Self-Supervised Video Interaction Classification using Image Representation of Skeleton Data

Farzaneh Askari Ruixi Jiang* Zhiwei Li* Jiatong Niu* Yuyan Shi* James J. Clark

McGill University, Montreal, QC, Canada

{farzaneh.askari,ruixi.jiang,zhiwei.li2,jiatong.niu,yuyan.shi, james.clark1}@mai1.mcgill.ca

Abstract

Recognizing interactions from sports games broadcast videos is an application of Interaction Recognition from Videos (IRV), that offers many challenges due to complex interactions that are often recorded from a suboptimal view point. Annotating large scale sports specific datasets is expensive and time-consuming. Therefore, in this study, we propose to demonstrate the effectiveness of applying Self-Supervised Learning (SSL) methods for building useful representations from human skeleton pose data (pose for short) without requiring costly annotations for a large scale dataset. Given the numerous well established image-based SSL methods, we demonstrate how to adapt them for sequences of pose through data transformation and a series of pose-based augmentations. We specifically adapt the Relational Reasoning SSL (Relational-SSL for short) [27] and achieve $68.18 \pm 0\%$ *and* $76.62 \pm 2.7\%$ *in linear evaluation and finetuning protocols, respectively, for the downstream task of IRV from sports broadcast videos. Lastly, we run ablation studies on different components of the method, including the effect of using estimated pose (versus ground truth) on the performance of the downstream task.*[1]

1. Introduction

Video understanding is a popular field in computer vision owing to its diverse applications such as surveillance, health care, and entertainments. Understanding human action and their interactions with the environment (e.g., other humans and objects) is a crucial component of video understanding. In recent years, the groundbreaking advancement of Deep Learning (DL) methods marked a new era in advancement of computer vision applications, including video understanding. The increased usage of social media platform and advanced broadcast technologies provided the researchers with enormous amount of data in various setups such as social setups, retail environments, outdoors, and sports games. Among these applications, sports analytics gained significant attentions because of challenges and promises it offers from a scientific and practical point of view.

The research contributions in sports analytics from a video understanding point of view, involve recognizing and distinguishing different sports activities (e.g., running vs swimming vs basketball) [15, 19], and recognizing individual players' actions and/or group activities from team sports such as volleyball [29]. The former is often solvable through using inter-class appearance and/or dynamic differences (e.g., water vs basketball court). The latter is less studied due to requiring more sophisticated feature extractions and sport-specific large scale datasets. These datasets involve a costly and time-consuming data collection and annotation process.

The research community has been consistently seeking new approaches to reduce or eliminate expensive data annotation. The goal of Self-Supervised Learning (SSL) is to obtain useful representations without depending on a large annotated dataset. A characteristic of SSL methods is their reliance on defining and learning a surrogate objective (pretext task). The objectives are defined on the unlabeled training set such that the model disregards the obvious information and focuses on fine-grained non-obvious features. Once the pretext objective is achieved, the trained backbone is used for the downstream tasks (e.g., activity classification). Given the abundance of unlabeled data from the sports games and the high cost of data collection and annotation, SSL approaches are beneficial for the field of sports analytics.

Despite the large literature on human action recognition from videos, many studies focus on single person actions; however, in reality many scenes consist of multiple persons and their interactions with the environment [1]. Therefore,

*These authors contributed equally.

[1]The use of images from broadcast of NHL games was made pursuant to Fair Dealing Guidelines for non-commercial research purpose. The use of this copy may require the permission of copyright owners.

Figure 1. Frames from the hockey penalty dataset with players' poses and stick annotations; classes from top to bottom: tripping, slashing, and no penalty. Reprinted with permission. [1]

in recent years a few datasets focused on human-human and human-object interactions [31, 33, 45]; however, most of the videos include simple interactions (e.g., giving/taking objects) with inter-class high variance recorded in laboratory setups. The sports broadcast videos often include complex interactions, varied camera view point, frequent occlusions, and blurry scenes due to camera motion. Sports such as Ice hockey (hockey for short) present even more challenges, such as fast movements of the players and rapid transitions between the game events. In addition to the aforementioned challenges, hockey games involve frequent penalties, which are complex interactions among the players. The penalties are inevitable parts of the game due to players' speed and density in the small hockey rink (i.e., $\frac{1}{4}$ of a soccer field). The penalties are challenging to recognize and distinguish due to substantial (self) occlusions, varying and suboptimal camera view point, and low inter-class variance of penalty scenes.

Although hockey penalty classification is an interesting application in the intersection of sports analytics and IRV, the sports data collection and annotation are expensive and time-consuming processes. This makes the use of unlabeled sports data in conjunction with SSL methods appealing. Even though most SSL approaches use large-scale datasets to pretrain and report their results [4, 5, 12, 13], Cao et al. demonstrates the effectiveness of SSL algorithms on small datasets [3]. Therefore, in this paper, we study the advantage of using SSL pretraining on the unlabeled portion of a hockey penalty dataset [1]. We then examine the quality of the learned representations through linear evaluation and fine-tuning protocols on the labeled portion of the dataset for the downstream task of two-person hockey penalty classification from broadcast videos.

The hockey penalty dataset includes ground truth pose annotations for all the visible players in the scene [1]. The location of body joints of interacting individuals in a frame and how they shift over the time has proven to be valuable yet concise information for the task of IRV [1, 6, 16, 30, 40]. In our study, we aim to use the SSL method to capture the semantics of human skeleton data in a video. Given the abundance of well established image-based SSL methods, we propose to convert the temporal pose data to spatio-temporal image-like inputs. This enables us to take advantage of successful existing SSL algorithms developed for image data [4, 5, 12, 13, 27]. In this study, we specifically utilize the Relational-SSL method proposed by Patacchiola et al. [27] and adapt it for our IRV problem. We evaluate the effects of the different algorithmic decisions such as choice of augmentations, dataset partition, aggregation methods, and so forth on the performance of the downstream task.

Although off-the-shelf pose estimators provide us with pose annotation for cheap, they often fail to capture poses from complex scenes such as penalties in hockey games [1]. The struggle is due to occlusions, camera motions, unusual poses of the players, and the color of the jerseys blending with the background (i.e., ice and side boards) [1]. The ground truth pose annotation is more accurate and yet expensive. Therefore, we study the effect of obtaining

poses using an off-the-shelf pose estimator (versus using the ground truth pose) on the quality of SSL pretrained representations and the performance of the downstream task.

To summarize, our contributions are as follows:

- Demonstrating the effectiveness of SSL approaches on a small-scale dataset for the downstream task of penalty classification from hockey broadcast videos as an application of IRV
- Adapting image-based Relational-SSL for temporal skeleton data by proposing an effective data transformation method and suitable augmentations relevant to the downstream task
- Evaluating the robustness of Relational-SSL to estimated poses and the effect of estimated poses on the performance of the downstream task
- Lastly, elaborating on the adaptivity of the Relational-SSL to our task through ablation studies

To the best of our knowledge this is the first paper to study the effectiveness of SSL approaches for the downstream task of IRV from sports broadcast videos, specifically hockey, using sequence of pose data. Our paper is structured as follows. Sec. 2 reviews the current literature on the topic. Sec. 3 presents the dataset and different components of our methodology in details. Sec. 4 elaborates on our experimental setup, results, and ablation studies. Finally, Sec. 5 discusses our findings and concludes our work.

2. Related work

Pose is a popular feature for the task of human interaction recognition, either as the main [16, 30, 40] or complimentary feature [1, 6]. Liu et al. [22] feed the restructured skeleton data into a gated spatio-temporal LSTM network to recognize actions and interactions. Their input is restructured using a tree traversal algorithm. Another study [6] proposes to use pose as a guide for motion and appearance features from part patches. These features are input to two streams of RGB and flow convolutional neural network. A group of studies models the problem of interaction recognition as interaction between body joints (or limbs) of the actors [14, 28, 45]. Perez et al. [28], benefit from the Relation Network (RN) [32] architecture and proposes to solve interaction recognition from videos through reasoning about the relations between the actors' joints. They specifically define two types of relations, namely, intra-person and inter-person. Intra and inter-person relations capture the interaction between an actors' joints to his/her own joints and the other actors' joints over time, respectively.

The Relational Network (RN) architecture was first introduced by Santoro et al. [32] to explicitly solve problems that involve relational reasoning using neural networks. Similar to Convolutional Neural Networks (CNNs) that capture spatial, translation invariant features from grid like inputs; RN can reason about relations. RN has been used in several applications of computer vision, such as interaction recognition [28], pose estimation [26], video question answering [20], and SSL [27].

Among available approaches in SSL, contrastive learning methods gained much attention with the methodology being extended to many of the existing applications [5, 11, 35, 46] including human action recognition from videos. Gao et al. [11] propose a contrastive SSL method to learn useful representations from unlabeled pose data for the task of human action recognition from videos. Following the principle of contrastive learning, they generate augmentations of pose data from videos by applying scale and rotation transformations. The correlated pairs are then used to train a base encoder network by minimizing a contrastive loss. The goal of contrastive learning is to maximize the agreement between an example (i.e., data point) and its augmentations (positive pairs) while maximizing the disagreement between different examples (negative pairs). Once they achieve the contrastive objective, they use the trained encoder for a downstream task.

The limitations of contrastive learning, such as sensitivity to the type of augmentation [38], and its reliance on a large quantity of negative pairs led the community to discover other alternatives. A subset of these alternatives attempts to achieve more efficient objectives by eliminating the reliance on the negative pair [12] or replacing the contrastive loss with other objectives [27]. Patacchiola et al. [27] leverage the RN architecture, which in turn allows optimizing the binary cross-entropy loss as a more efficient objective to learn a pretext task. Their method initiates by augmenting a mini-batch of images and passing them through a CNN backbone. They then train a relation head to discriminate the negative pairs (inter-reasoning) from the positive pairs (intra-reasoning). After the pretext objective is achieved, the relation head is discarded, and the backbone is used for a classification task. We will discuss this method in more depth in Sec. 3.

Most of the sports analytics studies on hockey focus on player identification, localization (of player and/or puck), and tracking [10, 18, 25, 41–44]. Vats et al. [43] proposed a transformer based approach along with a weakly supervised learning framework to identify players in hockey broadcast videos. Ludwig et al [24] demonstrated the effectiveness of using SSL with unlabeled videos and small set of labeled images for the task of 2D pose estimation from sports video.

In terms of action and interaction recognition in hockey, the study by Tora et al. [39] classifies multi-person puck possession events such as shot, dump in, using a CNN-RNN architecture. A group of studies propose methods for clas-

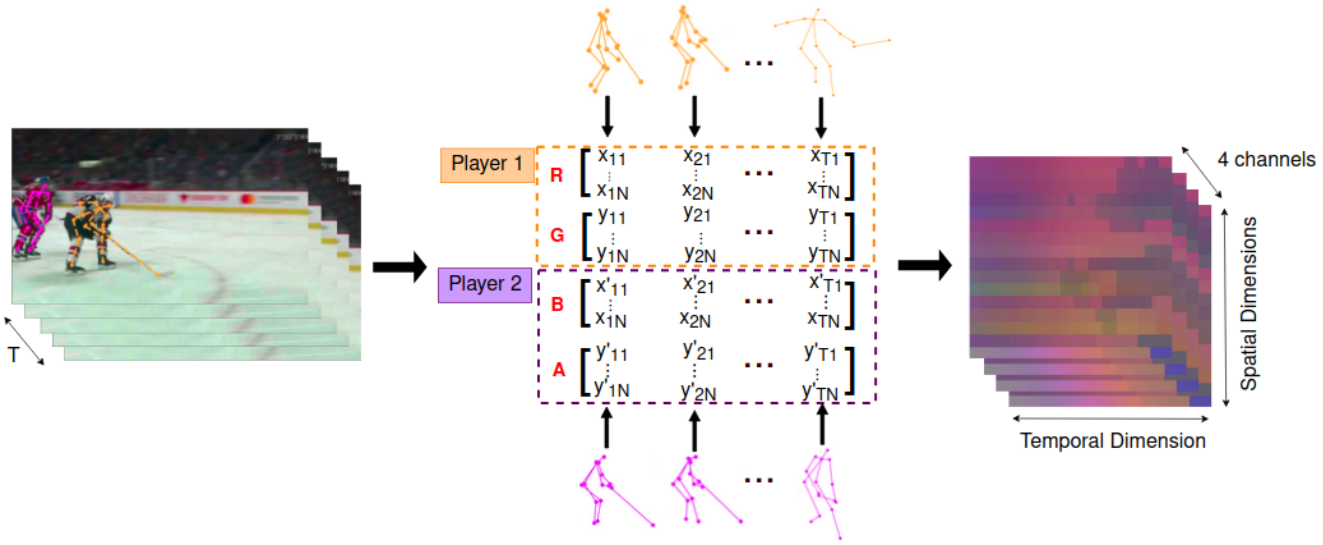

Figure 2. Data transformation pipeline. Pose skeleton sequences of two players for each video (with N keypoints and T frames) are transformed to spatio-temporal image-like representations with $N \times T \times 4$ dimensions.

sification of single-person actions [2, 9] or multi-person interactions [1], using pose and temporal components such as optical flow and/or recurrent neural networks. There are several papers [23, 34] on using SSL for activity recognition by using datasets such as UCF [19]; however, they are not specifically geared toward sports analytics. To the best of our knowledge, we are the first study applying SSL techniques for the downstream task of interaction recognition from hockey broadcast videos.

3. Method

Our goal in this paper is to profit from existing SSL approaches to achieve good performance on the complicated tasks of IRV from sports broadcast videos. Despite the popular usage of SSL approaches on large-scale datasets, we propose to use them on small-scale dataset and demonstrate their effectiveness. We are hoping leveraging SSL techniques opens a new door to sports analytics research by alleviating the need for large-scale sports specific dataset and eliminating expensive annotations.

Dataset: We use the hockey penalty dataset from the study by Askari et al. [1]. The dataset includes three classes of No penalty, Tripping, and Slashing with 98, 80, and 76 videos in each class respectively. The clips in this dataset are two to six seconds long with 30 fps, that are presented in either actual speed or slow-motion replays. Each penalty is completely encapsulated within the duration of the clip, meaning the clip starts several frames before the start of the penalty and ends several frames after the end of it. The dataset offers challenges such as significant view variation in terms of scale and angle, camera motion, (self) occlusions, blurry frames, and complex interactions. Fig. 1 demonstrates a few examples from the dataset.

The hockey penalty dataset includes ground-truth pose annotation for all the players in each clip. The annotations include 14 body key-points as well as two key-point for both ends of the hockey stick. The dataset marks the two main interacting players (i.e., P_1 and P_2) in each frames of each video. In the penalty scenes the interacting players are the two players directly involved in the penalty and in terms of the No penalty class, this assignment depends on the scenes. For example, for a goal event, the offensive player and the goaltender are considered as interacting players. Finally, each player's ID is unique and tracked for the duration of video [1].

Data transformation: given that most of the popular SSL methods (including our candidate) have been developed for image inputs, our first task is to transform temporal pose data from the videos into spatio-temporal image-like inputs. There are studies on different ways of transforming temporal pose data to image-like inputs [7, 8]; however, most of them are focused on single-person actions. Therefore, in our study, we propose a novel data transformation for interaction videos such that first, the transformation output is image-like (grid) data to enable us utilizing CNN backbones, second, the output preserves the original temporal dynamic and spatial structure of video, and third it captures information for more than one actor, so that it is suitable for the task of interaction recognition.

In order to extract spatio-temporal representations from each video, we define a matrix with four channels, similar to an RGBA image. Each of the channels is of size $T \times N$ with T and N representing the number of frames and number of

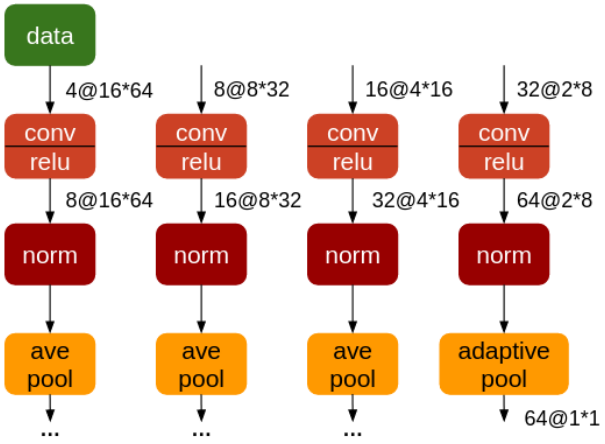

Figure 3. Conv-4 architecture. The network has four blocks of 8,16,32 and 64 feature maps. The convolutional layers are (3×3, stride=1, padding=1). The average pooling are (2×2, stride=2). Last layer has an adaptive average pooling to generate $c \times 1 \times 1$ maps (c number of channels).

joints, respectively. For instance, with (i, j) representing matrix indices, $R_i = [x_{11}, x_{21}, ..., x_{T1}]$ represents the location of the x coordinate of the first joint over T frames; and $R_j = [x_{11}, x_{12}..., x_{1N}]$ represents the x coordinates of all the N joints in the first frame (forming the first channel R_{ij}). The RGBA channels represent the same structure for x, y coordinates of P_1 and, P_2 respectively. Fig. 2 elaborates on our data transformation method.

SSL architecture: following the transformation explained above, our data is now ready to be the input of an image-based SSL method. Among the available methods, Relational-SSL approach proposed by Patacchiola et al. [27] offers great performance through an efficient objective of minimizing a binary cross entropy loss; in comparison with other computationally expensive methods using contrastive losses. In order for our work to be self-contained we briefly elaborate on the Relational-SSL approach; for more details on the method please refer to the original paper [27].

Relational-SSL follows the idea of many SSL methods [5, 35, 46], where the surrogate objective is to minimize the distance between the representations of positive pairs and maximize it for negative pairs. Relational-SSL combines this underlying idea with relational reasoning method [32] and proposes to model the relation between pairs as intra-reasoning for positive pairs and inter-reasoning for negative pairs. Employing the RN allows them to train a relation head through a simple classification task using the efficient objective of binary cross entropy loss.

Consider an unlabeled dataset that consists of sequences of poses for two people (P_1 and P_2) $D = \{S_m\}_{m=1}^{M}$ and $S_m = \{\{P_1\{x_n, y_n\}_{n=1}^{N}, P_2\{x_n, y_n\}_{n=1}^{N}\}_t\}_{t=1}^{T}$, where N, T, and M refers to number of body keypoints, frames, and data samples in the dataset, respectively. On each sample, several stochastic data augmentations are applied, $S_m^i = A(S_m)$, resulting in multiple augmented instances of each sample, K indicates the number of augmentations. After data transformation c, a CNN backbone, f_θ extracts representations from each instance. The representations are aggregated by the aggregation function (e.g., concatenation, summation, maximum) denoted by a to form pairs. These pairs can include augmented instances of a data point (positive pairs) or different data points (negative pairs). These pairs are input to a non-linear function with learnable parameters, r_ϕ, which is the building block of relational reasoning network. This function takes in each pair and outputs a relation score 0 or 1, with 1 indicating positive pair (intra-reasoning) and 0 indicating negative pairs (inter-reasoning). Finally, the loss is calculated between the relation score and target (denoted as g). Eq. (1) is the formulation of Relational-SSL, notations partially used as [27].

$$argmin_{\Theta,\Phi} \sum_{m=1}^{M} \sum_{i=1}^{K} \sum_{j=1}^{K} L(r_\Phi(a(z_m^{(i)}, z_m^{(j)})), g = 1) + L(r_\Phi(a(z_m^{(i)}, z_{m'}^{(j)})), g = 0) \text{ with } z_m = f_\Theta(c(S_m^{(i)})) \quad (1)$$

Augmentation: although we transformed the data into image-like inputs, it is important to note that ours are different from regular images (e.g., apples, cats, etc.) found in image classification datasets (e.g., CIFAR10). Consequently, the popular data augmentations used in image based SSL, including the ones used in the Relational-SSL (e.g., random crop-size and color distortion), are not meaningful and relevant in our case. Therefore, we propose to apply a set of augmentations in the time and skeleton space that are meaningful for sequence of pose data representing interactions. Specifically, for each batch augmentation we generate augmented versions of 2D skeleton data (i.e., positive pairs for intra-reasoning) through randomly applying augmentations such as rotation, translation, and shear followed by transforming the results to their spatio-temporal image-like representations. We will elaborate on the additional details about the augmentation in the Sec. 4.

Pose extraction: As mentioned in Sec. 1, given that pose annotation is an expensive process, we aim to evaluate the effect of using estimated poses (versus ground truth annotated poses) on the performance of SSL method and the downstream task. We chose the HRNet [36] frame-based top-down pose estimator, which takes bounding boxes for target people in each frame and output poses. Since our method is focused on recognizing the interaction between two main players, we only extract the poses for $P1$ and $P2$.

The hockey penalty dataset does not include bounding boxes for players; therefore, we automatically deduce

	Linear Evaluation			Finetune		
Method/Pose	Random Weights (lower bound)	Supervised (upper bound)	Relational-SSL	Random Weights (lower bound)	Supervised (upper bound)	Relational-SSL
GT pose	48.37 ± 0.45 %	89.28 ± 1.37 %	68.18 ± 0.0 %	n/a	100 ± 0.0 %	76.62 ± 2.7 %
Estimated pose	37.90 ± 1.845 %	80.125 ± 1.87 %	53.26 ± 2.3 %	n/a	84.76 ± 0.94 %	65.25 ± 4.3 %

Table 1. Experiment results of linear evaluation and finetune protocols using ground truth and estimated pose. We report percentage test set accuracy (mean and standard deviation of two runs) on hockey penalty dataset for the downstream task of penalty classification from hockey broadcast videos. Random weights and supervised setting represent the lower and upper bounds respectively

Backbone	Linear Evaluation
Relational-SSL (Conv-4)	**68.18 ± 0.0 %**
Relational-SSL (Resnet-8)	51.61 ± 1.37 %
Relational-SSL (Resnet-32)	53.56 ± 1.37 %

Table 2. Comparison of different backbones

Aggregation	Linear Evaluation
Relational-SSL (concatenation)	**68.18 ± 0.0 %**
Relational-SSL (summation)	66.66 ± 4.8 %
Relational-SSL (maximum)	64.28 ± 2.7 %

Table 3. Comparison of aggregation functions in the Relational-SSL

bounding boxes from ground truth pose data. The boxes are extracted to encapsulate all the joints' annotation, plus a correction margin to count for approximate location of key points annotation. The margin is set to twice the distance between neck and head key points to ensure each player is fully encapsulated by the box.

Since we use a pretrained model on the COCO dataset [21], the output includes 17 body keypoints; whereas, the hockey penalty dataset includes 14 body keypoints and two key points for the hockey stick. Therefore, we average the extra head keypoints (from the COCO format annotation) to one head keypoint and add the neck keypoint by averaging the shoulder keypoints. This process results to 14 estimated body keypoints (with the same format as hockey penalty dataset); we then add back the ground truth stick annotations (from the dataset) to the estimated pose to obtain the final pose with 16 keypoints. In the cases where HRNet completely fails to capture the poses of one or both of the players, we fill the missing pose with zeros, following the out-of-frame joint annotation protocol from the hockey penalty dataset [1]. Fig. 4 demonstrates a few examples of estimated poses using HRNet [36] (before adding the sticks back).

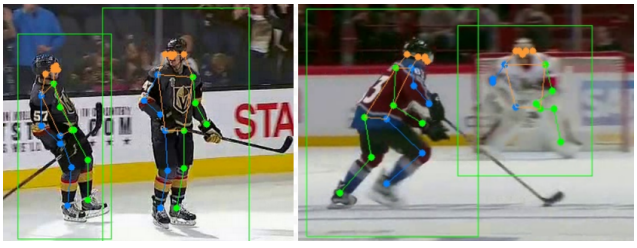

Figure 4. Outputs from HRNet [36] top-down pose estimator pretrained on COCO dataset [21]. Left: correct pose prediction. Right: failing to capture all the joints in motion-blurred frame.

4. Experimental Evaluation

4.1. Experiment setup

We uniformly sample 64 frames from each clip; unless the video is slow-motion, then we sample every third frame, given the motion between consecutive frames is negligible. We augment the dataset to three times its size using scale and horizontal flip. We utilize 50%, 30%, and 20% of the dataset as unlabeled set, labeled set for training (i.e., linear evaluation and finetuning), and test set respectively.

Mini-batch size and number of augmentations per batch (K) are 64. We use 2D skeleton rotation (0°-180°), translation (0% to 20% of image dimensions), and shear, with equal chances. As mentioned in Sec. 3 given our primary input data is skeleton, we apply augmentations in the skeleton space, followed by transforming the results to spatio-temporal image-like inputs.

The models are optimized using binary cross-entropy loss with the Adam optimizer and learning rate of 10^{-3} for 600 epochs for each of the backbone training and evaluation protocols. We used Conv-4 and Resnet-8 as our backbones. The Conv-4 backbone consists of four blocks of 8,16,32 and 64 feature maps with kernel size of three, stride and padding of one. Each block includes BatchNorm, ReLu, and average pooling with kernel size of two and stride of two. The last layer includes an adaptive average pooling to generate the maps with $c \times 1 \times 1$ dimensions (c number of channels) [27]. Fig. 3 demonstrates the network architecture. The relation head is a 256 unit fully connected layer with BatchNorm, leaky-ReLu and sigmoid output. The output representations from the relation head are aggregated using a concatenation function.

We follow standard evaluation protocols for SSL studies, meaning linear-evaluation and finetune [17]. The linear-evaluation includes training the backbone for 600 epochs on the unlabeled set, followed by training a layer of linear classifier on top of the backbone with labeled portion of the dataset for the downstream task of interaction classification in our study. In the linear-evaluation setting, during the classification training, the backbone weights are frozen. The finetune protocol is similar to linear-evaluation, except with backpropagation to the backbone weights during the training. For both protocols, the final metric is the accuracy of the downstream task on the test set portion of the dataset. It is important to note our data transformation method does not add significant computational cost; therefore, the overall cost is equivalent to the computational cost of the Relational-SSL model [27].

4.2. Results and ablation studies

In this section, we follow evaluation methods from [27] and discuss the results of our experiments. Tab. 1 demonstrates the test set accuracy using linear evaluation and finetuning protocols. In random weight experiments, the backbone weights are initialized randomly followed by linear evaluation, which sets the lower bound with $48.37 \pm 0.45\%$ test set accuracy. The upper bound is defined by the supervised setting, where the model has access to all the labels, which demonstrates $89.28 \pm 1.37\%$ and $100 \pm 0.0\%$ test set accuracy for linear evaluation and finetunning, respectively. Despite the small size of the hockey penalty dataset and using only 30% of labeled data (which is 50% less data compared to the supervised setting, considering the test set portion), the Relational-SSL method achieves test accuracy of $68.18 \pm 0.0\%$ for linear evaluation and $76.62 \pm 2.7\%$ for finetunning.

As demonstrated in Tab. 1, the results when using estimated pose across all the settings is worse compared to using ground truth pose. Across all the models, the performance drops by $10 - 15\%$. This is expected given the complexity, specially occlusions and motion blur, that challenge the HRNet pose estimator. In several cases, where one of the players is majorly occluded by another player, despite the bounding boxes provided, the pose estimator is unable to extract any keypoints for the occluded player. Given that the sequence of pose is the only input modality to the model, the low quality poses from the pose estimator significantly affect the performance of Relational-SSL and the downstream task.

We perform ablation studies on different components of the model. for all the ablation studies we present the results by reporting test set accuracy percentage (mean and standard deviation of two runs) after running Relational-SSL pretraining followed by linear evaluation. Among the CNN backbones we tested, we gained our best performance by using Conv-4 backbone (see Sec. 3), followed by Resnet-8 with $51.61 \pm 1.37\%$ and Resnet-32 $53.56 \pm 1.37\%$ (see Tab. 2). We observe the performance decreases as the number of backbone parameters increase, which is due to small size of the dataset and overfitting.

Additionally, we experimented with different aggregation methods of Relational-SSL. Similar to what several studies using relational network [27, 28, 37] report, we also find concatenation the most effective method for aggregating representations. Tab. 3 shows the maximum yields the lowest performance with $64.28 \pm 2.7\%$ accuracy. Moreover, we study the effect of amount of available labeled data and augmentations on the performance of the downstream task. Expectedly, the best performance is achieved when the Relational-SSL has access to all the labeled data with $74.02 \pm 0.91\%$. Finally, to demonstrate our approach is applicable to other image-based SSL method, we couple our data transformation approach (using GT pose) with deepinfomax [13] and report $51.64 \pm 1.4\%$ for linear evaluation and $53.26 \pm 0.93\%$ for finetunning setups.

5. Discussion and conclusion

In this paper, we demonstrated how to adapt the image-based Relational-SSL for the task of interaction recognition from sports broadcast videos. Specifically, we showed by using an effective data transformation and suitable augmentations, it is possible to adapt existing image-based SSL methods on sequence of pose data. Askari et al. [1] reports 80.6% accuracy on the interaction recognition from the hockey penalty dataset using a Recurrent Neural Network method fully supervised on sequence of pose data. In our method, however, we convert sequences of pose to image-like representations and use a CNN as backbone, which is not inherently an architecture to capture the temporal dynamics. Additionally, we only use 30% of the dataset as labeled data. Given these, $68.18 \pm 0.0\%$ and $76.62 \pm 2.7\%$ in linear evaluation and finetuning protocols demonstrate the effectiveness of our proposal.

Despite the popular approach of using SSL methods with large-scale dataset, we demonstrate the effectiveness on small-scale datasets as well; which is supported in another study by Cao et al. [3] as well. In many image-based SSL studies [27], when using large-scale complex datasets (such as CIFAR100), there is often $10 - 20\%$ of accuracy gap (using linear evaluation) between upper bound and SSL method. This is the case where the methods are specifically designed for image-based downstream task and trained on hundred thousand data points. However, in our study with less than a thousand data points, we achieve the same accuracy gap with the upper bound. Although, the number of classes in hockey penalty dataset is significantly less, but, the task is more complex and there is low inter-class variance.

Additionally, we observed the effect of using lower quality estimated poses, compared to ground truth pose, on the performance of the models. This emphasizes on the importance of the quality of pose data, specially, when it is used as the primary input form. Our experiments with estimated poses also serve as quantitative proofs of the claim by Askari et al. [1] that current off-the-shelf pose estimators struggle to extract high quality poses from complex sports scenes. Although, our method leverages the ground truth annotations of hockey stick, but it is important to note that, owing to the recent annotation technologies, annotating stick ends is much cheaper and easier compared to annotating the skeleton data. Our ablation studies also demonstrate that smaller backbones yield better results with small-scale datasets, and similar to many other studies that use relational reasoning architecture [27, 28, 37] concatenation is the most successful aggregation function.

In conclusion, using SSL methods is specially beneficial for the field of sports analytics, where there are abundant unlabeled data available through sports broadcast but creating sports-specific labeled datasets is time-consuming and expensive. Therefore, by taking advantage of available SSL approaches we can tackle complex tasks in sports analytics by leveraging the unlabeled data and requiring only a fraction of labeled data, bearing a small trade-off in performance compared to supervised learning. As the future work, our proposed method can be evaluated on the other publicly available skeleton-based interaction recognition datasets. Finally, this study can be further expanded to include more than two actors/players.

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. 1, 2, 3, 4, 6, 7, 8

[2] 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. 4

[3] Yun-Hao Cao and Jianxin Wu. Rethinking self-supervised learning: Small is beautiful. *arXiv preprint arXiv:2103.13559*, 2021. 2, 7

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3, 5

[6] 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. 2, 3

[7] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7024–7033, 2018. 4

[8] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, pages 579–583. IEEE, 2015. 4

[9] Mehrnaz Fani, Helmut Neher, David A Clausi, Alexander Wong, and John Zelek. Hockey action recognition via integrated stacked hourglass network. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 29–37, 2017. 4

[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] Xuehao Gao, Yang Yang, and Shaoyi Du. Contrastive self-supervised learning for skeleton action recognition. 3

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 3

[13] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 2, 7

[14] 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. 3

[15] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 1

[16] 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. 2, 3

[17] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1920–1929, 2019. 7

[18] 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

[19] 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. 1, 4

[20] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 3

[21] 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. 6

[22] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 816–833. Springer, 2016. 3

[23] 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. 4

[24] Katja Ludwig, Sebastian Scherer, Moritz Einfalt, and Rainer Lienhart. Self-supervised learning for human pose estimation in sports. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021. 3

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

[26] Sungheon Park and Nojun Kwak. 3d human pose estimation with relational networks. *arXiv preprint arXiv:1805.08961*, 2018. 3

[27] Massimiliano Patacchiola and Amos J Storkey. Self-supervised relational reasoning for representation learning. *Advances in Neural Information Processing Systems*, 33:4003–4014, 2020. 1, 2, 3, 5, 6, 7, 8

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

[29] Mauricio Perez, Jun Liu, and Alex C Kot. Skeleton-based relational reasoning for group activity analysis. *Pattern Recognition*, 122:108360, 2022. 1

[30] 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. 2, 3

[31] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Encouraging lstms to anticipate actions very early. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 280–289, 2017. 2

[32] 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. 3, 5

[33] Yuge Shi, Basura Fernando, and Richard Hartley. Action anticipation with rbf kernelized feature mapping rnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 301–317, 2018. 2

[34] 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. 4

[35] Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9631–9640, 2020. 3, 5

[36] 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 (CVPR)*, June 2019. 5, 6

[37] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 7, 8

[38] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020. 3

[39] 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

[40] 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. 2, 3

[41] 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. 3

[42] 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

[43] 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. 3

[44] 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

[45] 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. 2, 3

[46] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 3, 5