Human Scene Transformer

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

Abstract

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 robot 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 where our approach leverages multi-modal data to provide a error reduction of up-to 11%. For links to an extended paper, data and code: human-scene-transformer.github.io

1 Introduction

We present the Human Scene Transformer (HST) which leverages different feature streams: Historic positions of each human, vision-based features such as skeletal keypoints 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 Ngiam et al. [2022] 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. (II) We present a prediction architecture (HST), which flexibly processes and includes detailed vision-based human features such as skeletal keypoints and head orientation. (III) We evaluate the system's capabilities on a dataset recorded from a service robot's sensors and re-purposed for the prediction task.

2 Related Work

Prior works in *trajectory prediction* commonly target the autonomous driving use-case Sun et al. [2021], Salzmann et al. [2020], Ngiam et al. [2022], Nayakanti et al. [2022], Yuan et al. [2021], Czech et al. [2022], Kooij et al. [2019] and rely on GANs Gupta et al. [2018], Sadeghian et al. [2019] or CVAEs Mangalam et al. [2020], Ivanovic et al. [2020], Salzmann et al. [2020], Ivanovic and Pavone [2019], Ivanovic et al. [2018], this work follows the recent trend towards Transformers Ngiam et al. [2022], Yuan et al. [2021], Nayakanti et al. [2022] 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 Corona et al. [2020], Yuan and Kitani [2020], Zhang et al. [2021], Mao et al. [2020], Salzmann et al. [2022]. 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.Prior works *combine pose estimation with trajectory prediction*, but operate on motion capture datasets which do not exhibit diverse positional movement of the human Kuderer et al. [2012], Corona et al. [2020], Mahdavian et al. [2022], Schreiter et al. [2022] or are limited to prediction in 2D image space Yagi et al. [2018], Chen et al. [2020], Czech et al. [2022]. For robotic navigation, we instead solely rely on onboard sensor information of a robotic platform and predict in the metric frame rather than in image space.

NeurIPS 2023 Workshop on Robot Learning: Pretraining, Fine-Tuning, and Generalization with Large Scale Models, New Orleans, United States of America

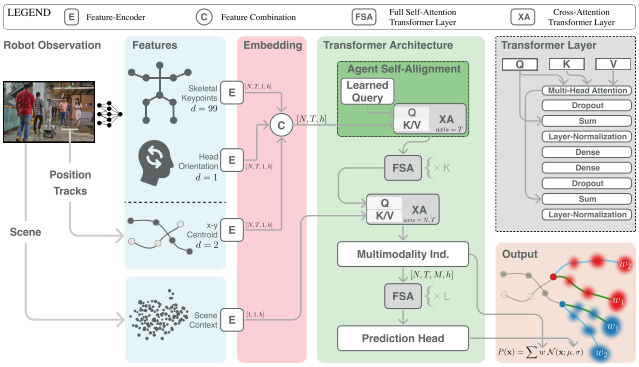


Figure 1: **HST architecture.** From the robot's sensors we extract the scene context, the agent tracks, and skeletal keypoints/head orientation when feasible. All features are encoded individually before the agent features are combined via cross-attention (XA) using a learned query tensor. The resulting 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 over all possible futures.

3 Human Scene Transformer

3.1 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 1 - 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 produce 3D keypoints, we apply the work of Grishchenko *et al*. Grishchenko et al. [2022] to estimate 3D keypoints from images using a pre-trained model.

3.2 Model Architecture

Transformer Layer. The primary building block of the model's architecture is the Transformer layer (Figure 1 - top right), which itself is comprised of a Multi-Head Attention layer Vaswani et al. [2017] and multiple dense and normalization layers. For a comprehensive explanation on the self-attention (SA) and cross-attention (XA) mechanisms and their inputs we refer the reader to Vaswani *et al*. Vaswani et al. [2017].

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, common technique in missing-data problems Vaswani et al. [2017], Ngiam et al. [2022], Yuan et al. [2021]. 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 Ngiam et al. [2022] 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.

Table 1: **Comparison against Scene Transformer on JRDB prediction dataset.** HST outperforms the original Scene Transformer on all metrics.

Model Configuration			*minADE*	*MLADE*	*NLL*
Scene Transformer Ngiam et al. [2022]			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**

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*.

3.3 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):

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

where m is the m-th future mode.

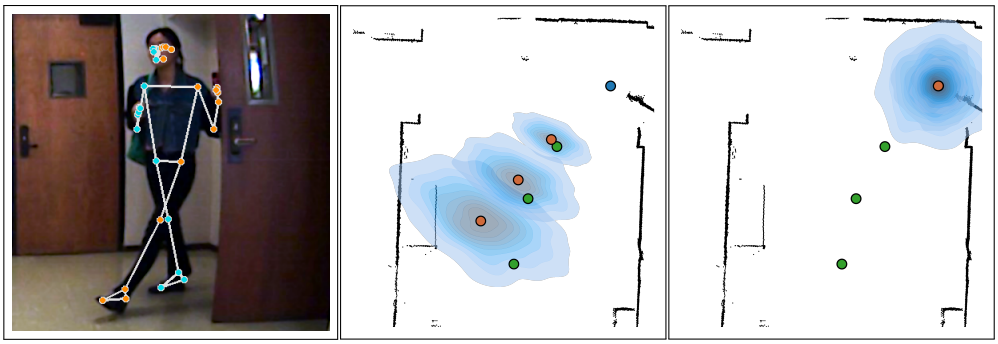
We adopt a joint future loss function, that is, the cumulative negative log-likelihood of the Gaussian mode (m^*) with the smallest mean negative log-likelihood using the ground-truth agent position. The resulting prediction represents M possible realizations of all agents at once in a consistent manner, where the mode mixture weights w are shared by all agents in the scene.

4 Experiments

Datasets. A dataset which is recorded in diverse human-centric environments using sensors on a mobile robotic platform is the JackRabbot Dataset and Benchmark (JRDB) Martin-Martin et al. [2021]. 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 Shenoi et al. [2020] 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 3) 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.

Trajectory Prediction in Human-centric Environments. In Table 1 we show quantitative 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 ability to capture interactions across time, leading to improvements across all metrics.

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,



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

Figure 2: 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.

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 2 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% . 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.

5 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.

References

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.

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.

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

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.