DySeT: a Dynamic Masked Self-distillation Approach for Robust Trajectory Prediction

Mozhgan Pourkeshavarz, Junrui Zhang, and Amir Rasouli

Noah's Ark Lab, Huawei, Canada
firstname.lastname@huawei.com

Abstract. The lack of generalization capability of behavior prediction models for autonomous vehicles is a crucial concern for safe motion planning. One way to address this is via self-supervised pre-training through masked trajectory prediction. However, the existing models rely on uniform random sampling of tokens, which is sub-optimal because it implies that all components of driving scenes are equally informative. In this paper, to enable more robust representation learning, we introduce a dynamic masked self-distillation approach to identify and utilize informative aspects of the scenes, particularly those corresponding to complex driving behaviors, such as overtaking. Specifically, for targeted sampling, we propose a dynamic method that prioritizes tokens, such as trajectory or lane segments, based on their informativeness. The latter is determined via an auxiliary network that estimates token distributions. Through sampler optimization, more informative tokens are rewarded and selected as visible based on the policy gradient algorithm adopted from reinforcement learning. In addition, we propose a masked self-distillation approach to transfer knowledge from fully visible to masked scene representations. The distillation process not only enriches the semantic information within the visible token set but also progressively refines the sampling process. Further, we use an integrated training regime to enhance the model's ability to learn meaningful representations from informative tokens. Our extensive evaluation on two large-scale trajectory prediction datasets demonstrates the superior performance of the proposed method and its improved prediction robustness across different scenarios.

1 Introduction

Trajectory prediction, as one of the crucial components for autonomous driving, involves forecasting future positions of road users based on their observed behavior and the environment's layout. Recently, learning-based approaches, in particular supervised learning methods [17, 33, 49, 55, 65, 66, 83], have dominated the trajectory prediction domain, setting new benchmarks on large-scale datasets [12, 70]. However, the applicability of these methods to real-world scenarios is still limited [4, 13, 56]. To be effective in practice, these methods not only need to be accurate but must also be robust and adaptable to various conditions, ensuring their effectiveness across different environments.

Self-supervised learning (SSL) has become increasingly popular as a way of addressing the aforementioned challenges. Most recently, building on the success of masked language modeling, *masked prediction* is considered a key strategy in SSL to learn generalizable and adaptable representations. This technique involves transforming input data into tokens, masking a portion of them, and then pre-training the model using the visible (unmasked) tokens to predict the missing (masked) ones. This dual-phase learning procedure not only enhances the model's performance but also increases its robustness and reduces overfitting risk in downstream applications [29]. Such characteristics are highly desirable for trajectory prediction, which relies on large-scale annotated data and manually crafted HD maps which are very costly to produce.

To develop a generalizable prediction model, some methods have utilized masked prediction within mask autoencoder (MAE) architectures, using a *uniform random masking* strategy [14,15]. However, since not all tokens are equally informative, assuming a uniform probability distribution over them for selecting visible tokens is sub-optimal [21, 41, 76]. In other words, random masking often leads to selecting visible tokens from less informative parts of the input (driving scene), such as simpler cruising trajectories rather than more complex overtaking maneuvers, or simple straight lanes instead of intricate curved ones. Hence, during reconstruction, this approach reduces the inclusion of informative components of the scene, such as those corresponding to more complex driving behaviors. This leads to less accurate reconstructions and hampers the model's ability to learn meaningful representations.

Motivated by this, we propose a dynamic masked self-distillation approach termed **DySeT**, which is a novel self-supervised pre-training framework that adopts a dynamic masking mechanism and a self-distillation approach within the MAE architecture. Our key insight is to prioritize learning from behaviorally rich segments of driving scenes, refining the model's focus towards scene segments that are more critical for understanding complex driving behaviors, and, thereby, enhancing behavior representation learning.

To this end, we first design a dynamic sampling method that optimizes both an MAE and a token sampling network based on *token informativeness*. Our approach, unlike uniform random sampling, utilizes an auxiliary network to estimate the token distribution for sampling. To address non-differentiability of the sampling, we employ an auxiliary objective grounded in REINFORCE algorithm [69] adopted from reinforcement learning. Furthermore, we propose a novel masked self-distillation approach to distill knowledge (i.e. representation) from a fully visible scene to the representation predicted from a masked scene, ensuring that the sampled visible token set is semantically rich. The integrated training of the sampling network and the MAE via the masked self-distillation approach gradually guides the model towards sampling informative visible tokens and, thus, improves representation learning.

In summary, our **main contributions** are as follows: 1) We propose a dynamic end-to-end trainable token sampling strategy for masked trajectory prediction, focusing on selecting fewer yet highly informative visible tokens to facil-

itate behavior representation learning; 2) To ensure a semantically rich selection of visible tokens, we propose a masked self-distillation approach designed to transfer knowledge from a complete visible scene to its masked counterpart's predicted representation; 3) We conduct extensive experimental evaluations on common benchmark datasets and demonstrate our model's effectiveness in learning transferable and adaptable scene representations; 4) We further conduct comprehensive studies on the robustness of our approach against various contextual perturbations followed by ablation studies to underscore the significance of our proposed components in enhancing overall performance.

2 Related Work

Trajectory Prediction. For safe motion planning, autonomous driving systems rely on trajectory prediction methods, with a broad range of architectures [2, 6, 11, 38, 52, 55, 59, 83]. Current approaches utilize various representations and models, such as CNNs with rasterized images [11, 23, 59], and GNNs [22, 27, 50, 55, 57, 78] or transformers [25, 34, 37, 47, 84] in conjunction with vectorized representations. The goal of these methods is to enhance map encoding and interaction modeling. However, they require complex architectures with extensive parameterization, leading to increased overfitting risks and potentially limited generalizability.

SSL in Prediction. There exist a few works that leverage SSL techniques to improve prediction model's generalizability. The method in [71] generates additional map patches for pre-training maps (limited rasterized maps) and trajectory encoders using contrastive learning. The authors of [8] demonstrate that well-designed pretext tasks could improve feature richness without extra data, although it would not combine tasks for better representation learning. In [55] a set of SSL tasks on HD maps, managed by a meta-learning technique, is proposed to predict diverse and admissible trajectories. Inspired by masked language modeling, some methods propose adopting masked autoencoder (MAE) for pre-training to enhance representation learning [14, 15]. The authors in [15] propose a complementary masking approach which affects the agents' past and future trajectories differently. Alternatively, the method in [14] uses multiple masking strategies in a continual learning framework. However, for masking, both methods adopt a random selection approach, which does not necessarily result in an optimal outcome. In this paper, we aim to enhance upon these masking strategies by moving beyond naive random masking.

Masked Prediction. Due to the success of masked prediction approaches in models for natural language processing, mask prediction has been adopted in other domains for enhancing pre-training and representation learning. In computer vision, the integration of mask prediction with widely used vision transformers (ViTs) [20] has led to the development of the BEIT model [7]. In this model, the input image is divided into patches, some of which are removed. The ViT is then pre-trained to reconstruct the missing patches based on the remaining visible ones. This technique has become a popular method for pre-training

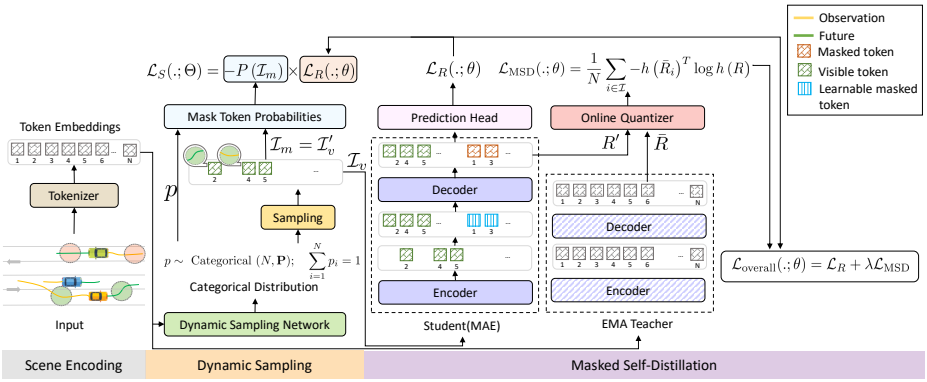

Fig. 1: Overview of the proposed self-supervised pre-training framework. Given the token embeddings, we sample visible tokens using a categorical distribution from a dynamic sampling network, optimized to favor tokens that express a more complex driving behavior. We then pass sampled visible tokens to an MAE to predict the complete scene. Beyond reconstruction accuracy, we further supervise the model with a masked self-distillation approach to transfer knowledge from the fully visible to the masked scene, inducing the model to focus on sampled tokens while maintaining awareness of the entire scene and, thereby, gradually guiding the sampling network towards selecting informative tokens.

in various modalities and has significantly improved performance in many downstream tasks [30, 46, 64, 68].

Given the widespread use of masked prediction, efforts have been made to design methods for improving it [5, 32, 41, 76, 79]. The masking strategy has been recognized as one of the major factors that directly influence the success of downstream tasks. For example, the authors in [36, 43] propose to utilize attention maps to refine masking. Here, one map focuses on preserving the key patches by masking nonessential areas and the other aims to mask highly attended tokens. To optimize the sampling, the method in [61] employs an adversarial training technique to identify optimal masking patches. Alternatively, some works adopt task-specific knowledge to improve masking. For instance, for video classification, the methods in [32, 41] learn the semantic parts of the frames, i.e. motions, and utilize them through the masking procedure.

Although masked prediction has been shown to improve representation learning for downstream tasks, its use in motion forecasting is not well explored. To address this gap, we introduce a novel masked trajectory prediction method that prioritizes sampling from scene elements indicative of more complex driving behaviors.

Knowledge Distillation is a technique to transfer knowledge from one model, i.e. teacher, to another, i.e. a student [31] . The main objective of this approach is to induce a smaller student model to match the teacher model's performance by learning from the soft targets (such as class probabilities) that the teacher produces on a dataset.

Self-distillation uses the model as both teacher and student, utilizing its own outputs rather than relying on posterior distributions for knowledge transfer, thus framing it as a discriminative self-supervised objective [1,19,35,39,77]. This approach typically relies on multiple iterative training phases. In each phase, the model learns from its own previous iterations and gradually refines its predictions and representations. The aim is to improve the model's performance and efficiency without the need for an external pre-trained teacher model. Self-distillation leverages the concept of learning from soft targets generated by the model itself, often leading to enhanced generalization and robustness in tasks across different domains [3,10,16,42]. In this paper, we aim to distill knowledge from a fully visible driving scene representation to its masked counterpart, inducing the model to focus on the informative parts of the scene, i.e. the parts that express a driving behavior.

3 Methodology

We propose a novel self-supervised pre-training framework that utilizes trajectory data for efficient yet effective pre-training, aiming to enhance behavior representation learning for a more transferable and generalizable prediction model. Utilizing an autoencoder architecture, our method features a novel masking strategy for masked trajectory prediction, focusing on selecting informative visible tokens (i.e. those that are depicting more complex driving behavior) during the masking phase to ensure learning of meaningful representations specifically tailored for the downstream task of trajectory prediction.

To do so, we propose a *dynamic sampling* step in which a sampling network is optimized to take tokens' informativeness into account to ensure a focus on fewer, yet more meaningful visible tokens. In addition, we propose a *masked self-distillation* approach to distill knowledge (i.e. representation) of a fully visible scene to the sampled tokens. As such, during the reconstruction, the model is induced to focus on learning the sampled tokens while maintaining awareness of the entire scene. During the integrated training, the model learns to reconstruct detailed scene representations from the masked scene and simultaneously guides the sampling network to favor more informative tokens. Such pre-training not only enhances the model's ability to generalize from limited or complex data but also strengthens its robustness and adaptability to unseen data. Fig. 1 provides an overview of the proposed self-supervised pre-training framework.

3.1 Problem Formulation

The goal of trajectory prediction is to forecast future locations of surrounding agents according to their observed behavior. Specifically, at time step t, let the past trajectory of the i-th vehicle be a set of $2D$ coordinates in bird's eye view over some observation horizon O time steps $X_i = \{(x_i, y_i)^{t-O+1}, \cdots, (x_i, y_i)^t\}$. Accordingly, the objective is to predict future trajectory $Y_i = \{(x_i, y_i)^{t+1}, \cdots, (x_i, y_i)^{t+T}\}$, where T is the prediction horizon. In addition, the model is provided with the road information extracted from the driving scenes represented

as an HD map in the vector space. For simplicity, in the rest of the paper, we refer to inputs, e.g., observations and the map, and future predictions as x and y, respectively.

3.2 Scene Representation

For scene encoding, we use vectorized representations and employ a tokenizer that models agents' trajectories and lane segments as polylines [44], which serve as *tokens* in our framework.

To find token embeddings for trajectory polylines, we utilize a feature pyramid network (FPN) [45] to merge multi-scale agent motion features. We then apply a $1D$ neighborhood attention to extract local motion features. To embed lane polylines, we use a simplified PointNet model [58] and use learnable embeddings for semantic attributes, including agent categories or lane types. We follow the agent-centric coordinate system to normalize data as [15]. To maintain the positional information of the tokens, we embed global position data into tokens using a two-layer multi-layer perceptron (MLP).

3.3 Dynamic Sampling

The effectiveness of masked prediction heavily depends on the chosen masking strategy, which influences the type of information learned [21, 67, 79]. Using a naive strategy, such as random masking may overlook variations in information density among tokens, potentially leading to inefficient learning.

To address this shortcoming, we propose a dynamic sampling step where a sampling network is optimized to identify and select highly informative trajectory and lane tokens. These tokens represent complex driving behaviors and intricate, curved lanes within the driving scene.

Our approach processes tokens' embeddings $\mathbf{E}$ obtained from the tokenizer through a multi-head attention network (MHA). The output of the MHA is fed into a MLP block Φ followed by a Softmax activation to calculate the probability score for each token as follows:

$$\begin{aligned} \mathbf{Z} &= \mathrm{MHA}(\mathbf{E}); \quad \mathbf{Z} \in \mathbb{R}^{N \times d}, \\ \mathbf{P} &= \mathrm{Softmax}(\Phi(\mathbf{Z})); \quad \mathbf{P} \in \mathbb{R}^{N}, \end{aligned} \tag{1}$$

where N is the number of tokens. We then assign an N-dimensional categorical distribution to $\mathbf{P}$ ($p \sim$ Categorical $(N, \mathbf{P})$; $\sum_{i=1}^{N} p_i = 1.$) and select a set of visible tokens $\mathcal{I}_V$ without replacement, creating a masking set $\mathcal{I}_M$. The number of visible tokens n_v is determined by a predefined masking ratio $\rho \in (0, 1)$, calculated as $N \times (1 - \rho)$.

Sampling Objective As mentioned earlier, the proposed dynamic sampling network learns to identify and select informative trajectory and lane tokens for MAE pre-training by learning a categorical distribution over all tokens. Due to the non-differentiability of sampling, we define an auxiliary objective function for the sampler's learning process. This method is grounded in the REINFORCE

policy gradient algorithm [69] in reinforcement learning. In specific, for the sampler with parameters Θ, we sample action a and calculate probabilities $p(a)$. Hence, having a reward function $\mathcal{R}$, we apply policy gradient as follows:

$$\Delta\Theta = \alpha \cdot \mathcal{R} \cdot \frac{\partial \log p(a)}{\partial \Theta}, \tag{2}$$

where α is the learning rate. As such, treating token sampling as an action within the environment of MAE, we define the masked reconstruction loss $\mathcal{L}_R$, which serves as the reward. Following the expected reward maximization in the REINFORCE algorithm, our goal is to optimize the sampling network by *increasing the expected reconstruction error* over the masked trajectory and lane tokens $\mathbb{E}[\mathcal{L}_R]$.

In specific, our observations indicate a notable disparity in reconstruction errors, with tokens depicting complex driving behaviors, such as overtaking or turning showing higher errors, contrasted with lower errors for tokens related to simpler behaviors such as cruising. This discrepancy suggests that tokens indicative of complex driving scenarios are more challenging to accurately reconstruct, pointing towards their higher information content and significance in understanding driving dynamics. Therefore, via optimizing the sampler by maximizing the expected reconstruction error over the masked tokens, the sampling network predicts high probability scores for the tokens from complex driving behaviors compared to the tokens from the simpler ones. This targeted sampling, doing non-uniform sampling, enhances behavior representation learning by leveraging the complexity levels of the tokens. In addition, it achieves a better accuracy comparable to random sampling with fewer visible tokens (i.e. higher masking ratio), leading to reduced computational costs and faster pre-training.

Formally speaking, we define the sampling objective as below:

$$\mathcal{L}_S(\Theta) = -\mathbb{E}_\Theta\left[\mathcal{L}_R(\theta)\right] = -\sum_{i=1}^{|\mathcal{I}_m|} P_\Theta^i \cdot \mathcal{L}_R^i(\theta), \tag{3}$$

where $|\mathcal{I}_m|$ indicates the number of masked trajectory and lane tokens. P_Θ^i and $\mathcal{L}_R^i$ stand for the probability and the reconstruction error of the i^{th} masked token, respectively. Note that $\mathcal{L}_R(\theta)$ represents the reconstruction error of MAE with parameters θ. Here, we ensure that gradient updates from the sampling network $\mathcal{L}_S(\Theta)$ do not affect the MAE.

3.4 Masked Self-distillation

The core idea of masked self-distillation is to distill representation from a full driving scene to the representation predicted from a masked one. By doing so, we prompt the model to utilize informative scene elements, such as driving behaviors, during reconstruction and subtly direct the sampling network to select trajectory and lane tokens that reflect these behaviors. To achieve this goal, we adopt a masked self-distillation approach in trajectory prediction. Below, we discuss the distillation objective and then provide a detailed explanation of the learning procedure in the autoencoder framework.

Distillation Objective Knowledge distillation involves training a student model to mimic a teacher model's output, thereby enhancing the student's performance. In this context, we aim to distill knowledge from a teacher with full input access to a student with partial (i.e. masked) observation. Here, instead of bringing in an external teacher, our method utilizes the model itself as the teacher by using a mean teacher model technique [26, 63]. In this approach, the teacher is derived from the student, i.e. the original model itself. In detail, the teacher model mirrors the student's structure but uses the student's parameters' exponential moving averages (EMA) as its own parameters. Formally, the mean teacher parameters are formulated as follows:

$$\bar{\theta}_t = \alpha \bar{\theta}_{t-1} + (1-\alpha)\theta_t, \tag{4}$$

where θ and $\bar{\theta}$ indicate the original and mean teacher model's parameters, respectively. α is a hyper-parameter for smoothing updates. We refer to the mean teacher as the EMA teacher.

With the EMA teacher as the teacher and the original model as the student, we initiate the masked self-distillation procedure. Initially, the teacher processes the input driving scene, in token embeddings, to generate *distillation targets* for N tokens, $\bar{R}_i|_{i=1}^{N}$. Meanwhile, we feed the visible tokens $\mathcal{I}_v$ with size $n_v << N$ obtained from the dynamic sampling step to the student, resulting in encoded visible tokens. Next, to form a complete representation set compared with distillation targets with size N, we merge encoded visible tokens with a shared, learnable feature vector that stands for masked tokens. This set then feeds the decoder to reconstruct the *distillation prediction* $R'_i|_{i=1}^{N}$. During the reconstruction phase, the EMA teacher has a full view of the driving scene (with no masking) while the student has a partial observation. We, therefore, define our masked self-distillation objective as the means of distilling knowledge from a complete scene representation $\bar{R}_i$ to one derived from a masked version R'_i.

Learning Procedure Given the absence of categorical labels in MAE pretraining, distilling logits is ineffective for learning semantically meaningful representations. We, therefore, resort to distilling the intermediate representations. For this, we employ an online quantizer to convert output features into a soft codewords distribution [82]. Hence, we define a masked self-distillation loss $\mathcal{L}_{\text{MSD}}$ by minimizing the cross-entropy between target and predicted representations as bellow:

$$\mathcal{L}_{\text{MSD}} = \frac{1}{N} \sum_{i \in \mathcal{I}} -h\left(\bar{R}_i\right)^T \log h\left(R'_i\right), \tag{5}$$

where $h(.)$ stands for the online quantizer technique. As such, using the masked self-distillation loss $\mathcal{L}_{\text{MSD}}$, the student will be optimized to distill the representation obtained by the EMA teacher. Note that, in each step, the EMA teacher's weight is updated by the exponential moving average of the original model and does not require a gradient.

Autoencoder We use a transformer-based encoder-decoder architecture as in [15]. The encoder consists of several transformer blocks and only encodes the visible trajectory and lane tokens obtained through the sampling procedure. Positional embeddings are added to the entire input sequence, including the mask tokens. Each mask token type corresponds to a learned vector specific to its masked element type.

Note that for the reconstruction error we only consider masked tokens. Hence, we feed the decoded masked tokens to a prediction head, implemented as a linear projection layer, to predict the normalized 2-dimensional coordinates of masked trajectory and lane polylines. As for the reconstruction loss $\mathcal{L}_R$, we use L1 loss for trajectory reconstruction and mean squared error (MSE) loss for lane polyline reconstruction.

3.5 Overall Objective

Pre-training We pre-train the proposed DySeT end-to-end. In specific, in each step, first, the dynamic sampling network with parameter Θ is optimized w.r.t $\mathcal{L}_S$. Next, the MAE, parameterized with θ, is updated by the following objective:

$$\mathcal{L}_{overall}(.;\theta) = \mathcal{L}_R(.;\theta) + \lambda \mathcal{L}_{\text{MSD}}(.;\theta), \tag{6}$$

where $\mathcal{L}_R$ stands for the reconstruction loss and λ is a hyperparameter to control the distillation objective.

Fine-tuning on the Downstream Task For the motion forecasting task, we implement an end-to-end fine-tuning process on the pre-training model similar to [15]. This involves discarding the MAE decoder, sampling network and masking, and substituting the pretext prediction heads with a multi-modal future decoder. We adopt the commonly used Huber loss for trajectory regression and cross-entropy loss for confidence classification with equal weights. All agents present in the scene are used to compute the loss.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on the recently released large-scale Argoverse2 [12] and widely used pedestrian ETH-UCY [40, 54] datasets. The Argoverse2 contains 250K non-overlapping scenarios, recorded at 10 Hz, divided into 200K, 25K, and 25K samples for training, validation, and testing, respectively. The task is to make $6s$ predictions based on $5s$ observations. The ETH-UCY dataset comprising five subsets (ETH, HOTEL, UNIV, ZARA1, ZARA2) at 2.5 Hz, is used to predict future movements over 12 time steps (4.8 seconds) based on 8 time steps (3.2 seconds) of observed human trajectories.

Table 1: Quantitative results on the Argoverse2 motion forecasting test set. For each metric, the best result is in **bold** and the second best result is underlined.

Method		b-FDE$_6$	minADE$_6$	minFDE$_6$	MR$_6$	minADE$_1$	minFDE$_1$	MR$_1$
Supervised-Learning	THOMAS [24]	2.16	0.88	1.51	0.20	1.95	4.71	0.64
	GoRela [18]	2.01	0.76	1.48	0.22	1.82	4.62	0.66
	GANet [65]	1.96	0.72	1.35	0.17	1.77	4.47	**0.59**
	FRM [53]	2.47	0.89	1.81	0.29	2.37	5.93	0.71
	ProphNet [66]	**1.88**	0.68	1.33	0.18	1.80	4.74	-
	QCNet [83]	1.91	**0.65**	1.29	**0.16**	**1.69**	**4.30**	**0.59**
	HPTR [81]	2.03	0.73	1.43	0.19	1.84	4.61	0.61
SSL-Lane [8]	Lane Masking	2.38	0.83	1.70	0.25	2.17	5.67	0.67
	Dist. to Inter.	2.39	0.84	1.71	0.25	2.18	5.71	0.67
	S/F Cls.	2.35	0.83	1.67	0.25	2.22	5.90	0.69
T-MAE [14]	Cont. finetune	2.11	0.73	1.51	0.18	1.94	4.48	0.64
Forecast-MAE [15]	Scratch	2.06	0.73	1.43	0.19	1.84	4.60	0.62
	finetune	2.03	0.71	1.39	0.17	1.74	4.35	0.61
DySeT (Ours)		1.93	0.67	**1.28**	**0.16**	1.76	4.41	0.61

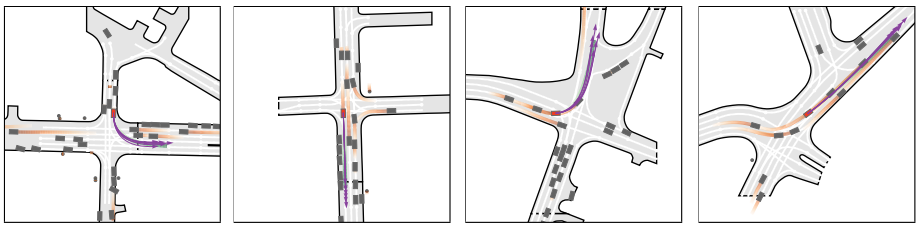

Fig. 2: Qualitative results on Argoverse2. The Agent of interest is shown in red color and the observation, ground truth, and prediction are shown as orange, green, and purple, respectively.

Metrics. We employ official benchmark metrics including minimum average displacement error (minADE$_K$), minimum final displacement error (minFDE$_K$), (b-minFDE$_K$), and miss rate (MR$_K$), where K stands for the number of predicted trajectories in the multimodal setting.

To assess robustness, we use the perturbation resistance score (PRS), calculated by first determining the per-sample absolute change in a trajectory prediction error metric m as $\text{abs}(\Delta) = \frac{1}{n}\sum_{i=1}^{n} |m_{\text{perturbed}}\ (i) - m_{\text{original}}\ (i)|$, comparing errors on perturbed versus original data. PRS is then equal to $[1 - (\text{abs}(\Delta)/\ m_{\text{original}}\ )] * 100$, indicating the model's perturbation resistance.

Implementation Details. We train the model using AdamW Optimizer with a 1e-4 weight decay, batch size of 128, and 1e-3 learning rate. The model operates in an agent-centric coordinate system, focusing on agents and lanes within 100 meters of the focal agent, with a latent feature dimension of 128. The distillation smoothing parameter λ is empirically set to $1e^{-1}$. We use an encoder and decoder with depth 5.

4.2 Comparison to SOTA

In this subsection, we pre-train and fine-tune the model on the same benchmark to demonstrate the advantages of our proposed self-supervised pre-training framework.

Table 2: Comparison with state-of-the-art methods on $\text{minADE}_{20}/\text{minFDE}_{20}$ metrics on ETH/UCY dataset. For each metric, the best result is in **bold** and the second best result is underlined

Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	Average
S-STGCNN [51]	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
PECNet [48]	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48
RSBG [62]	0.80/1.53	0.33/0.64	0.59/1.25	0.40/0.86	0.30/0.65	0.48/0.99
SGCN [60]	0.63/1.03	0.32/0.55	0.37/0.70	0.29/0.53	0.25/0.45	0.37/0.65
MID [28]	0.39/0.66	0.13/0.22	**0.22**/0.45	**0.17**/0.30	0.13/0.27	0.21/0.38
GP-Graph [73]	0.43/0.63	0.18/0.30	0.24/**0.42**	**0.17**/0.31	0.15/0.29	0.23/0.39
GroupNet [72]	0.46/0.73	0.15/0.25	0.26/0.49	0.21/0.39	0.17/0.33	0.25/0.44
EqMotion [74]	0.40/0.61	**0.12**/**0.18**	0.23/0.43	0.18/0.32	0.13/**0.23**	0.21/0.35
DySet (Ours)	**0.32**/**0.46**	0.14/0.21	0.24/0.45	**0.17**/**0.28**	**0.12**/0.25	**0.20**/**0.33**

Argoverse. We compare our method with the state-of-the-art SSL-based and supervised prediction approaches as shown in Table 1. For SSL-lane, similar to [15], we report three of its pretext tasks suitable for the Argoverse2 dataset including lane marking, distance to the intersection (dist. to inter.), and success-failure classification. The table shows our method surpassing all SSL-based approaches on most metrics. Particularly in the multimodal setting, it outperforms Forecast-MAE [15], the recent masked prediction model, by up to 8% on minFDE. Moreover, our model's performance matches that of the leading supervised models, ranking first on two k = 6 metrics and closely second on others with a small margin. Some qualitative examples of our proposed approach are shown in Fig. 2.

ETH/UCY. In Table 2, we compare our model with state-of-the-art approaches on the ETH/UCY dataset. The table indicates that, on average, our method outperforms other methods on both metrics. As for the five subsets, our model has competitive results, especially, on the ETH subset, outperforming the second best model by 18%(0.39 → 0.32) on minADE and 25%(0.61 → 0.46) on minFDE metrics. These results further verify the effectiveness of the proposed method.

4.3 Robustness and Generalization

Generalizability. In the context of autonomous driving, to understand how well trajectory prediction models perform across different settings, the models can be trained in one environment and tested in another. This approach reveals the challenges of varying conditions between training and testing scenarios, such as differences in pedestrian trajectories between shopping malls and streets or in vehicle motions across cities with unique road layouts and traffic patterns.

Motivated by this, we conduct two experiments to evaluate our method's generalizability for vehicle and pedestrian trajectory prediction tasks as explained in details below.

Vehicle Motion Forecasting. Following the experiment in [15], we categorize the Argoverse2 training and validation datasets into two distinct groups based on the six cities that the data was collected from. Initially, we train our

Table 3: Analyzing generalizability to an *unseen domain* on the Argoverse2 validation.

Method		minADE_6	minFDE_6	MR_6
Forecast-MAE [15]	Scratch	0.910	1.645	0.235
	Finetune	0.897	1.613	0.216
DySet (Ours)	Scratch	0.739	1.406	0.206
	Finetune	0.724	1.328	0.179

Table 4: Analyzing generalizability to *unseen domains* on the ETH/UCY dataset. The metrics are reported as $\text{minADE}_{20}/\text{minFDE}_{20}$.

Method / Seen domain	ETH	Hotel	UNIV	ZARA1	ZARA2
S-STGCNN [51]	1.51/2.77	2.15/3.85	0.81/1.62	0.82/1.62	0.83/1.66
PECNet [48]	1.56/2.79	2.19/3.87	0.93/1.59	0.93/1.66	0.93/1.59
RSBG [62]	1.68/2.91	2.28/4.04	0.94/1.82	0.92/1.81	1.07/1.91
SGCN [60]	1.44/2.74	2.20/3.83	0.80/1.58	0.83/1.65	0.83/1.64
T-GNN [75]	**1.09**/**2.01**	1.82/3.14	0.59/1.30	0.64/1.32	0.65/1.33
GroupNet [72]	1.17/2.42	1.80/3.39	0.56/**1.29**	0.52/1.38	0.60/1.38
GP-Graph [73]	1.21/2.54	1.96/3.58	**0.54**/**1.29**	0.59/1.46	0.38/1.45
DySet (Ours)	1.19/2.13	**1.77**/**3.01**	0.61/1.33	**0.49**/**1.20**	**0.57**/**1.20**

model using data exclusively from the first group of cities, specifically Miami, Pittsburgh, and Austin, to test its performance on data from cities not encountered during training. We refer to these as *unseen* cities.

As shown in Table 3, our method surpasses Forecast-MAE [15], which aims to learn generalizable representations using masked prediction approach. Notably, in the experiments utilizing a finetuning regime, our model achieves a performance improvement of up to 17%. These results confirm that even when the test domain is not seen during the training phase, the proposed model can effectively learn more generalizable and adaptable representations.

Human Trajectory Prediction. Each subset within the ETH/UCY dataset captures unique aspects of urban pedestrian behavior. For instance, UNIV showcases intricate social interactions and group dynamics, ZARA1 and ZARA2 highlight navigational strategies in crowded shopping areas, HOTEL reflects more leisurely, less dense walking patterns, and ETH combines rapid academic commutes with leisurely walks. Hence, we treat each subset as one domain. As such, we train the model on one domain, as *seen* domain, and test it on four other *unseen* domains following [75].

Given a seen domain, we average the results across four tested domains and report the results in Table 4. As the table indicates, our method ranked first in three subsets on both metrics. Particularly, given ZARA2 as the source domain, our method outperforms the second-best method by up to 12%($0.65 \rightarrow 0.57$) and 10%($1.33 \rightarrow 1.20$) on minADE and minFDE metrics, respectively.

As an exception, in the UNIV subset, our method underperforms compared to GP-Graph [73] and GroupNet [72]. This is likely due to dense scenes with extensive group-wise social interactions involving large numbers of pedestrians (20 or more). Such scenarios are generally rarer in other subsets. Dense scenes favors these models, which emphasize group representation learning. Group-wise masking, tailored for human trajectory prediction, can be considered as a future extension to our work.

Table 5: Analyzing robustness against *adversarial perturbations* on the Argoverse2 validation. (↓) and (↑) indicate lower and higher values are better.

Attack	Method	Original/Perturbed	abs(Δ)(↓)	PRS (%)(↑)
Targeted	Forecast-MAE [15]	0.712 / 0.879	0.153±0.06	78.51
	DySet (Ours)	0.706 / 0.863	0.093±0.02	86.83
Non-Targeted	Forecast-MAE [15]	0.712 / 0.755	0.132±0.09	81.46
	DySet (Ours)	0.706 / 0.741	0.068±0.05	90.37

Adversarial Robustness. Adversarial robustness is one of the key concerns for reliability of trajectory prediction models in autonomous driving [9, 80]. Here, we perform an experiment to evaluate our method's resistance to adversarial examples.

For this purpose, we use adversarial perturbations to agent dynamics method of [80]. Following the original configuration, we employ a white-box method and train the models with an Adam optimizer at a 0.01 learning rate. To ensure fairness, we limit iterations to 100 and perturbation deviation to a maximum of 1 for both methods. We compare our model to Forecast-MAE [15], aiming for robust and generalizable feature learning. The results are reported in Table 5 and are organized by whether the perturbations are non-targeted or targeted, which are determined by the optimization of adversarial objectives with respect to the reported metric, in this case minADE.

From the table, when applying the adversarial perturbation, our method faces less degradation compared to Forecast-MAE (∼13% vs ∼21%) in the more challenging targeted scenarios. In the non-targeted setting, both methods achieve better results, however, our method still significantly outperforms Forecast-MAE on the robustness metric PSR, achieving more than ∼90%. These findings further verify the superior robustness of the proposed framework against existing self-supervised pre-training methods.

4.4 Ablation Study

Masking Ratio. We evaluate the impact of the masking ratio on performance using both random sampling and our proposed strategy, as depicted in Fig. 3. Our method, which selects visible tokens using a sampling network, outperforms in both minADE (0.724) and minFDE (1.278) metrics. Notably, this enhancement occurs at a higher masking ratio of 70% where memory usage is reduced and pre-training duration is shortened. This indicates that our strategy efficiently selects a larger number of informative tokens related to driving behavior and fewer less informative ones. As a result, our method requires fewer visible tokens compared to random sampling while achieving better performance.

Controlling Distillation Impact. Hyperparameter λ is used to balance the two terms in Eq. 6. Setting λ too high, emphasizes too much on the distillation objective and leads the model to a wrong convergence direction and reducing performance. Hence, we set different values to find the most suitable λ. We show results of the evaluation on the Argoverse2 validation set on the left side of Table 6. As the table indicates, when we set $\lambda = 1e^{-1}$, the model can achieve the best

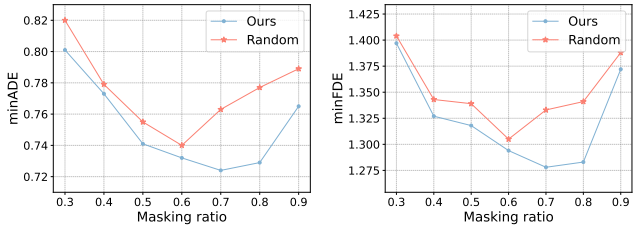

Fig. 3: The impact of masking ratio on the model's performance.

Table 6: Ablation studies on the Argoverse2 validation set to examine the impact of the self-distillation objective on overall performance **(Left)** and the related design choices of the proposed method **(Right)**.

	Distillation loss weight			
Metric	$\lambda = 1e^{-2}$	$\lambda = 1e^{-1}$	$\lambda = 1e^{0}$	$\lambda = 1e^{+1}$
minADE_6	0.747	**0.724**	0.783	0.804
minFDE_6	1.347	**1.278**	1.372	1.425

Model	**minADE$_6$**	**minFDE$_6$**
Baseline	0.822	1.486
+ Dyn. samp.	0.751	1.306
+ Self-dist.	**0.724**	**1.278**

performance, with a decrease in performance observed for significantly smaller or larger λ values.

Contributions of Each Component. We conduct an ablation study on the proposed method using a masked autoencoder with a random masking strategy as a baseline. We incrementally add the proposed components to the baseline and report the results on the Argoverse2 validation set and report the results on the right side of Table 6. It can be observed from the table that adopting the proposed dynamic sampling in the baseline can significantly improve performance, especially on minFDE by $\sim 12\%(1.486 \rightarrow 1.306)$. This finding validates the effectiveness of our sampling module in selecting informative tokens for enhancing behavior learning and, consequently, achieving better performance in predicting the final goal, as reflected by the minFDE metric. Moreover, adding the proposed masked self-distillation module further improves the baseline by up to 14% on both metrics. Note that for each model, we report on the masking ratio that gives the best result.

5 Conclusion

In this work, we proposed a novel self-supervised pre-training framework that leverages trajectory data for efficient yet effective representation learning, leading to more transferable and generalizable predictions. Based on a masked autoencoder architecture, we proposed a dynamic sampling step and a masked self-distillation approach that focuses on selecting semantically rich visible tokens. These tokens depict more complex driving behaviors, thus ensuring the acquisition of meaningful representations tailored for trajectory prediction tasks. We conducted an extensive evaluation on large-scale autonomous driving and human trajectory prediction datasets and demonstrated that our approach not only outperforms existing SSL-based and supervised learning-based methods but is also significantly more robust in varied conditions.

References

1. Andonian, A., Chen, S., Hamid, R.: Robust cross-modal representation learning with progressive self-distillation. In: CVPR (2022) 5
2. Aydemir, G., Akan, A.K., Guney, F.: Adapt: Efficient multi-agent trajectory prediction with adaptation. In: ICCV (2023) 3
3. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. In: ICML (2022) 5
4. Bahari, M., Saadatnejad, S., Rahimi, A., Shaverdikondori, M., Shahidzadeh, A.H., Moosavi-Dezfooli, S.M., Alahi, A.: Vehicle trajectory prediction works, but not everywhere. In: CVPR (2022) 1
5. Bandara, W.G.C., Patel, N., Gholami, A., Nikkhah, M., Agrawal, M., Patel, V.M.: Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In: CVPR (2023) 4
6. Bansal, M., Krizhevsky, A., Ogale, A.: ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst. In: RSS (2019) 3
7. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. ICLR (2022) 3
8. Bhattacharyya, P., Huang, C., Czarnecki, K.: SSL-Lanes: Self-supervised learning for motion forecasting in autonomous driving. In: CoRL (2022) 3, 10
9. Cao, Y., Xiao, C., Anandkumar, A., Xu, D., Pavone, M.: Advdo: Realistic adversarial attacks for trajectory prediction. In: ECCV (2022) 13
10. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021) 5
11. Chai, Y., Sapp, B., Bansal, M., Anguelov, D.: MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In: CoRL (2019) 3
12. Chang, M.F., Lambert, J.W., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J.: Argoverse: 3d tracking and forecasting with rich maps. In: CVPR (2019) 1, 9
13. Chen, C., Pourkeshavarz, M., Rasouli, A.: Criteria: a new benchmarking paradigm for evaluating trajectory prediction models for autonomous driving. ICRA (2024) 1
14. Chen, H., Wang, J., Shao, K., Liu, F., Hao, J., Guan, C., Chen, G., Heng, P.A.: Traj-mae: Masked autoencoders for trajectory prediction. In: ICCV (2023) 2, 3, 10
15. Cheng, J., Mei, X., Liu, M.: Forecast-mae: Self-supervised pre-training for motion forecasting with masked autoencoders. In: ICCV (2023) 2, 3, 6, 9, 10, 11, 12, 13
16. Cheng, R., Wu, B., Zhang, P., Vajda, P., Gonzalez, J.E.: Data-efficient language-supervised zero-shot learning with self-distillation. In: CVPR (2021) 5
17. Choi, S., Kim, J., Yun, J., Choi, J.W.: R-pred: Two-stage motion prediction via tube-query attention-based trajectory refinement. In: ICCV (2023) 1
18. Cui, A., Casas, S., Wong, K., Suo, S., Urtasun, R.: Gorela: Go relative for viewpoint-invariant motion forecasting. In: ICRA (2023) 10
19. Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al.: Maskclip: Masked self-distillation advances contrastive language-image pretraining. In: CVPR (2023) 5
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020) 3

21. Feichtenhofer, C., Li, Y., He, K., et al.: Masked autoencoders as spatiotemporal learners. In: NeurIPS (2022) 2, 6
22. Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., Schmid, C.: VectorNet: Encoding HD maps and agent dynamics from vectorized representation. In: CVPR (2020) 3
23. Gilles, T., Sabatini, S., Tsishkou, D., Stanciulescu, B., Moutarde, F.: GOHOME: Graph-oriented heatmap output for future motion estimation. In: ICRA (2022) 3
24. Gilles, T., Sabatini, S., Tsishkou, D., Stanciulescu, B., Moutarde, F.: THOMAS: Trajectory heatmap output with learned multi-agent sampling. In: ICLR (2022) 10
25. Girgis, R., Golemo, F., Codevilla, F., Weiss, M., D'Souza, J.A., Kahou, S.E., Heide, F., Pal, C.: AutoBot: Latent variable sequential set transformers for joint multi-agent motion prediction. In: ICLR (2022) 3
26. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: NeurIPS (2020) 8
27. Gu, J., Sun, C., Zhao, H.: DenseTNT: End-to-end trajectory prediction from dense goal sets. In: ICCV (2021) 3
28. Gu, T., Chen, G., Li, J., Lin, C., Rao, Y., Zhou, J., Lu, J.: Stochastic trajectory prediction via motion indeterminacy diffusion. In: CVPR (2022) 11
29. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. In: NeurIPS (2019) 2
30. Hess, G., Jaxing, J., Svensson, E., Hagerman, D., Petersson, C., Svensson, L.: Masked autoencoder for self-supervised pre-training on lidar point clouds. In: CVPR (2023) 4
31. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015) 4
32. Huang, B., Zhao, Z., Zhang, G., Qiao, Y., Wang, L.: Mgmae: Motion guided masking for video masked autoencoding. In: CVPR (2023) 4
33. Huang, Z., Liu, H., Lv, C.: Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In: ICCV (2023) 1
34. Huang, Z., Mo, X., Lv, C.: Multi-modal motion prediction with transformer-based neural network for autonomous driving. In: ICRA (2022) 3
35. Ji, M., Shin, S., Hwang, S., Park, G., Moon, I.C.: Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In: CVPR (2021) 5
36. Kakogeorgiou, I., Gidaris, S., Psomas, B., Avrithis, Y., Bursuc, A., Karantzalos, K., Komodakis, N.: What to hide from your students: Attention-guided masked image modeling. In: ECCV (2022) 4
37. Karim, R., Shabestary, S.M.A., Rasouli, A.: Destine: Dynamic goal queries with temporal transductive alignment for trajectory prediction. In: ICRA (2024) 3
38. Khandelwal, S., Qi, W., Singh, J., Hartnett, A., Ramanan, D.: What-if motion prediction for autonomous driving. arXiv:2008.10587 (2020) 3
39. Kim, K., Ji, B., Yoon, D., Hwang, S.: Self-knowledge distillation with progressive refinement of targets. In: ICCV (2021) 5
40. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: Computer Graphics Forum (2007) 9
41. Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C.: Semmae: Semantic-guided masking for learning masked autoencoders. In: NeurIPS (2022) 2, 4

42. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS (2021) 5
43. Li, Z., Chen, Z., Yang, F., Li, W., Zhu, Y., Zhao, C., Deng, R., Wu, L., Zhao, R., Tang, M., et al.: Mst: Masked self-supervised transformer for visual representation. In: NeurIPS (2021) 4
44. Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., Feng, S., Urtasun, R.: Learning lane graph representations for motion forecasting. In: ECCV (2020) 6
45. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017) 6
46. Liu, J., Huang, X., Zheng, J., Liu, Y., Li, H.: Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In: CVPR (2023) 4
47. Liu, Y., Zhang, J., Fang, L., Jiang, Q., Zhou, B.: Multimodal motion prediction with stacked transformers. In: CVPR (2021) 3
48. Mangalam, K., Girase, H., Agarwal, S., Lee, K.H., Adeli, E., Malik, J., Gaidon, A.: It is not the journey but the destination: Endpoint conditioned trajectory prediction. In: ECCV (2020) 11, 12
49. Mao, W., Xu, C., Zhu, Q., Chen, S., Wang, Y.: Leapfrog diffusion model for stochastic trajectory prediction. In: CVPR (2023) 1
50. Mercat, J., Gilles, T., El Zoghby, N., Sandou, G., Beauvois, D., Gil, G.P.: Multi-head attention for multi-modal joint vehicle motion forecasting. In: ICRA (2020) 3
51. Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C.: Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In: CVPR (2020) 11, 12
52. Nayakanti, N., Al-Rfou, R., Zhou, A., Goel, K., Refaat, K.S., Sapp, B.: Wayformer: Motion forecasting via simple & efficient attention networks. In: ICRA (2023) 3
53. Park, D., Ryu, H., Yang, Y., Cho, J., Kim, J., Yoon, K.J.: Frm: Leveraging future relationship reasoning for vehicle trajectory prediction. In: ICLR (2023) 10
54. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV (2009) 9
55. Pourkeshavarz, M., Chen, C., Rasouli, A.: Learn tarot with mentor: A meta-learned self-supervised approach for trajectory prediction. In: ICCV (2023) 1, 3
56. Pourkeshavarz, M., Sabokrou, M., Rasouli, A.: Adversarial backdoor attack by naturalistic data poisoning on trajectory prediction in autonomous driving. In: CVPR (2024) 1
57. Pourkeshavarz, M., Zhang, J., Rasouli, A.: Cadet: a causal disentanglement approach for robust trajectory prediction in autonomous driving. In: CVPR (2024) 3
58. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3d classification and segmentation. In: CVPR (2017) 6
59. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: ECCV (2020) 3
60. Shi, L., Wang, L., Long, C., Zhou, S., Zhou, M., Niu, Z., Hua, G.: Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In: CVPR (2021) 11, 12
61. Shi, Y., Siddharth, N., Torr, P., Kosiorek, A.R.: Adversarial masking for self-supervised learning. In: ICML (2022) 4

62. Sun, J., Jiang, Q., Lu, C.: Recursive social behavior graph for trajectory prediction. In: CVPR (2020) 11, 12
63. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017) 8
64. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: NeurIPS (2022) 4
65. Wang, M., Zhu, X., Yu, C., Li, W., Ma, Y., Jin, R., Ren, X., Ren, D., Wang, M., Yang, W.: Ganet: Goal area network for motion forecasting. In: ICRA (2023) 1, 10
66. Wang, X., Su, T., Da, F., Yang, X.: Prophnet: Efficient agent-centric motion forecasting with anchor-informed proposals. In: CVPR (2023) 1, 10
67. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: CVPR (2022) 6
68. Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., Revaud, J.: Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In: CVPR (2023) 4
69. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning (1992) 2, 7
70. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv:2301.00493 (2023) 1
71. Xu, C., Li, T., Tang, C., Sun, L., Keutzer, K., Tomizuka, M., Fathi, A., Zhan, W.: Pretram: Self-supervised pre-training via connecting trajectory and map. In: ECCV (2022) 3
72. Xu, C., Li, M., Ni, Z., Zhang, Y., Chen, S.: Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In: CVPR (2022) 11, 12
73. Xu, C., Li, M., Ni, Z., Zhang, Y., Chen, S.: Learning pedestrian group representations for multi-modal trajectory prediction. In: ECCV (2022) 11, 12
74. Xu, C., Tan, R.T., Tan, Y., Chen, S., Wang, Y.G., Wang, X., Wang, Y.: Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In: CVPR (2023) 11
75. Xu, Y., Wang, L., Wang, Y., Fu, Y.: Adaptive trajectory prediction via transferable gnn. In: CVPR (2022) 12
76. Yuan, J., Zhang, X., Zhou, H., Wang, J., Qiu, Z., Shao, Z., Zhang, S., Long, S., Kuang, K., Yao, K., et al.: Hap: Structure-aware masked image modeling for human-centric perception. In: NeurIPS (2024) 2, 4
77. Yun, S., Park, J., Lee, K., Shin, J.: Regularizing class-wise predictions via self-knowledge distillation. In: CVPR (2020) 5
78. Zeng, W., Liang, M., Liao, R., Urtasun, R.: LaneRCNN: Distributed representations for graph-centric motion forecasting. In: IROS (2021) 3
79. Zhang, Q., Wang, Y., Wang, Y.: How mask matters: Towards theoretical understandings of masked autoencoders. In: NeurIPS (2022) 4, 6
80. Zhang, Q., Hu, S., Sun, J., Chen, Q.A., Mao, Z.M.: On adversarial robustness of trajectory prediction for autonomous vehicles. In: CVPR (2022) 13
81. Zhang, Z., Liniger, A., Sakaridis, C., Yu, F., Van Gool, L.: Real-time motion prediction via heterogeneous polyline transformer with relative pose encoding. In: NeurIPS (2023) 10
82. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. In: ICLR (2022) 8

83. Zhou, Z., Wang, J., Li, Y.H., Huang, Y.K.: Query-centric trajectory prediction. In: CVPR (2023) 1, 3, 10
84. Zhou, Z., Ye, L., Wang, J., Wu, K., Lu, K.: HiVT: Hierarchical vector transformer for multi-agent motion prediction. In: CVPR (2022) 3