Three Steps to Multimodal Trajectory Prediction: Modality Clustering, Classification and Synthesis

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

Abstract

Multimodal prediction results are essential for trajectory prediction task as there is no single correct answer for the future. Previous frameworks can be divided into three categories: regression, generation and classification frameworks. However, these frameworks have weaknesses in different aspects so that they cannot model the multimodal prediction task comprehensively. In this paper, we present a novel insight along with a brand-new prediction framework by formulating multimodal prediction into three steps: modality clustering, classification and synthesis, and address the shortcomings of earlier frameworks. Exhaustive experiments on popular benchmarks have demonstrated that our proposed method surpasses state-of-the-art works even without introducing social and map information. Specifically, we achieve 19.2% *and* 20.8% *improvement on ADE and FDE respectively on ETH/UCY dataset.*

1. Introduction

Trajectory prediction is one of the cornerstones of autonomous driving and robot navigation [22, 9, 27, 28, 35, 37, 12], which investigates reasonable future states of traffic agents for the following decision-making process. Considering the uncertainty of human behaviors and the multimodal nature of the future [8, 14], one great challenge of trajectory prediction lies in predicting all possible future trajectories of high probabilities.

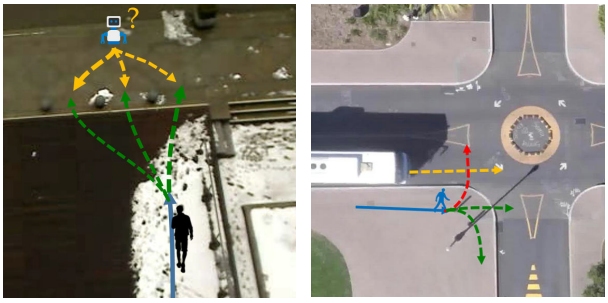
To tackle this problem, previous research mainly follows three lines. The first line adds extra randomness for regression frameworks [36, 18], while the second generation line [8, 14, 21, 29] models the multimodal nature by learning a distribution of the future. But both lines have two defects as shown in Fig. 1: i) lack of probability corresponding to each modality which may leave the decision process confusing, and ii) the prediction results are not deterministic which may leave potential safety risks.

The third line of classification frameworks [4, 26] gets rid of the two defects by classifying the observation to predefined future trajectories, as the classification operation can give probabilities and ensure determinacy. However, the classification frameworks still face certain weaknesses. First, the predefined trajectories are obtained by hand-crafted principles, thus it is difficult to capture comprehensive representations for future behaviors. Second, the predicted deterministic trajectories will be the same for different inputs classified to a same class, which fails to explore fine-grained motions for traffic agents deterministically. Due to these weaknesses, the performance of a classification framework lags behind state-of-the-art regression and generation models. Further, a highly annotated scene raster is required as the input of classifier which is difficult to access in many cases.

In this paper, we aim to explore a distinct formulation for trajectory prediction framework to address the shortcomings discussed above. We present the insight of **Prediction via modality Clustering, Classification and Synthesis** (PCCSNet) by solving multimodal prediction with a *classification-regression* approach. In our vision, the modalities of the future are usually centralized around a few different behaviors which can be revealed by a series of learned modality representations. We can apply a deep clustering process [3] on training samples and each center of clusters could represent a modality. Naturally, such a modality can be formulated into a class, and a classification network can be adopted to distinguish and score the modalities according to the observed trajectory in this manner. Finally, a synthesis process is used to regress prediction results for highly probable modalities with historical states and the modality representations.

We propose a modular designed framework to model this novel insight summarized in Fig. 2. States of agents are first fed into feature encoders to get deep historical and future representations for better clustering, classification and syn-

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



(a) (b)

Figure 1. Examples of why probability and determinacy of a prediction algorithm is important. In figure (a), a probabilistic prediction (indicated by the thickness of the lines) will significantly reduce the probability of collisions. Figure (b) illustrates a potential safety hazard of stochastic predictions, since the red path should always be predicted for safety purposes. A detailed discussion is in Sec. 2.

thesis [2]. These deep features are clustered for modality representations and used to train a classifier where the cluster assignments are seen as pseudo-labels. The classifier will score the modalities according to historical representations in test phase for probabilistic prediction. A synthesis module is then introduced to regress pseudo future representations for each modality, and finally both historical and synthesized future representations are decoded to get fine-grained deterministic predictions. Moreover, we newly propose a Modality Loss to enhance the capability of the classifier to identify multiple highly probable future modalities.

We conduct exhaustive experiments on multiple popular trajectory prediction benchmarks. In these experiments, our novel prediction framework exhibits high accuracy, great robustness and adequate projections for the future. Specifically, we achieve 19.2% and 20.8% improvement in average on ADE and FDE respectively on ETH [25]/UCY [13] datasets comparing with state-of-the-art method [36].

2. Background and Related Work

Trajectory prediction [1, 8, 31, 29, 14, 10, 18, 36, 21, 7] is proposed to forecast possible trajectories of an agent. It takes advantages from tracking [28, 24] and human interactions [37, 35, 16], and has many applications in the field of robotics and autonomous driving [9, 27, 32, 33], as the predicted result is an important guidance for decision making. Observing the multimodal nature of the future which can be interpreted as no single correct answer for the future [8], an important point lies in how to predict multiple highly probable trajectories. This task is named as multimodal trajectory prediction. Note that a small part of methods [5, 17] re-formulate this task by predicting probabilistic maps in pixel level. We mainly discuss prevailing approaches that outputs multiple possible trajectories of spatial coordinate system (meters) in real world in this paper.

Multimodal prediction task is non-trivial as a single input may map to multiple outputs. Early works [1, 22, 34, 12] ignore the multimodality of the future and only aim at predicting the most possible future trajectory. Recently, a great number of research proposes various frameworks to formulate this non-functional relationship. They mainly follow three common practices, regression, generation and classification.

Regression Frameworks. Regression models [1, 19] are first proposed to solve unimodal prediction tasks and show great performance. However, these encode-decode structures are not able to give multimodal predictions, and some methods address this defect by adding noise [36] or using random initialization [18]. Although multiple different predictions can be obtained by imposing randomness on the model, it is difficult for randomness to accurately model the multimodal nature of future.

Generation Frameworks. Some research considers the multimodality of the future as a distribution, formulates trajectory prediction as a distribution fitting and sampling problem, and introduces generative models to solve it. DESIRE [14] first introduces stochastic model to learn the distribution of future states, and generates diverse predictions by sampling plausible hypotheses from that distribution. Following this formulation, plenty of research [8, 21, 29] aims at designing different generative structures to pursue more reasonable outcomes and achieve the state-of-the-art performance.

Classification Frameworks. Some research [4, 26] attempts to use a classification network to solve this problem by classifying on predefined artificial modalities. Multipath [4] clusters a fixed set of anchor trajectories with mean square error distance, and classifies the input to these anchors. CoverNet [26] revises Multipath by manually designing anchors. Approaches under this framework face three main weakness. First, the predefined trajectories are obtained by subjectively designated clustering distance or manually designed anchors, thus it is difficult to capture the full range of future behaviors. Second, it is hard for these predefined trajectories to capture fine-grained motions. Further, both methods require a highly annotated scene raster as input which is difficult to access in many cases.

Probability and Determinacy. The properties of probability and determinacy are important for a multimodal prediction approach. i) Probability. The probabilistic property of future [11] are extremely helpful to improve the effectiveness of the ensuing decision-making process. In Fig. 1 (a), a probabilistic prediction can tell that the pedestrian is most unlikely to take the left path (on the reader's side). Therefore, the robot can follow the left path (on the reader's side)

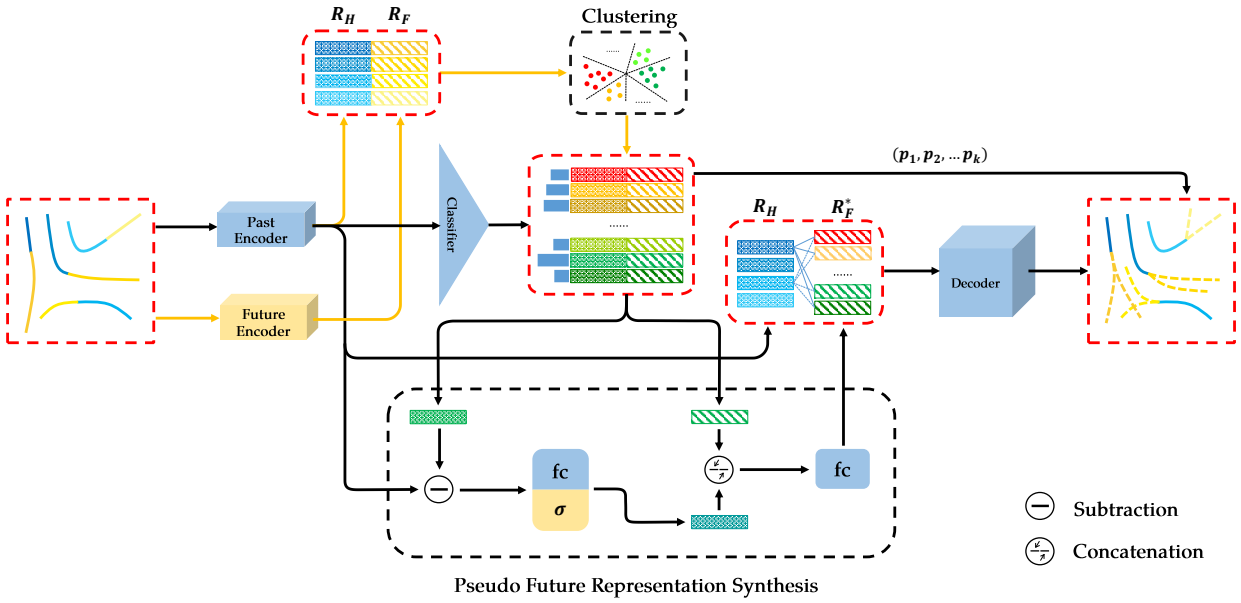


Figure 2. Overview of our proposed prediction pipeline. The arrows in yellow are only present in training, when both historical and future representations are concatenated and clustered for future usage. At test time, historical representations are fed into a classifier to score different modalities. The modality representations are then processed with the historical representations to synthesize multimodal predictions.

to avoid collision as far as possible. ii) Determinacy. A stochastic framework may leave huge potential safety risks. Fig. 1 (b) gives a common case in which a bus (autonomous vehicle) perceives a pedestrian at the crossroads when moving forward. A stochastic model cannot be proved to predict the trajectory in red every time. And if it fails, a traffic accident will happen. But most regression and generation models miss these two points, as equiprobable randomness is introduced. Previous approaches with classification frameworks can give probabilistic and deterministic predictions, but these frameworks fail in exploiting deep behavior representations and fine-grained motions as discussed above.

We propose a probabilistic and deterministic framework in this work which can still capture deep behavior representations and give predictions at fine-grained level. The great differences between our proposed framework and previous classification frameworks will be discussed in Sec. 3.7.

3. Approach

In this section, we introduce a newly proposed PCCSNet prediction pipeline, which is illustrated in Fig. 2. Our main insight is to formulate the multimodal prediction framework as a classification-regression process. It is designed to model multimodal trajectory task more comprehensively by addressing the shortcomings of prior frameworks.

3.1. Problem Definition

Following previous works [1, 8], we assume that each video is preprocessed by detection and tracking algorithms to obtain the spatial coordinates for each person at each timestep. We take the coordinate sequences X in time step $[1, T_{obs}]$ as input, and predict top k reasonable coordinate sequences $\hat{\mathbb{Y}} = \{\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k\}$ in $[T_{obs+1}, T_{obs+pred}]$ along with their probabilities $\mathbb{P} = \{p_1, p_2, \ldots, p_k\}$ as output.

3.2. Overview

In PCCSNet, we introduce an intermediate variable named modality representation M to formulate multimodal prediction framework into three steps of deep clustering, classification and synthesis. All of the possible modality representations can be obtained by clustering deep historical representations R_H and future representations R_F of training samples and gathered into a modality set $\mathbb{M}$

$$R_H = f_H(X), R_F = f_F(Y) \tag{1}$$

$$\begin{aligned}\mathbb{M} =& clustering(\{[R_H^i, R_F^i] | i \in trainset\}) \\ =& \{M_1, M_2, \ldots, M_n\}\end{aligned} \tag{2}$$

where past encoder $f_H(\cdot)$ and future encoder $f_F(\cdot)$ are trained to learn better representations of historical trajectory X and future trajectory Y following [2].

Then, through modality classification and modality synthesis, we can acquire $\hat{\mathbb{Y}}$ along with its probability $\mathbb{P}$ by

$$\mathbb{P} = g_{\mathbb{M}}(R_H) \tag{3}$$

$$\hat{\mathbb{Y}} = \{\hat{Y}_i = h([R_H, M_i]) | i \in [1, n]\} \tag{4}$$

where $g_{\mathbb{M}}(\cdot)$ represents modality classification on $\mathbb{M}$ and $h(\cdot)$ represents modality synthesis. In this manner, we can

predict probabilistic multimodal future trajectories deterministically. Note that we often predict k ($k < n$) future paths with the top probabilities in practical terms to reduce the test time.

In the following sections, we will introduce how we cluster and train the classification network $g_{\mathbb{M}}(\cdot)$ in Sec. 3.3. In Sec. 3.4, we propose a novel Modality Loss to encourage the classifier to recognize multiple reasonable futures comprehensively instead of just the most likely one. Finally, we show how to synthesize one prediction result by $h(\cdot)$ in Sec. 3.5.

3.3. Classification with Modality Clustering

Following Eq. 2 and Eq. 3, we need to construct modality set $\mathbb{M}$ and corresponding classifier $g_{\mathbb{M}}(\cdot)$.

Feature Encoder. To capture better representations for deep clustering, classification and further synthesis procedure, we first encode the historical and future trajectories for each agent. Given that a trajectory is a time series and has a strong dependency and consistency between each time step according to [30], we adopt the BiLSTM architecture as our feature encoders.

Clustering. We believe that each modality of trajectory indicates behaviors and movements of the same kind, and in turn, we can express a modality representation M with the average of a series of deep trajectory features, which can be written as

$$M = AVG(\{[R_H^i, R_F^i] | i \in C\}) \tag{5}$$

where C is a cluster represented by trajectory ids. In our implementation, we use the clustering center of C to represent the AVG operation, and Eq. 5 can be rewritten as

$$M = [R_H^c, R_F^c] \tag{6}$$

where R_H^c and R_F^c are the values of historical and future representations in the clustering center. In this way, we create a bridge between modality construction and clustering.

To generate distinct Cs, we introduce a clustering algorithm. Considering the definition of M, we use $[R_H^i, R_F^i]$ as features for path id i and a weighted L2 distance for clustering. Specifically, the distance is written as

$$\mathcal{D} = w_H ||R_H^1 - R_H^2||_2 + w_F ||R_F^1 - R_F^2||_2 \tag{7}$$

where w_H and w_F represent the weight for historical representations and future representations respectively.

Here we assume that the mean value and distribution of the training set are similar as those of the test set, which is usually how it works. The modality set $\mathbb{M}$ we construct will very nicely cover different kinds of future possibilities for test samples, as shown in Fig. 6.

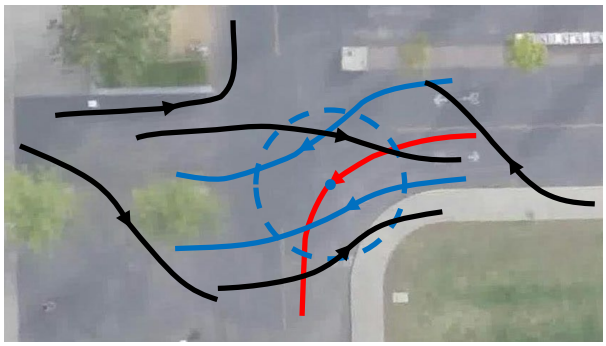


Figure 3. A schematic on how to estimate other reasonable modalities (in blue) for a target path (in red).

Classifier. Eq. 3 illustrates the functionality of our classifier. $g_{\mathbb{M}}(\cdot)$ receives the encoded feature R_H of input trajectory X and outputs possibilities for each modality. In our implementation, $g_{\mathbb{M}}(\cdot)$ is a three-layer MLP (Multilayer Perceptron) with a $tanh$ activation. We train the classifier by treating the cluster assignments as pseudo-labels.

3.4. Enhancing Diversity with Modality Loss

Samples in traditional classification tasks have only one-class assignments (ground truth). This goes against the optimization goal of a multimodal classification task that the classifier should figure out a series of futures with high probabilities to happen. In this regard, we propose a statistical method to estimate reasonable and feasible pseudo future modalities for a target path.

Specifically, Fig. 3 illustrates an example, where the red curve denotes the target path. We draw a circle o centered on the end of its observed part with a radius r and count other trajectories truncated by this circle in the entire time period of this scene. Then we group trajectories that have similar speeds and directions as the ground truth path. Qualified paths are highlighted in blue, and can be seen as other potential movements $Y_1^*, Y_2^*, \ldots, Y_N^*$.

Then, the pseudo possibility of each modality is calculated according to these potential movements by

$$p_j^* = \frac{|\{Y_i^* | Y_i^* \in M_j\}|}{N} \tag{8}$$

and we define Modality Loss as

$$\mathcal{L}_M = CrossEntropy(\mathbb{P}, \mathbb{P}^*) \tag{9}$$

where $\mathbb{P}$ denotes the classification result acquired from Eq. 3 and $\mathbb{P}^*$ denotes the pseudo label acquired from Eq. 8. In this way, we can use these statistical probabilities to supervise the classifier, which brings great diversity to our model.

3.5. Prediction with Modality Synthesis

So far, we have calculated each modality M and historical representation R_H for an input observed path. In this

section, we will discuss how $h(\cdot)$ in Eq. 4 synthesizes a prediction result for each modality.

Pseudo Future Representation for Each Modality. In order to synthesize a future trajectory in line with a given modality, we need a corresponding pseudo future representation to indicate the future trajectory propensity. We propose a regression model to fit a pseudo future representation R_F^* centered on M for given R_H.

A diagram of the regression is shown at the bottom of Fig. 2. We first subtract R_H^c from R_H and use an MLP with a *sigmoid* activation to encode the difference. After that, the encoded features are concatenated with corresponding R_F^c to form a new vector, which indicates the average of the future propensity as well as its bias to the input trajectory. Finally, the vector is fed into a fully connected layer to extract the pseudo future representation R_F^*. Each R_F^* summarizes a behavior of modality M and is able to reflect the tendency of future trajectory.

For training, we only compute R_F^* of the cluster which input R_H is assigned to. Corresponding R_F and L2 loss are used to supervise the generation of R_F^*.

Decoder. With historical representation R_H and pseudo future representation R_F^* of different modalities, we can synthesize future trajectories using an LSTM decoder. We input $h_0 = [R_H, R_{F_i}^*]$ and output predicted $\hat{Y}_i$ for modality M_i. All modalities share the same parameters and we use exponential L2 Loss in [30] for better performance.

3.6. Implementation Details

In our implementation, both BiLSTM encoders have a hidden size of 48, while the LSTM decoder has a hidden size of 96. We choose the classic K-Means [20] algorithm for clustering where the hyper-parameter K is set to 200. For weight coefficients in Eq. 7, we let $w_H = w_F = 0.5$. To find the qualified paths for Modality Loss as shown in Fig. 3, we set the radius $r = 1$, and employ a 10% limit on speed differences Δv and a 0.1π limit on direction differences $\Delta\theta$.

3.7. Discussion

Comparing with previous classification framework MultiPath [4], our insight and approach show great differences. i) We encode deep representations for behaviors and cluster modalities on them. In our vision, human behaviors are too complicated to be represented by simple coordinate series [4]. This deep clustering process can explore much better representations for modalities. Further, the idea of metric learning is implied in this process while human behavior is rather complex to be clustered by manually designed distance [4]. ii) Our classifier does not require extra scene raster as input. Actually, our framework can work well with only historical paths as input and thus our approach has a strong ability to be generalized into most prediction cases. iii) A synthesis step is proposed to provide deterministically predicted trajectories at fine-grained level for each modality, while the deterministic trajectories will be the same for inputs classified to a same class in [4] and the prediction space will be severely constrained.

We only introduce historical paths as past features in this paper for clarity since the highlight of this research is prediction framework rather than 'social or contextual information'. However, it would be easy to incorporate other information into our proposed framework. Similar to Past Encoder, one can use a social/map encoder to encode these features and concatenate them with R_H.

4. Experiment

4.1. Datasets

Performance of our method is evaluated on popular datasets, including ETH [25]/UCY [13] Dataset and Stanford Drone Dataset [27]. The ETH/UCY dataset is widely used for trajectory prediction benchmark [1, 8, 18, 29], which consists five different sub-datasets (ETH, HOTEL, UNIV, ZARA1 and ZARA2). The Stanford Drone Dataset is a large-scale dataset including various agents. These trajectories are recorded by drone cameras in bird's eye view with sufficient diversity.

In our experiments, we follow the same data preprocessing procedure and evaluation configuration as previous work [36, 21]. To evaluate the accuracy of our prediction results, we use Average Displacement Error (ADE) and Final Displacement Error (FDE) as metrics. we observe historical trajectories for 8 frames and predict future trajectories for 12 frames. 20 samples of the future trajectories are predicted.

4.2. Quantitative Evaluation

ETH/UCY. Experimental results on ETH/UCY benchmark against competing methods are shown in Tab. 1, including state-of-the-art STAR [36] and PECNet [21]. Note that the input information varies from different baselines, where P denotes historical path, S denotes social information and M denotes map information. To present the power of our framework more clearly, we only use historical paths as the information source. Results demonstrate that the performance of trajectory prediction is further elevated with our PCCSNet framework. We reach improvement of 19.2% (0.05/0.26) and 20.8% (0.11/0.53) on ADE and FDE in average respectively comparing with SOTA performance achieved by STAR. Notably, we achieve such improvement by using historical paths as the only input while STAR uses both paths and social information.

Our method fails comparing with PECNet on FDE on some subsets. We attribute this to the major differences

Method	Input	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
SGAN [8]	P+S	0.81 / 1.52	0.72 / 1.61	0.60 / 1.26	0.34 / 0.69	0.42 / 0.84	0.58 /1.18
Sophie [29]	P+S+M	0.70 / 1.43	0.76 / 1.67	0.54 / 1.24	0.30 / 0.63	0.38 / 0.78	0.54 / 1.15
Next [18]	P+M	0.73 / 1.65	0.30 / 0.59	0.60 / 1.27	0.38 / 0.81	0.31 / 0.68	0.46 / 1.00
Social STGCNN [23]	P+S	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 [21]	P+S	0.54 / 0.87	0.18 / 0.24	0.35 / 0.60	0.22 / **0.39**	0.17 / **0.30**	0.29 / 0.48
STAR [36]	P+S	0.36 / 0.65	0.17 / 0.36	0.31 / 0.62	0.26 / 0.55	0.22 / 0.46	0.26 / 0.53
PCCSNet	P	**0.28 / 0.54**	**0.11 / 0.19**	**0.29 / 0.60**	**0.21** / 0.44	**0.15** / 0.34	**0.21 / 0.42**

Table 1. Comparison on ETH and UCY dataset for $T_{obs} = 8$ and $T_{pred} = 12$ (ADE/FDE), including SOTA STAR and PECNet. P denotes historical path, S denotes social information and M denotes map information. [18] also uses pose information from AlphaPose [6, 15]. It is worth mentioning that our approach outperforms other methods in average without using social and map information.

Method	SGAN [8]	Sophie [29]	PECNet [21]	PCCSNet
Input	P+S	P+S	P+S	P
ADE	27.23	16.27	9.96	**8.62**
FDE	41.44	29.38	**15.88**	16.16

Table 2. Comparison with baseline methods on SDD for $T_{obs} = 8$ and $T_{pred} = 12$, including SOTA PECNet.

Method	minADE_1	minFDE_1	minADE_5
MultiPath [4]	28.32	58.38	17.51
PCCSNet	**18.14**	**36.32**	**12.54**

Table 3. Comparison with MultiPath on SDD for $T_{obs} = 5$ and $T_{pred} = 12$. minADE_k and minFDE_k measures the displacement error against the closest trajectory in top k samples.

Method	KM w/o deep	KM	HAC	GMM
ADE	0.24	**0.21**	0.21	0.21
FDE	0.45	**0.42**	0.43	0.42
Time/min	0.6	0.7	18	90

Table 4. Comparison between different clustering methods on ETH/UCY Dataset. Results are the average of five sub-datasets.

K	100	200	500	1000
ADE	0.22	**0.21**	0.21	0.22
FDE	0.44	**0.42**	0.43	0.45

Table 5. Comparison between different parameter K in K-means on ETH/UCY Dataset. Results are the average of five sub-datasets.

that our method is ADE-prioritized while PECNet is FDE-prioritized, which means that PECNet has a tendency to achieve a lower FDE than a lower ADE. Further, the social information is absent in our method.

SDD. We also report the prediction performance on SDD dataset in Tab. 2. Comparing with the SOTA framework PECNet, we achieve remarkable improvement of 13.5% (1.34/9.96) increase on ADE. There is a little FDE decline of 1.8% (0.28/15.88). Considering the differences we have discussed above and the trade-off between huge ADE improvement and a minor FDE decline, we believe that our prediction results are promising.

4.3. Analysis

Comparing with MultiPath [4]. We compare with MultiPath in Tab. 3. Note that the frames of observation 5 in MultiPath is different from commonly used 8 [29, 21]. Our method greatly outperforms MultiPath.

Clustering on Deep Features. Our clustering process is applied on deep features to explore better representations for different modalities. Comparison between k-means on trajectory coordinates (KM w/o deep) and deep features (KM) in Tab. 4 proves that our deep clustering process can capture modality representations effectively.

Clustering Algorithm. We compare the performance of three common clustering algorithms when they are used for constructing $\mathbb{M}$ in Eq. 2, including K-means (KM), hierarchical agglomerative clustering (HAC) and Gaussian mixture model (GMM). Results in Tab. 4 show that the simple K-means algorithm exhibits a huge advantage in speed with no decline in accuracy. Therefore, we use K-means as the clustering algorithm in our implementation.

K in K-means. We also study the effect of different K in K-means algorithm on the results, shown in Tab. 5. A larger K can let each modality M reveal more fine-grained representations, yet it may cause a reduction in classification accuracy. We assign 200 for K in our experiments for better overall performance.

Weights for Clustering Distance. For clustering distance as described in Eq. 7, both historical and future representations are taken into consideration. We add weights w_H and w_F to balance them for clustering. According to experiments in Tab. 6, a proper weighting ratio near $1:1$ will get much better performance. If the weight of future representation grows, the performance drops a little. And if the weight of historical representation grows, the performance drops a lot. This reveals two facts: 1) Future features take a leading role in modality clustering. 2) Historical features play a supporting role to differentiate confusing situations for better clustering results.

Analysis for Modality Loss. Tab. 7 shows a comprehensive analysis of Modality Loss. After applying it to enhance the diversity, our performance is further improved. Some hyper-parameters will decide the scale of restrictions of similar movements, including the radius of circle r, and

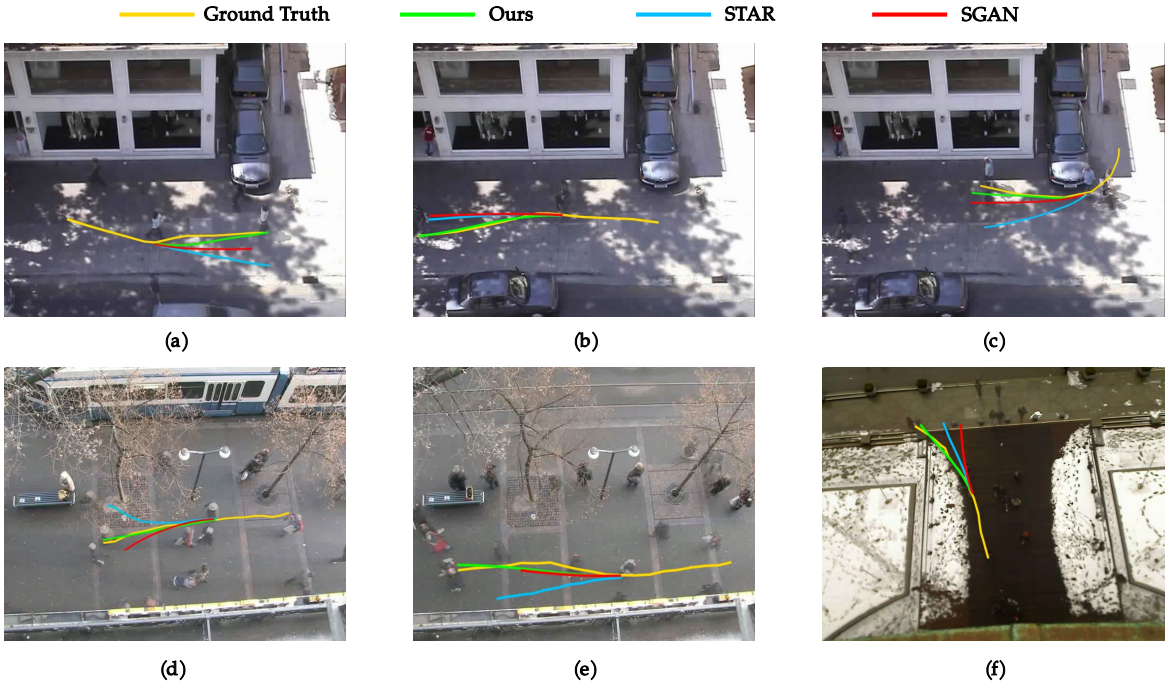


Figure 4. Accuracy analysis of PCCSNet where yellow lines indicate ground truth paths and our predicted results are in green. Predictions of two comparison models, STAR and SGAN, are denoted as blue and red. Examples are from different scenes including ZARA (a-c), HOTEL (d,e) and ETH (f).

$(W_H : W_F)$	1:3	1:2	1:1	2:1	3:1
ADE	0.22	0.21	**0.21**	0.23	0.24
FDE	0.43	0.42	**0.42**	0.47	0.49

Table 6. Comparison between different weight configurations for clustering distance in Eq. 7 on ETH/UCY Dataset. The results are the average of five sub-datasets.

r/m	Δv	$\Delta\theta/\pi$	ADE	FDE
0	0	0	0.221	0.443
1	**10%**	**0.1**	**0.208**	**0.422**
0.5	10%	0.1	0.212	0.426
2	10%	0.1	0.215	0.432
1	5%	0.1	0.213	0.426
1	20%	0.1	0.216	0.438
1	10%	0.05	0.212	0.427
1	10%	0.2	0.213	0.427

Table 7. Sensitive analysis on differet hyper-parameter configurations for Modality Loss on ETH/UCY Dataset. The results are the average of five sub-datasets.

Method	w/o synthesis	PCCSNet	Δ
ADE	0.23	**0.21**	8.7%
FDE	0.45	**0.42**	6.7%

Table 8. Contribution of modality synthesis on ETH/UCY. Results are the average of five sub-datasets.

thresholds for speed Δv and angle $\Delta\theta$. When these parameters vary, the performance fluctuates, where speed constraint is more sensitive than others.

Contribution of Modality Synthesis. Modality synthesis is proposed to optimize predictions at fine-grained level. Tab. 8 shows the synthesis process brings huge improvement in accuracy.

4.4. Qualitative Evaluation

Accuracy Analysis. We compare PCCSNet with other approaches on various challenging cases including turning and acceleration. Predicted results of the best modality in Fig. 4 illustrate that our model can capture more accurate future modalities in terms of both speed and direction. (c) and (f) are two typical cases of turning. Our model predicts more accurate angles with a proper speed. STAR gives a much smaller angle in both cases while SGAN predicts a wrong turning direction in (f). Another challenging case (e) shows that even though PCCSNet ignores a jitter in the future, the predictions of speed, direction and destination are remarkable. Other approaches fail to estimate neither the speed nor the destination.

Multimodality with Probability. Fig. 5 demonstrates the probabilistic property of our proposed framework. In these three specific scenes, not only can our model give accurate predictions, probabilities of reasonable futures also tend to be much higher than average. Therefore, our model is less likely to predict trajectories that are improbable while maintaining the accuracy of the best-match. However, each prediction sampled by a generative model can only be interpreted as a average probability of $\frac{1}{k}$ when taking k predic-

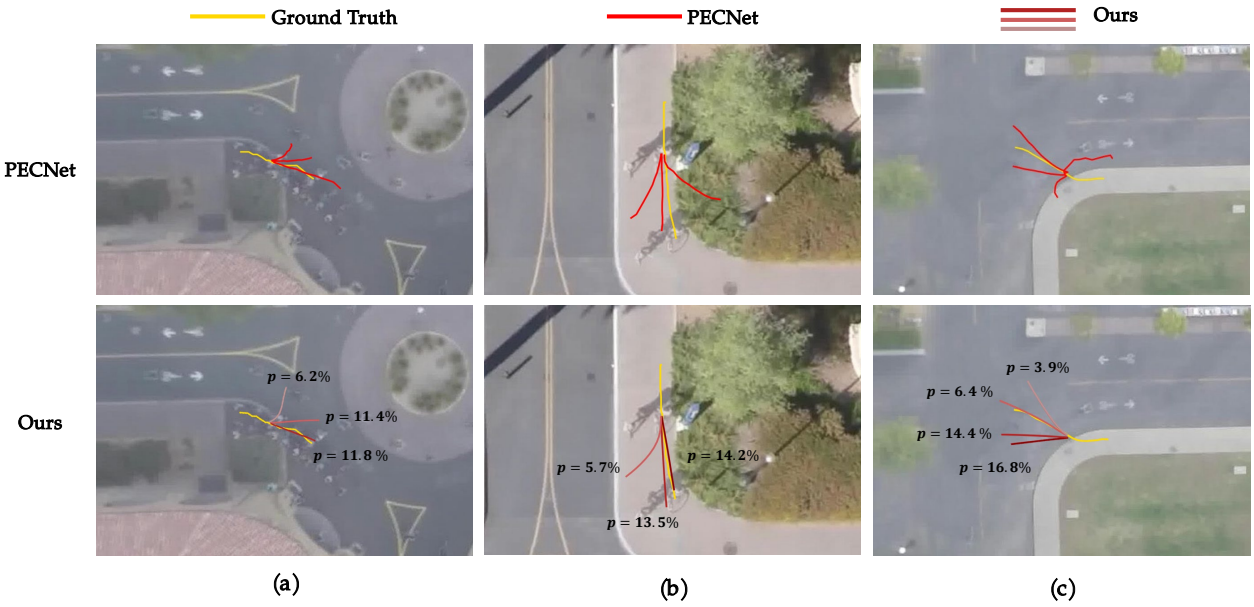


Figure 5. Illustration of probabilistic multimodal predictions where yellow curves indicate ground truth paths and our predicted results are in red (darker refers to higher probability). PECNet is used for comparison and its results are denoted as the same dark red since each result it outputs has equal probability. The probabilities of our method are marked out while they take the same value of 5% for PECNet (1/20). We only visualize some representative trajectories for a clear view.

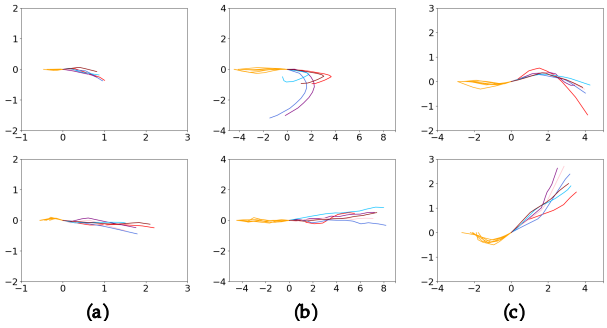


Figure 6. Visualization of clusters. Past trajectories are in yellow and future ones are in various colors for a clear view. Each column refers to a pair with similar past trajectories and different future modalities. Values of x,y-axis are coordinates in meters.

tions. Thus, they lose probabilistic information for each modality, which is important to the following decision-making process.

Clustering Analysis. To illustrate the effectiveness of the clustering algorithm in exploring different potential modalities, we visualize samples in some different clusters in Fig. 6. Pair (a) indicates two types of possible futures that both come after a slow, linear past trajectory. The top one remains slow whereas the bottom one begins to accelerate. In pair (b), the past trajectories are still linear but faster. We show an 180° turn and a straight path that may happen in the future. Pair (c) shows curved past trajectories unlike former ones. Corresponding future modalities are right and left turning. These cases give a strong proof that our clustering algorithm is sensitive enough to capture differences between potential modalities. Note that although the clustering is performed in a high dimensional space following Eq. 2, the visualization is in 2D space for readability.

5. Conclusion

In this paper, we formulate the multimodal prediction framework into three steps of modality clustering, classification and synthesis to address major weaknesses in previous works and present a brand-new pipeline PCCSNet to solve it. Considering that the future are usually centralized around several different behaviors, we first cluster encoded historical and future representations to identify potential behavior modalities. A classifier then is trained to figure out the probability of occurrence for each modality given a historical path with a novel Modality Loss. Further, a modality synthesis mechanism is proposed to get fine-grained prediction results deterministically. Exhaustive experiments demonstrate the superiority of our elaborately designed framework in accuracy, diversity and reasonableness, even without introducing social and map information.

Acknowledgment This work is supported in part by the National Key R&D Program of China, No. 2017YFA0700800, National Natural Science Foundation of China under Grants 61772332 and Shanghai Qi Zhi Institute, SHEITC (2018-RGZN-02046)

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[2] Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *International conference on machine learning*, pages 552–560, 2013.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

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

[5] Chiho Choi and Behzad Dariush. Looking to relations for future trajectory forecast. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017.

[7] 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.

[8] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

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

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

[11] James Joyce. Bayes' theorem. 2003.

[12] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012.

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

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

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

[16] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020.

[17] Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10508–10518, 2020.

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

[19] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6120–6127, 2019.

[20] 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.

[21] 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.

[22] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942. IEEE, 2009.

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

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

[25] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009.

[26] 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.

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

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

[29] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019.

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

[31] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018.

[32] Hanqing Wang, Wenguan Wang, Wei Liang, Caiming Xiong, and Jianbing Shen. Structured scene memory for vision-language navigation. *arXiv preprint arXiv:2103.03454*, 2021.

[33] Hanqing Wang, Wenguan Wang, Tianmin Shu, Wei Liang, and Jianbing Shen. Active visual information gathering for vision-language navigation. In *European Conference on Computer Vision*, pages 307–322. Springer, 2020.

[34] Dan Xie, Sinisa Todorovic, and Song-Chun Zhu. Inferring "dark matter" and "dark energy" from videos. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[35] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *CVPR 2011*, pages 1345–1352. IEEE, 2011.

[36] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. *arXiv preprint arXiv:2005.08514*, 2020.

[37] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2871–2878. IEEE, 2012.