Towards Robust and Adaptive Motion Forecasting: A Causal Representation Perspective

Yuejiang Liu **Riccardo Cadei** **Jonas Schweizer** **Sherwin Bahmani** **Alexandre Alahi**
École Polytechnique Fédérale de Lausanne (EPFL)
{firstname.lastname}@epfl.ch

Abstract

Learning behavioral patterns from observational data has been a de-facto approach to motion forecasting. Yet, the current paradigm suffers from two shortcomings: brittle under covariate shift and inefficient for knowledge transfer. In this work, we propose to address these challenges from a causal representation perspective. We first introduce a causal formalism of motion forecasting, which casts the problem as a dynamic process with three groups of latent variables, namely invariant mechanisms, style confounders, and spurious features. We then introduce a learning framework that treats each group separately: (i) unlike the common practice of merging datasets collected from different locations, we exploit their subtle distinctions by means of an invariance loss encouraging the model to suppress spurious correlations; (ii) we devise a modular architecture that factorizes the representations of invariant mechanisms and style confounders to approximate a sparse causal graph; (iii) we introduce a style consistency loss that not only enforces the structure of style representations but also serves as a self-supervisory signal for test-time refinement on the fly. Experiment results on synthetic and real datasets show that our three proposed components significantly improve the robustness and reusability of the learned motion representations, outperforming prior state-of-the-art motion forecasting models for out-of-distribution generalization and low-shot transfer.

1 Introduction

Motion forecasting is essential for autonomous systems running in dynamic environments. Yet, it is a challenging task due to strong spatial-temporal interactions. These interactions arise from two major sources: (i) physical laws (*e.g.*, inertia, goal-directed behaviors) that govern general dynamics; (ii) social norms (*e.g.*, social distance, left or right-hand traffic) that influence motion styles. Classical models attempt to describe these interactions based on domain knowledge but often fall short of social awareness in complex scenes [1, 2]. As an alternative, learning motion representations from observational data has become a *de-facto* approach [3–5]. In light of the rapid progress over the past few years, solving motion forecasting is seemingly just around the corner by pursuing this fashion at larger scales.

However, the promise of the current learning paradigm for motion forecasting is shadowed by two shortcomings:

- struggle to discover the physical laws from data, *e.g.*, output inadmissible solutions under spurious shifts [6];
- inefficient for knowledge transfer, *e.g.*, require a large number of observations to adapt from one environment to another even if the underlying change is sparse [7].

These issues do not become any less severe with larger models [8]. Instead, they are profoundly rooted in the principle of statistical learning that only seeks correlations for the prediction task at hand, regardless of their *robustness* and *reusability* under distributional shifts that may occur in practice .

Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

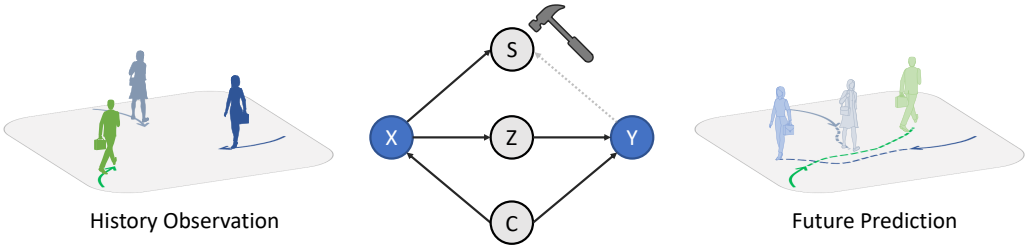


Figure 1: Our causal formalism of motion forecasting. We cast the problem as a dynamic process with three groups of latent variables: domain-invariant physical laws ($\boldsymbol{z}$), domain-specific style confounders ($\boldsymbol{c}$), and non-causal spurious features ($\boldsymbol{s}$). The spurious features are not parents of future movements ($\boldsymbol{y}$) in the causal factorization, *e.g.*, no edges or anti-causal (dotted line), and their statistical correlations may vary drastically under different conditions. This formalism motivates our design and learning method to promote the robustness and reusability of the learned motion representation.

In this work, we aim to tackle these challenges from a causal representation perspective. Incorporating causal relations into statistical modeling has garnered growing interest lately, as it not only offers a mathematical language to articulate distribution changes [9, 10] but also brings critical insights to representation learning [11–13]. Studies in cognitive science have also revealed its paramount importance in the motion context: few-month-old infants are already able to reason sensibly about physical and social causalities [14]; they can even learn that by sorely observing adult behaviors, without any hands-on experience of their own [15]. How can we build motion learning algorithms that are able do the same? To this end, we propose a method that explictly incorporates causal invariance and structure into motion representation learning in order to promote its robustness and transferability under common types of distributional shifts.

2 Method

2.1 Formalism of Motion Forecasting

Preliminary. Consider a motion forecasting problem in multi-agent environments. For a scene of M agents, let $s_t = \{s_t^1, \cdots, s_t^M\}$ denote their joint state and $s_t^i = (x_t^i, y_t^i)$ denote the state of an individual agent i at time t. The model takes an input sequence of past observations $\boldsymbol{x} = (s_1, \cdots, s_t)$ in order to predict their states in the future $\boldsymbol{y} = (s_{t+1}, \cdots, s_T)$ up to time T. Modern forecasting models are largely built with encoder-decoder neural networks, where the encoder $\phi(\cdot)$ extracts a compact motion representation $\boldsymbol{z}$ of the past observations and the decoder $g(\cdot)$ rolls out the predicted trajectory $\hat{\boldsymbol{y}}$.

The training data $\mathcal{D}$ is often collected from a set of K environments $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$. Previous work typically merges them into a large dataset and assumes the mixture as a representative of the unseen test environment $\tilde{e}$. Under this assumption, the model is trained to minimize the empirical risk:

$$\mathcal{R}(\phi, g) := \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \mathcal{L}_{\text{task}}(g(\phi(\boldsymbol{x})), \boldsymbol{y}). \tag{1}$$

where $\mathcal{L}_{\text{task}}$ is the loss function of the motion forecasting task, such as mean square error (MSE) or negative log-likelihood (NLL). However, the i.i.d. assumption does not always hold in practice. In fact, recent work [16] has shown that the test environment can be significantly different from the training ones in the widely used ETH-UCY benchmark. We will next introduce a causal formalism of motion forecasting that allows us to formulate this challenge and design solutions to address it.

Causal formalism. Motion behaviors are essentially dynamic processes governed by latent variables, such as physical laws, traffic rules and social norms. To build accurate predictive models, the conventional learning paradigm typically aims to discover these latent variables and model their correlations with the observed future states. However, the learned correlations may vary across environments and thus fail to generalize at test time.

To tackle this fundamental challenge, we introduce a new formalism of motion forecasting through the lens of causality. As shown in Figure 1, we categorize the latent variables into three groups:

- invariant variables: physical laws that are universal to everyone at any place;

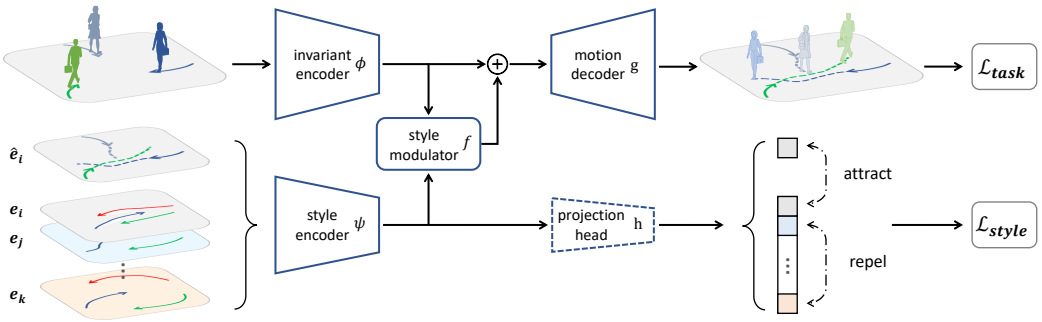


Figure 2: Our modular forecasting model contains two separate encoders for universal laws and style confounders, respectively. The model is built in three steps: (i) learn an invariant predictor based on the encoder ϕ with the goal to be equally optimal in all training environments (Sec 2.2), (ii) learn an embedding space based on the encoder ψ to capture the style relation between different scenes (Sec 2.4), (iii) incorporate domain-specific style features into the forecasting model by training ψ, f, g and h on the main task and the auxiliary style contrastive task jointly (Sec 2.3-2.4).

- hidden confounders: motion styles that may vary from site to site in a local and sparse manner;
- spurious features: other variables, *e.g.*, level of noises, that are not direct causes of future motion.

In principle, neither the second group nor the third have stable correlations with the target future motion. Yet, these two groups may lead to distinctive effects on forecasting models in the presence of environmental changes. The relations with spurious features can become largely different and even completely reversed in new domains, resulting in catastrophic errors at test time. In contrast, the change of motion styles is often more restricted. While failing to capture the correct style yields inaccurate predictions, it should still output plausible solutions subject to physical laws. We will next describe three algorithmic components that treat spurious features and hidden confounders differently in order to promote the robustness and reusability of the learned motion representations.

2.2 Causal Invariant Forecasting

Invariant principle. By definition, invariant features should have identical joint distributions with the target variable (future motion) across different environments, whereas the non-invariant ones are the opposite. This distinction can be formulated as a necessary condition for the domain invariant predictor, *i.e.*, $g \circ \phi$ is equally optimal in every environment [17]. More formally, our goal is to solve the following problem:

$$\begin{aligned} \min_{\phi,g} \quad & \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathcal{R}^e(\phi, g) \\ \text{s.t.} \quad & g \in \arg\min_{g^*} \mathcal{R}^e(\phi, g^*) \quad \forall e \in \mathcal{E}, \end{aligned} \tag{2}$$

where g^* is an optimal predictor built on top of the extracted features in an individual environment e. Intuitively, if a learned forecasting model can perform similarly well across multiple training environments, it is more likely to generalize to another related test environment $\tilde{e}$ as well.

Invariant loss. The exact form of the invariant learning principle (Eq. 2), however, leads to a bi-level optimization problem, which is difficult to solve in practice. Recent work [18, 19] proposed to relax it to a gradient norm penalty over the empirical risk $\mathcal{R}^e$ in each training environment:

$$\min_{\phi,g} \quad \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \left[\mathcal{R}^e(\phi, g) + \lambda \|\nabla_g \mathcal{R}^e(\phi, g)\|_2^2\right]. \tag{3}$$

This objective prevents the forecasting model from learning an *average* effect of spurious features on future trajectories and enforces the model to solve it the hard way by seeking universal mechanisms behind motion behaviors. We will show in Sec. B.1 that this technique can greatly improve the robustness of the forecasting model against the domain shifts to spurious correlations.

However, the strength of suppressing spurious features comes with a clear drawback, *i.e.*, the learned representation tends to erroneously drop the motion styles that change across environments. This may cause inaccurate predictions in both training and test environments. To cope with this issue, we next introduce a modular architecture that allows the model to properly structure the knowledge and strategically adapt from one style to another.

2.3 Modular Forecasting Model

Most recent forecasting models are built with dense connections at their core, albeit with some detailed differences. On the one hand, this design principle is very powerful when the training data is sufficient; on the other hand, it often lands in a highly inter-twined architecture that lack semantic structure. As such, in the presence of a style shift, one may have to update the whole model without any focus. This common practice inevitably leads to low sample efficiency for transfer learning. Ideally, a forecasting model could preserve a clear structure of the learned knowledge, separate the impacts of physical laws and motion styles on motion behaviors, and approximate the sparse causal graph in Figure 1.

To achieve this goal, we devise a modular network that consists of two encoders and one decoder. The first encoder ϕ is trained to compute domain-invariant features, as described in Sec 2.2. Subsequently, we introduce a second encoder ψ which captures a representation of motion styles varying across domains. Given a long sequence of scene observations $\boldsymbol{o}$ from a particular environment e, the role of the style encoder ψ is to infer the latent features of the style confounders $\boldsymbol{c}$. More formally, our modular network predicts the future trajectory as follows:

$$\boldsymbol{z} = \phi(\boldsymbol{x}), \qquad \boldsymbol{c} = \psi(\boldsymbol{o}), \qquad \tilde{\boldsymbol{z}} = f(\boldsymbol{z}, \boldsymbol{c}) + \boldsymbol{z}, \qquad \hat{\boldsymbol{y}} = g(\tilde{\boldsymbol{z}}), \tag{4}$$

where $\tilde{\boldsymbol{z}}$ is the latent feature that incorporates both $\boldsymbol{z}$ and $\boldsymbol{c}$, and the style modulator f can be modeled by a small MLP. Here, we can also compute $\boldsymbol{c}$ based on multiple scene observations from the same environment, *e.g.*, averaging several style feature vectors, to obtain a more robust estimate of the motion style. As shown in Figure 2, our modular design allows us to precisely localize and fine-tune a small subset of parameters to account for the underlying style shift.

2.4 Style Consistency Loss

Our modular forecasting model composed of multiple sub-networks can be practically difficult to train, especially in the few-shot transfer setting where data collected from the new environment is limited. To overcome this challenge, we introduce a style consistency loss, which aims to not only strengthen the modular structure of motion representation during training but also allows for reusing the encoded style knowledge at test time.

Style Contrastive Learning. Ideally, the feature vector produced by the style encoder should not only provide the basic style information for predicting the future motion accurately but also properly capture the style relation between different scenes. We formulate this intuition into an auxiliary task in the form of supervised contrastive learning. Specifically, we consider two scene observations from the same environment as a pair of positive samples, whereas those from different environment as negative pairs. We map the style feature $\boldsymbol{c}$ to a projected embedding $\boldsymbol{p}$ by a small head $h(\cdot)$. The style contrastive loss for a positive pair of samples (i, j) is as follows,

$$\mathcal{L}_{\text{style}} = -\log \frac{\exp(\text{sim}(\boldsymbol{p}_i, \boldsymbol{p}_j)/\tau)}{\sum_k \mathbb{1}_{[k=j \vee \boldsymbol{e}_k \neq \boldsymbol{e}_i]} \exp(\text{sim}(\boldsymbol{p}_i, \boldsymbol{p}_k)/\tau)}, \tag{5}$$

where $\mathbb{1}_{[\boldsymbol{e}_k \neq \boldsymbol{e}_i]}$ is an indicator function equal to 1 if and only if the two samples i and k are drawn from the same environment, τ is a temperature parameter and $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$ denotes the dot product between normalized $\boldsymbol{u}$ and $\boldsymbol{v}$ (*i.e.*, cosine similarity).

One key advantage of the proposed style contrastive loss over the conventional classification loss is that it does not impose any assumptions about the number of domain classes in the design of the projection head h. This property allows the model to incrementally bootstrap from the knowledge already learned about the existing styles to some additional ones without changing the shape of h or learning any parameters from scratch. This is particularly beneficial as in the transfer setting where the number of additional styles is not known a priori.

Test-time Refinement. One common phenomenon in transfer learning is that the model fine-tuned on only a few samples remains sub-optimal in the new environment. To alleviate this performance gap, we reuse the style contrastive loss as a self-supervisory signal for test-time refinement on the fly. Concretely, we feed the predicted output back as an input to the style encoder and iteratively adjust the internal feature $\tilde{\boldsymbol{z}}$. Here, the parameters to optimize are no longer the model weights but rather the feature activations per inference case. The refinement process gradually reduces the distances between the predicted output and the reference examples of style in the learned embedding space. By tightly coupling the modular architecture design with the style consistency loss, our method enables the effective use of the auxiliary contrastive task during both training and deployment.

Acknowledgments

This work is supported by the Swiss National Science Foundation under the Grant 2OOO21-L92326. We thank Bastien Van Delft, Brian Alan Sifringer and Yifan Sun for valuable feedback on early drafts.

References

[1] Dirk Helbing and Peter Molnar. Social Force Model for Pedestrian Dynamics. *Physics Review E*, May 1998.

[2] Jur van den Berg, Stephen J. Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-Body Collision Avoidance. In Cédric Pradalier, Roland Siegwart, and Gerhard Hirzinger, editors, *Robotics Research*, Springer Tracts in Advanced Robotics, pages 3–19, Berlin, Heidelberg, 2011. Springer.

[3] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, June 2016. ISSN: 1063-6919.

[4] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 683–700, Cham, 2020. Springer International Publishing.

[5] Osama Makansi, Julius von Kügelgen, Francesco Locatello, Peter Gehler, Dominik Janzing, Thomas Brox, and Bernhard Schölkopf. You Mostly Walk Alone: Analyzing Feature Attribution in Trajectory Prediction. *arXiv:2110.05304 [cs]*, October 2021. 00000 arXiv: 2110.05304.

[6] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, June 2011. ISSN: 1938-7228.

[7] Todor Davchev, Michael Burke, and Subramanian Ramamoorthy. Learning Structured Representations of Spatial and Interactive Dynamics for Trajectory Prediction in Crowded Scenes. *IEEE Robotics and Automation Letters*, 6(2):707–714, April 2021. Conference Name: IEEE Robotics and Automation Letters.

[8] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, November 2020. ISSN: 2640-3498.

[9] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018.

[10] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning series. MIT Press, Cambridge, MA, USA, November 2017.

[11] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards Causal Representation Learning. *arXiv:2102.11107 [cs]*, February 2021. arXiv: 2102.11107.

[12] Yixin Wang and Michael I. Jordan. Desiderata for Representation Learning: A Causal Perspective. *arXiv:2109.03795 [cs, stat]*, September 2021. arXiv: 2109.03795.

[13] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *arXiv:2109.00725 [cs]*, September 2021. arXiv: 2109.00725.

[14] Anne Schlottmann, Deborah Allen, Carina Linderoth, and Sarah Hesketh. Perceptual Causality in Children. *Child Development*, 73(6):1656–1677, 2002.

[15] Anna Waismeyer and Andrew N. Meltzoff. Learning to make things happen: Infants' observational learning of social and physical causal events. *Journal of Experimental Child Psychology*, 162:58–71, October 2017.

[16] Guangyi Chen, Junlong Li, Jiwen Lu, and Jie Zhou. Human Trajectory Prediction via Counterfactual Analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9824–9833, 2021. 00000.

[17] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12167.

[18] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv:1907.02893 [cs, stat]*, March 2020. arXiv: 1907.02893.

[19] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The Risks of Invariant Risk Minimization. In *International Conference on Learning Representations*, September 2020. 00046.

[20] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, June 2018. ISSN: 2575-7075.

[21] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6272–6281, 2019.

[22] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human Trajectory Forecasting in Crowds: A Deep Learning Perspective. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–15, 2021. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[23] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social NCE: Contrastive Learning of Socially-Aware Motion Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15118–15129, 2021.

[24] Deyao Zhu, Mohamed Zahran, Li Erran Li, and Mohamed Elhoseiny. Motion Forecasting with Unlikelihood Training in Continuous Space. In *5th Annual Conference on Robot Learning*, June 2021.

[25] Christina Heinze-Deml, Marloes H. Maathuis, and Nicolai Meinshausen. Causal Structure Learning. *arXiv:1706.09141 [stat]*, June 2017. arXiv: 1706.09141.

[26] Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like DAGs? A Survey on Structure Learning and Causal Discovery. *arXiv:2103.02582 [cs, stat]*, March 2021. arXiv: 2103.02582.

[27] Bryon Aragam and Qing Zhou. Concave penalized estimation of sparse gaussian bayesian networks. *J. Mach. Learn. Res.*, 16(1):2273–2328, 2015.

[28] Mauro Scanagatta, Giorgio Corani, Cassio P De Campos, and Marco Zaffalon. Learning treewidth-bounded bayesian networks with thousands of variables. In *In Advances in Neural Information Processing Systems*, pages 1462–1470, 2016.

[29] Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560, 2018.

[30] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *In Advances in Neural Information Processing Systems*, volume 33, 2020.

[31] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data with independent changes. *J. Mach. Learn. Res.*, 21(89):1–53, 2020.

[32] Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *Proceedings of the International Conference on Machine Learning*, pages 5541–5550, 2018.

[33] Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *J. Mach. Learn. Res.*, 16(1):2147–2205, 2015.

[34] Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. In *In Advances in Neural Information Processing Systems*, 2019.

[35] Chengchun Shi, Tianlin Xu, Wicher Bergsma, and Lexin Li. Double generative adversarial networks for conditional independence testing. *arXiv:1412.1442*, 2020.

[36] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *In Advances in Neural Information Processing Systems*, volume 20, pages 489–496, 2007.

[37] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv:1202.3775*, 2012.

[38] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[39] Huanfei Ma, Kazuyuki Aihara, and Luonan Chen. Detecting causality from nonlinear dynamics with short-term time series. *Scientific reports*, 4(1):1–10, 2014.

[40] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *In 2021 IEEE/CVF Conferenceon Computer Vision and Pattern Recognition*, pages 9593–9602, 2021.

[41] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

[42] Junsouk Choi, Robert Chapkin, and Yang Ni. Bayesian causal structural learning with zero-inflated poisson bayesian networks. In *In Advances in Neural Information Processing Systems*, volume 33, 2020.

[43] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[44] Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[45] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2172–2180. Curran Associates, Inc., 2016.

[46] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv:1804.03599 [cs, stat]*, April 2018. arXiv: 1804.03599.

[47] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a Definition of Disentangled Representations. *arXiv:1812.02230 [cs, stat]*, December 2018. arXiv: 1812.02230.

[48] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning Independent Causal Mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, July 2018. ISSN: 2640-3498.

[49] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness. In *International Conference on Machine Learning*, pages 6056–6065. PMLR, May 2019. ISSN: 2640-3498.

[50] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *arXiv:1811.12359 [cs, stat]*, June 2019. arXiv: 1811.12359.

[51] Peter Bühlmann. Invariance, Causality and Robustness. *arXiv:1812.08233 [stat]*, December 2018. 00057 arXiv: 1812.08233.

[52] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant Risk Minimization Games. *Proceedings of the International Conference on Machine Learning*, 1, 2020.

[53] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, pages 5815–5826. PMLR, July 2021. ISSN: 2640-3498.

[54] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from Several Related Classification Tasks to a New Unlabeled Sample. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. 00168.

[55] Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*, September 2020. 00109.

[56] Erick Delage and Yinyu Ye. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research*, 58(3):595–612, June 2010. 01272 Publisher: INFORMS.

[57] John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, June 2021. 00090 Publisher: Institute of Mathematical Statistics.

[58] Hamed Rahimian and Sanjay Mehrotra. Distributionally Robust Optimization: A Review. *arXiv:1908.05659 [cs, math, stat]*, August 2019. 00150 arXiv: 1908.05659.

[59] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*, February 2018. 04340.

[60] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to Unseen Domains via Adversarial Data Augmentation. In *Advances in Neural Information Processing Systems 31*, January 2018. 00251.

[61] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[62] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep Transfer Learning with Joint Adaptation Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2208–2217. PMLR, July 2017. ISSN: 2640-3498.

[63] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation Alignment for Unsupervised Domain Adaptation. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pages 153–171. Springer International Publishing, Cham, 2017.

[64] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning. *International Conference on Learning Representations*, International Conference on Learning Representations, November 2016. 00283.

[65] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, January 2016.

[66] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial Discriminative Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[67] Mei Wang and Weihong Deng. Deep Visual Domain Adaptation: A Survey. *arXiv:1802.03601 [cs]*, May 2018. arXiv: 1802.03601.

[68] Gabriela Csurka. Deep Visual Domain Adaptation. *arXiv:2012.14176 [cs]*, December 2020. arXiv: 2012.14176.

[69] Sicheng Zhao, Bo Li, Colorado Reed, Pengfei Xu, and Kurt Keutzer. Multi-source Domain Adaptation in the Deep Learning Era: A Systematic Survey. *arXiv:2002.12169 [cs, stat]*, February 2020. arXiv: 2002.12169.

[70] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing Machines. *arXiv:1410.5401 [cs]*, December 2014. 02045 arXiv: 1410.5401.

[71] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-Value Memory Networks for Directly Reading Documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas, November 2016. Association for Computational Linguistics. 00631.

[72] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv:1703.03400 [cs]*, March 2017. arXiv: 1703.03400.

[73] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and Jose M. F. Moura. Few-Shot Human Motion Prediction via Meta-Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 432–450, 2018. 00073.

[74] Chuanqi Zang, Mingtao Pei, and Yu Kong. Few-shot Human Motion Prediction via Learning Novel Motion Dynamics. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, volume 1, pages 846–852, July 2020. 00005 ISSN: 1045-0823.

[75] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It Is Not the Journey But the Destination: Endpoint Conditioned Trajectory Prediction. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 759–776, Cham, 2020. Springer International Publishing. 00072.

[76] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings. In *Computer Vision – ECCV 2010*, Lecture Notes in Computer Science, pages 452–465, Berlin, Heidelberg, 2010. Springer.

[77] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by Example. *Computer Graphics Forum*, 26(3):655–664, 2007.

[78] Changan Chen, Yuejiang Liu, Sven Kreiss, and Alexandre Alahi. Crowd-Robot Interaction: Crowd-Aware Robot Navigation With Attention-Based Deep Reinforcement Learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6015–6022, May 2019. ISSN: 2577-087X.

[79] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 01006.

[80] Yisong Yue, Patrick Lucey, Peter Carr, Alina Bialkowski, and Iain Matthews. Learning Fine-Grained Spatial Models for Dynamic Sports Play Prediction. In *2014 IEEE International Conference on Data Mining*, pages 670–679, December 2014. 00093 ISSN: 2374-8486.

[81] Daniel McDuff, Yale Song, Jiyoung Lee, Vibhav Vineet, Sai Vemprala, Nicholas Gyde, Hadi Salman, Shuang Ma, Kwanghoon Sohn, and Ashish Kapoor. CausalCity: Complex Simulations with Agency for Causal Discovery and Reasoning. *arXiv:2106.13364 [cs]*, June 2021. arXiv: 2106.13364.

A Related Work

Motion Forecasting. Modern motion forecasting models [3, 4, 20–22] are largely built with neural networks and trained with the maximum likelihood principle. Despite strong performance for short-range predictions within the training domain, they often struggle to generalize under covariate shift. Recently, a couple of works proposed to promote their robustness using negative data augmentation [23, 24]. However, designing negative examples of high-dimension, *e.g.*, long sequences, can be difficult in practice. Our work explores a causality-inspired alternative that does not require hand-engineered interventions over training data and is hence more theoretically grounded and algorithmically generic.

Closely related to ours, another recent work [16] attempts to mitigate biases in motion datasets through counterfactual analysis. Our method differs from theirs in three aspects: (i) their approach learns to estimate dataset biases before subtracting them in the feature space, whereas our method aims to directly suppress biased features; (ii) their approach inherits the merge-and-shuffle convention, which destroys some critical information for bias estimation; in contrast, we keep each subset separately and exploit unstable correlations across environments; (iii) the counterfactual problem formulated in their approach is generally difficult to solve (*cf.* Pearl's ladder of causation [9]); conversely, we consider spurious correlations from the interventional perspective, which is easier to tackle in practice.

Causal Learning. The intersection of causal inference and machine learning has been a vibrant area of research in the past few years [10, 12, 13]. Some earlier works attempted to identify causal structures from observational or interventional data [25, 26]. Examples include score-based [27–29], constraint-based [30–33], conditional independence test [34–37], continuous optimization [38–40] and many others [41–43]. While these methods are theoretically appealing, they are often practically restricted to classical problems that assume direct access to high-level causal variables rather than the low-level observations present in modern problems [11].

More recently, several different approaches have been proposed to automatically discover causal variables of interest from low-level data. One notable line of work lies in disentangled representation

learning [44–47], which is closely tied to independent causal mechanisms [48, 49]. However, separating independent factors of variation in an unsupervised manner is often exceedingly challenging without strong assumptions [50]. As an alternative, a few other recent works seek casually invariant representations by exploiting observational data collected under different setups [18, 19, 51–53]. Our work also falls into this category: we reveal the strengths and weaknesses of the invariant learning principle in the motion context and propose to tightly integrate invariant representation with structural architectural design based on domain knowledge.

Distributional Shifts. Previous methods tackle the challenge of distributional shifts from three main paradigms: domain generalization, domain adaptation, and transfer learning. Domain generalization is the most ambitious one, which aims to learn models that can directly function well in related but unseen test distributions [54, 55]. Recent literature has proposed a variety of solutions, such as distributionally robust optimization [56–58], adversarial data augmentation [59, 60]. Yet, these techniques often rely on strong assumptions on the test distribution, which may not hold in practice.

Domain adaptation is another popular approach that relaxes these assumptions by allowing a learning algorithm to observe a set of unlabelled test samples. Modern methods of this kind typically attempt to learn an embedding space where the training and test samples are subject to similar feature distributions through divergence minimization [61–64] or adversarial training [65, 66]. While this approach has been shown effective in a variety of supervised tasks [67–69], it is not well suited for motion forecasting where labels in the form of future trajectories are fairly easy to acquire without human annotation but sample efficiency matters crucially.

Previous work in the third category – transfer learning given limited data – often leverages special architecture designs, *e.g.*, external memory [70, 71], or transfer-oriented objectives, *e.g.*, meta-learning [72]. Some of these techniques have also been applied to motion forecasting [73, 74]. Our work differs from them in that we adopt a causal approach and propose a unified learning framework that facilitates both robust generation and fast adaptation to common types of distributional shifts in motion forecasting.

B Experiments

We evaluate our proposed method on two types of forecasting models (recurrent STGAT [21] and feedforward PECNet variant [75]) under distributional shifts of spurious features or style confounders. In the considered forecasting task, a model processes the past 8 time steps (3.2 seconds) of human trajectories in the scene to then predict their future movements in the following 12 (4.8 seconds) time steps. Identical to many prior works [3, 4, 20], we evaluate forecasting models on two metrics:

- *Average Displacement Error* (ADE): the average Euclidean distance between the predicted output and the ground truth over all predicted time steps.
- *Final Displacement Error* (FDE): the Euclidean distance between the predicted final destination and the true final destination at the end of the prediction horizon.

We evaluate each method over five experiments with different random seeds. More implementation details are summarized in Appendix E.

B.1 Spurious Shifts

We first evaluate the robustness of the forecasting model trained by our invariant loss under different ranges of spurious shifts. In particular, we compare our method against the two following baselines:

- *Vanilla ERM*: the conventional learning method that minimizes the average prediction error on all training samples (Eq. 1);
- *Counterfactual Analysis* [16]: a causality-inspired trajectory forecasting method that estimates and subtracts biased features through counterfactual interventions.

For a fair comparison with the recent counterfactual approach [16], we implement our method based on the same open-sourced code. Specifically, we follow their choice of the base model, *i.e.*, STGAT [21], in our experiments. The encoder of the STGAT contains two LSTMs and one graph attention network (GAT) to account for the historical trajectory and social interaction clues, while the decoder is modeled by an LSTM to rollout the future trajectory.

Setup The original ETH-UCY dataset contains five subsets collected at different locations [76, 77]. While recent work [16] has highlighted the intrinsic differences between these subsets, it is still not trivial to pinpoint the detailed biases in each environment. To clearly examine the robustness

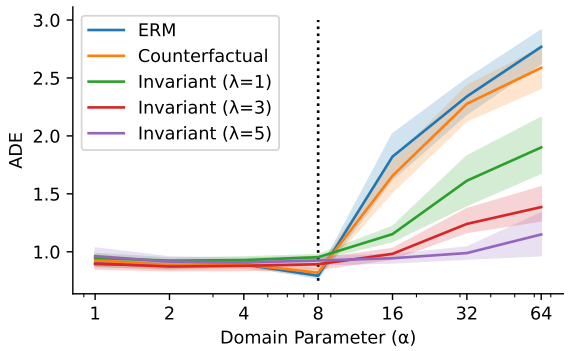


Figure 3: Comparison of different methods on the ETH-UCY dataset with controlled spurious correlations. Our invariant learning approach substantially outperforms the conventional ERM and the counterfactual approach [16] in the out-of-distribution regime $\alpha \in (8, 64]$, while being on par within the training domains.

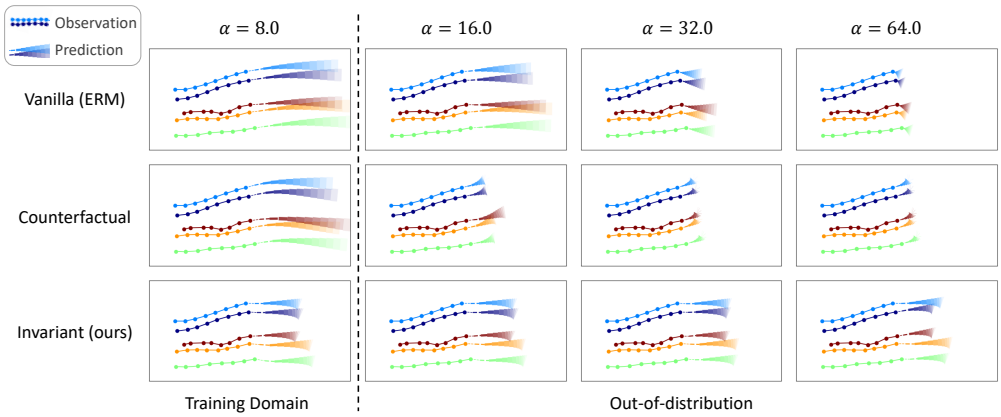


Figure 4: Visualization of the predicted trajectories from different methods in a particular test case of the ETH-UCY dataset with controlled spurious features. Despite the *same* past trajectory observation and ground truth future, the predicted trajectories from the two baselines abruptly slow down, when the strength of the spurious feature at test time is larger than that in the training domains. In comparison, our invariant learning approach results in much more robust solutions, even under substantial spurious shifts (*e.g.*, $\alpha = 64.0$).

of a motion forecasting model against non-causal biases, we modified the ETH-UCY dataset by introducing a third input variable measuring the level of observation noises, given that variations of spurious noise often occur in real-world problems [18]. Specifically, for each agent at time step t, we assign the observational uncertainty $\boldsymbol{\sigma}_t$ based on a linear function of the local trajectory curvature:

$$\begin{aligned} \boldsymbol{c}_t &:= (\dot{x}_{t+\delta t} - \dot{x}_t)^2 + (\dot{y}_{t+\delta t} - \dot{y}_t)^2, \\ \boldsymbol{\sigma}_t &:= \boldsymbol{\alpha} \cdot (\boldsymbol{c}_t + 1), \end{aligned} \tag{6}$$

where $\dot{x}_t = x_{t+1} - x_t$ and $\dot{y}_t = y_{t+1} - y_t$ reflect the velocity of the agent within the temporal window of length $\delta t = 8$, and $\boldsymbol{\alpha}$ is a domain specific parameter to control the strength of spurious features. We train the model in four environments ('hotel','univ','zara1' and 'zara2') with $\alpha \in \{1, 2, 4, 8\}$ and test it on the remaining one ('eth') with $\alpha \in \{1, 2, 4, 8, 16, 32, 64\}$.

Results In Figure 3 we show the prediction accuracy on the test sets resulting from different learning methods. All the methods perform strongly in the training domains, *i.e.*, $\alpha \in [1, 8]$. However, in the out-of-distribution regime, the accuracies of the two baseline methods significantly drop with an increased value of the domain parameter. Notably, when the strength of the spurious feature is 8 times of the maximum strength seen during training, the ADE of the vanilla ERM rises to 3.0, approximately three times worse than its performance in the training domains. While the counterfactual approach [16] is slightly better than the vanilla ERM, it also suffer suffers from a large ADE at $\sim$2.5. In comparison, the forecasting models trained by our invariant method are clearly less sensitive to the changes of the domain parameter. It is also visually distinct that a large emphasis (λ in Eq. 3) on the invariant penalty term during training leads to a more robust model under spurious shifts.

Table 1: Quantitative comparison of different methods under style shifts. The performance is measured by ADE (lower is better) over 5 seeds. Both the vanilla baseline and our invariant approach alone suffer from large errors, since they either *average* the domain-varying styles or *ignore* them. Our modular network incorporates *distinctive* style features into prediction and hence yields much better results. In particular, enforcing the causal invariance of the first encoder ψ leads to the best OOD robustness, while being highly competitive in the training environments.

Method	IID	OOD-Inter	OOD-Extra
Vanilla (ERM)	0.113 ± 0.004	0.112 ± 0.003	0.192 ± 0.013
Invariant (ours)	0.115 ± 0.005	0.114 ± 0.004	0.191 ± 0.007
Modular (ours)	$\mathbf{0.063 \pm 0.005}$	0.070 ± 0.006	0.112 ± 0.004
Inv + Mod (ours)	0.065 ± 0.007	$\mathbf{0.069 \pm 0.007}$	$\mathbf{0.107 \pm 0.007}$

In Figure 4 we visualize the qualitative results on a particular test example of the augmented ETH-UCY dataset. While the input trajectories remain the same across all domains, a growing strength of the spurious feature causes a dramatic shrinkage of the predicted trajectories from the baseline methods. In contrast, the outputs of our method stay almost constant under spurious shifts.

B.2 Style Shifts

We further evaluate the forecasting models trained by our method in the presence of style shifts. As elaborated in Sec 2.4, it is often impractical for the model to directly generalize to new styles. We therefore consider two different scenarios: robustness in the zero-shot and transfer learning results in the low-shot setting.

Setup The motion styles of existing real-world data are often largely unknown. We thus create some synthetic trajectories using ORCA [2], a popular multi-agent simulator, in the circle-crossing scenarios [78] with varied style parameters. Specifically, we consider four training styles where the simulated agents keep different minimum separation distances from each other, *i.e.*, $\{0.1, 0.3, 0.5\}$ meters. For each training domain, we generate 10,000 trajectories for training, 3,000 trajectories for validation and 5,000 trajectories for test. We evaluate each model on the training environments (IID) as well as the two new test environments with the minimum separation distance of 0.4 (OOD-Inter) and 0.6 (OOD-Extra). We use a variant of PECNet [75] as our base model which employs a MLP as the basic building block for the encoders and decoder in our modular design. More implementation details are reported in Appendix E.

B.2.1 Out-of-distribution Generalization

Results In Table 1 we report the results of different forecasting models in both the training domain and the out-of-distribution regimes. Similar to Sec B.1, the vanilla baseline suffers from much larger prediction errors in the OOD test sets than in the training ones. Given style changes, our invariant method alone does not yield clear advantages neither, as it tends to ignore the domain-specific style confounder. In contrast, our modular architecture design allows the model to effectively incorporate the domain-specific style features and thus achieves superior performance in all environments. In particular, training the first encoder ϕ in our modular network with the invariant loss results in the best robustness in the OOD regime while being competitive in the training domains. It is also evident that there remains a clear performance gap between the IID and OOD-Extra domain, which suggests the importance of building adaptive models investigated next.

B.2.2 Low-shot Transfer

As shown above, it is practically unrealistic for a forecasting model to directly generalize to all kinds of distributional shifts. We next evaluate the effectiveness of our proposed adaptation method in the context of low-shot transfer. We again consider the challenging OOD-Extra style shift scenario and compare the following options:

(a) conventional approach fine-tuning all parameters;
(b) our modular adaptation strategy fine-tuning f only;
(c) our test-time refinement on top of our method (b).

We evaluate all methods given a limited number of samples, *i.e.*, $\{1, 2, \ldots, 6\} \times \mathrm{BS}$, where $\mathrm{BS} = 64$ is the batch size.

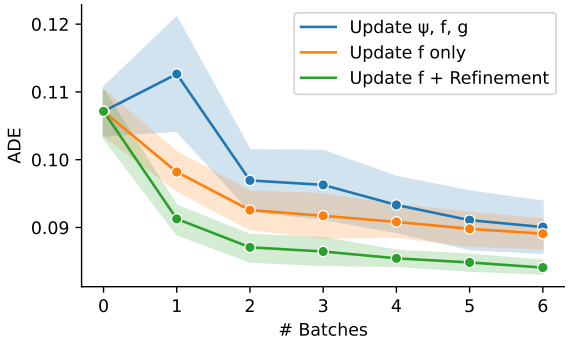


Figure 5: Quantitative results of different methods for transfer learning to a new motion style, given limited batch of samples. Our modular adaptation strategy (updating the style modulator f) yields higher sample efficiency than the conventional counterpart in the low-data regime. Moreover, refining the predicted output for 3 iterations further reduces the prediction error on the fly.

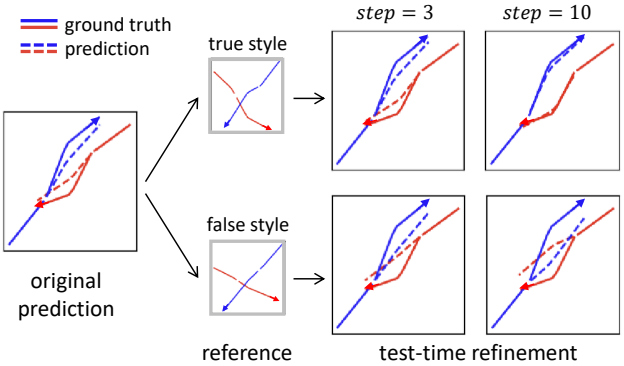


Figure 6: Qualitative effects of test-time refinement in a two-agent scenario. The initial predicted output suffers from a clear prediction error. Given a scene observation of the true style (large separation distance) as a reference, our method gradually closes the discrepancy between the predicted trajectory and the ground truth. Conversely, when conditioned on a scene of a different style (small separation distance) as the reference, our method manages to drive the output towards the corresponding false style as well.

Effect of Modular Adaptation. Figure 5 shows the results of different adaptation methods in the low-shot setting. In the case of only one batch of observations, fine-tuning all style-related parameters (ψ, f and g) leads to noisy outcomes and worsens the results on average. In comparison, updating f while keeping the remaining majority of the parameters fixed yields clearly better performance in the low-data regime. For instance, fine-tuning f on 2 batches of the new style achieves the same level of prediction accuracy as fine-tuning the whole model on 5 batches.

Effect of Test-time Refinement Finally, we evaluate the effectiveness of test-time refinement based on the style consistency loss. As shown in Figure 5, our refinement techniques leads to substantial error reductions on top of the fine-tuned models. Figure 6 shows the qualitative effect on a two-agent scenario, where the predicted trajectories gradually get closer to the ground truth based on a scene observation of the target style as a reference. This result suggests a strong promise of reusing the structural knowledge learned in our modular forecasting model at test-time.

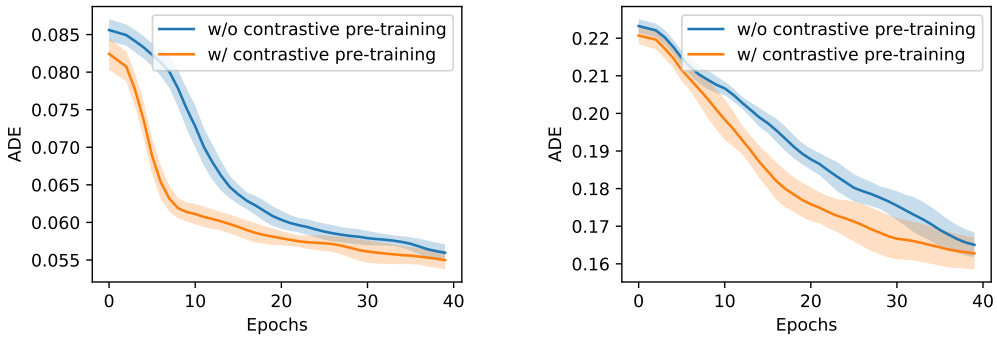
Additional Results and Discussions Please refer to Appendix D for additional ablation results and Appendix F for detailed discussions about limitations and societal impact.

C Conclusion

In this work, we present a causality-inspired learning method for motion forecasting. Given observational data collected from multiple locations, our invariant loss yields superior generalization ability over the previous statistical and counterfactual methods in the presence of spurious features. In addition, our modular architecture design coupled with the proposed style consistency loss greatly enhances the robustness and transferability of the learned forecasting model under style shifts. Our

Table 2: ADE scores of different methods on OOD-Extra domains. The full version (invariant + modular) of our method yields more performance gains with an increasing degree of style shifts.

Method	$d = 0.6$	$d = 0.7$	$d = 0.8$
Vanilla (ERM)	0.192 ± 0.013	0.246 ± 0.020	0.309 ± 0.025
Invariant (ours)	0.191 ± 0.007	0.245 ± 0.009	0.309 ± 0.011
Modular (ours)	0.112 ± 0.004	0.169 ± 0.011	0.242 ± 0.020
Inv + Mod (ours)	**0.107** ± 0.007	**0.156** ± 0.013	**0.221** ± 0.020



(a) IID style $d = 0.1$ (b) OOD style $d = 0.7$

Figure 7: Comparison of the models with and without style contrastive pre-training. The model pre-trained on the style contrastive task converges faster than the counterpart during the end-to-end training in both domains.

results suggest that incorporating causal invariance and structure into representation learning can be a promising direction towards robust and adaptive motion forecasting.

D Additional Experiments

Larger Style Shifts. As a supplement to Table 1, we summarize the detailed results under larger style shifts in Table 2. Among these OOD test domains ($d > 0.5$), the farther the style parameter is from the training ones, the larger improvement we obtain from using the full version of our method. This result confirms the advantage of our modular design with an enforced structure of the invariant and style knowledge for robust generalization.

Effect of Style Contrastive Pre-training. As described in Sec 2.4, one advantage of incorporating the proposed style contrastive loss is to ease the training of our modular model that consists of multiple sub-networks. In Figure 7 we compare the performance of the models during training with and without the style contrastive pre-training. The model pre-trained on the style contrastive task learns significantly faster than the counterpart without it.

E Implementation Details

E.1 Spurious Shift Experiments

Architecture. For the experiments reported in Sec B.1, we use the standard STGAT [21] architecture for a fair comparison with the counterfactual analysis approach [16]. In order to take the x and y coordinate as well as the observational uncertainty $\boldsymbol{c}_t$ as inputs, we adjust the input dimension of the first LSTM module to three. All the remaining configurations align with the original STGAT. Following the previous work [16], we train the model in three steps: (i) pre-train the first LSTM, (ii) pre-train the GAT together with the second LSTM, (iii) train the whole model.

Hyper-parameters. We use the same hyper-parameters as in the original STGAT [21]. For the invariant penalty coefficient λ, we run grid search in a range from 0.001 to 100, and report the results of 1, 3 and 5 in Figure 3. Since the focus of our experiments is on the robustness and adaptability under distributional shifts rather than the performance on training domains, we only predict one trajectory output per instance instead of multiple ones [20] during training, which reduces computational expenses for the comparison of different methods. Other detailed hyper-parameters are summarized in Table 3.

Table 3: Hyper-parameters in spurious shift experiments.

config	value
batch size	64
epochs per stage	150, 100, 150
learning rate	0.001

Table 4: Hyper-parameters in style shift experiments.

config	value
batch size	64
epochs per stage	100, 50, 20, 300
contrastive loss coefficient	1.0
learning rate baseline	0.001
learning rate style encoder	0.0005
learning rate projection head	0.01
learning rate style modulator (train)	0.01
learning rate style modulator (adapt)	0.001

E.2 Style Shift Experiments

Architecture. For the experiments in Sec B.2, we use a PECNet-like [75] feedforward network as our base model. Specifically, we model all components using MLPs. We train the modular network in four detailed steps: (i) train the invariant encoder together with the decoder, (ii) subsequently pre-train the style encoder and the projection head, (iii) followed by the style modulator, and (iv) finally train the entire model end-to-end.

Hyper-parameters. We keep most hyper-parameters identical to the setup in Sec E.1. We further tune the learning rates for each module separately due to their distinct properties. Detailed settings for training, adaptation and refinement are summarized in Table 4.

E.3 Other Details

We train all of our models on a single NVIDIA Tesla V100 GPU. Each run takes around one hour. The source code of our method as well as baselines can be found at `https://github.com/vita-epfl/causalmotion`.

F Additional Discussions

To the best of our knowledge, our work provides the first attempt to incorporate causal invariance and structure into the design and learning of motion forecasting models. Despite encouraging results, our work is still subject to a couple of limitations.

Limitations & Future Work. One major technical limitation lies in the granularity of the considered causal representations. While our method places great emphasis on three prominent groups of high-level latent features, we have largely overlooked the structure of fine-grained features. One interesting direction for future work is to further exploit detailed causal structure for motion forecasting, for instance, (i) disentangling the left or right-hand traffic rules from social distance conventions within the group of style confounders, (ii) encouraging sparse interplay between sub-modules, *e.g.*, pruning the connections between inertia features and left or right-hand traffic rules, given their presumably minute significance.

Another limitation of our work is tied to the scale and diversity of experiments. Thus far, we have demonstrated the strengths of our method on two human motion datasets and two base models as proofs of concept. Nevertheless, our method is highly generic and we hypothesize that it can also bring similar benefits to other types of motion problems and datasets, *e.g.*, vehicles [79], sports [80] and driving simulations [81]. Extending the current empirical findings to more contexts can be another valuable avenue for future work.

Societal Impact. Out-of-distribution robustness remains a salient weakness of motion forecasting models while having a crucial impact on the safety of autonomous systems, especially in the context of autonomous driving. Even though these machines operate accurately in their training environments, deploying them in unseen test conditions can result in undesired behavior, which may ultimately lead to fatal consequences in specific scenarios. With our work, we contribute to reducing this

performance gap. However, we are aware of the remaining deficits of our motion forecasting approach in significantly changing conditions that should not be neglected when utilizing such systems in real-world applications.