GigaTraj: Predicting Long-term Trajectories of Hundreds of Pedestrians in Gigapixel Complex Scenes

Haozhe Lin[1]*, Chunyu Wei[1]*, Yuchen Guo[1]*, Li He[1]*, Yunqi Zhao[1], Shanglong Li[1], Lu Fang[1]✉
[1]Tsinghua University
*These authors contributed equally to this work. ✉Corresponding author: fanglu@tsinghua.edu.cn

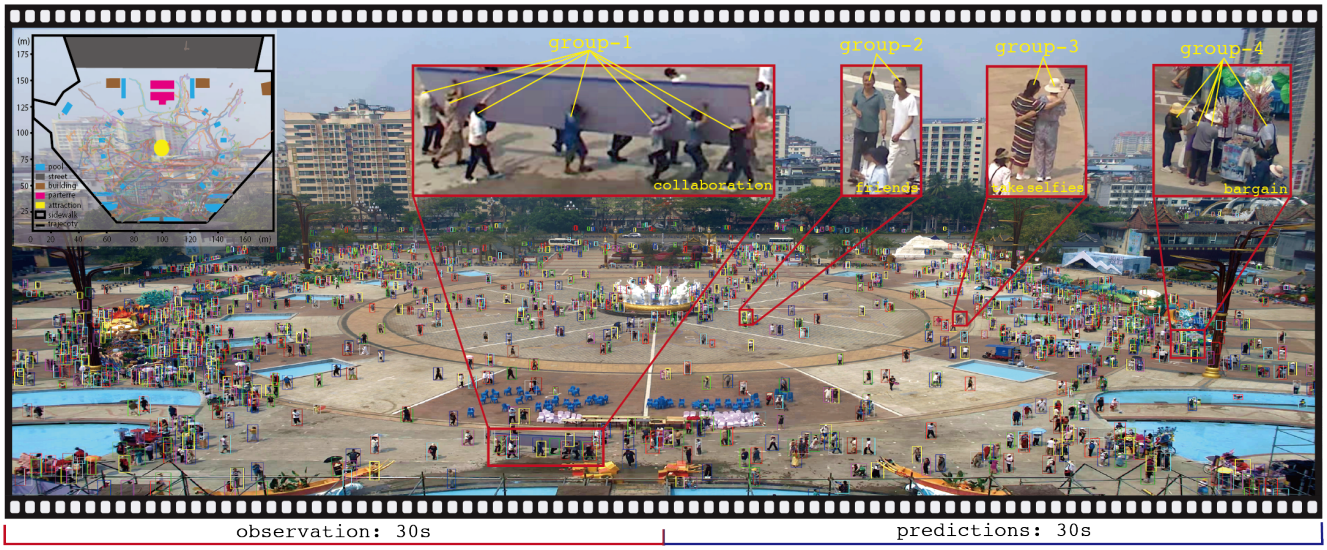


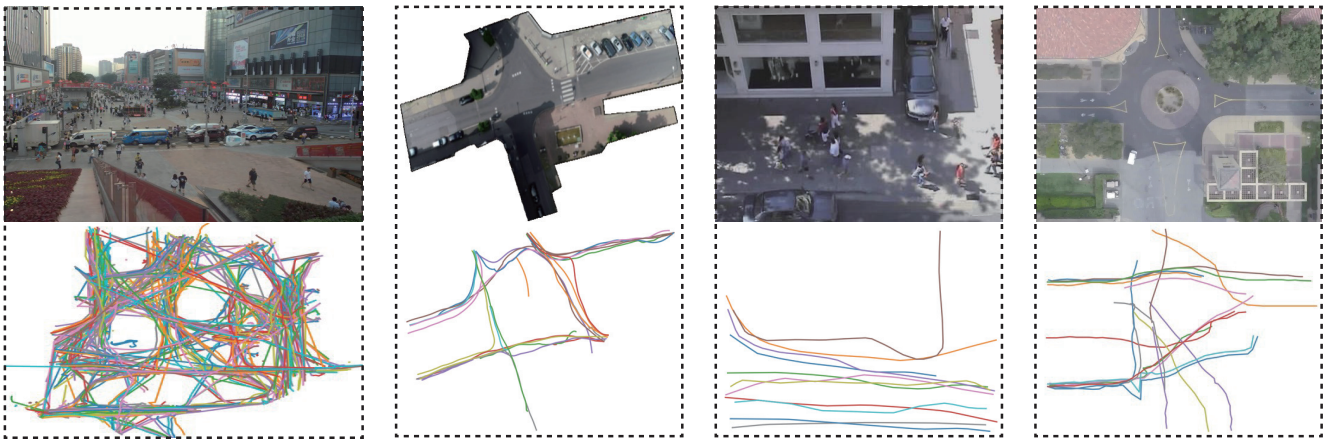
Figure 1. A representative scene *Water Sprinkling Festival* from GigaTraj dataset. Hundreds of pedestrians interacted within a $\sim 4\times10^4\ m^2$ area, observed through a gigapixel-level camera array and annotated with minute-level long-term trajectories. Bounding boxes, IDs, world coordinates, group and interaction relationships, and scene semantics are annotated for predicting the minute-level long-term trajectory of hundreds of pedestrians in gigapixel complex scenes.

Abstract

Pedestrian trajectory prediction is a well-established task with significant recent advancements. However, existing datasets are unable to fulfill the demand for studying minute-level long-term trajectory prediction, mainly due to the lack of high-resolution trajectory observation in the wide field of view (FoV). To bridge this gap, we introduce a novel dataset named GigaTraj, featuring videos capturing a wide FoV with $\sim 4 \times 10^4\ m^2$ and high-resolution imagery at the gigapixel level. Furthermore, GigaTraj includes comprehensive annotations such as bounding boxes, identity associations, world coordinates, group/interaction relationships, and scene semantics. Leveraging these multimodal annotations, we evaluate and validate the state-of-the-art approaches for minute-level long-term trajectory prediction in large-scale scenes. Extensive experiments and analyses have revealed that long-term prediction for pedestrian trajectories presents numerous challenges, indicating a vital new direction for trajectory research. The dataset is available at www.gigavision.ai.

1. Introduction

Pedestrian trajectory prediction is a crucial problem that has significant implications for many industries including unmanned system planning, smart city, human behavior understanding, and service systems. Currently, there are numerous datasets [1–7] available for trajectory prediction research, and previous work [8–13] has achieved promising results. In such context, attention has shifted towards long-term trajectory prediction [14–17]. However, current datasets may not be well-suited to facilitate this challenge. On the one hand, datasets with narrow field of view (FoV), such as ETH [1]/UCY [2], cannot observe long-term trajectories. On the other hand, datasets with a wide FoV usually lack high-resolution details, such as SDD [3] and InD [4], which are uninformative to support long-term trajectory prediction. Therefore, there is a pressing need for a new dataset that provides videos with both wide FoV and high-resolution, and abundant annotations to advance long-term trajectory prediction research.



(a) GigaTraj (b) InD (c) ETH/UCY (d) SDD

Figure 2. Dataset comparison. GigaTraj offers gigapixel videos in large-scale complex scenes, showcasing numerous pedestrians engaging in interactions. In contrast, existing widely used trajectory datasets suffer from relatively small-scale scenes or relatively low resolution, where the trajectories may be easier to predict.

Dataset	covered area	resolution	#traj.	ATD	ABBS	APCF	#scene
ETH [1]	$\sim 9m \times 7m$	640×480	749	$7.4s$	✗	13	2
UCY [2]	$\sim 13m \times 10m$	720×576	909	$10.2s$	✗	46	3
SDD [3]	$\sim 52m \times 33m$	1400×1904	11,240	$14.8s$	16×5	40	8
InD [4]	$\sim 93m \times 52m$	1170×780	11,500	$46.4s$	✗	9	4
GigaTraj	$\sim 200m \times 200m$	23972×13484	15,520	$57.3s$	196×453	313	14

Table 1. Comparison of GigaTraj with other datasets. The following numbers are from the description from the original paper and statistics from the datasets. '#' represents 'the number of'; 'ATD' represents 'average trajectory duration'; 'ABBS' represents 'average bounding box size'; 'APCF' represents 'average pedestrian count per frame'.

In recent years, significant progress has been made in the field of imaging. By leveraging array cameras, capturing gigapixel-level outdoor scenes has become effortless, signifying a burgeoning trend for the future [19**?** , 20]. We have noticed that the PANDA dataset [21] is derived from such imaging devices, offering videos featuring wide FoV and gigapixel-level high resolution. This characteristic holds potential for long-term trajectory prediction research. However, the state-of-the-art research [22] can only make short-term predictions based on non-visual inputs based on PANDA. The aforementioned circumstances motivate us to enhance the annotation of the original PANDA dataset specifically tailored for long-term trajectory prediction purposes. By analyzing the existing PANDA dataset, we have identified three primary weaknesses that require attention. Firstly, there is a dearth of homography matrices necessary to acquire world coordinates corresponding to the original videos. Secondly, there is an absence of scene semantics annotation. Lastly, there is a pressing need for an expanded collection of videos and annotations to facilitate a more appropriate division of training and testing data.

To overcome these limitations, we have developed a new dataset called GigaTraj. This dataset addresses the shortcomings by taking the following measures. We have collected 6 additional complex scenes at the gigapixel level, one of which is the captivating *Water Sprinkling Festival* scene showcased in Figure 1. These scenes have been divided into 8 training sets and 8 testing sets. Notably, the testing sets consist of both seen and unseen scenes, allowing for the subtle evaluation of long-term trajectory prediction performance. We have also utilized laser scanners to reconstruct the scenes according to original videos. This process has enabled us to estimate homography matrices and obtain world coordinates of pedestrians and scenes with centimeter-level precision. Additionally, we have supplemented the dataset with an increased quantity and dimension of annotations. Overall, the GigaTraj dataset comprises 16 expansive scenes, covering a $\sim 4 \times 10^4\ m^2$ area with high-resolution imagery at the gigapixel level. It includes a total of 15,520 trajectories. Furthermore, the dataset offers an abundance of annotations, including bounding boxes, IDs, pedestrian interactions, and semantic information.

Leveraging the comprehensive multimodal annotations in the GigaTraj dataset, we have implemented and refined state-of-the-art methods for trajectory prediction, evaluating their performance in long-term trajectory prediction within large-scale scenes. The experimental results indicate that existing methods are inadequate for addressing the minute-level long-term trajectory prediction challenges presented by GigaTraj, and incorporating multimodal annotations presents a non-trivial task. Building on these findings, we have analyzed several potential research directions based on GigaTraj for future exploration. Our main contributions can be summarized as follows:

- We have constructed the GigaTraj dataset to facilitate the predictions of minute-level long-term trajectories in complex scenes. GigaTraj contains 16 videos with $\sim 4\times10^4\ m^2$ wide FoV and gigapixel-level high-resolution. Additionally, it includes detailed annotations such as bounding boxes, long-term ID associations, complex group and interaction information, world coordinates of pedestrians and important scenes, as well as scene semantics.
- We have conducted an empirical study to evaluate the performance of existing trajectory prediction models. Through extensive experiments, we have found that the state-of-the-art methods are inadequate for predicting minute-level long-term trajectories in complex scenes.
- We have identified several critical factors that are essential for training successful long-term trajectory prediction models in large-scale scenes. Additionally, we have outlined promising research directions that can be explored further by leveraging the GigaTraj dataset.

2. Related work

Trajectory prediction datasets. Trajectory prediction is a crucial task in computer vision, and many datasets [1–7] have been developed for this purpose. For example, based on the ETH [1]/UCY [2] dataset, Yue el al. proposed the NSP-SFM [23] model achieving remarkable Average Displacement Error (ADE) and Final Displacement Error (FDE) for predicting 12-frame trajectories with an 8-frame input in, which closely match the ground truth. Consequently, there has been a shift in focus towards studying long-term trajectory prediction tasks [14–17]. However, to the best of our knowledge, there is currently no dataset available specifically designed to support minute-level trajectory prediction research. By analyzing the widely used trajectory prediction datasets as shown in Table 1, we realized that the reason behind this is that the videos in these datasets cannot simultaneously provide a wide FoV and high-resolution imaging, making it challenging to record long trajectories or observe detailed pedestrian behavior. For example, the ETH [1]/UCY [2] dataset has a relatively small FoV (shown in Figure 2), resulting in short average trajectory lengths of only a few seconds. On the other hand, datasets like SDD [3] and InD [4] offer a wider FoV, but cannot observe detailed pedestrian actions and group interactions due to the use of bird-eye imaging. We notice that the PANDA [21] offers a wide FoV and high-resolution details, making it highly promising for long-term trajectory prediction research. However, the PANDA dataset suffers from a lack of essential homography matrices and ground semantics annotations. Furthermore, there is an urgent need to increase the quantity of data and improve the rationality of the dataset's division into training and testing sets. To address this limitation, we introduce the GigaTraj dataset, which resolves the shortcomings of the PANDA and provides a solution for conducting long-term trajectory prediction research.

Long-term trajectory predictions. Current trajectory prediction methods have been extensively explored using various techniques, including force models [24, 25], recurrent neural networks [10, 26, 27], generative adversarial networks [9, 28, 29], variational auto-encoder [16, 30], and many more. Despite the high performance achieved by these methods, there is currently a lack of approaches specifically designed for minute-level long-term trajectory prediction. In order to achieve the goal of long-term trajectory prediction, researchers have also explored intention-oriented methods. However, due to the highly complex nature of pedestrian intentions in large-scale outdoor scenes, understanding and predicting the intentions and trajectories of pedestrians is challenging. It seems that relying solely on world coordinates for long-term trajectory prediction is uninformative. There are also some works currently focusing on multi-modal trajectory prediction. For instance, Social-biGAT [31] utilizes both the world coordinates and image features for trajectory prediction, while Bae et al. [**?**] process interactions into multiple modalities for cluster analysis and prediction. However, considering GigaTraj's videos consist of billions of pixels, methods like Social-biGAT based on VGG's temporal approach are not directly applicable. Additionally, the rich multimodal information provided by GigaTraj has not been considered by current methods. Therefore, the future direction of building multi-modal models is an important research focus.

3. GigaTraj Benchmark

The GigaTraj dataset contains gigapixel-level and minute-level videos with abundant annotations, including bounding boxes, IDs, group and interaction relationships, world coordinates, and scene semantics. Given these multimodal annotations, a trajectory $\mathcal{T}_n$ of a pedestrian n with multiple observed points $\mathcal{P}_n^t$ in GigaTraj can be represented as $\mathcal{T}_n = \{\mathcal{P}_n^0, \cdots, \mathcal{P}_n^t\}$, with $\mathcal{P}^t = \{x_{tl}, y_{tl}, x_{br}, y_{br}, x_w, y_w\}$, where $x_{tl}, y_{tl}, x_{br}, y_{br}$ represent the top-left and bottom-

right pixel coordinates of a bounding box, x_w, y_w represent the corresponding world coordinates.

3.1. Data Collection

We build GigaTraj partially based on the publicly available PANDA dataset, which is collected in public areas where videography was officially approved and is published under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 license. PANDA has already anonymized the facial and other personal information in the images to protect individual privacy [**?**]. So no privacy issues are involved. Following PANDA's precedent, we used a gigapixel camera array to capture an additional 6 new scenes to supplement the GigaTraj dataset. By thoroughly analyzing the data characteristics of research and industries related to human behavior, public safety, and smart cities, we selected scenes that align with these domains. We recorded 2-4 hours of videos using a gigapixel camera array, capturing scenes from high-rise buildings. Afterward, we carefully selected representative video segments and performed post-processing to create four new gigapixel-level large-scale complex scenes.

3.2. Data Annotations

Basic Annotations. We borrow some important annotations used in this paper from the PANDA dataset, which are important for trajectory predictions. First, pixel coordinates $x_{tl}, y_{tl}, x_{br}, y_{br}$ index a high-resolution bounding box in a gigapixel-level image, which reflects the pose and behavior of pedestrians. Notably, when pedestrians are severely occluded, the bounding boxes are estimated to encompass their entire bodies. Second, group graph $\mathcal{G}_{group} \in \mathbb{R}^{N\times N}$ and interaction graph $\mathcal{G}_{interaction} \in \mathbb{R}^{N\times N}$ represent the complex social relationships among hundreds of pedestrians, with pedestrian as nodes. Specifically, the edge e_G^{ij} of group graph and the edge e_I^{ij} of interaction graph represents the social and interaction relationship between the pedestrian i and j, where e_G^{ij} can be divided into 3 classes (*Acquaintance*, *Family*, *Business*) and e_I^{ij} can be divided into 5 classes (*Physical Contact*, *Body Language*, *Face Expressions*, *Eye Contact*, *Talking*). We strongly believe that this information is crucial for accurate trajectory predictions. The readers can refer to the PANDA paper for more information [21]. Following the annotation procedure of PANDA, we have annotated the 4 new scenes.

Homography and World Coordinates. World coordinates serve as an essential component in trajectory prediction, especially for the datasets with side-view videos. Therefore, without the homography matrix required to obtain world coordinates, the PANDA dataset cannot be directly utilized for trajectory prediction. To obtain the world coordinates, we first utilized a laser scanner to collect the cloud points of the real-world large-scale scenes. Then, we selected hundreds of marks in the scenes, and measured the distance between them. Finally, we employ direct linear transformation algorithms to estimate the homography matrix for each scene and determine the corresponding world coordinates. Based on the precision of the scanner, the discrepancy between the computed world coordinates and the actual ones is at the centimeter-level, which is more than sufficient for trajectory prediction.

Scene Semantic Annotations. Scene semantics is crucial for studies of human behaviors. For example, the probability of pedestrians walking in parterre is extremely low, while the probability of walking on the sidewalk is very high. We also notice that when a person walks on the lawn, it will affect the pedestrians nearby to walk on the lawn too. Therefore, we carefully segment the gigapixel images to obtain the scene semantics. We defined 9 class labels, including 'sidewalk', 'lawn', 'store', 'street', 'parterre', 'building', 'attraction', 'station', 'pool', and so on. After obtaining meticulous pixel semantics, we use the estimated homography matrix to map pixels to the ground.

3.3. Dataset Statistics

We construct the GigaTraj dataset for minute-level trajectory predictions. Although the original videos are beyond one minute, we set up a one-minute window to capture the original 2 FPS annotated image sequences, and obtain a sequence of 120 frames to construct the GigaTraj dataset. Overall, GigaTraj contains long-term complex trajectories with intricate interactions in large-scale scenes.

Trajectory Length. For small-scale scenarios, even with long observation times, it is not possible to record long-range trajectories. However, in the GigaTraj dataset, the scenes reach up to $4\times10^4\ m^2$, making it easy to observe trajectories for a minute. However, due to occlusions or pedestrians exiting the scene, some trajectories may be less than 1 minute in length. In the process of creating the dataset, we removed trajectories with more than 20% missing data and counted the number of available trajectories for each scene. It can be seen that in GigaTraj, the minimum number of trajectories per scene is 90, the maximum is 3078, and the average is 1014, which presents a significant advantage compared to the data in Table 1.

Trajectory Complexity. In some straightforward scenarios (such as those depicted in Figure 2bcd, where pedestrian paths and lane markings are clearly defined), the trajectory patterns are relatively uncomplicated and easily predictable. However, in GigaTraj, the scenes are expansive and semantically complex, resulting in a greater diversity of trajectories and posing challenges for prediction. We utilize yaw angles to assess the complexity of each scene in GigaTraj.

Statistics	Scene 01	Scene 02	Scene 03	Scene 04	Scene 05	Scene 06	Scene 07
#Traj.	275	633	212	358	271	189	101
Traj. Length (s)	58.7	57.9	57.0	57.7	57.8	57.9	57.7
Complexity	111.74°	96.67°	112.67°	120.99°	78.39°	107.06°	70.99°
Bbox width (px)	199±135	120±98	156±152	97±38	189±191	223±205	342±215
Bbox height (px)	481±316	298±233	388±366	244±78	504±448	514±463	973±577

Statistics	Scene 08	Scene 09	Scene 10	Scene 11	Scene 12	Scene 13	Scene 14
#Traj.	191	341	292	103	279	677	526
Traj. Length (s)	59.2	56.9	57.2	57.7	57.6	51.8	51.5
Complexity	89.56°	123.16°	79.11°	85.60°	88.29°	98.16°	162.59°
Bbox width (px)	123±74	97±38	189±191	354±218	132±95	100±47	297±166
Bbox height (px)	299±169	244±78	504±448	823±476	302±197	253±102	538±391

Table 2. Statistics for the specific scenes in GigaTraj. Following careful selection, all scenes have yielded over one hundred valid trajectories, with each trajectory observation exceeding 50 seconds. Across all scenes, the yaw angles of the trajectories are notably large, posing challenges for prediction. The mean value of the bounding boxes is relatively large, indicating sufficient visibility of pedestrian behavior, yet the significant standard deviation presents challenges for processing.

Specifically, the yaw angle is calculated using the formula $\eta = \arccos \frac{<\vec{u},\vec{v}>}{||\vec{u}||\cdot||\vec{v}||}$, where $\vec{u} = [x_2 - x_1, y_2 - y_1]$ and $\vec{v} = [x_3 - x_2, y_3 - y_2]$, given three points ($P_1(x_1, y_1)$, $P_2(x_2, y_2)$, and $P_3(x_3, y_3)$) in a trajectory as shown in Figure 3. We calculate the average of the top 5 yaw angles in a trajectory to represent its complexity. The mean complexity of trajectories in a scene reflects the complexity of that scene, which is presented in Table 2. When compared with the complexity of the *bookstore* in SDD, the complexity of GigaTraj is notably pronounced.

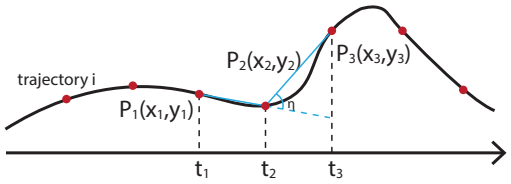


Figure 3. Yaw angle of a point in a trajectory.

Bounding Boxes. GigaTraj provides bounding boxes with visual information about pedestrians to improve trajectory prediction accuracy. As shown in Figure 4, it is easy for people to infer that the lady has a boyfriend, and her trajectory is closely related to his. They are also observed interacting with a store annotated in the dataset. Although their trajectories are not recorded for several seconds, it is expected that they will reappear after spending some time in the store. Given these informative observations, GigaTraj can support a new research direction in multimodal trajectory prediction. Besides, in practical applications, there is often a need for online trajectory prediction models, which means that trajectory prediction and object detection need to be iterated dynamically. In response to such requirements, the significant variation and distribution of multiple object scales in GigaTraj pose great challenges. Firstly, multi-scale detection itself is a huge challenge, with scale differences of the same type of objects exceeding 100 times in GigaTraj. Secondly, for trajectories, due to the imaging depth approaching 200 meters, a slight loss in distant detection may lead to significant errors in trajectory prediction. Therefore, it can be said that GigaTraj is an excellent dataset for studying joint modeling of long-range trajectories and multi-object detection in large-scale scenarios.

Interactions. The GigaTraj dataset contains more than 5.1k group and interaction annotations. However, when it comes to observing hundreds of pedestrians at a minute level, these annotations are still incomplete. During the annotation process, we observed that different annotators often held differing opinions regarding the classification of groups and the interactions among pedestrians within a scene. Consequently, we believe that proximity implies connections and mutual influence among individuals. By establishing a spatial distance threshold to delineate spatial relationships, we discovered that these relationships are highly intricate.

Quality Control Strategies. To ensure the quality of annotations, the annotation was jointly completed by the authors and a professional annotation company. We established strict annotation standards and set a minimum requirement of 98% annotation accuracy. The project maintained an annotation team of $\sim$ 30 people and a review team of $\sim$ 10 people. The project was carried out on a highly integrated annotation platform, where the annotation team, review team, and paper authors collaborated to complete the annotation and quality control process. Specifically, the review team randomly reviewed 20% of the annotations, and the authors addressed any misunderstandings or corner cases in the annotation process.

Figure 4. GigaTraj provides long-term bounding box annotation and estimation, as well as id associations.

3.4. Dataset Split

The GigaTraj dataset consisting of 16 scenes has been divided into 8 training sets and 8 test sets, according to the content of the videos. The thumbnails of the training and testing set are shown in Figure 5. Scene No.9 is extremely similar to Scene No.4, as they are in the same view and temporally adjacent. Scene No.10 and Scene No.5 are also in the same view, but there is a difference in lighting conditions. Scene No.5 is during the daytime, while Scene No.10 is during the nighttime. Scene No.6 and Scene No.11 are both set in the same location and both occur during the daytime, while their imaging perspectives are different. Due to the presence of these scenes in the training set, the predictions become relatively straightforward. Scene No.11-16 are unseen scenes for trajectory predictions, which makes them more challenging to predict. The models can be trained and verified using temporally divided 8 training sets.

4. Algorithms Analysis

In this section, we select representative algorithms for trajectory prediction, evaluate their performance on minute-level trajectory predictions using the GigaTraj dataset, and analyze the results.

4.1. Experimental Setup

Metrics. We adopt Average Displacement Error (ADE) and Final Displacement Error (FDE) to evaluate our methods, which are widely used in prior works [9, 29, 32]. ADE computes the mean square error of the overall estimated positions in the predicted and ground-truth trajectories, while FDE measures the distance between their respective final destinations:

$$ADE = \frac{\sum_{i=1}^{n} \sum_{t=T_{\text{start}}}^{T_{\text{end}}} \sqrt{(\hat{x}_i^t - x_i^t)^2 + (\hat{y}_i^t - y_i^t)^2}}{n\,(T_{\text{end}} - T_{\text{start}})} \quad (1)$$

$$FDE = \frac{\sum_{i=1}^{n} \sqrt{\left(\hat{x}_i^{T_{\text{end}}} - x_i^{T_{\text{end}}}\right)^2 + \left(\hat{y}_i^{T_{\text{end}}} - y_i^{T_{\text{end}}}\right)^2}}{n} \quad (2)$$

where x_i, y_i represnent the observed trajectory i at time t, $\hat{x}_i$, $\hat{y}_i$ represent the predicted ones, and T_{start}, T_{end} represent the start and end times of predicted trajectories. Lower ADE and FDE values indicate lower prediction errors.

Baselines. As discussed before, there are many technical routes for trajectory prediction, and we chose the most representative ones as the baselines. Specifically, social-LSTM [10] is the first method considering the complex interactions among pedestrians and uses a pooling layer allowing for information sharing among LSTMs, capturing interactions within neighboring trajectories. SGAN (short for social generative adversarial network) [9] is the most representative GAN-based trajectory prediction model, which uses a recurrent sequence-to-sequence model with a novel pooling mechanism and adversarial training to predict diverse and socially plausible human motion behaviors. Trajectron++ [8] is also a graph-structured recurrent model for trajectory prediction, and it especially uses agent dynamics and diverse data like semantic maps for more accurate forecasts. MemoNet [33] is the state-of-the-art method for trajectory prediction, which is replicated by the mechanism of retrospective memory in neuropsychology for trajectory prediction. PECNet [11] infers distant trajectory endpoints to assist in long-range multi-modal trajectory prediction. It incorporates a novel non-local social pooling layer to infer diverse yet socially compliant trajectories. Additionally, it employs a simple "truncation trick" for enhancing diversity and multi-modal trajectory prediction performance.

Implementation Details. For all the baseline methods, we re-implemented them in our scenarios according to the descriptions in the original papers, except PECNet and Memonet, for which we directly utilized the officially released code for experimentation. As for all baselines, we adhered strictly to the original methods' hyperparameters. Within the vicinity of the hyperparameters reported in the original papers, we employed grid search, and the settings that performed best on the validation set were used for the test setup. Regarding the training dataset, we used a 10-second sliding window to create training data of the required length from the original sequential data. Specifically, for PECNet, we sought to explore the impact of position semantics on the prediction results. We use the GT coordinates in the experiments. We believe that similar viewpoints indicate similar semantics, leading to similar trajec-

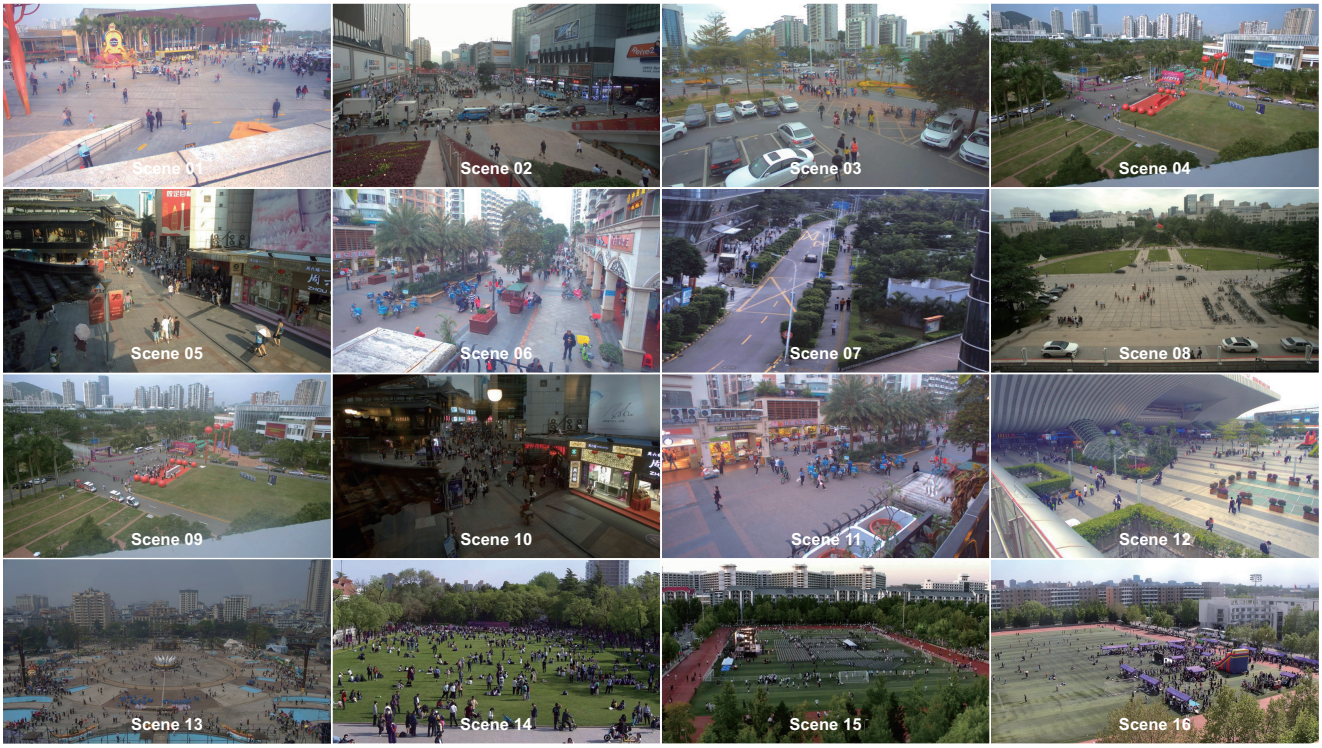


Figure 5. Dataset thumbnails. GigaTraj offers 16 gigapixel-level complex scenes featuring numerous pedestrians engaging in interactions. Scenes 1-8 are designated for training, while Scenes 9-16 are intended for testing. In particular, Scenes 4 and 9 share similar temporal and angle characteristics. Scenes 5 and 10 present the same perspective but with varying lighting conditions. Lastly, scene 6 and 11 capture identical scenes but from different viewing angles.

tory patterns and making prediction easier. All baselines use GT coordinates and BEV maps. These baselines except for Vanilla-LSTM use interaction graphs. Researching how to use detection, tracking, and pose estimation from raw images would be important topics in the future. We employed forward and backward interpolation to complete past trajectories, and we did not evaluate missing points in future trajectories. We conducted experiments with a default setting of 30s input and 30s output.

4.2. Results

Overall Performance. We present the prediction performance of the baselines in Table 3, and observe some several key phenomena as follows:

- In large-scale scenarios, hundreds of pedestrians have complex relationships, but leveraging these relationships to improve prediction accuracy is non-trivial. We observe that using simple pooling layers, social-LSTM, and SGAN results in larger errors compared to vanilla LSTM which does not consider relationships. This is clearly because the models cannot comprehend the complex associations within the scene. Therefore, it may be necessary to research how to construct an interaction graph among pedestrians that is easier for the model to understand. Besides, graph convolution tailored for complex relationships among multiple objects in large-scale scenarios is an important topic.
- Predicting 30s (60 frames) long trajectories is very challenging. The open scenes in GigaTraj imply that pedestrians may have more possibilities for their destinations and greater freedom in the trajectories between their starting point and destination. It can be observed that both ADE and FDE in the table are much larger than those on small-scale datasets (being 0.17 for NSP-SFM in ETH/UCY). Therefore, designing models that are more suitable for long-sequence prediction is an important topic.
- Models are difficult to generate for unseen scenes. Scenes 9-11 are somewhat similar in viewpoint and time to the samples in the training set, so each method performs much better in predicting them compared to scenes 12-13. How to make models learn the local semantics in scenes and improve their generalization is also one of the important directions for future research.
- We also investigate the performance of existing models when predicting only 10s trajectories. It can be observed that as the prediction time increases, the prediction error significantly increases. Long-term prediction in large-scale scenes poses a significant challenge.

Method	Scene 09	Scene 10	Scene 11	Scene 12	Scene 13	Scene 14	‡Overall ($T = 30s$)	‡Overall ($T = 10s$)
Vanilla-LSTM [34]	3.76/6.15	6.74/12.29	2.56/4.62	4.23/7.62	6.31/11.59	3.71/5.65	4.40/7.58	2.45/4.40
Social-LSTM [10]	2.68/4.48	9.37/13.63	3.10/5.41	4.15/7.60	8.00/18.49	5.15/7.34	5.71/9.78	2.19/3.26
SGAN [9]	2.71/5.22	7.74/15.04	2.41/4.87	3.78/7.66	5.00/9.57	3.30/5.97	4.13/8.02	1.03/2.84
PECNet [11]	1.51/2.50	8.30/15.33	3.22/6.70	15.72/33.83	5.39/10.49	1.61/2.78	2.05/2.88	1.09/2.77
Trajectron++ [8]	1.15/1.88	1.88/2.95	1.28/2.51	22.40/53.67	2.40/4.26	1.35/1.61	1.72/2.85	0.70/0.82
MemoNet [33]	1.56/1.92	2.12/2.73	1.22/1.60	1.50/2.23	2.53/3.15	1.70/1.81	2.04/2.87	0.89/1.00

‡ When submitting the article, Scene 15 & 16 have not been fully annotated yet, so only partial test set results are shown here. We will update the latest results on our website www.gigavision.cn.

Table 3. Performance comparison on minute-level trajectory prediction tasks in GigaTraj. $\min ADE_{20}$/$\min FDE_{20}$ are shown in this table, and the lower $\min ADE_{20}$/$\min FDE_{20}$ means the lower prediction error.

Ablation Study. Since PECNet has the characteristic of easily expandable modules, We chose PECNet as the base model and extended the inputs to accommodate multi-modal data input in line with GigaTraj. By concatenating these features after the two-dimensional position input, i.e., for each data point in the trajectory sequence, it was created in the form of (x, y, $semantic_{id}$) and (x, y, $orientation_{id}$). Furthermore, concerning social relations, we followed PECNet's approach, constraining all the neighbors of a pedestrian to be in the same mini-batch to perform the forward pass in mini-batches instead of processing all the pedestrians in the scene in a single forward pass, thereby avoiding memory overflow. We present the ablation results in Table 4. As we can see, although GigaTraj provides useful multimodal data input, utilizing them is a non-trivial task. A significant improvement occurred when using graph information. However, concatenating scene semantics without considering graph information did not directly work. Therefore, how to leverage multi-modal information for long-range trajectory prediction is an important research direction for the future.

Setup	$\min ADE_{20}$	$\min FDE_{20}$
with graph	2.18	4.71
without graph	3.85	6.85
with semantically	3.72	6.64

Table 4. Performance of Module Ablation in PECNet.

Collision in Large-scale Scenes. Existing methods for minute-level trajectory prediction in large scenes have encountered serious issues in collision prediction, leading to evidently unreasonable forecast results. Figure 6 illustrates a visual representation of the prediction results for the 10th frame by the state-of-the-art MemoNet algorithm, revealing significant collisions between objects in the scene. Therefore, there is an urgent need to develop predictive models that can effectively mitigate collision risks. In GigaTraj, we provide sufficiently long historical trajectories, bounding boxes (reflecting pedestrian behavior information), and scene semantics as features. These features play a crucial role in predicting long-term trajectories for multiple individuals and are worthy of further in-depth research.

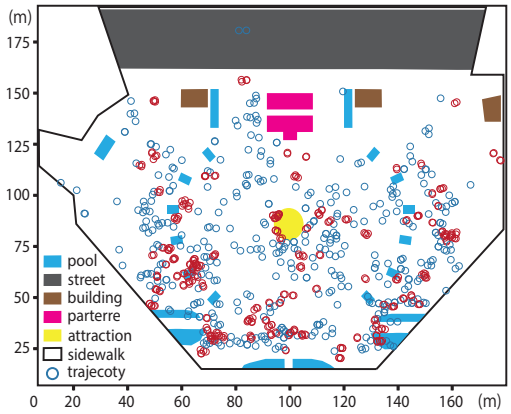


Figure 6. Memonet's 10th frame prediction visualization shows a severe collision marked by a red circle.

5. Conclusions

In this paper, we introduce the GigaTraj dataset designed for minute-level long-term trajectory predictions in complex scenes at the gigapixel level. GigaTraj consists of 14 videos covering a wide field of view of 4×10^4 m^2 and providing gigapixel-level high-resolution, accompanied by comprehensive annotations. These annotations include bounding boxes, long-term ID associations, complex group and interaction information, world coordinates of pedestrians and scenes, as well as scene semantics. Our empirical study has revealed the inadequacy of existing trajectory prediction models for such complex scenarios. Additionally, we have outlined promising research directions that can be pursued using the GigaTraj dataset, paving the way for further advancements in this area of study.

Acknowledgements

This work is supported in part by the Natural Science Foundation of China (NSFC) under contract No. 62125106, 62088102, and U21B2013 in part by the Ministry of Science and Technology of China under contract No. 2021ZD0109901, in part by Tsinghua-Zhijiang joint research center.

References

[1] 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.

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

[3] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 549–565. Springer, 2016.

[4] Julian Bock, Robert Krajewski, Tobias Moers, Steffen Runde, Lennart Vater, and Lutz Eckstein. The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1929–1934. IEEE, 2020.

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[6] Barbara Majecka. Statistical models of pedestrian behaviour in the forum. *Master's thesis, School of Informatics, University of Edinburgh*, 2009.

[7] 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.

[8] 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.

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

[11] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 759–776. Springer, 2020.

[12] Minye Wu, Haibin Ling, Ning Bi, Shenghua Gao, Qiang Hu, Hao Sheng, and Jingyi Yu. Visual tracking with multiview trajectory prediction. *IEEE Transactions on Image Processing*, 29:8355–8367, 2020.

[13] Ruijie Quan, Linchao Zhu, Yu Wu, and Yi Yang. Holistic lstm for pedestrian trajectory prediction. *IEEE transactions on image processing*, 30:3229–3239, 2021.

[14] Hung Tran, Vuong Le, and Truyen Tran. Goal-driven long-term trajectory prediction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 796–805, 2021.

[15] Harshayu Girase, Haiming Gang, Srikanth Malla, Jiachen Li, Akira Kanehara, Karttikeya Mangalam, and Chiho Choi. Loki: Long term and key intentions for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9803–9812, 2021.

[16] 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*, pages 2221–2230, 2022.

[17] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Fang Zheng, Nanning Zheng, and Gang Hua. Social interpretable tree for pedestrian trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2235–2243, 2022.

[18] Xiaoyun Yuan, Lu Fang, Qionghai Dai, David J Brady, and Yebin Liu. Multiscale gigapixel video: A cross resolution image matching and warping approach. In *2017 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2017.

[19] David J Brady, Michael E Gehm, Ronald A Stack, Daniel L Marks, David S Kittle, Dathon R Golish, EM Vera, and Steven D Feller. Multiscale gigapixel photography. *Nature*, 486(7403):386–389, 2012.

[20] Xiaoyun Yuan, Mengqi Ji, Jiamin Wu, David J Brady, Qionghai Dai, and Lu Fang. A modular hierarchical array camera. *Light: Science & Applications*, 10(1):37, 2021.

[21] Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, et al. Panda: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3268–3278, 2020.

[22] Xueyang Wang, Xuecheng Chen, Puhua Jiang, Haozhe Lin, Xiaoyun Yuan, Mengqi Ji, Yuchen Guo, Ruqi Huang, and Lu Fang. The group interaction field for learning and explaining pedestrian anticipation. *Engineering*, 2023.

[23] Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory prediction via neural social physics. In *European Conference on Computer Vision*, pages 376–394. Springer, 2022.

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

[25] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE conference on computer vision and pattern recognition*, pages 935–942. IEEE, 2009.

[26] Jeremy Morton, Tim A Wheeler, and Mykel J Kochenderfer. Analysis of recurrent neural networks for probabilistic mod-

eling of driver behavior. *IEEE Transactions on Intelligent Transportation Systems*, 18(5):1289–1298, 2016.

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

[28] Yue Hu, Siheng Chen, Ya Zhang, and Xiao Gu. Collaborative motion prediction via neural motion message passing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6319–6328, 2020.

[29] 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.

[30] A Bhattacharyya, M Hanselmann, M Fritz, B Schiele, and CN Straehle. Conditional flow variational autoencoders for structured sequence prediction. arxiv 2020. *arXiv preprint arXiv:1908.09008*, 1908.

[31] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6272–6281, 2019.

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

[34] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.