Bootstrap Motion Forecasting With Self-Consistent Constraints

Maosheng Ye[1]*, Jiamiao Xu[2], Xunnong Xu[2], Tengfei Wang[1], Tongyi Cao[2], Qifeng Chen[1]

[1]The Hong Kong University of Science and Technology [2]DeepRoute.AI

Abstract

*We present a novel framework to bootstrap **M**otion forecast**I**ng with **S**elf-consistent **C**onstraints (MISC). The motion forecasting task aims at predicting future trajectories of vehicles by incorporating spatial and temporal information from the past. A key design of MISC is the proposed Dual Consistency Constraints that regularize the predicted trajectories under spatial and temporal perturbation during training. Also, to model the multi-modality in motion forecasting, we design a novel self-ensembling scheme to obtain accurate teacher targets to enforce the self-constraints with multi-modality supervision. With explicit constraints from multiple teacher targets, we observe a clear improvement in the prediction performance. Extensive experiments on the Argoverse motion forecasting benchmark and Waymo Open Motion dataset show that MISC significantly outperforms the state-of-the-art methods. As the proposed strategies are general and can be easily incorporated into other motion forecasting approaches, we also demonstrate that our proposed scheme consistently improves the prediction performance of several existing methods.*

1. Introduction

Motion forecasting has been a crucial task for self-driving vehicles that aims at predicting the future trajectories of agents (e.g., cars, pedestrians) involved in traffic. The predicted trajectories can further help self-driving vehicles plan their future actions and avoid potential accidents. Since the future is not deterministic, motion forecasting is intrinsically a multi-modal problem with substantial uncertainties. This implies that an ideal motion forecasting method should produce a distribution of future trajectories or at least multiple most likely ones.

Due to the inherent uncertainty, motion forecasting remains challenging and unsolved yet. Recently, researchers have proposed different architectures based on various representations to encode the kinematic states and context information from HDMap in order to generate feasible multi-modal trajectories [2, 7, 15, 18, 28, 29, 36, 51, 55, 56, 58]. These methods follow a traditional static training pipeline, where frames of each scenario are split into historical frames (input) and future frames (ground truth) in a fixed pattern. Nevertheless, the prediction task is a streaming task in real-world applications, where the current state will become a historical state as time goes by, and the buffer of the historical state is a queue structure to make successive predicted trajectories. As a result, temporal consistency thus becomes a crucial requirement for the downstream tasks for fault and noise tolerance. To tackle this issue, trajectory stitching is widely applied in traditional planning algorithms [13] to ensure stability along the temporal horizon. However, as the trajectory stitching operation is non-differentiable, it cannot be easily incorporated into learning-based models. Though deep-learning-based models show unprecedented motion prediction performance compared with traditional counterparts, they do not explicitly consider temporal consistency, leading to unstable behaviors in downstream tasks such as planning.

Inspired by these phenomena, we raise a question: can we explicitly enforce consistency when training a deep motion prediction model? On the one hand, the predicted trajectories should be consistent given the successive inputs along the temporal horizon, namely temporal consistency. On the other hand, the predicted trajectories should be stable and robust against small spatial noise or disturbance, namely spatial consistency. In this work, we propose a self-supervised scheme, named as ***MISC***, to enforce consistency constraints in both spatial and temporal domains, namely *Dual Consistency Constraints*. *Dual Consistency Constraints* could be viewed as an inner-model consistency and can significantly improve the quality and robustness of motion forecasting, without the need for extra data.

On top of the inner-model consistency, we also exploit the intra-model consistency. Multi-modality is another core characteristic of the motion prediction task. Existing datasets [9, 49] only provide a single ground-truth trajectory for each scenario, which can not satisfy multi-choice situations such as junction scenarios. Most methods adopt the winner-takes-all (WTA) [26] or its variants [4, 35] to alleviate this situation. However, WTA tends to produce confused

*Work done during an internship at DeepRoute.AI.

predictions when two trajectories are very close. In contrast, our method addresses the multi-modality problem by using more robust teacher targets obtained from self-ensembling, which leverages intra-model consistency. Multiple teacher targets can be viewed as a special kind of intra-model distillation while alleviating the problem of multi-modality. Our contributions are summarized as follows,

- We propose self-consistent constraints in both intra and inner model aspects.
- For the inner-model consistency, Dual Consistency Constraints are proposed to enforce temporal and spatial consistency in our model, which is shown to be a general and effective way to improve the overall performance in motion forecasting.
- For the intra-model consistency constraints, a self-ensembling constraint is explicitly exploited to enforce self-consistency with teacher targets, which provides multi-modality supervision for training.
- Extensive experiments on the Argoverse [9] motion forecasting benchmark and Waymo Open Motion dataset [12] show that the proposed approach achieves state-of-the-art performance.

2. Related Work

Motion Forecasting. Traditional methods [22, 43, 54, 62] for motion forecasting mainly utilize HDMap information for the prior estimation and Kalman filter [23] for motion states prediction. With the recent progress of deep learning on big data, more and more works have been proposed to exploit the potential of data mining in motion forecasting. Early efforts [2, 7, 11, 15, 21, 28, 29, 44, 46, 55, 56, 61] explore different representations, including rasterized image, graph representation, point cloud representation and transformer to generate the features for the task and predict the final output trajectories by regression or post-processing sampling. Most of these works focus on finding more effective and compact ways of feature extraction on the surrounding environment (HDMap information) and agent interactions. Based on these representations, other approaches [6, 33, 46, 56, 57, 58] try to incorporate the prior knowledge with traditional methods, which take the predefined candidate trajectories from sampling or clustering strategies as anchor trajectories. To some extent, these candidate trajectories can provide better guidance and goal coverage for the trajectories regression due to straightforward HDMap encoding. Nevertheless, this extra dependency makes the stability of models highly related to the quality of the trajectory proposals. Goal-guided approaches [16, 18, 17] are therefore introduced to optimize goals in an end-to-end manner, paired with sampling strategies that generate the final trajectory for better coverage.

Consistency Regularization. Consistency regularization has been fully studied in semi-supervised and self-supervised learning. Temporally related works [53, 27, 60] apply pairwise matching to minimize the alignment difference through optical flow or correspondence matching to achieve temporal smoothness. Other works [1, 14, 38, 42, 52, 45] apply consistency constraints to predictions from the same input with different transformations in order to obtain perturbation-invariant representations. [8, 3] reverse the temporal order or mask some information and generate pairwise consistency between these predicted trajectories. [48] introduced consistency by examining the gap between agent-centric and scene-centric settings.

Multi-hypothesis Learning. Motion forecasting task inherently has multi-modality due to the future uncertainties and difficulties in acquiring accurate ground-truth labels. WTA [19, 47] in multi-choice learning and its variants [32, 41] incorporate with better distribution estimation to improve the training convergence, thus allowing more multi-modality. Some anchor-based methods [4, 7, 39, 56] introduce pre-defined anchors based on kinematics or road graph topology to provide guidance. However, these methods only allow one target per training stage. Other methods [4, 18] try to generate multi-target for supervision with heavy handcrafted optimizations. We propose a Teacher-Target-Constraints approach to provide more precise trajectory teacher labels by leveraging the power of self-ensembling [25, 59]. Multiple targets are explicitly provided to each agent to better model the multi-modality.

3. Approach

The overall architecture of MISC comprises three parts. 1) We first utilize a joint spatial and temporal learning framework TPCN [55] to extract pointwise features. Based on these features, we decouple the trajectory prediction problem as a two-stage regression task. The first stage performs goal prediction and completes the trajectory with the goal position guidance. The second stage takes the output of the first stage as anchor trajectories for refinement. 2) To enhance the spatial and temporal consistency of our MISC, we introduce *Dual Consistency Constraints* at the inner-model level, which helps regularize the predictions in a streaming task view. 3) We leverage self-ensembling to generate more precise teacher targets to provide intra-model level self-consistent **Teacher Targets Constraints** in Sec. 3.3

3.1. Architecture

Recently, TPCN [55] has gained popularity in this task due to its flexibility for joint spatial-temporal learning and scalability to adopt more techniques from point cloud learning. Considering its limitation in representing future uncertainty, we extend TPCN in a two-stage manner through goal position prediction for more accurate waypoints prediction as our baseline. The pipeline is shown in Fig. 1.

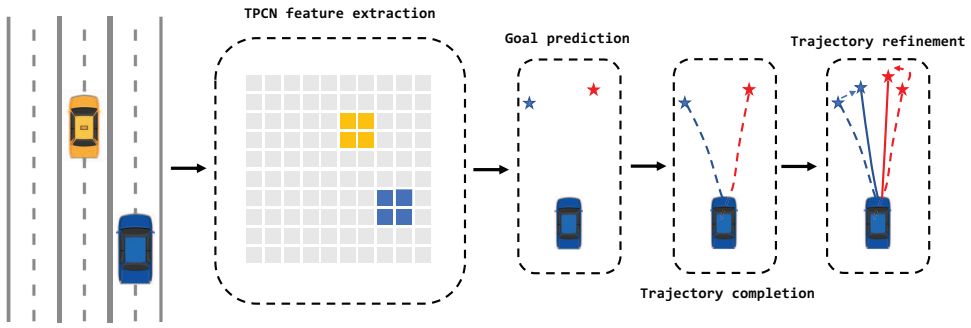

Figure 1. The overall architecture. We utilize TPCN [55] as a feature extraction backbone to model the spatial and temporal relationship among agents and map information. A goal prediction header is then used to regress the possible goal candidates; with the goal position, we apply trajectory completion to obtain full trajectories; finally, the trajectories are refined based on the output of the trajectory completion module as anchor trajectories.

Feature Extraction: TPCN utilizes dual-representation point cloud learning techniques with multi-interval temporal learning to model the spatial and temporal relationship. All the historical trajectories of input agents and map information are based on pointwise representation $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$, where $\mathbf{p}_i$ is the i-th point with N points in total, and then go through multi-representation learning framework to generate pointwise features $\mathcal{P} \in R^{N \times C}$, where C is the channel number.

Goal Prediction: With the pointwise features from the backbone, we also adopt the popular goal-based ideas [16, 18, 58] to find the optimal planning policy. Specifically, we first gather all corresponding pointwise agent features and then sum over features to get the agent instance feature $\phi \in R^{1 \times C}$. To generate K goal position prediction $G = \{G^k : (g_x^k, g_y^k) | 1 \leq k \leq K\}$, we use a simple MLP layer: $G = MLP(\phi)$. Instead of relying on heavy sampling strategies like previous goal-based methods, our method avoids generating extra proposals, which may lead to a large computation overhead.

Trajectory Completion: With the predicted goal positions, we need to complete each trajectory conditioned on these goals. We propose a simple trajectory completion module to generate K full trajectories $\left\{\tau_{reg}^k | 1 \leq k \leq K\right\}$ with a single MLP layer as follows:

$$\tau_{reg}^k = \{(x_1^k, y_1^k), (x_2^k, y_2^k), \ldots, (x_T^k, y_T^k)\}, \quad (1)$$

$$\tau_{reg}^k = MLP(concat(\phi, G^k)). \quad (2)$$

Trajectory Refinement: Inspired by Faster-RCNN [40] and Cascade-RCNN [5], we use the output trajectories from the Trajectory Completion as anchor trajectories to refine trajectories and predict the corresponding possibility of each trajectory. In particular, the input of the trajectory refinement module will be the whole trajectory with agent historical waypoints $\tau_{history}$. With a residual block followed by a linear layer Reg and Cls respectively, we regress the delta offset to the first stage outputs $\Delta_{\tau_{reg}} = Reg(\tau_{reg}, \tau_{history})$ and corresponding scores $\tau_{cls} = \left\{c^k | 1 \leq k \leq K\right\}$ respectively, where $\tau_{cls} = Cls(\tau_{reg}, \tau_{history})$. The final output trajectories will be $\tau_{reg'} = \Delta_{\tau_{reg}} + \tau_{reg}$.

3.2. Dual Consistency Constraints

Consistency regularization has been proved as an effective self-constraint that improves robustness against disturbances. We thus propose inner-model level **Dual Consistency Constraints** in both spatial and temporal domains to align predicted trajectories for continuity and stability.

3.2.1 Temporal Consistency

In motion forecasting, since each scenario contains multiple successive frames within a fixed temporal chunk, it is reasonable to assume that any two overlapping chunks of input data with a small time shift should produce consistent results. The motion forecasting task aims to predict K possible trajectories with T time steps for one scenario, given M frames historical information. Suppose the information at each history frame is I_i, where $1 \leq i \leq M$ and the k-th output future trajectories are $\left\{(x_i^k, y_i^k) | M < i \leq M + T\right\}$. We first apply time step shift s for the input for temporal consistency. Therefore, the input history frames information will be $\{I_i | 1 + s \leq i \leq M + s\}$ and then we apply the same network for the shifted history information with surrounding HDMap information to generate the k-th output trajectories $\left\{(x_i'^k, y_i'^k) | M + s < i \leq M + s + T\right\}$. When s is small, the driving intentions or behavior keeps stable for a short period. Since both trajectories have $T - s$ overlapping waypoints, they should be as close as possible and share consensus. Thus, we can construct self-constraints for a single scenario input due to the streaming property of the input data. Fig. 2 demonstrates the overall idea of the temporal consistency constraint.

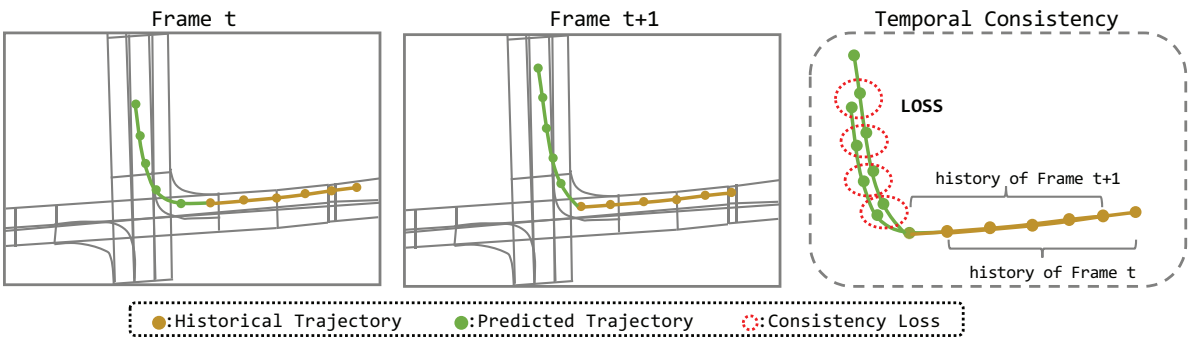

Figure 2. The overall idea of the temporal consistency. In the training stage, we first generate output prediction trajectory points as normal for each given scenario. Then we slide the input with a step in order to introduce the streaming nature to generate the consecutive output trajectory points. The proposed temporal consistency requires the overlap between these two outputs to be consistent

Trajectory Matching: Since we predict K future trajectories to deal with the multi-modality, it is crucial to consider the trajectory matching relationship between original predictions and time-shifted predictions when applying the temporal consistency alignment. For a matching problem, the metric on similarity criteria and matching strategies will be two key factors. Several ways can be used to measure the difference between trajectories, such as Average Displacement Error (**ADE**) and Final Displacement Error (**FDE**). We utilize **FDE** as the criteria since the last position error can partially reflect the similarity with less bias from averaging compared with **ADE**.

Matching Strategy: There are roughly four ways used for matching, namely forward matching, backward matching, bidirectional matching, and Hungarian matching. Forward matching takes one trajectory in the current frame and finds its corresponding trajectory in the next frame with the least cost or maximum similarity. Backward matching is the reverse way compared to forward matching. Furtherly, bidirectional matching consists of both forward and backward matching, which considers the dual relationship. Hungarian matching is a linear optimal matching solution based on linear assignment. Forward and backward matching only considers the one-way situation, which is sensitive to noise and unstable. Hungarian matching has a high requirement for cost function choice. Based on these observations, we choose bidirectional matching as our strategy. We also show its advantages over the other approaches in Sec. 4.3. After obtaining the optimal matching pairs $\{(m_k, n_k)|1 \leq k \leq K\}$, we can compute the consistency constraint by a simple smooth L_1 loss [40] $\mathcal{L}_{Huber}$:

$$\mathcal{L}_{\text{temp}} = \sum_{k=1}^{K} \sum_{t=s+1}^{T} \mathcal{L}_{\text{Huber}}((x_t^{m_k}, y_t^{m_k}), (x'^{n_k}_{t-s}, y'^{n_k}_{t-s})). \tag{3}$$

3.2.2 Spatial Consistency

Since our MISC is a two-stage framework, the second stage mainly aims for trajectory refinement. It will be more convenient to add spatial permutation in the second stage with less computational cost. First, we apply spatial permutation function Z, including flipping and random noise, to the trajectories from the first stage. The refinement module will process these augmented inputs and generate the offset to the ground truth and classification scores. Under the small spatial permutation and disturbance, we assume that the outputs of the network should also be self-consistent, meaning that the outputs have strong stability or tolerance to noise. Compared with data augmentation, it is explicit regularization. Then the spatial consistency constraint $\mathcal{L}_{spa}$ is as follows:

$$\mathcal{L}_{\text{spa}} = \mathcal{L}_{\text{Huber}}(\Delta_{\tau_{reg}}, Z^{-1}(Reg(Z(\tau_{reg}, \tau_{\text{history}})))). \tag{4}$$

Then the total loss for Dual Consistency Constraints module will be $\mathcal{L}_{cons} = \mathcal{L}_{spa} + \mathcal{L}_{temp}$.

3.3. Teacher-Target Constraints

Teacher-Target Constraints enforce intra-model consistency that not only leverages the power of knowledge distillation but also helps alleviate the multi-modality supervision problem. Existing datasets [9, 49] only provide a single ground-truth trajectory for the target agent, which is to be predicted in one scenario. In order to encourage the multi-modality of models, the winner-takes-all (WTA) strategy is commonly used to prevent the model from collapsing into a single domain. However, the WTA training strategy suffers from instability associated with network initialization. Some other approaches [4, 35] introduce robust estimation methods to select better hypotheses. To some extent, these methods can only implicitly model the multi-modality. Some other approaches [4, 58] generate several possible future trajectories based on the kinematics model and road graph topology. DenseTNT [18]

only uses teacher labels to predict goal sets through a hill-climbing algorithm. These optimization methods tend to impose strict constraints and handcrafted prior knowledge, resulting in inaccurate teacher-targets and inferior performance. In contrast, our approach aims to generate more accurate teacher targets to provide explicit multi-modality supervision through self-ensembling to leverage the power of semi-supervised learning and knowledge distillation.

Teacher-Target Generation. The key part of our approach lies in generating more accurate teacher labels for each agent. However, it is straightforward to apply model ensembling techniques [20, 24, 50] to obtain more powerful predictions. Compared with previous works [4, 7, 58], we do not rely on handcrafted anchor trajectory sampling, which is based on inaccurate prior knowledge, including motion estimation. Meanwhile, soft targets from ensembling can better finetune the predictions and reduce the gradient variance for better training convergence. As suggested in works [10, 37], the prediction error decreases when the ensemble approach is used once the model is diverse enough. Therefore, we apply k-means algorithm [31] to the predicted trajectories that are collected within different training procedures (for example, launched with different seeds of random number generators, optimized with different learning rates, etc.) of MISC without Teacher-Target Constraints to generate J trajectories with corresponding scores for each scenario. Fig. 3 shows the overall process of our approach. Then with the original ground-truth label, we will formulate $J+1$ target trajectories as follows:

$$\tau_{conf} = \{c_0, c_1, \ldots, c_J\}, \tag{5}$$

$$\tau_{tgt}^{j} = \{(x_1^{tgt_j}, y_1^{tgt_j}), (x_2^{tgt_j}, y_2^{tgt_j}), \ldots, (x_T^{tgt_j}, y_T^{tgt_j})\}, \tag{6}$$

where τ_{tgt}^{j} is the j-th trajectory with score c^j, among $J+1$ target trajectories. To simplify the notation, τ_{tgt}^{0} is the ground-truth trajectory with c_0 set to 1.

3.4. Learning

The total supervision of our MISC can be decoupled into several parts, as described in previous sections. For the regression and classification parts, we loop over all the possible $J+1$ targets τ_{tgt}. For each target τ_{tgt}^{j} with confidence τ_{conf}^{j}, we apply WTA strategy as described in Sec. 3.3. Suppose k^*-th trajectory from trajectory refinement output $\tau_{reg'}$ is the best trajectory which has the maximum similarity with target τ_{tgt}^{j}, the classification loss and regression loss are defined as:

$$\mathcal{L}_{cls}^{j} = \frac{1}{K}\sum_{k=1}^{K} \tau_{conf}^{j} \mathcal{L}_{Huber}(c^k, c^{k^*}), \tag{7}$$

$$\mathcal{L}_{reg}^{j} = \frac{1}{T}\sum_{t=1}^{T} \tau_{conf}^{j} \mathcal{L}_{Huber}((x_t^{k^*}, y_t^{k^*}), (x_t^{tgt_j}, y_t^{tgt_j})). \tag{8}$$

For classification loss design, we adopt the displacement prediction idea from TPCN [55] to alleviate the hard assignment phenomenon. As for converting the displacement into probability, we use the standard softmin function to distribute the scores. Since we have trajectory completion and refinement modules, the regression loss will be $\mathcal{L}_{reg} = \sum_{j=0}^{J}(\mathcal{L}_{reg}^{j} + \mathcal{L}_{\Delta reg}^{j})$, where $\mathcal{L}_{\Delta reg}^{j}$ is the regression loss for the refinement module. The final loss is $\mathcal{L} = \mathcal{L}_{reg} + \mathcal{L}_{cls} + \mathcal{L}_{cons}$.

4. Experiments

We conduct experiments on the Argoverse dataset [9], one of the largest publicly available motion forecasting datasets. We compare our MISC with other state-of-the-art methods. Furthermore, we provide ablation studies to evaluate the effectiveness and generalization ability of each proposed module and design experiments for some hyper-parameter choices.

4.1. Experimental Setup

Dataset. Argoverse [9] provides more than 300K scenarios with rich HDMap information. For each scenario, objects are divided into three types: agent, AV and others, where "agent" is the object to be predicted. Moreover, each scenario contains 50 frames sampled at 10 Hz, meaning that the time interval between successive frames is 0.1s. The whole dataset is split into training, validation, and test sets, with 205, 942, 39, 472, and 78, 143 sequences, respectively. Waymo open motion dataset (WOD) contains multiple types of agents including vehicles, pedestrians, and cyclists. A total of more than 100,000 segments are provided with more than 1500 km of roadway coverage.

Metrics. We use the standard evaluation metrics, including ADE and FDE. ADE is defined as the average displacement error between ground-truth trajectories and predicted trajectories over all time steps. FDE is defined as displacement error between ground-truth trajectories and predicted trajectories at the last time step. We predict K candidate trajectories for each scenario and calculate the metrics with the ground truth labels. Accordingly, minADE and minFDE are minimum ADE and FDE over the top K predictions. Moreover, miss rate (MR) is also considered, defined as the percentage of the best-predicted trajectories whose FDE is within a threshold (2m). Brier-minFDE is the minFDE plus

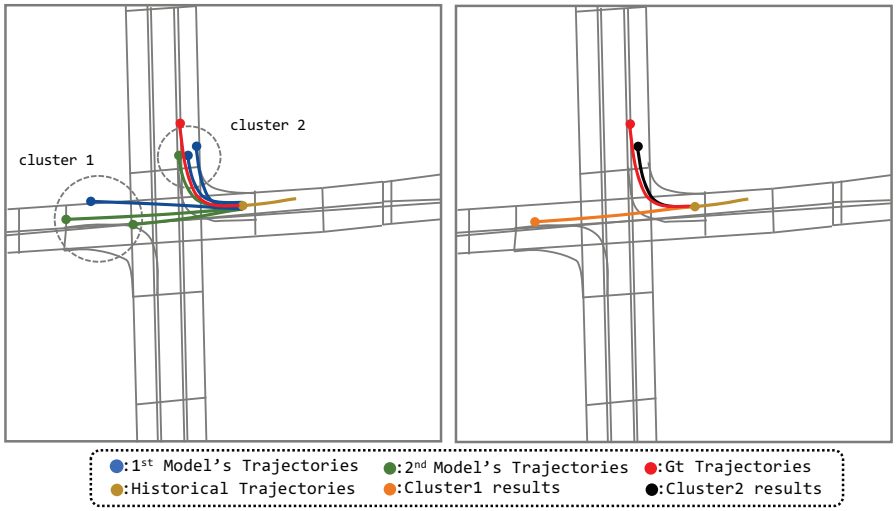

Figure 3. The overall procedure for the teacher-target generation. We obtain multiple predictions from outputs of different models for the target agents in each scenario; then we apply the K-means clustering algorithm to ensemble the trajectories

$(1-p)^2$, where p is the corresponding trajectory probability. Metrics for $K = 1$ and $K = 6$ are used in our experiments. Note that Brier-minFDE$_6$ is the ranking metric.

Experimental Details. We apply some data augmentation, including random flipping with a probability of 0.5 and global random scaling with a scaling ratio between $[0.8, 1.25]$ during the training stage. As for model settings, the time shift s for the temporal consistency constraint is set to 1. We adopt $K = 6$ to generate 6 trajectories and use $J = 6$ teacher targets for each scenario. Furthermore, we choose bidirectional matching for temporal consistency constraint. We finally use 10 models for ensembling due to computation resource limits. For more training details, we have included them in the supplementary materials.

4.2. Experimental Results

4.2.1 Results on Argoverse Dataset

Argoverse Leaderboard Results. We provide detailed quantitative results of our MISC on the Argoverse test set as well as public state-of-the-art methods in Tab. 1. Compared with previous methods, our MISC improves all the evaluation metrics except MR$_6$ by a large margin. Furtherly, since the proposed modules are all general training components, other existing motion forecasting models can also benefit greatly from these strategies.

Qualitative Results. We also present some qualitative results on the Argoverse validation set in Fig. 4. Compared with results without consistency, the Dual Consistency Constraints improve both the quality and smoothness of the predicted trajectories significantly, resulting in more feasible and stable results despite the input noise.

4.2.2 Results on Waymo Open Motion Dataset

We provide some quantitative results on the validation set of the Waymo Open dataset motion prediction task [12], shown in Tab. 2. Compared with KEMP [30] and Scene-Transformer [36], we also achieve very promising results and show comparable improvement, demonstrating the effectiveness of our approach. We also provide some ablation studies on WOD in the supplementary materials.

4.3. Ablation Studies

Component Study. As shown in Tab. 3, we conduct an ablation study for our MISC on the Argoverse validation set to evaluate the effectiveness of each proposed component. We adopt TPCN [55] as the baseline shown in the first row of Tab. 3 and add the proposed components progressively. The architecture modifications from the goal set prediction and trajectory refinement module show their promising improvements of about 2%. Dual consistency Constraints have the largest improvements of more than 5% among all the evaluation metrics. Especially for minFDE$_1$, temporal consistency can optimize 20 cm, indicating the temporal constraints can improve both final position and trajectory probability prediction. Compared with temporal consistency, spatial consistency has less effect on models since we only enforce this constraint in the trajectory refinement stage. Finally, the Teacher-Target Constraints significantly increase performance, manifesting its effectiveness in helping training convergence.

Temporal Consistency Factors. We study the factors in the matching problems, including similarity and matching strategies. As shown in Tab. 4, both Hungarian and Bidirectional matching show their advantages over the single

Models	minADE_1	minFDE_1	MR_1	minADE_6	minFDE_6	MR_6	b-FDE_6
Jean [9, 34]	1.74	4.24	0.68	0.98	1.42	0.13	2.12
LaneConv [28]	1.71	3.78	0.59	0.87	1.36	0.16	2.05
LaneRCNN [56]	1.68	3.69	0.57	0.90	1.45	0.12	2.15
mmTransformer [29]	1.77	4.00	0.62	0.87	1.34	0.15	2.03
SceneTransformer [36]	1.81	4.06	0.59	0.80	1.23	0.126	1.88
TNT [58]	1.77	3.91	0.59	0.94	1.54	0.13	2.14
DenseTNT [18]	1.68	3.63	0.58	0.88	1.28	0.125	1.97
PRIME [46]	1.91	3.82	0.59	1.22	1.55	0.12	2.09
TPCN [55]	1.58	3.49	0.56	0.88	1.24	0.13	1.92
HOME [16]	1.70	3.68	0.57	0.89	1.29	**0.08**	1.86
MultiPath++ [51]	1.623	3.614	0.564	0.790	1.214	0.13	1.793
Hivt++ [61]	1.56	3.44	0.563	0.767	1.146	0.12	1.817
Ours	**1.476**	**3.251**	**0.532**	**0.766**	**1.135**	0.11	**1.756**

Table 1. The detailed results of our MISC and other top-performing approaches on the Argoverse test set. And b-FDE_6 is the abbreviation of brier-minFDE_6

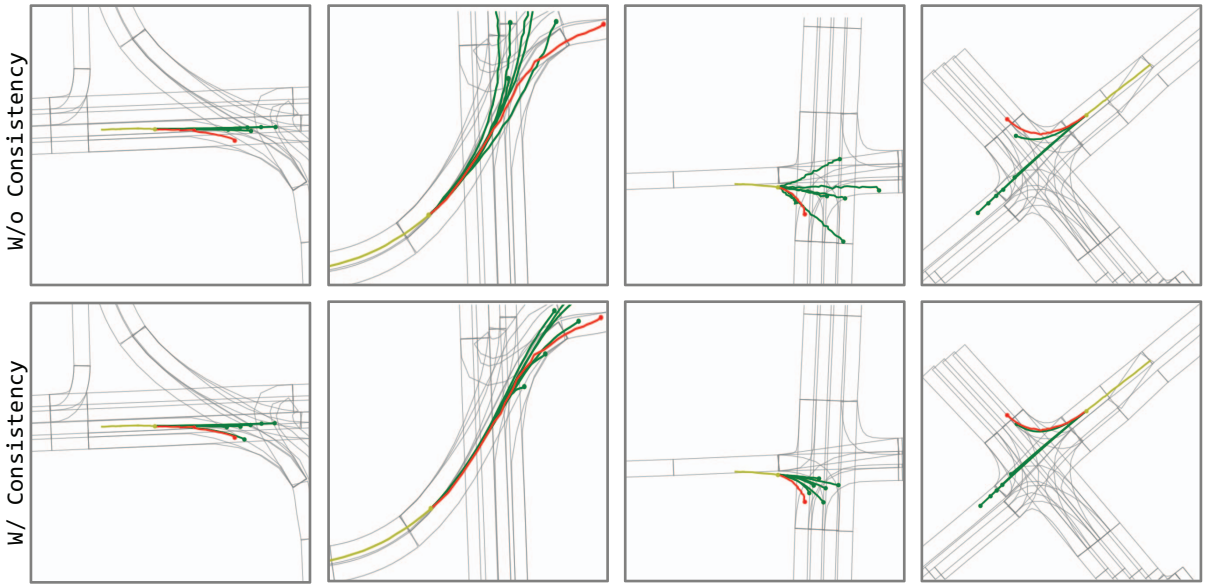

Figure 4. The past trajectory is in yellow, the predicted trajectory is in green, and the ground truth is in red. The top row of the figure shows the results without consistency, while the bottom row shows the results with consistency

Method	minADE↓	minFDE↓	Miss Rate↓	mAP↑
Baseline [12]	0.675	1.349	0.183	0.268
KEMP [30]	0.5691	1.1993	0.1458	0.394
SceneTrans [36]	0.613	1.22	0.157	0.284
Ours	**0.54**	**1.11**	**0.128**	**0.41**

Table 2. Quantitative results on the validation set of the Waymo Open dataset motion prediction task.

direction matching. Although Hungarian matching can ensure the one-to-one matching relationship, it is sensitive to the similarity metric and numerical precision, both of which are not stable in the early training stage. In contrast, bidirectional matching with the FDE similarity metric nearly achieves the best results across all the evaluation metrics. Meanwhile, we also conduct experiments to find the best time-shift value s in the temporal consistency. The details can be found in the supplement.

Reduction on Temporal Inconsistency. We use the average L2 distance among all predicted trajectory waypoints to measure the temporal consistency. As shown in Fig. 5, our model without temporal consistency will have large incon-

Architecture		Consistency		TTC	K=1		K=6	
Goal	Ref.	Temp.	Spatial		minADE	minFDE	minADE	minFDE
					1.34	2.95	0.73	1.15
✓					1.33	2.91	0.725	1.10
✓	✓				1.31	2.89	0.71	1.07
✓	✓	✓			1.24	2.70	0.662	0.981
✓	✓	✓	✓		1.22	2.67	0.653	0.954
✓	✓			✓	1.26	2.77	0.69	1.01
✓	✓	✓	✓	✓	**1.19**	**2.60**	**0.640**	**0.929**

Table 3. Ablation study results of modules. Goal refers to Trajectory completion with goal prediction. "Ref." is the trajectory refinement module, and the "Temp." is temporal consistency. TTC refers to Teacher-Target Constraints during training

Matching Strategy	Similarity	K=1			K=6		
		minADE	minFDE	MR	minADE	minFDE	MR
Forward	ADE	1.25	2.70	0.46	0.670	0.982	0.089
	FDE	1.24	2.69	0.46	0.668	0.980	0.088
Backward	ADE	1.25	2.70	0.46	0.670	0.982	0.089
	FDE	1.24	2.68	0.46	0.667	0.958	0.085
Bidirectional	ADE	**1.22**	**2.67**	0.446	0.666	0.972	0.087
	FDE	**1.22**	**2.67**	**0.445**	**0.653**	**0.954**	**0.084**
Hungarian	ADE	1.24	2.69	0.46	0.668	0.975	0.088
	FDE	1.23	2.69	0.45	0.660	0.968	0.088

Table 4. Ablation study on matching factor for temporal consistency. In this experiment, we remove the Teacher-Target Constraints to fairly study the effect

Teacher Target Num	K=1			K=6		
J	minADE	minFDE	MR	minADE	minFDE	MR
1	1.29	2.82	0.50	0.70	1.03	0.104
3	1.28	2.80	0.48	**0.69**	1.02	0.10
6	**1.26**	**2.77**	**0.47**	**0.69**	**1.01**	**0.09**

Table 5. Ablation study results on the teacher target number J

Method	Consistency	TTC	K=1		K=6	
			minADE	minFDE	minADE	minFDE
LaneGCN [28]	×	×	1.35	2.97	0.71	1.08
	✓	×	**1.29**	**2.80**	**0.68**	**1.00**
	×	✓	**1.30**	**2.88**	**0.69**	**1.04**
TPCN [55]	×	×	1.34	2.95	0.73	1.15
	✓	×	**1.27**	**2.79**	**0.69**	**1.04**
	×	✓	**1.30**	**2.86**	**0.69**	**1.09**
mmTransformer [29]	×	×	1.38	3.03	0.71	1.15
	✓	×	**1.31**	**2.83**	**0.68**	**1.02**
	×	✓	**1.29**	**2.80**	**0.68**	**1.04**
DenseTNT [18]	×	×	1.36	2.94	0.73	1.05
	✓	×	**1.25**	**2.81**	**0.68**	**0.98**
	×	✓	**1.30**	**2.82**	**0.69**	**1.00**

Table 6. Ablation study of consistency constraints and Teacher Target Constraints on different state-of-the-art methods on Argoverse validation set. Performance for methods without constraints is obtained from corresponding papers or our reproduction

sistency even though the time shift s is small, which may lead to unstable behavior for the downstream task such as planning. With temporal consistency constraints, there is a significant improvement in the L2 distance divergence, demonstrating the effectiveness of our method.

Number of Teacher Targets. As shown in Tab. 5, more teacher targets could bring better performance. Compared with $J = 1$, 6 teacher targets bring an extra nearly 1% improvements. However, the marginal improvement decreases

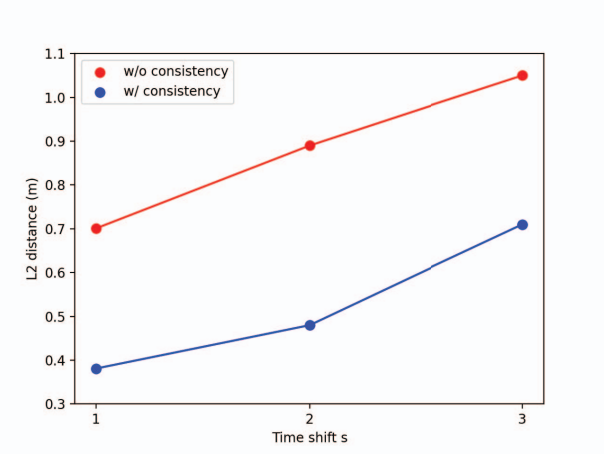

Figure 5. The L2 distance which varies with the time shift s.

significantly so we finally choose $J = 6$.

4.4. Generalization Capability

To verify the generalization capability of Dual Consistency Constraints and Teacher Targets Constraints, we also apply them to different models with state-of-the-art performance to show that they can be plugin-in training schemes.

Consistency Component. As shown in Tab. 6, our dual consistency constraints can effectively improve the performance of models regardless of their representations through the training phase. There is a noticeable improvement of over 5% on every metric, especially for minFDE.

Teacher Target. Teacher-Target Constraints is another general training trick that can be widely used in other frameworks. In Tab. 6, we also verify its effectiveness on other public methods. Methods with Teacher-Target Constraints have nearly over 3% improvement in all metrics. For the original DenseTNT [18], we replace its original handcrafted optimization for teacher goal targets with our self-ensembling teacher targets. This strategy brings an over 5% increase in performance, demonstrating the better quality of the self-ensembling teacher targets than handcrafted optimizations and estimation.

5. Conclusion

In this work, we propose MISC, an effective architecture for the motion forecasting task. We impose inner-model dual consistency regularization on both spatial and temporal domains to leverage the potential of self-supervision, which has been ignored by previous efforts. Besides, we explicitly model the multi-modality by providing supervision and constraints with powerful self-ensembling techniques in an intra-model aspect. Experimental results on the Argoverse motion forecasting dataset and Waymo dataset show the effectiveness of our approach and generalization capability to other methods.

References

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27:3365–3373, 2014. 2

[2] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet. In *Robotics: Science and Systems XV*, 2019. 1, 2

[3] Prarthana Bhattacharyya, Chengjie Huang, and Krzysztof Czarnecki. Ssl-lanes: Self-supervised learning for motion forecasting in autonomous driving. *arXiv preprint arXiv:2206.14116*, 2022. 2

[4] Antonia Breuer, Quy Le Xuan, Jan-Aike Termöhlen, Silviu Homoceanu, and Tim Fingscheidt. Quo vadis? meaningful multiple trajectory hypotheses prediction in autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 637–644. IEEE, 2021. 1, 2, 4, 5

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 3

[6] Sergio Casas, Wenjie Luo, and Raquel Urtasun. Intentnet: Learning to predict intention from raw sensor data. In *Conference on Robot Learning*, pages 947–956, 2018. 2

[7] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*, 2019. 1, 2, 5

[8] Titas Chakraborty, Akshay Bhagat, and Henggang Cui. Improving motion forecasting for autonomous driving with the cycle consistency loss. *arXiv preprint arXiv:2211.00149*, 2022. 2

[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 Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019. 1, 2, 4, 5, 7

[10] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000. 5

[11] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015. 2

[12] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021. 2, 6, 7

[13] Haoyang Fan, Fan Zhu, Changchun Liu, Liangliang Zhang, Li Zhuang, Dong Li, Weicheng Zhu, Jiangtao Hu, Hongye Li, and Qi Kong. Baidu apollo em motion planner. *arXiv preprint arXiv:1807.08048*, 2018. 1

[14] Peter Földiák. Learning invariance from transformation sequences. *Neural computation*, 3(2):194–200, 1991. 2

[15] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11533, 2020. 1, 2

[16] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Home: Heatmap output for future motion estimation. *arXiv preprint arXiv:2105.10968*, 2021. 2, 3, 7

[17] 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. 2

[18] 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, 2, 3, 4, 7, 8

[19] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *Advances in neural information processing systems*, 25, 2012. 2

[20] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3d object detection from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11873–11882, 2020. 5

[21] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015. 2

[22] Adam Houenou, Philippe Bonnifait, Véronique Cherfaoui, and Wen Yao. Vehicle trajectory prediction based on motion model and maneuver recognition. In *2013 IEEE/RSJ international conference on intelligent robots and systems*, pages 4363–4369. IEEE, 2013. 2

[23] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *J. Basic Eng*, 82(1):35–45, 1960. 2

[24] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 5

[25] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2

[26] Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. In *Advances in Neural Information Processing Systems*, pages 2119–2127, 2016. 1

[27] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[28] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representa-

tions for motion forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 541–556, 2020. 1, 2, 7, 8

[29] 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, 7, 8

[30] Qiujing Lu, Weiqiao Han, Jeffrey Ling, Minfa Wang, Haoyu Chen, Balakrishnan Varadarajan, and Paul Covington. Kemp: Keyframe-based hierarchical end-to-end deep model for long-term trajectory prediction. *arXiv preprint arXiv:2205.04624*, 2022. 6, 7

[31] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967. 5

[32] Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7144–7153, 2019. 2

[33] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. *arXiv preprint arXiv:2004.02025*, 2020. 2

[34] Jean Mercat, Thomas Gilles, Nicole El Zoghby, Guillaume Sandou, Dominique Beauvois, and Guillermo Pita Gil. Multi-head attention for multi-modal joint vehicle motion forecasting. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9638–9644. IEEE, 2020. 7

[35] Sriram Narayanan, Ramin Moslemi, Francesco Pittaluga, Buyu Liu, and Manmohan Chandraker. Divide-and-conquer for lane-aware diverse trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15799–15808, 2021. 1, 4

[36] 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 multi-task model for behavior prediction and planning. *arXiv preprint arXiv:2106.08417*, 2021. 1, 6, 7

[37] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198, 1999. 5

[38] Hao Ouyang, Tengfei Wang, and Qifeng Chen. Internal video inpainting by implicit long-range propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14579–14588, October 2021. 2

[39] Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. Covernet: Multimodal behavior prediction using trajectory sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2020. 2

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *International Conference on Neural Information Processing Systems*, 2015. 3, 4

[41] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3591–3600, 2017. 2

[42] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171, 2016. 2

[43] Jens Schulz, Constantin Hubmann, Julian Löchner, and Darius Burschka. Interaction-aware probabilistic behavior prediction in urban environments. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3999–4006. IEEE, 2018. 2

[44] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013. 2

[45] Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. Tangent prop-a formalism for specifying selected invariances in an adaptive network. *Advances in neural information processing systems*, 4, 1991. 2

[46] Haoran Song, Di Luan, Wenchao Ding, Michael Y Wang, and Qifeng Chen. Learning to predict vehicle trajectories with model-based planning. In *Conference on Robot Learning*, pages 1035–1045. PMLR, 2021. 2, 7

[47] NN Sriram, Gourav Kumar, Abhay Singh, M Siva Karthik, Saket Saurav, Brojeshwar Bhowrnick, and K Madhava Krishna. A hierarchical network for diverse trajectory proposals. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 689–694. IEEE, 2019. 2

[48] DiJia Andy Su, Bertrand Douillard, Rami Al-Rfou, Cheol Park, and Benjamin Sapp. Narrowing the coordinate-frame gap in behavior prediction models: Distillation for efficient and accurate scene-centric motion forecasting. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 653–659. IEEE, 2022. 2

[49] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 1, 4

[50] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 5

[51] 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. *arXiv preprint arXiv:2111.14973*, 2021. 1, 7

[52] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2001–2010, October 2021. 2

[53] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 2

[54] Guotao Xie, Hongbo Gao, Lijun Qian, Bin Huang, Keqiang Li, and Jianqiang Wang. Vehicle trajectory prediction by integrating physics-and maneuver-based approaches using interactive multiple models. *IEEE Transactions on Industrial Electronics*, 65(7):5999–6008, 2017. 2

[55] 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. 1, 2, 3, 5, 6, 7, 8

[56] Wenyuan Zeng, Ming Liang, Renjie Liao, and Raquel Urtasun. Lanercnn: Distributed representations for graph-centric motion forecasting. *arXiv preprint arXiv:2101.06653*, 2021. 1, 2, 7

[57] Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-to-end interpretable neural motion planner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8660–8669, 2019. 2

[58] 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. 1, 2, 3, 4, 5, 7

[59] Wu Zheng, Weiliang Tang, Li Jiang, and Chi-Wing Fu. Se-ssd: Self-ensembling single-stage object detector from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14494–14503, 2021. 2

[60] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 2

[61] 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. 2, 7

[62] Julius Ziegler, Philipp Bender, Markus Schreiber, Henning Lategahn, Tobias Strauss, Christoph Stiller, Thao Dang, Uwe Franke, Nils Appenrodt, Christoph G Keller, et al. Making bertha drive—an autonomous journey on a historic route. *IEEE Intelligent transportation systems magazine*, 6(2):8–20, 2014. 2