Human Trajectory Prediction with Momentary Observation

Jianhua Sun, Yuxuan Li, Liang Chai, Hao-Shu Fang, Yong-Lu Li, Cewu Lu[§]
Shanghai Jiao Tong University, China
{gothic,yuxuan_li,chailiang1234,yonglu_li,lucewu}@sjtu.edu.cn, fhaoshu@gmail.com

Abstract

Human trajectory prediction task aims to analyze human future movements given their past status, which is a crucial step for many autonomous systems such as self-driving cars and social robots. In real-world scenarios, it is unlikely to obtain sufficiently long observations at all times for prediction, considering inevitable factors such as tracking losses and sudden events. However, the problem of trajectory prediction with limited observations has not drawn much attention in previous work. In this paper, we study a task named ***momentary trajectory prediction****, which reduces the observed history from a long time sequence to an extreme situation of two frames, one frame for social and scene contexts and both frames for the velocity of agents. We perform a rigorous study of existing state-of-the-art approaches in this challenging setting on two widely used benchmarks. We further propose a unified feature extractor, along with a novel pre-training mechanism, to capture effective information within the momentary observation. Our extractor can be adopted in existing prediction models and substantially boost their performance of momentary trajectory prediction. We hope our work will pave the way for more responsive, precise and robust prediction approaches, an important step toward real-world autonomous systems.*

1. Introduction

Human trajectory prediction [2,9,15] plays an important role in the area of human behavior understanding [17, 18, 28, 40] and autonomous driving systems [3, 15, 31] by investigating future movements of traffic agents given their past status observed from a video. Despite the good performance achieved by existing methods, these approaches are developed on historical observations over several seconds. However, accurate tracking over long periods of time is quite difficult, especially in congested traffic scenarios [11, 16, 21, 25, 29]. Further, responsive and precise predictions are important and necessary for safety purposes when emergency situations suddenly come into view (see Fig. 1). Therefore, a prediction task with momentary observation is essential to be studied.

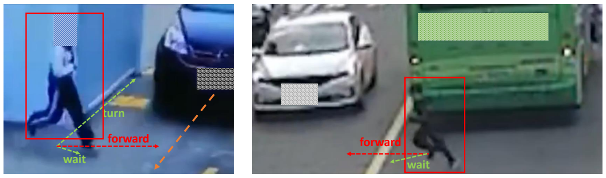

Figure 1. The necessity of momentary prediction in emergency situations. The figures illustrate some sudden events, often occurring in highly occluded areas such as underground parking lots and congested roads. The pedestrian's trajectory have to be precisely and immediately predicted to avoid the potential collision in a very short time.

In this paper, we focus on the most extreme case that only two frames are available for observation, one frame for social and scene context and both frames for the velocity of agents. We refer to the task as *Momentary Trajectory Prediction*. Note that we exclude the case of single-shot (one-frame observation) where the basic information, velocity magnitude, is missing. After reducing the observation horizon in previous approaches, we observe obvious degradation of their performance (see Fig. 2).

Despite the lack of temporal features, three common types of information still widely exist between two adjacent frames and provide a large number of leads for trajectory prediction: i) velocity of agents, ii) social contexts, and iii) scene contexts. The velocity provides motion basis of human short-term behavior, while context information implies enough clues to predict long-term trends and disturbances of movements. A fundamental aspect of momentary trajectory prediction is how to explore and integrate these potential information implicit in the limited observations. In this paper, we raise a unified input formulation to join the three types of information as a whole (see Fig. 4), integrating all information at data level. The major insight of this input formulation is to consider scene restrictions as static interactive objects while surrounding traffic agents as dynamic interactive objects with corresponding velocities. Then Momentary Observation feature Extractor (MOE) (see Fig. 5)

[§] Cewu Lu is corresponding author, member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China and Shanghai Qi Zhi Institute

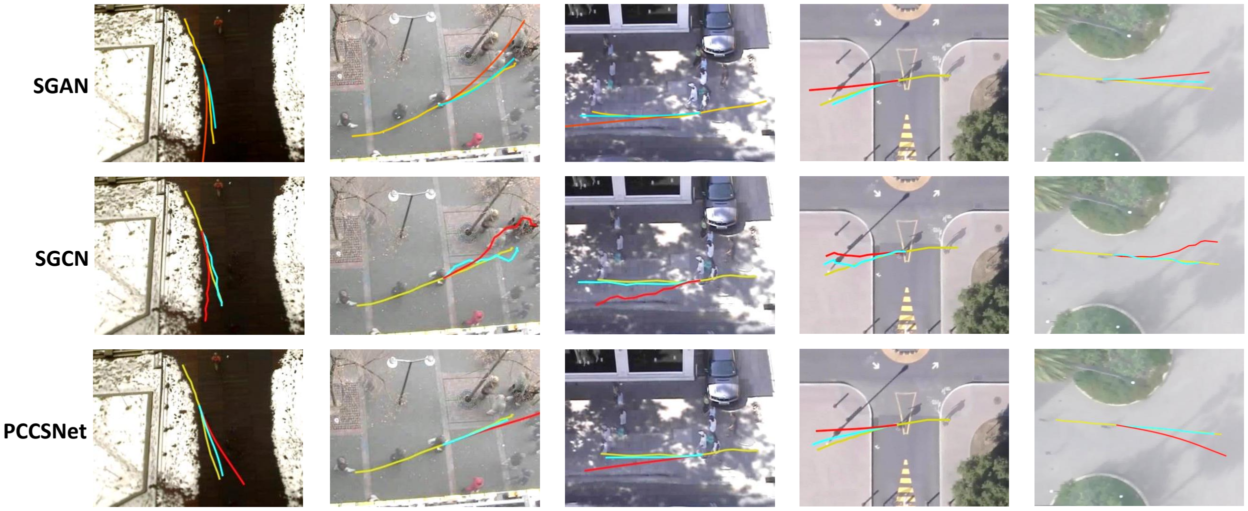

Figure 2. Qualitative comparison of trajectory prediction between traditional and momentary observation time on [6,32,35], including state-of-the-art PCCSNet [35]. Lines in yellow denote ground truth trajectories in the past (8 frames/3.2 sec) and future (12 frames/4.8 sec), lines in cyan denote predictions with traditional observation time (8 frames/3.2 sec) and lines in red denote predictions with momentary observation time (2 frames/0.8 sec). The frame rate is 2.5fps. Obvious degradation occurs as the observation time reduces.

is proposed to directly explore a joint historical representation from the input.

Our feature extraction process enjoys a significant advantage by avoiding the need for feature alignment between social and scene context. Previous work such as [12, 19, 30] models these different types of information by separately encoding them, then fusing them together with a fusion module or concatenation, and finally decoding the fused feature for prediction results. This paradigm in previous work adds a burden of feature alignment on the fusion module or the following decoder, since there is a big difference between the feature of trajectories (coordinate sequences) and scene contexts (RGB images). This burden is enlarged in the momentary observation setting, considering a good social feature is much harder to obtain due to the lack of historical trajectories. In comparison, by integrating these information at data level and encoding the input in a unified manner, our approach subtly avoids the need of feature alignment and learns a better momentary observation representation.

Moreover, another potential problem resulting from the reduced observation is that prediction models become more difficult to fit, since they have to map the observation space with a much lower dimension to the original prediction space. We alleviate this by introducing a novel pre-training mechanism for the MOE, named soft pre-training, leveraging ideas from multi-task learning and self-supervised learning. In soft pre-training, the supervisions are several sub-tasks related to trajectory prediction yet much easier. We raise masked trajectory complement and context restoration as two sub-tasks in our implementation. After pre-training, the MOE can be easily integrated into existing prediction frameworks to improve their momentary prediction performance.

Exhaustive experiments are conducted on ETH/UCY dataset [13,27] and SDD [28] (Stanford Drone Dataset). We first perform a rigorous study of the performance of existing state-of-the-art approaches on this challenging setting. Then, we adopt the MOE into multiple prediction frameworks to show substantial improvement can be brought with the aid of our approach.

2. Related Work and Motivation

Human Trajectory Prediction. Given a period of observation, the trajectory prediction task [2, 10, 15, 30, 38, 41] is proposed to forecast possible future movements of traffic agents. It has already been widely adopted in the field of robotics, security surveillance and autonomous driving. Previous studies have covered different aspects of trajectory prediction. A large number of work mainly focuses on mining and exploiting representations that are useful for prediction, including social interactions [2, 6], environment restrictions [19, 30] and human kinematics [19]. Meanwhile, some researchers study the multimodal nature of trajectory prediction [6, 15]. Pioneers mainly conduct human trajectory prediction research based on typical settings (i.e. observation lasts 3.2sec and predicting future 4.8sec). Recently, more challenging prediction settings have been studied to push a step forward to real-world autonomous systems. For example, [4, 23] are raised for long term trajectory prediction and achieve great performance. In this paper, we focus on the momentary observation setting, which is inevitable in many real-world scenarios. We perform a rigorous study of

Benchmark	New Agents	Tracking Losses	Station to Movement	Total	Duration/sec
SDD	10300 / 0.59	20335 / 1.17	8031 / 0.46	38666 / 2.22	17384
ETH/UCY	1950 / 0.98	-	83 / 0.04	2033 / 1.02	1980

Table 1. The appearance quantity (times) / frequency (times per second) of agents with only momentary observation.

the performance of existing approaches on this new setting in Sec. 5.

Feature Extraction for Trajectory Prediction. Noticing the observation clues for human trajectory prediction often come from different sources such as social interactions, environment restrictions and *etc*., approaches of feature extraction for trajectory prediction various a lot. Many studies encode social interactions by integrating pooling mechanisms [2, 6] or GNNs [12, 24, 34] into a temporal or recurrent pipeline. To take environment restrictions into consideration, Sophie [30] extracts rich information from images with deep convolutional neural networks. Next [19] analyzes person keypoints with CNNs to take a step further. To fuse deep features from different sources together, the majority of studies [12, 19, 30] use attention mechanisms or simple concatenations. However, in the momentary observation setting, the feature alignment between social and scene features becomes much harder, leading to unsatisfactory prediction results. In this paper, we integrate different types of observation clues at data level following a novel input formulation, and then design the MOE to jointly learn them in a unified manner.

Self-supervised Learning. Self-supervised learning is of great interest recently in many areas [5, 14, 39]. Traditional self-supervised learning approaches mainly follow two lines, generative [5, 8] and contrastive [7, 26]. A generative learning method uses reconstruction loss to supervise whether the decoder can accurately reconstruct the original input with learned representations. A contrastive method, on the other hand, aims to train the model to discriminate between different inputs on the feature space. In this paper, both sub-tasks, masked trajectory complement and context restoration, proposed for pre-training is performed in a generative self-supervised learning manner.

Multi-task Learning. Multi-task learning [36] aims to improve the performance and generalization of models by leveraging domain-specific information contained in the training signals of related tasks. One of its great advantages is that more comprehensive representations can be learned since associated tasks often share complementary information. Recently, multi-task learning models combining trajectory prediction and other vision tasks including pedestrian detection [20, 22, 42], tracking [20] and action recognition [19] have attracted a lot of interest. We introduce this idea in our pre-training process to encourage our model to learn compact and comprehensive representations revealing both human motions and surrounding contexts.

3. Momentary Trajectory Prediction

Problem Formulation. Given two frames of the video at time instants $[1, 2]$, speed of all agents and an image of the scene can be first pre-processed as observed information. Assume the displacement of an agent between the two frames is Δx and the frame rate is r, the speed of this agent can be estimated by $\Delta x \times r$. Based on these observations, the task aims to predict future possible trajectories $\hat{Y} = \{(\hat{x}_t, \hat{y}_t)\}$ for prediction horizon $t \in [3, 2 + T_{pred}]$.

A Prevalent Situation. Although the setting of momentary trajectory prediction is extreme, this situation is prevalent in real-world scenarios. In general, three widely existing common cases will lead to agents with only momentary observation: i) new agents coming into the view, ii) tracking losses, and iii) stationary agents suddenly starting to move. We count the three cases on two major large-scale human trajectory prediction benchmarks, SDD [28] and ETH/UCY [13, 27], see Tab. 1. The statistics show that agents with only momentary observations appear every second on average. Thus, momentary trajectory prediction is an essential topic that should not be ignored to develop safer autonomous systems. Note that ETH/UCY datasets do not give annotations for tracking losses, we only consider the other two situations.

Discussion. Comparing with traditional formulation that the observation time instants are $[1, T_{obs}]$ (T_{obs} is usually 8 in experiments of previous work), our formulation shortens the number of observed frames to an extreme level. In this situation, temporal motion behavior information implicit in historical paths is missing. Since these missing information usually includes clues for acceleration, deceleration, redirection and many others behaviors which are essential for trajectory prediction, momentary prediction is much more challenging. Still, there are three types of potential information still widely existing in momentary observations: velocity estimated by the displacement between two frames, social context and scene context. Other information such as skeletons and facial expressions may also be extracted from some momentary observations.

To develop a momentary prediction algorithm, there are two inevitable issues that have to be addressed. One is how to effectively extract observation representations of poten-

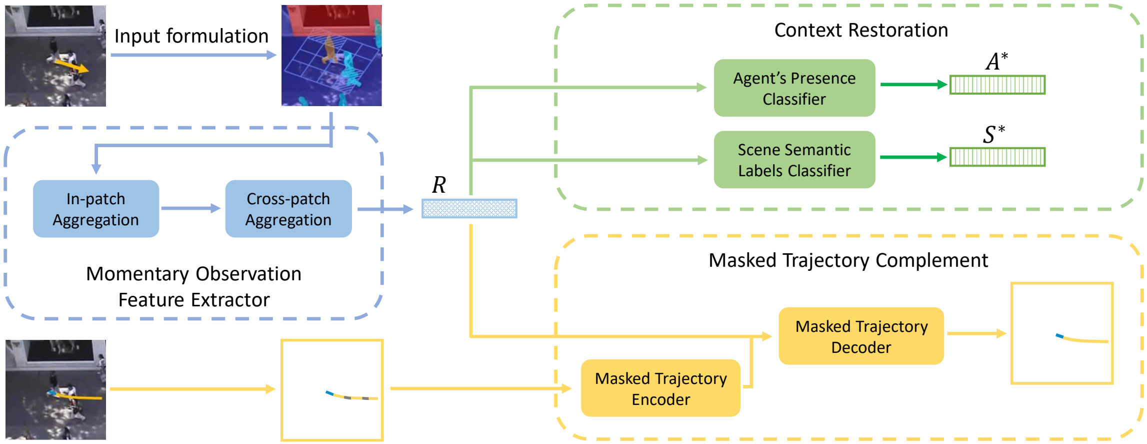
Input formulation
Context Restoration
Agent's Presence Classifier
A^*
Scene Semantic Labels Classifier
S^*
In-patch Aggregation
Cross-patch Aggregation
R
Momentary Observation Feature Extractor
Masked Trajectory Complement
Masked Trajectory Decoder
Masked Trajectory Encoder

Figure 3. An overview of our proposed approach including: i) feature extraction with Momentary Observation feature Extractor (blue) and ii) soft pre-training with Masked Trajectory Complement (yellow) and Context Restoration (green). R denotes the learned representation output by MOE, S^*, A^* are restored labels indicating scene semantics and the presence of traffic agents referring to Eq. 3 and 4. How we formulate the input is illustrated in Fig. 4, and the detailed structure of MOE is shown in Fig. 5. After pre-training the MOE according to the flowchart, it can be adopted in existing prediction frameworks as the feature extraction part.

tial information in momentary observations (*i.e.* social and scene context). The paradigm in previous work [12, 19, 30] that integrates social and scene information in feature level (with fusion modules or concatenation) shows certain defects in feature alignment, considering satisfactory social features become much harder to obtain because of the lack of historical trajectories. The other is how to effectively fitting the model. As the dimension of observation space greatly narrows, prediction models have to fill a larger gap between the observation space and the prediction space.

Note that we do not study the case of one-shot observation that only one frame is considered as input, because the magnitude of velocity cannot be estimated. Without such an important lead, models may be able to plan a reasonable future path but fail to align the time and locations accurately.

4. Approach

As mentioned above, there are two inevitable issues that have to be addressed in the momentary trajectory prediction configuration: i) effectively extracting momentary observation representations, and ii) facilitating better fitting of the model. An illustration of our approaches is in Fig. 3. We discuss about our proposed Momentary Observation feature Extractor (MOE) to address issue i in Sec. 4.1 and the soft pre-training to address issue ii in Sec. 4.2. Further, we show how to integrate our MOE into existing prediction frameworks and implementation details in Sec. 4.3.

4.1. Momentary Observation Feature Extraction

Different from previous approaches [12, 19, 30] that integrate social and scene context information at the feature level, our insight is to integrate them at the data level, which directly avoids the potential problems in feature alignment. To model this insight, we design a novel formulation for both social and scene information into a unified input format I, and then use the MOE to learn joint representations.

Input formulation. The formulation of our input is illustrated in Fig. 4. We organize it as a context map I by considering scene restrictions as static interactive objects and traffic agents as dynamic interactive objects with corresponding velocity. In detail, the input can be generated as following:

1. Pre-process the scene by semantic segmentation and map the semantic label to the coordinate system of trajectories by mapping each pixel to a coordinate.

2. Rotate the coordinate system so that the velocity direction of the target agent points to the positive X-axis, for normalization. Then mark an effective context area with the target as its center and split it into patches. In our implementation, the effective area is a square whose length is $m * l$, and it is split into $m \times m$ sub squares.

3. Represent each patch with an $n * 5$ dimensional patch tensor p of $n * (S, v_x, v_y, o_x, o_y)$, where n denotes the number of traffic agents in the patch, S denotes the semantic label of the patch, v_x, v_y indicate the velocity of the agent, and o_x, o_y indicate its offset from the patch center. For those patches outside the scene, we consider their semantic label as inaccessible areas. For patches containing areas with different semantic labels, we regard the semantic label of the patch center as S. For patches with no agents, n is one and (v_x, v_y, o_x, o_y) are all zeros as placeholder.

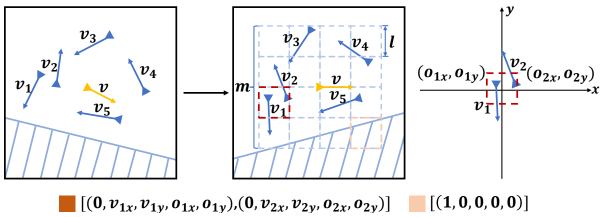

Figure 4. Schematic of the input formulation. a. The preprocessed scene in the trajectory coordinate system including the position and velocity of agents and scene semantics; b. The patch-form effective area on the rotated scene; c. A closer look in a patch.

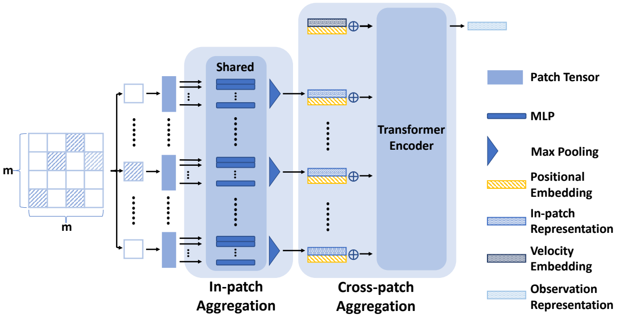

Figure 5. Detailed structure of the MOE. The architecture of the transformer encoder follows [37].

With this formulation, each 5D vector (S, v_x, v_y, o_x, o_y) can represent an interactive object surrounding the target agent and we generate context map $I = \{p_i | i \in [1, m^2]\}$ for the following encoding operation.

Feature extractor. The MOE receives the context map I as input and outputs a deep representation R for the target agent (see Fig. 3). A detailed architecture of MOE is in Fig. 5, consisting of two parts: in-patch and cross-patch feature aggregation. The in-patch feature aggregation part takes 5D vectors of n agents in a patch tensor as input, encodes each vector with multi-layer perceptrons (mlps) which share the same weights, and then aggregates the features by max pooling. The cross-patch feature aggregation part is designed as a transformer architecture, a popular feature extractor for patch form data. We adopt a learnable variable matrix for positional embedding. Further, the velocity feature of the target agent embedded by mlps is prepended to the sequence of embedded patches, whose state at the output of the transformer encoder will serve as our learned representation R. With both in-patch and cross-patch feature aggregation, the representation R contains rich information of the momentary observation, and can be then fed into any proper types of decoder to get prediction results.

4.2. Soft Pre-training

As the observed information greatly reduced, learning a mapping between the observation space and the prediction space becomes more difficult, making the model much harder to fit when directly using the whole future ground truth trajectory as supervision. To alleviate this issue, we propose the soft pre-training mechanism, which pre-trains the MOE on several sub-tasks related to trajectory prediction yet much easier in a multi-task learning manner. We raise masked trajectory complement and context restoration as sub-tasks, aiming at learning effective and comprehensive historical representation R incorporating two essential aspects for human trajectory prediction: human motion behaviors and context information.

Masked trajectory complement. Masked trajectory complement softens the prediction task by giving some clues of future trajectories. As shown in Fig. 3, some parts of the ground truth future trajectory are first masked (in grey) in random positions, and the task requires to complement the masked parts according to the representation R, which can be written as

$$Y^* = D_m([R, E_m(Y_m)]) \tag{1}$$

where Y_m and Y^* denotes the masked ground truth trajectory and complemented future trajectory. The E_m encodes Y_m into deep features, which then are concatenated with R to feed into the decoder D_m to complement Y_m. We adopt a transformer as E_m to better handle sequence Y_m which is in random length and has dynamic time intervals due to the random masks. D_m is an LSTM decoder. The loss of this task $\mathcal{L}_m$ is performed as

$$\mathcal{L}_m = MSE(Y, Y^*) \tag{2}$$

where Y denotes the ground truth trajectory and MSE is mean square error.

Context restoration. We introduce context restoration task to encourage the model to perceive comprehensive social and scene context information. Receiving R as input, the task restores the context information for each patch: i) scene semantic label S and ii) the presence of traffic agents A (1 if any agent exists in the patch and 0 otherwise), written as

$$S^* = D_s(R) \tag{3}$$

$$A^* = D_a(R) \tag{4}$$

D_s, D_a are mlp classifiers for scene semantic labels and agents' presence respectively, and S^*, A^* denote restored S, A. The loss of this task $\mathcal{L}_r$ is performed as

$$\mathcal{L}_r = CrossEntropy(S, S^*) + CrossEntropy(A, A^*) \tag{5}$$

Observation	Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG	SDD
Traditional	SGAN [6]	0.71/1.29	0.48/1.02	0.56/1.18	0.34/0.69	0.31/0.64	0.48/0.96	16.67/33.18
	Next [19]	0.73/1.65	0.30/0.59	0.60/1.27	0.38/0.81	0.31/0.68	0.46/1.00	-
	Social-STGCNN [24]	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75	15.50/26.66
	Trajectron++ [31]	0.58/1.04	0.16/0.25	0.31/0.55	0.22/0.42	0.17/0.32	0.29/0.52	11.19/19.17
	SGCN [32]	0.63/1.03	0.32/0.55	0.37/0.70	0.29/0.53	0.25/0.45	0.37/0.65	13.49/23.61
	PCCSNet [35]	0.28/0.54	0.11/0.19	0.29/0.60	0.21/0.44	0.15/0.34	0.21/0.42	8.62/16.16
Momentary	SGAN [6]	0.86/1.60	0.52/0.99	0.57/1.20	0.41/0.79	0.36/0.74	0.54/1.06	17.76/34.83
	Next [19]	0.77/1.81	0.32/0.62	0.61/1.31	0.39/0.82	0.34/0.76	0.49/1.06	-
	Social-STGCNN [24]	1.24/2.23	0.77/1.44	0.45/0.81	0.38/0.57	0.35/0.58	0.64/1.13	17.77/29.12
	Trajectron++ [31]	0.76/1.43	0.30/0.56	0.36/0.74	0.22/0.42	0.18/0.34	0.36/0.70	13.07/22.88
	SGCN [32]	0.88/1.66	0.55/1.16	0.38/0.71	0.30/0.54	0.25/0.46	0.47/0.91	15.40/25.69
	PCCSNet [35]	0.34/0.65	0.14/0.25	0.31/0.63	0.23/0.46	0.16/0.37	0.24/0.47	9.19/17.71

Table 2. Comparison of baseline methods' performance (ADE/FDE) between traditional and momentary prediction setting. Note that i) the reported results of SGAN are obtained from their official code [1], which are much better than those in their paper [6], ii) we are unable to conduct experiments of Next on SDD due to the lack of extra annotations such as human skeletons, and iii) the reported results of Trajectron++ is obtained by fixing the bug mentioned in issue #40 in their official implementation [33].

The overall loss function of our soft pre-training can be formulated as

$$\mathcal{L} = \mathcal{L}_m + \lambda \mathcal{L}_r \tag{6}$$

where λ being 0.3 in this work is used to balance the two losses.

4.3. Implementation

Implementing the MOE to prediction frameworks. After pre-training the MOE according to Fig. 3, it can be adopted in existing prediction frameworks as the feature extraction part. For example, the MOE can be used to replace the person interaction module in Next [19], history modeling part in Trajectron++ [31], and the past encoder in PCCSNet [35]. After being integrated into the framework, the MOE can be further tuned to learn adaptive representations for the framework.

Implementation details. For the input formulation, we split the square area into 36 sub-squares ($m = 6$). Each sub-square has a side length l of 1 meters in ETH/UCY, or 40 pixels in SDD. The transformer encoder in the MOE consists of 2 stacked transformer blocks. Each block has 8 attention heads. For the masked trajectory complement task, we use a mask size of 6. For pre-training, we use the Adam optimizer with a base learn rate of 0.0005 and a polynomial decay with the power of 0.95. The model is pre-trained with batchsize of 128 and max epoch of 100 with early stopping. The training is carried out on a single RTX 2080Ti.

5. Experiments

Our experiments are conducted from two aspects. First, we study the performance of existing prediction frameworks in the momentary prediction setting in Sec. 5.1. Then, we adopt the proposed MOE to multiple existing prediction frameworks to show the effectiveness of our approach in Sec. 5.2.

Benchmarks. Experiments are conducted on widely-used ETH/UCY dataset [13,27] and a large scale Stanford Drone Dataset [28] (SDD). The data pre-processing strategy follows previous work [35]. In the traditional prediction setting [6,35], the observation horizon is 8 frames (3.2 seconds with a frame rate of 2.5fps) and the prediction horizon is 12 frames (4.8 seconds). In our momentary prediction setting, the observation horizon is reduced to 2 frames (0.8 seconds) while the prediction horizon keeps the same. The prediction results are evaluated with Average Displacement Error (ADE) and Final Displacement Error (FDE). The number of prediction samples is 20.

Baselines. Experiments are conducted on the following prediction frameworks. Social GAN [6] is a classic GAN based prediction framework using a pooling module to model social interactions. Social-STGCNN [24], Trajectron++ [31] and SGCN [32] improves the modeling of social behaviors with various Graph Neural Networks. Next [19] takes a further step to exploit scene context and human action information for trajectory prediction. Recently, PCCSNet [35] proposes a new framework to solve the trajectory prediction problem in a classification and regression manner, and achieves state-of-the-art performance with only historical trajectory information.

5.1. Performance of Existing Frameworks

Quantitative results. We conduct experiments on six typical existing approaches to quantitatively demonstrate how their performance changes in the momentary prediction setting. Results in Tab. 2 indicate that all of these methods suffer from a decline of performance. However, the trend of performance degradation shows great differences between

Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG	SDD
Next [19]	0.77/1.81	0.32/0.62	0.61/1.31	0.39/0.82	0.34/0.76	0.49/1.06	-
+full MOE	**0.71/1.57**	**0.30/0.58**	**0.52/1.12**	**0.38/0.81**	**0.33/0.73**	**0.45/0.96**	-
Trajectron++ [31]	0.76/1.43	0.30/0.56	0.36/0.74	0.22/0.42	0.18/0.34	0.36/0.70	13.07/22.88
+full MOE	**0.64/1.12**	**0.20/0.33**	**0.33/0.62**	**0.22/0.42**	**0.17/0.32**	**0.31/0.56**	**11.71/19.54**
PCCSNet [35]	0.34/0.65	0.14/0.25	0.31/0.63	0.23/0.46	0.16/0.37	0.24/0.47	9.19/17.71
+full MOE	**0.31/0.57**	**0.13/0.21**	**0.25/0.53**	**0.20/0.41**	**0.14/0.31**	**0.20/0.41**	**8.40/16.08**

Table 3. Momentary prediction performance (ADE/FDE) of baseline methods after aiding with our proposed approach. *full* dentoes that MOE is first pre-trained with the soft pre-training mechanism.

Method	ETH/UCY		SDD	
	ADE	FDE	ADE	FDE
NEXT [19]	0.49	1.06	-	-
+MOE	0.47	1.00	-	-
+full MOE	**0.45**	**0.96**	-	-
PCCSNet [35]	0.24	0.47	9.19	17.71
+MOE	0.22	0.44	8.75	16.53
+MOE w/ CR	0.21	0.43	8.63	16.44
+MOE w/ MTC	0.21	0.42	8.51	16.28
+full MOE	**0.20**	**0.41**	**8.40**	**16.08**

Table 4. Ablation study for our approach. *CR* denotes context restoration and *MTC* denotes masked trajectory complement. The average results of ETH/UCY are reported.

approaches. For approaches without scene context modeling [6, 24, 31, 32, 35], their performance decreases dramatically after a reduction in observation time. Specifically, state-of-the-art approach PCCSNet [35] drops about 14.3%/11.9% on ETH/UCY. In comparison, the approach with scene context modeling [19] has a relatively low sensitivity of the reduction of observation. Its performance drops less, about 6.5%/6% on ETH/UCY. This proves that scene information can provide a large number of clues for future trajectory prediction, which can compensate for the lack of temporal information in momentary prediction.

Qualitative results. Fig. 2 gives qualitative comparisons of existing methods between traditional and momentary configurations. As the observation time reduces, all these models may make wrong predictions in both speed and direction aspects. To this end, it is necessary to propose additional modules to improve the momentary prediction performance of existing approaches.

5.2. Effectiveness of Proposed Approach

Main results. We implement the full MOE (with soft pre-training) as the new feature extractor on three typical prediction frameworks [19, 31, 35] to comprehensively show the effectiveness of our proposed approach. Results are in Tab. 3. With the aid of our approach, the momentary prediction performance of existing methods can be improved significantly. Specifically, for state-of-the-art PCCSNet [35], the ADE/FDE improve about 16.7%/12.8% on ETH/UCY benchmark and 8.6%/9.2% on SDD.

Mask Size		4	6	8	12
ETH/UCY	ADE	0.20	**0.20**	0.20	0.21
	FDE	0.41	**0.41**	0.42	0.43
SDD	ADE	8.44	**8.40**	8.47	8.63
	FDE	16.13	**16.08**	16.19	16.44

Table 5. Analysis of different mask sizes for masked trajectory complement on PCCSNet. The average results of ETH/UCY are reported.

Ablation study. We deliver comprehensive ablation studies in Tab. 4 to investigate the contribution of different components in our approach. i) We first compare the performance of NEXT before and after replacing its original feature extractor with MOE (w/o pre-training), according to line 1 and 2 of the NEXT part. Since the original feature extractor of NEXT integrates social and scene context at feature level, this comparison shows the superiority of the design of MOE which integrates the information at data level to avoid potential problems in feature alignment. ii) Major ablation studies are conducted on the state-of-the-art model PCCSNet. According to the results, the MOE can bring substantial improvement alone, and the performance can be further boosted benefiting from the soft pre-training mechanism which helps the MOE to learn better representations.

Mask size. Tab. 5 studies the sensitivity of the mask size for masked trajectory complement sub-task. When the size is in a suitable range, the performance keeps stable. As it goes larger, the sub-task becomes more difficult and it is much harder for MOE to learn good representations during pre-training.

Qualitative results. We qualitatively analyze the performance of baseline methods in the momentary prediction setting with the aid of our approach in Fig. 6. Results show that our approach can help existing prediction frameworks give accurate, social plausible and scene consistent momentary prediction results. Social behaviors such as encountering (row 4, col 2) and following (row 2, col 4) can be

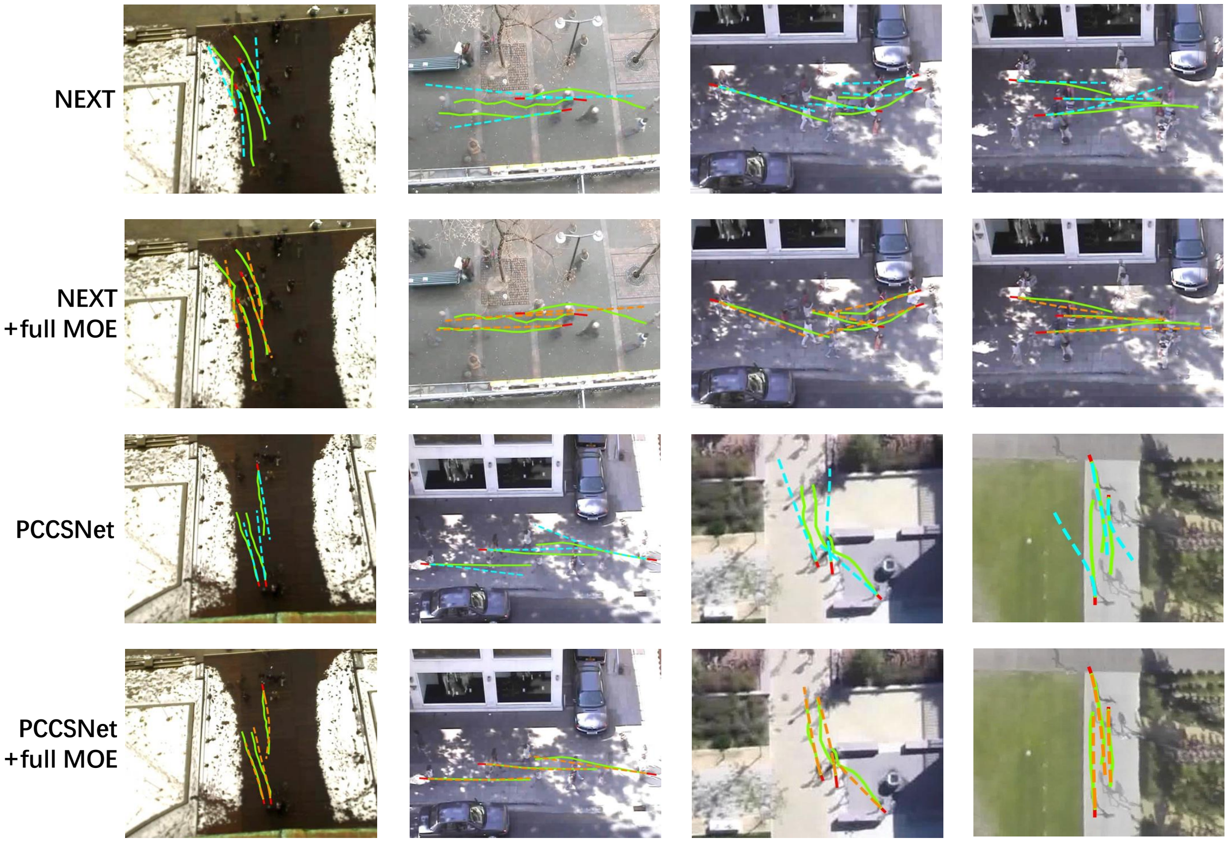

Figure 6. Qualitative comparison of baseline methods' performance on momentary prediction between w/ and w/o our approach. Red line denotes the observed trajectory, green line denotes the ground truth future trajectory, blue dashed line denotes predictions made without MOE and orange dashed line denotes predictions made with the aid of MOE. Zoom in for a clear view.

well handled. In a challenging case (row 2, col 2), although the deflection is still not predicted, the speed and direction of predicted movements are more accurate benefiting from our approach. Further, unwalkable areas can be successfully avoided when making predictions (row 4, col 3 and 4).

5.3. Limitations and Potential Negative Impact

In this paper, we propose an input formulation to integrate social and scene context at data level and design the MOE to better explore information for momentary prediction. Although the input formulation and MOE can handle social interactions and scene context, it is difficult to incorporate other information such as human skeletons into our current pipeline. We will list it as our future work. Since current trajectory prediction systems are usually data-driven, a potential negative impact of the systems is that they require expensive computational resources and large amount of high quality data, which could cost many financial and environmental resources. We will release our code and trained models to the community, as part of efforts to alleviate the repeated training of future work.

6. Conclusion

In this paper, we propose and study the task of momentary trajectory prediction. This task reduces the observation time to an extreme level and aims to resolve issues arising from tracking losses and sudden events. We propose the Momentary Observation feature Extractor to effectively exploit useful information implicit in momentary observations. A novel soft pre-training mechanism is further proposed to help MOE learn better representations, including two sub-tasks: masked trajectory complement and context restoration. We perform a rigorous study of existing methods and conduct exhaustive experiments to analyze the performance of our proposed approach on the momentary prediction setting. We hope this work will pave the way for more responsive, precise and robust prediction systems.

Acknowledgment We appreciate the support from National Natural Science Foundation of China (No.72192821, 72192820), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Qi Zhi Institute, and SHEITC (2018-RGZN-02046)

References

[1] agrimgupta92. Social gan. https://github.com/agrimgupta92/sgan, 2018. 6

[2] 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. 1, 2, 3

[3] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11533, 2020. 1

[4] 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 (ICCV)*, pages 9803–9812, October 2021. 2

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[6] 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. 2, 3, 6, 7

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[8] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 3

[9] Tsubasa Hirakawa, Takayoshi Yamashita, Toru Tamaki, and Hironobu Fujiyoshi. Survey on vision-based path prediction. In *International Conference on Distributed, Ambient, and Pervasive Interactions*, pages 48–64. Springer, 2018. 1

[10] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2375–2384, 2019. 2

[11] Chanho Kim, Fuxin Li, and James M Rehg. Multi-object tracking with neural gating using bilinear lstm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 200–215, 2018. 1

[12] 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*, 32:137–146, 2019. 2, 3, 4

[13] Laura Leal-Taixé, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, and Silvio Savarese. Learning an image-based motion context for multiple people tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3542–3549, 2014. 2, 3, 6

[14] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950. IEEE, 2019. 3

[15] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

[16] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2019. 1

[17] Yong-Lu Li, Liang Xu, Xijie Huang, Xinpeng Liu, Ze Ma, Mingyang Chen, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*, 2019. 1

[18] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019. 1

[19] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5725–5734, 2019. 2, 3, 4, 6, 7

[20] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[21] Cong Ma, Changshui Yang, Fan Yang, Yueqing Zhuang, Ziwei Zhang, Huizhu Jia, and Xiaodong Xie. Trajectory factory: Tracklet cleaving and re-connection by deep siamese bi-gru for multiple object tracking. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 1

[22] Srikanth Malla, Behzad Dariush, and Chiho Choi. Titan: Future forecast using action priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2020. 3

[23] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15233–15242, October 2021. 2

[24] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020. 3, 6, 7

[25] Bo Pang, Yizhuo Li, Yifan Zhang, Muchen Li, and Cewu Lu. Tubetk: Adopting tubes to track multi-object in a one-step training model. In *CVPR*, pages 6308–6318, 2020. 1

[26] Bo Pang, Yizhuo Li, Yifan Zhang, Jiajun Tang, Kaiwen Zha, Jiefeng Li, and Cewu Lu. Unsupervised representation for semantic segmentation by implicit cycle-attention contrastive learning. In *AAAI*, 2022. 3

[27] 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, 3, 6

[28] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016. 1, 2, 3, 6

[29] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 300–311, 2017. 1

[30] 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 Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019. 2, 3, 4

[31] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020. 1, 6, 7

[32] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8994–9003, 2021. 2, 6, 7

[33] StanfordASL. Trajectron++. https://github.com/StanfordASL/Trajectron-plus-plus/issues/40, 2020. 6

[34] Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–669, 2020. 3

[35] Jianhua Sun, Yuxuan Li, Hao-Shu Fang, and Cewu Lu. Three steps to multimodal trajectory prediction: Modality clustering, classification and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13250–13259, October 2021. 2, 6, 7

[36] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5

[38] 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 1–7. IEEE, 2018. 2

[39] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29:613–621, 2016. 3

[40] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021. 1

[41] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European Conference on Computer Vision*, pages 507–523. Springer, 2020. 2

[42] Zhishuai Zhang, Jiyang Gao, Junhua Mao, Yukai Liu, Dragomir Anguelov, and Congcong Li. Stinet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3