Best Practices for 2-Body Pose Forecasting

Muhammad Rameez Ur Rahman[*,1] Luca Scofano[*,2] Edoardo De Matteis[1]
Alessandro Flaborea[1] Alessio Sampieri[2]
Fabio Galasso[1]
Sapienza University of Rome, Italy
[1]{rahman, dematteis, flaborea, galasso}@di.uniroma1.it
[2]{scofano, sampieri}@diag.uniroma1.it

Abstract

The task of collaborative human pose forecasting stands for predicting the future poses of multiple interacting people, given those in previous frames. Predicting two people in interaction, instead of each separately, promises better performance, due to their body-body motion correlations. But the task has remained so far primarily unexplored.

In this paper, we review the progress in human pose forecasting and provide an in-depth assessment of the single-person practices that perform best for 2-body collaborative motion forecasting. Our study confirms the positive impact of frequency input representations, space-time separable and fully-learnable interaction adjacencies for the encoding GCN and FC decoding. Other single-person practices do not transfer to 2-body, so the proposed best ones do not include hierarchical body modeling or attention-based interaction encoding.

We further contribute a novel initialization procedure for the 2-body spatial interaction parameters of the encoder, which benefits performance and stability. Altogether, our proposed 2-body pose forecasting best practices yield a performance improvement of 21.9% over the state-of-the-art on the most recent ExPI dataset, whereby the novel initialization accounts for 3.5%. See our project page at https://www.pinlab.org/bestpractices2body

1. Introduction

Human 2-body pose forecasting predicts the future body poses of two people in interaction jointly. The task is relevant to long-term pose tracking [3], to understanding interacting pairs in sports such as dancing [17] and to the collaborative assembly in industry [12, 26], towards human-robot collaboration [43]. Considering the concurrent prediction of two bodies helps in cases where the people act synergistically. However, this task has remained mostly unexplored and limited to the dataset of [17][1]. Also, this differs from the related task of human trajectory forecasting, where social interaction has been key to most recent progress [27, 41, 42, 51].

There has been vast progress in single human pose forecasting [10, 19, 36], which has not transferred to the 2-body counterpart. Single-person techniques [6, 9, 18] tested on two-people data underperform, which is unsurprising, as they neglect the body-body motion correlations [17]. This motivates the current work, where the most recent modeling advancements are analyzed and integrated. Here, we refer to the best and complementary modeling aspects as *best practices*, which we leverage to bootstrap research on 2-body forecasting.

We propose a systematic analysis of single-person skeleton-based best practices by considering three processing stages (cf. Fig. 1): input representation, encoding, and decoding. For the first stage, we identify Discrete Cosine Transform (DCT) [7, 17, 37–39] as an asset to cope with the periodic body movements. For the second stage, we set to encode the body kinematics by Graph Convolutional Networks (GCN), which power the vast majority of most recent techniques [17, 19, 37–39, 45] and subsume general MLP-based formulations [19]. Here we evaluate as best practices the separability of space and time dynamics [45], the learnable adjacencies versus kinematic trees [50], attention [17], and hierarchical body representations [10]. Finally, for the third stage, we contrast the widely-adopted [36, 43, 45] decoding with convolutional networks (*a.k.a.* Temporal Convolutional Network–TCN [5]) with the simpler Fully Connected (FC) layers [19].

We propose a novel initialization technique for the learnable GCN parameters in the encoder. A large body of literature asserts the importance of initialization for performance, convergence speed, and robustness, and theory has

[*]Equal contribution.

[1]Beyond [17], another multi-body dataset has been introduced by [13], but annotations are only available for one individual at the time of writing.

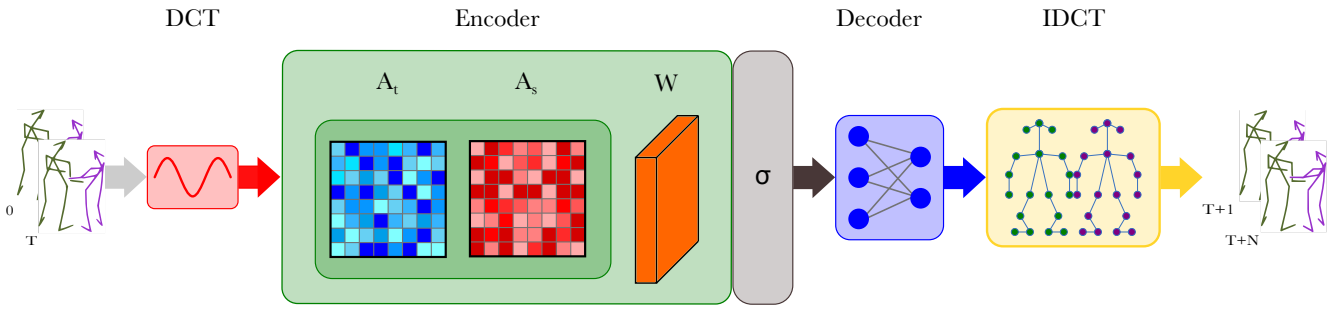

Figure 1. The general architecture of a 2-body pose forecasting model employing best practices. First, 3D joint coordinates are mapped to frequencies by DCT coefficients, a best input representation practice. Secondly, body kinematics are encoded by layers of a GCN $\sigma(A_s A_t X W)$, with separable space-time adjacency matrices $\sigma(A_t, A_s)$, learned unconstrainedly, upon our proposed parameter initialization. Thirdly, the FC-based decoder outputs future poses for the two people, mapped to 3D coordinates with inverse-DCT (IDCT).

been devised for MLP [16] and ConvNets [20, 28]. Up until recently, there has been a limited necessity for ad-hoc GCN initialization theories since techniques leveraged mainly shallow networks with fixed graphs structures (e.g., the people neighbors [4, 30, 46], the kinematic tree [10, 50]) or spectral normalizations [23, 25]. Since we determine that unconstrained learnable GCN affinities are best practices, we also develop a novel theory (See Sec. 3.4) and experimental study (See Sec. 4.3) on the initialization of GCN parameters.

Integrating the selected best practices into a 2-body pose forecasting model yields a large-margin improvement of 21.9% *wrt* the state-of-the-art (SoA) on the most recent ExPI dataset [17]. The best-practice model is also 5 times faster than the current best technique and only has 2% of its parameters (cf. Table 5). The improvement is similarly consistent in generalization tests, across unseen actions with an overall improvement of 14.7% (cf. Table 2) and 14.2% for unseen actors (cf. Table 3). And the same best-practice model performs on par (cf. Table 4) with the leading single-person pose forecasting techniques on the established Human3.6M dataset [22], without any hyper-parameter tuning. The novel initialization, proposed for the unconstrained learning of GCN affinities, contributes an average performance improvement of 3.5%, and it increases stability, as it reduces the long-term forecasting performance variance by (at least) a factor of 2.

The main contributions are summarized as follows:

- We thoroughly evaluate all leading best practices from single-person pose forecasting and bootstrap research on the 2-body task counterpart;
- We propose a novel theory and experimental study on the initialization of GCNs, applying to unconstrained learnable affinities, accounting for an increase in performance of 3.5% and a 2-fold increase in stability;
- On a closed-set dataset configuration, the best-practice model outperforms the 2-body forecasting SoA by a large margin of 21.9% while employing 2% of the parameters and running 5 times faster.

2. Related Work

Here we review related work from the field of human pose forecasting, specifically approaches of spatio-temporal pose modeling and hierarchical body representations. Additionally, we review relevant literature from initialization and multi-agent trajectory forecasting.

Human pose forecasting. Established methodologies for (single) human pose forecasting include Temporal Convolutional Network [30], Recurrent Neural Network [14, 36, 38, 48] and Transformer Networks [2, 17]. The MLP-based approach of [19] holds SoA performance.

Graph Convolutional Networks (GCN) [25, 50] are most popular on the task [10, 32, 43], due to their simplicity and effectiveness. GCNs model the kinematic body part interactions by a plain adjacency matrix at a fraction of the parameters of the otherwise required attention mechanism [17, 37]. In this realm, [37] integrates DCT to consider motion frequency; [10, 33] adopt multi-scale hierarchical representations, grouping joints to model relations between coarser body parts; [43, 45] factorize the spatial and temporal adjacency matrices, and they propose to learn them, unconstrainedly, without kinematic tree priors nor spectral normalization.

As we know, the only work that addresses multi-body pose forecasting is [49]. However, they utilize datasets that do not contain highly interactive actions. For comparison, we ran their model with our setup as a comparison with our proposed method (See Tab. 1). By contrast, for the task of 2-body pose forecasting, [17] provides the solely-available dataset (ExPI) and the only 2-body-specific technique, adaptation of [37] with cross-person attention. Not surprisingly, this outperforms single-person techniques.

Initialization. A proper initialization improves performance and accelerates convergence [29], limiting vanishing and exploding gradients [16, 20]. Techniques have been concerned with initializing the weights of linear [16] and convolutional [20, 28, 40] layers, generalizing from hyperbolic (tanh) to rectified-linear unit (ReLU) activations. For GCNs, spectral techniques [25, 34, 53] rely on the spectral normalization of the adjacency matrix to elude vanishing and exploding gradients, while spatial techniques [4] resort to degree-normalized transition matrices, derived from the adjacency. In all prior study cases, the graph connectivity is given. To the best of our knowledge, this work presents the first theoretical and empirical analysis of GCN initialization in the case of unconstrained learnable graph connectivity and edge weights.

Multi-agent trajectory forecasting. For trajectory forecasting, employed techniques include attention [21, 27, 51] and graph-based modeling [31, 42, 44]. The multi-agent relations may parallel the joint-joint interaction. However, nodes in a graph of joints have a fixed cardinality and a semantic meaning (head, torso, hand, etc.), which does not apply to general agent-agent graphs. Notably, best trajectory forecasting techniques model the agent-agent interaction [21, 27, 31, 42, 44, 51], which aligns with the motivation of this work, to forecast the poses of people jointly.

3. Methodology

We explore the best models for single-body pose forecasting [10, 19, 37, 38, 45] and select best practices for the 2-body task. We group and evaluate practices in three processing stages (cf. Fig. 1): 1) input representation (Sec. 3.1); 2) encoding of the body kinematics in the observed frames (Sec. 3.2); 3) decoding of the future poses (Sec. 3.3). In Sec. 3.4, we provide a theory for the proposed unconstrained-GCN initialization. To facilitate reading, we mark with a green check ✔ the selected *best practices* upon evaluation, cf. Sec. 4.

Problem formalization. Across T frames, we observe the motion of two human bodies $\mathcal{B}^1$ and $\mathcal{B}^2$, each consisting of J three-dimensional joints. At time t, the 3D body pose of each person is given by corresponding tensors $\mathcal{B}^1_t, \mathcal{B}^2_t \in \mathbb{R}^{3\times J}$. We define the concatenation of two bodies at timeframe t as $\mathbf{x}_t = \mathcal{B}^1_t || \mathcal{B}^2_t$, thus the observed motion history in T frames is $\mathcal{X}_{in} = [\mathbf{x}_1, \dots, \mathbf{x}_T] \in \mathbb{R}^{T\times 3\times 2J}$. Our goal is to predict the future N frames' poses $\mathcal{X}_{out} = [\mathbf{x}_{T+1}, \dots, \mathbf{x}_{T+N}] \in \mathbb{R}^{N\times 3\times 2J}$.

Preliminaries on the encoder-decoder baseline. We adopt an encoder-decoder architecture [43, 45], and following [50, 52], we encode the observed body parts and their kinematic interaction through a GCN, defined as

$$\mathbf{Y} = \sigma(A\mathbf{X}W), \tag{1}$$

where A is the adjacency matrix, W learnable weights and σ an activation function. Other encodings such as RNNs [8, 11] and MLPs [19] have been proposed, whereas we opt for a graph-based model to exploit the non-euclidean nature of graphs. As a decoder, we examine either a single fully connected layer as in [19] or a convolutional architecture [36, 45].

3.1. Input Representation

Most recent techniques [1, 19, 37, 38] use Discrete Cosine Transform (DCT) to represent 3D coordinate input as frequencies, under the claim that this captures the dynamic patterns of moving people better.

Frequency encoding ✔

Given the j-th body joint and the t-th timeframe we define the i-th DCT coefficient as

$$\mathcal{F}(\mathcal{X}^{in})_{j,i} = \sqrt{\frac{2}{T}} \sum_{t=1}^{T} x_{j,t} \frac{1}{\sqrt{1+\delta_{i1}}} \cos(\alpha) \tag{2}$$

$$\alpha = \frac{\pi}{2T}(2t-1)(i-1), \tag{3}$$

where the Kronecker delta function $\delta_{ij} \in 0, 1$ has null value if $i \neq j$ and 1 otherwise. After inference, frequencies are remapped to the pose representation via the inverse DCT decoding function $\mathcal{F}^{-1}$. Previous works [37, 39] truncate high frequencies to avoid jittery motion; we consider the impact of the number of retained DCT coefficients and discover that employing all of them yields the best performances. Studies on the impact of DCT coefficients are shown in Sec. 4 and Table 5.

3.2. Encoding Best Practices

Best-performing single-pose forecasting GCN encoders have considered two main aspects: the space-time separability of adjacency weight matrices and learning the body kinematic graph connectivity and weights. We detail these two aspects and empirically compare them in Table 5. Furthermore, we also consider hierarchical representations of the skeleton proposed by [10], but this is not a best practice, as we determine experimentally. Nor is it a good practice to add attention, as we discuss in this section and quantitatively evaluate in the next.

Space-time separability ✔

Each graph's intra-relations are expressed through a GCN-based framework that encodes the spatiotemporal motion

and the relationships between keypoints in one's skeleton [45, 50]. Tensor $\mathbf{X} \in \mathbb{R}^{T\times 2J\times C}$ represents a couple's skeleton pose and motion, adjacency matrices $A_s \in \mathbb{R}^{T\times 2J\times 2J}$ and $A_t \in \mathbb{R}^{2J\times T\times T}$ are responsible for learning spatial and temporal interactions respectively, as in [45]. Matrices are fully learnable, no kinematic tree is used, and the model is free to grasp the relation between body joints. Thus, this module is formulated as follows:

$$\mathbf{Y} = \sigma(A_s A_t \mathbf{X} W), \tag{4}$$

where σ is an activation function and $W \in \mathbb{R}^{C\times C'}$ is a tensor of learnable weights defined as a convolution with kernel dimension $k = 1$. Thus, it is conceptually similar to a fully connected layer. However, unlike the MLP design of [19], GCN shares the weights of W across all channels.

Learning the graph connectivity and weights ✔
Some works [10, 50] use inductive biases based on the human body, such as kinematic trees or specifically-devised connectivity weights. In contrast, others learn the graph adding a constraint on the optimization by spectral normalization [24]. Instead, we follow what is done in the most recent work [45]: unconstrained optimization of graph edges and weights i.e., we set A_{st} for nonseparable GCN and A_s, A_t in case of space-time separable GCN as a fully learnable matrix. This is effectively a best practice, experimentally proven in Table 5.

Attention
A GCN model equipped with attention is also known as a Graph Attention Network (GAT) [47]. In a GAT, attention re-defines the adjacency matrix terms as a function of the node embeddings. We employ attention to encode the relation between the two actor embeddings B_h^1 and B_h^2:

$$\mathcal{B}_{\mathbf{h}}^{\mathbf{1}} = \mathcal{B}^1 W_1, \mathcal{B}_{\mathbf{h}}^{\mathbf{2}} = \mathcal{B}^2 W_2, \tag{5}$$

Where $\mathcal{B}^1, \mathcal{B}^2 \in \mathbb{R}^{T\times J\times C}$ and $W_1, W_2 \in \mathbb{R}^{C\times C}$ are learnable weights to map features in a high-dimensional space. We use these features to calculate attention weights as follows:

$$\eta = softmax\Big(\sigma(\mathcal{B}_h^1 W_3 || (\mathcal{B}_h^2 W_4)^{\mathsf{T}})\Big), \tag{6}$$

Where $\mathcal{B}_h^1, \mathcal{B}_h^2 \in \mathbb{R}^{T\times J\times C}$, $W_3, W_4 \in \mathbb{R}^{C\times 1}$ and σ is a LeakyRelu activation function. We apply softmax to get attention weights $\eta \in \mathbb{R}^{T\times n\times m}$ constituiting n joints in $\mathcal{B}^1$ and m joints in $\mathcal{B}^2$ and reweight B_h^1 and B_h^2 as follows:

$$\mathcal{B}_{\mathbf{out}}^{\mathbf{1}} = \mathcal{B}_h^1 \eta, \mathcal{B}_{\mathbf{out}}^{\mathbf{2}} = \mathcal{B}_h^2 \eta^{\mathsf{T}}, \tag{7}$$

Where $\mathcal{B}_h^1, \mathcal{B}_h^2 \in \mathbb{R}^{T\times J\times C}$ and B_{out}^1, B_{out}^2 are the outputs of attention module. We observe that in its more common use [47], graph attention is used to estimate the interaction coefficients of the adjacency matrix A. This is done by learning a function (general MLP) of two node embeddings. By contrast, when the nodes of the graph are semantically given (body parts of a leader and follower person), one may learn the interaction coefficient (i.e., each term of A) directly, with a joint function of all nodes (not just pairs). The direct estimation results in better performance, as shown by the experiments in Sec. 4.3. Hence, the GCN with fully-learned parameters is selected as a best practice rather than attention.

Hierarchical body parts
To the best of our knowledge, a high-level motion representation improves the prediction of human poses [33]. [10] achieves this by concatenating the higher level as an extra node and hand-crafting ad-hoc neighborhoods of nodes.
We integrate a module within the model that enables it to decrease the number of skeleton keypoints for both bodies. We allow the model to naturally learn aggregations between nodes by excluding artificial aggregations while shifting between hierarchies. We employ a linear layer that learns an optimized aggregation when downscaling, and the same is done when upscaling to retrieve the original size skeleton. Although we gain a small improvement by adopting hierarchies, it becomes a limiting factor rather than a gain when combined with other best practices.

3.3. Decoding Best Practices

In earlier works, convolutions have been employed for the decoding stage [15, 35, 45]. However, the most recent SoA method chose a plain, fully connected layer [19]. In this section, we will analyze the two solutions, and in Sec. 4, we will show why we choose the latter.

Convolutional-based decoder
In the convolutional-based decoder, convolutional layers applied to the temporal dimension are responsible for estimating the pose. It aims to forecast the subsequent frames, $t{+}1$ to $t{+}n$, given the first t frames. This structure is known as Temporal Convolutional Network (TCN) [15, 35, 45].

FC-based decoder ✔
The decoder consists of a single linear layer [19] in charge of mapping the observed T frames to the predicted N.

3.4. Novel Adjacency Matrix Initialization

We propose a novel initialization methodology, aiming to preserve variance during the forward pass, which matches the preservation of gradients in the backward. Since over several layers a non-unit variance results in vanishing or exploding signals, and neither of those is good for

training, as they stall the gradient, we aim to preserve the variance. To do that, under the assumption of a neural network consisting of only linear layers and linear activation functions, [16] proposes to estimate the standard deviation by considering the number of neurons in both the current and previous layer.

It is particularly relevant for our model because it comprises 8 layers while GCNs are often shallow [25]. We propose to randomly initialize the fully learnable matrices A_s, A_t, and W according to a uniform distribution, whose bounds are defined in such a way that considers both the number of graph nodes and the number of timeframes.

Convolutions on graphs that adopt a normalized adjacency matrix [25, 46] use a well-known graph and do not let all nodes interact with each other. Furthermore, normalization avoids vanishing and exploding gradient, yet it limits the performance and, in the end, fully-learnable yields the best performances [43, 45]. Here is the importance of randomly initializing an *ad hoc* fully learnable adjacency matrix, avoiding exploding or vanishing gradients. The response from the Separable GCN at layer l, according to Eq. (4), is

$$\mathbf{X}^{l+1} = \sigma\left(A_s^l A_t^l \mathbf{X}^l W^l\right), \quad \forall l. \tag{8}$$

Let's assume matrices A_s, A_t, and W to be independent, have zero mean [16, 20] and uniformly distributed. To constrain variance, hence stabilize training and avoid exploding or vanishing gradient, constraining the variance of the output product of n^l neurons at layer l times W to 1 [20] is a sufficient condition, i.e.,

$$\frac{1}{k} n^l Var[W^l] = 1, \quad \forall l, \tag{9}$$

where $k = 2$ in the case of Re-LU activations, which are asymmetric [20] (while $k = 1$ for symmetric activations such as the `tanh`). For the spatial matrix, rather than the number of neurons n^l, we consider the number of nodes v, which A_s integrates

$$\frac{1}{k}(n_v^l) Var[A_s^l] = 1, \quad \forall l. \tag{10}$$

Similarly, we consider t time frames to initialize the temporal matrix A_t,

$$\frac{1}{k}(n_t^l) Var[A_t^l] = 1, \quad \forall l. \tag{11}$$

When initializing W with a zero-mean uniform distribution, the constraint of Eq. (9) yields the following distribution for the initialization:

$$W^l \sim U\left[-\sqrt{\frac{k}{n^l}}, \sqrt{\frac{k}{n^l}}\right], \quad \forall l. \tag{12}$$

The spatial and temporal matrix constraints of Eqs. (10) and (11) translate to the following initializing distributions for A_s and A_t respectively:

$$A_s^l \sim U\left[-\sqrt{\frac{k}{n_v^l}}, \sqrt{\frac{k}{n_v^l}}\right], \tag{13}$$

$$A_t^l \sim U\left[-\sqrt{\frac{k}{n_t^l}}, \sqrt{\frac{k}{n_t^l}}\right], \quad \forall l. \tag{14}$$

4. Experiments

We thoroughly evaluate the proposed best practices on the most recent and challenging 2-body pose forecasting dataset ExPI [17], comparing against the SoA and the best single-pose forecasting techniques adapted to the task. The selected best practices also perform on par with the SoA in single-pose forecasting on the established Human3.6M dataset [22].

4.1. Benchmark and baselines

Datasets. The dataset used for multi-body pose forecasting, ExPI [17], is a collection of two different dancing pairs performing Lindy Hop sessions, dubbed "extreme human interaction" by the authors [17]. Data were collected in a multi-camera platform with 68 synchronized and calibrated RGB cameras and a motion capture system with 20 mocap cameras. The missing points were manually fixed to ensure good data quality. ExPI contains 115 sequences at 25 fps with 18 body joints for each of the two persons involved. These agents are grouped in two couples, dubbed $(\mathcal{A}_c^1, \mathcal{A}_c^2)$, which perform 16 different actions. Actions A1 to A7 are common to both couples; A8 to A13 performed only by $\mathcal{A}_c^1$ and A14-A16 by $\mathcal{A}_c^2$. Based on this, ExPI provides three different splits to test the model on:

- **Common.** Training and test set are composed only of actions performed by both couples. The ones belonging to $\mathcal{A}_c^2$ define the train set, and $\mathcal{A}_c^1$'s the test set.
- **Unseen.** Differently from the previous one, this split has common actions to both $\mathcal{A}_c^1$ and $\mathcal{A}_c^2$ as the train set and couple-specific actions as the test one. This subset allows us to test for generalization.
- **Single.** In this split, a single action from couple $\mathcal{A}_c^2$ is used as a train set, and the same action from couple $\mathcal{A}_c^1$ as the test set. It allows testing how the model generalizes to a new couple for each action.

We also test on Human3.6M [22], an established dataset for single-person pose forecasting. It consists of a total of 3.6 million poses, acquired at 25 fps, depicting seven actors performing 15-day real-life actions, e.g., walking, sitting,

Action	A1				A2				A3				A4				A5				A6				A7				Average ↓			
Time (msec)	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*
LTD [38]	70	125	157	189	131	242	321	426	102	194	260	357	62	117	155	197	72	131	173	231	81	151	200	280	112	223	315	442	90	169	226	303
HisRep [37]	52	103	139	188	96	186	256	349	57	118	167	240	45	93	131	180	51	105	149	214	61	125	176	252	71	150	222	333	62	126	177	251
MSR-GCN [10]	56	100	132	175	102	187	256	365	65	120	166	244	50	95	127	172	54	100	138	202	70	132	182	258	82	154	218	321	69	127	174	248
MRT [49]	50	98	134	188	79	155	212	307	53	106	152	229	47	95	131	185	52	105	149	215	58	118	166	242	65	136	199	299	58	116	163	238
siMLPe [19]	49	102	137	177	88	180	244	336	57	122	174	254	45	100	137	182	50	103	144	206	59	126	175	250	77	164	134	348	60	128	178	250
XIA [17]	49	98	140	192	84	166	234	346	51	105	154	234	41	84	120	161	43	90	132	197	55	113	163	242	62	130	192	291	55	112	162	238
Ours	**34**	**71**	**105**	**159**	**56**	**121**	**181**	**292**	**36**	**78**	**118**	**195**	**30**	**66**	**98**	**145**	**35**	**74**	**113**	**171**	**41**	**88**	**129**	**193**	**47**	**108**	**166**	**261**	**39**	**86**	**129**	**202**

Table 1. Results in millimeters for ExPI Common actions split. Our model achieves state-of-the-art results in all actions considered, at each predicted time instant.

Action	A8			A9			A10			A11			A12			A13			A14			A15			A16			Average ↓		
Time (msec)	*400*	*600*	*800*	*400*	*600*	*800*	*400*	*600*	*800*	*400*	*600*	*800*	*400*	*600*	*800*	*400*	*600*	*800*	*400*	*600*	*800*	*400*	*600*	*800*	*400*	*600*	*800*	*400*	*600*	*800*
LTD [38]	252	333	387	174	228	268	139	184	217	239	324	394	175	226	259	148	191	220	176	240	286	143	178	192	146	193	226	177	233	272
HisRep [37]	157	219	257	134	190	233	96	146	187	195	283	358	121	169	206	92	129	**160**	129	193	245	80	104	121	112	154	187	124	176	218
MSR-GCN [10]	177	239	295	143	179	213	157	222	281	230	289	335	188	245	290	148	198	248	234	319	384	176	232	278	162	218	266	179	238	288
MRT [49]	170	231	308	145	199	270	141	245	338	225	327	481	131	180	253	120	169	238	165	229	322	110	151	209	105	144	201	146	205	291
siMLPe [19]	165	220	258	137	198	246	104	154	198	210	301	432	114	156	187	94	132	160	140	204	255	91	119	138	120	166	204	131	183	225
XIA [17]	156	216	256	126	175	213	96	152	205	191	287	377	118	165	203	91	129	162	122	183	232	81	**107**	**128**	106	150	185	121	174	218
Ours	**113**	**164**	**203**	**114**	**167**	**209**	**85**	**136**	**183**	**153**	**231**	**304**	**100**	**148**	**188**	**82**	**125**	162	**91**	**138**	**179**	**79**	109	132	**85**	**124**	**156**	**100**	**149**	**191**

Table 2. Results in millimeters for ExPI Unseen actions split. On average, we outperform the baseline considered over short and long time horizons.

and talking on the phone. Following [10, 37, 39], we train on subjects S1, S6, S7, S8, S9, we use S11 for validation, and S5 for testing.

Evaluation metrics. We validate performance by the *Mean per joint position Error*, defined as the MPJPE [22, 38] and renamed as JME in [17] at a future frame t:

$$L_{\text{JME}} = L_{\text{MPJPE}} = \frac{1}{V}\sum_{v=1}^{V} ||\hat{x}_{vt} - x_{vt}||_2, \qquad (15)$$

where $\hat{x}_{vt}$ and x_{vt} are the 3-dimensional vectors of a target joint and the ground truth, respectively. For the joint evaluation of the 2-body position error, the two body poses are normalized into the same reference system. In this work, we keep the MPJPE notation.

Baselines. We select the latest and best-performing single-body pose forecasting models, and we adapt them to predict the motion of two people. XIA-Transformer [17] is the only 2-body pose forecasting method in the literature. XIA uses a transformer to encode skeleton features and model the body-body interaction via attention. We consider [49] the only multi-body model based on a Transformer architecture. Due to the lack of multi-body pose forecasting models, we also compare them to single ones. LTD [38] consists of a cascade of GCN blocks acting on frequencies, and its extension, HisRep [37], inserts a motion attention mechanism based on DCT coefficients operating on sub-sequences of the input. MSR-GCN [10] is a hierarchical GCN-based technique that applies multi-scale aggregations, so coarser scales represent groups of body joints and coarser motion. In Table 4 we compare ourselves, again, to LTD [38], HisRep [37] and MSR-GCN [10] and, additionally, on two recent single-body models. SeS-GCN [43] adopts an all-separable GCN with a teacher-student approach, and the SoA [19], which consists of MLPs encoding spatial and temporal relationships.

4.2. Evaluation of human pose forecasting

We evaluate our model quantitatively and qualitatively on ExPI's [17] provided splits. We further test our model's generalization power on the single-body dataset Human3.6M.

ExPI Common Actions. Table 1 shows the results obtained from our best model with our selected best practices. These outperform every tested method by a large margin, both the SoA single-person and the SoA 2-body pose forecasting techniques. The overall mean improvement is 22% over all actions and all time horizons. In particular, on all actions, the improvement for short-term future predictions (200 msec) is 29% and 15% for the long-term.

ExPI Unseen Actions. Table 2 also showcases improvements using the proposed best practices. On average, across all forecasting horizons, the improvement is 14%.

ExPI Single Actions. In Table 3, also for the case of single actions, the best practices report an average improvement of 14.2%. They outperform all other tested techniques in 6 (out of 7) actions at all predicted time horizons. It confirms the generalization of our model to new people.

Action	A1				A2				A3				A4				A5				A6				A7			
Time (msec)	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*	*200*	*400*	*600*	*1000*
LTD [38]	70	126	155	183	131	243	312	415	102	194	252	338	62	117	153	203	71	131	171	231	81	151	199	299	112	223	306	411
HisRep [37]	66	118	153	190	128	231	308	417	74	143	205	295	64	120	159	191	63	121	166	227	90	168	232	312	88	166	232	332
MSR-GCN [10]	64	108	136	**170**	119	210	282	385	79	144	189	265	59	103	134	173	65	118	162	225	86	151	201	283	96	178	255	362
MRT [49]	63	120	160	218	97	190	249	346	77	148	193	240	51	102	139	186	61	118	163	226	58	115	151	198	82	172	244	340
siMLPe [19]	60	113	145	200	104	202	268	373	76	150	205	305	58	110	151	203	64	123	163	218	76	152	207	277	93	180	254	341
XIA [17]	64	120	160	199	109	200	275	381	59	117	174	277	60	116	162	209	53	106	152	221	65	122	166	223	74	144	**203**	**301**
Ours	**52**	**94**	**128**	179	**89**	**176**	**242**	**329**	**42**	**90**	**129**	**200**	**49**	**96**	**134**	**185**	**48**	**99**	**140**	**196**	**52**	**105**	**144**	**198**	**68**	**140**	204	305

Table 3. Results in millimeters for ExPI Single actions split. We outperform in 6 out of 7 stocks all baselines considered according to the MPJPE metric. For the other stocks our model is comparable with the current state of the art.

ExPI qualitative. In Fig. 2, the current SoA, ExPI [17], is compared against the best-practice model (*Ours*), qualitatively. The first three columns depict observations; the following four are future motion predictions. The light-colored pictograms represent ground-truth motion. The best practices provide, in general, better predictions. Best improvements are observed in the case of large motion displacements, cf. the last two rows, action "Cartwheel".

	MPJPE ↓			
Time Horizon (msec)	*160*	*400*	*560*	*1000*
LTD [38]	23.4	58.9	78.3	114.0
HisRep [37]	22.6	58.3	77.3	112.1
MSR-GCN [10]	25.5	63.3	81.1	114.1
SeS-GCN [43]	29.0	64.0	84.4	113.9
siMLPe [19]	**21.7**	**57.3**	**75.7**	**109.4**
Ours w/o init.	27.3	64.6	83.1	116.3
Ours	26.8	63.1	81.1	113.2

Table 4. Error in millimeters on Human3.6M dataset. We show how our method adapted to single-person human pose forecasting is comparable with the best-performing techniques on average.

Evaluation of single-person pose forecasting. We test how the 2-body best practices transfer back to single-person pose forecasting for a sanity check. In Table 4, observe that the best practices (*Ours*) yield results within a small margin compared to SoA. Note that, for the sake of this experiment, we just run the 2-body best-practice model *as is*. Without any hyper-parameter tuning. Furthermore, the initialization gives an overall 2.4% over the counterpart model that does not use it.

4.3. Evaluation of Best Practices

In this section, we refer to Table 5 and thoroughly assess each selected practice. First, we select a baseline GCN model. Secondly, we assess each practice's performance, added as a standalone extension. Thirdly, we integrate practices. Best practices are assessed based on their standalone performance improvement and complementarity. Finally, in Table 6, we evaluate the impact of the proposed initialization in more detail.

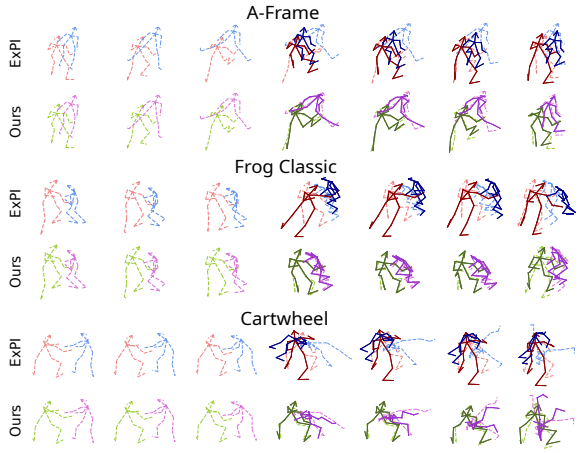

Figure 2. Visual comparison of our proposed best-practice model (*Ours*) against ExPI [17]. The first three columns are observed, and the last four are predicted poses. Light-colored and dashed skeletons are GT, and darker and solid ones are predictions. Note the improved larger-displacement motions (*Cartwheel*).

Baseline selection. We first select a baseline model on which we test each best practice. We identify three possible GCN-based encoder architectures:

- Space-time GCN [50]: this is a plain GCN model $\sigma(AXW)$ with learnable A (learnable connectivity and graph weights)
- Space-time separable GCN with learnable kinematic tree: inspired by [50] and [45] to factorize the adjacency matrix into two spatial and temporal learnable matrices, whereby the spatial connectivity is constrained to the kinematic tree
- Space-time separable GCN with fully-learnable connections: lastly, we evaluate a space-time separable GCN with fully-learnable adjacencies matrices taking inspiration from [45].

As shown in Table 5, the space-time GCN with separability (row 3) has an overall decrease of error by 25% compared to the base GCN (row 2). A considerable additional

	Model	Input Repr.	Encoding					Decoding	*MPJPE* ↓				Param. ↓
		Freq. Enc. ✓	Learn. ✓	Sep. ✓	Init. ✓	Att.	Hier.	FC ✓	*200*	*400*	*600*	*1000*	(M)
1	[17]	✓	✓			✓			55	112	162	238	8.5
2	Space-time GCN		✓						108	152	255	379	1.08
3	*(kin. tree)*			✓					81	129	183	260	0.18
4			✓	✓					55	112	156	224	0.18
5	Input repr. practice	✓	✓	✓					41	88	135	219	0.18
6			✓	✓	✓				53	106	148	216	0.18
7	Encoder practices		✓	✓		✓†			55	112	157	228	9.9
8			✓	✓			✓		51	104	148	223	0.18
9	Decoder practices		✓	✓				✓	51	104	145	212	**0.17**
10		✓	✓	✓				✓	41	89	133	208	**0.17**
11		✓	✓	✓			✓	✓	51	104	146	217	**0.17**
12	**Best model**	✓	✓	✓	✓			✓	**39**	**86**	**129**	**202**	**0.17**

Table 5. Combinations of best practices. From left to right, we have frequency encoding, fully learnable connections, Space-time separability, initialization, attention mechanism, hierarchy, fully connected layer as a decoder. †: we implement a Graph Attention Network (GAT) tailored for GCNs, similar in spirit to [17] designed for transformers.

performance boost (18% over all frames) s also given by using the separability and fully-learnable connections (row 4) instead of limiting the learning procedure on the kinematic tree. The simple space-time separable GCN already outperforms XIA [17] while having a fraction of the parameters, although XIA includes DCT representations and attention. Thus GCN with separability and fully-learnable connections is a good baseline to build upon.

Standalone best practices. Table 5 shows input representation (row 5), encoding (rows 6-8), and decoding practices (row 9). When considering the input representation and decoding techniques, DCT, and fully connected (FC) layer as decoder, it is clear that both have a considerable impact. The DCT provides a significant boost in short-term predictions, up to 25%, while the FC-based decoder offers a more substantial increase in long-term predictions, up to 7% against TCN (when the box is not ✓). Regarding the encoder practices, the novel initialization procedure and a hierarchical architecture improve the chosen baseline by 5% and 4%, respectively. On the other hand, using the attention technique did not lead to any gain in performance and is hence not considered a best practice.

Integrated best practices. Rows 10-12 in Table 5 refers to the combinations of techniques that performed best independently.

Integrating the input representation using DCT coefficients and the FC-based decoder indicates how these two methods can be used in addition to the standard method. Secondly, we include a Graph Attention Network as explained in Sec. 3.2 to account for the interaction. The performance does not benefit from it, and the number of parameters is considerably higher. Lastly, a hierarchical structure lowers performance when combined with other practices, so we do not consider it a best practice. Our proposed initialization improves our best practice model by another 3.5%.

Impact of initialization. Table 6 shows the average of multiple runs for different initialization methods and the corresponding standard deviation. We compare our strategy with the Uniform sampling and the two established methodologies of [16], and [20]. Our proposed initialization exceeds or is on par with the others on average, having more than 2.6% improvement over uniform sampling over the longer time horizon. Note also the lower standard deviation of performance for our proposed technique, especially for the most challenging long-term prediction horizon (at least 2x lower), which we interpret as improved stability.

	MPJPE ↓			
Time Horizon (msec)	*200*	*400*	*600*	*1000*
Uniform	39.7 ±0.7	87.6 ±0.7	132.2 ±0.5	207.7 ±1.1
Glorot et al. [16]	40.3 ±0.1	89.4 ±1.2	134.3 ±1.5	207.9 ±1.8
He et al. [20]	40.2 ±0.4	88.6 ±0.7	133.4 ±1.4	206.6 ±1.2
Ours	**39.2 ±0.4**	**86.4 ±0.6**	**129.4 ±1.0**	**202.2 ±0.5**

Table 6. Initialization procedures for best practices model.

5. Conclusion

This work has identified, reviewed, and experimentally evaluated best practices for 2-body pose forecasting, to bootstrap research in the mostly unexplored task. Best practices have a large impact on SoA performance, and the novel initialization adds further improvement in performance and stability. Notably, predicting the future of two people in interaction yields better estimates than considering each person separately, so 2-body forecasting is recommended for applications such as sports and collaborative assembly in factories.

References

[1] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, 2021.

[3] Mykhaylo Andriluka, Umar Iqbal, Anton Milan, Eldar Insafutdinov, Leonid Pishchulin, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018.

[4] James Atwood and Don Towsley. Diffusion-convolutional neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2016.

[5] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[6] Abdallah Benzine, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Deep, robust and single shot 3d multiperson human pose estimation from monocular images. In *2019 IEEE International Conference on Image Processing (ICIP)*, 2019.

[7] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, Ding Liu, Jing Liu, and Nadia Magnenat Thalmann. Learning progressive joint propagation for human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[8] Hsu-Kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[9] Rishabh Dabral, Nitesh Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Dynamic multiscale graph neural networks for 3d skeleton-based human motion prediction. In *Multi-person 3d human pose estimation from monocular images. In 2019 International Conference on 3D Vision (3DV)*, 2019.

[10] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[11] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] N. F. Duarte, M. Raković, J. Tasevski, M. I. Coco, A. Billard, and J. Santos-Victor. Action anticipation: Reading the intentions of humans and robots. *IEEE Robotics and Automation Letters*, 3(4):4132–4139, 2018.

[13] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[14] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[15] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR, 2017.

[16] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9. PMLR, 2010.

[17] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[18] Wen Guo, Enric Corona, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. Pi-net: Pose interacting network for multi-person monocular 3d pose estimation. In *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

[19] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Alameda-Pineda Xavier, and Moreno-Noguer Francesc. Back to mlp: A simple baseline for human motion prediction. *arXiv preprint arXiv:2207.01567*, 2022.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[21] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[22] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[23] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[24] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pages 2688–2697. PMLR, 2018.

[25] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[26] Hema Swetha Koppula and Ashutosh Saxena. Anticipating human activities for reactive robotic response. In *2013*

IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 2071–2071, 2013.

[27] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):7386–7400, 2022.

[28] Philipp Krähenbühl, Carl Doersch, Jeff Donahue, and Trevor Darrell. Data-dependent initializations of convolutional neural networks. *CoRR*, abs/1511.06856, 2016.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[30] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[31] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *arXiv: Computer Vision and Pattern Recognition*, 2020.

[32] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:3316–3333, 2022.

[33] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[34] Qimai Li, Zhichao Han, and Xiao-ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[35] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018.

[36] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[37] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[38] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[39] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Multi-level motion attention for human motion prediction. *International Journal of Computer Vision*, 129(9):2513–2535, 2021.

[40] Dmytro Mishkin and Jiri Matas. All you need is a good init. In *4th International Conference on Learning Representations, ICLR*, 2016.

[41] Abduallah A. Mohamed, Deyao Zhu, Warren Vu, Mohamed Elhoseiny, and Christian G. Claudel. Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[42] Alessio Monti, Alessia Bertugli, Simone Calderara, and Rita Cucchiara. Dag-net: Double attentive graph neural network for trajectory forecasting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2551–2558, 2021.

[43] Alessio Sampieri, Guido Maria D'Amely di Melendugno, Andrea Avogaro, Federico Cunico, Francesco Setti, Geri Skenderi, Marco Cristani, and Fabio Galasso. Pose forecasting in industrial human-robot collaboration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[44] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. Sgcn:sparse graph convolution network for pedestrian trajectory prediction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[45] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[46] Zekun Tong, Yuxuan Liang, Changsheng Sun, David S. Rosenblum, and Andrew Lim. Directed graph convolutional network, 2020.

[47] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018.

[48] Borui Wang, Ehsan Adeli, Hsu-Kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[49] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers, 2021.

[50] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[51] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[52] Songyang Zhang, Xiaoming Liu, and Jun Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 148–157, 2017.

[53] Xitong Zhang, Yixuan He, Nathan Brugnone, Michael Perlmutter, and Matthew Hirn. Magnet: A neural network for

directed graphs. In *Advances in Neural Information Processing Systems*, 2021.