# Supplementary Material for
# *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

## 1. Contents

In this supplementary material, we provide:

- More about our model details Sec. 2.
- More ablation Study in Sec. 3.
- Quantitative results of our approach in Sec. 4.

## 2. Model Details

We provide the detailed network architecture of our MISC in Fig. 1. We use TPCN [9] as our backbone. The feature extraction consists of 4 spatial modules and 4 dynamic temporal learning layers same as TPCN. Before the prediction header, we calculate the mean features and remove map instances features. For the spatial module, the point representation utilizes PointNet++ [8] with neighborhood radius of $[0.2m, 0.4m, 0.8m]$, while the voxel representation uses Sparse BottleNeck. We use all the points in this process without any sampling. More details about backbone can be found in TPCN [9].

### 2.1. Training Details

We train MISC for 50 epochs using a batch size of 32 with Adam [4] optimizer with an initial learning rate of 0.001, which is decayed every 15 epochs in a ratio of 0.1.

### 2.2. Model complexity

| Method | Param (M) | Speed (ms) |
|---|---|---|
| LaneGCN | 3.7 | 55 |
| DenseTNT | **1.1** | 40 |
| mmTransformer | 2.6 | **34** |
| Ours | 3.6 | 36 |

Table 1. The number of parameters and running time.

We provide detailed runtime speed evaluated in a single RTX2080Ti with the model parameters shown in Tab. 1. Compared with other state-of-the-art models, we achieve decent performance without introducing more computation cost.

*Work done during an internship at DeepRoute.AI.

| Time shift | K=1 | | | K=6 | | |
|---|---|---|---|---|---|---|
| $s$ | minADE | minFDE | MR | minADE | minFDE | MR |
| 1 | **1.22** | **2.67** | **0.444** | **0.653** | **0.954** | 0.084 |
| 2 | 1.23 | **2.67** | **0.444** | 0.654 | 0.958 | **0.082** |
| 3 | 1.25 | 2.69 | 0.445 | 0.662 | 0.964 | 0.085 |
| 4 | 1.25 | 2.70 | 0.446 | 0.667 | 0.969 | 0.086 |

Table 2. Ablation study results of time-shift $s$ used by temporal consistency

## 3. Ablation study

### 3.1. Temporal consistency

Meanwhile, we also conduct experiments to find the best time-shift value $s$ in the temporal consistency. As shown in Tab. 2, choosing time shift $s = 1$ has already achieved decent performance, with five out of six metrics ranking the first. Further increasing the $s$ will not bring much performance gain since the driving behavior could change a lot with large $s$.

### 3.2. Spatial consistency

Furthermore, we also measure the spatial inconsistency against flipping and Gaussian noise with zero mean and standard deviation of 15cm. The average spatial inconsistency will be 19.3cm, while the number decreases to 10.2cm with our spatial consistency constraint.

### 3.3. Component Study

We provide a controlled experiment to verify the effectiveness of the proposed method when turning both Dual Consistency Constraints and Teacher-Target Constraints on at the same time shown in Tab. 3. With both modules on, the performance of all the methods benefits a lot, about nearly 7%, demonstrating the generalization capability and effectiveness of our approach. It also shows that these two modules can be independently helpful.

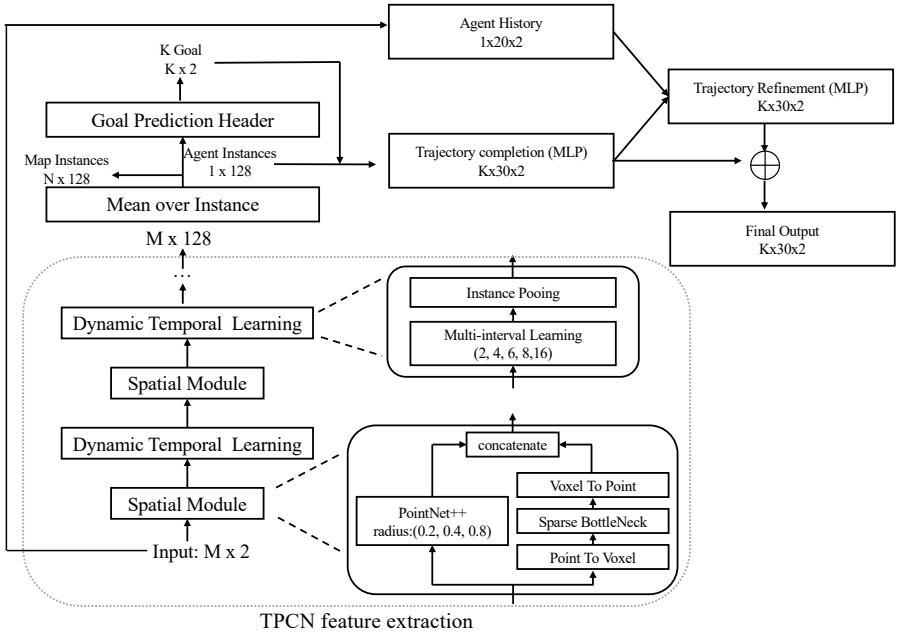


Figure 1. Detailed illustration of our MISC.

| Method | Consistency & TTC | K=1 | | K=6 | |
|---|---|---|---|---|---|
| | | minADE | minFDE | minADE | minFDE |
| LaneGCN [5] | × | 1.35 | 2.97 | 0.71 | 1.08 |
| | ✓ | **1.25** | **2.71** | **0.66** | **0.98** |
| TPCN [9] | × | 1.34 | 2.95 | 0.73 | 1.15 |
| | ✓ | **1.23** | **2.70** | **0.67** | **1.00** |
| mmTransformer [6] | × | 1.38 | 3.03 | 0.71 | 1.15 |
| | ✓ | **1.25** | **2.77** | **0.67** | **0.99** |
| DenseTNT [3] | × | 1.36 | 2.94 | 0.73 | 1.05 |
| | ✓ | **1.23** | **2.71** | **0.66** | **0.95** |

Table 3. Results of consistency constraints and Teacher-Target Constraints (TTC) supervision on different state-of-the-art methods on Argoverse validation set. Performance for methods without consistency constraints is obtained from corresponding papers or our reproduction.

### 3.4. Ablation Study on Waymo Dataset

Since the scale and object types in waymo dataset and argoverse dataset are different, we conduct experiments to find the best time shift $s$ for each class on Waymo Dataset. As shown in Tab. 4, best time shift for vehicle and cyclist will be 1, while the value will be 2 for pedestrian class. To achieve the best performance for the overall metrics, we finally choose $s = 1$ in our setting.

### 3.5. Results on ETH Dataset

To verify the temporal consistency on the low framerate dataset, we conduct experiments on the ETH [7] dataset. We report the **ADE** and **FDE** metrics for $t_{pred} = 8$ and $t_{pred} = 12$ respectively. Following the common settings used by previous methods [2], we use $K = 1$ and $K = 20$. As shown in Tab. 5, our temporal consistency significantly improves the performance. Choosing $s = 1$ works well in most of the evaluation metrics.

## 4. Qualitative Analysis

We provide some visual results of MISC on the the Argoverse [1] validation set in Fig. 3 as well as the Argoverse test set in Fig. 4. These qualitative results demonstrate the effectiveness and the high-quality predicted trajectories of our method.

### 4.1. Failure Cases

We also present some failure cases on the validation set in Fig. 2. Some possible reasons are:

- The ground-truth labels contain some noises. Since the ground-truth labels are obtained from tracking, there may be some id switches, leading to the sudden perturbation of the agents' location (e.g., the first and third example in the second row of Fig. 2). Under these scenarios, the predicted trajectories from MISC are more reasonable and stable without large jerks.

- The multi-modality problem. In some situations, MISC can not predict the intention perfectly without enough motion and map information. The first and third example in the first row of Fig. 2 demonstrate this phenomenon. The agent makes a lane change decision without many hints in the historical information. Thus, this can be furtherly improved by introducing more map constraints.

| Time shift | minADE↓ | | | minFDE↓ | | | MR↓ | | | mAP↑ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | veh | ped | cyc | veh | ped | cyc | veh | ped | cyc | veh | ped | cyc |
| 1 | **0.622** | 0.34 | **0.654** | **1.262** | 0.663 | 1.294 | **0.135** | 0.085 | **0.197** | 0.285 | **0.252** | **0.214** |
| 2 | 0.625 | **0.33** | 0.660 | 1.263 | **0.662** | 1.296 | **0.135** | **0.084** | 0.200 | **0.283** | **0.252** | 0.215 |
| 3 | 0.632 | 0.34 | 0.667 | 1.274 | 0.666 | 1.302 | 0.136 | 0.086 | 0.198 | 0.290 | 0.254 | 0.217 |
| 4 | 0.634 | **0.33** | 0.672 | 1.278 | 0.670 | 1.303 | 0.137 | 0.086 | 0.199 | 0.288 | 0.253 | 0.217 |

Table 4. Ablation study results of time-shift $s$ used by temporal consistency on Waymo Open Motion Dataset motion prediction

| Time shift | Dataset | K=1 | | K=20 | |
|---|---|---|---|---|---|
| | | ADE | FDE | ADE | FDE |
| 0 | **ETH** | 0.69 / 0.98 | 1.30 / 1.98 | 0.51 / 0.79 | 1.05 / 1.66 |
| | **HOTEL** | 0.27 / 0.33 | 0.46 / 0.55 | 0.20 / 0.25 | 0.36 / 0.44 |
| 1 | **ETH** | **0.65** / 0.93 | **1.22 / 1.86** | **0.47 / 0.73** | **0.97 / 1.55** |
| | **HOTEL** | **0.23** / 0.29 | **0.42** / 0.50 | **0.18 / 0.23** | **0.33 / 0.42** |
| 2 | **ETH** | **0.65 / 0.92** | 1.23 / 1.88 | 0.48 / **0.73** | 1.00 / 1.56 |
| | **HOTEL** | 0.24 / **0.27** | 0.43 / **0.49** | **0.18** / 0.25 | 0.34 / **0.42** |
| 3 | **ETH** | 0.66 / 0.93 | 1.24 / 1.89 | 0.48 / **0.73** | 0.98 / 1.57 |
| | **HOTEL** | 0.24 / 0.30 | 0.43 / 0.52 | 0.19 / 0.24 | 0.34 / 0.44 |
| 4 | **ETH** | 0.66 / 0.94 | 1.23 / 1.89 | 0.49 / 0.74 | 0.99 / 1.58 |
| | **HOTEL** | 0.25 / 0.31 | 0.44 / 0.51 | 0.20 / 0.25 | **0.33** / 0.44 |

Table 5. Ablation study results of time-shift $s$ used by temporal consistency on ETH Dataset

## References


[1] 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. 2

[2] Liangji Fang, Qinhong Jiang, Jianping Shi, and Bolei Zhou. Tpnet: Trajectory proposal network for motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6797–6806, 2020. 2

[3] 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. 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[5] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 541–556, 2020. 2

[6] 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. 2

[7] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European conference on computer vision*, pages 452–465. Springer, 2010. 2

[8] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017. 1

[9] 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

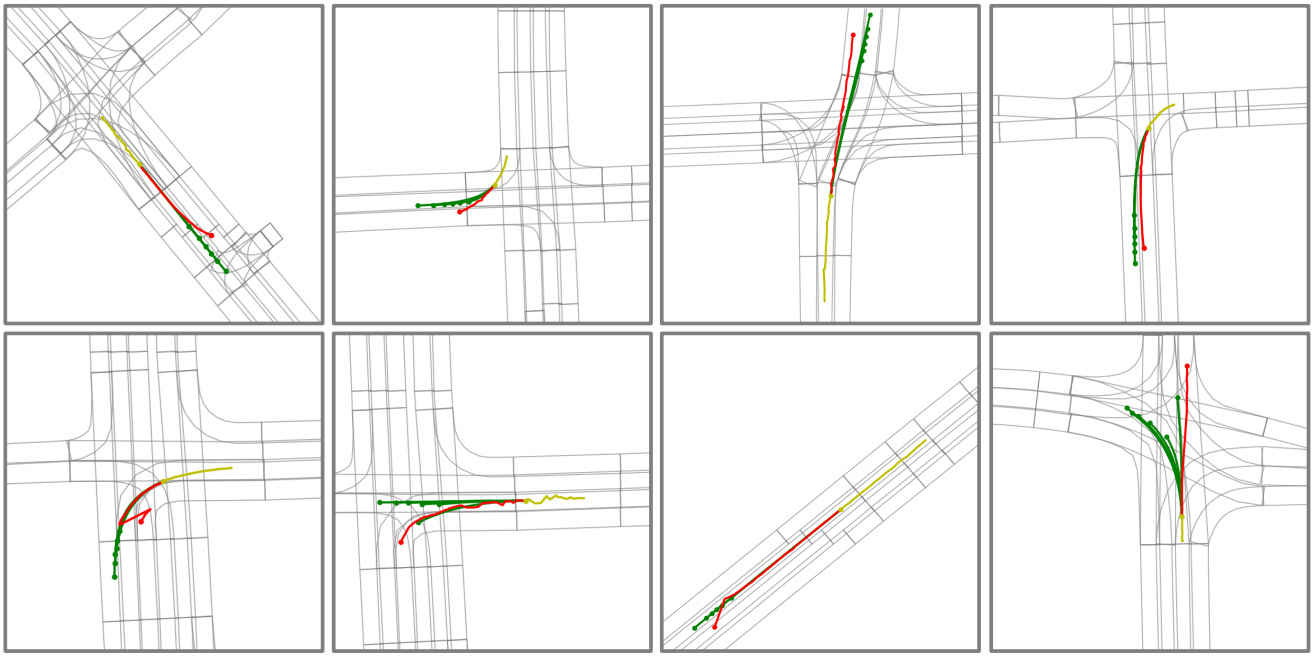

Figure 2. Failure cases on the Argoverse validation set. The target agent's past trajectory is in yellow, predicted trajectory in green, and ground truth in red.

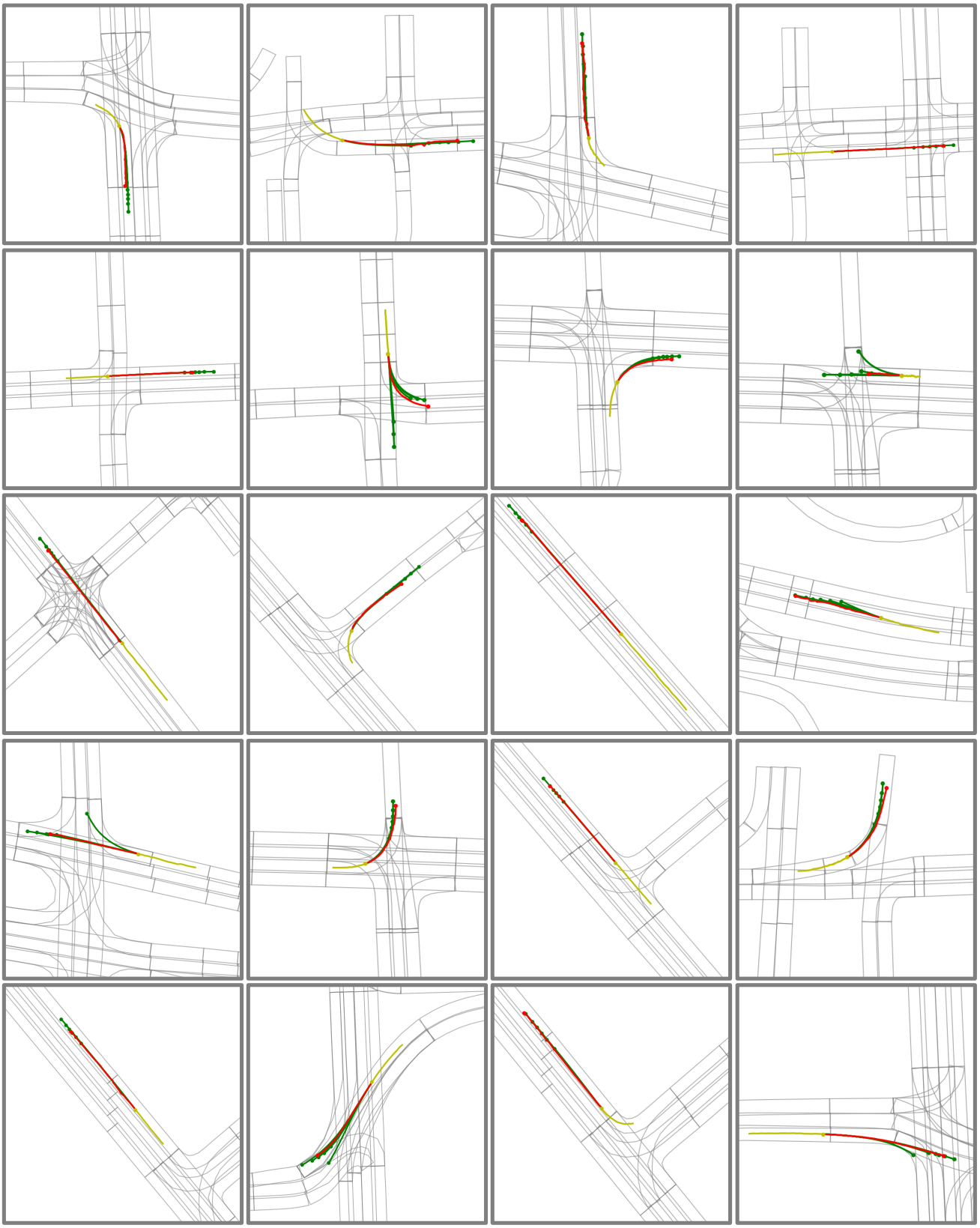

Figure 3. The motion forecasting results on the Argoverse validation set. The target agent's past trajectory is in yellow, predicted trajectory is in green, and ground truth is in red.

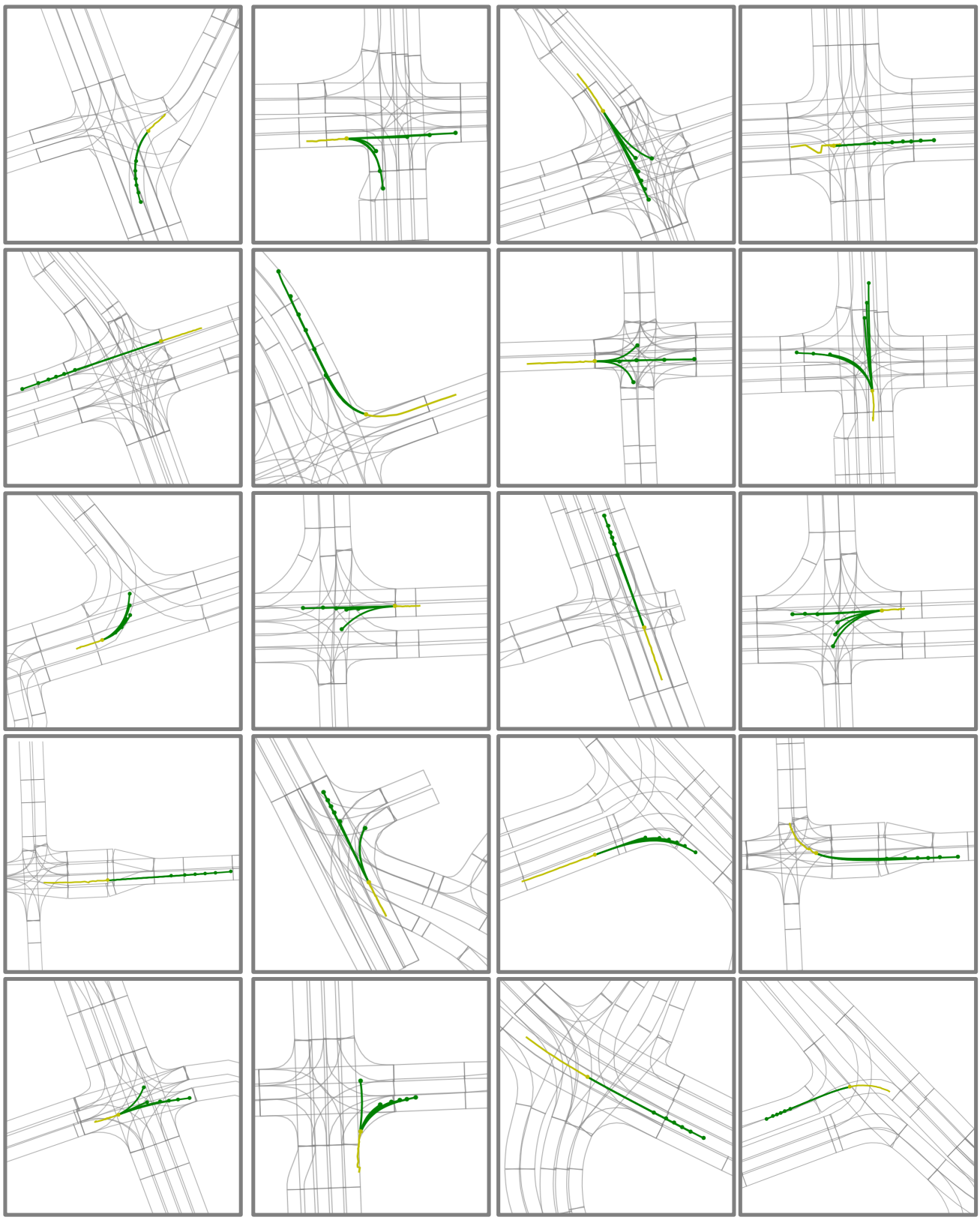

Figure 4. The motion forecasting results on the Argoverse test set. The target agent's past trajectory is in yellow and predicted trajectory in green.