Human Scene Transformer

Tim Salzmann, Lewis Chiang, Markus Ryll, Dorsa Sadigh, Carolina Parada, and Alex Bewley

***Abstract*—Robot navigation in human centric environments, such as homes or office spaces, remains a challenging task. In such spaces humans do not follow strict rules of motion and there are often multiple occluded entry points such as corners and doors that create opportunity for sudden encounters. In this work, we present a human-centric scene transformer to predict human future trajectories from input features including human positions, and 3D skeletal keypoints from onboard in-the-wild sensory information. The resulting model captures the inherent uncertainty for future human trajectory prediction and achieves state-of-the-art performance on common prediction benchmarks and a human tracking dataset captured from a mobile robot. Furthermore, we identify agents with limited historical data as a major contributor to error and demonstrate that our approach achieves a displacement error reduction of up-to 11% using 3D skeletal poses perceived by a mobile robot.**

I. Introduction

Predicting human trajectories within indoor environments such as offices, homes and care facilities could have a profound impact on service robotics. These environments are narrow with multiple occluded entry-points resulting in close proximity upon first observation. Our goal is to enable more natural, safe, smooth, and predictable navigation by anticipating where humans will be moving in the near future using the robot's onboard sensors.

We present the Human Scene Transformer (HST) which leverages different feature streams: Historic positions of each human, vision-based features such as skeletal keypoints (see Figure 1, joints of the human skeleton) or head orientation when available. We specifically focus on demonstrating the usefulness of noisy in-the-wild human skeletal information from a 3D human pose estimator. While prior Transformer prediction architectures [25] implicitly model interactions between humans at individual timesteps using single-axis attention, we allow for attention between humans at differing time — historic actions can directly influence another humans position at later time — by offering a simple alignment mechanism. As such our contribution is threefold:

(I) To the best of our knowledge, we are the first to demonstrate that detailed human 3D vision-based features improve predictions in a human-centric service robot context notwithstanding imperfect in-the-wild data. Especially, we showcase the benefits of our approach in critical situations such as close proximity between robot and human on early observation.

(II) We present a prediction architecture (HST), which flexibly processes and includes detailed vision-based human features such as skeletal keypoints and head orientation. To target crowded human-centric environments, HST builds upon ideas from trajectory prediction in autonomous driving. We demonstrate HST's capability to consistently model interactions which is critical in human-centric environments.

(III) We evaluate the system's capabilities on a dataset recorded from a service robot's sensors and re-purposed for the prediction task. Simultaneously, we display state-of-the art performance on a common outdoor pedestrian dataset.

Fig. 1: A service robot navigating a busy office space. To do so it anticipates human motion using human-position and visual 3D skeletal keypoints.

II. Related Work

Prior works in *trajectory prediction* commonly target the autonomous driving use-case [32, 28, 25, 24, 41, 5, 15] and rely on GANs [9, 27] or CVAEs [21, 14, 28, 12, 13], this work follows the recent trend towards Transformers [25, 41, 24] as they naturally lend themselves to the set-to-set prediction problems such as multi-agent trajectory prediction and are invariant to a varying number of agents. Another related area is *human pose forecasting* in 3D [4, 40, 42, 22, 29]. However, these approaches commonly consider a single human motion relying on ground truth pose information from a motion capture system, while we target multi-human in-the-wild scenarios.There have been prior efforts to *combine pose estimation with trajectory prediction*, i.e., informing forecasted trajectories by incorporating historic pose information. However, these works are either operating on motion capture datasets which do not exhibit diverse positional movement of the human [16, 4, 18, 30] or are limited to prediction in 2D image space [39, 3, 5]. However, for robotic navigation it is desired to obtain predictions for agents across multiple sensors and in a 3D or bird's-eye metric space. We follow these requirements by solely relying on onboard sensor information of a robotic platform and predict in the metric frame rather than in image space.

III. Human Scene Transformer

To incorporate vision-based human features and achieve state-of-the-art trajectory prediction performance, we present Human Scene Transformer (HST). HST follows the concept of masked sequence to sequence prediction using an architecture with Transformer blocks. This concept has shown promising vehicle prediction results in the autonomous driving domain [25]. HST introduces multiple important ideas extending the general Transformer architecture which makes it suitable for human trajectory prediction. These include the utilization of vision-based human features, a feature attention mechanism to merge multiple, potentially incomplete features, an improved attention mechanism facilitating a more complete information flow, and a self-alignment layer which elegantly solves the problem of discriminating between multiple masked agent timesteps while keeping permutation equivariance.

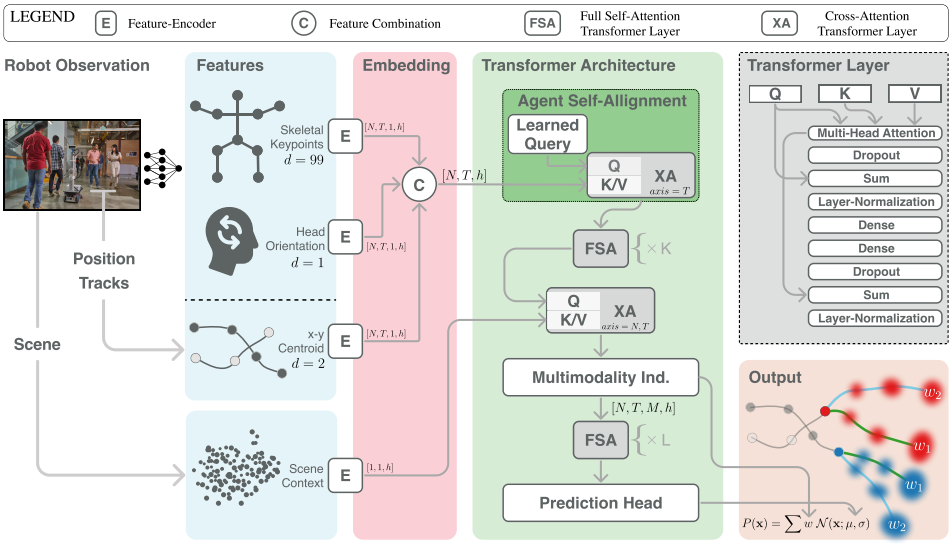


Fig. 2: **Overview of the HST architecture.** From the robot's sensors we extract the scene context, the historic tracks of each agent, and vision based skeletal keypoints/head orientation when feasible. All features are encoded individually before the agent features are combined via cross-attention (XA) using a learned query tensor. The resulting hidden vector passes to our Agent Self-Alignment layer which enables the use of subsequent full self-attention (FSA) layers. Embedded scene context is attended to via cross-attention (XA). After multimodality is induced and further FSA layers the model outputs the parameters of a Normal distribution for each agent at each prediction timestep. We can represent the full output structure as a Gaussian Mixture Model (formula in bottom right) over all possible futures where the mixture coefficients w come from the Multimodality Induction. Both cross-attention (XA) and full self-attention layers use the Transformer layer (top right) with different input configurations.

A. Model Inputs: Incorporating Vision-based Features

We process the robot's observations at each timestep $O(t), \ldots, O(t-H)$ into agent features and scene context (Figure 2 - blue box). Scene context can be an occupancy grid or a LiDAR point cloud at the current timestep, containing information common to nearby agents (e.g. static obstacles). Agent features include the centroid position and vision-based features: skeletal keypoints, and head orientation for each agent. For each detected N nearby humans (equivalent agents) in the scene, we project the 3D bounding box into the 360 degree image using ex- and intrinsic camera calibrations. This results in an associated image patch for all agents. To extract 2D skeletal key points from these patches, one could choose from a plethora of off-the-shelf skeletal keypoints extractor for images [2, 7, 20, 34, 33]. To produce 3D keypoints, we apply the work of Grishchenko *et al.* [8] to estimate 3D keypoints from images using a pre-trained model. As existing datasets commonly only include 2D keypoint annotations, the 3D label required for supervised pre-training is generated by fitting a parametric human shape model to available 2D keypoints solving the following optimization problem:

$$\text{argmin}_{\mathbf{k}} \left(\|r(\mathbf{k}) - \hat{\mathbf{k}}_2\|_2 + \lambda p(\mathbf{k}) \right), \tag{1}$$

where $\mathbf{k}$ are the 3D skeletal keypoints, $\hat{\mathbf{k}}_2$ is the 2D keypoints label, $r : \mathbb{R}^{33\times 3} \rightarrow \mathbb{R}^{33\times 2}$ is the re-projection function of 3D points into the 2D image space using the camera calibrations. To capture both head information and limb articulation, we choose a 33 keypoints skeleton representation [8, 38]. The learned prior distribution over human pose configuration $p(\mathbf{k})$ penalizes infeasible poses which can arise in optimization for the underdetermined 3D-2D-projection problem.

B. Model Architecture

Transformer Layer. The primary building block of the model's architecture is the Transformer layer (Figure 2 - top right), which itself is comprised of a Multi-Head Attention layer [35] and multiple dense and normalization layers. For a comprehensive explanation on the Attention mechanism and its inputs we refer the reader to Vaswani *et al.* [35]. We define a self-attention (SA) operation as a Transformer layer where inputs Query (Q), Key (K), and Value (V) are the same tensor: The tensor attends to itself and conveys it's information along one or more dimensions. Similarly, we define cross-attention (XA) as a Transformer layer where the Q input is distinct from the K/V inputs. Intuitively the query attends to additional information from a different tensor as means of merging multiple streams of information.

Input Embedding. The input agent features (blue) are tensors of shape $[N, T, d]$, where $d = 2$ for the x-y centroid position, $d = 99$ for the x-y-z position of 33 skeletal keypoints, and $d = 1$ for the head orientation. We mask all future as well as unobserved agent timesteps by setting their feature value to 0, making only available historical and current information accessible to the model. This masking approach is a well known technique in missing-data problems such as future prediction using Transformer based architectures [35, 25, 41]. Masking exploits the inductive bias inherent in the prediction problem, which allows for the filling of missing information using available context in vicinity of the gaps. As such, our approach allows for missing keypoints in frames due to bad lighting or other influences as the Transformer effectively "fills" in for the missing information. The agent features are encoded independently and are combined by a learned attention query. This masked attention mechanism offers scalability to systems with large number of features with limited availability.

Full Self-Attention Via Agent Self-Alignment. Previous methods [25] rely on factorized attention, where information is alternately propagated along the time and along the agent dimension. In social interactions, however, a change in action such as adjustment in walking direction does not have an immediate influence on other humans in proximity but rather influences their future. Following this illustration, an agent's latent representation at a given timestep in our Transformer architecture should be able to attend not just to other agents at the current timestep (factorized attention) but to *all* agents at *all* timesteps. This operation, which we name *full self-attention* (FSA), can propagate the same information flow across both agents and time with a single operation leading to improved performance and a smaller model.

After embedding, all future timesteps of all agents are masked out to not hold any information. Naïvely applying full self-attention results in two agents that inevitably have

the same masked future timesteps to also have the same input (Query) representation to the Transformer layer (Figure 2 - top right). This results in the same attention to historic input information across all agents. Intuitively, using this naïve approach, when filling the masked future timesteps in a full self-attention step, the model can not associate future timesteps of an agent with its history as *all* future agents' timesteps "look" the same (masked). The problem could be addressed by enforcing an innate order on perceived agents, where all agents are enumerated. This, however, would eliminate the permutation invariant set-to-set prediction capabilities; one of the core strengths of Transformers.

Instead, we achieve full self-attention via a simple approach that we refer to as *agent self-alignment* mechanism (dark green box in Figure 2). After the agent embeddings are combined, we cross-attend with a learned query tensor *only* in the time dimension. This query, a weight matrix jointly optimized with all other network weights during training, learns to propagate available historic information for each agent to future timesteps, enabling the model to align future masked timesteps of an agent with historic ones during full self-attention without an explicit enumeration embedding. This agent self-alignment mechanism preserves agents' permutation invariance and enables full self-attention without restricting information flow along matching timesteps [25] or utilizing special attention matrices which explicitly separates agents [41]. The output tensors of the agent self-alignment then passes through K Transformer layers with full self-attention across agents and time before cross-attending to the encoded scene features.

Multimodality Induction. Our architecture can predict multiple consistent futures (modes) for a scene. To do so, the Multimodality Induction module repeats the hidden vectors by the number of future modes (M), resulting in a tensor of shape $[N, T, M, h]$. To discriminate between modes it is combined with a learned *mode-identifier* tensor of shape $[1, 1, M, h]$. Each future's logit probability $w_m;\ m \in 1, \ldots, M$ is inferred by having the *mode-identifier* attend to the repeated input.

Prediction Head. The hidden vectors updated with the learned mode-identifier go through L Transformer layers, again with full self-attention, before predicting per mode parameters μ, σ using a dense layer as *prediction head*.

C. Producing Multimodal Trajectory Distributions

Combining μ and σ with the mode likelihoods w_m from the multimodality induction, the distribution of the i-th agent's position at each timestep t is modeled as a Gaussian Mixture Model (GMM):

$$P_\theta^i(\mathbf{x}_t|O(t), ..., O(t-H)) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{x}; \sigma_{m,i,t}, \mu_{m,i,t}), \quad (2)$$

where m is the m-th future mode.

We adopt a joint future loss function, that is, the cumulative negative log-likelihood of the Gaussian mode (m^*) with the smallest mean negative log-likelihood:

$$\mathcal{L}_{\text{minNLL}} = \sum_{i,t} -\log(\mathcal{N}(\mathbf{x}_{i,t}^*; \sigma_{m^*,i,t}, \mu_{m^*,i,t})), \quad (3)$$

where $\mathbf{x}_{i,t}^*$ is the ground truth agent position.

The resulting prediction represents M possible realizations of all agents at once in a consistent manner, where the mode mixture weights w are shared by all agents in the scene.

IV. Experiments

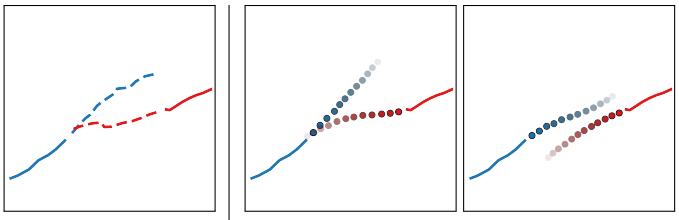
We structure our experiments to support our contributions: First, we will highlight a gap in prior work by showing the limitations of existing datasets for human trajectory prediction in indoor navigation and propose an adaptation of existing datasets. Further, we qualitatively and quantitatively demonstrate that our architecture provides accurate predictions for the human-centric service robot domain, where HST can leverage and model interactions between humans consistently over multiple possible futures. We especially demonstrate how HST can leverage vision-based features in human-centric environments to improve prediction accuracy, specifically in short history situations where prediction errors are high. Finally, we demonstrate that our approach is cross-domain compatible with unconstrained outdoor pedestrian prediction.

Datasets. Many of existing datasets are collected from a single top-down camera in a limited number of environments, such as the ETH [26] and UCY [17] pedestrian datasets. Others are specific to the autonomous driving domain [6, 1, 37, 10], mostly focusing on predicting vehicles. While none of these datasets provide labels for skeleton keypoints, other datasets [11, 19, 36] which are collected using a motion capture system or wearable IMU devices, do offer such labels. However, these datasets are limited to artificial environments and often feature stationary or scripted motions.

One dataset which is recorded in diverse human-centric environments using sensors on a mobile robotic platform is the JackRabbot Dataset and Benchmark (JRDB) [23]. However, JRDB was created as a detection and tracking dataset rather then a prediction dataset. To make the data suitable for a prediction task, we first extract the robot motion from the raw sensor data to account for the robot's movement over time. Tracks are generated for both train and test split using the JRMOT [31] detector and tracker. The ground truth labeled bounding-boxes on the train set were disregarded as they were exposed to filtering during the labeling process to the point where the smoothness eases the prediction task. We were able to increase the number human tracks for training by associating the JRMOT detections to ground truth track labels via Hungarian matching, while on the test split we solely use JRMOT predictions. Due to factors such as distance, lighting and occlusion the pre-trained 3D pose estimator model (Section III) is not guaranteed to produce keypoints for all agents at all timesteps. We observed human keypoints information in $\sim 50\%$ of all timesteps for all agents.

In addition, we also compare our model to the ETH [26] and UCY [17] datasets. These are standard benchmarks for pedestrian trajectory prediction and enable a fair comparison of our architecture against other methods.

Trajectory Prediction in Human-centric Environments. In Table I and Figure 3 we show quantitative and qualitative results of HST's predictions in the human-centric environment. We show that in crowded human-centric environments the influence of interaction between humans has large benefits on the prediction accuracy of each individual. To show this, we compare against a version of our model which is trained to predict a single human at a time ignoring interactions with other agents. Subsequently, adding our full self-attention via self-alignment mechanism additionally increases the model's



(a) (b)

Fig. 3: **Consistently modeled interactions in different predicted futures for a single scene in the x-y-plane [m].** Two humans approaching each other head on. (a) History (solid) and ground truth future (dashed - increasing transparency with time) of both humans. (b) Two of the M predicted futures (dots) of the scene by HST. Within each mode the influence and reaction of both agents is consistent and reasonable. The humans' futures are predicted without collisions giving each other space to navigate within the specific predicted future mode of the scene.

TABLE I: **Comparison against Scene Transformer on JRDB prediction dataset.** HST outperforms the original Scene Transformer on all metrics.

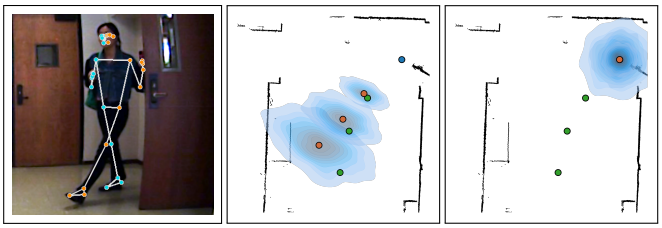
Model Configuration			*minADE*	*MLADE*	*NLL*
Scene Transformer [25]			0.53	0.86	0.25
	Full Self-Attention	Interaction Attention			
HST	✗	✗	0.57	0.93	0.89
HST	✗	✓	0.50	0.84	−0.02
HST	✓	✓	**0.48**	**0.80**	**−0.13**

ability to capture interactions across time, leading to improvements across all metrics. The capability to *consistently* account for interactions between humans is qualitatively demonstrated in Figure 3 where we show multiple predicted futures for a scene of interacting humans.

Vision-based Features. We consider the adversarial setting, where the robot encounters a human unexpectedly, i.e., the robot observes a new human with little historical observations. We note that prediction architectures solely relying on historic position information struggle in scenarios where no or only a limited amount of history of the human position is available to the model. Specifically, at the first instance of human detection, experimentally the error is up to 200% higher compared to full historic information over 2 s. Given the specifics of our targeted human-centric environment, where we are mostly interested in humans close to the robot, we are likely able to extract vision-based features for the human in addition to the position. Specifically, we target the research question: *"Can information from human visual features lead to improved prediction accuracy?"*

Before answering this question quantitatively we show a clarifying visual example in Figure 4 where a human just entered the scene through a door and is first detected. When solely relying on historic position information the most likely prediction by the model is stationary. However, when we employ the pre-trained skeleton keypoints estimator to provide pose keypoints as additional input to our model it correctly recognizes the human's walking motion and how the human is oriented, accurately predicting the most likely future trajectory.

Quantitatively, during evaluation, when keypoints are available on the first detection we observe a substantial prediction improvement of up to 11% (Figure 5). When additional timesteps with position information are available the improvement using keypoints vs not using keypoints averages between 5% and 10%. The relative improvement generally increases with the number of timesteps with keypoints in the history and decreases with the number of historic position information.



(a) First detection of person entering the scene. (b) Prediction with keypoints. (c) Prediction without keypoints.

Fig. 4: A visualization of the predicted trajectory distributions for a new human agent entering the scene through the door on the right as viewed in (a). For *both* (b) and (c) the HST model does not have any historic information here and only has access to the current frame. The plot of future trajectory distributions in (b) and (c) show the effect of using and not using skeletal keypoints (respectively) as input in that single frame. Without pose keypoints the HST model predicts the agent to be most-likely stationary while, with keypoints as input, it can reason that the human is moving and correctly anticipates the direction. Blue dot is the detected human at the initial frame, orange dots are most likely mode predictions with corresponding distribution shown in blue shading, green dots are the ground truth human future.

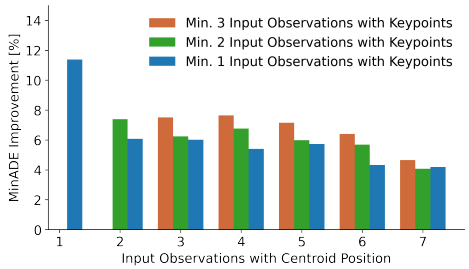


Fig. 5: Impact of vision-based features conditioned on different number of consecutive non occluded input timesteps.

Pedestrian Dataset. We further validate our architecture against a range of state-of-the-art prediction methods on a dataset which has been used by the community for several years: On the ETH/UCY dataset (Table II), we either improve current state-of-the-art methods or are on par with them on 4 out of the 5 scenes leading to the best overall average.

V. Conclusion

While concepts originally designed for trajectory prediction in autonomous driving are generally transferable to the domain of human-centric service robot environments, they suffer in challenging settings where the history of a human is limited. Specifically in these situations we demonstrate how the HST can leverage vision-based features to improve prediction accuracy. Beyond scenarios such as when robot and human encounter each other in blind corners, general improvement trends using in-the-wild skeletal pose detections were also observed with more observations. Our architecture finds state-of-the-art prediction results on a common pedestrian prediction dataset and improves upon existing autonomous driving prediction models in the domain of human-centric service robot environments.

TABLE II: Overall results on ETH and UCY datasets.

Method	$minADE_{20}$ / $minFDE_{20}$
SoPhie [27]	0.54 / 1.15
Trajectron++ [28][1]	0.32 / 0.55
AgentFormer [41][1]	0.23 / 0.39
Scene Transformer [25]	0.31 / 0.40
HST	**0.22** / **0.38**

References

[1] 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.

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[3] Kai Chen, Xiao Song, and Xiaoxiang Ren. Pedestrian Trajectory Prediction in Heterogeneous Traffic Using Pose Keypoints-Based Convolutional Encoder-Decoder Network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1764–1775, 2020.

[4] 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*, pages 6992–7001, 2020.

[5] Phillip Czech, Markus Braun, Ulrich Kreßel, and Bin Yang. On-Board Pedestrian Trajectory Prediction Using Behavioral Features. *arXiv preprint arXiv:2210.11999*, 2022.

[6] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aur'elien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large Scale Interactive Motion Forecasting for Autonomous Driving: The Waymo Open Motion Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9710–9719, October 2021.

[7] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[8] Ivan Grishchenko, Valentin Bazarevsky, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, Richard Yee, Karthik Raveendran, Matsvei Zhdanovich, Matthias Grundmann, and Cristian Sminchisescu. BlazePose GHUM Holistic: Real-time 3D Human Landmarks and Pose Estimation. *Sixth Workshop on Computer Vision for AR/VR*, 2022.

[9] 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.

[10] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *Conference on Robot Learning*, pages 409–418. PMLR, 2021.

[11] 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, jul 2014.

[12] 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*, pages 2375–2384, 2019.

[13] Boris Ivanovic, Edward Schmerling, Karen Leung, and Marco Pavone. Generative modeling of multimodal multi-human behavior. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3088–3095. IEEE, 2018.

[14] Boris Ivanovic, Karen Leung, Edward Schmerling, and Marco Pavone. Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach. *IEEE Robotics and Automation Letters*, 6(2):295–302, 2020.

[15] Julian FP Kooij, Fabian Flohr, Ewoud AI Pool, and Dariu M Gavrila. Context-based path prediction for targets with switching dynamics. *International Journal of Computer Vision*, 127(3):239–262, 2019.

[16] Markus Kuderer, Henrik Kretzschmar, Christoph Sprunk, and Wolfram Burgard. Feature-based prediction of trajectories for socially compliant navigation. In *Robotics: science and systems*, 2012.

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

[18] Mohammad Mahdavian, Payam Nikdel, Mahdi TaherAhmadi, and Mo Chen. STPOTR: Simultaneous Human Trajectory and Pose Prediction Using a Non-Autoregressive Transformer for Robot Following Ahead. *arXiv preprint arXiv:2209.07600*, 2022.

[19] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*, pages 5442–5451, October 2019.

[20] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2646, 2022.

[21] 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.

[22] 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.

[23] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[24] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. *arXiv preprint arXiv:2207.05844*, 2022.

[25] Jiquan Ngiam, Vijay Vasudevan, Benjamin Caine, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, David J. Weiss, Ben Sapp, Zhifeng Chen, and Jonathon Shlens. Scene Transformer: A unified architecture for predicting future trajectories of multiple agents. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=Wm3EA5OlHsG.

[26] 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.

[27] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1349–1358, 2019.

[28] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pages 683–700. Springer, 2020.

[29] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal Probabilistic Human Motion Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6457–6466, 2022.

[30] Tim Schreiter, Tiago Rodrigues de Almeida, Yufei Zhu, Eduardo Gutierrez Maestro, Lucas Morillo-Mendez, Andrey Rudenko, Tomasz P Kucner, Oscar Martinez Mozos, Martin Magnusson, Luigi Palmieri, et al. The Magni Human Motion Dataset: Accurate, Complex, Multi-Modal, Natural, Semantically-Rich and Contextualized. *arXiv preprint arXiv:2208.14925*, 2022.

[31] Abhijeet Shenoi, Mihir Patel, JunYoung Gwak, Patrick Goebel, Amir Sadeghian, Hamid Rezatofighi, Roberto Martin-Martin, and Silvio Savarese. Jrmot: A real-time 3d multi-object tracker and a new large-scale dataset. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10335–10342. IEEE, 2020.

[32] Pei Sun, Weiyue Wang, Yuning Chai, Gamaleldin Elsayed, Alex Bewley, Xiao Zhang, Cristian Sminchisescu, and Dragomir Anguelov. Rsn: Range sparse net for efficient, accurate lidar 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5725–5734, 2021.

[33] tensorflow.org. Real-time human pose estimation in the browser with tensorflow.js, 2018. URL https://blog.tensorflow.org/2018/05/real-time-human-pose-estimation-in.html.

[34] tensorflow.org. MoveNet: Ultra fast and accurate pose detection model., 2022. URL https://www.tensorflow.org/hub/tutorials/movenet.

[35] 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.

[36] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

[37] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hart-

nett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next Generation Datasets for Self-driving Perception and Forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.

[38] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020.

[39] Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. Future person localization in first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7593–7602, 2018.

[40] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020.

[41] 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.

[42] 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.