Interpretable Self-Aware Neural Networks for Robust Trajectory Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract: Although neural networks have seen tremendous success as predictive models in a variety of domains, they can be overly confident in their predictions on out-of-distribution data. To be viable for safety-critical applications, like autonomous vehicles, neural networks must accurately estimate their epistemic or model uncertainty, achieving a level of system *self-awareness*. Popular techniques in epistemic uncertainty quantification often require out-of-distribution data during training or multiple forward passes through the neural network during inference. These approaches may be not be suitable to achieve real-time performance on high-dimensional inputs. Furthermore, existing methods lack interpretability of the estimated uncertainty, which limits their usefulness both to engineers for further system development and to downstream modules in the autonomy stack. We propose the use of evidential deep learning to estimate the epistemic uncertainty over a low-dimensional, interpretable latent space in a trajectory prediction setting. We introduce an interpretable paradigm for trajectory prediction that distributes the uncertainty amongst the semantic concepts: past agent behavior, social context, and road structure. We validate our approach on real-world autonomous driving data, demonstrating superior performance over state-of-the-art baselines.

Keywords: Autonomous vehicles, trajectory prediction, distribution shift

1 Introduction

Deep learning techniques have had success across a multitude of domains, including human trajectory prediction in the context of autonomous vehicles (AVs) [1, 2]. However, a key challenge in deploying these systems is their lack of *self-awareness* about the quality of the predictions they make. Deep learning models often overestimate their confidence in unfamiliar situations [3, 4, 5, 6, 7, 8, 9]. For robots deployed in human environments, this over-confidence could result in dangerous maneuvers and safety concerns. Flagging unfamiliar situations encountered in the real world could be helpful to downstream autonomy stack components, such as path planners to execute a fail-safe maneuver, and to engineers for further system development.

Real-world systems are subject to *aleatoric* and *epistemic* uncertainty. The former is irreducible data uncertainty, which, for trajectory prediction, could be represented as a distribution over future trajectories given past observations (e.g., turning or continuing straight given a straight past trajectory). Aleatoric uncertainty can be modelled explicitly during training of learning-based systems, for example, using variational approaches [10, 11, 12, 13, 14, 15, 16]. Epistemic uncertainty reflects *what the model does not know*, which can arise due to the limited ability of a model to represent data and data distribution shift. Neural networks often fail to correctly calibrate this uncertainty, resulting in unreliable predictions for out-of-distribution (OOD) inputs [3, 4, 5, 6, 7, 8, 9]. In this paper, we focus on epistemic uncertainty quantification for trajectory prediction.

Submitted to the 6th Conference on Robot Learning (CoRL 2022). Do not distribute.

There has been growing interest in estimating epistemic uncertainty for deep learning models [17]. However, the majority of the proposed methods consider small benchmark datasets or require OOD data for training [18, 19, 20, 21, 22, 23]. A few recent papers (e.g., [24, 25, 26, 27, 28, 29, 30, 31]) have began exploring epistemic uncertainty estimation for learning-based robot perception and prediction. Identifying OOD inputs for tasks like trajectory prediction is particularly difficult due to the high dimensionality of the data. Since it is impossible to foresee all possible scenarios encountered on the road, an OOD dataset cannot be manually curated for training. Furthermore, successful epistemic uncertainty estimation techniques, such as ensembles [32] and Monte Carlo (MC) dropout [18], require multiple forward passes during inference, potentially hindering real-time performance. Finally, most methods output a single value for epistemic uncertainty. In a cluttered, dynamic setting (e.g., urban driving), where there are multiple possible sources of uncertainty, this value could be challenging to interpret. For example, it may be unclear whether high epistemic uncertainty stems from an unfamiliar observed trajectory maneuver, new intersection type, or strange behavior of surrounding agents due to an external disturbance (e.g., road construction).

Modeling epistemic uncertainty for trajectory prediction has three key challenges: (1) lack of explicit OOD data during training, (2) efficiency requirements during inference, and (3) interpretability of the computed uncertainty. This paper addresses these challenges for trajectory prediction through an evidential deep learning approach that estimates the epistemic uncertainty in one-shot with only a single inference forward pass through the model and requires no OOD data for training. We allocate this uncertainty among interpretable, low-dimensional latent variables. *Evidential deep learning* estimates parameters for a second-order distribution, (e.g., a Dirichlet distribution, which is a distribution over categorical distributions) to capture the epistemic uncertainty [19, 20, 33]. To avoid OOD data for training, normalizing flows may be used to constrain the learned Dirichlet distribution parameters by enforcing the densities of the individual classes to integrate to the number of training samples in that class, as in the Posterior Network (PostNet) architecture [34].

In trajectory prediction, human behavior is commonly modeled using discrete modes that can represent high-level maneuvers like accelerating, braking, driving straight, and turning [35, 36, 37, 12, 38, 39]. We apply ideas from PostNet [34] for modeling epistemic uncertainty to trajectory prediction architectures with discrete modes. To address interpretability in the learned epistemic uncertainty, we propose the following semantic insight. High epistemic uncertainty for trajectory prediction may originate from unfamiliar road structures, input behaviors, or social contexts. We distribute the learned epistemic uncertainty over these interpretable categories encoded in a low-dimensional latent space, thus encouraging tractability. Hence, we introduce a trajectory prediction architecture paradigm that is self-aware of its prediction confidence and is able to provide rich, interpretable information to downstream components of the autonomy stack and to engineers for development. We call this paradigm *Interpretable Self-Aware Prediction (ISAP)*. Our key contributions are as follows:

- We propose a novel application of evidential deep learning methods to the task of epistemic uncertainty quantification for trajectory prediction.
- We infuse interpretability into the uncertainty estimate by distributing it over interpretable, low-dimensional latent spaces. The epistemic uncertainty source is decomposed as coming from unfamiliar agent behavior, social context, or road configuration.
- We demonstrate superior uncertainty estimation performance of the proposed ISAP framework over state-of-the-art approaches on the real-world NuScenes dataset [40].

2 Related Work

Interpretable Trajectory Prediction. A popular means to encourage interpretability in trajectory prediction architectures is through discrete modes [37, 39, 12, 36, 38, 41]. For example, Chai et al. [37], Mangalam et al. [42], and Hu et al. [43] induce interpretability by learning distributions over discrete intention goals. Kothari et al. [41] learn a probability distribution over possibilities in an interpretable discrete choice model. Due to the ubiquity of discrete modes in trajectory prediction literature, we develop an epistemic uncertainty quantification approach assuming the presence of

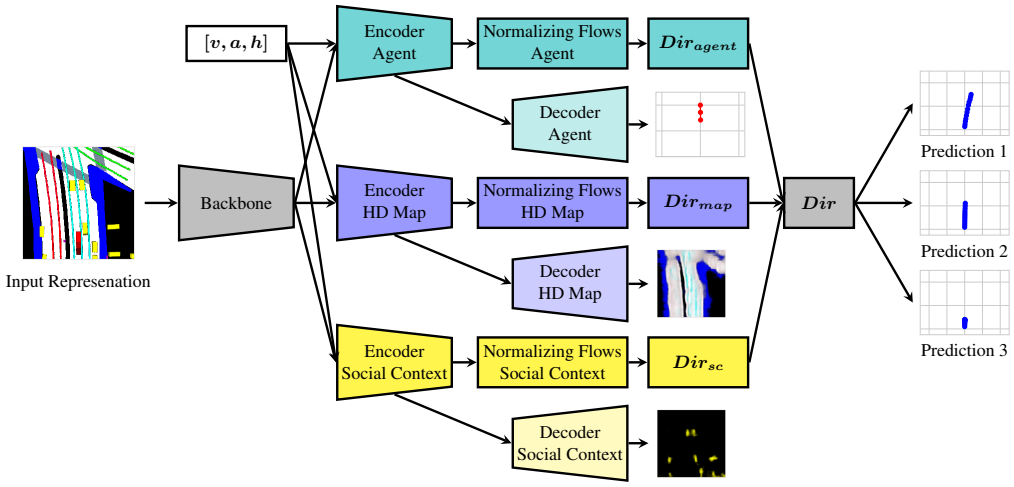

Figure 1: Interpretable self-aware prediction (ISAP) system overview. Our approach learns the uncertainty distributed over semantically interpretable latent concepts: the agent's past behavior, map, and social context. The network outputs parameters for three corresponding Dirichlet distributions, learned using normalizing flows, which can then be combined through a uniform prior to form the output distribution. The aleatoric uncertainty for the trajectory prediction task is modeled by the expected value of the Dirichlet parameters, while their sum is an indicator of epistemic uncertainty. A high Dirichlet parameter sum indicates evidence for an input, and corresponds to low epistemic uncertainty.

discrete modes in the architecture. Another common method for facilitating interpretability in a latent space is through an encoder-decoder structure. Neumeier et al. [44] use a decoder with expert knowledge to produce an interpretable latent space in a trajectory prediction model. Inspired by this idea, we enforce the epistemic uncertainty to be learned over three separate latent encodings that correspond to observed agent behavior, social context, and road configuration. This interpretable structure is achieved via decoder components that are learned through self-supervised signals.

Epistemic Uncertainty for Learning-Based Perception and Prediction. Gawlikowski et al. [17] survey modern uncertainty quantification methods. A common approach to compute epistemic uncertainty in learning-based autonomous systems is to use MC dropout [18] due to its simple implementation and because it does not require OOD data for training. This approach is ubiquitously used in tasks like trajectory prediction [45, 46], pedestrian bounding box prediction [47], semantic segmentation and depth regression [48], and inverse sensor model learning [49]. However, MC dropout requires multiple forward passes through the model during inference and has been found to be less robust to distribution shift than ensembles [50]. Furthermore, MC dropout requires a dropout architecture element in the model, which may not always be desirable. Ensembles [32] are regarded as a robust epistemic uncertainty estimation technique, but can be expensive to train and perform inference, hindering their widespread use in robotics applications. One-shot evidential uncertainty estimation methods, like evidential regression, have recently shown promising results in autonomous perception tasks, like depth estimation [19, 25]. In this work, we investigate the use of the PostNet evidential deep learning technique [34] within a trajectory prediction task. PostNet does not require OOD data during training and only needs one forward pass during inference time to obtain both the model output and the epistemic uncertainty estimate. We are interested in how this technique scales to high-dimensional data necessary for trajectory prediction.

3 Methods

Problem Definition. We consider the following trajectory prediction setting. We assume the input representation x consists of a combination of the road structure (e.g., a high-definition (HD) map), the past trajectory and current state for the agent of interest, and the past trajectory information for the surrounding agents. The agent's state consists of speed v, acceleration a, and heading change rate h. The goal of the trajectory prediction task is to predict a 2D position vector $y \in \mathbb{R}^{2\times T}$ for a

time horizon of T time steps into the future. We make the assumption that the trajectory prediction architecture has a set of discrete anchors with ground truth labels $d \in \{1, \ldots, C\}$. For example, the MultiPath [37] architecture uses k-means to cluster trajectories into a discrete latent space and the CoverNet [38] model constructs a set of possible trajectories with a specified level of coverage.

Posterior Network (PostNet). To estimate epistemic uncertainty in the trajectory prediction setting, we use ideas from PostNet [34]. PostNet is an evidential deep learning approach [19, 20] that employs normalizing flows [51] to predict a closed-form posterior distribution over predicted probabilities. We analyze its performance as applied to the discrete anchor space of the complex trajectory prediction task. PostNet's posterior distribution encompasses both aleatoric and epistemic uncertainties without requiring OOD data during training. The uncertainty is parameterized by a Dirichlet distribution, which is the conjugate prior of the categorical distribution. This approach is one-shot as it takes one network pass to compute the epistemic distribution $q^{(i)}$ and the aleatoric distribution $\bar{p}^{(i)}$, for an input $x^{(i)}$,

$$q^{(i)} = \text{Dir}(\alpha^{(i)}) \quad \text{and} \quad \bar{p}^{(i)} = \text{Cat}(\bar{\xi}^{(i)}) \quad \text{with} \quad \bar{\xi}_c^{(i)} = \mathbb{E}_{q^{(i)}}[\xi^{(i)}]_c = \frac{\alpha_c^{(i)}}{\alpha_0^{(i)}}, \tag{1}$$

where $\alpha^{(i)} \in \mathbb{R}_+^C$ are the Dirichlet parameters and $\alpha_0^{(i)} = \sum_{c=1}^{C} \alpha_c^{(i)}$ reflects the total amount of evidence or certainty allocated to the input (higher α_0 indicates lower epistemic uncertainty). The parameters $\xi^{(i)} \in \left\{[0,1]^C \mid \sum_c \xi_c^{(i)} = 1\right\}$ of a categorical distribution $p^{(i)} = \text{Cat}(\xi^{(i)})$ can be sampled from the epistemic distribution: $\xi^{(i)} \sim q^{(i)}$. An anchor prediction $\hat{d}^{(i)}$ is made according to: $\hat{d}^{(i)} = \arg\max_c \bar{\xi}_c^{(i)}$. The Dirichlet parameters $\alpha^{(i)}$ are constructed via $\alpha^{(i)} = \beta^{\text{prior}} + \beta^{(i)}$, where $\beta^{\text{prior}} \in \mathbb{R}_+^C$ is a fixed prior and $\beta^{(i)} \in \mathbb{R}_+^C$ represents learned pseudo-counts as evidence for an input $x^{(i)}$. Following Charpentier et al. [34], we use an uninformative prior with $\beta^{\text{prior}} = 1$.

The pseudo-counts $\beta^{(i)} \in \mathbb{R}_+^C$ are defined as follows,

$$\beta_c^{(i)} = N_c \cdot r(z^{(i)} \mid c; \phi), \tag{2}$$

where z is a low-dimensional latent space, c is the discrete anchor class, ϕ are neural network parameters, and N_c is related to the ground truth count for an anchor class c, serving as a certainty budget for each class. The normalized probability density $r(z^{(i)} \mid c; \phi)$ in Eq. (2) is learned over the low-dimensional latent space z to encourage tractability and scaling of the algorithm to high-dimensional inputs. First, a neural network encodes the input $x^{(i)}$ into the latent space, $z^{(i)} = f_\theta(x^{(i)})$. Then, due to its representational capacity, a normalizing flow is used to learn the distribution over this latent space [51, 52]. It is important that $r(z^{(i)} \mid c; \phi)$ be a normalized density to encourage the α_0 parameters to be high for high-density, in-distribution (ID) regions (low epistemic uncertainty) and low for low-density, OOD regions (high epistemic uncertainty). As $r(z^{(i)} \mid c; \phi)$ goes to zero, the Dirichlet parameters $\alpha^{(i)}$ reduce to the uninformative prior $\beta^{\text{prior}} = 1$. To optimize PostNet, we use the evidence lower bound (ELBO) loss [53],

$$\mathcal{L}_{\text{ELBO}} = \frac{1}{N} \sum_{i=1}^{N} -\mathbb{E}_{q^{(i)}} \left[\log p^{(i)}(d^{(i)})\right] + \text{KL}(q^{(i)} \mid\mid \text{Dir}(1)). \tag{3}$$

Interpretable Self-Aware Prediction (ISAP). The PostNet approach to uncertainty quantification naturally transfers to trajectory prediction models with supervised, discrete latent anchors. However, to make the uncertainty estimates more informative in complex settings (e.g., urban scenes), we infuse interpretability into the latent space z, forming the proposed ISAP approach. We encode the input x into three separate latent variables: z_{agent}, z_{map}, and z_{sc}, representing the agent's past behavior, the road structure map, and the social context surrounding the agent of interest, respectively. Such a decomposition has been shown to be effective at the input level for trajectory prediction, supporting our choice of interpretable structure [54]. These semantic concepts are encoded into the latent space using accompanying decoder components. The decoders are self-supervised with the input data x split into the agent's past trajectory, the road structure representation, and the past trajectories of other agents. The decoder weights are learned through associated reconstruction losses.

The ISAP network then outputs parameters for three Dirichlet distributions corresponding to the three semantic concepts. These Dirichlet distributions are combined through an equally weighted average of their parameters, signifying a uniform prior over the three categories,

$$\alpha = \frac{1}{3}\alpha_{\text{agent}} + \frac{1}{3}\alpha_{\text{map}} + \frac{1}{3}\alpha_{\text{sc}}. \tag{4}$$

The full loss for the ISAP architecture is then,

$$\mathcal{L} = \mathcal{L}_{\text{ELBO}} + \lambda_{\text{agent}}\mathcal{L}_{\text{rec,agent}} + \lambda_{\text{map}}\mathcal{L}_{\text{rec,map}} + \lambda_{\text{sc}}\mathcal{L}_{\text{rec,sc}}, \tag{5}$$

where λ_{agent}, λ_{agent}, and λ_{agent} are scaling coefficients for each reconstruction loss term. Additional details on the reconstruction loss terms can be found in Appendix A. Postels et al. [55, 27] demonstrate that regularization of the latent space in terms of reconstruction capability improves epistemic uncertainty estimation. These findings further support our choice of interpretable architecture for the epistemic uncertainty quantification task. The full ISAP architecture is illustrated in Fig. 1.

4 Experiments

We experimentally validated the epistemic uncertainty estimation and OOD detection capabilities of the proposed ISAP paradigm. All models were trained on a single NVIDIA GeForce RTX 2080 Ti GPU. Appendix A provides further details.

Data. To test our approach in a real-world setting, we use data from the prediction task in the NuScenes [40] autonomous driving dataset. The predictions are made for $6\,\text{s}$ in the future based on $1\,\text{s}$ of past information collected at $2\,\text{Hz}$. We use the same data and input representation as Phan-Minh et al. [38], but we filter out any data with less than $1\,\text{s}$ of past trajectory information to enable decoding of the agent's past trajectory. The input representation $x \in [0, 1]^{500\times500\times3}$ combines the past agent trajectory, the HD map, and the past trajectories of other agents into a bird's-eye view rendering of the scene (see Fig. 1). The agent's state $[v, a, h]$ serves as input to each network branch. We split the data into ID and OOD examples according to the agent's past trajectory. ID input trajectories are chosen to be slower than OOD ones. We use the l_2 distance between the most recent and the earliest waypoints as a heuristic of the relative trajectory 'speed'. We threshold the ID data to have an l_2 distance of less than $10\,\text{m}$, leaving faster trajectories for OOD data. This split allowed for sufficient ID training (25,669), validation (7,344), and test (7,270) examples, while still having a reasonable number of OOD examples (validation: 2,521, test: 3,267). We further validate our approach on another OOD data split based on the input road structure in Appendix B.

Architecture and Training Details. For our trajectory prediction architecture, we chose the CoverNet [38] model, which is the baseline technique for the NuScenes prediction task. This model is convenient because it frames trajectory prediction as classification over a predefined set of trajectories. We can directly integrate our ISAP approach with this architecture. For our experiments, we used a trajectory anchor set of size 64 for classification [38]. The latent variables z_{agent}, z_{map}, and z_{sc} are set to dimension 4, as this low dimensionality results in good uncertainty estimation performance and computational efficiency for the normalizing flows. The probability density $r(z^{(i)} \mid c; \phi)$ for each of the three latent variables is modeled with radial normalizing flows consisting of eight layers as done by Charpentier et al. [34]. The map and social context decoders take as input features one layer upstream of z_{map} and z_{sc} of dimension $4,096$ to enable higher representational capacity for the decoded outputs. Further details on the architecture and the decoder outputs are in Appendix A.

We modified the ELBO loss to use CoverNet's constant lattice loss in the reconstruction term. The ground truth classification labels are the trajectory anchor classes with the smallest l_2 distance to the ground truth trajectories. We found the coefficients: $\lambda_{\text{agent}} = 1$, $\lambda_{\text{map}} = 1$, and $\lambda_{\text{sc}} = 10$ to work well in training. The coefficient for the social context decoding had to be increased as this representation is spatially sparse, and otherwise the decoding collapsed to a stable local minimimum of predicting no other agents in the scene. We trained the model for 25 epochs using the Adam [56] optimizer with a 0.001 learning rate, a batch size of 16, and a weight decay of 5×10^{-4}.

Baselines. We consider three baselines to our approach: CoverNet [38], Post-CoverNet, and ensembles [32]. We benchmark against the CoverNet model to form a comparison for the trajectory prediction and calibration performance that it can achieve without our modifications. Post-CoverNet is an ablation on our ISAP approach without the interpretability element. Finally, ensembles are the state-of-the-art for estimating epistemic uncertainty for neural network models. Gustafsson et al. [26] show that ensembles [32] consistently outperform MC dropout. Thus, we baseline against the more performant of the two approaches with $N = 5$ and $N = 10$ models in the ensemble. All baselines were trained according to the same training set-up as ISAP.

Metrics. We employ a wide variety of metrics to investigate the performance of the proposed approach in terms of (1) estimating epistemic uncertainty and (2) maintaining trajectory prediction performance. To measure trajectory prediction performance, we use standard trajectory prediction metrics employed by Phan-Minh et al. [38]: minimum average displacement error over the most likely k modes (minADE$_k$) and final displacement error (FDE). Lower is better for these metrics.

We then evaluate the uncertainty estimation performance. Following Charpentier et al. [34], we use the area under the receiver operating characteristic (AUROC) to compute the confidence in the correct predictions (higher is better). The ground truth labels are 0 for incorrect predictions and 1 for correct predictions. The scores to compute the aleatoric confidence are $\max_c \bar{p}_c^{(i)}$. For epistemic confidence, the scores are $\max_c \alpha_c$ for Post-CoverNet and ISAP and $1/Var_c$ for ensembles where Var_c is the empirical variance of the predicted class probability across the ensemble. The Brier score measures how well the output distribution is calibrated. It is computed as $\frac{1}{N}\sum_{i=1}^{N} \|\bar{p}^{(i)} - d^{(i)}\|$ (lower is better) where $d^{(i)}$ are the classification labels represented as one-hot encodings.

To determine the performance of the considered approaches at distinguishing OOD data, we use AUROC and average precision (APR) with labels 0 for OOD and 1 for ID data (higher is better). For OOD detection based on aleatoric uncertainty, the scores are computed according to $\max_c \bar{p}_c^{(i)}$. When based on epistemic uncertainty, the scores are $\alpha_0^{(i)}$ for Post-CoverNet and ISAP and $1/Var_c$ for ensembles. To provide further intuition for Post-CoverNet and ISAP, we also report the ratio of the average sums of the Dirichlet parameters for ID and OOD data $\bar{\alpha}_{0,OOD}/\bar{\alpha}_{0,ID}$ (lower is better).

5 Results

Quantitative Results. The trajectory prediction performance is reported in Table 1. As could be expected, the best performing approaches were the ensemble baselines. Ensembles tend to be more robust than their single network counterparts, in this case represented by CoverNet [38]. Interestingly, the smaller ensemble ($N = 5$) slightly outperformed the bigger ensemble ($N = 10$), indicating that the higher variability amongst models may cause a small drop in performance. As PostCoverNet and ISAP add competing terms to the trajectory prediction objective, it is not surprising that the trajectory prediction performance is mildly compromised as a result. However, it is encouraging that the performance drop is only slight.

The focus of our work is to accurately estimate the epistemic uncertainty of the trajectory prediction model. The uncertainty quantification results are presented in Table 2. The ISAP model outperformed the baseline approaches across almost all metrics. We make two interesting observations in Table 2. The first is that ISAP outperforms Post-CoverNet in epistemic uncertainty estimation. It appears that the interpretability encoded into the latent space within ISAP helps its performance, particularly in OOD data detection. We hypothesize that distributing the uncertainty over simpler, interpretable latent variables makes the uncertainty estimation task easier. The second observation is that the ISAP framework outperforms the ensemble baseline in OOD detection. Ensembles are often the canonical method for OOD detection; however, ISAP and Post-CoverNet outperform the smaller ensemble baseline by a large margin. The larger ensemble approaches Post-CoverNet performance, but ISAP still outperforms. This result also supports the findings of Charpentier et al. [34] where a similar trend was observed for a PostNet architecture on smaller classification problems.

Table 1: Trajectory prediction metrics (lower is better). The best performance is highlighted in bold. Our methods perform comparably to ensembles and the original CoverNet model.

	CoverNet [38]	Ensemble [32] ($N = 5$)	Ensemble [32] ($N = 10$)	Post-CoverNet (Ours)	ISAP (Ours)
minADE$_1$	4.327	**4.241**	4.246	4.529	4.711
minADE$_5$	1.885	1.867	**1.859**	1.951	2.004
minADE$_{10}$	1.545	**1.529**	1.539	1.581	1.599
minADE$_{15}$	1.413	**1.421**	1.423	1.440	1.474
FDE	9.474	**9.270**	9.293	10.009	10.177

Table 2: Uncertainty estimation metrics. The best performance is highlighted in bold. Our methods outperform across most metrics.

	CoverNet [38]	Ensemble [32] ($N = 5$)	Ensemble [32] ($N = 10$)	Post-CoverNet (Ours)	ISAP (Ours)
Alea. Confidence ↑	0.638	0.638	0.636	0.630	**0.652**
Epi. Confidence ↑	–	0.455	0.434	0.573	**0.621**
Brier Score ↓	0.837	**0.835**	0.840	0.857	0.850
Alea. OOD (APR) ↑	0.542	0.530	0.538	0.833	**0.930**
Epi. OOD (APR) ↑	–	0.810	0.961	0.960	**0.976**
Alea. OOD (AUROC) ↑	0.241	0.218	0.240	0.652	**0.871**
Epi. OOD (AUROC) ↑	–	0.693	0.913	0.919	**0.955**
$\bar{\alpha}_{0,OOD}/\bar{\alpha}_{0,ID}$ ↓	–	–	–	0.171	**0.145**

We investigate how well the learned pseudo-counts α_0 reflect the true data distribution on the test set in Fig. 2. We plot the true data distribution as a histogram over the ℓ_2 distance between the earliest and most recent waypoints in the agent's past trajectory. We distinguish between ID (green) and OOD (orange) examples. The data peaks at $0\,\mathrm{m}$ (stopped) and at around $5\,\mathrm{m}$ ($18\,\mathrm{km/h}$), and then slopes off for OOD data. The learned $\alpha_{0,\text{agent}}$ parameters reflect these trends. As the ID versus OOD distinction is less obvious for the map and social context latent variables, the $\alpha_{0,\text{map}}$ and $\alpha_{0,\text{sc}}$ trends are more flat across both data types, although $\alpha_{0,\text{map}}$ still shows a pretty clear distinction. We hypothesize that the road configuration (e.g., multi-lane highway versus roundabout) is correlated with agent speed, while the surrounding social context may not always influence the agent's speed.

Qualitative Results. Fig. 3 presents the qualitative results produced by ISAP for ID and OOD examples. The figure shows the input to the network, the decoded latent variables, and their associated α_0 values. The ID example has a slower (shorter) trajectory input than the OOD example. The ISAP network produces a clear distinction between the ID and OOD examples in the Dirichlet parameter reflecting the agent's past trajectory, as expected. The ID $\alpha_{0,\text{agent}}$ value is much higher (lower uncertainty) than that for OOD. The OOD $\alpha_{0,\text{agent}}$ value reaches almost total uncertainty as the uniform prior over 64 possible anchors would produce $\alpha_0 = 64$. The $\alpha_{0,\text{map}}$ and $\alpha_{0,\text{sc}}$ values are also lower for the OOD input than the ID one. We hypothesize that intersections are more likely in the ID data, where the agent of interest is traveling at a slow speed, than large four-lane roads

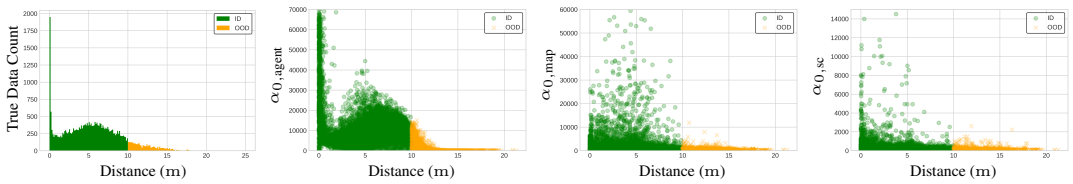

Figure 2: Learned $\alpha_{0,\text{agent}}$, $\alpha_{0,\text{map}}$, and $\alpha_{0,\text{sc}}$ parameters in comparison to a histogram of the true data distribution. All data is from the ID and OOD test sets, aside from the histogram ID data, which reflects the training distribution. In the scatter plots, outliers with large α_0 values were omitted for better visualization. The $\alpha_{0,\text{agent}}$ parameters accurately reflect the data distribution according to the input speed of the agent of interest. The speed is represented as the distance between the earliest and most recent waypoints in the input trajectory.

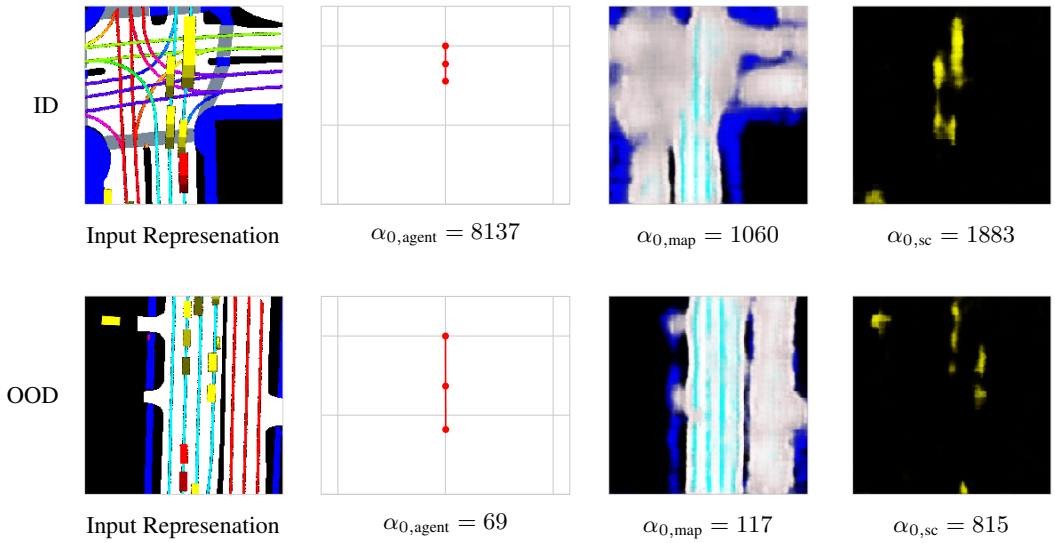

Figure 3: Qualitative results from the proposed ISAP framework on in-distribution (ID) and out-of-distribution (OOD) examples. The OOD example has a faster input trajectory than the ID one (plot scales for the decoded trajectories are the same). Thus, the learned epistemic uncertainty in the agent behavior component of the ISAP network is high ($\alpha_{0,\text{agent}}$ is low) for the OOD case and low ($\alpha_{0,\text{agent}}$ is high) for the ID case.

that resemble highways as in the OOD example. A surprising observation in Fig. 3 is that the OOD $\alpha_{0,\text{sc}}$ value is quite high, suggesting relatively low epistemic uncertainty. In both the ID and OOD examples, the agent of interest is traveling in traffic with many cars on the road. Such scenarios are likely common for the slow-speed training data, indicating ID characteristics. Thus, our ISAP paradigm provides insight into the interpretable sources of epistemic uncertainty.

6 Conclusions

In this paper, we propose a new trajectory prediction paradigm termed Interpretable Self-Aware Prediction (ISAP). ISAP learns the aleatoric and epistemic uncertainty over a discrete set of supervised anchors in a trajectory prediction architecture. These uncertainties are estimated in one-shot by using ideas from evidential deep learning. We infuse interpretability into the epistemic uncertainty estimate by subdividing the uncertainty into semantic concepts: past agent behavior, road structure map, and social context. Our approach maintains comparable trajectory prediction performance to an unmodified trajectory prediction architecture and outperforms established techniques, like ensembles, in uncertainty estimation while requiring only a single network pass during inference.

Limitations. Although we show that normalizing flows in the ISAP framework are performant for uncertainty estimation, they can be brittle and slow to train. They also struggle to scale to large latent spaces, restricting the representational capacity of the latent space. Furthermore, although the drop in trajectory prediction metrics is small with the addition of epistemic uncertainty estimation, it is prudent to consider how this gap could be closed in future work.

Future Work. Another promising avenue for future work is extending our proposed ISAP paradigm to map-centric spatiotemporal environment prediction settings [57, 58, 59]. In map-centric prediction, the inputs and outputs are sequential occupancy grid maps, which are higher dimensional than the object-centric input and output representations in traditional trajectory prediction architectures. Map-centric representations are useful since they are robust to partial occlusions, can handle an arbitrary number of agents in the scene, and do not require significant preprocessing. Thus, scaling epistemic uncertainty estimation to these settings is an exciting open research problem.

References

[1] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. Human motion trajectory prediction: A survey. *International Journal of Robotics Research*, 39(8): 895–935, 2020.

[2] F. Leon and M. Gavrilescu. A review of tracking and trajectory prediction methods for autonomous driving. *Mathematics*, 9(6):660, 2021.

[3] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, et al. The limits and potentials of deep learning for robotics. *International Journal of Robotics Research*, 37(4-5):405–420, 2018.

[4] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR)*, 2017.

[5] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436. IEEE, 2015.

[6] K. Nguyen and B. O'Connor. Posterior calibration and exploratory analysis for natural language processing models. *Conference on Empirical Methods in Natural Language Processing*, 2015.

[7] J. Provost Foster, F. Tom, and K. Ron. The case against accuracy estimation for comparing induction algorithms. In *International Conference on Machine Learning (ICML)*, pages 445–453, 1998.

[8] D. Yu, J. Li, and L. Deng. Calibration of confidence measures in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2461–2473, 2011.

[9] J. Nitsch, M. Itkina, R. Senanayake, J. Nieto, M. Schmidt, R. Siegwart, M. J. Kochenderfer, and C. Cadena. Out of distribution detection for automotive perception. In *International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2021.

[10] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision (ECCV)*, pages 835–851. Springer, 2016.

[11] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine. Stochastic variational video prediction. In *International Conference on Learning Representations (ICLR)*, 2018.

[12] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision (ECCV)*, 2020.

[13] M. Itkina, B. Ivanovic, R. Senanayake, M. J. Kochenderfer, and M. Pavone. Evidential sparsification of multimodal latent spaces in conditional variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

[14] P. Chen, M. Itkina, R. Senanayake, and M. J. Kochenderfer. Evidential softmax for sparse multimodal distributions in deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

[15] H. Cheng, W. Liao, M. Y. Yang, B. Rosenhahn, and M. Sester. AMENet: Attentive maps encoder network for trajectory prediction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 172:253–266, 2021.

[16] M. Itkina, Y.-J. Mun, K. Driggs-Campbell, and M. J. Kochenderfer. Multi-agent variational occlusion inference using people as sensors. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2022.

[17] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv*, 2021.

[18] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016.

[19] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3179–3189, 2018.

[20] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7047–7058, 2018.

[21] Q. Yu and K. Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *International Conference on Computer Vision (ICCV)*, pages 9518–9526. IEEE, 2019.

[22] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 14707–14718, 2019.

[23] A. Vyas, N. Jammalamadaka, X. Zhu, D. Das, B. Kaul, and T. L. Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *European Conference on Computer Vision (ECCV)*, pages 550–564, 2018.

[24] A. Amini, W. Schwarting, A. Soleimany, and D. Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[25] A. Malinin, S. Chervontsev, I. Provilkov, and M. Gales. Regression prior networks. *arXiv*, 2020.

[26] F. K. Gustafsson, M. Danelljan, and T. B. Schon. Evaluating scalable Bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 318–319, 2020.

[27] J. Postels, M. Segu, T. Sun, L. Van Gool, F. Yu, and F. Tombari. On the practicality of deterministic epistemic uncertainty. *arXiv*, 2021.

[28] R. McAllister, G. Kahn, J. Clune, and S. Levine. Robustness to out-of-distribution inputs via task-aware generative uncertainty. In *International Conference on Robotics and Automation (ICRA)*, pages 2083–2089. IEEE, 2019.

[29] A. Filos, P. Tigkas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *International Conference on Machine Learning (ICML)*, pages 3145–3153. PMLR, 2020.

[30] A. Farid, S. Veer, and A. Majumdar. Task-driven out-of-distribution detection with statistical guarantees for robot learning. In *Conference on Robot Learning (CoRL)*, pages 970–980. PMLR, 2021.

[31] J. Lee, J. Feng, M. Humt, M. G. Müller, and R. Triebel. Trust your robots! Predictive uncertainty estimation of neural networks with sparse Gaussian processes. In *Conference on Robot Learning (CoRL)*, pages 1168–1179. PMLR, 2021.

[32] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[33] M. Sensoy, L. Kaplan, F. Cerutti, and M. Saleki. Uncertainty-aware deep classifiers using generative models. In *Conference on Artificial Intelligence*. AAAI, 2020.

[34] B. Charpentier, D. Zügner, and S. Günnemann. Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1356–1367, 2020.

[35] E. Schmerling, K. Leung, W. Vollprecht, and M. Pavone. Multimodal probabilistic model-based planning for human-robot interaction. In *International Conference on Robotics and Automation (ICRA)*, pages 1–9. IEEE, 2018.

[36] B. Ivanovic and M. Pavone. The Trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *International Conference on Computer Vision (ICCV)*. IEEE, 2019.

[37] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov. MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Conference on Robot Learning (CoRL)*, pages 86–99, 2019.

[38] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff. CoverNet: Multi-modal behavior prediction using trajectory sets. In *Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14074–14083. IEEE, 2020.

[39] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid, et al. TNT: Target-driven trajectory prediction. In *Conference on Robot Learning (CoRL)*, 2020.

[40] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. NuScenes: A multimodal dataset for autonomous driving. In *Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631. IEEE, 2020.

[41] P. Kothari, B. Sifringer, and A. Alahi. Interpretable social anchors for human trajectory forecasting in crowds. In *Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15556–15566. IEEE, 2021.

[42] K. Mangalam, Y. An, H. Girase, and J. Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *International Conference on Computer Vision (ICCV)*, pages 15233–15242. IEEE, 2021.

[43] Y. Hu, W. Zhan, L. Sun, and M. Tomizuka. Multi-modal probabilistic prediction of interactive behavior via an interpretable model. In *Intelligent Vehicles Symposium (IV)*, pages 557–563. IEEE, 2019.

[44] M. Neumeier, M. Betsch, A. Tollkühn, and T. Berberich. Variational autoencoder-based vehicle trajectory prediction with an interpretable latent space. In *International Conference on Intelligent Transportation Systems (ITSC)*, pages 820–827. IEEE, 2021.

[45] S. Capobianco, N. Forti, L. M. Millefiori, P. Braca, and P. Willett. Uncertainty-aware recurrent encoder-decoder networks for vessel trajectory prediction. In *International Conference on Information Fusion (FUSION)*, pages 1–5. IEEE, 2021.

[46] P. Dijt and P. Mettes. Trajectory prediction network for future anticipation of ships. In *International Conference on Multimedia Retrieval (ICMR)*, pages 73–81. ACM, 2020.

[47] A. Bhattacharyya, M. Fritz, and B. Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4194–4202. IEEE, 2018.

[48] A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5574–5584, 2017.

[49] D. Bauer, L. Kuhnert, and L. Eckstein. Deep, spatially coherent inverse sensor models with uncertainty incorporation using the evidential framework. In *Intelligent Vehicles Symposium (IV)*, pages 2490–2495. IEEE, 2019.

[50] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 13991–14002, 2019.

[51] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning (ICML)*, pages 1530–1538. PMLR, 2015.

[52] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016.

[53] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[54] B. F. de Brito, H. Zhu, W. Pan, and J. Alonso-Mora. Social-VRNN: One-shot multi-modal trajectory prediction for interacting pedestrians. In *Conference on Robot Learning (CoRL)*, pages 862–872. PMLR, 2020.

[55] J. Postels, H. Blum, Y. Strümpler, C. Cadena, R. Siegwart, L. Van Gool, and F. Tombari. The hidden uncertainty in a neural networks activations. *arXiv*, 2020.

[56] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[57] M. Itkina, K. Driggs-Campbell, and M. J. Kochenderfer. Dynamic environment prediction in urban scenes using recurrent representation learning. In *Intelligent Transportation Systems Conference (ITSC)*, pages 2052–2059. IEEE, 2019.

[58] M. Toyungyernsub, M. Itkina, R. Senanayake, and M. J. Kochenderfer. Double-prong occupancy ConvLSTM: Spatiotemporal prediction in urban environments. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

[59] B. Lange, M. Itkina, and M. J. Kochenderfer. Attention augmented ConvLSTM for environment prediction. In *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021.

[60] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6306–6315, 2017.