitting Down	Standing Up	Wear Jacket	Take off Jacket	Top-1 Accuracy
Source only	24.4 %	74.2 %	33.3 %	97.8 %	95.1 %	83.7 %	92.7 %	71.6 %
With adaptation	66.7 %	82.8 %	73.0 %	97.6 %	99.8 %	92.3 %	94.4 %	86.3 %

Table 3. Performance variation related to different proportions of the target data samples used for domain adaptation. (100% denotes that 799 unlabeled samples from the target domain are used)

Percentage of data use	0% (source only)	5%	20%	30%	50%	70%	100%
Top-1 Accuracy	71.6 %	79.1%	82.5%	84.6%	85.1%	85.9%	86.3 %

Table 4. Experimental comparison of the discriminator implemented with different network structures. Memory denotes the allocated GPU memory resources during training upon the entire framework. Time/epoch denotes the time cost per epoch for training the framework on the source domain dataset.

Discriminator	Top-1 Accuracy	Memory	Time/epoch
CNN	86.3 %	4.1GB	55.7s
MLP	77.5%	3.5GB	49.4s

the ratios (varying from 0% to 100%) of samples used for training that were randomly selected from the target domain. Note that the number of test samples remains the same (i.e., 370) in the experiment. Overall experimental results are presented in Table 3. First, when no (0%) target domain samples are used, the model degrades into a source only model. Second, the model performance tends to increase monotonically along with the ratio of the target domain samples involved in adaptation. It is worth noting that even while using a very small number (e.g., 5%) of data samples from the target domain, the model will achieve a significant performance gain of 7.5% from the source only model (71.6% in Table 3), suggesting the capability of the proposed method for learning critical skeletal action semantics from heterogeneous domains.

2) Efficient implementation of the discriminator

The experimental results in Tables 1 and 2 suggest that the proposed method guarantees a performance boost to the target domain. Nevertheless, the introduction of a CNN-structured discriminator **D** inevitably imposes computational overhead to the conventional GCN based training pipeline. To circumvent the costs of the computational overhead, this research also examines an efficient implementation of **D** for domain adaptation. Specifically, it adopts a multi-layer perceptron (MLP) to fulfill the discrimination space. The MLP is composed of a spatial pooling layer and a temporal

Table 5. Performance variation related to hyperparameter λ .

λ	0	0.01	0.05	0.08	0.1	0.5
Top-1 Accuracy	71.6%	77.9%	85.8%	86.3%	86.0%	77.9%

pooling layer, followed by two linear layers equipped with a ReLU activation after the first linear layer. MLP aims to make binary classification upon the feature encodings generated by the encoder **E** while imposing a lower computational cost than a CNN network.

The experimental results are summarized in Table 4. First, the MLP-based discriminator reaches inferior performance compared to the CNN implementation in the task “NTU RGB+D to Northwestern UCLA”. It is conjectured that the dimension reduction imposed by the first two pooling layers could lead to information loss in spatial and temporal dimensions, compromising the eventual effectiveness of MLP in the discrimination space. On the other hand, even though the CNN-based discriminator, as proposed in Section III, preserves better adaptation performance, the convolution operations in the discrimination space involve an exponential number of network weights, leading to a heavier computation cost compared to MLP (e.g., extra 0.6GB memory use and 6.3s per epoch of training). Consequently, the experimental results demonstrated in Table 4 would suggest a practical trade-off between the model performance and computational cost in the implementation of the discriminator.

E. Parameter Learning

The adversarial interaction between the generation space and the discrimination space is crucial to the proposed method and is controlled by the hyperparameter λ , as formulated in Eq. (6). Setting an appropriate value of λ not only involves sufficient adversarial supervision for domain adaptation learning, but also alleviates the over-adaptation problem in the target domain. To this end, experimental investigation analyzes the choice of λ by conducting an experiment where it learns a GCN model with the same domain adaptation

implementation detailed in the task “**NTU RGB+D to Northwestern UCLA**” but altering λ over the range [0.01 0.5]. Experimental results are reported in Table 5. When $\lambda = 0$, the adversarial learning degrades into a traditional fully supervised GCN training upon the source domain where the resulting model is equivalent to a source-only model, which achieved 71.6% on the task “**NTU RGB+D to Northwestern UCLA**” in Table 2. Otherwise, the adaptation performance is clearly affected as λ varies with an eventual degradation in performance when the hyperparameter places too much weight on adversarial learning. Overall, it is observed that the best performance is achieved when λ is around 0.08.

CONCLUSIONS

This research investigates the scalability of existing human action recognition models as they are significantly limited by discrepancies in the skeletal representation that surface when the acquired data from a target domain is encoded in a different skeleton graph compared to that of the source training domain. This paper proposes the first domain adaptation framework for cross skeleton graph-based human action recognition. The proposed method utilizes adversarial learning to enable a model learning action-aware representations from a labeled source dataset, while extracting graph invariant representations from an unlabeled target domain. In its implementation, it employs a generation space and a discrimination space where a GCN model aggregates skeleton joint correlation from heterogeneous skeletal graphs while a discriminator predicts overall domain labels with respect to the respective domain aggregations encoded by the generation space. Experiments are conducted on two cross-graph domain adaptation tasks to demonstrate the validity of the proposed adaptation method and evaluate its effectiveness.

In future work, the aim will be to extend the proposed method to multi-variation domain adaptation. For instance, it will consider the skeletal data shift in both extrinsic (environment-specific) and intrinsic (sensor-specific) variations derived from different sources and devices, to attempt to transpose the proposed method for formulating more adaptive skeletal data-based human action recognition models.

ACKNOWLEDGMENTS

This research was supported in part by MITACS Accelerate and NSERC Discovery grants. The authors also acknowledge the collaboration of Spectronix Inc.

REFERENCES

- [1] Y. Kong and Y. Fu, “Human action recognition and prediction: A survey,” *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.
- [2] B. Lange, C.-Y. Chang, E. Suma, B. Newman, A. S. Rizzo, and M. Bolas, “Development and evaluation of low cost game-based balance rehabilitation tool using the Microsoft Kinect sensor,” in 2011 annual international conference of the IEEE engineering in medicine and biology society, 2011, pp. 1831–1834.
- [3] W. Niu, J. Long, D. Han, and Y.-F. Wang, “Human activity detection and recognition for video surveillance,” in 2004 IEEE international conference on multimedia and expo (ICME)(IEEE Cat. No. 04TH8763), 2004, vol. 1, pp. 719–722.
- [4] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, “Deep learning-based human pose estimation: A survey,” *ACM Computing Surveys*, 56, no. 1 (2023): 1–37.
- [5] Q. Wang, G. Kurillo, F. Ofli, and R. Bajcsy, “Evaluation of pose tracking accuracy in the first and second generations of microsoft kinect,” in 2015 international conference on healthcare informatics, 2015, pp. 380–389.
- [6] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, “Actional-structural graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3595–3603.
- [7] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32, no. 1.
- [8] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, “Channel-wise topology refinement graph convolution for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13359–13368.
- [9] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, “A theory of learning from different domains,” *Machine learning*, vol. 79, pp. 151–175, 2010.
- [10] S. Ma, Z. Yuan, Q. Wu, Y. Huang, X. Hu, C. H. Leung, D. Wang, and Z. Huang, “Deep into The Domain Shift: Transfer learning through dependence regularization,” *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [11] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, “Analysis of representations for domain adaptation,” *Advances in neural information processing systems*, 19, 2006.
- [12] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [13] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, “Domain-adversarial training of neural networks,” *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.
- [14] T.H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, “Advent: adversarial entropy minimization for domain adaptation in semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [15] Y. Mitsuzumi, G. Irie, A. Kimura, and A. Nakazawa, “Phase Randomization: A data augmentation for domain adaptation in human action recognition,” *Pattern Recognition* 146 (2024): 110051.
- [16] H. Liu, Y. Liu, T.-J. Mu, X. Huang, and S.-M. Hu, “Skeleton-CutMix: Mixing Up Skeleton with Probabilistic Bone Exchange for Supervised Domain Adaptation,” *IEEE Transactions on Image Processing*, 2023.
- [17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
- [18] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” *Advances in neural information processing systems*, vol. 30, 2017.
- [19] J. Hoffman, D. Wang, F. Yu, and T. Darrell, “Fcns in the wild: Pixel-level adversarial and constraint-based adaptation,” *arXiv preprint arXiv:1612.02649*, 2016.
- [20] K. Fukushi, Y. Nozaki, K. Nishihara, and K. Nakahara, “Few-Shot Generative Model for Skeleton-Based Human Action Synthesis Using Cross-Domain Adversarial Learning,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3946–3955, 2024.
- [21] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, “Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.
- [22] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, “Cross-view action modeling, learning and recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2649–2656.