CaDeT: a Causal Disentanglement Approach for Robust Trajectory Prediction in Autonomous Driving

Mozhgan Pourkeshavarz, Junrui Zhang, Amir Rasouli
Noah's Ark Lab, Huawei, Canada
firstname.lastname@huawei.com

Abstract

For safe motion planning in real-world, autonomous vehicles require behavior prediction models that are reliable and robust to distribution shifts. The recent studies suggest that the existing learning-based trajectory prediction models do not posses such characteristics and are susceptible to small perturbations that are not present in the training data, largely due to overfitting to spurious correlations while learning.

In this paper, we propose a causal disentanglement representation learning approach aiming to separate invariant (causal) and variant (spurious) features for more robust learning. Our method benefits from a novel intervention mechanism in the latent space that estimates potential distribution shifts resulted from spurious correlations using uncertain feature statistics, hence, maintaining the realism of interventions. To facilitate learning, we propose a novel invariance objective based on the variances of the distributions over uncertain statistics to induce the model to focus on invariant representations during training. We conduct extensive experiments on two large-scale autonomous driving datasets and show that besides achieving state-of-the-art performance, our method can significantly improve prediction robustness to various distribution shifts in driving scenes. We further conduct ablative studies to evaluate the design choices in our proposed framework.

1. Introduction

Vehicle trajectory prediction is one of the main building blocks of autonomous driving systems. Prediction captures how the future might unfold based on the road structure and the behavior of the road users. Accurate prediction of nearby traffic agents, however, is a daunting task due to the complex spatiotemporal interactions between the road users and the environment. Recently, learning-based methods [12, 21, 36, 43, 59, 60, 72] have become increasingly prevalent in the trajectory prediction domain, achieving state-of-the-art performance on the existing behavior prediction benchmarks [9, 53, 63].

To be deployed in real-world, besides being accurate, prediction models must be robust and reliable under different conditions, i.e. they must be insensitive to spurious features. However, recent evidence [3, 6, 45, 69] suggests that the existing trajectory prediction models for autonomous driving are susceptible to distributions that are even slightly different from their training data. Such lack of generalizability is mainly attributed to the tendency of the models to overfit to spurious correlations in the training data [45]. The difficulties pertaining generalizability cannot effectively be mitigated by utilization of large-scale models and data [47]. These issues are intrinsically rooted in statistical learning, which prioritizes the identification of correlations, exclusively for the prediction task, without the consideration for their robustness and reusability under distribution shifts that may occur in practice.

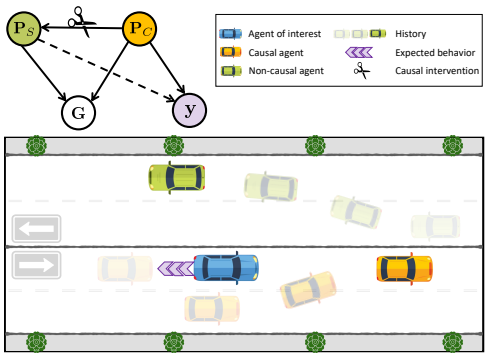

Figure 1. Illustration of a driving scenario showing two vehicles surrounding the agent of interest where one represents a causal factor causing the agent of interest to slow down and the other is non-causal, hence poses no impact. Here, $\mathbf{P_S} \leftarrow \mathbf{P}_C$ would open a backdoor (a non-causal dashed link) creating a spurious correlation. The proposed approach first separates causal and spurious factors through the spatiotemporal environment of scene $\mathbf{G}$ and then cuts the backdoor path via an uncertainty-driven causal intervention, inducing the model to focus only on causal patterns $\mathbf{P_C}$ when predicting future behavior $\mathbf{y}$, and discard the spurious patterns $\mathbf{P_S}$.

To this end, we frame the prediction problem through the lens of disentangled representation learning. Specifically, we seek to force the model to learn disentangled representations that split the underlying sources of variation in the data. This could pave the way for separating invariant (causal) from variant (spurious) representations (see Figure 1). In turn, the model can rely on the invariant features when making predictions, leading to more robust and generalizable inference.

We formulate representation learning from a causal perspective in a dynamic heterogeneous information network

which models the spatiotemporal interaction patterns between the agents and the agents and the environment. We argue that invariant representations correspond to causal variables that have cause→effect relation with the correct prediction. Motivated by this insight, we propose a causal disentanglement approach to discover invariant and variant factors via an uncertainty-aware intervention mechanism designed to create multiple intervened distributions. To maintain the realism of the perturbations induced by the intervention, we estimate the distributions based on uncertain features statistics in the latent space. We further propose an invariance objective based on the variances of the distributions to induce the model to focus on invariant representations in the training time. We conduct extensive empirical evaluations on two large-scale benchmark datasets and show that besides achieving state-of-the-art performance, our proposed method can significantly improve prediction robustness to various distribution shifts.

Our contributions are as follows: (1) We propose a novel causal disentanglement approach for trajectory prediction that enhances model robustness and generalization by effectively separating causal from spurious factors. (2) We simulate potential distribution shifts via a novel intervention mechanism by incorporating uncertainty modeling in the latent space to maintain the realism. (3) We design a new invariance training objective that focuses on leveraging causal factors for predictions while mitigating the negative effects of spurious correlations, thereby enhancing the model's resilience against potential distribution shifts. (4) We conduct experiments on common benchmark datasets and show that our model achieves state-of-the-art performance on various metrics. (5) We further conduct comprehensive studies on the robustness of our approach against various contextual perturbations followed by ablative analyses highlighting the contributions of proposed components on the overall performance.

2. Related Works

2.1. Trajectory Prediction

In autonomous driving, trajectory prediction is about forecasting future behavior of road users for safe motion planning. The literature in this domain is vast offering a variety of solutions for the prediction task [2, 4, 7, 25, 37, 48]. Given the highly dynamic and interactive nature of driving scenes, graph neural networks [18, 43, 46, 68] and transformer-based models [2, 15, 23, 37, 49, 72] are more prevalent. These approaches take advantage of heterogeneous contextual information, such as agents' dynamics and maps to create rich representations, which in turn are used for inferring future trajectories.

Causality in Prediction. The key consideration for the safety-critical autonomous driving application is robustness and generalizability of the prediction models. One way to achieve these is to induce the models to learn the underlying cause→effect relationships between driving scene elements while minimizing the effect of spurious correlations. Recently, in the domain of human motion prediction, a number of approaches have surfaced that are inspired by causal theory [10, 16, 33]. The method in [33] employs a dual-encoder architecture to capture invariant (i.e.physical laws) and style (i.e. domain-specific) representations learned via two objectives optimized in a sequential fashion. The authors of [10] perform an analysis by intervention in the feature space by adding counterfactual features, such as uniform rectilinear motion or random trajectory. In [16] a backdoor adjustment mechanism is used to incorporate social environment patterns into prediction using a social cross-attention module, hence removing the confounding effects of spurious correlations.

In this paper, we explore causality in the domain of autonomous driving and propose an intervention mechanism to disentangle invariant (causal) and variant (spurious) representations. To maintain realism, we approximate different variant representational distributions in the latent space. Unlike the past works, our approach relies on a more general mechanism that approximates the distributions based on statistical uncertainty of spurious patterns resulted from the data.

2.2. Domain Generalization

In recent years, there has been a growing focus on the development of models that generalize well to related but unseen test domains. One of the common techniques for this objective is ensembling in which a collection of diverse models or modules are used to improve the generalization and robustness of the predictions [71]. Although effective in trajectory prediction [51, 58, 72], ensembling comes at the cost of increasing the model's complexity, making this approach less practical for real-world applications.

Invariant representation learning is another technique to achieve generalizability by enabling models to learn features that are invariant to domain changes. For instance, domain alignment [28, 30] is used to minimize distances between different distributions, hence, forcing the model to learn invariant features. Disentangled representation learning (DLR) [29, 42, 65] is a more general approach that decouples features into variant and domain-invariant components in the observed data and learns their representations simultaneously without the need for direct knowledge of the adopted domain.

Disentangled Representation Learning. DLR is an unsupervised learning method that aims to characterize latent explanatory factors behind the observed data. This method can potentially lead to more robust, explainable, and transferable knowledge, as evident in a wide range of domains [14, 26, 35, 54, 62, 66, 67], in particular, in continual learning works where DRL is used to train a network incrementally to mimic human perception and cognition [1, 31, 64]. In computer vision, DRL is used to disentangle the identity of faces from their views or pose information in order to improve face recognition and anti-spoofing models [52, 56, 61], and 3D facial expression modeling [52]. DRL has also been used in the video analysis domain, for instance, for retrieving subtle human actions from co-occurring contextual elements [57, 65]. This method of learning is also employed in domains, such as natural language generation for separating writing style from text content to facilitate text-style transfer [24]. Despite hav-

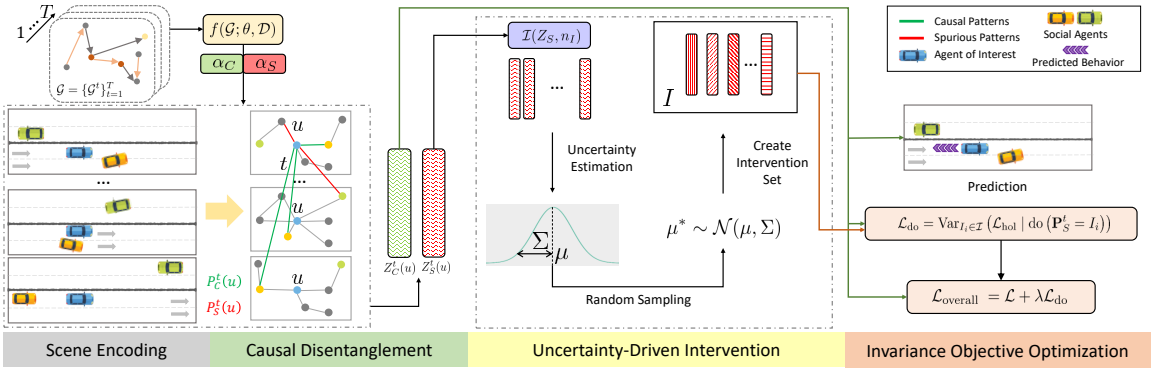

Figure 2. Overview of the proposed approach. We represent the scene content as a DyHIN and then using the uncertainty-driven intervention mechanism, disentangle the features into causal and spurious groups. Using the invariant loss, we suppress spurious feature forcing the model to rely on causal features to make predictions.

ing many desirable properties [34], DRL remains largely unexplored in the trajectory prediction domain.

3. Methodology

3.1. Problem Formulation

The goal of trajectory prediction is to forecast future locations of surrounding agents according to their observed behavior. Specifically, at time step t, let the past trajectory of the i-th vehicle be a set of $2D$ coordinates in bird's eye view over some observation horizon O time steps $X_i = \{(x_i, y_i)^{t-O+1}, \cdots, (x_i, y_i)^t\}$. Accordingly, the goal is to predict future trajectory $Y_i = \{(x_i, y_i)^{t+1}, \cdots, (x_i, y_i)^{t+H}\}$, where H is the prediction horizon. The road information extracted from the driving scenes in the form of an HD map represented in the vector space is also provided. For simplicity, in the rest of the paper, we refer to inputs, e.g., observations and the map, and future predictions as x and y, respectively.

3.2. Scene Encoding.

Recent studies show that explicitly modeling heterogeneity of driving scenes can improve models' ability to interpret complex interactions, for instance, between different types of agents, and agents and lanes [13, 22, 43, 70]. Following this approach, temporal and spatial information are modeled sequentially to learn heterogeneous relations. We argue that adding the time dimension to heterogeneity can uncover interactions that happened in non-uniform intervals that might be hidden when temporal and spatial steps are separated. In this regard, we model the driving scene as a Dynamic Heterogeneous Information Network (DyHIN) to model both structural and time-related characteristics of the scene simultaneously.

Definition. Dynamic Heterogeneous Information Network (DyHIN) is defined as a dynamic graph $\mathcal{G} = \{\mathcal{G}^t\}_{t=1}^T$, where T is the number of time stamps and $\mathcal{G}^t = (V^t, E^t)$ is the heterogeneous graph slice at timestamp t where $V = \bigcup_{t=1}^T V^t, E = \bigcup_{t=1}^T E^t$. At each timestamp, V is the set of nodes and E is the set of edges, each representing a binary relation between two nodes in V. $\mathcal{G}$ consists of two mappings: 1) node type mapping function $\phi : V \rightarrow T$ and 2) edge type mapping function $\psi : E \rightarrow R$, where T and R denote sets of node and edge types, respectively. In DyHIN, each node has a dynamic neighborhood $\mathcal{N}_u^{1:t}$ within a time horizon t which includes all neighbor nodes that have l-order ($1 < l < L$) interaction with the target node.

We encode the scene as a directed DyHIN with node types $T = \{\text{lane, agent}\}$ and $R = \{\text{left, right, successor, predecessor, lane-agent}\}$ as basic relations between adjacent lanes and between the lanes and agents, e.g., vehicle, cyclist, and pedestrians. To initialize node features in the DyHIN, we use a simplified PointNet model [44] with a multi-layer perception (MLP) to process polyline features and a 1D convolution with a feature pyramid network [32] to process the observations at each time step. For heterogeneous structures, we use the relative temporal encoding technique [19] to capture the dynamic structural dependencies with arbitrary durations within DyHINs. For the sake of brevity, the details and formulation are presented in the supplementary materials. One advantage of the proposed method is that any types of graph neural network (GNN) can be used to process the DyHIN. Here, we employ Dynamic Heterogeneous Graph Transformer (DyHGT) as a simple yet effective model [19].

3.3. Causal Disentanglement

Learning disentangled representations to split underlying sources of variation in high dimensional data is essential for data efficient and robust use of data-driven models. In trajectory prediction, as in our case, there is a high degree of variation in the input space, hence, overfitting and sensitivity to spurious correlations are highly probable. Therefore, disentanglement can be utilized to discover invariant representations that are more generalizable across potential distribution shifts.

In multi-agent environments, such as driving scenes, the agents' behaviors affect one another, hence, producing cause→effect relations. In the existing datasets, however, such causal relations are not explicitly labeled, therefore, supervised learning methods cannot be applied. Instead, by coupling causality with representation learning, we propose a causal disentanglement framework to discover and utilize invariant representations, e.g., causal factors, through the latent space in an unsupervised manner, while eliminating the effect of variant representation, e.g., spurious factors. Below, we explain how

the representation is modeled, followed by describing the proposed causal disentanglement objective and a detailed look at the representation disentangling mechanism.

In dynamic driving environments, temporal interactions among heterogeneous elements, such as different agent types, agents and static context (i.e. lanes), form complex patterns. Potential perturbations in these patterns are the source of variations in the driving scenes. Thus, exploring spatiotemporal patterns provides fine-grained representations of the scenes' dynamics that help learn how the interactions, governed by physical laws, evolve through time.

However, to forecast the future behavior of an agent, not all patterns within the scenes have causal effects. For instance, in an overtaking scenario, patterns resulted from road constraints and movements of the passing vehicle are causal, whereas patterns created by an approaching vehicle in the opposite lane are non-causal, i.e. spurious. Therefore, we aim to disentangle representations obtained from spatiotemporal patterns.

To this aim, for each agent, we establish a spatiotemporal environment which is a dynamic neighboorhood in the DyHIN. The spatiotemporal pattern, then, is a subset of the agent's spatiotemporal environment as $P_u^t = \alpha(\mathbf{G}_u^{1:t})$ where α identifies structures and features. These patterns are referred to as spatiotemporal patterns in the remainder of the paper.

3.3.1 Causality-inspired Objective

From the causal perspective, we formulate the disentangled representation learning in the DyHIN with a structural causal model (SCM) [8]. Let $\mathbf{P}_C$ and $\mathbf{P}_S$ denote the invariant and variant representations formed by causal and spurious patterns. As such, we define the causal links as $\mathbf{P}_S \rightarrow \mathbf{G} \leftarrow \mathbf{P}_C \rightarrow \mathbf{y}$ and $\mathbf{P}_S \leftarrow \mathbf{P}_C$. In the former, $\mathbf{P}_S \rightarrow \mathbf{G} \leftarrow \mathbf{P}_C$ denotes that causal and spurious patterns together construct the agent's spatiotemporal environment within the DyHIN and $\mathbf{P}_C \rightarrow \mathbf{y}$ implies that causal patterns determine the expected, i.e. ground truth (GT), behavior $\mathbf{y}$, *no matter how the spurious patterns change due to potential perturbations*. Recalling the overtaking scenario example, road constraints and the movements of the passing vehicle should induce a slow-down behavior in the agent of interest, whereas changes in the behavior of the approaching vehicles in the opposite direction should not.

At times, the associations between spurious patterns and the GT may occur forming a backdoor path [40] (i.e., non-causal path) as $\mathbf{P}_S \leftarrow \mathbf{P}_C \rightarrow \mathbf{y}$, leading to a statistical relation termed spurious correlation. Models that highly rely on such backdoor links may fail under potential perturbations on non-causal factors. From this insight, backed by causal theory [40, 41], we assume that if the driving scene representations are disentangled, there exists a causal subset that is sufficient to predict the correct behavior (i.e., GT).

Theory. Causal Invariance. Let assume that for a given task there is a predictor $f(\cdot)$ for input samples $(\mathcal{G}_u^{1:t}, y^t)$ derived from a distribution. There exists causal patterns $P_C^t(u)$ and non-causal, e.g., spurious, patterns $\mathbf{P}_S^t(u)$ such that $y^t = f(\mathbf{P}_C^t(u)) + \epsilon$ and $\mathbf{P}_C^t(u) = \mathcal{G}_v^{1:t} \backslash \mathbf{P}_S^t(u)$, i.e., $\mathbf{y}^t \perp \mathbf{P}_S^t(u) \mid \mathbf{P}_C^t(u)$. In the presence of spurious patterns, therefore, the model should rely on causal patterns to achieve better generalizability. Hence, we define our objective as below:

$$\begin{aligned} &\min_{\theta_1,\theta_2} \mathbb{E}_{(\mathbf{y}^t,\mathcal{G}_u^{1:t})} \mathcal{L}(f_{\theta_1}(\tilde{\mathbf{P}}_C^t(u)), \mathbf{y}^t) \\ &\text{s.t} \quad \Phi_{\theta_2}(\mathcal{G}_u^{1:t}) = \tilde{\mathbf{P}}_C^t(u), \quad \mathbf{y}^t \perp \tilde{\mathbf{P}}_S^t(u) \mid \tilde{\mathbf{P}}_C^t(u). \end{aligned} \tag{1}$$

where $\mathbf{y}^t$ indicate the GT. $\Phi_{\theta_2}(\cdot)$ seeks to find spurious patterns and $f_{\theta_1}(\cdot)$ makes predictions based on causal patterns. This objective is used as the means of eliminating spurious correlations and is minimized by reducing the effect of spurious patterns on the GT. From the causal perspective, this objective corresponds to *block the backdoor path* in the causal graph $\mathbf{P}_S \leftarrow \mathbf{P}_C \rightarrow \mathbf{y}$.

For optimization, we use a do-calculus $\mathrm{do}(\mathbf{P}_S)$ to intervene spurious patterns and therefore cutting the causal links from causal patterns to spurious patterns [40]. In this way, the model can learn the direct causal effects from causal patterns to the GT in the intervened distributions $p(\mathbf{y}, \mathbf{G} \mid \mathrm{do}(\mathbf{P}_S))$. Since the risks should be the same across the distributions, we can minimize the variance of empirical risks under different intervened distributions to help the model focus on the relations between the causal patterns and GT [40, 55]. The objective in Eq. 1, therefore, can be transformed into,

$$\begin{aligned} &\min_{\theta_1,\theta_2} \mathbb{E}_{(y^t,\mathcal{G}_u^{1:t})} \mathcal{L}(f_{\theta_1}(\Phi_{\theta_2}(\mathcal{G}_u^{1:t})), \mathbf{y}^t) + \\ &\mathrm{Var}_{I\in\mathcal{I}}\, \lambda \mathbb{E}_{(\mathbf{y}^t,\mathcal{G}_v^{1:t}|\mathrm{do}(\mathbf{P}_S^t=I))} \mathcal{L}(f_{\theta_1}(\Phi_{\theta_2}(\mathcal{G}_u^{1:t})), \mathbf{y}^t) \end{aligned} \tag{2}$$

where λ is a balancing hyperparameter. As such, minimizing the variance term in Eq. 2 helps the model satisfy the constraint $\mathbf{y}^t \perp \tilde{\mathbf{P}}_S^t(u) \mid \tilde{\mathbf{P}}_C^t(u)$ in Eq. 1, which corresponds to,

$$p(\mathbf{y}^t \mid \tilde{\mathbf{P}}_C^t(u), \tilde{\mathbf{P}}_S^t(u)) = p(\mathbf{y}^t \mid \tilde{\mathbf{P}}_C^t(u)). \tag{3}$$

Here, if we have the optimal predictor $f_{\theta_1}^*$ and pattern finder $\Phi_{\theta_2}^*$ according to Eq. 1, then the variance term in Eq. 2 is minimized because variant patterns will not affect the predictions of $f_{\theta_1}^* \circ \Phi_{\theta_2}^*$ across different intervened distributions. We will discuss the proposed intervention mechanism in Sec. 3.4.

3.3.2 Disentangled Attention Block

From the viewpoint of causality, we aim to decouple the encoded representation as two disjoint spurious and causal sets. To do so, we propose a disentangled attention block as an addition to the transformer-based GNN (DyHGT). Thus, for node u at timestamp t and its dynamic neighbors $w \in \mathcal{N}^{t'}(u), \forall t' \leq t$, we calculate the Query-Key-Value vectors as:

$$\begin{aligned} &\mathbf{q}_u^t = \mathbf{W}_q(\mathbf{Z}_u^t \| \,\mathrm{R}(t)), \\ &\mathbf{k}_v^{t'} = \mathbf{W}_k(\mathbf{Z}_v^{t'} \| \,\mathrm{R}(t')), \mathbf{v}_v^{t'} = \mathbf{W}_v(\mathbf{Z}_v^{t'} \| \,\mathrm{R}(t')), \end{aligned} \tag{4}$$

where $\mathbf{Z}_u^t$ denotes the representation of node u at timestamp t, and $\mathbf{q}, \mathbf{k}$ and $\mathbf{v}$ represent the query, key and value vectors, respectively. Here, we omit the bias term for simplicity. $\mathrm{R}(t)$ denotes the relative temporal technique for encoding the time when interactions happen. Next, we calculate the attention scores among the nodes in the dynamic neighborhood of each node to obtain the disentanglement masks as below:

$$\alpha_C = \text{Softmax}(\mathbf{q} \cdot \mathbf{k}^T / \sqrt{d}),$$
$$\alpha_S = \text{Softmax}(-\mathbf{q} \cdot \mathbf{k}^T / \sqrt{d}), \tag{5}$$

where d denotes the feature dimension, and α_C and α_S represent the disentangled masks of causal and spurious patterns, respectively. In this way, dynamic neighbors with higher attention scores in causal patterns will have lower attention scores in spurious ones, meaning that the causal and spurious patterns have a negative correlation. We also selectively re-weight the causal representations through a learnable representational mask $\alpha_r = \text{Softmax}(\mathbf{w}_r)$. Hence, the dynamic neighborhoods' messages in the GNN can be summarized with the obtained disentangled masks as follows:

$$\tilde{\mathbf{Z}}_C^t(u) = \sum_i \alpha_{C,i}(\mathbf{r}_i \odot \alpha_r),$$
$$\tilde{\mathbf{Z}}_S^t(u) = \sum_i \alpha_{S,i}\mathbf{r}_i, \tag{6}$$
$$\mathbf{Z}_{C/S}^t(u) = \phi(\tilde{\mathbf{Z}}_{C/S}^t(u) + \mathbf{Z}(u)^t),$$

where r stands for representation and $\phi(\cdot)$ is a transformation layer. Lastly, the disentangled representations are aggregated to be fed into subsequent layers,

$$\mathbf{Z}(u)^t \leftarrow \mathbf{Z}_C^t(u) + \mathbf{Z}_S^t(u). \tag{7}$$

Note that, similar to classic message-passing in GNNs, the disentanglement mechanism enables each node to indirectly access high-order dynamic neighborhoods, where $\mathbf{Z}_C^t(u)$ and $\mathbf{Z}_S^t(u)$ at l-th layer in our method is a summarization of causal and spurious patterns in l-order dynamic neighborhood.

3.4. Uncertainty-driven Causal Intervention

To identify causal connections in the graph, one way is to directly intervene by generating or altering the connections in the DyHIN. This approach, however, is infeasible since the alterations can potentially make the driving scene representation unrealistic. Here, we propose an alternative intervention mechanism that creates multiple distributions by intervening in variant representations formed by spurious patterns. To maintain the realism, we approximate the distributions in the latent space using the embedding statistics of the data.

In general, embedding statistics, including mean and standard deviation, contain informative domain characteristics of the data [20, 27, 29]. This means that small perturbations in data points may cause uncertain statistics shifts with varying directions and magnitudes [20, 29]. In trajectory prediction, there is a wide range of variations in the context due to the dynamic nature of the driving scenes, therefore, such perturbations are highly probable. As a result, prediction models that are trained unaware of the potential shifts in uncertain statistics tend to be more sensitive to perturbations in spurious patterns.

We estimate the potential uncertain statistics shifts as a way of approximating intervened distributions. Here, the feature statistics are hypothesized to follow a multivariate Gaussian distribution after considering potential uncertainties. Specifically, having each representation's original statistics values in the center, the distribution scope determines the level of intervention considering potential distribution shifts.

In our method, we estimate the distribution of feature statistics based on the variances of the mini-batch statistics in a non-parametric manner [29]. Here, by referring to a batch of encoded embeddings of patterns $\mathbf{Z} \in \mathbb{R}^{b \times d}$, we denote the feature statistics mean and standard deviation as $\mu(\mathbf{Z})$ and $\sigma^2(\mathbf{Z})$, respectively, and define the non-parametric model for uncertainty estimation as follows:

$$\Sigma_\mu^2(\mathbf{Z}) = \frac{1}{b}\sum_{j=1}^{b}(\mu(\mathbf{Z}) - \mathbb{E}_j[\mu(\mathbf{Z})])^2$$
$$\Sigma_\sigma^2(\mathbf{Z}) = \frac{1}{b}\sum_{j=1}^{b}(\sigma(\mathbf{Z}) - \mathbb{E}_j[\sigma(\mathbf{Z})])^2, \tag{8}$$

where b denotes batch size and d the feature dimension. Here, feature statistics variants are randomly sampled from the estimated Gaussian distribution and then are used to construct the intervention set $\mathcal{I}$ in Eq. 2. In this way, the magnitudes of uncertainty estimation can reveal the possibility that the corresponding latent space may potentially change by perturbing spurious patterns. Although the underlying distribution of changes in the latent space is unpredictable, the uncertainty estimation captured from the mini-batch can provide an appropriate and meaningful variation range for the latent space.

3.5. Training Objective

Based on the multiple-intervened data distributions, we can optimize the model to focus on causal patterns for the prediction. We present an invariance training objective to instantiate Eq. 2. Let $\mathbf{Z}_C$ and $\mathbf{Z}_S$ be the summarized causal and spurious patterns' representations resulting from the disentanglement step. We first calculate the prediction loss $\mathcal{L}$ by using only $\mathbf{Z}_C$, allowing the model to utilize causal patterns for predictions. We follow the commonly used regression-head plus classification-head combination [22, 38, 43] for trajectory prediction.

Next, we calculate a holistic loss $\mathcal{L}_{\text{hol}}$ to measure the model's prediction ability when spurious patterns are exposed to it. At the end, the invariance training objective is given by,

$$\mathcal{L}_{\text{do}} = \text{Var}_{I_i \in \mathcal{I}}(\mathcal{L}_{\text{hol}} \mid \text{do}(\mathbf{P}_S^t = I_i)). \tag{9}$$

Here, the objective measures the variance of the model's prediction ability under multiple intervened distributions. The final training objective is therefore as follows:

$$\mathcal{L}_{\text{overall}} = \mathcal{L} + \lambda\mathcal{L}_{\text{do}}, \tag{10}$$

where the task loss $\mathcal{L}$ is minimized to exploit causal patterns while the invariance loss $\mathcal{L}_{do}$ helps the model to discover causal and variant patterns. Here, λ is a hyperparameter to balance the two objectives. In the inference stage, we only use causal patterns to make predictions.

4. Experiments

We compare the proposed method against state-of-the-art trajectory prediction approaches on autonomous driving benchmarks. We specifically seek to examine the robustness of our

Table 1. Quantitative results on the AGV2 motion forecasting leaderboard. The "†" indicates an ensemble version. For each metric, the best result is in **bold** and the second best result is underlined.

Method	Reference	b-minFDE$_6$	minADE$_6$	minFDE$_6$	MR$_6$	minADE$_1$	minFDE$_1$	MR$_1$
THOMAS [17]	ICLR 2022	2.16	0.88	1.51	0.20	1.95	4.71	0.64
MTR [50]	NeurIPS 2022	1.98	0.73	1.44	**0.15**	1.74	4.39	**0.58**
GANet [59]	ICRA 2023	1.96	0.72	1.35	0.17	1.77	4.47	0.59
GoRela [13]	ICRA 2023	2.01	0.76	1.48	0.22	1.82	4.62	0.66
FRM [39]	ICLR 2023	2.47	0.89	1.81	0.29	2.37	5.93	0.71
QCNet [72]	CVPR2023	1.91	**0.65**	1.29	0.16	**1.69**	**4.30**	0.59
ProphNet [60]	CVPR 2023	1.88	0.68	1.33	0.18	1.80	4.74	-
Forecast-MAE [11]	ICCV 2023	2.02	0.70	1.39	0.17	1.74	4.35	0.60
HPTR [70]	NeurIPS 2023	2.03	0.73	1.43	0.19	1.84	4.61	0.61
Forecast-MAE†[11]	ICCV 2023	1.91	0.69	1.33	0.17	1.65	4.14	0.59
QCNet†[72]	CVPR 2023	1.78	0.62	1.19	0.14	1.56	3.95	0.55
CaDeT (Ours)	-	**1.86**	0.67	**1.24**	**0.15**	1.74	4.33	**0.58**

Figure 3. Qualitative results of CaDeT on AGV2. The Agent of interest is shown in red color and color intensity on other vehicles indicate causality score (darker is higher). The observation, ground truth, and prediction are shown as orange, green, and purple, respectively.

method against various contextual perturbations in order to highlight the benefits of the proposed causal disentanglement mechanism. We refer to our model as Causal Disentanglement for Trajectory prediction or short for **CaDeT**.

4.1. Experimental Setup

Datasets. We evaluate our method on two large-scale motion forecasting datasets including Argoverse-2 (AGV2) [9] and Waymo Open Motion Dataset (WOMD) [53]. AGV2 contains 250K scenarios split into 200K, 25K, and 25K samples for training, validation, and testing, respectively. The task is to make $6s$ predictions based on $5s$ observations. WOMD consists of 487K training scenes and validation and testing set each with 44K scenes. Here, the objective is to predict $8s$ into the future based on $1s$ observation.

Metrics. We use official benchmark metrics, minimum average displacement error (minADE$_K$), minimum final displacement error (minFDE$_K$), b-minFDE$_K$, miss rate (MR$_K$), and mean average precision (mAP). Here, K refers to the number of predicted trajectories in the multimodal setting.

For robustness evaluation, we adopt the perturbation resistance score (PRS) metric computed as follows: Given a trajectory prediction error metric m, we first measure the per-sample absolute change in the metric as $\text{abs}(\Delta) = \frac{1}{n}\sum_{i=1}^{n} \mid m_{\text{perturbed}}(i) - m_{\text{original}}(i) \mid$ where $m_{\text{perturbed}}$ and $m_{original}$ refer to the errors on perturbed and original data respectively. PRS is then calculated as $[1 - (\text{abs}(\Delta)/m_{\text{original}})] * 100$ showing how robust the model is against specific perturbations.

Implementation Details. For data representation, we use all agent types and lanes whose distance from the agent is smaller than 100 meters. For the architecture, our model has six layers of DyHGT with 128 hidden dimension. For the transformation layer in the attention block, we use layer normalization, an MLP and a skip connection as $\alpha \cdot \text{MLP}(\text{LayerNorm}(\mathbf{x})) + (1-\alpha)\cdot \mathbf{x}$ where α is a learnable parameter. For training, we set the size of intervention set n_I to 1000 and λ to 1 and 0.1 for AGV2 and WOMD datasets, respectively (See Fig. 4). We train using batch size of 32, the AdamW optimizer with an initial learning rate of $5e-4$, and weight decay $1e{-}4$. For the objective, we use the crossentropy loss for the classification-head and negative log likelihood loss for the regression-head.

4.2. Comparison to State-of-the-art

We begin by evaluating the proposed model, **CaDeT** on the AGV2 dataset. As shown in Table 1, our approach achieves state-of-the-art performance on the majority of metrics and stands second best on the rest with small margins. Performance of our model is also comparable to significantly more complex model ensembles and in some cases surpasses them. For instance, on $k = 6$ metrics, our model performs better on all metrics compared to ensemble Forecast-MAE by up to 7% (see Figure 3 for qualitative examples).

Table 2. Analyzing robustness against *adversarial perturbations* on AGV2 validation. (↓) and (↑) indicate lower and higher values are better.

Training-size (%)	Method	Targeted			Non-Targeted		
		Original / Perturbed	abs(Δ)(↓)	PRS (%)(↑)	Original / Perturbed	abs(Δ)(↓)	PRS (%)(↑)
d = 100	Forecast-MAE [11]	0.712 / 0.879	0.153±0.06	78.51	0.712 / 0.755	0.132±0.09	81.46
	CaDeT (Ours)	0.701 / 0.841	0.087±0.04	87.59	0.701 / 0.737	0.061±0.02	91.30
d = 80	Forecast-MAE [11]	0.733 / 0.924	0.198±0.03	72.99	0.733 / 0.863	0.174±0.07	76.26
	CaDeT (Ours)	0.721 / 0.859	0.094±0.02	86.96	0.721 / 0.822	0.082±0.04	88.62
d = 50	Forecast-MAE [11]	0.761 / 1.177	0.259±0.04	65.96	0.761 / 1.167	0.257±0.06	66.23
	CaDeT (Ours)	0.759 / 0.918	0.117±0.05	84.58	0.759 / 0.886	0.095±0.08	87.48

Table 3. Analyzing generalizability to an *unseen domain*.

Test Domain	Method	minADE_6	minFDE_6	MR_6
Unseen	Forecast-MAE [11]	0.897	1.613	0.216
	CaDeT (Ours)	0.638	1.182	0.157
Seen	Forecast-MAE [11]	0.837	1.489	0.173
	CaDeT (Ours)	0.713	1.186	0.168

4.3. Robustness and Generalization

Adversarial Perturbations. DRL involves learning to represent input data in such a way that various sources of variation in the data are separated, or disentangled, into distinct, non-overlapping features. As a result, DRL can potentially enhance data efficiency in data-driven models.

To investigate the practical impact of our method, we conduct an experiment on the AGV2 validation set for which we train the models using $d\%$ of the data. To evaluate the robustness, we apply adversarial perturbations to the observed dynamics of the agents using the method in [69]. Following the recommended setting, we use a white box approach and train the models using the Adam optimizer with learning rate of 0.01. For a fair comparison, we set the maximum iteration to 100 with maximum deviation of 1 to optimize perturbations.

We compare our model against Forecast-MAE [11] which is state-of-the-art model trained using a self-supervised learning technique for robust and generalizabile contextual feature learning. The results are reported in Table 2 and divided into targeted and non-targeted perturbations depending on whether the adversarial objectives were optimized according to the reported metric, in this case minADE. The findings suggest that the proposed model is significantly more robust as the performance degradation is drastically lower. Comparing original and perturbed results, the degradation for our model is as low as 20% at $d = 100\%$ and as high as 21% at $d = 50\%$ whereas for Forecast-MAE the values are 30% and 55%, respectively. These values not only show the robustness of our model but also its lack of sensitivity to the amount of data used for training. In fact, as also shown in the changes in PRS metric, our model maintains a similar level of robustness even when trained on only 50% of the data but this is not the case for Forecast-MAE as its PSR value dropped by more than 12%. In non-targeted perturbations, as one would expect, both models perform better, however, our model gains more at the two extreme cases of training with 100% and 50% of data.

Generalizability. One way to evaluate generalizability of trajectory prediction models for autonomous driving is to train the models on data collected in one environment, e.g. a city, and test them in another one with different road types and traffic styles. To this end, we conduct an experiment in which we split the train and validation data of AGV2 into two non-overlapping sets based on the six cities that the data was collected from as in [5]. Then we train the model on the data from the first set of the cities, namely Miami, Pittsburgh and Austin, and evaluate on **unseen** cities. We also evaluate the models using the original train and validation set that are sampled from all the cities and refer to these experiments as **seen**.

The results are summarized in Table 3 and show that our method significantly outperforms Forecast-MAE in both settings on all metrics. Of particular interest is the performance gap in the unseen experiment, where our model performs up to 27% better. This suggests that even when the test domain is not seen during training, the proposed model can effectively learn causal factors and predict the correct trajectories.

Causal Perturbations. Not all agents in traffic scenarios play equally pivotal roles in influencing the future behavior. Some agents, termed non-causal, do not significantly impact predictions directly. Hence, if misinterpreted as causal, these agents can lead to spurious correlations. One way to evaluate their negative impact is by removing non-causal agents from the scenes to determine how the models resist the perturbation.

In this regard, we conduct an experiment using the newly released causal annotations on the validation subset of WOMD[1] [45]. These labels are provided for social agents surrounding the autonomous vehicle (AV) indicating whether the agents have causal impacts on the future behavior of the AV. We follow the evaluation protocol in [45] and report the prediction error by minADE and robustness by PSR. In addition, we train the models under two conditions, *trained-All* where all agents are used and *trained-AV* where only the AV is used.

As shown in Table 4, the proposed model, besides being comparable or better on the test set, demonstrates a notably higher level of robustness by achieving 12% higher PRS compared to the best model, MTR. It is worth noting that the degradation of our model's robustness is minimal, by only 3%, even when reducing the training data by 50%, maintaining its top position compared to other models. This observation further confirms the data efficiency of our method as well as its robustness against removal of non-causal agents.

Comparing the results across Trained-All and -AV, overall the accuracy of all models improve as expected, since the models are optimized with respect to the AV. However, at the same time their robustness has also dropped when perturbations are applied. Despite such a change, our model still performs best maintaining its performance gap with MTR.

[1]The dataset is available at https://github.com/google-research/causal-agents

Table 4. Robustness of prediction methods against *causal perturbations* in WOMD. (↓) and (↑) indicate lower and higher values are better.

	Validation set without non-causals						Test Set		
	Trained-All			Trained-AV					
Method / Metrics	**Original / Perturbed**	**abs(Δ)(↓)**	**PRS(%)(↑)**	**Original / Perturbed**	**abs(Δ)(↓)**	**PRS(%)(↑)**	**mAP**	**minADE$_6$**	**minFDE$_6$**
MTR [50]	0.384 / 0.407	0.075±0.18	80 (%)	0.339 / 0.360	0.072±0.06	78 (%)	0.412	0.605	1.221
HDGT [22]	0.567 / 0.582	0.125±0.26	78 (%)	0.407 / 0.425	0.098±0.16	76 (%)	0.357	0.768	1.108
Multipath++ [58]	0.900 / 0.945	0.226±0.32	75 (%)	0.376 / 0.395	0.141±0.21	62 (%)	0.409	0.556	1.158
SceneTransformer [38]	0.305 / 0.328	0.081±0.14	73 (%)	0.250 / 0.265	0.067±0.12	73 (%)	0.279	0.612	1.212
CaDeT (d= 50%)	0.346 / 0.358	0.037±0.18	89 (%)	0.298 / 0.336	0.042±0.17	86 (%)	0.373	0.570	1.169
CaDeT (d= 100%)	0.312 / 0.327	0.026±0.12	**92** (%)	0.253 / 0.254	0.026±0.09	**90** (%)	0.390	0.545	1.136

Table 5. Ablation studies and related design choices of the proposed method on the validation set of AGV2. $\mathcal{D}$ stands for the causal disentanglement framework and $\mathcal{I}$ for intervention. x in $\mathcal{I}_x$ represents s : spatial, t : temporal, and u : uncertainty-driven, respectively.

Model	$\mathcal{D}$	$\mathcal{I}_s$	$\mathcal{I}_t$	$\mathcal{I}_u$	**b-FDE$_6$**	**minADE$_6$**	**minFDE$_6$**	**MR$_6$**
Baseline					2.07	0.93	1.57	0.21
M1	✓				2.03	0.88	1.48	0.20
M2	✓	✓		✓	1.92	0.77	1.33	0.17
M3	✓		✓	✓	1.89	0.75	1.29	0.17
M4	✓	✓	✓		1.95	0.80	1.37	0.18
CaDeT	✓	✓	✓	✓	**1.87**	**0.71**	**1.22**	**0.16**

4.4. Ablation Study

We conduct an ablation study on the proposed model using DyHIN as baseline and the AGV2 validation set.

Representation Disentanglement. The proposed representation disentanglement $\mathcal{D}$ paves the way for separating causal and spurious factors. Particularly, the disentanglement step is the prerequisite to accomplishing robust learning in prediction. We show this by adding the disentanglement attention block to the baseline and refer to it as **M1** in Table 5. Here, we can observe a performance boost of up to 6% across all metrics verifying the benefit of disentangling representations.

Spatial vs. Temporal Intervention. The dynamic nature of traffic in driving scenes is the primary source of variations. We model these variations as spatiotemporal patterns in order to capture both structural and temporal relations. Here, we conduct experiments using two versions of our baseline, namely **M2** and **M3**. In **M2**, we impose a constraint that the variable patterns utilized for intervention must originate from the same timestamp ($\mathcal{I}_s$), thereby prohibiting interventions across different time steps. In **M3**, we set a constraint that requires the variable patterns for the intervention to be sourced exclusively from the same node, i.e., the agent, across temporal dimension ($\mathcal{I}_t$). As the findings in Table 5 suggest, adding either form of intervention, the performance improves on all metrics, indicating the benefit of the proposed intervention mechanism. Here, higher improvement gain using **M3** indicates that temporal patterns have a higher potential to create spurious correlations in traffic scenes.

Effect of Uncertainty-driven Intervention. Lastly, we validate the impact of the proposed uncertainty-driven intervention mechanism ($\mathcal{I}_u$) and create a variation of our model, **M4**, in which we approximate the intervention process by sampling and replacing the variant pattern summarizations at random. We achieve this by gathering different patterns from all agents at every timestamp, and then, choosing one and use it to substitute the patterns found in other nodes throughout time.

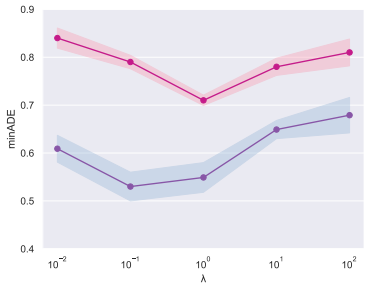

Figure 4. Sensitivity of λ in WOMD (top) and AGV2 (bottom).

As shown in Table 5, using the alternative intervention mechanism, there is a drop across all metrics compared to standalone interventions using our uncertainty mechanism in **M2** and **M3**. This suggests that using uncertainty sampling is necessary to maintain the realism of the interventions in order to simulate potential distribution shifts effectively.

Controlling Intervention Impact. As noted in Eq. 10, we use hyperparameter, λ, to control the intervention objective, i.e. its influence on the overall prediction task objective. To determine the sensitivity of this parameter, we conduct a study on AGV2 and WOMD using the minADE metric. As shown in Figure 4, there is a drop in the performance when the λ parameter is either too small or too large. If λ is set too small, the model does not suppress spurious features effectively, and if set too large, the model fails to learn the causal features.

5. Conclusion

We proposed a novel causal disentanglement approach in which we formulated both the spatial and temporal relations in the scenes through spatiotemporal patterns and used a causal disentanglement approach to separate causal and spurious factors. We proposed an intervention mechanism to simulate the potential distribution shifts at inference time by generating multiple intervened distributions based on spurious factors in the latent space using feature statistics, hence, maintaining the realism of interventions. Lastly, we proposed an invariance training objective to leverage causal factors and intervened distributions to induce the model to focus on causal relations. As a result, the influence of spurious correlations are mitigated making the model more robust against distribution shifts during inference time. We conducted extensive empirical studies on two large-scale autonomous driving datasets and demonstrated that our approach not only achieves state-of-the-art performance but also significantly improves upon prediction robustness against various distribution shifts.

References

[1] Alessandro Achille, Tom Eccles, Loic Matthey, Chris Burgess, Nicholas Watters, Alexander Lerchner, and Irina Higgins. Life-long disentangled representation learning with cross-domain latent homologies. *NeurIPS*, 2018. 2

[2] Gorkay Aydemir, Adil Kaan Akan, and Fatma Guney. Adapt: Efficient multi-agent trajectory prediction with adaptation. In *ICCV*, 2023. 2

[3] Mohammadhossein Bahari, Saeed Saadatnejad, Ahmad Rahimi, Mohammad Shaverdikondori, Amir Hossein Shahidzadeh, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Vehicle trajectory prediction works, but not everywhere. In *CVPR*, 2022. 1

[4] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst. In *RSS*, 2019. 2

[5] Prarthana Bhattacharyya, Chengjie Huang, and Krzysztof Czarnecki. SSL-Lanes: Self-supervised learning for motion forecasting in autonomous driving. In *CoRL*, 2022. 7

[6] Yulong Cao, Chaowei Xiao, Anima Anandkumar, Danfei Xu, and Marco Pavone. Advdo: Realistic adversarial attacks for trajectory prediction. In *ECCV*, 2022. 1

[7] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *CoRL*, 2019. 2

[8] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 2017. 4

[9] Ming-Fang Chang, John W Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019. 1, 6

[10] Guangyi Chen, Junlong Li, Jiwen Lu, and Jie Zhou. Human trajectory prediction via counterfactual analysis. In *ICCV*, 2021. 2

[11] Jie Cheng, Xiaodong Mei, and Ming Liu. Forecast-mae: Self-supervised pre-training for motion forecasting with masked autoencoders. In *ICCV*, 2023. 6, 7

[12] Sehwan Choi, Jungho Kim, Junyong Yun, and Jun Won Choi. R-pred: Two-stage motion prediction via tube-query attention-based trajectory refinement. In *ICCV*, 2023. 1

[13] Alexander Cui, Sergio Casas, Kelvin Wong, Simon Suo, and Raquel Urtasun. Gorela: Go relative for viewpoint-invariant motion forecasting. In *ICRA*, 2023. 3, 6

[14] Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. Debiasing graph neural networks via learning disentangled causal substructure. *NeurIPS*, 2022. 2

[15] Shaoheng Fang, Zi Wang, Yiqi Zhong, Junhao Ge, and Siheng Chen. Tbp-former: Learning temporal bird's-eye-view pyramid for joint perception and prediction in vision-centric autonomous driving. In *CVPR*, 2023. 2

[16] Chunjiang Ge, Shiji Song, and Gao Huang. Causal intervention for human trajectory prediction with cross attention mechanism. In *AAAI*, 2023. 2

[17] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. THOMAS: Trajectory heatmap output with learned multi-agent sampling. In *ICLR*, 2022. 6

[18] Daniel Grimm, Philip Schörner, Moritz Dreßler, and J.-Marius Zöllner. Holistic graph-based motion prediction. In *ICRA*, 2023. 2

[19] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *The World Wide Web Conference*, 2020. 3

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *CVPR*, 2017. 5

[21] Zhiyu Huang, Haochen Liu, and Chen Lv. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *ICCV*, 2023. 1

[22] Xiaosong Jia, Penghao Wu, Li Chen, Hongyang Li, Yu Liu, and Junchi Yan. Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding. *arXiv:2205.09753*, 2022. 3, 5, 8

[23] Chiyu "Max" Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, and Dragomir Anguelov. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *CVPR*, 2023. 2

[24] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. *arXiv:1808.04339*, 2018. 2

[25] Siddhesh Khandelwal, William Qi, Jagjeet Singh, Andrew Hartnett, and Deva Ramanan. What-if motion prediction for autonomous driving. *arXiv:2008.10587*, 2020. 2

[26] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 2

[27] Boyi Li, Felix Wu, Ser-Nam Lim, Serge Belongie, and Kilian Q Weinberger. On feature normalization and data augmentation. In *CVPR*, 2021. 5

[28] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 2

[29] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. *ICLR*, 2022. 2, 5

[30] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018. 2

[31] Zhiyuan Li, Xiajun Jiang, Ryan Missel, Prashnna Kumar Gyawali, Nilesh Kumar, and Linwei Wang. Continual unsupervised disentangling of self-organizing representations. In *ICLR*, 2022. 2

[32] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3

[33] Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. Towards robust and adaptive motion forecasting: A causal representation perspective. In *CVPR*, 2022. 2

[34] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*. PMLR, 2019. 3

[35] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018. 2

[36] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *CVPR*, 2023. 1

[37] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S. Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. In *ICRA*, 2023. 2

[38] Jiquan Ngiam, Vijay Vasudevan, Benjamin Caine, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene Transformer: A unified architecture for predicting future trajectories of multiple agents. In *ICLR*, 2022. 5, 8

[39] Daehee Park, Hobin Ryu, Yunseo Yang, Jegyeong Cho, Jiwon Kim, and Kuk-Jin Yoon. Frm: Leveraging future relationship reasoning for vehicle trajectory prediction. In *ICLR*, 2023. 6

[40] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 4

[41] Judea Pearl et al. Models, reasoning and inference. *Cambridge University Press*, 2000. 4

[42] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *ICML*. PMLR, 2020. 2

[43] Mozhgan Pourkeshavarz, Changhe Chen, and Amir Rasouli. Learn tarot with mentor: A meta-learned self-supervised approach for trajectory prediction. In *ICCV*, 2023. 1, 2, 3, 5

[44] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 3

[45] Rebecca Roelofs, Liting Sun, Ben Caine, Khaled S Refaat, Ben Sapp, Scott Ettinger, and Wei Chai. CausalAgents: A robustness benchmark for motion forecasting using causal relationships. *arXiv:2207.03586*, 2022. 1, 7

[46] Luke Rowe, Martin Ethier, Eli-Henry Dykhne, and Krzysztof Czarnecki. Fjmp: Factorized joint multi-agent motion prediction over learned directed acyclic interaction graphs. In *CVPR*, 2023. 2

[47] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *ICML*, 2020. 1

[48] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *ECCV*, 2020. 2

[49] Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S. Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *ICCV*, 2023. 2

[50] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr: Motion transformer with global intention localization and local movement refinement. *NeurIPS*, 2022. 6, 8

[51] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *arXiv:2306.17770*, 2023. 2

[52] Hao Sun, Nick Pears, and Yajie Gu. Information bottlenecked variational autoencoder for disentangled 3d facial expression modelling. In *CVPR*, 2022. 2

[53] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 1, 6

[54] Ayush Tewari, Xingang Pan, Ohad Fried, Maneesh Agrawala, Christian Theobalt, et al. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *CVPR*, 2022. 2

[55] Jin Tian, Changsung Kang, and Judea Pearl. A characterization of interventional distributions in semi-markovian causal models. In *AAAI*, 2006. 4

[56] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017. 2

[57] Beatrice van Amsterdam, Abdolrahim Kadkhodamohammadi, Imanol Luengo, and Danail Stoyanov. Aspnet: Action segmentation with shared-private representation of multiple data sources. In *CVPR*, 2023. 2

[58] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. MultiPath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *ICRA*, 2022. 2, 8

[59] Mingkun Wang, Xinge Zhu, Changqian Yu, Wei Li, Yuexin Ma, Ruochun Jin, Xiaoguang Ren, Dongchun Ren, Mingxu Wang, and Wenjing Yang. Ganet: Goal area network for motion forecasting. In *ICRA*, 2023. 1, 6

[60] Xishun Wang, Tong Su, Fang Da, and Xiaodong Yang. Prophnet: Efficient agent-centric motion forecasting with anchor-informed proposals. In *CVPR*, 2023. 1, 6

[61] Yu-Chun Wang, Chien-Yi Wang, and Shang-Hong Lai. Disentangled representation with dual-stage feature learning for face anti-spoofing. In *CVPR*, 2022. 2

[62] Xin Wen, Junsheng Zhou, Yu-Shen Liu, Hua Su, Zhen Dong, and Zhizhong Han. 3d shape reconstruction from 2d images with disentangled attribute flow. In *CVPR*, 2022. 2

[63] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv:2301.00493*, 2023. 1

[64] Maciej Wolczyk, Michal Zajkac, Razvan Pascanu, Lukasz Kucinski, and Piotr Milos. Disentangling transfer in continual reinforcement learning. *NeurIPS*, 2022. 2

[65] Kun Xia, Le Wang, Sanping Zhou, Nanning Zheng, and Wei Tang. Learning to refactor action and co-occurrence features for temporal action localization. In *CVPR*, 2022. 2

[66] Yinghao Xu, Menglei Chai, Zifan Shi, Sida Peng, Ivan Skorokhodov, Aliaksandr Siarohin, Ceyuan Yang, Yujun Shen, Hsin-Ying Lee, Bolei Zhou, et al. Discoscene: Spatially disentangled generative radiance fields for controllable 3d-aware scene synthesis. In *CVPR*, 2023. 2

[67] Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *NeurIPS*, 2022. 2

[68] Wenyuan Zeng, Ming Liang, Renjie Liao, and Raquel Urtasun. LaneRCNN: Distributed representations for graph-centric motion forecasting. In *IROS*, 2021. 2

[69] Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Qi Alfred Chen, and Z Morley Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. In *CVPR*, 2022. 1, 7

[70] Zhejun Zhang, Alexander Liniger, Christos Sakaridis, Fisher Yu, and Luc Van Gool. Real-time motion prediction via heterogeneous polyline transformer with relative pose encoding. *NeurIPS*, 2023. 3, 6

[71] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *Transactions on Image Processing*, 2021. 2

[72] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *CVPR*, 2023. 1, 2, 6