EqMotion: Equivariant Multi-agent Motion Prediction with Invariant Interaction Reasoning

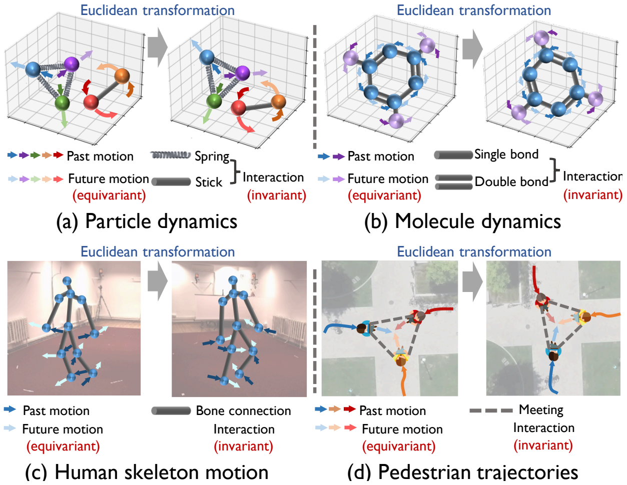
Chenxin Xu[1,2], Robby T. Tan[2], Yuhong Tan[1], Siheng Chen[1,3*],
Yu Guang Wang[1], Xinchao Wang[2], Yanfeng Wang[3,1]
[1]Shanghai Jiao Tong University, [2]National University of Singapore, [3]Shanghai AI Laboratory
{xcxwakaka,tyheeeer,sihengc,yuguang.wang,wangyanfeng}@sjtu.edu.cn, {robby.tan,xinchao}@nus.edu.sg

Abstract

Learning to predict agent motions with relationship reasoning is important for many applications. In motion prediction tasks, maintaining motion equivariance under Euclidean geometric transformations and invariance of agent interaction is a critical and fundamental principle. However, such equivariance and invariance properties are overlooked by most existing methods. To fill this gap, we propose EqMotion, an efficient equivariant motion prediction model with invariant interaction reasoning. To achieve motion equivariance, we propose an equivariant geometric feature learning module to learn a Euclidean transformable feature through dedicated designs of equivariant operations. To reason agent's interactions, we propose an invariant interaction reasoning module to achieve a more stable interaction modeling. To further promote more comprehensive motion features, we propose an invariant pattern feature learning module to learn an invariant pattern feature, which cooperates with the equivariant geometric feature to enhance network expressiveness. We conduct experiments for the proposed model on four distinct scenarios: particle dynamics, molecule dynamics, human skeleton motion prediction and pedestrian trajectory prediction. Experimental results show that our method is not only generally applicable, but also achieves state-of-the-art prediction performances on all the four tasks, improving by 24.0/30.1/8.6/9.2%. *Code is available at* https://github.com/MediaBrain-SJTU/EqMotion.

1. Introduction

Motion prediction aims to predict future trajectories of multiple interacting agents given their historical observations. It is widely studied in many applications like physics [3, 28], molecule dynamics [7], autonomous driving [35] and human-robot interaction [38,68]. In the task of motion prediction, an often-overlooked yet fundamental principle is that a prediction model is required to be equivariant under the Euclidean geometric transformation (including translation, rotation and reflection), and at the same time maintain the interaction relationships invariant. Motion equivariance here means that if an input motion is transformed under a Euclidean transformation, the output motion must be equally transformed under the same transformation. Interaction invariance means that the way agents interact remains unchanged under the input's transformation. Figure 1 shows real-world examples of motion equivariance and interaction invariance.



(a) Particle dynamics (b) Molecule dynamics
(c) Human skeleton motion (d) Pedestrian trajectories

Figure 1. Motion equivariance and interaction invariance under the Euclidean geometric transformation is a fundamental principle for a prediction model, but this principle is often overlooked by previous works. In this work, we propose EqMotion to fill this gap.

Employing this principle in a network design brings at least two benefits. First, the network will be robust to arbitrary Euclidean transformations. Second, the network will have the capability of being generalizable over rotations and translations of the data. This capability makes the network more compact, reducing the network's learning burden and contributing to a more accurate prediction.

Despite the motion equivariance property being important

[*] Corresponding author.

and fundamental, it is often neglected and not guaranteed by most existing motion prediction methods. The main reason is that these methods transform the input motion sequence directly into abstract feature vectors, where the geometric transformations are not traceable, causing the geometric relationships between agents to be irretrievable. Random augmentation will ease the equivariance problem, but it is still unable to guarantee the equivariance property. [30] uses non-parametric pre and post coordinate processing to achieve equivariance, but its parametric network structures do not satisfy equivariance. Some methods propose equivariant parametric network structures utilizing the higher-order representations of spherical harmonics [14, 61] or proposing an equivariant message passing [58], but they focus on the state-to-state prediction. This means that they use only one historical timestamp to predict one future timestamp. Consequently, these methods have limitations on utilizing motion's temporal information and modeling interaction relationships since a single-state observation is insufficient for both interaction modeling and temporal dependency modeling.

In this paper, we propose EqMotion, the first motion prediction model that is theoretically equivariant to the input motion under Euclidean geometric transformations based on the parametric network. The proposed EqMotion has three novel designs: equivariant geometric feature learning, invariant pattern feature learning and invariant interaction reasoning. To ensure motion equivariance, we propose an equivariant geometric feature learning module to learn a Euclidean transformable geometric feature through dedicated designs of equivariant operations. The geometric feature preserves motion attributes that are relevant to Euclidean transformations. To promote more comprehensive representation power, we introduce an invariant pattern feature learning module to complement the network with motion attributes that are independent of Euclidean transformations. The pattern features, cooperated into the geometric features, provide expressive motion representations by exploiting motions' spatial-temporal dependencies.

To further infer the interactions during motion prediction, we propose an invariant interaction reasoning module, which ensures that the captured interaction relationships are invariant to the input motion under Euclidean transformations. The module infers an invariant interaction graph by utilizing invariant factors in motions. The edge weights in the interaction graph categorize agents' interactions into different types, leading to better interaction representation.

We conduct extensive experiments on four different scenarios to evaluate our method's effectiveness: particle dynamics, molecule dynamics, 3D human skeleton motion and pedestrian trajectories. Comparing to many task-specific motion prediction methods, our method is generally applicable and achieves state-of-the-art performance in all these tasks by reducing the prediction error by 24.0/30.1/8.6/9.2% respectively. We also present that EqMotion is lightweight, and has a model size less than 30% of many other models' sizes. We show that EqMotion using only 5 % data can achieve a comparable performance with other methods that take full data. As a summary, here are our contributions:

• We propose EqMotion, the first motion prediction model that theoretically ensures sequence-to-sequence motion equivariance based on the parametric network. With equivariance, EqMotion promotes more generalization ability of motion feature learning, leading to more robust and accurate prediction.

• We propose a novel invariant interaction reasoning module, in which the captured interactions between agents are invariant to the input motion under Euclidean geometric transformations. With this, EqMotion achieves more generalization ability and stability in the interaction reasoning.

• We conduct experiments on four types of scenarios and find that EqMotion is applicable to all these different tasks, and importantly outperforms existing state-of-the-art methods on all the tasks.

2. Related Work

Equivariant Networks. Equivariance first draws high attention on the 2D image domain. Since CNN structure is sensitive to rotations, researchers start to explore rotation-equivariant designs like oriented convolutional filters [8, 49], log-polar transform [11], circular harmonics [67] or steerable filters [66]. Meanwhile, GNN architectures [73–75] exploring symmetries on both rotation and translation have been emerged. Specifically, [57, 62] achieves partial symmetries by promoting translation equivariance. [14, 61] builds filters using spherical harmonics allowing transformations between high-order representations, achieving the rotation and translation equivariance. [12, 24] construct a Lie convolution to parameterize transformations into Lie algebra form. [10] proposes a series of equivariant layers for point cloud networks. [27] propose geometric vector perceptions for protein structure learning. Recently, EGNN [58] proposes a simple equivariant message passing form without using computationally expensive high-order representations. [23] further extends it by considering geometrical constraints. However, most existing methods are only applicable to state prediction, limiting models from exploiting sequence information. [30] uses pre and post coordinate processing to achieve motion equivariance but its network structure does not satisfy equivariance. In this work, we propose an equivariant model based on the parametric network which is generally applicable to motion prediction tasks and achieves a more precise prediction.

Motion Prediction. Motion prediction has wide application scenarios. [3, 53, 56] proposes graph neural networks for learning to simulate complex physical systems. [17, 28, 37, 71] both explicitly infer the interactions relationships and perform prediction in physical systems. For human or vehicle trajectory prediction, social forces [21, 51], Markov [29, 65] process, and RNNs [1, 52, 63] are first methods to employ. Multi-model prediction methods are further proposed like using generator-discriminator structures [20, 22], multi-head output [43, 59], conditional variational autoencoders [32, 45, 55, 69, 71, 72, 76], memory mechanisms [48, 70], Gaussian mixture distribution prediction [17, 37]. The HD map information is specifically considered in the autonomous driving scenario [5, 6, 15, 22, 42, 43, 77]. For 3D human skeleton motion prediction, early methods are based on the state prediction [33, 60]. Later RNN-based models considering the sequential motion states are proposed [13, 26, 50, 64]. [19, 36] use spatial graph convolutions to directly regress the whole sequences. Besides, some methods [4, 40, 46, 47] specifically exploit the correlations between body joints. Multi-scale graphs are built by [9, 39, 41] to capture different body-level dependencies. In this work, we propose a generally applicable motion prediction network which promotes Euclidean equivariance, a fundamental property but neglected by previous methods, to have a more robust and accurate prediction.

3. Background and Problem Formulation

3.1. Motion Prediction

Here we introduce the general problem formulation of motion prediction, which aims to generate future motions given the historical observations. Mathematically, consider M agents in a multi-agent system. Let $\mathbf{X}_i = [\mathbf{x}_i^1, \mathbf{x}_i^2, \cdots \mathbf{x}_i^{T_\mathrm{p}}] \in \mathbb{R}^{T_\mathrm{p} \times n}$ and $\mathbf{Y}_i = [\mathbf{y}_i^1, \mathbf{y}_i^2, \cdots \mathbf{y}_i^{T_\mathrm{f}}] \in \mathbb{R}^{T_\mathrm{f} \times n}$ be the ith agent's past and future motion, where T_p and T_f are the past and future timestamps and n is the dimension of the system space. n usually equals to 2 or 3 (corresponding to 2D or 3D case). The tth timestamp locations $\mathbf{x}_i^t$ and $\mathbf{y}_i^t$ are n-dimension vectors. The whole system's past and future motion is represented as $\mathbb{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M] \in \mathbb{R}^{M \times T_\mathrm{p} \times n}$ and $\mathbb{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_M] \in \mathbb{R}^{M \times T_\mathrm{f} \times n}$. We aim to propose a prediction network $\mathcal{F}_{\mathrm{pred}}(\cdot)$ so that the predicted future motions $\widehat{\mathbb{Y}} = \mathcal{F}_{\mathrm{pred}}(\mathbb{X})$ are as close to the ground-truth future motions $\mathbb{Y}$ as possible.

3.2. Equivariance and Invariance

The Euclidean geometric transformation has three basic forms: translation, rotation and reflection. The translation is modeled by a translation vector and rotation (or reflection) is modeled by an orthogonal rotation (or reflection) matrix. Let $\mathbf{t} \in \mathbb{R}^n$ be a translation vector and $\mathbf{R} \in \mathrm{SO}(n)$ be a $n \times n$ rotation matrix, we have the following definitions for equivariance and invariance:

Definition 1 *Let $\mathbb{X}$ be an input, $\mathcal{F}(\cdot)$ be an operation and $\mathbb{Z} = \mathcal{F}(\mathbb{X})$ be the corresponding output. The operation $\mathcal{F}(\cdot)$ is called equivariant under Euclidean transformation if*

$$\mathbb{Z}\mathbf{R} + \mathbf{t} = \mathcal{F}(\mathbb{X}\mathbf{R} + \mathbf{t}) \quad \forall \mathbf{R} \in \mathrm{SO}(n), \forall \mathbf{t} \in \mathbb{R}^n.$$

Definition 2 *Let $\mathbb{X}$ be an input, $\mathcal{F}(\cdot)$ be an operation and $\mathbb{Z} = \mathcal{F}(\mathbb{X})$ be the corresponding output. The operation $\mathcal{F}(\cdot)$ is called invariant under Euclidean transformation if*

$$\mathbb{Z} = \mathcal{F}(\mathbb{X}\mathbf{R} + \mathbf{t}) \quad \forall \mathbf{R} \in \mathrm{SO}(n), \forall \mathbf{t} \in \mathbb{R}^n.$$

Here, when a motion prediction network $\mathcal{F}_{\mathrm{pred}}(\cdot)$ satisfies Definition 1, $\mathcal{F}_{\mathrm{pred}}(\cdot)$ is said motion equivariant. When an interaction reasoning model $\mathcal{F}_{\mathrm{reason}}(\cdot)$ satisfies Definition 2, $\mathcal{F}_{\mathrm{reason}}(\cdot)$ is considered interaction invariant. We also say the output $\mathbb{Z}$ is equivariant/invariant (to the input motion under Euclidean transformation) if Definition 1/2 is satisfied.

In the following section, we will introduce our motion prediction network along with geometric features that are equivariant under Euclidean transformations, and pattern features along with the interaction reasoning module are invariant under Euclidean transformations.

4. Methodology

In this section, we present EqMotion, a motion prediction network which is equivariant under Euclidean geometric transformations. The whole network architecture is shown in Figure 2. The core of EqMotion is to successively learn equivariant geometric features and invariant pattern features by mutual cooperation in designed equivariant/invariant operations, which not only provide expressive motion and interaction representations, but also preserve equivariant/invariant properties. For agents' motions $\mathbb{X} \in \mathbb{R}^{M \times T_\mathrm{p} \times n}$, the overall procedure of the proposed EqMotion is formulated as

$$\mathbb{G}^{(0)}, \mathbf{H}^{(0)} = \mathcal{F}_{\mathrm{IL}}(\mathbb{X}), \tag{1a}$$

$$\{\mathbf{c}_{ij}\} = \mathcal{F}_{\mathrm{IRM}}(\mathbb{G}^{(0)}, \mathbf{H}^{(0)}), \tag{1b}$$

$$\mathbb{G}^{(\ell+1)} = \mathcal{F}_{\mathrm{EGFL}}^{(\ell)}(\mathbb{G}^{(\ell)}, \mathbf{H}^{(\ell)}, \{\mathbf{c}_{ij}\}), \tag{1c}$$

$$\mathbf{H}^{(\ell+1)} = \mathcal{F}_{\mathrm{IPFL}}^{(\ell)}(\mathbb{G}^{(\ell)}, \mathbf{H}^{(\ell)}), \tag{1d}$$

$$\widehat{\mathbb{Y}} = \mathcal{F}_{\mathrm{EOL}}(\mathbb{G}^{(L)}). \tag{1e}$$

Step (1a) uses an initialization layer $\mathcal{F}_{\mathrm{IL}}(\cdot)$ to obtain initial geometric features $\mathbb{G}^{(0)}$ and pattern features $\mathbf{H}^{(0)}$. For interaction relationship unavailable cases, Step (1b) uses an invariant interaction reasoning module $\mathcal{F}_{\mathrm{IRM}}(\cdot)$ to infer an interaction graph $\{\mathbf{c}_{ij}\}$ whose edge weight $\mathbf{c}_{ij}$ is the interaction category between agent i and j. Step (1c) uses the ℓth equivariant geometric feature learning layer $\mathcal{F}_{\mathrm{EGFL}}^{(\ell)}(\cdot)$ to learn the $(\ell{+}1)$th geometric feature $\mathbb{G}^{(\ell+1)}$. Step (1d) uses the ℓth invariant pattern feature learning layer $\mathcal{F}_{\mathrm{IPFL}}^{(\ell)}(\cdot)$ to learn the $(\ell{+}1)$th pattern feature $\mathbf{H}^{(\ell+1)}$. Step (1c) and Step (1d) will repeat L times. Step (1e) uses an equivariant output layer $\mathcal{F}_{\mathrm{EOL}}(\cdot)$ to obtain the final prediction $\widehat{\mathbb{Y}}$.

Note that i) to incorporate the geometric feature's equivariance, we need to design equivariant operations for the initialization layer $\mathcal{F}_{\mathrm{IL}}(\cdot)$ and the equivariant geometric feature learning layer $\mathcal{F}_{\mathrm{EGFL}}^{(\ell)}(\cdot)$; ii) to introduce the pattern feature's invariance, we need to design invariant operations for the initialization layer $\mathcal{F}_{\mathrm{IL}}(\cdot)$ and the invariant pattern feature

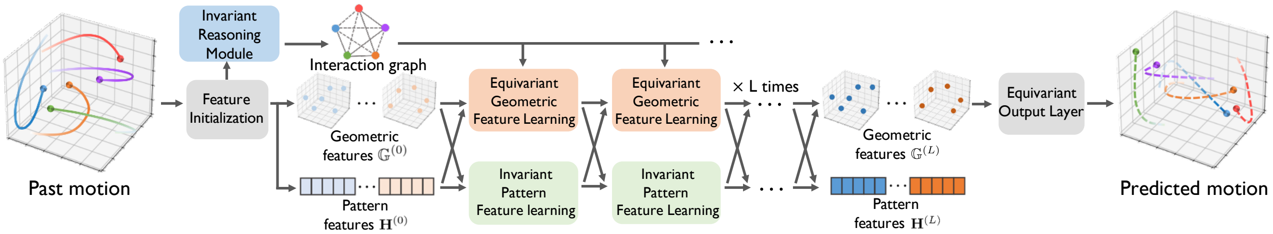

Invariant Reasoning Module
Interaction graph
Past motion
Feature Initialization
Geometric features $\mathbb{G}^{(0)}$
Pattern features $\mathbf{H}^{(0)}$
Equivariant Geometric Feature Learning
Invariant Pattern Feature learning
Equivariant Geometric Feature Learning
Invariant Pattern Feature Learning
× L times
Geometric features $\mathbb{G}^{(L)}$
Pattern features $\mathbf{H}^{(L)}$
Equivariant Output Layer
Predicted motion

Figure 2. EqMotion architecture. In EqMotion, we first use a feature initialization layer to initialize geometric features and pattern features. We then successively update the geometric features and the pattern features by the equivariant geometric feature learning and invariant pattern feature learning layers, obtaining expressive feature representation. We further propose an invariant reasoning module to infer an interaction graph used in equivariant geometric feature learning. Finally, we use an equivariant output layer to obtain the final prediction.

learning layer $\mathcal{F}_{\rm IPFL}^{(\ell)}(\cdot)$; and iii) the interaction graph categorizes agent's spatial interaction into different categories for better interaction representing. The interaction graph is invariant due to the reasoning module $\mathcal{F}_{\rm IRM}(\cdot)$ design. In subsequent sections, we elaborate the details of each step.

4.1. Feature Initialization

The feature initialization layer aims to obtain initial geometric features and pattern features while equipping them with different functionality. The initial geometric feature is denoted as $\mathbb{G}^{(0)} = [\mathbf{G}_1^{(0)}, \cdots, \mathbf{G}_M^{(0)}] \in \mathbb{R}^{M \times C \times n}$ whose ith agent's geometric feature consists of C geometric coordinates. The initial pattern feature is denoted as $\mathbf{H}^{(0)} = [\mathbf{h}_1^{(0)}, \cdots, \mathbf{h}_M^{(0)}] \in \mathbb{R}^{M \times D}$ whose ith agent's pattern feature is a D-dimensional vector. Given the past motions $\mathbb{X}$ whose the ith agent's motion is $\mathbf{X}_i \in \mathbb{R}^{T_{\rm P} \times n}$, we obtain two initial features of the ith agent:

$$\mathbf{G}_i^{(0)} = \phi_{\rm init\_g}(\mathbf{X}_i - \overline{\mathbb{X}}) + \overline{\mathbb{X}},$$
$$\mathbf{V}_i = \triangle \mathbf{X}_i, \rho_i^t = ||\mathbf{V}_i^t||_2, \theta_i^t = {\rm angle}(\mathbf{V}_i^t, \mathbf{V}_i^{t-1}),$$
$$\mathbf{h}_i^{(0)} = \phi_{\rm init\_h}([\,\rho_i; \theta_i\,]),$$

where $\phi_{\rm init\_g}(\mathbf{X}) = \mathbf{W}_{\rm init\_g}\mathbf{X}, \mathbf{W}_{\rm init\_g} \in \mathbb{R}^{C \times T_p}$ is a linear function, $\overline{\mathbb{X}}$ is the mean coordinate of all agents all past timestamps. $\triangle$ is the difference operator to obtain the velocity $\mathbf{V}_i \in \mathbb{R}^{T_{\rm P} \times n}$. $\rho_i \in \mathbb{R}^{T_{\rm P}}$ is the velocity magnitude sequence and $\theta_i \in \mathbb{R}^{T_{\rm P}}$ is the velocity angle sequence. The superscript $(\cdot)^t$ denotes the tth element, $||\cdot||_2$ is vector 2-norm, ${\rm angle}(\cdot)$ is the function calculating the angle between two vectors and $\phi_{\rm init\_h}(\cdot)$ is an embedding function implemented by MLP or LSTM. $[\cdot\,;\cdot]$ represents concatenation.

The geometric feature preserves both equivariant property and motion attributes that are sensitive to Euclidean geometric transforms since we linearly combine the equivariant locations. The pattern feature remains invariant and is sensitive to motion attributes independent of Euclidean geometric transforms.

4.2. Invariant Reasoning Module

For most scenarios, the interaction relationship is implicit and unavailable. Thus, we propose an invariant reasoning module for inferring the interaction category between agents. Note that we design the reasoning module to be invariant as the interaction category is independent of Euclidean transformations. The output of the reasoning module is an invariant interaction graph whose edge weight $\mathbf{c}_{ij} \in [0,1]^K$ is a categorical vector representing the type of interaction between agent i and j. K is the interaction category number. To obtain the interaction categorical vector, we perform a message passing operation using the agent's initial pattern feature $\mathbf{h}_i^{(0)}$ and geometric feature $\mathbf{G}_i^{(0)}$:

$$\mathbf{m}'_{ij} = \phi_{\rm rm}\big([\mathbf{h}_i^{(0)}; \mathbf{h}_j^{(0)}; ||\mathbf{G}_i^{(0)} - \mathbf{G}_j^{(0)}||_{2,{\rm col}}]\big),$$
$$\mathbf{p}'_i = \sum_{j \in \mathcal{N}_i} \mathbf{m}'_{ij}, \quad \mathbf{h}'_i = \phi_{\rm rh}\big([\mathbf{p}'_i; \mathbf{h}_i^{(0)}]\big),$$
$$\mathbf{c}_{ij} = {\rm sm}\Big(\phi_{\rm rc}\big([\mathbf{h}'_i; \mathbf{h}'_j; ||\mathbf{G}_i^{(0)} - \mathbf{G}_j^{(0)}||_{2,{\rm col}}]\big)/\tau\Big),$$

where $||\cdot||_{2,{\rm col}}$ is a column-wise ℓ_2-distance, $\phi_{\rm rm}(\cdot)$, $\phi_{\rm rh}(\cdot)$, $\phi_{\rm rc}(\cdot)$ are learnable functions implemented by MLPs. ${\rm sm}(\cdot)$ is the softmax function and τ is the temperature to control the smoothness of the categorical distribution. The interaction edge weights $\{\mathbf{c}_{ij}\}$ will be end-to-end learned with the whole prediction network.

4.3. Equivariant Geometric Feature Learning

The equivariant geometric feature learning process aims to find more representative agents' geometric features by exploiting their spatial and temporal dependencies, while maintaining the equivariance. During the process, we perform i) equivariant inner-agent attention to exploit the temporal dependencies; ii) equivariant inter-agent aggregation to model spatial interactions; and iii) equivariant non-linear function to further enhance the representation ability.

Equivariant inner-agent attention To exploit the temporal dependency, we perform an attention mechanism on the coordinate dimension of the geometric feature since the coordinate dimension originates from the temporal dimension. Given the ith agent's geometric feature $\mathbf{G}_i^{(\ell)} \in \mathbb{R}^{C \times n}$ and pattern feature $\mathbf{h}_i^{(\ell)}$, we have

$$\mathbf{G}_i^{(\ell)} \leftarrow \phi_{\rm att}^{(\ell)}(\mathbf{h}_i^{(\ell)}) \cdot (\mathbf{G}_i^{(\ell)} - \overline{\mathbb{G}}^{(\ell)}) + \overline{\mathbb{G}}^{(\ell)}, \qquad (2)$$

where $\phi_{\rm att}^{(\ell)}(\cdot) : \mathbb{R}^D \to \mathbb{R}^C$ is a function which learns the attention weight for per coordinate, and $\overline{\mathbb{G}}^{(\ell)}$ is the mean coordinates summing up over all agents and all coordinates.

The above operation will bring the following benefits: i) The learned attention weight is invariant and the learned dependencies between different timestamps of the motion will not be disturbed by irrelevant Euclidean transformations; and ii) The learned geometric feature is equivariant because of the coordinate-wise linear multiplication.

Equivariant inter-agent aggregation The equivariant inter-agent aggregation aims to model spatial interactions between agents. The key idea is to use the reasoned or provided interaction category to learn aggregation weights, and then use the weights to aggregate neighboring agents' geometric features. The aggregation operation reads

$$\begin{aligned} \mathbf{e}_{ij}^{(\ell)} &= \sum_{k=1}^{K} \mathbf{c}_{ij,k}\phi_{e,k}^{(\ell)}([\mathbf{h}_i^{(\ell)};\mathbf{h}_j^{(\ell)};||\mathbf{G}_i^{(\ell)}-\mathbf{G}_j^{(\ell)}||_{2,\mathrm{col}}]), \\ \mathbf{G}_i^{(\ell)} &\leftarrow \mathbf{G}_i^{(\ell)} + \sum_{j\in\mathcal{N}_i} \mathbf{e}_{ij}^{(\ell)}\cdot(\mathbf{G}_i^{(\ell)}-\mathbf{G}_j^{(\ell)}), \end{aligned} \tag{3}$$

where $\mathbf{e}_{ij}^{(\ell)} \in \mathbb{R}^C$ is the learned aggregation weights between agent i and j, $||\cdot||_{2,\mathrm{col}}$ is a column-wise ℓ_2-distance, $\cdot$ is dot product and $\mathcal{N}_i$ is the ith agent's neighboring set. For the kth interaction category, we assign a function $\phi_{e,k}^{(\ell)}(\cdot)$ that is implemented by MLP to model how the interaction works.

The social influence from agent j to agent i is modeled by the coordinates' difference between the two agents. The intuition behind this design is that the mutual force between two objects is always in the direction of the line they formed.

Note that for most scenarios, we do not have an explicit definition of interacted "neighbors", thus we use a fully-connected graph structure, which means every agent's neighbors include all other agents. For some special cases with massive nodes, like point clouds, we construct a local neighboring set by choosing neighbors within a distance threshold.

Equivariant non-linear function According to Eq. (2) and Eq. (3), the operation of inner-agent attention and inter-agent aggregation are linear combinations with agents' coordinates. It is well known that non-linear operation is the key to improving neural networks' expressivity. Therefore, we propose an equivariant non-linear function to enhance the representation ability of our prediction network while preserving equivariance. The key idea is proposing a criterion with the invariance property for splitting different conditions, and for each condition design an equivariant equation. Mathematically, the non-linear function is

$$\mathbf{Q}_i^{(\ell)} = \mathbf{W}_{\mathrm{Q}}^{(\ell)}(\mathbf{G}_i^{(\ell)} - \overline{\mathbb{G}}^{(\ell)}),\ \mathbf{K}_i^{(\ell)} = \mathbf{W}_{\mathrm{K}}^{(\ell)}(\mathbf{G}_i^{(\ell)} - \overline{\mathbb{G}}^{(\ell)}),$$

$$\mathbf{g}_{i,c}^{(\ell+1)} = \begin{cases} \mathbf{q}_{i,c}^{(\ell)} + \overline{\mathbb{G}}^{(\ell)} & \text{if } \langle \mathbf{q}_{i,c}^{(\ell)}, \mathbf{k}_{i,c}^{(\ell)}\rangle \geq 0, \\ \mathbf{q}_{i,c}^{(\ell)} - \langle \mathbf{q}_{i,c}^{(\ell)}, \frac{\mathbf{k}_{i,c}^{(\ell)}}{||\mathbf{k}_{i,c}^{(\ell)}||_2}\rangle \frac{\mathbf{k}_{i,c}^{(\ell)}}{||\mathbf{k}_{i,c}^{(\ell)}||_2} + \overline{\mathbb{G}}^{(\ell)} & \text{otherwise,} \end{cases}$$

where $\mathbf{Q}_i^{(\ell)}$ and $\mathbf{K}_i^{(\ell)}$ is the learned query coordinates and key coordinates, $\mathbf{W}_{\mathrm{K}}^{(\ell)} \in \mathbb{R}^{C\times C}$ are learnable matrices for queries and keys. $\overline{\mathbb{G}}^{(\ell)}$ is the mean coordinates over all agents and all coordinates. $\mathbf{q}_{i,c}^{(\ell)}$, $\mathbf{k}_{i,c}^{(\ell)}$ is the cth coordinate of $\mathbf{Q}_i^{(\ell)}$ and $\mathbf{K}_i^{(\ell)}$. $\langle\cdot,\cdot\rangle$ is the vector inner product.

For every geometric coordinate $\mathbf{g}_{i,c}^{(\ell)}$, we learn a query coordinate $\mathbf{q}_{i,c}^{(\ell)}$ and a key coordinate $\mathbf{k}_{i,c}^{(\ell)}$. We set the criterion as the inner product of the query coordinate and the key coordinate. If the inner product is positive, we directly take the value of the query coordinate as output; otherwise, we clip the query coordinate vector by moving out its parallel components with the key coordinate vector. Finally, we obtain the ith agent's geometric features of the next layer $\mathbf{G}_i^{(l+1)}$ by gathering all the coordinates $\mathbf{g}_{i,c}^{(l+1)}$. The non-linear function is equivariant, since the criterion is invariant and the two equations under two conditions are equivariant.

4.4. Invariant Pattern Feature Learning

The invariant pattern feature learning aims to obtain a more representative agent pattern feature through interacting with neighbors. Here we learn the agent's pattern feature with an invariant message passing mechanism. Specially, we add the relative geometric feature difference into the edge modeling in the message passing to complement the information of relationships between agent absolute locations, which cannot be obtained by solely using the pattern features. Given the ith agent's pattern feature $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^D$ and geometric feature $\mathbf{G}_i^{(\ell)} \in \mathbb{R}^{C\times n}$, the next layer's pattern feature $\mathbf{h}_i^{(l+1)}$ is obtained by

$$\begin{aligned} \mathbf{m}_{ij}^{(\ell)} &= \phi_m^{(\ell)}([\mathbf{h}_i^{(\ell)};\mathbf{h}_j^{(\ell)};||\mathbf{G}_i^{(\ell)}-\mathbf{G}_j^{(\ell)}||_{2,\mathrm{col}}]), \\ \mathbf{p}_i^{(\ell)} &= \sum_{j\in\mathcal{N}(i)} \mathbf{m}_{ij}^{(\ell)},\ \mathbf{h}_i^{(l+1)} = \phi_h^{(\ell)}([\mathbf{h}_i^{(\ell)};\mathbf{p}_i^{(\ell)}]). \end{aligned}$$

Here $\mathbf{m}_{ij}^{(\ell)}$ is the edge features and $\mathbf{p}_i^{(\ell)}$ is the aggregated neighboring features in the message passing. $||\cdot||_{2,\mathrm{col}}$ is a column-wise ℓ_2-distance and $[\cdot\,;\,\cdot]$ denotes concatenation. Functions $\phi_m^{(\ell)}(\cdot)$ and $\phi_h^{(\ell)}(\cdot)$ are implemented by MLPs.

4.5. Equivariant Output Layer

After total L feature learning layers, we obtain the final ith agent's geometric feature $\mathbf{G}_i^{(L)}$ and pattern feature $\mathbf{h}_i^{(L)}$. We use the geometric feature for the final prediction by a linear operation for equivariance: for the ith agent:

$$\widehat{\mathbf{Y}}_i = \mathbf{W}_{\mathrm{out}}(\mathbf{G}_i^{(\ell)} - \overline{\mathbb{G}}^{(\ell)}) + \overline{\mathbb{G}}^{(\ell)},$$

where $\mathbf{W}_{\mathrm{out}} \in \mathbb{R}^{T_{\mathrm{f}}\times D}$ is a learnable weight matrix. Finally, we gather all the agent prediction $\widehat{\mathbf{Y}}_i$ to have the final predicted motions $\widehat{\mathbb{Y}}$.

4.6. Theoretical Analysis

In this section, we analyze our network's equivariance property and the interaction reasoning module's invariance property, as stated in the following theorem. Let $\mathbb{G}^{(\ell)} \in \mathbb{R}^{M\times C\times n}$ be all agents' geometric features, $\mathbf{H}^{(\ell)} \in \mathbb{R}^{M\times D}$ be all agents' pattern features at the ℓth layer, and $\{\mathbf{c}_{ij}\}$ be the set of all the interaction categorical vectors.

Theorem 1 *For arbitrary translation vector* $\mathbf{t} \in \mathbb{R}^n$ *and rotation (or reflection) matrix* $\mathbf{R} \in \mathrm{SO}(n)$*, the modules with equivariance and invariance in our network satisfy:*

1. For the initialization layer $\mathcal{F}_{\mathrm{IL}}(\cdot)$*, the initial geometric feature is equivariant and the initial pattern feature is invariant:*

$$\mathbb{G}^{(0)}\mathbf{R}+\mathbf{t},\ \mathbf{H}^{(0)} = \mathcal{F}_{\mathrm{IL}}(\mathbb{X}\mathbf{R}+\mathbf{t}).$$

2. The reasoning module $\mathcal{F}_{\mathrm{IRM}}(\cdot)$ *along with reasoned interaction categorical vectors* $\{\mathbf{c}_{ij}\}$ *is invariant:*

$$\{\mathbf{c}_{ij}\} = \mathcal{F}_{\mathrm{IRM}}(\mathbb{G}^{(0)}\mathbf{R}+\mathbf{t}, \mathbf{H}^{(0)}).$$

3. The ℓ*th geometric feature learning layer* $\mathcal{F}^{(\ell)}_{\mathrm{EGFL}}(\cdot)$ *is equivariant:*

$$\mathbb{G}^{(l+1)}\mathbf{R}+\mathbf{t} = \mathcal{F}^{(\ell)}_{\mathrm{EGFL}}(\mathbb{G}^{(\ell)}\mathbf{R}+\mathbf{t}, \mathbf{H}^{(\ell)}, \{\mathbf{c}_{ij}\}).$$

4. The ℓ*th pattern feature learning layer* $\mathcal{F}^{(\ell)}_{\mathrm{IPFL}}(\cdot)$ *is invariant:*

$$\mathbf{H}^{(l+1)} = \mathcal{F}^{(\ell)}_{\mathrm{IPFL}}(\mathbb{G}^{(\ell)}\mathbf{R}+\mathbf{t}, \mathbf{H}^{(\ell)}).$$

5. The output layer $\mathcal{F}_{\mathrm{EOL}}(\cdot)$ *is equivariant:*

$$\widehat{\mathbb{Y}}\mathbf{R}+\mathbf{t} = \mathcal{F}_{\mathrm{EOL}}(\mathbb{G}^{(L)}\mathbf{R}+\mathbf{t}).$$

See the detailed proof in Appendix. Theorem 1 presents the equivariance/invariance properties with the Euclidean transformation for each operation in the proposed network. Based on Theorem 1, by combining all network operations together, we can show that the whole network is equivariant:

Corollary 1 *For arbitrary translation vector* $\mathbf{t} \in \mathbb{R}^n$ *and rotation matrix* $\mathbf{R} \in \mathrm{SO}(n)$*, our whole network EqMotion* $\mathcal{F}_{\mathrm{pred}}(\cdot)$ *satisfies:*

$$\widehat{\mathbb{Y}}\mathbf{R}+\mathbf{t} = \mathcal{F}_{\mathrm{pred}}(\mathbb{X}\mathbf{R}+\mathbf{t}).$$

5. Experiment

In this section, we validate our method on four different scenarios: particle dynamics, molecule dynamics, 3D human skeleton motion, and pedestrian trajectories. The detailed dataset description, implementation details, and additional experiment results are elaborated in the Appendix.

Metric We use: i) Average Displacement Error (ADE) and Final Displacement Error (FDE). ADE/FDE is the ℓ_2 distance of the predicted whole motion/endpoint to the ground truth of the whole motion/endpoint; ii) Mean Per Joint Position Error (MPJPE). It records average ℓ_2 distance between predicted joints and target ones at each future timestamp.

5.1. Scenario 1: Particle Dynamic Prediction

We use the particle N-body simulation [28] in a 3D space similar to [14, 58]. In the Springs simulation, particles are randomly connected by a spring. In the Charged simulation, particles are randomly charged or uncharged.

Validation on interaction reasoning Since we can get the ground-truth interaction through simulation, we evaluate the ability of interaction reasoning as a category recognition task. To evaluate the robustness of the reasoning result with the Euclidean transformation, we introduce recognition "consistency", which is the ratio of the same recognition result under 20 random Euclidean transformations. The "Supervised" represents an upper bound that uses the ground-truth category to train the model. Table 1 reports the comparison of reasoning results on both springs and charged simulations, for recognizing whether there is a spring connection or electrostatic force. We see that i) our method achieves significant improvement on the recognition accuracy and has more robust reasoning results due to the invariance of our reasoning module; and ii) our method achieves a close performance to the upper bound, reflecting that our method is capable of reasoning a robust and accurate interaction category.

Table 1. Interaction recognition accuracy and consistency (mean ± std in % in 5 independent runs) on the physical simulation.

Model	**Springs**		**Charged**	
	Accuracy	Consistency	Accuracy	Consistency
Corr.(path) [28]	58.1 ± 0.0	99.8 ± 0.1	57.5 ± 0.1	87.9 ± 0.1
Corr.(LSTM) [28]	53.5 ± 0.5	92.4 ± 2.1	57.2 ± 0.4	91.7 ± 1.1
EGNN [58]	61.0 ± 1.3	**100.0** ± 0.0	58.2 ± 1.4	**100.0** ± 0.0
NRI [28]	93.0 ± 1.1	93.7 ± 1.2	70.0 ± 0.6	88.5 ± 1.3
dNRI [17]	93.3 ± 2.0	89.6 ± 2.0	70.4 ± 1.7	83.6 ± 1.8
Ours	**97.6** ± 1.1	**100.0** ± 0.0	**80.9** ± 3.4	**100.0** ± 0.0
Supervised	98.7 ± 0.2	100.0 ± 0.0	97.4 ± 0.2	100.0 ± 0.0

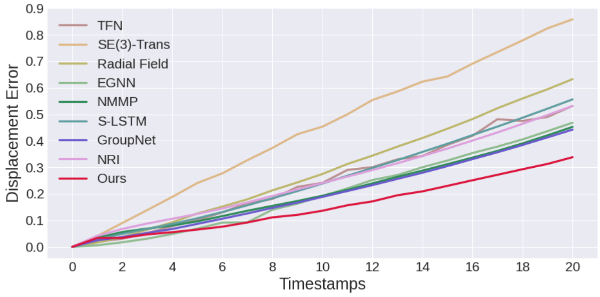


Figure 3. Comparison of displacement error across different timestamps on the physical simulation.

Validation on future prediction We conduct the experiment to evaluate the prediction performance on the charged settings. Figure 3 compares the displacement error at all future timestamps of different methods. We see that our method (red line) achieves state-of-the-art prediction performance, reflecting our model's efficiency in future prediction.

5.2. Scenario 2: Molecule Dynamic Prediction

We adopt the MD17 [7] dataset which contains the motions of different molecules generated via a molecular dynamics simulation environment. The goal is to predict the motions of every atom of the molecule. We randomly pick four kinds of molecules: Aspirin, Benzene, Ethanol and Malonaldehyde and learn a prediction model for each molecule.

Table 4 presents the comparison of different motion prediction methods. We achieve state-of-the-art prediction performance on all four molecules. The ADE/FDE across four molecules is decreased by 34.2%/30.1% on average.

Table 2. Comparisons of short-term skeleton motion prediction on 11 representative actions and average results across all actions on H3.6M.

Motion	Walking				Eating				Smoking				Discussion				Directions				Phoning			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res-sup. (CVPR'17)	29.4	50.8	76.0	81.5	16.8	30.6	56.9	68.7	23.0	42.6	70.1	82.7	32.9	61.2	90.9	96.2	35.4	57.3	76.3	87.7	38.0	69.3	115.0	126.7
Traj-GCN (ICCV'19)	12.3	23.0	39.8	46.1	8.4	16.9	33.2	40.7	7.9	16.2	31.9	38.9	12.5	27.4	58.5	71.7	9.0	19.9	43.4	53.7	10.2	21.0	42.5	52.3
DMGNN (CVPR'20)	17.3	30.7	54.6	65.2	11.0	21.4	36.2	43.9	9.0	17.6	32.1	40.3	17.3	34.8	61.0	69.8	13.1	24.6	64.7	81.9	12.5	25.8	48.1	58.3
MSRGCN (ICCV'21)	12.2	22.7	38.6	45.2	8.4	17.1	33.0	40.4	8.0	16.3	31.3	38.2	12.0	26.8	57.1	69.7	8.6	19.7	43.3	53.8	10.1	20.7	41.5	51.3
PGBIG (CVPR'22)	10.2	19.8	34.5	40.3	7.0	15.1	30.6	38.1	6.6	14.1	28.2	34.7	10.0	23.8	53.6	66.7	7.2	17.6	40.9	51.5	8.3	18.3	38.7	48.4
SPGSN (ECCV'22)	10.1	19.4	34.8	41.5	7.1	14.9	30.5	37.9	6.7	13.8	28.0	34.6	10.4	23.8	53.6	67.1	7.4	17.2	39.8	50.3	8.7	18.3	38.7	48.5
EqMotion (Ours)	**9.0**	**17.5**	**32.6**	**39.2**	**6.3**	**13.6**	**28.9**	**36.5**	**5.5**	**11.3**	**23.0**	**29.3**	**8.2**	**18.9**	**42.1**	**53.9**	**6.3**	**15.8**	**38.9**	**50.1**	**7.4**	**16.7**	**36.9**	**47.0**
Motion	Posing				Sitting				Sitting Down				Waiting				Walking Together				Average			
millisecond	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Res-sup. (CVPR'17)	36.1	69.1	130.5	157.1	42.6	81.4	134.7	151.8	47.3	86.0	145.8	168.9	30.6	57.8	106.2	121.5	26.8	50.1	80.2	92.2	34.7	62.0	101.1	115.5
Traj-GCN (ICCV'19)	13.7	29.9	66.6	84.1	10.6	21.9	46.3	57.9	16.1	31.1	61.5	75.5	11.4	24.0	50.1	61.5	10.5	21.0	38.5	45.2	12.7	26.1	52.3	63.5
DMGNN (CVPR'20)	15.3	29.3	71.5	96.7	11.9	25.1	44.6	50.2	15.0	32.9	77.1	93.0	12.2	24.2	59.6	77.5	14.3	26.7	50.1	63.2	17.0	33.6	65.9	79.7
MSRGCN (ICCV'21)	12.8	29.4	67.0	85.0	10.5	22.0	46.3	57.8	16.1	31.6	62.5	76.8	10.7	23.1	48.3	59.2	10.6	20.9	37.4	43.9	12.1	25.6	51.6	62.9
PGBIG (CVPR'22)	10.7	25.7	60.0	76.6	8.8	19.2	42.4	53.8	13.9	27.9	57.4	71.5	8.9	20.1	43.6	54.3	8.7	18.6	34.4	41.0	10.3	22.7	47.4	58.5
SPGSN (ECCV'22)	10.7	25.3	59.9	76.5	9.3	19.4	42.3	53.6	14.2	27.7	56.8	**70.7**	9.2	19.8	43.1	54.1	8.9	18.2	33.8	40.9	10.4	22.3	47.1	58.3
EqMotion (Ours)	**8.2**	**18.9**	**43.4**	**57.5**	**8.1**	**18.0**	**41.2**	**52.9**	**13.0**	**26.5**	**56.2**	**70.7**	**7.6**	**17.4**	**39.9**	**51.1**	**7.8**	**16.1**	**30.6**	**37.1**	**9.1**	**20.1**	**43.7**	**55.0**

Table 3. Comparisons of long-term skeleton motion prediction on 8 representative actions and average results across all actions on H3.6M.

Motion	Walking		Eating		Smoking		Discussion		Greeting		Phoning		Posing		Walking Together		Average	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res-Sup. [50]	81.7	100.7	79.9	100.2	94.8	137.4	121.3	161.7	156.3	184.3	143.9	186.8	165.4	236.8	173.6	202.3	129.2	165.0
Traj-GCN [47]	54.1	59.8	53.4	77.8	50.7	72.6	91.6	121.5	115.4	148.8	69.2	103.1	114.5	173.0	55.0	65.6	81.6	114.3
DMGNN [41]	71.4	85.8	58.1	86.7	50.9	72.2	81.9	138.3	144.5	170.5	71.3	108.4	125.5	188.2	70.5	86.9	93.6	127.6
MSRGCN [9]	52.7	63.0	52.5	77.1	49.5	71.6	88.6	117.6	116.3	147.2	68.3	104.4	116.3	174.3	52.9	65.9	81.1	114.2
PGBIG [44]	48.1	56.4	51.1	76.0	46.5	69.5	87.1	118.2	110.2	143.5	65.9	102.7	106.1	164.8	51.9	64.3	76.9	110.3
SPGSN [40]	46.9	53.6	49.8	73.4	46.7	68.6	89.7	118.6	111.0	143.2	66.7	102.5	110.3	165.4	49.8	60.9	77.4	109.6
EqMotion (Ours)	**43.4**	**52.8**	**48.4**	**73.0**	**41.0**	**63.4**	**75.3**	**105.6**	**108.7**	**142.0**	**64.7**	**101.0**	**84.9**	**139.4**	**44.5**	**56.0**	**73.4**	**106.9**

Table 4. Prediction ADE/FDE ($\times 10^{-2}$) on the MD17 dataset.

	Aspirin	Benzene	Ethanol	Malonaldehyde
Radial Field [31]	17.98/26.20	7.73/12.47	8.10/10.61	16.53/25.10
TFN [61]	15.02/21.35	7.55/12.30	8.05/10.57	15.21/24.32
SE(3)-Trans [14]	15.70/22.39	7.62/12.50	8.05/10.86	15.44/24.47
EGNN [58]	14.61/20.65	7.50/12.16	8.01/10.22	15.21/24.00
LSTM	17.59/24.79	6.06/9.46	7.73/9.88	15.14/22.90
S-LSTM [1]	13.12/18.14	3.06/3.52	7.23/9.85	11.93/18.43
NRI [28]	12.60/18.50	1.89/2.58	6.69/8.78	12.79/19.86
NMMP [22]	10.41/14.67	2.21/3.33	6.17/7.86	9.50/14.89
GroupNet [69]	10.62/14.00	2.02/2.95	6.00/7.88	7.99/12.49
EqMotion(Ours)	**5.95/8.38**	**1.18/1.73**	**5.05/7.02**	**5.85/9.02**

5.3. Scenario 3: Human Skeleton Motion Prediction

We conduct experiments on the Human 3.6M (H3.6M) dataset [25] for 3D human skeleton motion prediction. H3.6M contains 7 subjects performing 15 actions. Following the standard paradigm [41, 50], we train the models on 6 subjects and tested on the specific clips of the 5th subject.

Short-term motion prediction Short-term motion prediction aims to predict future poses within 400 milliseconds. We compare EqMotion with six state-of-the-art methods. Table 2 presents the comparison of prediction MPJPEs. See the overall result in the Appendix. We see that i) EqMotion obtains superior performance at most timestamps across all the actions; and ii) compared to the baselines, EqMotion achieves much lower MPJPEs by 8.6% on average.

Long-term motion prediction Long-term motion prediction aims to predict the poses over 400 milliseconds. Table 3 presents the prediction MPJPEs of various methods. We see that EqMotion achieves more effective prediction on most actions and has lower MPJPEs by 3.8 % on average.

Visualization results Figure 4 visualizes and compares the predictions on H3.6M. Baselines fail to predict future movements. EqMotion completes the action more precisely.

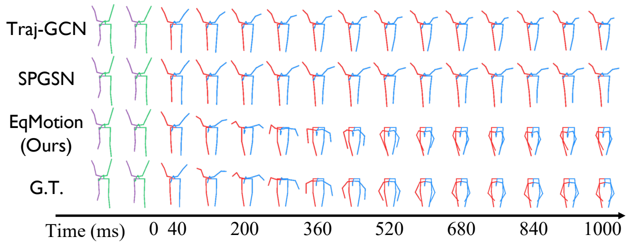


Figure 4. Qualitative comparison on the action 'Sitting Down' of H3.6M. EqMotion completes the action more accurately.

The quantitative and visualization results reveal that EqMotion outperforms many previous methods that are task-specific, reflecting the effectiveness of EqMotion.

5.4. Scenario 4: Pedestrian Trajectory Prediction

We conduct experiments on pedestrian trajectory prediction using the ETH-UCY dataset [34, 54], which contains 5 subsets, ETH, HOTEL, UNIV, ZARA1, and ZARA2. Following the standard setting [1, 20, 76], we use 3.2 seconds (8 timestamps) to predict the 4.8 seconds (12 timestamps).

Here we apply our EqMotion to two prediction modes: deterministic and multi-prediction. Deterministic means the model only outputs a single prediction for each input motion. Multi-prediction means the model has 20 predictions for each input motion. Under multi-prediction, ADE and FDE will be calculated by the best-performed predicted motion. To adapt to multi-prediction, we slightly modify EqMotion to repeat the last feature updating layer and the output layer 20 times in parallel to have a multi-head prediction, see the details in the Appendix. Table 5 compares our method with sixteen baselines. We observe that i) EqMotion achieves state-of-the-art performance with the lowest average ADE and FDE, outperforming many baseline methods that are specifically designed for this task. EqMotion reduces the FDE by 10.4% and 7.9% under deterministic and multi-prediction settings; and ii) EqMotion achieves the best or the

Table 5. Prediction performance on the ETH-UCY dataset. The **bold**/underline font denotes the best/second best result.

	Performance (ADE/FDE)					
Deterministic	ETH	Hotel	Univ	Zara1	Zara2	Average
S-LSTM [1]	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
SGAN-ind [20]	1.13/2.21	1.01/2.18	0.60/1.28	0.42/0.91	0.52/1.11	0.74/1.54
Traj++ [55]	1.02/2.00	0.33/0.62	0.53/1.19	0.44/0.99	0.32/0.73	0.53/1.11
TransF [16]	1.03/2.10	0.36/0.71	0.53/1.32	0.44/1.00	0.34/0.76	0.54/1.17
MemoNet [70]	1.00/2.08	0.35/0.67	0.55/1.19	0.46/1.00	0.37/0.82	0.55/1.15
EqMotion(Ours)	**0.96/1.92**	**0.30/0.58**	**0.50/1.10**	**0.39/0.86**	**0.30/0.68**	**0.49/1.03**
Multi-prediction	ETH	Hotel	Univ	Zara1	Zara2	Average
SGAN [20]	0.87/1.62	0.67/1.37	0.76/0.52	0.35/0.68	0.42/0.84	0.61/1.21
NMMP [22]	0.61/1.08	0.33/0.63	0.52/1.11	0.32/0.66	0.43/0.85	0.41/0.82
Traj++ [55]	0.61/1.02	0.19/0.28	0.30/0.54	0.24/0.42	0.18/0.31	0.30/0.51
PECNet [45]	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48
Agentformer [76]	0.45/0.75	0.14/0.22	0.25/0.45	0.18/**0.30**	0.14/0.24	0.23/0.39
GroupNet [69]	0.46/0.73	0.15/0.25	0.26/0.49	0.21/0.39	0.17/0.33	0.25/0.44
MID [18]	**0.39**/0.66	0.13/0.22	**0.22**/0.45	**0.17/0.30**	**0.13**/0.27	**0.21**/0.38
GP-Graph [2]	0.43/0.63	0.18/0.30	0.24/0.42	**0.17**/0.31	0.15/0.29	0.23/0.39
EqMotion(Ours)	0.40/**0.61**	**0.12/0.18**	0.23/**0.43**	0.18/0.32	**0.13/0.23**	**0.21/0.35**

Table 6. Ablation study on key modules of EqMotion on H3.6M.

EGFL	IPFL	IRM	80ms	160ms	320ms	400ms	Average
			12.9	31.9	68.2	82.4	48.9
✓			10.1	22.6	48.7	60.7	35.5
✓	✓		9.2	20.8	45.4	57.0	33.1
✓	✓	✓	**9.1**	**20.1**	**43.7**	**55.0**	**32.0**

Table 7. Ablation study on operations in the equivariant geometric feature learning on H3.6M dataset.

Ablation	80ms	160ms	320ms	400ms	Average
w/o Inner att	9.2	20.5	44.3	55.7	32.4
w/o Inter agg	9.7	22.0	47.2	58.9	34.5
w/o Non-linear	9.4	21.3	46.7	58.6	34.0
EqMotion	**9.1**	**20.1**	**43.7**	**55.0**	**32.0**

second best ADE/FDE result on most subsets, reflecting its effectiveness on the pedestrian trajectory prediction.

5.5. Ablation Studies

Effect of network modules We explore the effect of three proposed key modules in EqMotion on the H3.6M dataset, including the equivariant geometric feature learning (EGFL), the invariant pattern feature learning (IPFL) and the invariant reasoning module (IRM). Table 6 presents the experimental result. It is observed that i) the proposed three key modules all contribute to an accurate prediction; and ii) the equivariant geometric feature learning module is most important since learning a comprehensive equivariant geometric feature directly for prediction is the most important.

Effect of equivariant operations We explore the effect of three proposed operations in the equivariant feature learning module in EqMotion, including the inner-agent attention (Inner att), inter-agent aggregation (Inter agg) and non-linear function (Non-linear). Table 7 presents the results. We see that the proposed three key operations all contribute to promoting an accurate prediction.

Different amounts of training data Figure 5 presents the comparison of model performance under different amounts of training data on H3.6M. We see that i) our method achieves the best prediction performance under all training data ratios; and ii) our method even outperforms some full-data using baselines by only using 5% of training data since the equivariant design promotes the network generalization ability under Euclidean transformations.

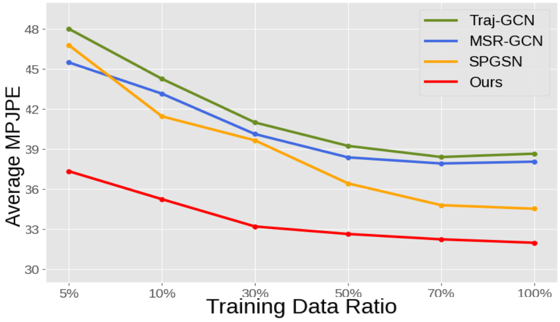


Figure 5. Comparison of model performance on different amounts of data in short-term prediction on H3.6M dataset.

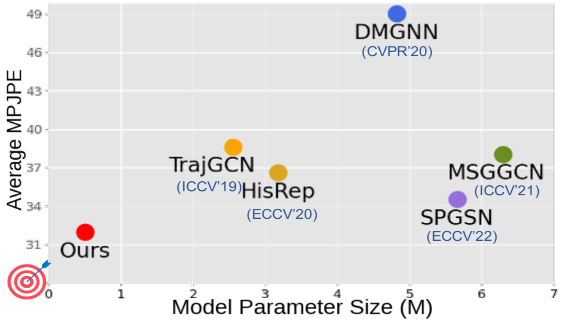


Figure 6. Comparison of model size and MPJPE in short-term prediction on H3.6M dataset. The target means the ideal model.

Model size Figure 6 compares EqMotion to existing methods in terms of the model size and prediction results in short-term prediction on H3.6M. We can observe that EqMotion has the smallest model size (less than 30% of other models' sizes) with the lowest MPJPE thanks to the equivariant design that compacts the model free from generalizing over rotations and translations of the data.

6. Conclusion

In this work, we present EqMotion, a motion prediction network that is theoretically equivariant under Euclidean transformations. EqMotion includes three novel designs: the equivariant geometric feature learning, the invariant pattern feature learning and the invariant reasoning module. We evaluate our method on four different scenarios and our method achieves state-of-the-art prediction performance.

Acknowledgements. This project is partially supported by National Natural Science Foundation of China under Grant 62171276, the Science and Technology Commission of Shanghai Municipal under Grant 21511100900 and 22DZ2229005, and the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG2-RP-2021-023). Robby T. Tan's work is supported by MOE AcRF Tier, A-0009455-01-00.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 3, 7, 8

[2] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Learning pedestrian group representations for multi-modal trajectory prediction. In *European Conference on Computer Vision*, pages 270–289. Springer, 2022. 8

[3] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016. 1, 2

[4] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*, pages 226–242. Springer, 2020. 3

[5] Sergio Casas, Wenjie Luo, and Raquel Urtasun. Intentnet: Learning to predict intention from raw sensor data. In *Conference on Robot Learning*, pages 947–956. PMLR, 2018. 3

[6] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*, 2019. 3

[7] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017. 1, 6

[8] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 2

[9] 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*, pages 11467–11476, 2021. 3, 7

[10] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12200–12209, 2021. 2

[11] Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. Polar transformer networks. *arXiv preprint arXiv:1709.01889*, 2017. 2

[12] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pages 3165–3176. PMLR, 2020. 2

[13] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015. 3

[14] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020. 2, 6, 7

[15] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11533, 2020. 3

[16] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10335–10342. IEEE, 2021. 8

[17] Colin Graber and Alexander G Schwing. Dynamic neural relational inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8513–8522, 2020. 2, 3, 6

[18] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17113–17122, 2022. 8

[19] Xiao Guo and Jongmoo Choi. Human motion prediction via learning local structure representations and temporal dependencies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2580–2587, 2019. 3

[20] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 3, 7, 8

[21] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 3

[22] Yue Hu, Siheng Chen, Ya Zhang, and Xiao Gu. Collaborative motion prediction via neural motion message passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6319–6328, 2020. 3, 7, 8

[23] Wenbing Huang, Jiaqi Han, Yu Rong, Tingyang Xu, Fuchun Sun, and Junzhou Huang. Equivariant graph mechanics networks with constraints. *arXiv preprint arXiv:2203.06442*, 2022. 2

[24] Michael J Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In *International Conference on Machine Learning*, pages 4533–4543. PMLR, 2021. 2

[25] 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, 2013. 7

[26] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh 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. 3

[27] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020. 2

[28] 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. 1, 2, 6, 7

[29] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European conference on computer vision*, pages 201–214. Springer, 2012. 3

[30] Miltiadis Kofinas, Naveen Nagaraja, and Efstratios Gavves. Roto-translated local coordinate frames for interacting dynamical systems. *Advances in Neural Information Processing Systems*, 34:6417–6429, 2021. 2

[31] Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: sampling configurations for multi-body systems with symmetric energies. *arXiv preprint arXiv:1910.00753*, 2019. 7

[32] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017. 3

[33] Andreas M Lehrmann, Peter V Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1314–1321, 2014. 3

[34] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 7

[35] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 163–168. IEEE, 2011. 1

[36] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5226–5234, 2018. 3

[37] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3

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

[39] 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(6):3316–3333, 2021. 3

[40] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3d human motion prediction. In *European Conference on Computer Vision*, 2022. 3, 7

[41] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2020. 3, 7

[42] Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10508–10518, 2020. 3

[43] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *European Conference on Computer Vision*, pages 541–556. Springer, 2020. 3

[44] 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*, pages 6437–6446, 2022. 7

[45] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European Conference on Computer Vision*, pages 759–776. Springer, 2020. 3, 8

[46] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020. 3

[47] 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*, pages 9489–9497, 2019. 3, 7

[48] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Mantra: Memory augmented networks for multiple trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7143–7152, 2020. 3

[49] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5048–5057, 2017. 2

[50] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 3, 7

[51] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942. IEEE, 2009. 3

[52] Jeremy Morton, Tim A Wheeler, and Mykel J Kochenderfer. Analysis of recurrent neural networks for probabilistic modeling of driver behavior. *IEEE Transactions on Intelligent Transportation Systems*, 18(5):1289–1298, 2016. 3

[53] Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li F Fei-Fei, Josh Tenenbaum, and Daniel L Yamins. Flexible neural representation for physics prediction. *Advances in neural information processing systems*, 31, 2018. 2

[54] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009. 7

[55] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Multi-agent generative trajectory forecasting with heterogeneous data for control. 2020. 3, 8

[56] Alvaro Sanchez-Gonzalez, Victor Bapst, Kyle Cranmer, and Peter Battaglia. Hamiltonian graph networks with ode integrators. *arXiv preprint arXiv:1909.12790*, 2019. 2

[57] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, pages 8459–8468. PMLR, 2020. 2

[58] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021. 2, 6, 7

[59] Bohan Tang, Yiqi Zhong, Ulrich Neumann, Gang Wang, Ya Zhang, and Siheng Chen. Collaborative uncertainty in multiagent trajectory forecasting. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[60] Graham W Taylor and Geoffrey E Hinton. Factored conditional restricted boltzmann machines for modeling motion style. In *Proceedings of the 26th annual international conference on machine learning*, pages 1025–1032, 2009. 3

[61] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 2, 7

[62] Benjamin Ummenhofer, Lukas Prantl, Nils Thuerey, and Vladlen Koltun. Lagrangian fluid simulation with continuous convolutions. In *International Conference on Learning Representations*, 2019. 2

[63] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE international Conference on Robotics and Automation (ICRA)*, pages 4601–4607. IEEE, 2018. 3

[64] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision*, pages 3332–3341, 2017. 3

[65] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):283–298, 2007. 3

[66] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2018. 2

[67] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017. 2

[68] Chenxin Xu, Siheng Chen, Maosen Li, and Ya Zhang. Invariant teacher and equivariant student for unsupervised 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3013–3021, 2021. 1

[69] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6507, 2022. 3, 7, 8

[70] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2022. 3, 8

[71] Chenxin Xu, Yuxi Wei, Bohan Tang, Sheng Yin, Ya Zhang, and Siheng Chen. Dynamic-group-aware networks for multiagent trajectory prediction with relational reasoning. *arXiv preprint arXiv:2206.13114*, 2022. 2, 3

[72] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, 2022. 3

[73] Xingyi Yang, Daquan Zhou, Songhua Liu, Jingwen Ye, and Xinchao Wang. Deep model reassembly. In *Conference on Neural Information Processing Systems*, 2022. 2

[74] Yiding Yang, Zunlei Feng, Mingli Song, and Xinchao Wang. Factorizable graph convolutional networks. In *Conference on Neural Information Processing Systems*, 2020. 2

[75] Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. Distilling knowledge from graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[76] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021. 3, 7, 8

[77] Yiqi Zhong, Zhenyang Ni, Siheng Chen, and Ulrich Neumann. Aware of the history: Trajectory forecasting with the local behavior data. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 393–409. Springer, 2022. 3