Cross-Conditioned Recurrent Networks for Long-Term Synthesis of Inter-Person Human Motion Interactions

Jogendra Nath Kundu* Himanshu Buckchash* Priyanka Mandikal Rahul M V
Anirudh Jamkhandi R. Venkatesh Babu
Video Analytics Lab, CDS, Indian Institute of Science, Bangalore, India
jogendrak@iisc.ac.in, hbuckchash@cs.iitr.ac.in, priyanka.mp2808@gmail.com,
rmvenkat@andrew.cmu.edu, anirudh.g.jamkhandi@gmail.com, venky@iisc.ac.in

Abstract

Modeling dynamics of human motion is one of the most challenging sequence modeling problem, with diverse applications in animation industry, human-robot interaction, motion-based surveillance, etc. Available attempts to use auto-regressive techniques for long-term single-person motion generation usually fails, resulting in stagnated motion or divergence to unrealistic pose patterns. In this paper, we propose a novel cross-conditioned recurrent framework targeting long-term synthesis of inter-person interactions beyond several minutes. We carefully integrate positive implications of both auto-regressive and encoder-decoder recurrent architecture, by interchangeably utilizing two separate fixed-length cross person motion prediction models for long-term generation in a novel hierarchical fashion. As opposed to prior approaches, we guarantee structural plausibility of 3D pose by training the recurrent model to regress latent representation of a separately trained generative pose embedding network. Different variants of the proposed frameworks are evaluated through extensive experiments on SBU-interaction, CMU-MoCAP and an inhouse collection of duet-dance dataset. Qualitative and quantitative evaluation on several tasks, such as Short-term motion prediction, Long-term motion synthesis and Interaction-based motion retrieval against prior state-of-the-art approaches clearly highlight superiority of the proposed framework.

1. Introduction

Human motion analysis has gained significant attention in the past years [4, 39, 35, 30] as a result of the availability of huge MoCAP (motion capture) datasets (such as CMU-MoCAP, DanceDB) and success of deep recurrent models [34, 2, 12] to analyze such highly non-linear structured temporal patterns. However, almost all prior approaches [25, 27, 9] completely ignore complex interperson motion interactions and focus solely on modeling 3D pose sequences of individual human characters for long-term synthesis and motion forecasting. Though interperson action recognition has been studied in certain prior works [38, 33], such methods are limited to simple short-term actions like kicking, punching, hugging, etc. However, interactions in duet performances like Salsa, Samba, Cha-cha, etc. are highly complex across both temporal and spatial aspects as compared to the above discussed simple motion interaction categories [6]. Directly extending available single-person motion modeling approaches for such complex duet interactions delivers suboptimal results without explicit design modifications [25]. Therefore, realizing an effective framework capable of modeling such complex spatiotemporal interactions among two human beings is a highly challenging task with applications in diverse industry domains, beyond graphical animation and video games. Such systems would be capable of simulating motion patterns of human looking virtual agents for applications in Virtual and Augmented Reality frameworks.

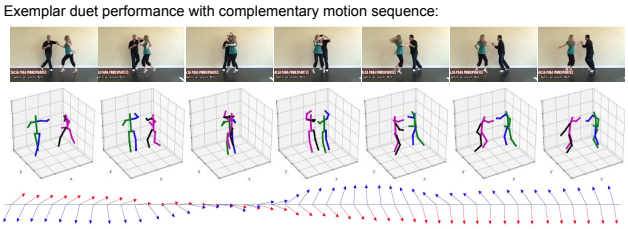

Figure 1. Illustration of complex 3D human motion as a result of coordinated complementary interaction between two characters in an exemplar duet dance performance (a sample from the in-house collection of Duet-Dance dataset). In last row, we show variations in absolute *facing-view* direction depicting complex emotional interaction beyond physical 3D pose.

Human motion is one of the most complex stochastic temporal process. For a given short-term temporal pose pattern, there exists a set of plausible future motion dynamics

*Equal contribution

as a result of various external stochastic factors. Therefore, long-term predictions in future may not match with the samples from the dataset. Acknowledging this, there is a clear distinction between quality assessment of a) short-term prediction versus b) long-term synthesis. In literature, a direct distance metric between the prediction and the corresponding ground-truth is used to benchmark models targeting short-term prediction task. However, quantitative assessment of methods targeting long-term synthesis is harder as result of the inherent stochasticity in future dynamics.

Challenges. Broadly, challenges in long-term synthesis can be attributed as follows:

a) Firstly, human pose in 3D is known to be highly structured [29, 28] as it is constrained by various kinematic factors like joint angle limits, relative bone-length constraints, limits imposed by earth gravitation, etc. Therefore, the predicted 3D pose should be bounded by the structural feasibility limits [1]. Most of the prior motion modeling approaches [25, 27, 10] ignore this simple yet crucial constraint, by aiming to directly regress the 3D joint locations. Such approaches end up generating implausible pose patterns during long-term synthesis as a result of recursive error accumulation [25].

b) Secondly, as compared to simple actions like walking, running, hand-shaking, kicking, etc., long-term synthesis of complex interaction-based acts like duet dance, martial arts, etc. constitutes highly complex pose sequences (operating close to the boundaries of kinematic structural limits as seen in gymnastics), which are often highly non-periodic or repeated after a very high periodicity gap.

c) Thirdly, employing Recurrent Neural Networks (RNN) in a fully auto-regressive setting [27, 5] has been shown to result in motion stagnancy or convergence to unrealistic motion patterns. The accumulation of feedback error in an auto-regressive recurrent framework is one of the major challenges in any sequence modeling task (not specific to human motion synthesis) [9, 20]. Unrolling the RNN for a much longer time-step by taking its own noisy predictions as input conditioning inevitably derails the network's learned capability [25], as a result of the input test sequences being far away from the training set distribution.

Our contributions. In this paper, we aim to address each of the concerns discussed by formalizing novel architectural modifications. Our major contributions in this paper are outlined below.

a) Aiming to impose a strong structural constraint on 3D pose estimation, we learn a view-invariant generative latent representation after disentangling external global variations. Following this, the recurrent model is restricted to regress latent representations satisfying the predefined prior, which guarantees generation of realistic 3D pose.

b) As opposed to prior arts, we leverage positive implications of both auto-regressive and encoder-decoder recurrent architecture, by interchangeably utilizing two separate fixed-length cross person motion prediction models for long-term generation in a novel hierarchical fashion. As a result, we are able to realize a model, which can perform long-term generation of inter-person interactions for several minutes (300+ Seconds) devoid of the issues related to motion stagnancy and divergence to unrealistic pose sequence.

c) Different variants of the proposed framework are evaluated through extensive experiments on several tasks, such as Short-term motion prediction, Long-term motion synthesis and Interaction-based semantic motion retrieval against prior state-of-the-art approaches, which clearly highlights superiority of the proposed framework.

2. Related work

In the past, the task of modeling multi-person social interactions has largely been addressed for very simple human action categories [24] like kicking, pushing, punching, etc. and human-human interactions [37]. Representations such as relative joint distance [38, 14, 15] were employed by these methods, which fail in cases of large variation in relative distances between the two performers over time (a common scenario in complex duet human interactions). Moreover, none of these approaches are designed to address long-term synthesis [25], which is the primary goal in the proposed framework.

Recently, RNNs have shown significant progress in various sequence modeling tasks, such as generation of text [34], image captioning [36] and also generation of hand-written characters [11, 13]. However, for human motion generations [4, 20, 18, 17, 22] one of the major challenges is to preserve the structural regularity among individual joints in the generated poses thus avoiding unrealistic pose frames. Fragkiadaki *et al*. [9] propose to jointly learn structural skeleton embedding along with the temporal motion sequence by employing an encoder-recurrent-decoder (ERD) architecture. Jain *et al*. [20] proposed structural-RNN, a spatio-temporal graph architecture to effectively capture the interaction among skeleton joints across the temporal dimension. Martinez *et al*. [27] employ a sequence-to-sequence architecture to model short-term human motion forecasting. However, all the above approaches fail to generalize for long-term synthesis, as a result of motion stagnancy or convergence to unrealistic motion pattern [25] after just several seconds.

acRNN [25] proposed an auto-conditioned RNN framework to explicitly address the problem of long-term synthesis. Instead of only feeding in ground-truth instances, they use sub-sequences of the network's own outputs at periodic intervals. Though acLSTM [25] demonstrates long-term realistic motion synthesis, such architecture is not suitable for modeling interaction-based complex motion patterns.

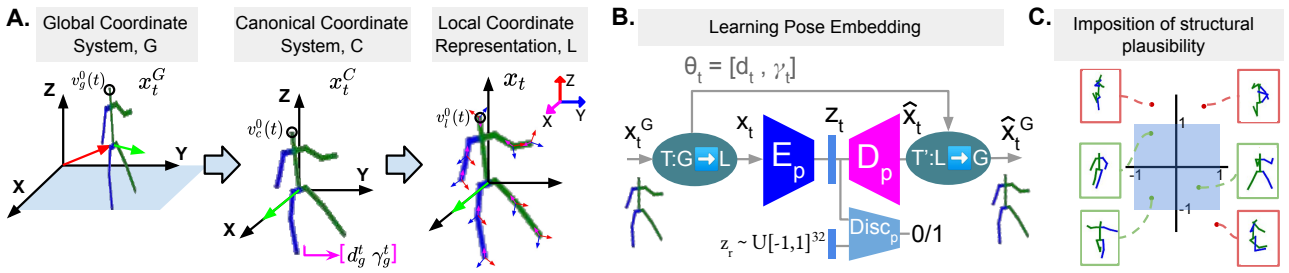

Figure 2. Transformation for obtaining view-invariant skeleton pose followed by generative learning setup for obtaining the pose-embedding representation. On the rightmost figure, we show how the learned embedding restricts estimation of implausible pose patterns (red-box: implausible pose, green-box: plausible pose).

3. Approach

This section describes different components of the proposed cross-conditioned recurrent network architecture. It first provides details of the proposed pose embedding framework to explicitly impose strong structural constraints on the predicted pose sequence in the later stage. Next, we detail different variants of the cross-conditioned recurrent architecture followed by a hierarchical auto-regressive scheme for efficient long-term synthesis.

3.1. Learning pose embedding

The core objective behind separately training a pose embedding representation is to disentangle enforcement of structural constraints in the subsequent steps for the final motion synthesis framework. While generating human motion, the model should completely restrain from generating unrealistic pose patterns. Moreover, the generated pose should strictly follow the structural constraints [1] such as joint angle limits, fixed bone length, limb interpenetration restrictions etc. To realize this in the most effective form, we plan to completely disentangle global variations from the final pose representation.

View-invariant skeleton representation. Formally, let the pose in Global coordinate system (G) for a skeleton sample x_t^G be, $v_g^j(t) = [a_t^j, b_t^j, c_t^j]^T$, where $j \in (1, .., J)$ with J being the total number of skeleton joints. A canonical direction, v_t^n is defined to uniquely represent view-invariant pose coordinates, $v_c^j(t)$. v_t^n is obtained as the vector perpendicular to the XY-projection of the line segment joining the left-hip and right-hip point of $v_g^j(t)$. Following this, the global translation, $d_t^g = [a_t^g, b_t^g, c_t^g]^T$ and azimuth view-angle, γ_t^g (angle between v_t^n and the positive X-axis) is separated to form a new set of joints $v_c^j(t) = \mathcal{R}_{\gamma_t^g} \times (v_g^j(t) - d_t^g)$, with $\mathcal{R}_{\gamma_t^g}$ being the corresponding rotation matrix. Here d_t^g is the coordinate of the pelvis point in G. To achieve scale-invariance, we then normalize the bone lengths to that of a chosen canonical skeleton, and convert v_c^j to local parent coordinates for selected end-limb joints according to the kinematic structure of human skeleton. We follow Gram-Schmidt based approach inline with Akhter *et al*. [1] to obtain the pose representation for limb joint locations in the Local Coordinate System, L as, $v_l^j(t)$. We denote it as, x_t (i.e. $v_l^j(t)$ for $j = 1, ..J$), which effectively disentangles difficulties of modeling global position, camera-view and joint-location with respect to a single pelvis root.

Training of pose embedding network. We define encoder E_p, decoder D_p and discriminator $Disc_p$ following similar architecture inspired from the kinematic tree of limb connections [8]. The inference network, E_p takes the previously defined x_t as input representation as shown in Figure 2B. Motivated from adversarial auto-encoder framework [26], to effectively leverage both continuity of pose embedding along with reduced reconstruction error, we employ an adversarial learning strategy inline with [23] against a predefined prior distribution $z \sim U[-1, 1]^{32}$. We denote the final pose embedding as $z_t = E_p(x_t)$ for each time t.

3.2. Cross-conditioned recurrent architecture

Before describing network architecture of the proposed cross-conditioned recurrent network, we discuss pros and cons of the two varied school of thoughts addressing sequence modeling problems.

a) Auto-regressive recurrent architecture. In general, auto-regressive recurrent networks are the best candidate for motion synthesis as utilized by various previous arts [27, 25]. Here a recurrent network can be interpreted as a function with hidden memory representation h_t i.e. $AB_{t+1} = f(AB_{1:t}, h_t)$, where AB_{t+1} is a combined representation of A_{t+1} and B_{t+1} at $(t+1)$th time-step as shown in Figure 3A. Intuitively, h_t stores important semantic information till tth time-step, to help prediction at $(t+1)$th time-step. One can use LSTM [16] or GRU [7] based architecture as special type of RNNs with improved training capability. However, such approaches exhibit several short-comings, as a result of *error accumulation* while unrolling for longer time-steps [27, 25] during synthesis. Note that, such models are trained on ground-truth input sequence. However, during testing they are expected to work on input sequence obtained from the noisy prediction of the previous time-

step. The discrepancy between prediction and ground-truth sample distribution progressively derails the model from its expected behavior, resulting in stagnated motion or divergence to unrealistic pattern [25]. Previous approaches employing auto-regressive techniques [27, 9, 20] show similar failure trends while generating long-term sequences beyond 1-2 seconds.

b) Encoder-decoder recurrent architecture. Aiming to alleviate the above problems, certain prior approaches [19, 18] decided to use auto-encoder like architecture (convolutional or recurrent) by employing separate encoder and decoder model instead of a single RNN as used in auto-regressive framework. However, such models can only work for generation of a *fixed-length sequence* as a result of the separate decoder architecture. Unlike auto-regressive, here the decoder is trained to take the compact embedding obtained from a fixed length sequence (encoded representation), to predict a sequence of the exact same length, without explicitly accessing the expected ground-truth sequence. Though such models are devoid of the shortcomings exhibited by auto-regressive models (i.e. discrepancy in training versus testing scenarios), they are not suitable for long-term synthesis as a result of the restricted sequence length constraint.

Our architecture. We plan to exploit the positive implications of both the above approaches i.e. a) fixed-length decoding and b) auto-regression, by formalizing the problem as an alternating cross-motion prediction pipeline. Aiming to effectively model duet inter-personal motion patterns, we design a novel recurrent base-model which is not only capable of predicting one's motion from the other but also forecasts pose sequence of the predicted motion without accessing corresponding input from the other (see Figure 3B). We refer the 3D pose sequence of person-1 and person-2 for m time-steps as $[A_1, A_2, .., A_m]$ and $[B_1, B_2, .., B_m]$ respectively. To realize this, we first introduce two separate RNNs for prediction and forecasting, i.e. *pRNN* and *fRNN* respectively. We refer the 3D pose sequence of person-1 and person-2 for m time-steps as $[A_1, A_2, .., A_m]$ and $[B_1, B_2, .., B_m]$ respectively. As shown in Figure 3B, $pRNN^{A\rightarrow B}$ takes a sequence of A as input representation, i.e. $A_{1:m}$, while outputting the sequence of B, i.e. $B_{1:m}$ at the corresponding time-steps. Following this, the hidden state of $fRNN^{A\rightarrow B}$ is initialized through a two-layer neural network transformation of the final hidden-sate of $pRNN^{A\rightarrow B}$ to forecast the future motion sequence of B for n-steps, denoted as $B_{m+1:m+n}$. We refer this combination of $pRNN^{A\rightarrow B}$ followed by $fRNN^{A\rightarrow B}$ as $ccRNN^{A\rightarrow B}$ as shown in Figure 3B. Note that, in certain duet dance forms there is a notion of leader and follower, where one of the two persons acts as a leader with the other person complementing the first during transition in dance steps. Motivated from this, we plan to model two separate recurrent models

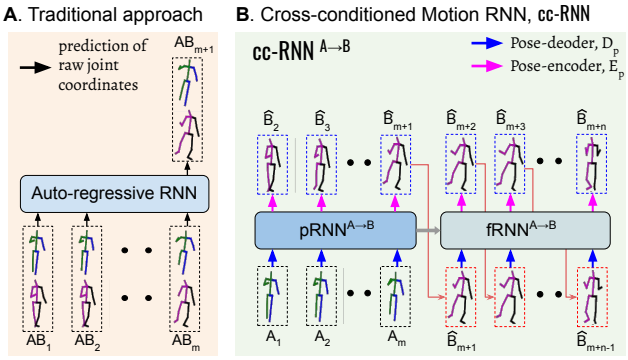

Figure 3. A. Illustration of extending traditional motion modeling framework for interaction-based duet motions. B. An overview of the proposed *ccRNN* framework operating on the pose-embedding using D_p and E_p (see Section 3.2).

for prediction of one's motion sequence from the other's, i.e. $ccRNN^{A\rightarrow B}$ and $ccRNN^{B\rightarrow A}$ respectively.

3.2.1 End-to-end training on frozen pose embedding

As explained in Section 3.1, the forward kinematic transformation $\mathcal{T}$ converts the raw joint locations in G to unit vectors in local coordinates after disentangling translation and view-angle as shown in Figure 2B. Similarly the inverse transformation $\mathcal{T}'$ performs inverse kinematics to obtain the skeleton coordinates back in to the Global coordinate system. We model the view-angle γ as a two dimensional feature $[\sin\gamma, \cos\gamma]$ (with *tanh* non-linearity) where $\gamma = \tan^{-1}(\sin\gamma/\cos\gamma)$ with unit magnitude normalization, maintaining cyclic property of the angle prediction.

Unlike traditional approaches (black arrows in Figure 3A), in the proposed *ccRNN* architecture the RNN operates on the pose embedding representation as both input and output sequence, obtained from the frozen E_p and D_p followed by $\mathcal{T}$ and $\mathcal{T}'$ transformations (blue and pink arrows in Figure 3B). Essentially the effective input and output sequence to *ccRNN* is represented as a tuple, $[z_t, d_t, \gamma_t]$ at each time-step t. Considering the fully differentiable transformations E_p, D_p, $\mathcal{T}$ and $\mathcal{T}'$, we formalize a single end-to-end loss directly on the output skeletons (i.e. $v_g^j(t)$) as, $\mathcal{L}_i^A = \sum_{j=1}^{J} |v_g^j(t) - \hat{v}_g^j(t)|$. This greatly stabilizes the training procedure by automatically balancing gradients among the disentangled factors as opposed to having independent losses with additional hyper-parameter for loss balancing. Note that, the *tanh* non-linearity on the logits predicting an estimate of z_t, constraints the model to predict a plausible human pose pattern as a result of the identical predefined input prior distribution of D_p, during the generative pose-embedding training (see Figure 2C).

Considering a particular canonical structure, we always make $\gamma_t^A \leftarrow \gamma_t^A - \gamma_0^A$, $d_t^A \leftarrow d_t^A - d_0^A$ and $\gamma_t^B \leftarrow \gamma_t^B - \gamma_0^A$, $d_t^B \leftarrow d_t^B - d_0^A$ to normalize both view-direction and

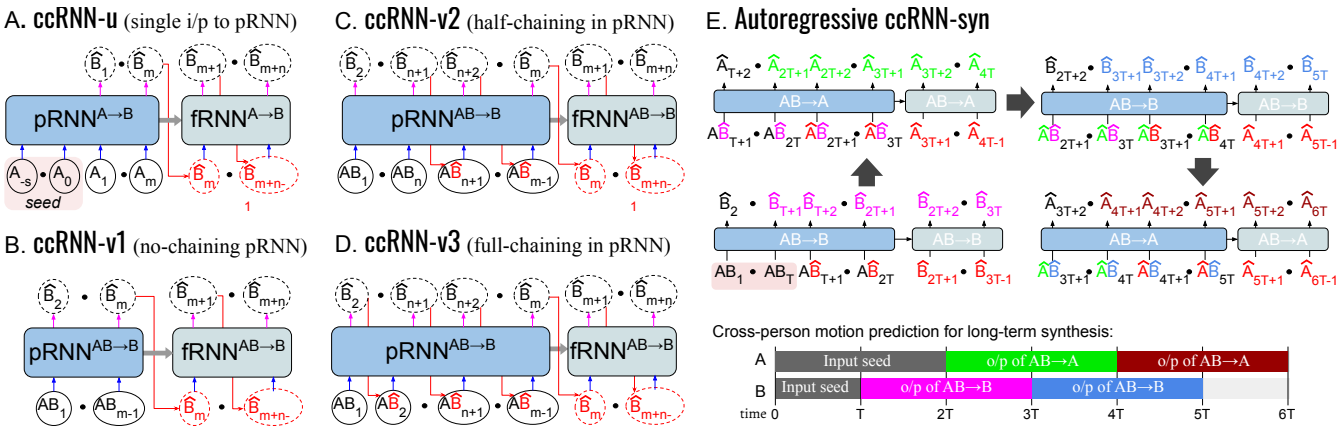

Figure 4. Illustration of different variants of *ccRNN* framework followed by the proposed hierarchical autoregressive pipeline to regularize the model for long-term synthesis (best viewed in color).

translation of A and B. In both Figure 3 and 4, $A_t = x_t^G$ for person-1 and similarly $B_t = x_t^G$ for person-2.

3.2.2 Variants of cross-conditioned recurrent model

We analyze 4 different cross-motion prediction frameworks as shown in Figure 4. Note that, for illustration we only show $ccRNN^{A\rightarrow B}$, however we also train the other counterpart i.e. $ccRNN^{B\rightarrow A}$ for all the 4 variants discussed below (i.e. for Figure 4A, 4B, 4C and 4D). The two-way cross prediction models are required for effective long-term synthesis as explained in Section 3.3.

ccRNN-u. Here we consider a single person as the input sequence with a seed of s time-steps as shown in Figure 4A. For the $A \rightarrow B$ model, *pRNN* is used to regress the pose sequence of B, $\hat{B}_{1:m}$ conditioned on the input seed of A ($-s$ to 0 time-step) along with the pose sequence of A at the corresponding time-steps i.e. $A_{1:m}$. Whereas *fRNN* is employed in a autoregressive setting with chained input from past prediction (shown in red), even during training.

ccRNN-v1. Avoiding the use of seed-sequence, we plan to model *pRNN* as $AB \rightarrow B$ unlike $A \rightarrow B$ used in *ccRNN-u*. This design structure is incorporated to reduce uncertainty in the prediction of B from A, as here each prediction of B is conditioned on the past dynamics of a combined representation of both A and B (see Figure 4B).

ccRNN-v2, v3 and v4. One of the major short-coming of both *ccRNN-u* and *ccRNN-v1* is that, the *pRNN* employed in both these architectures always takes ground-truth sequence as input, which makes them unsuitable for the hierarchical auto-regressive framework. Note that, a suitable candidate for the hierarchical auto-regression should support noisy predictions as input sequence to the corresponding *pRNN* model, adapting it for the test scenario as required for long-term synthesis. Following this, we propose two different variants of *ccRNN-v1* denoted as *ccRNN-v2* and *ccRNN-v3* as shown in Figure 4C and 4C respectively. Note that, the only difference here is the input to *pRNN*, which is modeled as a half-chaining and full-chaining setting for *ccRNN-v2* and *ccRNN-v3* respectively (see the red lines in Figure 4C-D). We also define another model named as *ccRNN-v4*, which is simultaneously trained on all the three architectural settings at alternate training iterations i.e. *ccRNN-v1, v2 and v3* to realize a unified model suitable for both short-term and long-term prediction.

3.3. Hierarchical auto-regressive, *ccRNN-syn*

All the individual variants of *ccRNN* discussed above are trained to operate on a sequence of fixed length for both *pRNN* and *fRNN*. However, we plan to use them as building blocks for a hierarchical auto-regressive framework, capable of synthesizing duet motion sequence for much longer (ideally indefinite) time-steps. As shown in Figure 4E, we require an initial seed (ground-truth) of $A_{1:2T}$ and $B_{1:T}$ to start the hierarchical autoregression pipeline. We first obtain a prediction of $\hat{B}_{T:3T}$ (pink) using the half-chaining architecture discussed for *ccRNN-v2*. Note that, $\hat{B}_{T:3T}$ is a temporal concatenation of $\hat{B}_{T+1:2T}$ and $\hat{B}_{2T+1:3T}$ as output of *pRNN* and *fRNN* respectively (see the bar-plot at the bottom of Figure 4E). Following this, the next $AB \rightarrow B$ model takes the previously predicted $\hat{B}_{T:3T}$ for further synthesis of $\hat{A}_{2T+1:4T}$. However, this model still have access to some seed information i.e. $A_{T+1:2T}$ which will not be the scenario for the synthesis of further time-steps. Such an auto-regressive reuse strategy is repeated for 4 times with 2 time usage of both $AB \rightarrow A$ and $AB \rightarrow B$ during training by constructing random mini-batches of 6T length as shown in Figure 4E. To further regularize the model and to prepare it for the test scenarios encountered in long-term synthesis, we design a weighted loss function to fine-tune the model with parameters initialized from the fixed-length training of *ccRNN-v4*, as

$$L_{syn.} = \sum_{i=2T+1}^{6T} w_{i-T}\mathcal{L}^{A_i} + \sum_{i=T+1}^{5T} w_{i-T}\mathcal{L}^{B_i}; w_i = e^{1/2T}$$

While training, we also consider the other counterpart, where seed starts from a longer sequence of B i.e. $B_{1:2T}$ with a smaller sequence of A i.e. $A_{1:T}$ to balance out all cross conditional synthesis possibilities one can encounter during testing. In this case, the above loss function is modified by replacing A with B and vice-versa. We denote this model as *ccRNN-syn* in further sections of this paper.

4. Our *DuetDance* Dataset

Existing datasets such as CMU-MoCAP and SBU-interaction only cater to model dynamics of comparatively simpler (highly periodic) single-person activities such as walking, running, exercising, etc. Aiming to address the deficiency of training data for multi-person interactions, we release a multi-person interaction dataset, which we refer to as '*DuetDance*'. Our dataset comprises of a variety of dance performances, and encompasses long-term correlations between physical motions, and hence serves as a benchmark to evaluate methods that aim to model multi-person interactions.

We carefully curate dance tutorials on YouTube to collect a diverse set of complex dances such as Cha-cha, Jive, Rumba, Salsa, and Samba. To obtain the ground truth 3D skeletons from the collected videos, we run LCRNet++ [31] on each individual frame. Here, we outline the issues we encountered in the skeleton outputs and ways in which we address them: **(1) Person Tracking**: While LCRNet detects people in the scene, it does not track them throughout the video. Tracking the identity of the two dancers however is crucial for our task. We annotate the collected videos for the male and female dancers across all frames. **(2) Occlusion**: In cases of inter-person occlusion, LCRNet++ often detects a single person as it does not capture any temporal consistency in the data. We therefore automatically detect such frames and adopt an interpolation technique to generate the ground truth for the missing dancer. **(3) Jitter**: Additionally, we found that a lack of temporal consistency leads to the problem of motion jitter in successive frames. We apply the savgol filter [32] on the pose sequence, to obtain smooth ground truth motion.

5. Experiments

In this section, we present additional training details, qualitative comparisons on long-term duet-dance synthesis, and quantitative comparisons to state-of-the-art methods on motion forecasting task. Furthermore, we evaluate robustness of the learned motion representation on a cross-person motion retrieval task.

Training Details. We make use of LSTMs implemented in the tensorflow framework, with a hidden state size of 256. All videos are normalized to 30 fps with $T = 10$. Our model converges in 300 epochs of training using the ADAM backpropogation algorithm [21] with a learning rate of 0.0001, and a batch size of 32.

5.1. Comparison against prior arts

Here, we outline the various distinguishing characteristics of our method in comparisons to previous arts and present quantitative comparisons on various tasks such as Short-term motion prediction, Cross-person retrieval, and Long-term synthesis.

Characteristic Comparisons. Table 1 depicts the different tasks that our method addresses, in comparison to previous approaches. In contrast to previous methods which separately address the task of learning a motion embedding and performing long term motion synthesis, ours is the first method that is capable of addressing both these tasks using a single network as a result of the architecture, with mixed characteristics of both auto-regressive and encoder-decoder as discussed in Section 3.2.

Dataset Selection. To validate our model generalizability for both simple as well as complex two person interactions, we evaluate our method on a) our in-house duet dance dataset, and the publicly available b) CMU-Interaction dataset and c) SBU-Interaction dataset. Both these publicly available datasets consist of simple interactions such as shaking hands, walking together, kicking, punching, etc as opposed to complex long-term interactions present in duet dance forms.

Quantitative Comparison on short-term prediction. In Table 2, we compare the different variants of *ccRNN* to previous methods such as [27, 25, 3] considering the task of motion forecasting. For the short-term forecasting task, an average prediction error (at the output of *fRNN*) on a held out test set is obtained using the corresponding $A \rightarrow B$ and $B \rightarrow A$ variants of *ccRNN-u* and *ccRNN-v1*, whereas for *ccRNN-syn*, $AB \rightarrow A$ and $AB \rightarrow B$ is entangled as discussed Section 3.3. The results in Table 2 portrays the superiority of *ccRNN*, as a result of our carefully chosen design choices. We also observe that, the imposition of structural constraints via the pose embedding helps us gain a signifi-

Table 1. Characteristic comparison of *ccRNN* against the previous motion modeling approaches.

Methods	Autoencoding Methods [19]	Autoregressive (acLSTM [25])	*ccRNN* Ours
Motion Embedding	✓	×	✓
Content-based Retrieval	✓	×	✓
Long Term Synthesis	×	✓	✓
Variable Length Synthesis	×	✓	✓
Cross-person Retrieval	×	×	✓

Table 2. Quantitative results on short-term prediction task. Here, we find that ccRNN-syn's capability to perform long-term synthesis comes at the cost of inferior motion forecasting performance when compared to the other variants.

DuetDance Dataset															
Methods	Cha-Cha			Jive			Rumba			Salsa			Samba		
	80ms	320ms	640ms	80ms	320ms	640ms	80ms	320ms	640ms	80ms	320ms	640ms	80ms	320ms	640ms
Short Term Forecasting Methods															
ERD [9]	0.17	0.45	1.33	0.20	0.39	1.44	0.24	0.66	0.98	0.15	0.61	1.21	0.22	0.51	1.22
seq2seq [27]	0.20	0.42	1.38	0.23	0.35	1.39	0.21	0.69	1.08	0.19	0.55	1.31	0.19	0.42	1.01
sch. smp. [3]	0.25	0.45	1.09	0.19	0.30	1.28	0.26	0.60	1.18	0.21	0.58	1.35	0.23	0.47	0.92
ccRNN-u	0.15	0.40	1.09	0.19	0.35	1.03	0.21	0.69	0.90	0.12	0.58	1.15	0.20	0.49	1.17
ccRNN-v1	**0.09**	**0.32**	**0.97**	**0.14**	**0.29**	**0.92**	**0.09**	**0.56**	**0.82**	**0.07**	**0.44**	**0.97**	**0.11**	**0.37**	**1.02**
Long Term Synthesis Methods															
acLSTM [25]	1.22	2.10	2.92	1.44	2.56	3.02	1.09	2.33	2.54	0.98	2.76	3.23	1.02	2.35	2.88
ccRNN-v2	**1.01**	**2.02**	**2.73**	**1.21**	**2.36**	**2.91**	**0.98**	**2.23**	**2.49**	**0.91**	**2.26**	**3.01**	**0.92**	**2.01**	**2.50**
ccRNN-syn	1.14	2.06	2.86	1.30	2.42	2.98	1.05	2.28	2.53	0.96	2.52	3.19	0.98	2.20	2.68

SBU Dataset													**CMU Dataset**		
Methods	Kicking			Punching			Hugging			Shaking Hands			All classes		
	80ms	320ms	640ms	80ms	320ms	640ms	80ms	320ms	640ms	80ms	320ms	640ms	80ms	320ms	640ms
Short Term Forecasting Methods															
ERD [9]	0.25	0.69	1.83	0.35	0.45	1.78	0.29	0.76	1.18	0.25	0.69	1.15	0.19	0.59	1.42
seq2seq [27]	0.21	0.55	1.65	0.31	0.48	1.95	0.33	0.66	1.18	0.29	0.59	1.01	0.29	0.52	1.51
sch. smp. [3]	0.25	0.45	1.09	0.19	0.30	1.28	0.26	0.60	1.18	0.21	0.58	1.35	0.23	0.47	0.92
ccRNN-u	0.23	0.64	1.49	0.39	0.41	1.63	0.31	0.73	0.98	0.20	0.57	1.01	0.24	0.55	1.10
ccRNN-v1	**0.11**	**0.44**	**1.32**	**0.24**	**0.33**	**1.22**	**0.19**	**0.59**	**0.87**	**0.09**	**0.49**	**0.92**	**0.15**	**0.44**	**0.99**
Long Term Synthesis Methods															
acLSTM [25]	1.34	2.27	2.98	1.56	2.63	3.42	1.29	2.36	2.51	1.09	2.89	3.21	1.33	2.29	2.72
ccRNN-syn	**1.28**	**2.15**	**2.53**	**1.41**	**2.49**	**2.82**	**1.15**	**2.29**	**2.48**	**0.98**	**2.53**	**3.11**	**1.05**	**2.14**	**2.66**

cant advantage over the prior arts.

Quantitative Comparison on long-term synthesis. To quantify the efficacy of the proposed *ccRNN* framework against prior arts for the task of long-term motion synthesis, we train a critic model (bi-directional LSTM with 256 hidden units) to discriminate between ground-truth versus predicted motion sequence of 60 sequence length. The predicted sequences of both prior arts and different variants of our method are taken as negative samples against the positive ground-truth sequences for training a critic model. After the critic training accuracy has saturated, we can make use of it to ascertain the efficacy of our predictions. If a certain set of predictions achieves a higher critic accuracy then it means that the critic is easily able to classify the predictions as fake, whereas a lower accuracy (close to 50%) would mean that the predictions are realistic enough to fool the discriminator, Table 3 clearly shows our superiority on long-term synthesis as *ccRNN* generates high-quality realistic non-stagnating motion patterns as compared to prior arts. A qualitative analysis of our long-term generation results against acLSTM [25] is presented in Figure 5.

Table 3. Quantitative results on long-term synthesis. Comparison of critic accuracy for binary discrimination of real versus predicted motion sequence of a chunk length 2 seconds taken around the time-steps 1s, 5s, 10s and 20s. (classifier accuracy on real test samples: 50.2%). A higher value indicates worse performance, as the critic is able to discriminate between the real versus predicted motion patterns.

Methods	Critic Accuracy			
	1s	5s	10s	20s
Short Term Forecasting Methods				
ERD [9]	0.63	0.78	0.84	0.85
seq2seq [27]	0.64	0.77	0.82	0.83
sch. smp. [3]	0.62	0.79	0.84	0.84
ccRNN-u	0.56	0.70	0.77	0.83
ccRNN-v1	**0.54**	**0.69**	**0.76**	**0.81**
Long Term Synthesis Methods				
acLSTM [25]	0.60	0.68	0.75	0.86
ccRNN-syn (w/o pose embedding)	0.58	0.65	0.72	0.79
ccRNN-syn	**0.51**	**0.56**	**0.58**	**0.64**

5.2. Application to Cross-Motion Retrieval

To evaluate robustness of the learned motion embedding, we setup two different skeleton retrieval tasks.

Retrieval settings Consider $q_a^{\alpha:\alpha+T}$ to be a given query skeleton sequence taken as a chunk of frames from time $\alpha(q_a) : \alpha(q_a) + T$ of a video-index $v(q_a)$ which belongs to a motion class denoted by $mc(q_a)$. The retrieval database consists of all the complementary skeletons, $\mathcal{D}_r = d_b(i)$. We consider, $r_b = argmin_{d_b}(||q_a - d_b(i)||_2)$ as the retrieved video-clip.

a) Motion-class retrieval First, we consider the task of retrieving a video-clip of the same motion class as the query sequence, from a database of video-clips. Here, we consider a retrieval as True Positive if $mc(q_a) = mc(r_a)$.

b) Semantic-motion retrieval Next, we consider the task of retrieving a video-clip which has the most similar pose

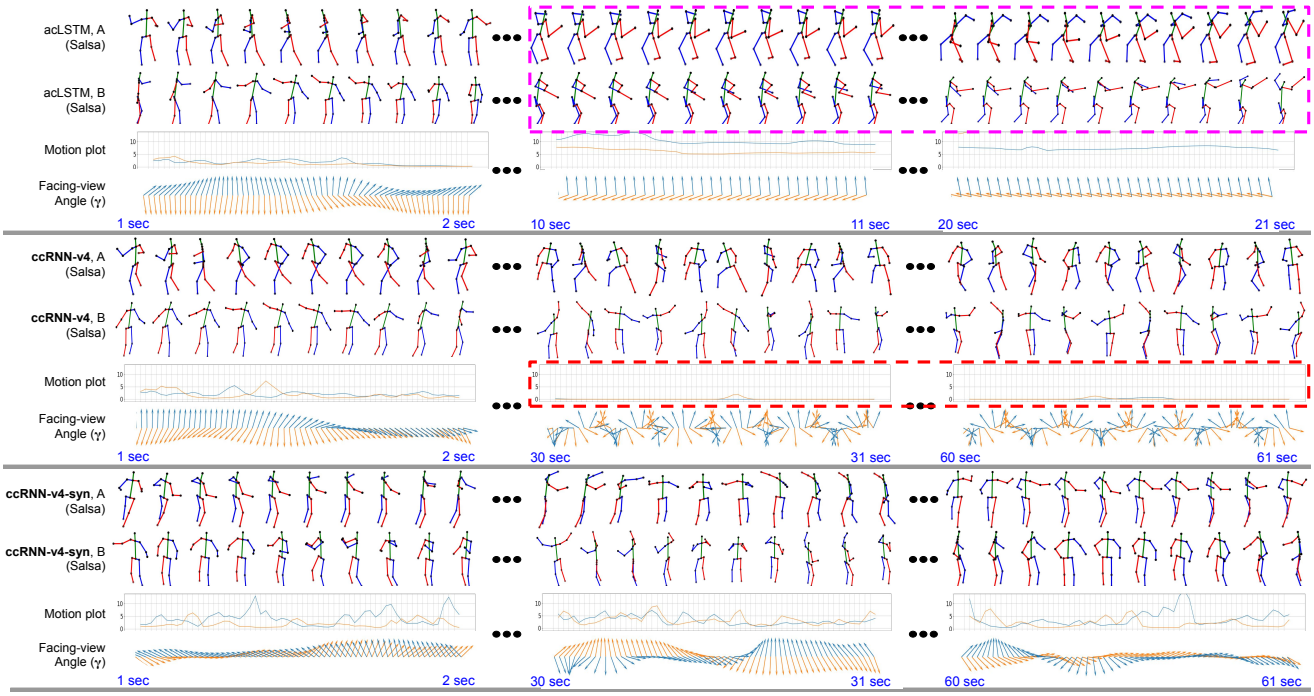

Figure 5. Qualitative comparison. Here, motion plot shows change in view-independent local pose between subsequent time-steps. **acLSTM**. The purple dotted box indicates the estimation of implausible pose pattern beyond just several seconds (i.e. 10 sec). **ccRNN-v4**. Without the hierarchical auto-regressive training our model produces stagnated local-motion beyond 30 seconds as shown in red-box. **ccRNN-v4-syn**. After hierarchical auto-regressive training, our model produces complex inter-person dynamics beyond 60 seconds.

dynamics as the query. A retrieved-clip is considered as a True Positive if $v(q_a) = v(r_b), |\alpha(q_a) - \alpha(r_a)| < \delta$ with δ being the allowed time-shift. For this task, we set $\delta = 10$.

Precision-Recall curves and AUC (Area under the curve) numbers for these two tasks have been reported In Table. 6. Our method outperforms the previous state-of-the-art methods by a significant amount as a result of the encoder-decoder architecture using separate *pRNN* and *fRNN*.

5.3. Qualitative analysis of long-term synthesis

In Figure. 5, it can be observed that our model effectively predicts the difference in view of the complementary dancer, in addition to the disentangled local pose separately for two performers. Additionally, in comparison to acLSTM [25] our model is devoid of motion stagnation and estimation of unrealistic pose pattern. The comparison of *ccRNN-v4* against *ccRNN-v4-syn* shows the utility of the proposed auto-regressive regularization in maintaining inter-person motion dynamics beyond several minutes.

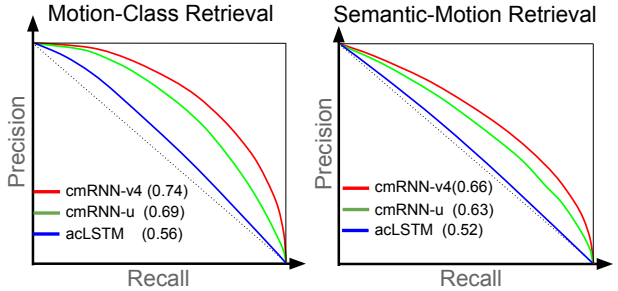

Figure 6. Cross Motion Retrieval on Duet-Dance dataset. Here, AUC values have been indicated in brackets.

6. Conclusion

We target duet-motions with synchronized complementary movements, where two actors perform nonidentical motions. However, in multi-person scenarios, such predictions are either highly uncertain in the long-term (e.g. random crowd movement) or highly identical (e.g. group dance) and hence less interesting to analyze. Additionally, in the absence of a publicly available dataset for modeling long-term multi-person interaction under the presence of meaningful interaction based temporal patterns, we collect a dataset in-house for studying duet-human motion interactions. Moreover, our state-of-the art results on short-term motion prediction, long-term duet-motion synthesis and cross-person retrieval tasks validate effectiveness of the proposed architectural setup. Probabilistic modeling, and RNN based Adversarial losses could result in increased realism, and is something that we plan to explore in future.

Acknowledgements. This work was supported by a Wipro PhD Fellowship (Jogendra), and a project grant from Robert Bosch Centre for Cyber-Physical Systems, IISc.

References

[1] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1446–1455, 2015.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015.

[4] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom. Deep representation learning for human motion prediction and classification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[5] J. Butepage, M. J. Black, D. Kragic, and H. Kjellström. Deep representation learning for human motion prediction and classification. In *30th IEEE/CVF Conference on Computer Vision and Pattern Recognition, JUL 21-26, 2016, Honolulu, HI*, pages 1591–1599. IEEE, 2017.

[6] X. Chen, T. Grossman, D. J. Wigdor, and G. Fitzmaurice. Duet: exploring joint interactions on a smart phone and a smart watch. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 159–168. ACM, 2014.

[7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Gated feedback recurrent neural networks. In *International Conference on Machine Learning*, pages 2067–2075, 2015.

[8] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[9] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[10] P. Ghosh, J. Song, E. Aksan, and O. Hilliges. Learning human motion models for long-term predictions. In *3D Vision, 2017 International Conference on*. IEEE, 2017.

[11] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[12] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing, IEEE International Conference on*, pages 6645–6649. IEEE, 2013.

[13] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

[14] E. S. Ho and T. Komura. Character motion synthesis by topology coordinates. In *Computer Graphics Forum*, volume 28, pages 299–308. Wiley Online Library, 2009.

[15] E. S. Ho, T. Komura, and C.-L. Tai. Spatial relationship preserving character motion adaptation. In *ACM Transactions on Graphics*, volume 29, page 33. ACM, 2010.

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] D. Holden, T. Komura, and J. Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics*, 36(4):42, 2017.

[18] D. Holden, J. Saito, and T. Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics*, 35(4):138, 2016.

[19] D. Holden, J. Saito, T. Komura, and T. Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, page 18. ACM, 2015.

[20] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016.

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] J. N. Kundu, M. Gor, and R. V. Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *AAAI Conference on Artificial Intelligence*, 2019.

[23] J. N. Kundu, M. Gor, P. K. Uppala, and V. B. Radhakrishnan. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In *2019 IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1467. IEEE, 2019.

[24] K. H. Lee, M. G. Choi, and J. Lee. Motion patches: building blocks for virtual environments annotated with motion data. In *ACM Transactions on Graphics*, volume 25, pages 898–906. ACM, 2006.

[25] Z. Li, Y. Zhou, S. Xiao, C. He, Z. Huang, and H. Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *International Conference on Learning Representations*, 2018.

[26] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[27] J. Martinez, M. J. Black, and J. Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4674–4683. IEEE, 2017.

[28] M. M. Pawar, G. N. Pradhan, K. Zhang, and B. Prabhakaran. Content based querying and searching for 3d human motions. In *International Conference on Multimedia Modeling*, pages 446–455. Springer, 2008.

[29] G. N. Pradhan, C. Li, and B. Prabhakaran. Hierarchical indexing structure for 3d human motions. In *International Conference on Multimedia Modeling*, pages 386–396. Springer, 2007.

[30] S. Raghuraman, K. Venkatraman, Z. Wang, B. Prabhakaran, and X. Guo. A 3d tele-immersion streaming approach using skeleton-based prediction. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 721–724. ACM, 2013.

[31] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[32] A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[33] A. Shahroudy, J. Liu, T.-T. Ng, and G. 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.

[34] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1017–1024, 2011.

[35] Y. Tian, Q. Ruan, G. An, and Y. Fu. Action recognition using local consistent group sparse coding with spatio-temporal structure. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 317–321. ACM, 2016.

[36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[37] J. Won, K. Lee, C. O'Sullivan, J. K. Hodgins, and J. Lee. Generating and ranking diverse multi-character interactions. *ACM Transactions on Graphics*, 33(6):219, 2014.

[38] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35. IEEE, 2012.

[39] A. Zunino, J. Cavazza, A. Koul, A. Cavallo, C. Becchio, and V. Murino. Predicting human intentions from motion cues only: a 2d+ 3d fusion approach. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 591–599. ACM, 2017.