Explainable Causal Representation Learning for Drone Navigation using Causal Graphs and Interventions

Aadya (2021370), Akanksha Singal (2021008)

Introduction

Causal representation learning is an emerging field at the intersection of machine learning and causal inference. It aims to uncover the underlying causal structures from data, enabling models to generalize beyond correlations and make robust, interpretable decisions. Traditional machine learning models, including deep learning, primarily rely on statistical correlations in data. However, correlation does not imply causation. Causal representation learning seeks to bridge this gap by discovering high-level causal variables and their relationships from raw observational data.

Causal theory, rooted in statistical and philosophical foundations, provides the framework for understanding cause-and-effect relationships. Structural Causal Model (SCM) framework uses directed acyclic graphs (DAGs) to represent causal structures. In this framework, each node represents a variable, and directed edges encode causal dependencies. The three fundamental principles of causal reasoning—intervention, counterfactual reasoning, and causal discovery—allow researchers to not only predict outcomes but also understand the effect of interventions and answer "what-if" questions (Narendra et al. 2018).

Causal representation learning involves extracting meaningful representations that align with the true causal generative processes of the data. Unlike conventional representation learning, which focuses on maximizing predictive performance, causal representation learning seeks representations that remain invariant under interventions, domain shifts, or distributional changes. This property, known as causal invariance, is crucial for out-of-distribution generalization, robust AI, and transfer learning.

Papers

Causal VAEs (CVPR 2021) The paper "CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models" (Yang et al. 2020) introduces a novel framework that integrates causal structure learning within Variational Autoencoders (VAEs) to achieve disentangled representation learning. Traditional approaches to disentanglement assume that latent factors are independent, but real-world data often exhibits complex causal relationships among variables. To address this, the authors propose CausalVAE, which incorporates a SCM layer that transforms independent exogenous factors into causally meaningful representations. This transformation enables the model to learn structured latent variables aligned with the true causal generative process of the data.

One of the key innovations of CausalVAE is the introduction of a Causal Layer, which enforces causal dependencies through a DAG, and a Mask Layer, which propagates causal effects from parent to child variables. This structure allows the model to perform interventions and generate counterfactual samples.

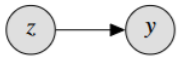

Figure 1: z causes y or y is the effect of z

Deep Reinforcement Learning based Multi-UAV Collision Avoidance with Causal Representation Learning (IEEE Big Data and Information Analytics BigDIA 2024) This paper (Han et al. 2024) improves drone navigation by making it smarter at avoiding obstacles in unseen environments using causal learning, reducing the risk of crashes when faced with new challenges. The researchers introduce Causal Representation Learning (CRL) to help the drone focus only on the key factors affecting navigation. Instead of memorizing obstacles, it learns the underlying cause-effect relationships (i.e., "An obstacle is something I need to avoid, no matter what it looks like"). Their approach outperformed previous methods, showing higher success rates, better efficiency, and improved generalization to new obstacles. This shows how we can use causal learning to develop models having higher success rates, better efficiency, and improved generalization.

Datasets

We are using our novel fly-to-target dataset using a stable IBVS-based controller similar to the one described in (Kumar et al. 2024) with expert demonstrations of flight trajectories. We consider scenarios where the target is static,

we use a proportional IBVS controller with feature error feedback to generate the control commands. To generate high-fidelity expert demonstrations, we consider the case of a quadrotor visual servoing to a static ground rover in a robot operating system (ROS) (Quigley et al. 2009) integrated simulator, GAZEBO (Koenig and Howard 2004). We create an environment consisting of a hector quadrotor with a camera (Meyer et al. 2012) and AprilTag-equipped husky ground rover (Clearpath Robotics).

To construct the dataset, we log complete flight trajectories, capturing sequences of RGB images alongside the corresponding control commands with a static target. For each trajectory, we also store the desired image that is recorded when the quadrotor and target are in the ideal geometric alignment in both position and orientation. This desired image is then subtracted from each frame in the recorded sequence to generate a difference image, as shown in Fig. 2. The difference images serve to highlight feature discrepancies between the current and desired views, enabling the network to learn meaningful visual cues for control. Each difference image of the flight trajectory, paired with the corresponding velocity reference commands $V_c = [v_x^b, v_y^b, v_z, \dot{\psi}]^T$, forms an input-output sample in the dataset as shown in Fig. 3. The recorded velocity reference commands include quadrotor's: (1) body-fixed linear velocities in the X and Y directions (v_x^b, v_y^b), (2) inertial-frame vertical velocity (v_z), and (3) heading rate $(\dot{\psi})$. Note that $\bar{v}$ and $\dot{\psi}$ represent the translational velocity and the heading rate of the virtual camera frame, also expressed in the virtual image frame. However, for the gazebo simulator, we have transformed these virtual frame velocities to V_c, a mixture of body fixed and inertial frame velocities. This structured data collection approach ensures a well-defined mapping between visual inputs and control outputs, facilitating effective learning for marker-free IBVS.

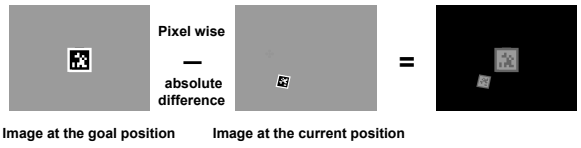

Figure 2: Absolute pixel-wise error between the image at the desired pose and the current image captured by the quadrotor's camera

Methodology

To enhance the explainability and interpretability of deep learning models in drone landing task, we propose a structured causal approach leveraging feature extraction, latent factor analysis, causal graph learning, and intervention-based explanations. Our methodology consists of the following key steps:

1. **Feature Extraction** We begin by extracting meaningful features from raw drone sensor data, focusing on orientation, corners, and area. These features capture essential geometric and positional information relevant to the drone's landing process on the goal (ground rover).

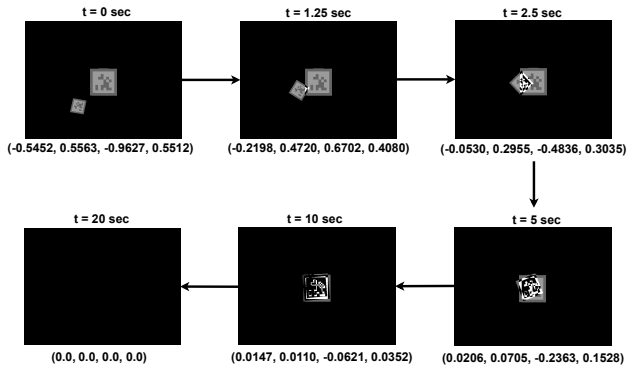

Figure 3: Evolution of the input sequence to the model for target equipped with a marker along with the velocity reference commands from IBVS controller

2. **Learning Latent Representations** Using an encoder network, we transform our raw sensor data into a lower-dimensional latent space. This latent space generates an embedding of the input data. The encoder is trained using our proposed encoder, given in Table 1 and 2.

Table 1: Proposed Encoder Block

Layer (Type)	Activation	Param #
InputLayer	-	0
Rescaling	-	0
Normalization	-	7
Conv2D	ReLU	1,824
Conv2D	ReLU	21,636
Conv2D	ReLU	43,248
Conv2D	ReLU	27,712
Conv2D	ReLU	9,232
Flatten	-	0
Dense, 128 units	Linear	159,872

3. **Causal Graph Learning** We employ a Causal Layer, as defined in Figure 4, which takes in the input embeddings, to discover and model the causal relationships among the latent factors. This step constructs a DAG, which represents the underlying causal dependencies between the latent factors. The causal structure is learned using SCMs that enforce acyclicity constraints (i.e., causal relationships).

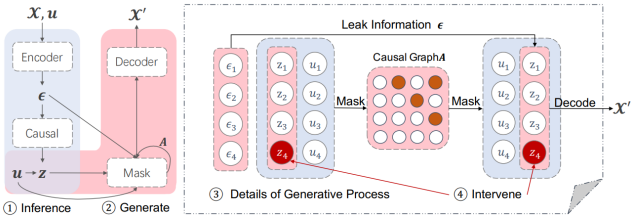

Figure 4: Learning the Causal Graph (Yang et al. 2020)

Table 2: Layer-wise Parameters of the Proposed Encoder

Layer Type	Input Dim.	Filters	Kernel	Stride
2D Conv.	144×256×3	24	5×5	2
2D Conv.	70×126×24	36	5×5	2
2D Conv.	33×61×36	48	5×5	2
2D Conv.	15×29×48	64	3×3	1
2D Conv.	13×27×64	16	3×3	2
Flatten	N/A	N/A	N/A	N/A
Fully Conn.	1248	N/A	N/A	N/A

4. **Mapping Latent Factors to Real-World Features** To improve interpretability, we map the learned latent factors back to real-world features (orientation, corners, and area), ensuring that each latent dimension has a meaningful semantic representation. This mapping is done using intervention and helps bridge the gap between abstract model representations and practical drone landing dynamics.
5. **Learning a Structural Causal Model-Based Neural Network (SCM-NN)** Once the causal relationships are established, we construct an SCM-based neural network to predict the success of the drone's landing on the ground rover. The SCM-NN incorporates the learned causal relationships, ensuring that predictions are robust to spurious correlations and align with domain knowledge. This step ensures that our model is not merely data-driven but also causally grounded.
6. **Intervention-Based Explainability** To validate and explain our model, we perform causal interventions on selected features and analyze their impact on landing success. These interventions help answer counterfactual questions such as, *"What happens if the drone's orientation changes while keeping other factors constant?"* By analyzing changes in landing success due to controlled interventions, we can derive meaningful insights into drone behavior and improve model transparency.

Explainability and Interpretability

The proposed approach significantly enhances the explainability and interpretability of deep learning models in multiple ways:

1. Latent Space to Real-World Features: By mapping latent representations to physical drone features (orientation, corners, and area), we ensure that the model's learned factors have real-world semantic meaning.
2. Causal Graph Representation: The learned DAG provides an explicit structure that outlines dependencies among features, helping to understand how different factors depend on each other.
3. Intervention-Based Insights: By actively modifying specific features and observing their effects on landing predictions, we can provide counterfactual explanations, making it clear how and why changes in inputs affect outcomes.

References

[Clearpath Robotics] Clearpath Robotics. Husky.

[Han et al. 2024] Han, G.; Wu, Q.; Wang, B.; Lin, C.; Zhuang, J.; Li, W.; Hao, Z.; and Fan, Z. 2024. Deep reinforcement learning based multi-uav collision avoidance with causal representation learning. In *2024 10th International Conference on Big Data and Information Analytics (BigDIA)*, 833–839. IEEE.

[Koenig and Howard 2004] Koenig, N., and Howard, A. 2004. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, volume 3, 2149–2154 vol.3.

[Kumar et al. 2024] Kumar, Y.; Shamsi, B. P.; Roy, S. B.; and P B, S. 2024. Tracking a planar target using image-based visual servoing technique. *IEEE Transactions on Intelligent Vehicles* 1–11.

[Meyer et al. 2012] Meyer, J.; Sendobry, A.; Kohlbrecher, S.; Klingauf, U.; and von Stryk, O. 2012. Comprehensive simulation of quadrotor uavs using ros and gazebo. In Noda, I.; Ando, N.; Brugali, D.; and Kuffner, J. J., eds., *Simulation, Modeling, and Programming for Autonomous Robots*, 400–411. Berlin, Heidelberg: Springer Berlin Heidelberg.

[Narendra et al. 2018] Narendra, T.; Sankaran, A.; Vijaykeerthy, D.; and Mani, S. 2018. Explaining deep learning models using causal inference. *arXiv preprint arXiv:1811.04376*.

[Quigley et al. 2009] Quigley, M.; Conley, K.; Gerkey, B.; Faust, J.; Foote, T.; Leibs, J.; Wheeler, R.; Ng, A. Y.; et al. 2009. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, 5. Kobe, Japan.

[Yang et al. 2020] Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; and Wang, J. 2020. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*.