Staged Contact-Aware Global Human Motion Forecasting

Luca Scofano[1]
scofano@diag.uniroma1.it
Alessio Sampieri[1]
sampieri@diag.uniroma1.it
Elisabeth Schiele[2]
elisabeth.schiele@tum.de
Edoardo De Matteis[1]
dematteis@di.uniroma1.it
Laura Leal-Taixé[2]
leal.taixe@tum.de
Fabio Galasso[1]
galasso@di.uniroma1.it

[1] Sapienza University of Rome
[2] Technical University of Munich

Abstract

Scene-aware global human motion forecasting is critical for manifold applications, including virtual reality, robotics, and sports. The task combines human trajectory and pose forecasting within the provided scene context, which represents a significant challenge.

So far, only Mao *et al*. NeurIPS'22 have addressed scene-aware global motion, cascading the prediction of future scene contact points and the global motion estimation. They perform the latter as the end-to-end forecasting of future trajectories and poses. However, end-to-end contrasts with the coarse-to-fine nature of the task and it results in lower performance, as we demonstrate here empirically.

We propose a STAGed contact-aware global human motion forecasting (*STAG*), a novel three-stage pipeline for predicting global human motion in a 3D environment. We first consider the scene and the respective human interaction as contact points. Secondly, we model the human trajectory forecasting within the scene, predicting the coarse motion of the human body as a whole. The third and last stage matches a plausible fine human joint motion to complement the trajectory considering the estimated contacts.

Compared to the state-of-the-art (SoA), *STAG* achieves a 1.8% and 16.2% overall improvement in pose and trajectory prediction, respectively, on the scene-aware GTA-IM dataset. A comprehensive ablation study confirms the advantages of staged modeling over end-to-end approaches. Furthermore, we establish the significance of a newly proposed temporal counter called the "time-to-go", which tells how long it is before reaching scene contact and endpoints. Notably, *STAG* showcases its ability to generalize to datasets lacking a scene and achieves a new state-of-the-art performance on CMU-Mocap, without leveraging any social cues. Our code is released at: https://github.com/L-Scofano/STAG.

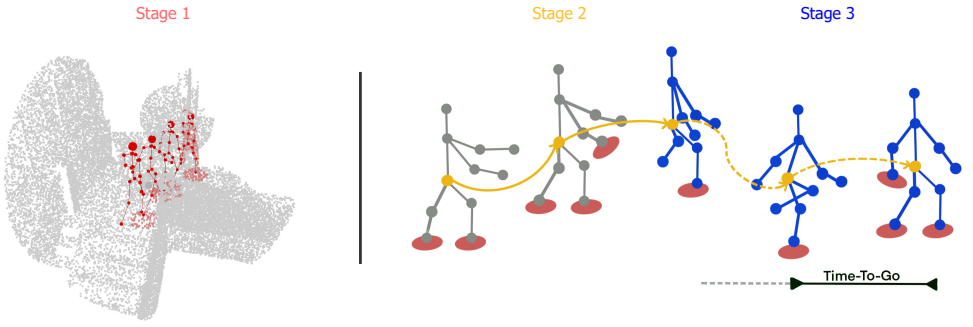

Figure 1: STAG forecasts scene-aware global human motion by three coarse-to-fine stages: (i) estimate the present and future contact points (light red) given the scene and the ground truth body joints (red); (ii) predict the future trajectory (dashed yellow), i.e. the future position of the root joints, given the past (solid yellow); (iii) predict the future body joints (blue) from the observed ones (gray). Each stage of STAG conditions on the previous, so trajectory forecasting leverages future estimated contact points, and pose forecasting leverages both other estimates. Awareness of the time-to-go (black arrow), the passing time between the current prediction and the end one, improves performance.

1 Introduction

Humans are inherently predicting the near future at all times [12, 13]. As humans and machines coexist more, predicting human motion in the immediate future becomes critical for human-robot interaction, e.g., in industrial environments [9, 29, 52] or breaking-in-time to avoid collisions [18, 47, 71]. Human motion forecasting generally includes *local* pose forecasting [20, 21, 39, 40, 41, 54, 64], in which the joint locations are predicted with respect to the root joint, and *global* pose forecasting [42, 69, 70], which takes into account the positions of joints and the root in relation to a global coordinate system.

One common issue in human motion forecasting is the omission of the environment. It leads to contrived motion when the model is used in more realistic scenarios, such as *ghost motions*, i.e., phasing through solid objects. To our knowledge, [42] is the only work that accomplishes scene-aware global human forecasting. They first process the scene and emphasize the human-scene interaction through contact points. Subsequently, they employ an end-to-end approach to model individuals' trajectories and poses. Although contact points have shown effectiveness, employing end-to-end modeling for both trajectory and pose is suboptimal. The pose of an individual is influenced by their motion trajectory, interaction with the surrounding scene, and the pose in previous frames. However, the pose is typically not the underlying cause of the pathway. Essentially, global motion forecasting naturally aligns with a coarse-to-fine methodology that considers the scene, the trajectory, and the human pose.

We propose a novel model for *STAGed contact-aware global human motion forecasting* (*STAG*) that cascades three coarse-to-fine processing stages: (i) predicting the contact points, (ii) using them to forecast the trajectories, (iii) estimating the body pose (see Fig.1). Our three-stage pipeline predicts the future motion autoregressively, conditioning each stage on the previous ones. We condition the global motion on end goals and propose a time-to-go

temporal encoding of the remaining duration until the endpoints are reached, i.e. informs the model on how many frames are missing.

We quantitatively evaluate the design choices of *STAG* and compare them to the SoA on the available GTA-IM dataset [42]. Overall, we get up to a 21.1% improvement on the path error with 16.2% on average, while on the pose error, we get up to 5.4% less error and 1.8% on average. We also show the generalizability of *STAG* by testing it on CMU-Mocap [10], a well-established multi-person dataset without scenes. To account for the missing scene, we only assume a planar ground. *STAG* sets a new SoA without leveraging social cues, which SoA methods use [1, 17, 59]. Overall, our contributions are threefold:

1. We introduce a novel three-stage, coarse-to-fine model, which cascadedly processes the contact points, the trajectories, and the poses of people.
2. We introduce a learnable temporal counter for the time-to-go to align the predictions with the missing time before the endpoint.
3. We perform a thorough analysis on GTA-IM [42], where we set a new SoA, and generalize *STAG* to the *scene-less* CMU-Mocap [10].

2 Related Work

We discuss literature relating to the three core aspects of contact-aware global human motion forecasting: human-scene interaction (Sec. 2.1), trajectory forecasting (Sec. 2.2), and human motion forecasting (Sec. 2.3).

2.1 Human-Scene Interaction

Human motion forecasting is inherently influenced by the scene context in which it occurs, thus, considering the interaction between humans and their surroundings is crucial for motion forecasting.

In motion synthesis, this shift towards including more contextual information can already be seen [7, 23, 26, 61, 69], and some works in trajectory forecasting also consider contextual information [8, 16, 30]. In human motion forecasting, scene information has been widely disregarded, with only a few works considering implicitly learning from the scene [5, 11]. However, this indirect modeling does not prevent *ghost motion*, i.e. body parts passing through objects or the scene. To the best of our knowledge, only [42] has investigated the explicit representation of human-scene interaction for human motion forecasting. [42] proposes a two-stage pipeline, first predicting future joint-scene distances, then using this information to predict the global pose.

Working with scene context requires data that enables the model to infer environmental clues. 3D point clouds provide dense information about surfaces and objects in the scene, which is ideal for human-scene interaction and trajectory forecasting. *STAG* elaborates on the idea of contact maps and adds a component of contextual knowledge through trajectory forecasting.

2.2 Trajectory Forecasting

Trajectory forecasting can be divided into two main categories: *model-based* and *model-free* approaches. Model-based approaches [14, 24, 37, 56] impose physical constraints di-

rectly, while model-free approaches typically rely on implicitly learned physical plausibility [27, 51]. Some recent works [2, 67] have combined the two. Model-free approaches employ a variety of deep learning techniques such as transformers [57, 66], RNNs [25, 43] or GCNs [28, 44]. Deep learning approaches define the SoA on multiple benchmarks [33, 34, 35, 45, 67]. *STAG* follows best practices of SoA methods, adopting GCNs, attention, and trajectory endpoints. Additionally, we are the first to propose a temporal encoding for the *time-to-go*, informing the current model prediction of how long it is missing before the endpoint.

2.3 Human Motion Forecasting

Human motion forecasting can be divided into (local) pose forecasting and global pose forecasting. Local pose forecasting [20, 21, 39, 40, 41, 54, 64] only considers the position of the agent in relation to its root, while global pose forecasting [42, 69, 70] takes the absolute position within the given scene into account. Thus, global pose forecasting can be viewed as combining the trajectory and the (local) pose. Many applications such as human-robot collaboration [9, 29, 52], autonomous driving [18, 47, 71], sports [63, 65], augmented reality [55] or animation [46, 58] require knowledge about the global position of the agent in the scene.

Many human pose forecasting works use 2D image data only [3, 15, 31, 49, 50, 68]. However, in tasks such as industrial human-robot collaboration, where the agent's and objects' exact position in the scene is crucial, 3D data is often used [4, 6, 63]. We consider a 3D point cloud for our task, as they offer a rich scene representation. The agent in the scene can be represented as a graph of body joints or more complex representations such as meshes [19, 22, 38, 48].

While delicate tasks may require a more specific human model, the skeletal representation (adopted in *STAG*) suffices for human motion forecasting [20, 21, 39, 40, 41, 54, 64].

3 Methodology

STAG is designed as a three-stage model, which we overview in Sec. 3.1. The modeling of each stage is detailed in Secs. 3.2-3.4.

3.1 Proposed STAGed contact-aware global motion modelling

Our novel approach for predicting contact-aware global human motion, named *STAG*, is designed in a coarse-to-fine manner by using a three-stage pipeline (see Fig. 2). *STAG* is composed of a first stage that computes the contact points between the 3D scene and the body (cf. Sec 3.2). The second stage uses the information from the previous stage and the past root trajectory to predict its future trajectory (cf. Sec. 3.3). In the third stage, the historical movement of the body, together with the contact points and human trajectory end goals, are used to predict the future global pose upon temporally encoding the time-to-go (cf. Sec 3.4).

Notation. We refer to $\mathbf{S}$ as the scene, $\mathbf{R}$ as the root joint trajectory, and to $\mathbf{M}$ as the global human motion. $\mathbf{S}$ represents a 3D scenes, where $\mathbf{S} \in \mathbb{R}^{N\times 3}$ contains N points, each expressed as a triplet (x,y,z). $\mathbf{R} = [R_0,\dots,R_F]$ is a sequence of root trajectories, where

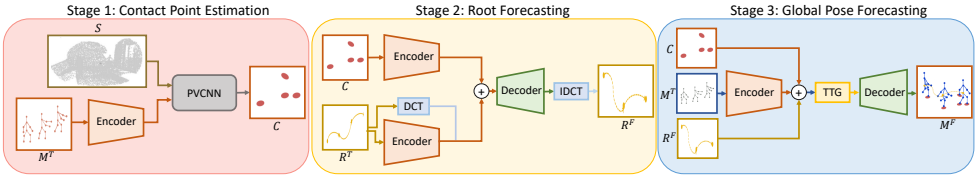

Figure 2: Overview of *STAG*'s three-staged pipeline. Stage 1 takes the scene and the human motion in input and predicts future interactions as contact points. Stage 2 feeds them to a trajectory forecasting model for a coarse prediction, and Stage 3 then refines it to predict future human poses.

$\mathbf{R}^T = [R_0, \ldots, R_{T-1}]$ represents the observed ones and $\mathbf{R}^F = [R_T, \ldots, R_F]$ the ones to predict. $R_i \in \mathbb{R}^3$ represents the 3D root coordinates. Similarly $\mathbf{M} = [M_0, \ldots, M_F]$ is a sequence of global body poses, where $\mathbf{M}^T = [M_0, \ldots, M_{T-1}]$ represents the observed ones and $\mathbf{M}^F = [M_T, \ldots, M_F]$ the ones to predict and $M_i \in \mathbb{M}^{V\times 3}$ represents the pose at timestamp i, consisting of V joints expressed as 3D coordinates.

Staged processing. In the first stage, the goal is to compute the contact points $\mathbf{C}$ defined as $[C_0, \ldots, C_T, \ldots, C_F]$, and $C_i \in \mathbb{R}^{V\times 4}$ [42] consisting of V points expressed as 4D coordinates, triplet (x, y, z) and one value $\{0, 1\}$ to indicate whether it is a contact point or not (cf. Sec 3.2). The second stage predicts the future trajectory $\mathbf{R}^F$ given the historical root coordinates $\mathbf{R}^T$, and the contact points $\mathbf{C}$ (cf. Sec 3.3). In the third stage, the objective is to predict the future body poses $\mathbf{M}^F$ by using $\mathbf{M}^T$, $\mathbf{C}$ and $\mathbf{R}^F$ (cf. Sec 3.4).

3.2 Contact Point Estimation

In the first stage, the goal is to predict the contact points $\mathbf{C}$ between the global human body motion $\mathbf{M}$ and the scene $\mathbf{S}$ (See Fig. 2). We use Point-Voxel CNN (PVCNN) [36] to model the scene $\mathbf{S}$ as a point cloud and encode $\mathbf{M}$ as a spatio-temporal graph [54], to capture the movement's proprieties. Following [42], we first compute the distance matrix $\mathbf{D} \in \mathbb{R}^{TV\times N}$ where each term represents the Euclidean distance between each joint in time TV and the N points in the scene. Since $\mathbf{D}$ is based on distances, it is smooth over time. We adopt a temporal encoding strategy of $\mathbf{D}$ based on the Discrete Cosine Transform (DCT) [39].

To leverage the DCT representation, we reformulate this problem by learning a mapping from the DCT coefficients of the past distance matrix $\mathbf{D}$ to those of the future one $\hat{\mathbf{D}}$. Following [42], we leverage PVCNN [36] to encode the 3D scene $\mathbf{S}$, as well as the encoded motion $\mathbf{M}^T$ and $\mathbf{D}$'s DCT coefficients.

Following leading pose forecasting literature, we use Graph Convolutional Networks [28] to encode the motion. Similarly to previous works [54, 62], we define a *spatial adjacency matrix* as $A_s \in \mathbb{R}^{T\times V\times V}$ to model the connections between joints and a *temporal adjacency matrix* $A_t \in \mathbb{R}^{V\times T\times T}$ to capture the temporal relationships.

$$\bar{\mathbf{M}}^T = \sigma(A_s A_t \mathbf{M}^T W) \tag{1}$$

We aim to obtain a latent vector representing the entire movement sequence and serving as a conditioning variable. To compress spatial and temporal information, we propose using

two separate MLPs, MLP_S, and MLP_T:

$$\tilde{\mathbf{M}}^T = MLP_S(MLP_T(\bar{\mathbf{M}}^T)) \tag{2}$$

Each MLP consists of two linear layers and an equal number of activation functions. From now on, we will refer to this encoding technique as *GCN-MLP*. In summary:

$$\hat{\mathbf{D}} = IDCT(\mathbf{M}^T + f(\mathbf{S}, DCT(\mathbf{D}), \tilde{\mathbf{M}}^T)) \tag{3}$$

Where f represents the trainable point-cloud encoder [36]. Ultimately, we reconvert the distance matrix $\hat{\mathbf{D}}$ to the contact points as in [42] resulting in the predicted contact points $\mathbf{C}$.

3.3 Root Forecasting

In the second stage, we propose to predict the person's trajectory to account for future global motion. We achieve it by predicting the future root joint $\mathbf{R}^F$ from the past $\mathbf{R}^T$. The second stage integrates the scene contacts $\mathbf{C}$, estimated in stage one. $\mathbf{R}^T$ is encoded twice, once by using DCT and secondly by using the encoder described in Sec. 3.2. The formulation is similar, however, $\mathbf{M}^T$ gets changed with $\mathbf{R}^T$ in Eq. (2) and the number of nodes $V = 1$, resulting in $\tilde{\mathbf{R}}^T$. The same encoding technique is used for the contact points, where $\mathbf{M}^T$ gets changed with $\mathbf{C}$ in Eq. (2) and results in $\tilde{\mathbf{C}}$. The latter encodings are concatenated and fed to an MLP, which decodes the feature dimension C and outputs $\hat{\mathbf{R}} \in \mathbb{R}^{T \times V \times C}$.

$$\hat{\mathbf{R}}^T = MLP(DCT(\mathbf{R}^T) \,\|\, \tilde{\mathbf{R}}^T \,\|\, \tilde{\mathbf{C}})), \tag{4}$$

where $\|$ indicates a concatenation operation. Lastly, the *IDCT* reverts the transformation process to trajectories so that $\mathbf{R}^F = IDCT(\hat{\mathbf{R}}^T)$.

3.4 Global Pose Forecasting

For the third stage, we utilize the forecasted root trajectory $\mathbf{R}^F$ and the contact points $\mathbf{C}$ obtained from the preceding stages (see Fig. 2) as inputs. It enables us to predict the future pose and refine the trajectory, ultimately yielding the future global motion of the agent. We also encode the past body motion $\mathbf{M}^T$ as in Sec. 3.2 and concatenate the latter information. The decoding occurs autoregressively, where each future timeframe $\{i\}_{i=T}^F$ of the predicted body motion $\mathbf{M}_i^F$ is computed sequentially. We propose to temporally encode the scene contact points and the trajectory endpoints to raise the model understanding of the time-to-go, i.e. how long before it reaches them. At each i, we also concatenate the root's position $\mathbf{R}_e^F$, and the contact points $\mathbf{C}_e^F$ at the last frame F as end goal conditioning variables. Where respectively, $\mathbf{R}_e^F \in \mathbb{R}^C$ and $\mathbf{C}_e^F \in \mathbb{R}^{V \times C}$.

$$HM_i^F = MLP(\tilde{M}^T \,\|\, R_{i+1}^F \,\|\, \mathbf{R}_e^F \,\|\, \tilde{C}_{i+1} \,\|\, \tilde{\mathbf{C}}_e^F), \tag{5}$$

where HM_i^F is the the embedding at time $\{i\}_{i=T}^F$. Then we add the TE and decode the global body pose.

Time-to-go Temporal Encoding To insert time context, we use a learnable temporal encoder $\mathcal{T}_e$ to encode the time-to-go and add it to HM_i^F. During the autoregressive process, $\mathcal{T}_e$ measures how long is missing before the contact and endpoints are reached. To decode the global motion we use an MLP layer. In summary:

$$M^i = MLP(HM^i + \mathcal{T}_e). \tag{6}$$

4 Experiments

In this section, we detail the dataset and metrics, compare it to the current SoA [42], and perform an extensive ablation on the staged modeling design and its components. Furthermore, we show how our model compares with SoA.

Dataset. The GTA-IM dataset [8] is a large-scale synthetic dataset that captures human-scene interactions, which consists of 50 different characters performing various activities in 7 scenes. We use [42]'s proposed pre-processing, employing 4 of the scenes as our training set, the remaining 3 as the test set, and 21 out of the 98 human joints provided by the dataset. Videos are recorded at 30fps, and we train our models to observe the past 30 frames and predict the future 60. We evaluate *STAG* in all its stages and outperform [42].

[42] also considers PROX [22] but they do not distribute the pre-processed scene-to-pointcloud nor the code for pre-processing. PROX is a real dataset captured using a Kinect-One sensor, and it contains noise at frames (e.g. jittering and corrupted pixels) and in time (missing frames). Upon best efforts, we could not replicate the pre-processed pointcloud, so we could not use it for comparison.

Metrics. The first stage is evaluated by the L_2-norm between our predicted contact point and the ground truth. For the second and third stages, we consider the *Mean Per Joint Positional Error* (MPJPE) across all joints and all the future timeframes [42]. The global movement is called Path Error, and the Pose Error represents the local body movement.

State-of-the-art models and selected baserows. We evaluate *STAG* on the GTA-IM dataset and compare it with the current leading techniques. LTD [39] utilizes a graph convolutional network to encode motion representations in frequencies. DMGNN [32] employs a *dynamic multiscale* GNN for sequence encoding, with a decoder based on GRU. SLT [60] focuses on motion synthesis and employs an autoencoder architecture consisting of a multilayer perceptron as the initial stage, followed by motion generation using LSTM. The top-performing technique is Mao *et al*. [42], which combines MLP and RNN for motion encoding and employs an iterative prediction approach.

4.1 Comparison against SoA

First Stage - Contact points estimation. Table 1 is not present in [42]; thus, we ran their first stage and compared it to ours (Sec. 3.2). We have an overall 9.2% improvement, and it is due to our body movement's encoder, which more accurately extracts the latent representations.

	L_2-norm (mm)				
	0.5s	1s	1.5s	2s	mean
Mao *et al*. [42]	26.2	45.5	67.5	96	47.8
STAG	**24.3**	**41.9**	**61.6**	**86,2**	**43.6**

Table 1: Distance between the predicted contact points and the ground truth ones.

Second Stage - Global pose error. This section focuses on our second stage's impact on the Path Error. As in [42], we experiment with three configurations of our model: (i) no contact point to condition on, (ii) conditioning on the predicted contact points, and (iii) on the ground truth ones. With the original configuration (i), we note a decline in performance, indicating that the second stage necessitates supplementary contextual information for accurate operation. This emphasizes the importance of considering the scene when predicting overall bodily motion. When the predicted contact points (ii) are added, *STAG* has a 16.2% more accurate prediction over the path error. Such improvement increases when considering the GT contact points (iii), reaching a 21% decrease in mean over path error. (iii) also highlights that having precise contact points coming from stage one can significantly improve the overall performance of the second stage.

Third Stage - Local pose error. As in the previous paragraph, we consider: (i) no contact point to condition on, (ii) conditioning on the predicted contact points, and (iii) on the ground truth ones. In this case, we outperform [42] in all settings, reaching 1.8% improvement in (iii) and 4% when considering GT contact points. It demonstrates how our body movement encoder is more capable of creating reasonable latent representations. While the improvement in pose may not be as pronounced as the improvement in path, it is crucial to consider the 3D nature of the scenario and ensure coherent body movements by accounting for the surroundings. With *STAG*, we observe an overall enhancement in both path and pose compared to the SoA methods. The staged pipeline assigns equal importance to both tasks, leading to these improvements.

	Path Error (mm)					Pose Error (mm)				
Models	**0.5s**	**1s**	**1.5s**	**2s**	**mean**	**0.5s**	**1s**	**1.5s**	**2s**	**mean**
LTD [39]	67.0	119.3	207.6	375.6	147.4	67.5	93.8	98.9	103.5	80.5
DMGNN [53]	82.7	158.0	227.8	286.9	156.2	47.5	69.1	85.6	95.3	64.9
SLT* [60]	**45.9**	117.0	186.7	267.1	121.8	70.8	181.4	150.2	196.0	112.6
Mao *et al.* [42] *w/o contact*	61.1	111.7	171.0	249.0	118.8	57.8	74.8	82.4	98.1	68.2
Mao *et al.* [42] *w/ pred contact*	58.0	103.2	154.9	221.7	108.4	50.8	67.5	75.5	**86.9**	61.4
Mao *et al.* [42] *w/ GT contact*	52.4	77.8	95.8	129.5	74.1	49.8	64.8	70.4	78.3	58.2
STAG w/o contact	64.0	133.0	210.4	302.0	141	55.8	72.9	82.8	96.2	67.1
STAG w/ pred contact	55.4	**89,6**	**127.9**	**179.3**	**92.3**	**48.1**	**65.3**	**75.6**	88.2	**60.3**
STAG w/ GT contact	50.3	65.1	70.1	99.2	60.0	46.9	61.5	68.0	76.3	55.6

Table 2: Path and pose error on the output obtained by pipelining the second and third stages on GTA-IM dataset.

4.2 Ablation study

We perform ablative studies to explore our model's components extensively. The results in Table 3 consider GT contact points and refer to the metrics used in Table 2. **stages** indicates the training mode of the second stage module: *2-stage e2e* means that stages two and three are learned in an e2e fashion, as is done in [42]; *2-stage ft.* indicates that stage two is pre-trained and fine-tuned during the training of stage thee; *3-stage (STAG)* is our proposed pipeline. **end** indicates whether the endpoint is used in the third stage. **TTG** is flagged if the proposed *time-to-go* is used in the third stage. Regarding how many joints are used to compute the contact points, we conducted a dedicated ablation study outlined in Table 3. The **cont.** column indicates which joints we consider for contact. With "all", every body part is considered to estimate contact with the scene. With "feet", only the feet can generate

contact points, while with "feet, wrist", we consider contact points involving both the feet and hands.

The performance comparison in rows 1-3 reveals that even without the inclusion of end goals or TTG, the three-stage pipeline surpasses the performance of the two-stage pipeline. When end goals are introduced (rows 4-6), the performance gap becomes more apparent as they contribute to improved global performance. Lastly, using TTG in autoregressive prediction introduces time context and significantly enhances the results (rows 7-9). Moreover, it is preferable to consider the entire skeleton when calculating contact points, as the 3D scene is complex and involves multiple joints. Merely focusing on ground contact points (e.g., feet) or the most probable contact points (e.g., feet and hands) leads to unsatisfactory outcomes, as indicated in rows 10-13.

4.3 Comparison against global motion SoA models

Here we are testing the generalization of *STAG* to predict global motion without a given scene. The task aims to be comparable to other scene-free methods. The original version of *STAG* is evaluated under the assumption of a ground surface beneath the individual. This assumption is implemented by converting the floor into a scene representation as a 3D point cloud. Based on this information, the model estimates future contact points. It is worth mentioning that unlike competing techniques such as [1, 59], our model does not include multi-person joint forecasting or consider social relationships among individuals.

Dataset. We evaluate the performance of our model on additional datasets such as CMU-Mocap [10], which is widely used for absolute pose forecasting. The CMU dataset is captured at a rate of 30 frames per second (fps) using a marker system. Each sequence in the dataset consists of three individuals randomly selected from different scenes and merged together [59].

Comparison with state-of-the-art. Our model is compared to SoA approaches, among which are HRI [40], SocialPool [1], and MR-Trans [59]. HRI utilizes a motion attention mechanism to encode motion in both spatial coordinates and frequencies. SocialPool, on the other hand, is an RNN-based model that employs multiple GRU modules independently for each person in the scene, followed by a social module that considers the features of all individuals in the scene. MR-Trans, currently considered the SoA model, is a transformer-based approach that employs a discriminator to determine the suitability of pose and motion. Lastly, we also adapt Mao *et al*. [42] to the additional dataset *as is*.

					Path Error (mm)					Pose Error (mm)				
	stages	**end.**	**TTG**	**cont.**	**0.5s**	**1s**	**1.5s**	**2s**	**mean**	**0.5s**	**1s**	**1.5s**	**2s**	**mean**
1	2-stage e2e	×	×	all	55.8	77.7	87.5	121.5	71.3	48.8	64.15	70.7	77.9	57.8
2	2-stage ft.	×	×	all	53.6	72.5	83.7	115.8	68.4	48.8	64.2	70.8	77.9	57.8
3	3-stage (*STAG*)	×	×	all	53.4	72.4	84.1	117.8	68.5	48.8	64.2	70.9	78	57.8
4	2-stage e2e	✓	×	all	55.7	79.2	95.2	128.3	75.3	47.1	61.8	68.5	76.7	56
5	2-stage ft.	✓	×	all	51.9	68.7	78.1	113.1	65	47.1	61.8	68.5	76.6	55.9
6	3-stage (*STAG*)	✓	×	all	51.6	68.3	76.8	108.4	64	47.1	61.9	68.4	76.8	56
7	2-stage e2e	✓	✓	all	53.8	74.6	87.2	122.4	70	47.2	61.9	68.5	76.9	56
8	2-stage ft.	✓	✓	all	50.8	66.1	72.6	104.7	61.6	47.1	61.8	68.4	76.6	55.9
9	3-stage (*STAG*)	✓	✓	all	**50.3**	**65.1**	**70.1**	**99.2**	**60**	**46.9**	**61.5**	**68**	**76.3**	**55.6**
10	3-stage (*STAG*)	✓	✓	feet	55.1	79.9	96.7	136.5	76.2	47.5	62.1	68.7	77.3	56.3
11	2-stage e2e	✓	✓	feet, wrist	56.0	81.4	99.0	135.7	77.7	47.1	61.8	68.4	77.1	56
12	2-stage ft.	✓	✓	feet, wrist	55.6	82.0	102.0	140.0	79	47.1	61.4	68.9	79.9	56
13	3-stage (*STAG*)	✓	✓	feet, wrist	56.9	84.4	104.9	143.0	81.0	46.2	61.4	68.3	77.1	55.9

Table 3: Ablation study on the staged modeling.

The proposed approach performs similarly to the current best technique [59] in terms of overall error. However, when predicting future trajectories on the most challenging longer-term horizon, *STAG* slightly underperforms compared to MR-Trans. On pose error, *STAG* outperforms the previous SoA model [59] by 33.5%, and by 8.8% with respect to [42] on the most challenging longer-term horizon.

	Path Error			Pose Error			Global Error		
Models	**1s**	**2s**	**3s**	**1s**	**2s**	**3s**	**1s**	**2s**	**3s**
LTD [59]	0.97	1.73	2.62	0.98	1.21	1.37	1.37	2.19	3.26
HRI [40]	0.96	2.06	3.11	1.05	1.37	1.58	1.49	2.60	3.07
SocialPool [1]	0.96	2.01	2.96	1.03	1.41	1.71	1.15	2.71	3.90
MR-Trans [59]	**0.60**	**1.12**	**1.71**	0.79	1.05	1.22	0.96	**1.57**	**2.18**
Mao *et al.* [42] *w/ pred contact*	0.78	2.19	3.99	0.59	0.93	0.95	1.01	2.47	4.16
STAG w/ pred contact	0.71	1.43	2.02	**0.57**	**0.76**	**0.87**	**0.95**	1.70	2.29

Table 4: Path, pose and global error in meters on CMU-Mocap dataset.

5 Conclusion

This paper has addressed the prediction of global pose in a three-dimensional environment as the staged modelling of three core elements: the scene, the human trajectory, and the pose. *STAG* is the first scene-aware global forecasting model which splits trajectory and pose motion to match the coarse-to-fine nature of the task. In fact, the pose of a person is the result of its motion pathway and the scene, rather than the cause of it.

STAG yields SoA performance on GTA-IM, the sole available for testing scene-aware global forecasting. *STAG* also sets the SoA on the CMU-Mocap dataset, under the assumption that the scene consists solely of a flat ground surface, therefore generalizing the task of global forecasting, which earlier methods have addressed without consideration of the scene.

Acknowledgements This work was supported by the MUR PNRR project FAIR (PE00000013).

References

[1] Vida Adeli, Ehsan Adeli, Ian D. Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5:6033–6040, 2020.

[2] Alessandro Antonucci, Gastone Pietro Rosati Papini, Paolo Bevilacqua, Luigi Palopoli, and Daniele Fontanelli. Efficient prediction of human motion for real-time robotics applications with physics-inspired neural networks. *IEEE Access*, 10:144–157, 2022.

[3] Stefan Becker, Ronny Hug, Wolfgang Hubner, and Michael Arens. Red: A simple but effective baseline predictor for the trajnet benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[5] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020.

[6] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019.

[7] Guangyi Chen, Zhenhao Chen, Shunxing Fan, and Kun Zhang. Unsupervised sampling promoting for stochastic human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[8] Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Scept: Scene-consistent, policy-based trajectory predictions for planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17103–17112, June 2022.

[9] Mi-Young Cho and Youngsook Jeong. Human gesture recognition performance evaluation for service robots. *2017 19th International Conference on Advanced Communication Technology (ICACT)*, pages 847–851, 2017.

[10] CMU Graphics Lab. CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu.

[11] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] Massimiliano Di Luca and Darren Rhodes. Optimal perceived timing: Integrating sensory information with dynamically updated expectations. *Sci. Rep.*, 6:28563, July 2016.

[13] Grace Edwards, Petra Vetter, Fiona McGruer, Lucy S Petro, and Lars Muckli. Predictive feedback to V1 dynamically updates with sensory input. *Sci. Rep.*, 7(1):16538, November 2017.

[14] David Ellis, Eric Sommerlade, and Ian Reid. Modelling pedestrian trajectory patterns with gaussian processes. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1229–1234, 2009. doi: 10.1109/ICCVW.2009.5457470.

[15] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.

[16] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13106–13115, June 2022.

[17] 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.

[18] David Sierra González, Joshué Pérez, Vicente Milanés Montero, and Fawzi Nashashibi. A review of motion planning techniques for automated vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 17:1135–1145, 2016.

[19] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022.

[20] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13053–13064, 2022.

[21] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4809–4819, 2023.

[22] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

[23] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021.

[24] Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, may 1995.

[25] Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37:388–427, jan 2021.

[26] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[27] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

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

[29] 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. doi: 10.1109/IROS.2013.6696634.

[30] Mihee Lee, Samuel S. Sohn, Seonghyeon Moon, Sejong Yoon, Mubbasir Kapadia, and Vladimir Pavlovic. Muse-vae: Multi-scale vae for environment-aware long term trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2221–2230, June 2022.

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

[32] 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.

[33] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Simaug: Learning robust representations from simulation for trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 275–292. Springer, 2020.

[34] 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.

[35] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social nce: Contrastive learning of socially-aware motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15118–15129, 2021.

[36] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *Neural Information Processing Systems (NeurIPS)*, abs/1907.03739, 2019.

[37] Linbo Luo, Suiping Zhou, Wentong Cai, Malcolm Yoke Hean Low, Feng Tian, Yongwei Wang, Xian Xiao, and Dan Chen. Agent-based human behavior modeling for crowd simulation. *Comput. Animat. Virtual Worlds*, 19(3–4):271–281, sep 2008. ISSN 1546-4261.

[38] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.

[39] 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.

[40] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 474–489. Springer, 2020.

[41] 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.

[42] Wei Mao, Richard I Hartley, Mathieu Salzmann, et al. Contact-aware human motion forecasting. *Advances in Neural Information Processing Systems*, 35:7356–7367, 2022.

[43] Xiaoyu Mo, Yang Xing, and Chen Lv. Graph and recurrent neural network-based vehicle trajectory prediction for highway driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1934–1939, 2021. doi: 10.1109/ITSC48978.2021.9564929.

[44] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14424–14432, 2020.

[45] Abduallah Mohamed, Deyao Zhu, Warren Vu, Mohamed Elhoseiny, and Christian Claudel. Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 463–479. Springer, 2022.

[46] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022.

[47] Brian Paden, Michal Cáp, Sze Zheng Yong, Dmitry S. Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles*, 1:33–55, 2016.

[48] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[49] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018.

[50] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017.

[51] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1349–1358, 2019. doi: 10.1109/CVPR.2019.00144.

[52] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 51–69. Springer, 2022.

[53] Luca Scofano, Alessio Sampieri, Giuseppe Re, Matteo Almanza, Alessandro Panconesi, and Fabio Galasso. About latent roles in forecasting players in team sports. *AI4ABM@ICLR Workshop*, 2023.

[54] 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.

[55] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209–1, 2019.

[56] Jur van den Berg, Ming Lin, and Dinesh Manocha. Reciprocal velocity obstacles for real-time multi-agent navigation. In *2008 IEEE International Conference on Robotics and Automation*, pages 1928–1935, 2008. doi: 10.1109/ROBOT.2008.4543489.

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

[58] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018.

[59] Jiashun Wang, Huazhe Xu, Medhini G. Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. In *Neural Information Processing Systems*, 2021.

[60] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9401–9411, June 2021.

[61] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022.

[62] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *European Conference on Computer Vision (ECCV)*, 2022.

[63] Jianru Xue, Jianwu Fang, Tao Li, Bohua Zhang, Pu Zhang, Zhen Ye, and Jian Dou. Blvd: Building a large-scale 5d semantics benchmark for autonomous driving. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6685–6691. IEEE, 2019.

[64] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[65] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9793–9803, 2021.

[66] 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.

[67] Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory prediction via neural social physics. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 376–394. Springer, 2022.

[68] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kummerle, Hendrik Konigshof, Christoph Stiller, Arnaud de La Fortelle, et al. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088*, 2019.

[69] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20481–20491, 2022.

[70] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021.

[71] Yu Zhang, Huiyan Chen, Steven L. Waslander, Jian wei Gong, Guang ming Xiong, Tian Yang, and Kai Liu. Hybrid trajectory planning for autonomous driving in highly constrained environments. *IEEE Access*, 6:32800–32819, 2018.