Video Interaction Recognition using an Attention Augmented Relational Network and Skeleton Data

Farzaneh Askari[1] Cyril Yared[1] Rohit Ramaprasad[2] Devin Garg[2] Anjun Hu[3]
James J. Clark[1]

[1]University of McGill, Montreal, QC, Canada
[2]University of California San Diego, San Diego, CA, United States
[3]University of Oxford, Oxford, United Kingdom

{farzaneh.askari,cyril.yared,james.clark1}@mail.mcgill.ca, {rramaprasad,degarg}@ucsd.edu,
anjun.hu@eng.ox.ac.uk

Abstract

Recognizing interactions in multi-person videos, known as Video Interaction Recognition (VIR), is crucial for understanding video content. Often the human skeleton pose (skeleton, for short) is a popular feature for VIR as the main feature, given its success for the task in hand. While many studies have made progress using complex architectures like Graph Neural Networks (GNN) and Transformers to capture interactions in videos, studies such as [33] that apply simple, easy to train, and adaptive architectures such as Relation reasoning Network (RN) [37], yield competitive results. Inspired by this trend, we propose the Attention Augmented Relational Network (AARN), a straightforward yet effective model that uses skeleton data to recognize interactions in videos. AARN outperforms other RN-based models and remains competitive against larger, more intricate models. We evaluate our approach on a challenging real-world Hockey Penalty Dataset (HPD), where the videos depict complex interactions between players in a non-laboratory recording setup, in addition to popular benchmark datasets demonstrating strong performance. Lastly, we show the impact of skeleton quality on the classification accuracy and the struggle of off-the-shelf pose estimators to extract precise skeleton from the challenging HPD dataset.

1. Introduction

Recognizing human activity, and their interactions with each other and their environment, is a crucial component of video understanding. Utilizing the skeleton as the main feature (versus combining it with appearance features) has been popularized in recent years. It is due to the fact that skeleton is a compact, concise, effective feature, while alleviating scene and objects biases and reducing privacy concerns owing to its anonymity [13, 15, 19, 22, 50]. The advancement of human pose estimation methods as well as large scale interaction recognition skeleton-annotated datasets marked a new era for using skeleton data. In our study we use skeleton as the input to our VIR model.

Some studies [12, 33] model the coarse interaction between the individuals in terms of finding fine discriminative relations between individuals' body joints. From this point of view, VIR is a *relational reasoning* problem. Santoro *et al*. [37] introduced RN, which explicitly solve relational reasoning problems in neural networks. Similar to CNNs that capture spatial, translation invariant features from grid like inputs; RNs are simple, extendable, and powerful architectures capable of reasoning about relations [37]. We will briefly review the RN architecture in Sec. 3. In this study, we propose to equip the RN model with a Self Attention (SA) mechanism that allows for better integration of relational representations.

Graph Convolutional Networks (GCNs) are popular for VIR on account of their success [6, 21, 24, 25, 52]. However, as discussed by [8], GCNs come with drawbacks. First, they are not easily extendable, meaning that it is difficult to fuse the skeleton graph with other structured modalities, such as RGB and optical flow, especially during early fusion. Second, they lack scalablility; meaning adding new individual or joint linearly increases model complexity, which is undesirable for multi-person videos [8]. In contrast, AARN is easily extendable to integrate different modalities and scalable to multiple person without significant cost. More specifically, any modality can be presented

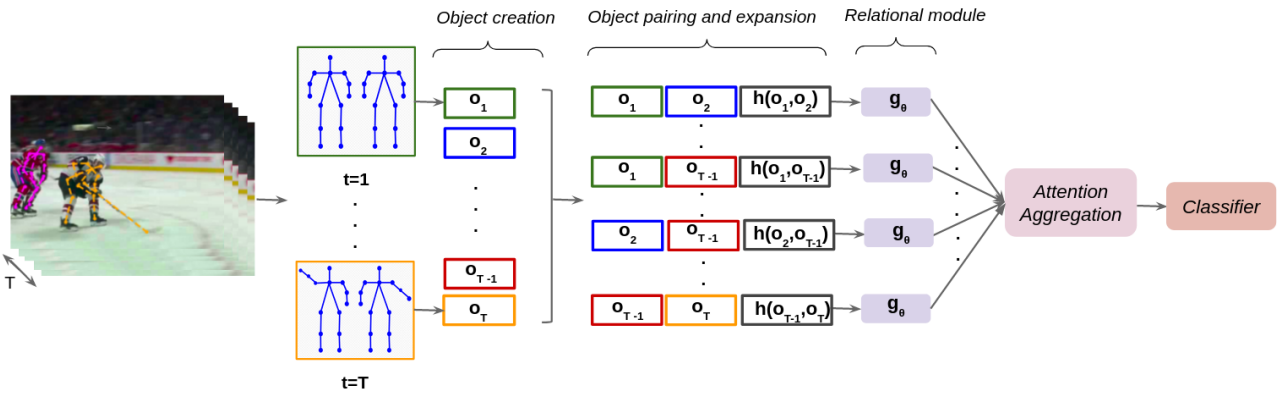

Figure 1. AARN architecture. The skeleton is extracted for every person in each frame. In HPD when there are multiple players, only the two main players are selected (using available annotations). Objects are generated from the skeletons of both actors within individual frames and then concatenated (in a unidirectional manner) to form object pairs. The object pairs serve as input to a relational module, which is then followed by a self attention module responsible for aggregating the relationships. The final relational representation is subsequently fed to a classifier to output class membership. See Sec. 3 for more details. HPD frames reprinted with permission [1].

as a form of object and concatenated along with other objects in RN framework.

Another recent trend for VIR is combining GCNs with Transformers [31]. Although, they mark the state-of-the-art, they suffer from the GCN shortcomings in addition to large number of parameters, heavy computational cost, and difficult training of Transformers. For example, IGFormer [31] uses three Transformer layers ($\sim$ 20M parameters) and requires pre-training on pseudo images of the skeleton. In contrast, AARN achieves competitive results by only using MLPs while training from scratch. Therefore, our proposed model offers a good balance between the computational cost, simplicity, and performance.

Skeleton data is effective for VIR; however, it poses some challenges. Similar actions may not yield similar numerical values in skeleton data due to varying angles and scales in each scene [51]. Normalization is a logical solution, but obtaining depth coordinates and camera features in real-world scenarios can be impractical or costly. Although different skeleton values for identical interactions is a problem, relationships between joints and poses formed by them can help the task [20, 28, 33]. Therefore, we expand our input with relative features (i.e., distance and inter/intra motion) from the objects using non-parametric h function. For more details see Sec. 3 and Fig. 2.

Although the RN architecture is known for its simplicity and adaptability, our research demonstrates how the performance of this architecture can be influenced by the structure of the input data. Consequently, in this study we propose to design our input data with two main purposes in mind: firstly, maintaining the inherent simplicity of the input, and secondly, enabling the RN to effectively reason about the spatio-temporal dynamics in a video.

Despite the great advancement in the field of VIR, many of the current datasets focus on simple interactions (e.g., hand shaking, etc) recorded in simplified recording setups. However, in many real world applications, the scene includes complex interactions, varied camera viewpoints (e.g., scale and angle), (self) occlusions, and blurry frames due to camera motions. These factors affect and challenge both the pose estimator and VIR models. Therefore, in our study, aside from the well-established skeleton-based VIR datasets, we evaluate AARN on the challenging HPD introduced by Askari *et al.* [1]. Additionally, we quantitatively and qualitatively analyze the performance of the state-of-the-art pose estimators on this dataset. The summary of our contributions in this work is:

- **Our approach:** We present an effective, easy to train and expand skeleton-based approach for VIR tasks. By integrating an SA module into the established RN architecture, we improve relational representation aggregation.
- **Input structure's impact:** We show the effect of the input structure on the performance and propose suitable inputs to maximize AARN potential. We enrich the inputs with static and dynamic features in line with our structure.
- **Ablation studies:** We show the effectiveness of our proposed modules through comprehensive ablation studies, showcasing their positive impact on performance.
- **Pose estimation study:** We evaluate the ability of current off-the-shelf pose estimators to extract precise skeletons from the challenging HPD and study the effect of skeleton quality on the performance of the VIR task.

The remaining of our paper is structured as follows. Sec. 2 reviews the related works, Sec. 3 elaborates on our methods in details and presents our pipeline for pose estimation study, Sec. 4 presents our experimental setup, and discuss

the results, ablation studies, and findings. Lastly, Sec. 5 concludes the paper.

2. Related Work

A group of studies use skeleton as the main feature [14, 36, 42]; often considering human skeleton structure while another group of studies, [1, 7], consider pose as a complementary and/or guiding feature for RGB appearance and motion features. Many studies on VIR leverage GNNs in combination with CNNs and LSTMs [6, 21, 24, 25, 52]. In [21] the actional and structural links are combined into generalized skeleton graph and fed to a GCN; resulting in enhanced representations and performance.

Several studies [33, 49, 54] incorporate the human skeleton data with relational networks (RNs) for the task of action classification. Although, all the studies above benefit from the idea of RNs, none of them explicitly focus on VIR. To the best of our knowledge, the study by Perez *et al*. [33] is the only study that employs RNs and skeleton for the task of interaction recognition. Perez *et al*. [33] propose "joint objects" defined as the coordinates of each joint through time. These objects are paired up together in a fashion representing a potential relation between the joints of each person throughout the interaction. They propose intra and inter person pairing, representing the relations between the body joints from the same person and different persons, respectively. They use RN [37] to model the relation between these objects and output VIR classification result.

While the innovative model of Perez *et al*. [33] presents interesting concepts, it struggles to effectively grasp the temporal dynamics of videos without resorting to the use of an LSTM. We theorize that this arises from their reliance on joint-centric (versus person-centric) object definition, which is local and only implicitly models the temporal aspect. Therefore, in our method, we take a more global person-centric approach to object definition with an emphasise on the temporal dimension.

The majority of sports analytics studies on ice hockey are on player tracking and identification, player/puck localization, and single person actions [4, 9, 10, 16, 30, 44–47]. There are some studies on multi-person action/interaction recognition from videos; such as [41] that proposes a CNN-RNN model to classify multi-person puck possession events (e.g., shot, dump) from videos. Askari et al. propose two studies on HPD; in one [1] they propose an RNN-CNN model for multi-person interaction recognition and key actor detection using skeleton and video frames. In another study [2], the authors propose a self supervised method based on the RN architecture where they derive image-like representations from skeleton sequence of unlabeled videos of hockey penalty dataset. They evaluate their method on the downstream task of two person interactions from videos. Aside from these two studies, most of sports activity recognition studies [27, 39] use datasets such as UCF [18]. As discussed earlier, given the interesting challenges ice hockey videos provide for VIR, there is a need for more studies on ice hockey specific datasets such as HPD.

The most well known categorization of pose estimators are top down and bottom up approaches. The top-down approaches involve a person detection as the first step, followed by the pose estimator which localize the body joints within the bounding box of the detected person. The bottom up approaches, on the other hand, first detect the body parts and then assemble them to form full body human poses. Researchers believe top-down pose estimators performance is bottle-necked by the person detector [5]. High-Resolution Net (HRNET) [40] is a top down successful pose estimator where they maintain high and low resolution representations in parallel through the whole process. The advantage of bottom up pose estimators [5, 17, 32, 34] is making the person detector dispensable, which in turns, makes the pose estimator run-time independent from number of people in the frame. Associative Embedding (AE) [29] is a successful bottom up model, where every detected human joint has an embedding vector. The distances between joints embedding are used to group the joints.

3. Method

In this section we elaborate on the architecture of our method as well as the pose evaluation study pipeline. Fig. 1 demonstrates our overall architecture. Some of the notations in this section are partially used as [33, 37].

3.1. Relational network architecture

Relational network definition: Our method is based on the RN architecture, which was first proposed by Santoro *et al*. [37]. Eq. (1) describes the underlying idea of the relational reasoning method.

$$RN(O) = f_\phi \left(\sum_{t,s=1}^{T} g_\theta \left(o_t, o_s \right) \right) \tag{1}$$

with O describing a set of T objects, where each object (e.g., o_t) is represented as a vector belonging to $\mathbb{R}^m$ containing the properties of an object. The definition of an object is flexible depending on the application; for example, an object can contain CNN features from an image, last hidden state of an LSTM describing a sentence, or restructured pose information from video frames. The relational model g_θ, is a function with learnable parameters that models the relationship between each pair of objects. The parameters are shared for all the object pairs. $\sum$ describes an aggregation function that is often non-parametric such as average. Finally, f_ϕ is a function with trainable weights that takes in the aggregated relational representations and output the reasoning , such as predicted class in a classification task.

In our work, we define a novel set of objects that are simple, elegant, yet effective to solve the task of interaction recognition in hand. Additionally, we propose to use a parametric aggregation function based on the transformer attention mechanism instead of the non-parametric ones. We demonstrate by applying the aforementioned changes we can increase the performance of RN models on interaction recognition from videos and reason about the temporal dynamics without requiring LSTMs as the studies such as [33]. We elaborate on the details of our approach below.

Objects definition: The underlying idea of employing RN in our application is to tackle the problem of video interaction classification through reasoning about the existing relations between the actors' body joints and how their development over time distinguishes the actions from each other. For instance, pushing and punching interactions share similarities in the beginning and we only can distinguish them by observing the action over time. Therefore we create our objects and relations with considering to capture the most spatio-temporal information from the skeleton data.

We define each object as:

$$o_t = (x_1^{p_1}, y_1^{p_1}, x_2^{p_1}, y_2^{p_1}, ..., x_N^{p_1}, y_N^{p_1}, x_1^{p_2}, y_1^{p_2}, x_2^{p_2}, y_2^{p_2}, ..., x_N^{p_2}, y_N^{p_2}, t) \tag{2}$$

where x_n and y_n are the 2D coordinates of the n^{th} joints, N indicates the total number of joints, p_1 and p_2 represent each of the actors, and t is the time index. Therefore, for each video frame we define an object that contains x, y coordinates of all the joints for both actors. It is important to note that for the datasets with 3D coordination we use x, y, z coordinates.

Our objects carry spatial information and by pairing them we form spatio-temporal inputs for the model. The pairwise inputs are formed by concatenating each two objects together. In order to avoid redundancy, the concatenation of each two objects is only unidirectional (e.g., if (o_3, o_4) exists (o_4, o_3) is not created).

Defining objects and object pairs as outlined earlier, offers the advantage of capturing the full skeleton structure of both actors per object, following a person-centric approach. This grants the relational module a more global scope of information in contrast to the joint-centric objects as in [33]. Furthermore, forming the object pairs temporally (i.e., between frame as opposed to between joints) allows for more holistic understanding of the dynamics present in the video.

Previous research utilizing skeleton data [33, 51, 53] has demonstrated that enhancing the input with relative features extracted from skeleton data significantly improves the performance. These features may vary based on the domain-specific knowledge. To this end, some studies leverage the bone and motion between the joints [8, 28]. In our study, we expand each object pair with relative features that are not only compatible with our input structure but also effective for VIR. Consequently, we extend the RN formulation to equation Eq. (3). We further analyze the impact of adding these features through ablation studies.

$$RN(O) = f_\phi \left(\sum_{t,s=1}^{T} g_\theta \left(o_t, o_s, h(o_t, o_s) \right) \right) \tag{3}$$

The h function outputs the concatenation of outputs from $D(o_t, o_s)$, $M(o_t, o_s)$, and $L(o_t, o_s)$ defined in Eq. (4), Eq. (5), and Eq. (6) respectively which are demonstrated in Fig. 2.

$$D(o_t, o_s) = (\| c_{1t}^{p1} - c_{1t}^{p2} \|, \| c_{2t}^{p_1} - c_{2t}^{p2} \|, ..., \| c_{Nt}^{p1} - c_{Nt}^{p2} \| \frown \| c_{1s}^{p1} - c_{1s}^{p2} \|, \| c_{2s}^{p1} - c_{2s}^{p2} \|, ..., \| c_{Ns}^{p1} - c_{Ns}^{p2} \|) \tag{4}$$

$$M(o_t, o_s) = (\| c_{1t}^{p1} - c_{1s}^{p1} \|, \| c_{2t}^{p1} - c_{2s}^{p1} \|, ..., \| c_{Nt}^{p1} - c_{Ns}^{p1} \| \frown \| c_{1t}^{p2} - c_{1s}^{p2} \|, \| c_{2t}^{p2} - c_{2s}^{p2} \|, ..., \| c_{Nt}^{p2} - c_{Ns}^{p2} \|) \tag{5}$$

$$L(o_t, o_s) = (\| c_{1t}^{p1} - c_{1s}^{p2} \|, \| c_{2t}^{p1} - c_{2s}^{p2} \|, ..., \| c_{Nt}^{p1} - c_{Ns}^{p2} \| \frown \| c_{1s}^{p1} - c_{1t}^{p2} \|, \| c_{2s}^{p1} - c_{2t}^{p2} \|, ..., \| c_{Ns}^{p1} - c_{Nt}^{p2} \|), \tag{6}$$

where the variable c stands for the coordinates of a joint (i.e., (x, y) (or (x, y, z) using 3D skeleton). Therefore $c_{Nt}^{p_1}$ is a vector of (x, y) locations of the N^{th} joint of person one (i.e., p_1) in frame t. And $c_{Ns}^{p_2}$ indicates the (x, y) locations of the N^{th} joint of person one (i.e., p_2) in frame s. $D(o_t, o_s)$ represents the distance between two actors within each timestep (or object); which adds extra spatial information in addition to raw coordinates; $M(o_t, o_s)$ captures the motion of each actor between t and s timesteps (a.k.a., intra-motion); and $L(o_t, o_s)$ is the motion between two actors in different timesteps (a.k.a., inter-motion). In other words M and L indicate how an actor's distance changes w.r.t. him/herself and the other actor across the two timesteps.

Attention based aggregation: Although most of the studies using RN architecture utilize non-parametric aggregation functions to aggregate the relational representations (i.e., representation out of g_θ), in this work we explore the benefits of a parametric aggregation module. When using average as the aggregation function, weighted sum of representations are calculated with equal weights assigned to all the representations. Our proposition is to use a SA mechanism and learn the assigned weights to relational representations instead. Since the interaction between the relational representations are desirable, we employ the scaled dot product SA from [43]. Additionally, as a baseline we replace the dot product SA with the additive SA from [3].

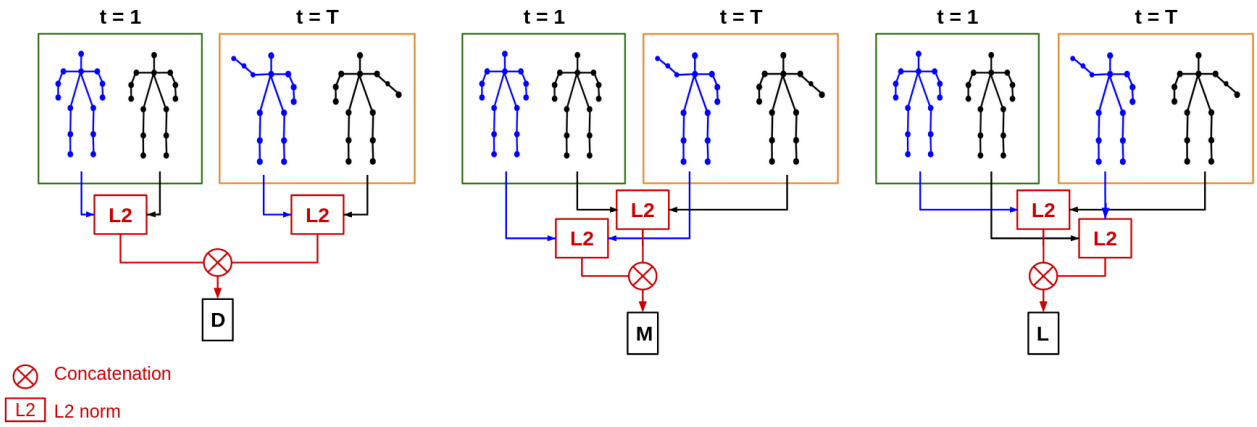

Figure 2. Relative features. Illustration of computation of D, M, and L (Eq. (4) to Eq. (6)) between two exemplar frames at time instances t=1 and t=T. Left: D signifies the distance between the two actors within each frame (i.e., within each object). Middle: M represents the movement of each actor between two time steps (A.K.A., intra-motion). Right: L represents the motion between the two actors across the same time interval (A.K.A, inter-motion) . h is the concatenation of all of them. See Sec. 3 for more details.

3.2. Baseline models:

Alternative Object definition: We hypothesize that defining objects for the RN architecture is not a trivial task and significantly affects the performance. In order to make the distinction, we term our original objects "temporal objects" (see Eq. (2)) and the alternative objects we define subsequently as "spatial objects". Throughout this paper, when referring to an object, we are referring to "temporal objects", unless specified otherwise.

To evaluate our hypothesis, we adopt a different approach where we individually define an object for each joint over time $s_i = (x_1^{p_1}, y_1^{p_1}, x_2^{p_1}, y_2^{p_1}, ..., x_T^{p_1}, y_T^{p_1}, x_1^{p_2}, y_1^{p_2}, ..., x_T^{p_2}, y_T^{p_2}, i)$ where x_t and y_t denote the 2D coordinates of the i^{th} joints at frame t. T represent the total number of frames, p_1 and p_2 correspond to the two actors, and i is the joint index. As a result, for each body joint we create an object that contains x, y coordinates of that joint for both actors across all the timesteps. It is important to note that when dealing with datasets that provide 3D coordinates we use x, y, z coordinates. This object definition bears resemblance to the one presented in [33], with the distinction that it includes the joints for both actors rather than just one. As a result, when constructing object pairs, the interaction among more joints are captured. Given the formulation of spatial objects above, we define appropriate distance and motion information. The details are available in Sec. 1 of supplementary materials.

Transmotion attention: Building on promising initial results with the addition of relative features (function h), we investigated directly integrating these features into the attention model. This new module, called "Transmotion" (short for Transformer + motion) attention, generates two sets of attention coefficients, averages them for final coefficients, and produces aggregated representations. The first attention set is from scaled dot product attention. The second set comes from summing values from Eq. (5) (intra-motion) for object pairs. We theorize greater motion between joints across timesteps increases attention coefficients for those object pairs, as more motion frames are assumed to carry more useful classification information. For a detailed module architecture, refer to Sec. 3 in the supplementary materials.

3.3. Pose evaluation study

Askari et al [1] mention in their study that the HPD imposes challenges to state-of-the-art pose estimators due to complex scenes, unusual players poses, and the generic large scale datasets that current pose estimators are trained on. As part of our experiment, we qualitatively analysis the performance of two state-of-the-art pose estimators on this dataset. The dataset uses a custom 14 body key-points annotations which a modification of COCO [23] annotation by averaging the five head keypoints to one key-point. Given that many pose estimators are pre-trained on COCO dataset with 18 keypoints, we first extrapolate 4 extra keypoints for eyes and ears. This is necessary in order to be able to inference poses using an available pre-trained model. We place an axis along passing through the nose and perpendicular to the nose-neck axis. The eyes are placed one quarter of the nose-neck distance and the ears half of that distance on both sides of the nose. Since the direction the player is facing is unknown, the left/right annotation is assigned based on the ear/eye keypoint distance from the left/right shoulder. Following this procedure we convert from dataset custom

keypoints to COCO format. Additionally, the evaluation of any top-down pose estimation method requires ground-truth bounding boxes. Since the HPD does not include bounding box annotations, we automatically infer them using the available keypoint annotations. We extract the boxes in a way that each player is fully encapsulated in the box. We add a margin, equal to twice of head-neck distance, to correct for estimation of bounding box and the approximate location of keypoints.

4. Experimental Evaluation

4.1. Datasets

HPD: [1] includes clips of penalties from National Hockey League (NHL) ice hockey broadcast videos. There are three classes of *tripping, slashing, and no penalty* with 80, 76, and 98 clips per class respectively. The clips are two to six second long, fully encapsulating the penalty event. Each video includes annotation of main interacting players (i.e., P_1 and P_2), ground truth pose annotations of 14 body keypoints, and two additional keypoints for each end of the hockey stick for every player in the frame. In our study, we only consider the two main interacting players as the input to our model. Additionally, we augment the dataset using the horizontal flip and affine transformation (e.g., scale). For evaluation on this dataset, we use 5-fold cross validation.

SBU [53] includes a total of 282 short videos of two person interactions. There are eight classes and the videos are 2-3 seconds long, recorded in a laboratory setup with static background. The dataset includes 3D skeleton data with 15 joints per person for each frame. The poses are noisy and inaccurate in some frames. For evaluation on this dataset, we use 5-fold cross validation (suggested by the authors).

NTU RGB+D [26, 38] is originally action recognition datasets; but they include 11 of interaction. Following the protocol in [33], we evaluate our method on the interaction classes only. The NTU RGB+D contains 10347 videos respectively. Despite using only the interaction classes, it surpass the scale of SBU and UT-Interaction datasets. Compared to SBU dataset, the scenes are more complex with varied viewpoints. The NTU dataset includes 3D skeleton of 25 joints for each person in all the frames. The dataset is evaluated using Cross-Subject and Cross-View protocols.

4.2. Experimental details

Pose estimation: We use the frame-based Associative Embedding (AE) [29] with the HRNet backbone [40] as the bottom-up model. We use Contextual Instance Decoupling (CID) [48] as the top-down model with the HRNet backbone [40]. In the pretrain-only evaluation model we use a model pre-trained on the COCO dataset. For the fine-tune evaluation phase, we initialize the network with pre-trained weights and fine-tune on the HPD with a learning rate of 0.0005 and 0.001 for the bottom-up and top-down models respectively. It is important to note for the fair evaluation of pose estimation methods we do not include the hockey stick keypoints given that it is specific to this sport and not the human body structure.

Method	AP	AP_{50}	AP_{75}	AR
CID (COCO pretrain only)	0.57	0.88	0.61	0.39
CID (fine-tune)	0.68	0.92	0.75	0.71
AE (COCO pretrain only)	0.50	0.84	0.55	0.56
AE (fine-tune)	0.70	0.95	0.77	0.75

Table 1. The result of pose estimations (pretraining and fine-tuned mode) on HPD.

Method	HPD (%)
Ground truth pose:	
LRCN (reported by [1])	63.64
PoseC3D (reported by [1])	81.63
KAD [1]	93.93
AARN-wo/RF	91.41 ± 0.62
AARN	94.54 ± 0.08 (94.6)
Estimated pose:	
AARN	87.24 ± 0.45 (87.57)

Table 2. Penalty classification accuracy. Comparison of AARN with previous works on HPD. Observing the effect of quality of estimated pose data on the classification performance.

AARN: we use a four layer MLP with 1000 units per layer for g_θ. The linear layer in dot product self-attention have 1000 units as well. We set the dropout rate to $p = 0.1$. The classifier has three fully connected layers with 500 units for the first layer and 250 for the last two layers followed by a $softmax$ to generate class membership. The additive self-attention (for baseline) consists of three fully connected layer with 500 units and $\tanh$ activation per unit, followed by a $softmax$ after the final layer. The layer weights are initialized with truncated normal distribution $(0, 0.045)$ for SBU and $(0, 0.09)$ for NTU RGB+D. We use the Adam optimizer with learning rate of $1e-4$ to minimize a binary cross entropy loss with early stopping. During training we randomly swap the order between the persons' joints to improve generalization.

4.3. Results and discussion

Pose evaluation study: Table 1 demonstrates the results for this study. In both the top-down and bottom-up pose estimation cases we observe significant improvements across all metrics after fine-tuning the pose estimator. This confirms the claim and qualitative analysis by Askari *et al.* [1] that despite their abundant benefits, current pose estimators

Method	SBU (%)
2s-GCA [24]	94.9
IGFormer [31]	98.4
LSTM-IRN [33]	98.2
AARN-wo/RF	94.22 ± 1.23
AARN	97.97 ± 0.46 (98.30)

Table 3. Interaction classification accuracy. Comparison AARN with previous works on SBU dataset.

Method	NTU RGB+D	
	X-Sub (%)	X-View (%)
GCA [25]	85.9	89.0
2s-GCA [24]	87.2	89.9
AS-GCN [21]	89.3	93.0
CTR-GCN [6, 31]	91.6	94.3
IGFormer [31] ($\sim$ 20M)	93.6	96.5
LSTM-IRN [33]	90.5	93.5
AARN-wo/RF	87.69 ± 0.35	90.57 ± 0.61
AARN ($\sim$ 3.5M)	90.79 ± 0.65 (91.26)	93.42 ± 0.65 (93.88)

Table 4. Interaction classification accuracy. Comparison AARN with previous works on NTU RGB+D dataset (interaction classes only). For fair comparison with Transformer-based model, IGFormer, we note the *approximate* number of parameters in parenthesis.

struggle to estimate precise poses of complex scenes and interactions. This is due to firstly, the large-scale pretraining datasets mostly containing generic poses (versus complex sports poses); secondly, the bulky hockey jersey affecting the overall human skeleton shape; and lastly, the variety of viewpoint scale and angle in sports dataset. Fig. 2 of supplementary material demonstrates the qualitative results.

AARN: Tab. 2 to Tab. 4 show our results in comparison with other methods. We report average accuracy and standard deviation of two/three runs and note the best run in parentheses. On the HPD, our model outperforms the KAD [1] that is an LSTMs based model equipped with additive SA mechanism [3], only when we use the relative features. This underlines the importance of including relative features (e.g., distance, motion, joints angle, etc) when using skeleton; which is emphasized in other studies as well [11, 33, 35, 53]. Another noteworthy observation is that our model offers good performance on videos without requiring RNNs; owning to our global temporal object definition as observed in Tab. 6. Additionally, We compare the results of using AARN with estimated vs ground-truth poses on HPD. In this scenario we use the top-down pose estimator to extract skeletons and pad the missing poses with zero. As part of our observation, many videos contained some missing poses. We observe a performance drop using estimated poses. Tab. 2 demonstrates our results.

On the public benchmark, our method outperforms LSTM-IRN, which is the most successful RN-based network for interaction recognition, without requiring the LSTM and with fewer objects/relations. AARN also outperforms several GNN-based architectures and offers competitive results with others. CTR-GCN [6] is a channel-wise topology refinement graph convolutional network that consists of ten spatio-temporally modeling block with residual connections. In their model they set the neighbourhood of each joint as the entire human skeleton which leads to scalability challenge. Unsurprisingly, the higher overall model capacity and dense skeleton graph definition benefit the performance. IGFormer [31] consists of three Transformers blocks initialized by pre-trained weights of ViT-based model. In comparison our method offers a light architecture, consisting of two MLP blocks and one SA block, with only a few percentage trade-off in performance. AARN is easy to train, scalable, easily expandable to other modalities, and has significantly less number of parameters.

Aggregation: We perform ablation study on the effect of aggregation function. The popular aggregation with RN-based approaches is the average, however, this is a non-parametric aggregation function that assigns equal weights to all representation. We additionally experiment with popular additive self-attention [3]. Although additive self-attention is a learnable function, the representation only interacts through the layer weights (and not directly), which makes it difficult to generate appropriate attention coefficients. We observe that the dot product attention is the most effective aggregation. Given that the dot product attention models direct interactions between the representations in order to assign appropriate weights, these results are expected. Furthermore, we evaluate the Transmotion attention (see Sec. 3) both with (noted as w/RF) and without (noted as wo/RF) expanding the input with the relative features. See Tab. 5. Transmotion attention combines scaled dot-product attention and motion data, it underperforms compared to using scaled dot-product attention alone. This could be due to central frame focus, and short video lengths, resulting in motion information being insufficient or excessively sparse as an attention factor.

Objects and relative features: The influence of object definition on the performance of the AARN model is depicted in Tab. 6. Despite retaining identical components except for the objects, a noticeable decline in performance is observed, particularly pronounced in larger-scale datasets. As discussed before, this decline is attributed to the localized and joint-centric input definition, which presents a difficulty for the model in adequately grasping comprehensive temporal dynamics. Consequently, techniques such as [33] resort to utilizing LSTMs to effectively capture temporal dynamics. Furthermore, we note that leveraging the rela-

Aggregation	HPD (%)	SBU (%)	NTU RGB+D	
			X-Sub (%)	X-View (%)
Average	90.61 ± 1.01	95.56 ± 0.069	89.98 ± 0.48	92.06 ± 0.388
Additive attention	90.29 ± 0.37	96.07 ± 0.07	90.31 ± 0.21	93.20 ± 0.3
Transmotion attention (wo/RF)	88.94 ± 0.32	95.04 ± 1.5	87.77 ± 0.21	89.98 ± 0.26
Transmotion attention (w/RF)	90.89 ± 0.46	96.8 ± 0.069	89.97 ± 0.20	92.35 ± 0.22
AARN	**94.54 ± 0.08 (94.6)**	**97.97 ± 0.46 (98.30)**	**90.79 ± 0.65 (91.26)**	**93.42 ± 0.65 (93.88)**

Table 5. The effect of different relational representation aggregation methods on the interaction classification accuracy.

Object type	HPD (%)	SBU (%)	NTU RGB+D	
			X-Sub (%)	X-View (%)
AARN (spatial objects)	90.41 ± 0.56	97.45 ± 0.08	87.12 ± 0.09	90.71 ± 0.12
AARN (default-temporal objects)	**94.54 ± 0.08 (94.6)**	**97.97 ± 0.46 (98.30)**	**90.79 ± 0.65 (91.26)**	**93.42 ± 0.65 (93.88)**

Table 6. The effect of object definition on the interaction classification accuracy.

tive features to expand inputs proves more effective than directly using them as attention coefficients, as observed in Tab. 5 (Transmotion w/RF and Wo/RF).

Lastly, the result of our ablation study is demonstrated in 2 to 4. (wo/RF) represents the same architecture without the enhancing the input with relative features. The result of our experiment demonstrates that using the relative skeleton features consistently improves the performance across all the datasets. It is important to note that the inclusion of these features are among the key contributors to the superiority of our approach. This enables us to outperform LSTM-IRN [33] and capture temporal dynamics without requiring an LSTM. We further analyze the effect of each component of relative features (e.g., Eq. (4) to Eq. (6)) and their combinations through ablation studies. Due to space limitation, these results are available in Sec. 2 in our supplementary materials.

5. Conclusion

We summarize our contributions to VIR by showcasing the potential of carefully designed inputs and models for the task. To the best of our knowledge, we are the first RN-based method to reach strong performance in VIR by solely utilizing skeleton data without requiring multi-modality (e.g., RGB) and CNN/RNN. We define skeleton objects that match our task, expand them with robust relative features and equip our RN-based model with a dot-product SA mechanism. Our architecture is light and easy to train (vs Transformers), extendable to multi-person, and easily expandable to other modalities (vs GCN). Additionally, we highlight the challenges posed by real-world datasets for off-the-shelf pose estimators and explore the impact of pose quality on downstream tasks relying on skeletons.

References

[1] Farzaneh Askari, Rohit Ramaprasad, James J Clark, and Martin D Levine. Interaction classification with key actor detection in multi-person sports videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3580–3588, 2022. 2, 3, 5, 6, 7

[2] Farzaneh Askari, Ruixi Jiang, Zhiwei Li, Jiatong Niu, Yuyan Shi, and James J Clark. Self-supervised video interaction classification using image representation of skeleton data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5228–5237, 2023. 3

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 4, 7

[4] Zixi Cai, Helmut Neher, Kanav Vats, David A Clausi, and John Zelek. Temporal hockey action recognition via pose and optical flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 3

[6] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. 1, 3, 7

[7] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3218–3226, 2015. 3

[8] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 1, 4

[9] Mehrnaz Fani, Helmut Neher, David A Clausi, Alexander Wong, and John Zelek. Hockey action recognition via in-

tegrated stacked hourglass network. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 29–37, 2017. 3

[10] Mehrnaz Fani, Pascale Berunelle Walters, David A Clausi, John Zelek, and Alexander Wong. Localization of ice-rink for broadcast hockey videos. *arXiv preprint arXiv:2104.10847*, 2021. 3

[11] Xiaoke Hao, Jie Li, Yingchun Guo, Tao Jiang, and Ming Yu. Hypergraph neural network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 30:2263–2275, 2021. 7

[12] Yanli Ji, Guo Ye, and Hong Cheng. Interactive body part contrast mining for human interaction recognition. In *2014 IEEE international conference on multimedia and expo workshops (ICMEW)*, pages 1–6. IEEE, 2014. 1

[13] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14:201–211, 1973. 1

[14] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Leveraging structural context models and ranking score fusion for human interaction prediction. *IEEE Transactions on Multimedia*, 20(7):1712–1723, 2017. 3

[15] Łukasz Kidziński, Bryan Yang, Jennifer L Hicks, Apoorva Rajagopal, Scott L Delp, and Michael H Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature communications*, 11(1): 4054, 2020. 1

[16] Maria Koshkina, Hemanth Pidaparthy, and James H Elder. Contrastive learning for sports video: Unsupervised player classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4528–4536, 2021. 3

[17] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11977–11986, 2019. 3

[18] Tian Lan, Yang Wang, and Greg Mori. Discriminative figure-centric models for joint action localization and recognition. In *2011 International conference on computer vision*, pages 2003–2010. IEEE, 2011. 3

[19] Alexander C Li, Alexei A Efros, and Deepak Pathak. Understanding collapse in non-contrastive siamese representation learning. In *European Conference on Computer Vision*, pages 490–505. Springer, 2022. 1

[20] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 3d human action representation learning via cross-view consistency pursuit. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4741–4750, 2021. 2

[21] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019. 1, 3, 7

[22] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9572–9581, 2019. 1

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[24] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2017. 1, 3, 7

[25] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1647–1656, 2017. 1, 3, 7

[26] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 6

[27] Guillaume Lorre, Jaonary Rabarisoa, Astrid Orcesi, Samia Ainouz, and Stephane Canu. Temporal contrastive pretraining for video action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 662–670, 2020. 3

[28] Yunyao Mao, Wengang Zhou, Zhenbo Lu, Jiajun Deng, and Houqiang Li. Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. In *European Conference on Computer Vision*, pages 734–752. Springer, 2022. 2, 4

[29] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017. 3, 6

[30] Kenji Okuma, David G Lowe, and James J Little. Self-learning for player localization in sports video. *arXiv preprint arXiv:1307.7198*, 2013. 3

[31] Yunsheng Pang, Qiuhong Ke, Hossein Rahmani, James Bailey, and Jun Liu. Igformer: Interaction graph transformer for skeleton-based human interaction recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pages 605–622. Springer, 2022. 2, 7

[32] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–286, 2018. 3

[33] Mauricio Perez, Jun Liu, and Alex C Kot. Interaction relational network for mutual action recognition. *IEEE Transactions on Multimedia*, 24:366–376, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[34] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt

Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016. 3

[35] Zhenyue Qin, Yang Liu, Pan Ji, Dongwoo Kim, Lei Wang, RI McKay, Saeed Anwar, and Tom Gedeon. Fusing higher-order features in graph neural networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 7

[36] Michalis Raptis and Leonid Sigal. Poselet key-framing: A model for human activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2650–2657, 2013. 3

[37] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30, 2017. 1, 3

[38] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 6

[39] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9787–9795, 2021. 3

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3, 6

[41] Moumita Roy Tora, Jianhui Chen, and James J Little. Classification of puck possession events in ice hockey. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 147–154. IEEE, 2017. 3

[42] Arash Vahdat, Bo Gao, Mani Ranjbar, and Greg Mori. A discriminative key pose sequence model for recognizing human interactions. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1729–1736. IEEE, 2011. 3

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[44] Kanav Vats, William McNally, Chris Dulhanty, Zhong Qiu Lin, David A Clausi, and John Zelek. Pucknet: Estimating hockey puck location from broadcast video. *arXiv preprint arXiv:1912.05107*, 2019. 3

[45] Kanav Vats, Mehrnaz Fani, David A Clausi, and John Zelek. Multi-task learning for jersey number recognition in ice hockey. In *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*, pages 11–15, 2021.

[46] Kanav Vats, William McNally, Pascale Walters, David A Clausi, and John S Zelek. Ice hockey player identification via transformers and weakly supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3451–3460, 2022.

[47] Kanav Vats, Pascale Walters, Mehrnaz Fani, David A Clausi, and John S Zelek. Player tracking and identification in ice hockey. *Expert Systems with Applications*, 213:119250, 2023. 3

[48] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11060–11068, 2022. 6

[49] Wei Wang, Jinjin Zhang, Chenyang Si, and Liang Wang. Pose-based two-stream relational networks for action recognition in videos. *arXiv preprint arXiv:1805.08484*, 2018. 3

[50] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision*, 129(5):1675–1690, 2021. 1

[51] Huimin Wu, Jie Shao, Xing Xu, Yanli Ji, Fumin Shen, and Heng Tao Shen. Recognition and detection of two-person interactive actions using automatically selected skeleton features. *IEEE Transactions on Human-Machine Systems*, 48 (3):304–310, 2017. 2, 4

[52] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1, 3

[53] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 28–35. IEEE, 2012. 4, 6, 7

[54] Jiagang Zhu, Wei Zou, Zheng Zhu, and Yiming Hu. Convolutional relation network for skeleton-based action recognition. *Neurocomputing*, 370:109–117, 2019. 3