Traj-MAE: Masked Autoencoders for Trajectory Prediction

Hao Chen[1*] Jiaze Wang[1*] Kun Shao[4] Furui Liu[2] Jianye Hao[4,6] Chenyong Guan[5]
Guangyong Chen[2†] Pheng-Ann Heng[1,3]
[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]Zhejiang Lab [3]Institute of Medical Intelligence and XR, The Chinese University of Hong Kong
[4]Huawei Noah's Ark Lab [5]Gudsen Technology Co. Ltd [6]Tianjin University

Abstract

Trajectory prediction has been a crucial task in building a reliable autonomous driving system by anticipating possible dangers. One key issue is to generate consistent trajectory predictions without colliding. To overcome the challenge, we propose an efficient masked autoencoder for trajectory prediction (Traj-MAE) that better represents the complicated behaviors of agents in the driving environment. Specifically, our Traj-MAE employs diverse masking strategies to pre-train the trajectory encoder and map encoder, allowing for the capture of social and temporal information among agents while leveraging the effect of environment from multiple granularities. To address the catastrophic forgetting problem that arises when pre-training the network with multiple masking strategies, we introduce a continual pre-training framework, which can help Traj-MAE learn valuable and diverse information from various strategies efficiently. Our experimental results in both multi-agent and single-agent settings demonstrate that Traj-MAE achieves competitive results with state-of-the-art methods and significantly outperforms our baseline model. Project page: https://jiazewang.com/projects/trajmae.html.

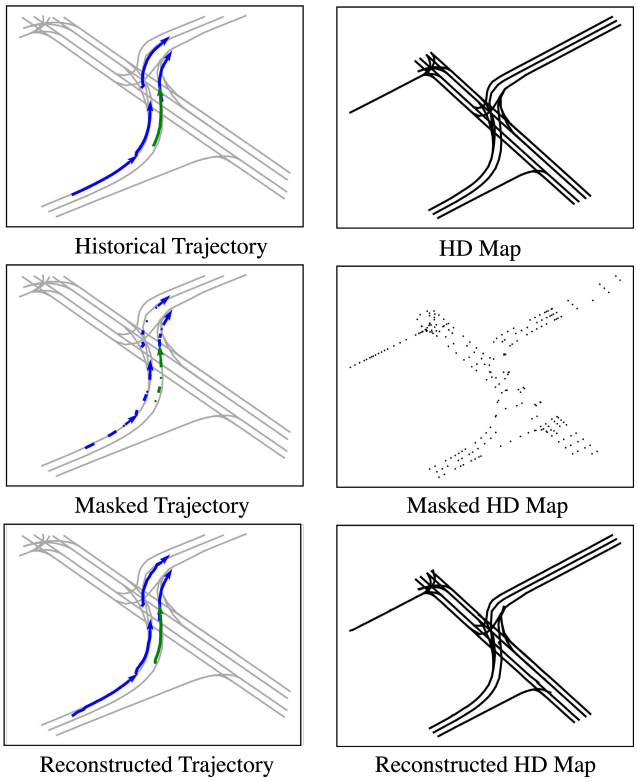


Figure 1: **An example of our masking and reconstruction strategy for trajectory prediction.** A portion of the historical trajectory and HD map is masked, the trajectory autoencoder and the map autoencoder are separately trained to recover the masked parts from the corrupted input. The green curve denotes the ego agent, the blue curves denote the surrounding agents, the same in the following figures in this paper.

1. Introduction

The goal of trajectory prediction is to predict the future trajectories of moving agents (e.g., *pedestrians and vehicles*), which is a crucial problem for building a safe, comfortable, and reliable autonomous driving system [32, 66, 37, 12, 51]. Many promising works [21, 7, 49, 26, 70, 60] have been proposed with great interest and demand from academia and industry. It has been demonstrated that modeling complex interactions between agents [47, 44, 46, 6, 27] is of great importance in trajectory prediction. On this basis, to address the colliding prediction problem and generate consistent trajectory predictions, it is essential to model social and temporal relations between agents and to have a global understanding of maps [2]. In this paper, we investigate this issue using self-supervised learning.

*Equal contribution
† Corresponding author: gychen@zhejianglab.com

Self-supervised learning aims to learn latent semantics from unlabeled data rather than building representations based on human annotations. Recent years have witnessed noteworthy advancements in the application of self-supervised learning to natural language processing [14, 62] and computer vision [57, 38, 4]. One of the most promising self-supervised methods is the masked autoencoders (MAE) [23] which achieve success in various tasks [39, 53, 73, 72, 22]. Furthermore, pre-train and fine-tune on the same small-scale datasets are also essential to learn a good representation [15]. Inspired by these works, we aim to explore the complex interactions between agents and the multiple granularities of maps using masked autoencoders. *How to design an efficient masked autoencoder to generate consistent trajectory predictions?* We attempt to answer the question from the following perspectives:

(i) The information density of trajectory and high definition (HD) maps differs significantly from that of images. While images are natural signals with high spatial redundancy, trajectories represent continuous temporal sequential signals with complex social interactions between agents, and HD maps contain highly structured information. Given the differences, models aimed at trajectory prediction require corresponding adjustments to capture informative features. Therefore, we investigate various masking strategies and suitable masking ratios for trajectories and HD maps. We develop both social and temporal masking to enable the trajectory encoder to capture information from diverse perspectives. We also study multiple granularities masking to enforce the map encoder to capture structural information from HD maps. Furthermore, we find that regardless of the masking strategy adopted, a high masking ratio (50% $\sim$ 60%) yields favorable results, which demands the encoders to acquire a holistic understanding of historical trajectories and HD maps.

(ii) The absence of an efficient framework for pre-training multiple strategies poses a challenge for effective multimodal trajectory prediction. While traditional multi-task learning from scratch [74] may struggle to converge due to the complex nature of this task, traditional continual learning methods [11, 40] are limited by their inability to train the network with multiple tasks without forgetting previously learned knowledge. To address this issue, we propose a novel approach that trains the new strategy simultaneously with the original strategies, utilizing previously learned parameters to initialize the network. Therefore, we ensure that our network can acquire new knowledge while retaining previously obtained knowledge.

Driven by the analysis, we present *Masked Trajectory Autoencoder (Traj-MAE)*, an efficient and practical framework for self-supervised trajectory prediction. As depicted in Figure 1, Traj-MAE leverages partial masking of the input trajectory and HD map, utilizing the trajectory encoder and map encoder to reconstruct the masked segments, respectively. Through employing diverse masking strategies to reconstruct missing parts of the input trajectory and HD map, the trajectory encoder and map encoder can acquire a comprehensive understanding of the latent semantics of the inputs from various perspectives. Moreover, we introduce a novel continual pre-training framework, which is a highly-efficient learning approach that trains the model with multiple strategies simultaneously, mitigating the issue of catastrophic forgetting.

Our core contributions are as follows:

- To our best knowledge, we are the first to present a neat and efficient masked trajectory autoencoder for self-supervised trajectory prediction.
- We explore different masking strategies which fully utilize MAE to exploit the latent semantics of historical trajectory and HD map. Meanwhile, a continual pre-training framework is proposed to efficiently train the model with multiple strategies.
- We conduct extensive experiments on the Argoverse, and INTERACTION for autonomous driving trajectory prediction, and the synthetic partition of TrajNet++ for pedestrian trajectory prediction. Our Traj-MAE achieves competitive results on these benchmarks and outperforms our baseline model by a notable margin.

2. Related Works

Trajectory Prediction is widely considered a sequence modeling task with many RNN-based methods [1, 71, 34, 8] proposed to model the trajectory pattern of agents' future locations, as RNN (e.g., LSTM [24]) have achieved remarkable success in sequence modeling. Due to the strong ability of Transformers [55, 59] to capture long-range dependencies, many transformer-based methods have emerged and flourished. STAR [65] is proposed to capture complex spatio-temporal interactions by interleaving between spatial and temporal Transformers. mmTransformer [32] is designed to hierarchically aggregate the past trajectories, the road information, and the social interaction. For predicting multi-agents future trajectories, AgentFormer [66] and AutoBots [20] have given solutions to model the time dimension and social dimension simultaneously. The enhancement of the encoder's ability to model information in both dimensions is an interesting and central focus of this work.

Self-supervised Learning has shown significant success in natural language processing and computer vision fields recently, especially the autoencoding method. Denoising autoencoders (DAE) [56] is a learning representation method that reconstructs original signals from corrupted inputs. BERT [14] can be seen as a development of DAE, which masks input tokens and trains the model to predict the miss-

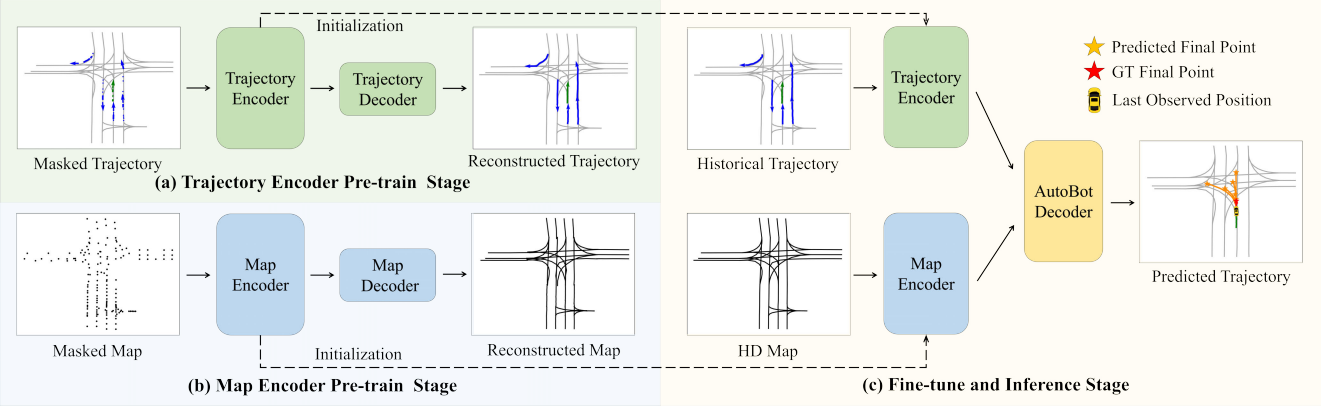


Figure 2: **Overview of Traj-MAE.** Traj-MAE is mainly composed of three stages: (a) Trajectory encoder pre-train stage with continual trajectory masking and reconstruction strategies. (b) Map encoder pre-train stage with continual map masking and reconstruction strategies. (c) Fine-tune and inference stage where the encoders are initialized by the pre-trained models' parameters.

ing content. With the Masked Language Modeling (MLM) task proposed in BERT, many MLM variants [62, 5] are proposed to improve the performance of transformer pre-training. Similarly, in computer vision, the autoencoding method often focuses on different pretext tasks for pre-training [38, 4, 23, 58]. One of the most popular methods is MAE [23], which randomly masks input patches and trains the model to recover masked patches in pixel space. Continuous progress [16, 39, 3] based on MAE has verified its effectiveness. Following the concept of MAE, our approach concentrate on utilizing MAE as a tool to pre-train model encoders with powerful feature extraction capability.

Continual Learning is a method to tackle the catastrophic forgetting problem that happens in sequentially learning samples of different input patterns. The methods can be roughly categorized into replay, regularization-based, and parameter isolation approaches [13]. With respect to replay methods [42, 43, 25, 10, 52, 31], previous task samples are replayed while learning a new task to alleviate forgetting. Instead, when learning new data, regularization-based methods [50, 41, 68, 29, 61] often introduce a regularization term in the loss function to consolidate previous knowledge. Parameter isolation methods [33, 48] dedicate different model parameters to each task to prevent any possible forgetting. In this work, we propose a continual pre-training framework to tackle the forgetting problem, in which way we are able to improve the generalization of model encoders by leveraging the specific information contained in the training samples of related masking strategies.

3. Approach

Our Traj-MAE is a sophisticated yet efficient self-supervised approach. Figure 2 provides an overview of the Traj-MAE framework. In this section, we begin by introducing our network backbone. We then delve into our analysis of the masking strategies for trajectory and HD-map reconstruction. Finally, we discuss how we incorporate Traj-MAE into our continual pre-training framework.

3.1. Network Backbone

In this work, We use Autobots [20] that has a transformer encoder-decoder architecture (detailed in supplementary material) as the baseline model to verify the effectiveness of the proposed method. Our Traj-MAE masks random parts from the input trajectory and HD map, then reconstructs the missing parts respectively. Following MAE [23] and VideoMAE [53], we adopt the asymmetric encoder-decoder design to reduce computation.

Traj-MAE Encoder. In Autobots, historical trajectories are encoded into context tensors, together with learnable seed parameters and map context, are passed to the decoder to predict future trajectories. Inspired by this design, we adopt the Autobots encoder as our trajectory encoder. However, in Autobots, the HD map is directly fed to the decoder, making it challenging for the model to capture the inherent information of the HD map. To address this limitation, we introduce a map encoder with a similar architecture to the trajectory encoder to better reconstruct the masked HD map. However, we observed that directly adding the map encoder to the Autobots results in little improvement (see supplementary material). Nevertheless, we found that pre-training the map encoder with our proposed masking and reconstruction strategy can further improve the accuracy, validating the effectiveness of our pre-training strategy.

Traj-MAE Decoder. The encoder in Traj-MAE processes only the unmasked parts of the input, while the decoder reconstructs the missing parts from the latent representation and mask tokens. Mask tokens are shared vectors that indi-

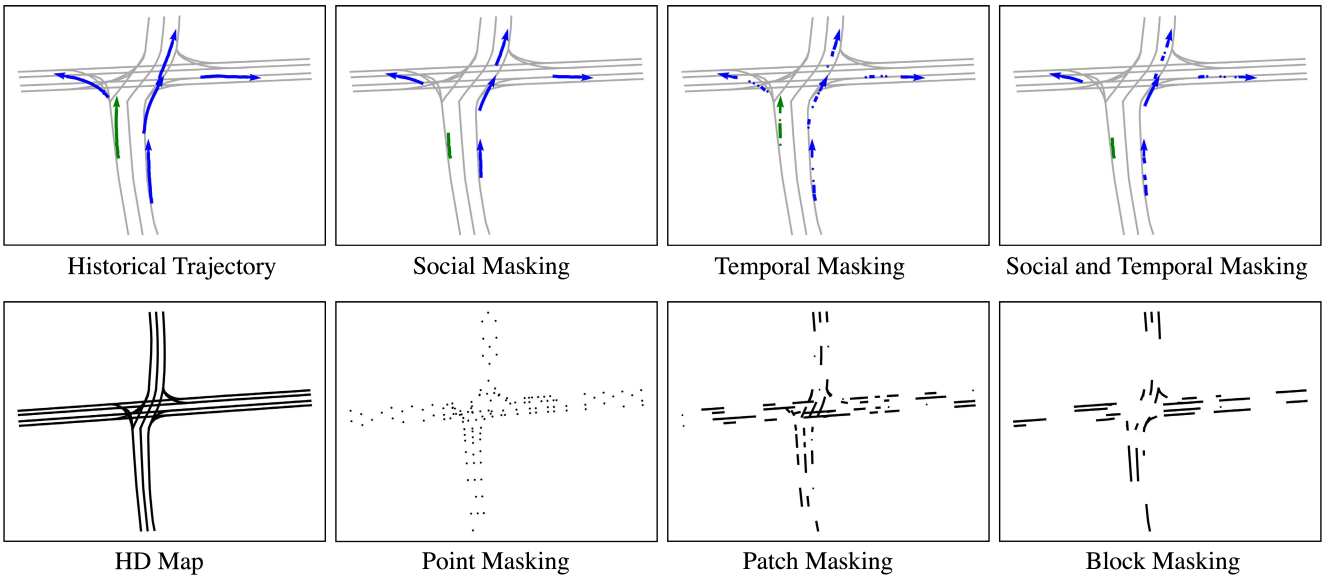


Figure 3: **Masking strategies for historical trajectory and HD map.** We show three masking strategies for the same input of trajectory and HD map, respectively. The leftmost column is input, the other three columns are different masking strategies performed on the input for historical trajectory (top) and HD map (bottom).

cate the presence of the missing parts that need to be predicted. Additionally, positional embeddings are added to all tokens to provide location information. Traj-MAE decoder is designed with Transformer blocks that are shallower than the encoder and are used solely during pre-training to perform the trajectory and map reconstruction strategies. This enables the decoder architecture to be flexible and independent of the encoder architecture. Pre-training with a lightweight decoder can notably reduce pre-training time.

3.2. Masking Strategy.

Different masking strategies determine the pretext task with different latent information that the network encoder can learn. To capture the social and temporal information from historical trajectories and multiple granularity representations from HD maps, we devise three masking strategies for the trajectory encoder and map encoder, respectively. Each strategy masks different scales and components of the input, with the goal of reconstructing the missing parts of the input.

Trajectory Masking Strategy. We introduce three effective masking and reconstruction strategies to enhance representation learned by the trajectory encoder. The three different strategies are illustrated in Figure 3.

Social Masking. Understanding the social relationships between agents is a fundamental concern when predicting trajectories. Social masking aims to reconstruct each agent's trajectory from surrounding agents. We mask the ego-agent's trajectory in the last observed time and nearby agents' trajectories at the beginning of the observed time. By utilizing this strategy, the network is able to leverage continuous trajectories that are observable to reconstruct trajectories that are unobservable for other agents. This approach improves the network's ability to model interactions among agents and generate socially consistent predictions.

Temporal Masking. Temporal masking strategy endeavors to reconstruct the trajectories of all agents that have been randomly masked in the time domain. By inferring the positions of agents at various temporal intervals based on their positions at specific times, our model is able to efficiently capture the temporal dynamics of historical trajectories.

Social and Temporal Masking. We also construct a masking strategy that reconstructs the historical trajectories in both temporal and social aspects. Specifically, half of the trajectories of surrounding agents are randomly masked in the time dimension, and half of that are masked in the last observed time, while the trajectory of the ego-agent in the last observed time is masked. By reconstructing these trajectories simultaneously, the social and temporal masking can further enhance the trajectory encoder to obtain the temporal and social information.

Map Masking Strategy. The input vector of the map encoder is based on the VectorNet approach [17], which selects a starting point and direction, and uniformly samples key points from the splines at the same spatial distance. To capture the latent semantics of the HD map, we propose a mask and reconstruction strategy that operates at multiple granularities.

Point Masking. Our point masking strategy randomly samples and masks key points from the input map vector

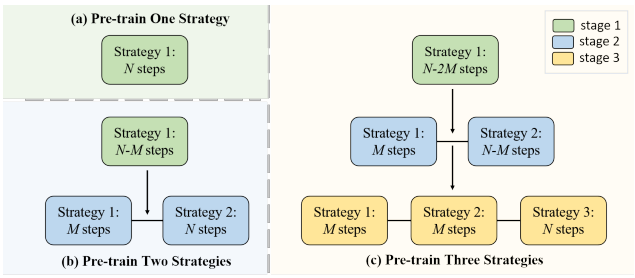


Figure 4: **Illustration of our continual pre-training framework.** For each strategy, we guarantee it has the same training steps throughout the pre-training process.

and reconstructs the whole map by predicting the missing points. This fine-grained learning approach is shown to have the best effect in our experiments.

Patch Masking. Patch masking refers to masking the map vector at the patch level, where patches are randomly sampled from the map vector. Patch masking and reconstruction are more complex than point masking due to the unknown surrounding points of the middle point. Consequently, the map encoder must infer the map architecture from longer distance points, allowing it to learn the long-term distance relation of the map.

Block Masking. The block masking approach removes large blocks from the input and predicts the missing parts in the input HD map, enabling the map encoder to have a better global understanding of the whole map. Block masking and patch masking differ in their granularities. For each polyline in HD map, patch masking masking several consecutive line segments, whereas block masking only masks a single continuous line segment.

Reconstruction Targets. Traj-MAE reconstructs the input by predicting the coordinates of location points in every masked part. Our loss function comprises the Huber loss between the reconstructed trajectory and original trajectory to pre-train the trajectory encoder, and the Huber loss between the reconstructed HD map and original HD map to pre-train the map encoder. Similar to MAE [23] and Point-MAE [39], we compute the loss only on masked parts.

3.3. Continual Pre-training Framework

We propose a continual pre-training framework to learn diverse information from multiple masking strategies efficiently.

Traditional continual learning method [11, 40] trains the model on only one strategy at each stage, and the model may suffer from the catastrophic forgetting problem, forgetting the previously learned knowledge. Multi-task learning method [74] trains the model with multiple tasks at the same time. In this way, we find it hard for our model to converge after many steps, and the performance could be even worse than the model pre-trained on a single strategy.

Method	minADE	minFDE	MR
TNT [75]	0.94	1.54	0.13
LaneRCNN [67]	0.90	1.45	0.12
mmTransformer [32]	0.84	1.34	0.15
GOHOME [19]	0.94	1.45	**0.11**
TPCN [63]	0.87	1.38	0.16
Scene Transformer [37]	0.80	1.23	0.13
MultiPath++ [54]	0.79	1.21	0.13
DenseTNT [21]	0.88	1.28	0.13
HiVT [76]	**0.77**	1.17	0.13
Wayformer [36]	**0.77**	1.16	0.12
DCMS [64]	**0.77**	**1.14**	**0.11**
GANet [60]	0.81	1.16	0.12
DSP [70]	0.82	1.22	0.13
Autobot-Ego [20]	0.89	1.41	0.16
Traj-MAE	0.81 ↓ 9%	1.25 ↓ 11%	0.137 ↓ 14%

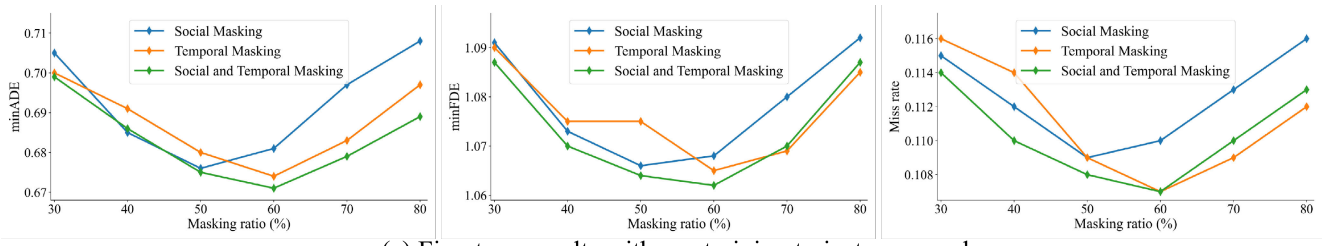
Table 1: **Comparison with state-of-the-art methods on the Argoverse test set**.

Therefore, there are two issues to solve. The first issue is to learn the strategies without forgetting the previous knowledge. The second issue is how to pre-train the model with multiple strategies efficiently. To overcome these issues, we propose a practical continual pre-training framework that enables model training with satisfactory efficiency and alleviates catastrophic forgetting. The core idea is to train the model over multiple stages with cross-stage parameter sharing. That means, for each pre-training stage except the first one, we use the parameters learned in previous stage to initialize the model. Then we pre-train a new strategy together with the previous strategies simultaneously in this stage. As shown in Figure 4, the number of pre-training stages is equal to that of strategies. Our framework allocates each strategy a fixed number of pre-training steps (N) and distributes the steps for each strategy over the whole pre-training stages. M is the steps to pre-train the previous strategy in the new stage. To pre-train different strategies in a single stage, we randomly select training samples from each strategy. In this way, the effectiveness of our continual pre-training framework can be guaranteed.

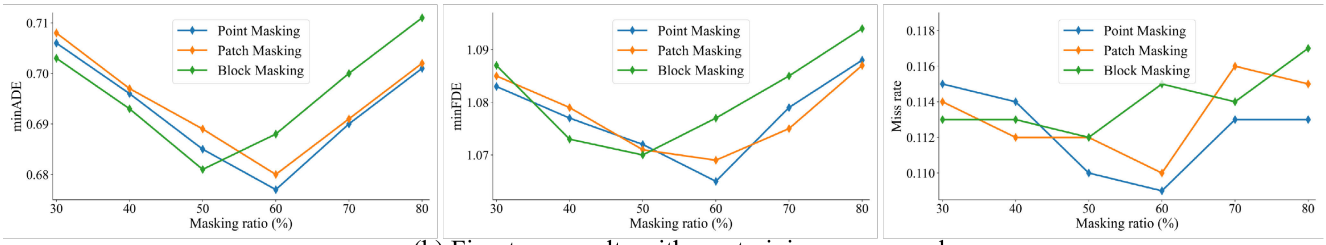
4. Experiments

4.1. Experimental Setup.

Datasets. Argoverse motion forecasting dataset [9] provides 333K real-world driving sequences, which are sampled at 10Hz, with 2 seconds history and 3 seconds future. The whole dataset is split into train, validation and test sets, with 205942, 39472, and 78143 sequences, respectively. The Interaction dataset [69] consists of various highly interactive driving situations, and each trajectory has 1 second history and 3 seconds future sampled at 10Hz. In the multi-agent prediction track, the target is to predict multiple target agents' coordinates and yaw jointly. The synthetic partition of the TrajNet++ dataset [45] has 54513 scenes, which



(a) Fine-tune results with pre-training trajectory encoder



(b) Fine-tune results with pre-training map encoder

Figure 5: **Ablation study on masking strategy with different ratios.**

Method	MinJointADE	MinJointFDE	MinJointMR	Cross CollisionRate	Ego CollisionRate	Consistent MinJointMR
DenseTNT [21]	0.412	1.129	0.224	**0**	0.014	0.224
HDGT [28]	**0.303**	**0.958**	0.194	0.163	0.005	0.236
THOMAS [18]	0.416	0.968	**0.179**	0.128	0.011	0.252
ReCoG2 [35]	0.467	1.160	0.238	0.069	0.011	0.268
L-GCN	0.393	1.249	0.284	0.060	**0.004**	0.297
MoliNet	0.729	2.554	0.444	0.075	0.042	0.473
HGT-Joint	0.307	1.056	0.186	0.016	0.005	0.190
AutoBot [20]	0.312	1.015	0.193	0.043	0.010	0.207
Traj-MAE	0.306 ↓ 2%	0.966 ↓ 5%	0.183 ↓ 5%	0.021 ↓ 51%	0.006 ↓ 40%	**0.188** ↓ 9%

Table 2: **Comparison with state-of-the-art methods on the INTERACTION Multi-Agent Track**.

Model	Ego minADE	Scene minADE	Scene minFDE
Linear [20]	0.439	0.409	0.897
AutoBot-AS [20]	0.196	0.316	0.632
AutoBot-Ego [20]	0.098	0.214	0.431
AutoBot [20]	0.095	0.128	0.234
Traj-MAE	**0.074** ↓ 22%	**0.093** ↓ 27%	**0.181** ↓ 23%

Table 3: **Results on TrajNet++ for a multi-agent forecasting scenario**.

is specifically designed to have a high level of interactions [30]. Given the state of all agents of the past 9 timesteps, the goal is to predict the next 12 timesteps for all agents. Due to TrajNet++ does no have HD map, we pre-train and fine-tune the trajectory encoder only.

4.2. Experimental Results.

In this subsection, we pre-train and fine-tune the model on the same benchmark to validate the benefit brought by our proposed self-supervised learning method. We forecast 6 future trajectories on all datasets. The meaning of the used metrics is presented in the supplementary material and the lower metrics indicate better performance.

Results on Argoverse. The trajectory prediction results on Argoverse are reported in Table 1. On the Argoverse test set, our Traj-MAE decreases the minADE by 9%, minFDE by 11% and MR by 14%, respectively. Although Traj-MAE does not achieve state-of-the-art, the performance demonstrates that our Traj-MAE significantly improves our baseline model (Autobot-Ego). Moreover, it uses much less computation, which is an effective and GPU-friendly pre-training mechanism that only pre-trained on a single GPU (V100) in under *48h*. Additionally, the performance improvement achieved by Traj-MAE is also demonstrated on the validation set, as shown in 4.3.

Results on INTERACTION. In Table 2, we evaluate our method on the INTERACTION multi-agent track and achieve state-of-the-art performance on this benchmark.

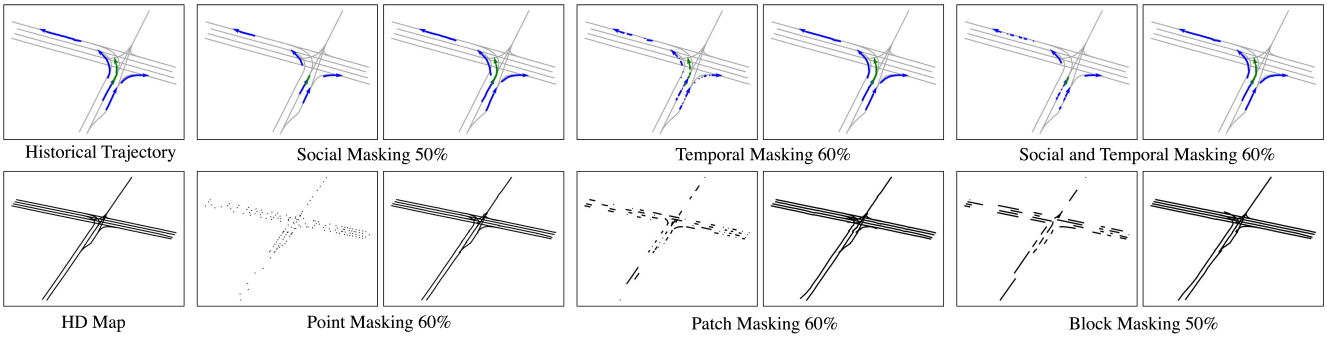


Figure 6: **Reconstruction results on historical trajectory and HD map with different masking strategies.** The leftmost side is input, and the masked input (left) and reconstruction (right) with different masking strategies are shown in the following columns.

	minADE	minFDE	MR
S	0.676	1.066	0.109
T	0.674	1.065	0.107
ST	0.671	1.062	0.107
S $\rightarrow$ T	0.642	1.042	0.103
T $\rightarrow$ S	0.651	1.051	0.105
S $\rightarrow$ T$\rightarrow$ ST	**0.621**	**1.027**	**0.099**
ST $\rightarrow$ T $\rightarrow$ S	0.636	1.038	0.101

(a) Continual pre-training for **trajectory reconstruction**.

	minADE	minFDE	MR
Po	0.677	1.065	0.109
Pa	0.680	1.069	0.110
B	0.681	1.070	0.112
Po $\rightarrow$ Pa	0.649	1.047	0.106
Pa $\rightarrow$ Po	0.651	1.053	0.108
Po $\rightarrow$ Pa $\rightarrow$ B	**0.627**	**1.033**	**0.102**
B $\rightarrow$ Pa $\rightarrow$ Po	0.641	1.046	0.104

(b) Continual pre-training for **HD map reconstruction**.

Table 4: **Ablation study on continual pre-training framework**. We show the results using different pre-train strategies with their best masking ratio. Note that 'S', 'T', 'ST' represent social masking, temporal masking, social and temporal masking strategy, respectively. 'Po', 'Pa', 'B' represent point masking, patch masking, block masking strategy, respectively.

The table shows that Traj-MAE outperforms all other approaches in terms of Consistent MinJointMR, the ranking metric. This metric encourages the model to make consistent predictions, thus the best result on this metric indicates that our Traj-MAE can well capture multi-agents' interaction. As for other metrics, our model still has competitive results. Especially, the reductions of 51% and 40% in CrossCollisionRate and EgoCollisionRate demonstrate that our Traj-MAE can better utilize social information and avoid collisions between agents.

Results on TrajNet++. To further demonstrate that our Traj-MAE has the ability to capture the social and temporal information for multi-agent trajectory prediction. We evaluate our method on the synthetic partition of TrajNet++ dataset. We compare our model with linear extrapolation (Linear), our baseline Autobot, and its variants: AutoBot-AntiSocial (AutoBot-AS) and AutoBot-Ego. The experiment results of agent-level metric and scene-level metrics defined by [8] are reported in Table 3. Our Traj-MAE achieves large reductions of 22%, 27%, 23% with respect to Ego Agent's minADE, Scene-level minADE, and Scene-level minFDE respectively.

4.3. Ablation Studies.

To investigate the properties of our method, we perform in-depth ablation studies on the Argoverse validation set. For these experiments, all models share the same experiment settings and architecture. We evaluate our models on minADE, minFDE, and miss rate (MR) of the predicted 6 trajectories. Our baseline model (Autobot-Ego with Map Encoder) achieves **0.732**, **1.096** and **0.119** on **minADE**, **minFDE** and **MR**, respectively. When fine-tuning the trajectory encoder with the pre-trained model, the map encoder is trained from scratch and vice versa.

Trajectory Masking Strategy. In Figure 5(a), we compare the performance of different trajectory masking strategies and ratios. As the masking ratio increases from 30% to a threshold of 50-60%, the performance of all three strategies improves simultaneously. However, the performance degrades as the masking ratio increases beyond the threshold to 80%. Furthermore, our experiments show that social and temporal masking outperforms the other two strategies, particularly in terms of minADE and minFDE. This suggests that incorporating social and temporal information is crucial for effective trajectory prediction. We attribute this to the fact that social and temporal masking encourages the

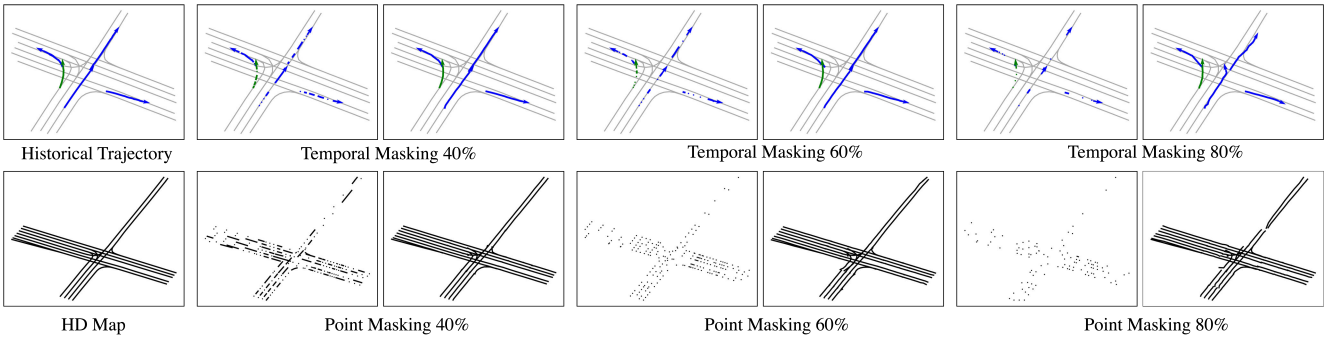


Figure 7: **Reconstruction results on historical trajectory and HD map with different masking ratios.** The leftmost side is input, and the masked input (left) and reconstruction (right) with different masking ratios are shown in the following columns.

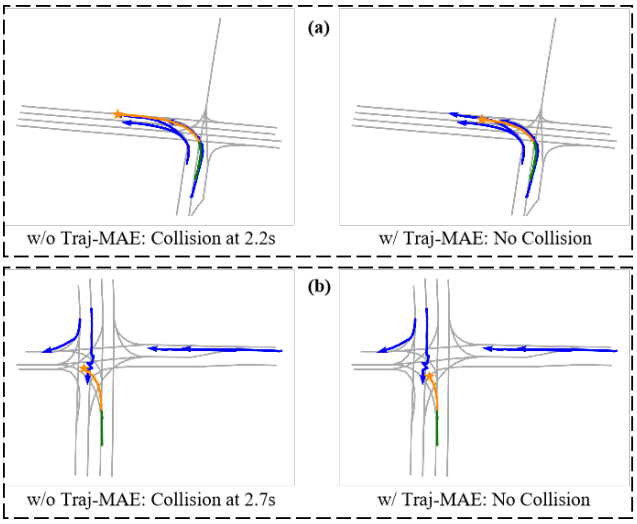


Figure 8: **Examples of collision avoidance using Traj-MAE.** The result in the left side is predicted by our baseline model and the right side is the future motion predicted using Traj-MAE. The predicted trajectory is denoted by orange curve.

trajectory encoder to capture more essential temporal and social information, thereby making Traj-MAE a challenging self-supervised trajectory prediction framework.

Map Masking Strategy. Figure 5(b) shows the influence of masking strategies with different masking ratios for HD map. Point masking works best for pre-training our map encoder. Besides, point masking and patch masking allow for a higher masking ratio (60%) compared to block masking (50%), which can provide a more significant speedup benefit. We suppose that with the increase of connected masked parts, the mask and reconstruct task is more challenging, and the optimal mask ratio is smaller.

Continual Pre-training Strategy. Table 4 shows the effectiveness of our Traj-MAE. First, we find that our continual pre-training framework further improves the network performance than pre-training on a single strategy. With continual pre-training, the trajectory encoder is better equipped to capture social and temporal information, while the map encoder excels at capturing structured information from multiple granularities. Moreover, our experiments reveal that the sequence of pre-training strategies can also impact model performance. For trajectory encoder pre-training, the best sequence is *social masking* $\rightarrow$ *temporal masking* $\rightarrow$ *social and temporal masking*. For map encoder pre-training, the best sequence is *point masking* $\rightarrow$ *patch masking* $\rightarrow$ *block masking*. By integrating our pre-trained trajectory and map encoders, we achieve impressive results on the Argoverse validation set, with a **minADE** of **0.604**, a **minFDE** of **1.003**, and an **MR** of **0.092**.

Qualitative Analysis. We visualize the reconstruction results on the Argoverse validation set for qualitative analysis, where Figure 6 illustrates the different masking strategies. Although social masking is challenging for trajectory reconstruction and block masking is difficult for map reconstruction, adopting a moderate masking ratio with these strategies can still yield satisfactory reconstruction results. We compare the different temporal masking ratios and point masking ratios in Figure 7. Our Traj-MAE can produce satisfying reconstructed trajectories and HD maps even under a high masking ratio (e.g., 80%), which indicates that our Traj-MAE can learn useful high-level representations. Furthermore, in Figure 8, we present two examples to demonstrate how Traj-MAE can effectively reduce the occurrence of collisions. Specifically, in a scene where our baseline model predicts a collision, Traj-MAE leverages the interaction relations among agents' historical trajectories to avoid collision happening.

5. Conclusion

This paper proposes a novel masked trajectory autoencoder (Traj-MAE) for self-supervised trajectory prediction learning. Our Traj-MAE incorporates diverse masking

strategies that facilitate the trajectory encoder learning the social and temporal information and map encoder capturing structural information with multiple granularities. We also propose a continual pre-training framework that enables efficient pre-training of multiple strategies. Experimental results show that our Traj-MAE produces impressive results on various challenging datasets in both multiple-agent and single-agent settings. We hope that this work will inspire further investigation into self-supervised learning for trajectory prediction.

Acknowledgement

This work was supported by the National Key R&D Program of China (2022YFE0200700), the National Natural Science Foundation of China (Project No. 62006219), the Natural Science Foundation of Guangdong Province (2022A1515011579), in part by the InnoHK Clusters via Hong Kong Center for Logistics Robotics and by the Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems.

References

[1] 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. 2

[2] Caio Azevedo, Thomas Gilles, Stefano Sabatini, and Dzmitry Tsishkou. Exploiting map information for self-supervised learning in motion forecasting. *arXiv preprint arXiv:2210.04672*, 2022. 1

[3] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022. 3

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2, 3

[5] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR, 2020. 3

[6] Apratim Bhattacharyya, Daniel Olmeda Reino, Mario Fritz, and Bernt Schiele. Euro-pvi: Pedestrian vehicle interactions in dense urban centers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6408–6417, 2021. 1

[7] Yulong Cao, Chaowei Xiao, Anima Anandkumar, Danfei Xu, and Marco Pavone. Advdo: Realistic adversarial attacks for trajectory prediction. In *European Conference on Computer Vision*, pages 36–52. Springer, 2022. 1

[8] Sergio Casas, Cole Gulino, Simon Suo, Katie Luo, Renjie Liao, and Raquel Urtasun. Implicit latent variable model for scene-consistent motion forecasting. In *European Conference on Computer Vision*, pages 624–641. Springer, 2020. 2, 7

[9] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019. 5

[10] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019. 3

[11] Zhiyuan Chen and Bing Liu. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–207, 2018. 2, 5

[12] Dooseop Choi and KyoungWook Min. Hierarchical latent structure for multi-modal vehicle trajectory forecasting. In *European Conference on Computer Vision*, pages 129–145. Springer, 2022. 1

[13] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021. 3

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[15] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021. 2

[16] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 3

[17] 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. 4

[18] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Thomas: Trajectory heatmap output with learned multi-agent sampling. *arXiv preprint arXiv:2110.06607*, 2021. 6

[19] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Gohome: Graph-oriented heatmap output for future motion estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 9107–9114. IEEE, 2022. 5

[20] Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. In *In-*

ternational Conference on Learning Representations, 2022. 2, 3, 5, 6

[21] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15303–15312, 2021. 1, 5, 6

[22] Ziyu Guo, Xianzhi Li, and Pheng Ann Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. *arXiv preprint arXiv:2302.14007*, 2023. 2

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3, 5

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[25] David Isele and Akansel Cosgun. Selective experience replay for lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

[26] Xiaosong Jia, Liting Sun, Masayoshi Tomizuka, and Wei Zhan. Ide-net: Interactive driving event and pattern extraction from human data. *IEEE Robotics and Automation Letters*, 6(2):3065–3072, 2021. 1

[27] Xiaosong Jia, Liting Sun, Hang Zhao, Masayoshi Tomizuka, and Wei Zhan. Multi-agent trajectory prediction by combining egocentric and allocentric views. In *Conference on Robot Learning*, pages 1434–1443. PMLR, 2022. 1

[28] Xiaosong Jia, Penghao Wu, Li Chen, Hongyang Li, Yu Liu, and Junchi Yan. Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding. *arXiv preprint arXiv:2205.09753*, 2022. 6

[29] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 3

[30] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):7386–7400, 2021. 6

[31] Lin Li, Chao Chen, Lei Pan, Jun Zhang, and Yang Xiang. Sigd: A cross-session dataset for ppg-based user authentication in different demographic groups. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023. 3

[32] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7577–7586, 2021. 1, 2, 5

[33] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018. 3

[34] Kaouther Messaoud, Nachiket Deo, Mohan M Trivedi, and Fawzi Nashashibi. Multi-head attention with joint agent-map representation for trajectory prediction in autonomous driving. *arXiv preprint arXiv:2005.02545*, 2020. 2

[35] Xiaoyu Mo, Yang Xing, and Chen Lv. Recog: A deep learning framework with heterogeneous graph for interaction-aware trajectory prediction. *arXiv preprint arXiv:2012.05032*, 2020. 6

[36] 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. 5

[37] Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified architecture for predicting multiple agent trajectories. *arXiv preprint arXiv:2106.08417*, 2021. 1, 5

[38] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2, 3

[39] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. *ECCV*, 2022. 2, 3, 5

[40] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. 2, 5

[41] Amal Rannen, Rahaf Aljundi, Matthew B Blaschko, and Tinne Tuytelaars. Encoder based lifelong learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1320–1328, 2017. 3

[42] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 3

[43] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[44] Saeed Saadatnejad, Mohammadhossein Bahari, Pedram Khorsandi, Mohammad Saneian, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Are socially-aware trajectory prediction models really socially-aware? *Transportation research part C: emerging technologies*, 141:103705, 2022. 1

[45] Amir Sadeghian, Vineet Kosaraju, Agrim Gupta, Silvio Savarese, and Alexandre Alahi. Trajnet: Towards a benchmark for human trajectory prediction. *arXiv preprint*, 2018. 5

[46] 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. 1

[47] 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

[48] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR, 2018. 3

[49] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *arXiv preprint arXiv:2209.13508*, 2022. 1

[50] Daniel L Silver and Robert E Mercer. The task rehearsal method of life-long learning: Overcoming impoverished data. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 90–101. Springer, 2002. 3

[51] Qiao Sun, Xin Huang, Junru Gu, Brian C Williams, and Hang Zhao. M2i: From factored marginal trajectory prediction to interactive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6543–6552, 2022. 1

[52] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975, 2020. 3

[53] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 2, 3

[54] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022. 5

[55] 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. 2

[56] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 2

[57] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[58] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4807–4814. IEEE, 2021. 3

[59] Jiaze Wang, Xiaojiang Peng, and Yu Qiao. Cascade multi-head attention networks for action recognition. *Computer Vision and Image Understanding*, 192:102898, 2020. 2

[60] Mingkun Wang, Xinge Zhu, Changqian Yu, Wei Li, Yuexin Ma, Ruochun Jin, Xiaoguang Ren, Dongchun Ren, Mingxu Wang, and Wenjing Yang. Ganet: Goal area network for motion forecasting. *arXiv preprint arXiv:2209.09723*, 2022. 1, 5

[61] Yi Wang, Jiaze Wang, Jinpeng Li, Zixu Zhao, Guangyong Chen, Anfeng Liu, and Pheng-Ann Heng. Pointpatchmix: Point cloud mixing with patch scoring. *arXiv preprint arXiv:2303.06678*, 2023. 3

[62] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 2, 3

[63] Maosheng Ye, Tongyi Cao, and Qifeng Chen. Tpcn: Temporal point cloud networks for motion forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11318–11327, 2021. 5

[64] Maosheng Ye, Jiamiao Xu, Xunnong Xu, Tongyi Cao, and Qifeng Chen. Dcms: Motion forecasting with dual consistency and multi-pseudo-target supervision. *arXiv preprint arXiv:2204.05859*, 2022. 5

[65] 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

[66] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021. 1, 2

[67] Wenyuan Zeng, Ming Liang, Renjie Liao, and Raquel Urtasun. Lanercnn: Distributed representations for graph-centric motion forecasting. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 532–539. IEEE, 2021. 5

[68] Chen Zeno, Itay Golan, Elad Hoffer, and Daniel Soudry. Task agnostic continual learning using online variational bayes. *arXiv preprint arXiv:1803.10123*, 2018. 3

[69] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kummerle, Hendrik Konigshof, Christoph Stiller, Arnaud de La Fortelle, et al. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088*, 2019. 5

[70] Lu Zhang, Peiliang Li, Jing Chen, and Shaojie Shen. Trajectory prediction with graph-based dual-scale context fusion. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11374–11381. IEEE, 2022. 1, 5

[71] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12085–12094, 2019. 2

[72] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35:27061–27074, 2022. 2

[73] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21769–21780, 2023. 2

[74] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018. 2, 5

[75] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Benjamin Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. *arXiv preprint arXiv:2008.08294*, 2020. 5

[76] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022. 5