

Article

Improving Generalization in Collision Avoidance for Multiple Unmanned Aerial Vehicles via Causal Representation Learning [†]

Che Lin ^{1,‡} , Gaofer Han ^{1,‡}, Qingling Wu ¹, Boxi Wang ¹, Jiafan Zhuang ^{1,*} , Wenji Li ¹, Zhifeng Hao ¹ and Zhun Fan ^{2,*}

¹ Department of Electronic Engineering, Shantou University, Shantou 515063, China; 19clin1@stu.edu.cn (C.L.); 22gfhan@stu.edu.cn (G.H.); 22qlwu@stu.edu.cn (Q.W.); 19bxwang@stu.edu.cn (B.W.); liwj@stu.edu.cn (W.L.); haozhifeng@stu.edu.cn (Z.H.)

² Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518000, China

* Correspondence: jfzhuang@stu.edu.cn (J.Z.); fanzhun@uestc.edu.cn (Z.F.)

[†] This paper is an extension version of the conference paper: Han, G.; Wu, Q.; Wang, B.; Lin, C.; Zhuang, J.; Li, W. Deep Reinforcement Learning Based Multi-UAV Collision Avoidance with Causal Representation Learning. In Proceedings of the 2024 10th International Conference on Big Data and Information Analytics (BigDIA), Chiang Mai, Thailand, 25–28 October 2024.

[‡] These authors contributed equally to this work.

Abstract: Deep-reinforcement-learning-based multi-UAV collision avoidance and navigation methods have made significant progress. However, the fundamental challenge of those methods is their restricted capability to generalize beyond the specific scenario in which they are trained on. We find that the cause of the generalization failures is attributed to spurious correlation. To solve this generalization problem, we propose a causal representation learning method to identify the causal representations from images. Specifically, our method can neglect factors of variation that are irrelevant to the deep reinforcement learning task through causal intervention. Subsequently, the causal representations are fed into the policy network for action prediction. Extensive testing reveals that our proposed method exhibits better generalization results compared to state-of-the-art methods in different testing scenes.

Keywords: deep reinforcement learning; multi-UAV collision avoidance; generalization failure; causal representation learning; causal intervention



Academic Editor: Pasquale Daponte

Received: 7 April 2025

Revised: 20 May 2025

Accepted: 22 May 2025

Published: 24 May 2025

Citation: Lin, C.; Han, G.; Wu, Q.; Wang, B.; Zhuang, J.; Li, W.; Hao, Z.; Fan, Z. Improving Generalization in Collision Avoidance for Multiple Unmanned Aerial Vehicles via Causal Representation. *Sensors* **2025**, *25*, 3303. <https://doi.org/10.3390/s25113303>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The low-altitude economy, as a strategic emerging industry with broad prospects, is experiencing rapid development, with the application of unmanned aerial vehicles (UAVs) becoming increasingly widespread. In fields such as environmental monitoring [1], emergency rescue [2,3], intelligent logistics [4], and precision agriculture [5], UAVs have already demonstrated significant social and economic value. Collision avoidance and navigation capabilities are crucial in ensuring that UAVs operate safely and effectively. UAVs need to navigate around obstacles in intricate environments and find the best route from a starting point to a destination point.

Unmanned aerial vehicle collision avoidance and navigation have been widely studied. Conventional methods typically utilize simultaneous localization and mapping (SLAM) to build a local representation of the environment, which is then used for planning safe paths [6–8]. Another approach is path retracing [9], where UAVs follow routes that are either predefined or learned from demonstrations. Despite their success in structured

environments, these techniques often struggle to adapt to dynamic or unfamiliar scenarios. A key limitation is their reliance on pre-built maps and handcrafted features, which restricts their robustness, flexibility, and generalization in real-world applications.

To address the limitations of traditional methods, deep reinforcement learning (DRL) has emerged as a promising alternative [10–12]. DRL enables UAVs to autonomously learn effective navigation strategies in complex and dynamic environments, eliminating the need for manual rule design or parameter tuning. It works by extracting compact state representations from raw sensory inputs, selecting actions based on these states, and refining its policy or value estimates through feedback in the form of rewards or penalties from the environment. The ultimate goal is to learn an optimal strategy that maximizes cumulative long-term rewards and achieves the desired task performance.

Despite the potential of deep reinforcement learning (DRL), a key limitation lies in its poor generalization to scenarios beyond the training environment. As illustrated in Figure 1, consider a case where UAVs are trained in a simplified environment containing only UAV-like obstacles. In this setting, they learn to navigate from a starting point to a target location by avoiding these familiar obstacles. However, when deployed in a test environment with novel, previously unseen obstacles, the learned policy—while effective in the original training domain—shows a significant drop in performance. This example highlights a major challenge in current DRL-based navigation systems: their limited ability to generalize to environments that differ from the training distribution.

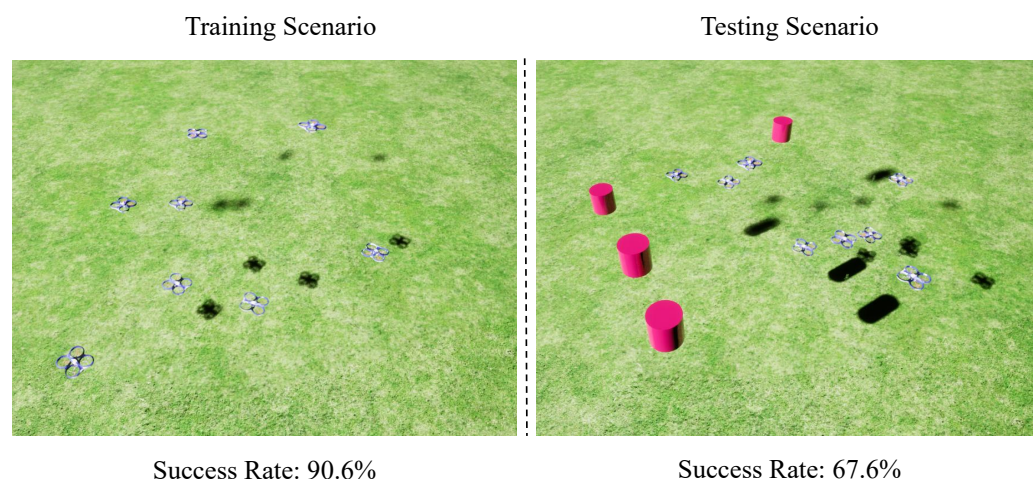


Figure 1. Illustration on the influence of spurious correlations. Due to the presence of spurious correlations, deep reinforcement learning algorithms often over fit to the specific obstacles encountered during training. As a result, these algorithms may achieve high success rates in the training scenario but develop a fragile policy that struggles to adapt to unseen obstacles in testing scenarios.

To reveal the reasons for the generalization failures in reinforcement learning, we review pioneering work in the field of multi-UAV collision avoidance based on deep reinforcement learning (i.e., SAC + RAE [13]). We have discovered that spurious correlations are the root of the generalization failures. Specifically, the DRL algorithms often overfit to the specific obstacles encountered during training. As a result, these algorithms may achieve a high success rate in the training scenario but develop a fragile policy that struggles to adapt to unseen obstacles in testing scenarios. Furthermore, learned policies often struggle to ignore non-causal factors (e.g., the obstacle shape) in their sensor observations and exhibit strong sensitivity to variations in these factors.

Recent studies also show that the generalization of DRL policies can be improved substantially by causal representation learning [14,15]. Causal representation learning [16] aims to find the high level causal variables from low-level observations. It can effectively

extract causal factors that impact the task, thereby reduce the effect of spurious correlations on the generalization performance of DRL. Therefore, we propose a causal representation learning (CRL) approach to identify the causal representations from images. Specifically, we only intervene the shape of obstacles within deep images, keeping the rest content of the image unchanged. Then, the images of spherical obstacles and cubic obstacles are used solely as auxiliary task for causal representation learning. By utilizing its auxiliary task, CRL enhances its generalization ability by maximizing the mutual information across the latent representations of different obstacle images. In addition, we apply supervision signals to the latent representations to reduce redundancy between dimensions. Intuitively, this encourages different dimensions to capture distinct information. After that, CRL can effectively identify and discard irrelevant variation factors, significantly enhancing the model's generalization capability.

To evaluate the generalization ability of CRL, we conduct several testing scenarios across different obstacles. In comparison with previous state-of-the-art (SOTA) methods, the results demonstrate that CRL outperforms the SOTA across different testing scenarios and demonstrate the efficacy of our CRL method. It is worth noting that an extended version of this paper was presented at the 2024 10th International Conference on Big Data and Information Analytics (BigDIA), Chiang Mai, Thailand, 25–28 October 2024 [17].

2. Related Work

2.1. Drl-Based UAV Collision Avoidance and Navigation

Deep reinforcement learning (DRL) offers a powerful framework for enabling autonomous obstacle avoidance and navigation in unmanned aerial vehicles (UAVs), particularly within complex and dynamic environments [18,19]. Foundational approaches often utilize Deep Q-Networks (DQNs) [20]; however, standard DQNs can encounter challenges related to training stability and convergence. To address these limitations, enhanced variants such as Double DQN [21,22] and Dueling DQN have been developed, demonstrating improved performance. Alternatively, policy gradient methods, notably Proximal Policy Optimization (PPO) [23,24], are frequently favored due to their relative stability and sample efficiency, making them robust choices for training reliable policies. For tasks involving the continuous control spaces typical of UAV flight, actor–critic architectures are particularly suitable. Algorithms like Deep Deterministic Policy Gradient (DDPG) [25] and Soft Actor–Critic (SAC) [13] effectively manage continuous actions by balancing policy exploitation with sufficient exploration. Furthermore, overcoming challenges in training complexity and bridging the simulation-to-reality gap often involves employing techniques such as curriculum learning, transfer learning, and sim-to-real strategies like domain randomization. Recognizing that real-world UAV operations often involve incomplete state information, we formulate the obstacle avoidance navigation task as a partially observable Markov decision process (POMDP). However, owing to the limitations in feature representation, the above methods are unable to generalize beyond the specific scenario in which they are trained on. Therefore, the collision avoidance navigation methods based on DRL face substantial challenges before they can be widely adopted and applied in practice.

2.2. Causal Representation Learning

Our work builds on the nascent field of causal representation learning [16], which is a method of modeling the causal relationships within data. It can reduce the effect of spurious correlations and transform the data into a structured representation that aligns with physical laws. Yang et al. [26] propose a method called CausalVAE, which is the first to introduce structural causal models into representation learning. By learning the causal structure of the data, CausalVAE can generate counterfactual data that differs from the

training data, which helps improve the model's generalization ability. However, CausalVAE introduces additional causal layers and structures, which increases the model's complexity and may make the training and optimization process more challenging. To address the problem of learning causal representation from multiple distributions, Zhang et al. [27] propose applying sparsity constraints to the latent variable graph structure. By learning the causal relationships of latent variables, it can reveal the underlying causal structure of the data to understand system behavior and make decisions; while the method can learn causal relationships, interpreting these relationships requires domain knowledge, especially when dealing with complex systems and multiple latent variables. To address the sparsity of supervisory signals and the long-tail problem in causal representation learning, Zhao et al. [28] propose a general framework called DCCL, which learns disentangled causal embeddings through contrastive learning based on causal mechanisms and causal graphs. This approach can notably improve the precision and robustness of model predictions. Although DCCL enhances the handling of sparse data through contrastive learning, its performance still relies to some extent on the quality and diversity of the training data.

In this work, we conduct causal intervention on the shape of obstacles to extract invariant causal representations inspired by the above studies. This method can effectively reduce the effect of spurious correlations on the generalization ability of DRL.

3. Approach

3.1. Definitions of Some Key Notations

In order to present our method more clearly, we define some key notations in this subsection. The definition results are shown in Table 1.

Table 1. Definitions of some key notations.

Notations	Definitions
S	State space in POMDP
A	Action space in POMDP
$P(s' s, a)$	Transition probability function in POMDP
$R(s, a)$	Reward function in POMDP
Ω	Observation space in POMDP
$O(o s)$	Observation function in POMDP
$\mathcal{O}$	UAV's observation space, $\mathcal{O} = [I, V, G]$
I	Accumulation of four consecutive depth images (part of $\mathcal{O}$)
V	Current velocity of the UAV (part of $\mathcal{O}$)
G	Euclidean distance to the target destination (part of $\mathcal{O}$)
a	Action vector, $a = [v_x^{\text{cmd}}, v_z^{\text{cmd}}, v_\omega^{\text{cmd}}]$
v_x^{cmd}	Commanded forward velocity
v_z^{cmd}	Commanded lateral velocity
v_ω^{cmd}	Commanded angular velocity
I_1	Depth images of UAVs and original obstacles
I_2	Depth images of UAVs and obstacles with modified shapes (after intervention)
z_1, z_2	Normalized latent representations from I_1, I_2 , respectively,
$\mathcal{L}_{\text{invariance}}$	Invariance loss, $\ z_1 - z_2\ ^2$
$\mathcal{L}_{\text{decorrelation}}$	Decorrelation loss to ensure different dimensions capture different information
$\mathcal{L}_{\text{rec}}$	Reconstruction loss for the autoencoder

3.2. Problem Formulation

3.2.1. Survey of Deep Reinforcement Learning in UAV Collision Avoidance

Deep reinforcement learning (DRL) provides an advanced methodological foundation for facilitating collision-free navigation of unmanned aerial vehicles (UAVs) in complex

environments. In this scenario, the UAV engages with the environment by executing actions informed by its real-time observations and subsequently receives feedback in the form of rewards or penalties, reflective of its operational efficacy. The primary goal is to maximize the cumulative reward over time, which drives the UAV to develop an optimal policy for navigation and collision avoidance.

Mathematically, the UAV collision avoidance problem is formalized as a partially observable Markov decision process (POMDP), which is defined by a 6-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \Omega, \mathcal{O})$, where:

- $\mathcal{S}$ denotes the state space, representing all possible configurations of the UAV and its environment.
- $\mathcal{A}$ represents the action space, consisting of all feasible actions that the UAV can execute.
- $\mathcal{P}$ is the transition probability function, $P(s'|s, a)$, which specifies the probability of transitioning to state s' given the current state s and action a .
- $\mathcal{R}$ is the reward function, $R(s, a)$, which assigns a scalar reward for taking action a in state s .
- Ω is the observation space, comprising all possible observations that the UAV can make.
- $\mathcal{O}$ is the observation function, $O(o|s)$, which provides the probability of observing o given the true state s .

The UAV's objective is to learn a policy $\pi(a|o)$ that maps observations to actions, with the goal of maximizing the expected cumulative reward.

3.2.2. Observation Space

The observation space $\mathcal{O}$ of the UAV is defined as $\mathcal{O} = [I, V, G]$, where:

- I represents the accumulation of four consecutive depth images captured by the UAV's onboard camera. These images provide spatial information about the environment and obstacles.
- V denotes the current velocity of the UAV, which includes its forward, turning, and climbing velocities.
- G represents the Euclidean distance from the UAV's current position to the target destination. This information helps the UAV navigate towards the goal.

3.2.3. Action Space

To ensure smoother and more controllable UAV movement, the action space is defined as a set of constrained velocities in continuous space. The action vector $a = [v_x^{\text{cmd}}, v_z^{\text{cmd}}, v_\omega^{\text{cmd}}]$, generated by the policy network $\pi(s)$, consists of two linear velocities and one angular velocity. Specifically, we constrain the forward velocity to $v_x^{\text{cmd}} \in (0.0, 2.0)$ m/s, the lateral velocity to $v_z^{\text{cmd}} \in (-0.5, 0.5)$ m/s, and the angular velocity to $v_\omega^{\text{cmd}} \in (-0.5, 0.5)$ rad/s. Notably, backward motion ($v_x^{\text{cmd}} < 0$ m/s) is disallowed, as each UAV is equipped with a front-facing camera only, which lacks rear visibility and cannot assist in avoiding collisions during backward movement.

3.2.4. Reward Function

The purpose of collision avoidance navigation is to control the UAVs to reach the target position in a complex environment without collision. This process can be divided into two subtasks: target approach r_g and obstacle avoidance r_c . Therefore, when the UAVs safely reaches the target positions, it should be given positive feedback as a reward. Conversely, when a collision occurs, negative feedback should be given as a punishment.

$$r = r_g + r_c \quad (1)$$

$$r_g = \begin{cases} r_{arrival} & \text{if } d_t < 0.5 \\ \alpha_{goal} \cdot (d_{t-1} - d_t) & \text{otherwise} \end{cases} \quad (2)$$

where d_t is the euclidean distance at time t . $r_{arrival}$ is the reward for UAVs that have reached the target positions. Furthermore, α_{goal} is the reward weight.

$$r_c = \begin{cases} r_{collision} & \text{if crash} \\ \alpha_{avoid} \cdot \max(d_{safe} - d_{min}, 0) & \text{otherwise} \end{cases} \quad (3)$$

where $r_{collision}$ is the collision penalty. α_{avoid} is the penalty weight. The d_{safe} is the safe distance of UAVs. The d_{min} is the minimum distance in a depth image.

3.2.5. Structural Composition of UAV Control in AirSim

The experimental platform utilizes AirSim's component-based architecture for quadrotor control, featuring decoupled subsystems that enable efficient motion planning [29]. As shown in Figure 2, the policy network $\pi(s)$ produces three-dimensional motion commands $a = [v_x^{cmd}, v_z^{cmd}, \omega^{cmd}]$, with each component being physically realized through rotor-generated aerodynamic forces.

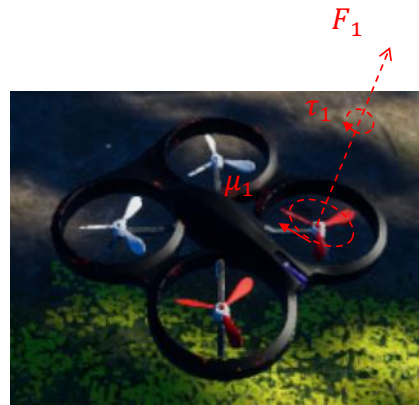


Figure 2. Quadrotor dynamic model visualization illustrating four rotors generating forces (F_1) and torques (τ_1) under control inputs u_1 .

The underlying physical model adheres to fundamental quadrotor dynamics described by:

$$\begin{cases} m \cdot \frac{d\mathbf{v}}{dt} = \mathbf{F} - m \cdot \mathbf{g} \\ I \cdot \frac{d\boldsymbol{\omega}}{dt} = \boldsymbol{\tau} \end{cases} \quad (4)$$

where m denotes UAV mass, $\mathbf{v}$ denotes the velocity vector, $\mathbf{F}$ denotes the total external force, $\mathbf{g}$ denotes the gravitational acceleration, I denotes the inertia matrix, $\boldsymbol{\omega}$ denotes the angular velocity vector, and $\boldsymbol{\tau}$ denotes the total external torque.

AirSim's native PID regulation module handles flight control execution, with parameter settings detailed in Table 2 and UAV specifications in Table 3. This architecture separates high-level decision making from low-level actuation: the policy network focuses on optimal command generation while AirSim's physics engine manages dynamic responses through PID-controlled motor outputs.

Table 2. PID parameter configurations for flight control.

Controller Type	K_P	K_I	K_D
Linear-velocity control	0.2	2.0	0.0
Angular-velocity control	0.25	0.0	0.0

Table 3. Characteristics of the quadrotor platform.

Item	Parameter
Weight	0.8 kg
Number of rotors	4
Minimum motor thrust	0
Maximum motor thrust	1
Minimum tilt throttle	0.05
Maximum angular velocity	2.5 rad/s
Dimensions (m)	$0.2413 \times 0.1143 \times 0.0762$

3.3. Architectural Overview

Our approach builds upon the previous work [13]. As shown in Figure 3, the framework comprises two parts: representation learning and policy learning. We construct a causal-representation-learning-structure-based autoencoder. It strives to learn representations that effectively reduce the information of obstacle shape. Subsequently, based on those causal representations along with the current speed and target position are inputted into policy network for strategy learning. Based on the current observation, the SAC algorithm will generate a series of actions for UAV control, which includes three-dimensional velocity adjustments. We adopted a learning paradigm of centralized training and distributed execution. Under this paradigm, the training phase has a global perspective, and the experience data collected by all drones during their interaction with the environment are aggregated into a central experience pool. This centralized learning enables the strategy to learn from the experience of all drones, which helps in the development of more effective and coordinated collision avoidance behaviors. In the execution phase, it is manifested as distributed operation, where each drone relies only on its own local observation information and independently uses this centrally trained and optimized shared policy network to make decisions and execute actions in real time, without the need for complex real-time communication or global state information between drones. This distributed execution method enhances the scalability and robustness of the system to single-point failures, allowing the trained strategy to be efficiently deployed in actual multi-drone operation scenarios.

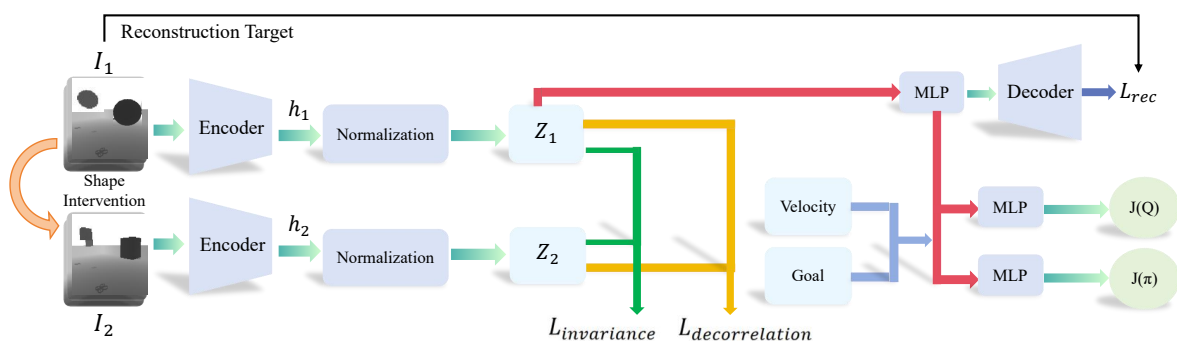


Figure 3. The illustration of our proposed method. To extract invariant causal representation, we intervene the shape of obstacles and maximize the mutual information through the $\mathcal{L}_{invariance}$. In addition, we introduce the $\mathcal{L}_{decorrelation}$ to ensure that different dimensions capture different information.

In our formulation, we consider a set of N UAVs operating in a shared environment. At each timestep t , the i -th UAV receives an observation o_i^t and computes an action a_i^t accordingly, with the objective of moving from its current position toward its goal position g_i^t . Throughout the navigation process, each UAV has access only to its own observation and does not have access to the states of other UAVs. The observation vector for each UAV consists of three components, $o^t = [o_i^t, o_v^t, o_g^t]$, where o_i^t represents images of the

surrounding environment captured by the onboard vision sensor, o_v^t denotes the current velocity of the UAV, and o_g^t indicates the relative position of the goal. Given this partial observation o^t , each UAV independently and simultaneously samples an action a^t which sampled from the policy π shared by all UAVs:

$$a^t \sim \pi_\theta(\cdot|o^t) \quad (5)$$

where θ is the parameter vector of the policy. The action a^t corresponds to a velocity command that guides the UAV toward its goal while avoiding collisions, and remains in effect over the time horizon Δt until the next observation o^t is received.

In the policy improvement step, we update the actor network by maximizing the loss function $J(\pi)$, defined as follows:

$$J(\pi) = \mathbb{E}_{o \sim B}[D_{KL}(\pi(a|o) \| Q(o, a))] \quad (6)$$

where the term $\pi(a|o)$ denotes the probability of taking action a given the observation o . The term $Q(o, a)$ represents the target action distribution under the observation o .

We update the critic network by minimizing the loss function $J(Q)$, formulated as follows:

$$J(Q) = \mathbb{E}_{(o, a, r, o') \sim B}[(Q(o, a) - r - \gamma \bar{V}(o'))^2] \quad (7)$$

where γ is the discount factor, $\bar{V}(o')$ is the target value function, and r denotes the obtained reward.

We adopt the loss function $\mathcal{L}(rec)$ to reconstruction image through updating the encoder p_ϕ and decoder ($z = g_x$):

$$\mathcal{L}(rec) = \mathbb{E}_x[\log p_\phi(x|z) + \lambda_z \|z\|^2 + \lambda_\phi \|\phi\|^2] \quad (8)$$

where x is the depth image extracted from observations. $\lambda_z \|z\|^2$ and $\lambda_\phi \|\phi\|^2$ are used to apply regularization constraints to z and model parameters ϕ . Both λ_z and λ_ϕ are hyperparameters.

3.4. Extract Invariant Causal Representation

This section focuses on the learning of causal representations from images, which makes the optimal policy based on these representations to remain robust across various training domains. Effectively, this method aims to uncover and utilize the factors that contribute to successful actions. Some works [30] have demonstrated that we need to discover invariant mechanisms from multiple source domain data and identify hidden causal variables. Therefore, in this work, we conduct causal intervention on the shape of obstacles to construct multiple source domain data with different obstacles. Specifically, we pause the state of the UAV at the same instant and then change the shape of the obstacle by controlling Airsim. Through this method, we can obtain depth images of obstacles with different shapes at the same position from the perspective of the UAV.

We propose a causal representation learning (CRL) method based on an autoencoder (AE) to identify causal representations, as depicted in Figure 3. Using a depth camera mounted facing forward on the UAV, we sample two distinct sets of observations from the environment: I_1 , containing depth images of the UAVs and original obstacles, and I_2 , containing depth images of the UAVs and obstacles with modified shapes. These depth images capture the distance of object surfaces from the camera's viewpoint. To incorporate temporal information, we stack four consecutive depth frames for each observation sequence. These stacked frames are then input into the encoder to obtain latent representations h_1 and

h_2 , corresponding to I_1 and I_2 , respectively. Subsequently, we apply instance-dimensional normalization to h_1 and h_2 to ensure each feature dimension has a mean of 0 and a standard deviation of $\frac{1}{\sqrt{d}}$ (where d is the feature dimensionality). The normalization is performed using the following formula:

$$z = \frac{h_i - \mu(h_i)}{\sigma(h_i) \times \sqrt{d}} \quad (9)$$

The obtained normalized z_1, z_2 are further used to maximize the mutual information through the invariance term:

$$\mathcal{L}_{invariance} = \|z_1 - z_2\|^2 \quad (10)$$

Intuitively, the invariance term is used to minimize the difference between two normalized representations.

Furthermore, previous work [31] has shown that multiple dimensions of representations share overlapping information. To ensure that different dimensions capture different information, we introduce the following decorrelation term:

$$\mathcal{L}_{decorrelation} = \mathcal{F}(z_1 z_1^T, I) + \mathcal{F}(z_2 z_2^T, I) \quad (11)$$

where $\mathcal{F}(\cdot, \cdot) = \|\cdot - \cdot\|_F^2$ denotes the Frobenius norm. The symbol “.” denotes the input to the function F and I is an identity matrix. The decorrelation term helps prevent the trivial collapse where the same vector is produced for all inputs by trying to equate the off-diagonal elements of the auto-correlation matrix of each representation to 0.

The causal intervention process can be formalized using do-calculus:

$$P(Y|do(X)) = \sum_z P(Y|X, Z=z)P(Z=z) \quad (12)$$

where $do(X)$ denotes the intervention operation, Z represents the confounding variables.

The invariance loss can be derived from mutual information:

$$I(Z_1; Z_2) = \mathbb{E}_{p(z_1, z_2)} \left[\log \frac{p(z_1, z_2)}{p(z_1)p(z_2)} \right] \quad (13)$$

Maximizing this mutual information ensures the learned representations are invariant to obstacle shapes.

The complete training procedure is formalized in Algorithm 1.

Algorithm 1 Causal representation learning process.

Require: Training environments $\mathcal{E}_1, \dots, \mathcal{E}_n$

- 1: **for** each training iteration **do**
 - 2: Sample batch x_i from training experience // Gather diverse states before intervention
 - 3: Apply shape intervention: $x_i^{intv} \leftarrow \text{Intervention}(x_i)$ // Online causal intervention
 - 4: Encode representations: $z_i, z_i^{intv} \leftarrow f_\theta(x_i), f_\theta(x_i^{intv})$
 - 5: Compute invariance loss: $\mathcal{L}_{invariance} = \|z_i - z_i^{intv}\|^2$ // Representation invariance enforcement
 - 6: Compute decorrelation loss: $\mathcal{L}_{decorrelation} = \mathcal{F}(z_1 z_1^T, I) + \mathcal{F}(z_2 z_2^T, I)$ // Dimensional decorrelation
 - 7: Compute total representation loss: $\mathcal{L}_{total} \leftarrow \mathcal{L}_{invariance} + \lambda \mathcal{L}_{decorrelation}$
 - 8: Update encoder: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{inv}$
 - 9: Train policy network with $\{z_i\}$
 - 10: **end for**
-

3.5. Causal Identifiability Analysis

As depicted in Figure 4, our structural causal model formalizes the key relationships between obstacle characteristics, sensor observations, and control policies. The identifiability of causal representation Z is guaranteed through our intervention mechanism and contrastive learning framework. We formally prove the identifiability via the following three aspects:

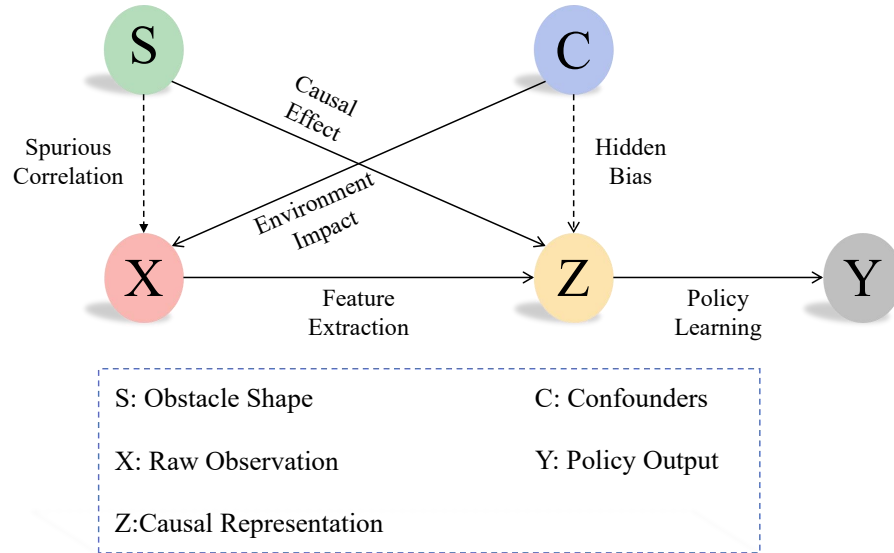


Figure 4. Structural causal model characterizing the collision avoidance system. Observed variables (S, X, Y) represent obstacle shapes, sensor inputs, and control commands, respectively. Latent variable Z denotes the causal representation learned through intervention $do(S)$, while C captures environmental confounders. Directed edges encode causal mechanisms: solid lines show the intended policy learning path $S \rightarrow Z \rightarrow Y$; dashed lines indicate spurious correlations $S \rightarrow X \rightarrow Z$ and confounding paths $C \rightarrow X \rightarrow Z$. The model satisfies identifiability via $Z \perp C | do(S)$, enabling generalization across environments.

Theorem 1 (Backdoor Path Blocking). *Under intervention $do(S = s)$, the confounding path $C \rightarrow X \rightarrow Z$ shown by dashed arrows in Figure 4 is blocked by cutting off the natural data-generating process of S . This satisfies the backdoor criterion:*

$$P(Z | do(S)) = \int P(Z | S, C) P(C) dC = P(Z | S) \quad (14)$$

The intervention severs the spurious correlation between S and C (visualized by the red dashed path in Figure 4), ensuring that the learned representation Z only captures the causal relationship $S \rightarrow Z$ (solid blue arrow). This is implemented through our obstacle shape randomization in AirSim’s rendering pipeline.

Theorem 2 (Conditional Independence). *The invariance loss $\mathcal{L}_{invariance}$ enforces the independence structure shown in Figure 4’s caption:*

$$Z \perp\!\!\!\perp C | do(S) \Rightarrow I(Z; C | do(S)) = 0 \quad (15)$$

where $I(\cdot)$ denotes mutual information. This conditional independence is achieved through the following:

- Instance-wise normalization (Equation (9)), removing environment-specific statistics.
- Intervention-invariant representation learning (Equation (10)), enforcing $\|z_1 - z_2\|^2$ minimization across environments under the same intervention.

The decorrelation loss $\mathcal{L}_{\text{decorrelation}}$ further ensures disentangled representations by enforcing orthogonality constraints on the latent space, corresponding to the absence of bidirected edges between Z nodes in Figure 4:

$$\mathbb{E}[z_i^T z_j] = \delta_{ij}, \quad \forall i \neq j \quad (16)$$

This eliminates redundant information channels that might encode confounding factors.

Theorem 3 (Identifiable Mechanism). *The structural equations satisfy the causal Markov condition shown in Figure 4:*

$$Z = f(S, \epsilon_z), \quad \epsilon_z \perp (S, C) \quad (17)$$

where f is an invertible nonlinear function learned by the encoder. The additive noise model ensures the identifiability up to component-wise transformations, corresponding to the solid causal pathway $S \rightarrow Z \rightarrow Y$ in our SCM.

The combination of intervention-based data generation and constrained representation learning establishes a minimal sufficient statistic Z for policy decisions. As shown in the causal graph (Figure 4), this enables generalization through the invariant causal mechanism $P(Y|do(Z))$ rather than spurious correlations $P(Y|X)$ mediated by the dashed paths.

4. Results

4.1. Simulation Environment and Experimental Setup

AirSim is a high-fidelity simulation platform widely used for UAV (unmanned aerial vehicle) testing. Given its robust capabilities, we have chosen AirSim as the foundation for our simulation framework to investigate UAV collision avoidance algorithms. The simulation experiments were conducted on a computer equipped with the Ubuntu 20.04 operating system and powered by a single NVIDIA RTX 4090 GPU. The hyperparameters employed in this study are thoroughly outlined in Table 4.

Table 4. Hyperparameters for policy training.

Parameter Name	Value
Batch size	128
Max episodes	150
Update times	400
Replay buffer $\mathcal{B}$ capacity	20,000
Discount γ	0.99
Learning Rate	10^{-3}
Critic's target update frequency	2
Critic's Q-function soft-update rate τ_Q	0.01
Critic's encoder soft-update rate τ_{enc}	0.05
Actor's update frequency	2
Actor's log stddev bounds	$[-10, 2]$
Optimizer	Adam

4.2. Evaluation Metrics and Experimental Scenarios

To evaluate the efficacy of the proposed methodology, the following performance indicators were established:

- **Success Rate:** This metric quantifies the proportion of UAVs that successfully reach their destinations within a predefined time interval, devoid of any collisions.
- **Excess Distance:** This denotes the supplementary distance traversed by UAVs beyond the direct linear distance between the origin and the target destination.

- Success-weighted Path Length (SPL): This metric amalgamates the success rate with the efficiency of the navigated path, offering a holistic assessment of task accomplishment and path efficacy.
- Average Velocity: This is computed as the quotient of the aggregate flight path length and the time elapsed for the UAV.

The performance metrics are reported as mean/standard deviation (mean/std), calculated over multiple simulation trials for each experimental condition. The mean represents the average performance achieved across these trials, providing a measure of central tendency. The standard deviation (std) quantifies the variability or spread of the results.

As illustrated in Figure 5, to further evaluate the proposed method's generalization capabilities, the CRL was tested against diverse obstacle shapes kept at a comparable size. The evaluated geometries encompassed simpler forms like cubes (side length—0.1 m) and spheres (diameter—0.1 m), alongside more complex structures including triangles, cylinders, pentahedrons, and cuboids (representative dimensions, 0.1 m × 0.1 m × 0.2 m). We randomly generate the initial and target positions of the UAVs and the specified positions of the obstacles in an area with a constant spatial dimension of 16 m × 16 m × 4 m configured in the simulation environment, containing eight UAVs and four stationary obstacles. In the training scenario, we choose the cube obstacle and the sphere obstacle as the seen obstacles of the UAV, while the others are the unknown obstacles during testing. In the training process, we fix the position of the obstacles in each frame; then, after the image data of the cube obstacle are obtained by the UAV, the original positions of the UAV and the obstacle are kept, and the shape of the obstacle is changed to that of the sphere obstacle, so as to achieve the purpose of shape intervention.

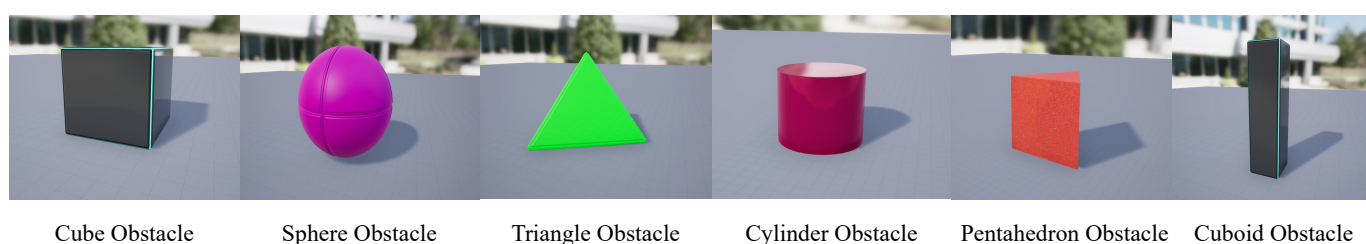


Figure 5. Simulation scenarios for evaluating the generalization ability. We set up six different shapes of obstacles in the playground scene during the testing phase.

4.3. Performance Comparison

4.3.1. Evaluation Obstacle Shapes

We compared our method against the SAC + RAE baseline across various obstacle scenarios under clear conditions to rigorously evaluate its generalization capabilities when faced with diverse object geometries. As quantitatively detailed in Table 5, our causal representation learning (CRL) approach significantly outperforms the baseline. This superiority is particularly evident when encountering previously unseen obstacle shapes such as triangles, cylinders, pentahedrons, and cuboids. For instance, in scenarios involving cuboid obstacles, CRL achieved a notable 6.2% higher success rate and a 3.7% improvement in SPL. These results highlight the effectiveness of learning true causal representations. By intervening on obstacle shape during training and enforcing invariance in the learned feature space, CRL learns features that correspond to the fundamental presence and properties of obstacles rather than overfitting to superficial characteristics like specific shapes (e.g., cubes or spheres) encountered during training, while the baseline method, SAC + RAE, struggles with novel geometries due to its reliance on potentially spurious correlations learned from the training distribution, CRL maintains robust and adaptive performance.

Furthermore, we challenged the models with randomly generated scenes where four different unseen obstacles were simultaneously present. The results from this “Mixed” scenario, also presented in Table 5, demonstrate CRL’s strong generalization capabilities, achieving a 7.1% higher success rate than the baseline. This underscores the ability of our method to handle complex environments with a variety of concurrently encountered novel objects, suggesting that the learned causal features provide a versatile foundation for decision making.

A nuanced observation pertains to the “Extra Distance” metric. For some unseen obstacles, such as cuboids, CRL resulted in a slightly longer flight path (1.851 m compared to 1.635 m for SAC + RAE). This suggests that the agent, guided by its understanding of core causal features, may adopt slightly more cautious or deliberative paths when navigating unfamiliar obstacles. This behavior prioritizes successful avoidance and mission completion over the shortest possible path, a strategy that might be favored by an overfitted baseline potentially underestimating risks associated with novel geometries. Nevertheless, the consistently higher average velocity observed for CRL across most scenarios indicates that, once decisions are made based on these robust causal features, the agent executes its actions efficiently. The ability to maintain good operational speed while navigating cautiously around novel objects points to a well-balanced policy learned by CRL. This careful yet efficient navigation is crucial for practical UAV applications where both safety and mission timeliness are paramount.

Table 5. Performance (as mean/std) comparison with different obstacle shapes.

Obstacle Shape	Unseen/Seen	Method	Success Rate (%)	SPL (%)	Extra Distance (m)	Average Velocity (m/s)
Cube	Seen	SAC + RAE	67.6	58.3	1.483/1.436	0.771/0.158
		Our method	74.4 (↑ 6.8)	62.5 (↑ 4.2)	1.542/1.519	0.806/0.125
Sphere	Seen	SAC + RAE	68.3	58.4	1.680/1.709	0.783/0.160
		Our method	73.4 (↑ 5.1)	62.2 (↑ 3.8)	1.735/1.718	0.814/0.132
Triangle	Unseen	SAC + RAE	72.5	62.1	1.535/1.622	0.775/0.164
		Our method	76.3 (↑ 3.8)	64.4 (↑ 2.3)	1.815/1.802	0.809/0.135
Cylinder	Unseen	SAC + RAE	67.6	58.6	1.646/1.621	0.783/0.155
		Our method	72.1 (↑ 4.5)	61.7 (↑ 3.1)	1.853/1.810	0.810/0.138
Pentahedron	Unseen	SAC + RAE	67.3	56.7	1.575/1.546	0.784/0.153
		Our method	73.8 (↑ 6.5)	61.8 (↑ 5.1)	1.651/1.623	0.813/0.130
Cuboid	Unseen	SAC + RAE	68.2	59.3	1.635/1.578	0.799/0.131
		Our method	74.4 (↑ 6.2)	63.0 (↑ 3.7)	1.851/1.820	0.803/0.136
Mixed	Unseen	SAC + RAE	67.1	60.4	1.646/1.601	0.801/0.129
		Our method	74.2 (↑ 7.1)	62.9 (↑ 2.5)	1.858/1.826	0.805/0.131

4.3.2. Evaluation Under Variable Weather Conditions

We conducted comprehensive testing of our method under diverse weather conditions while maintaining fixed cube-shaped obstacles, as shown in Figure 6.

The results in Table 6 underscore CRL’s robustness in challenging weather conditions. Compared to SAC + RAE, CRL shows marked improvements in success rate and SPL across snow, dust, and fog scenarios (e.g., an 11.6% increase in success rate in fog). This suggests that the causal representation learning helps the agent disentangle relevant obstacle information from sensor noise and visibility variations introduced by weather; while SAC + RAE’s performance degrades significantly as visual inputs become corrupted, CRL’s focus on invariant features provides a more stable basis for navigation policy. The consistently longer ‘Extra Distance’ reinforces the idea that CRL prompts more path adjustments, likely

to compensate for the increased uncertainty in perception under adverse conditions, but crucially, these adjustments lead to a much higher overall task success rate.

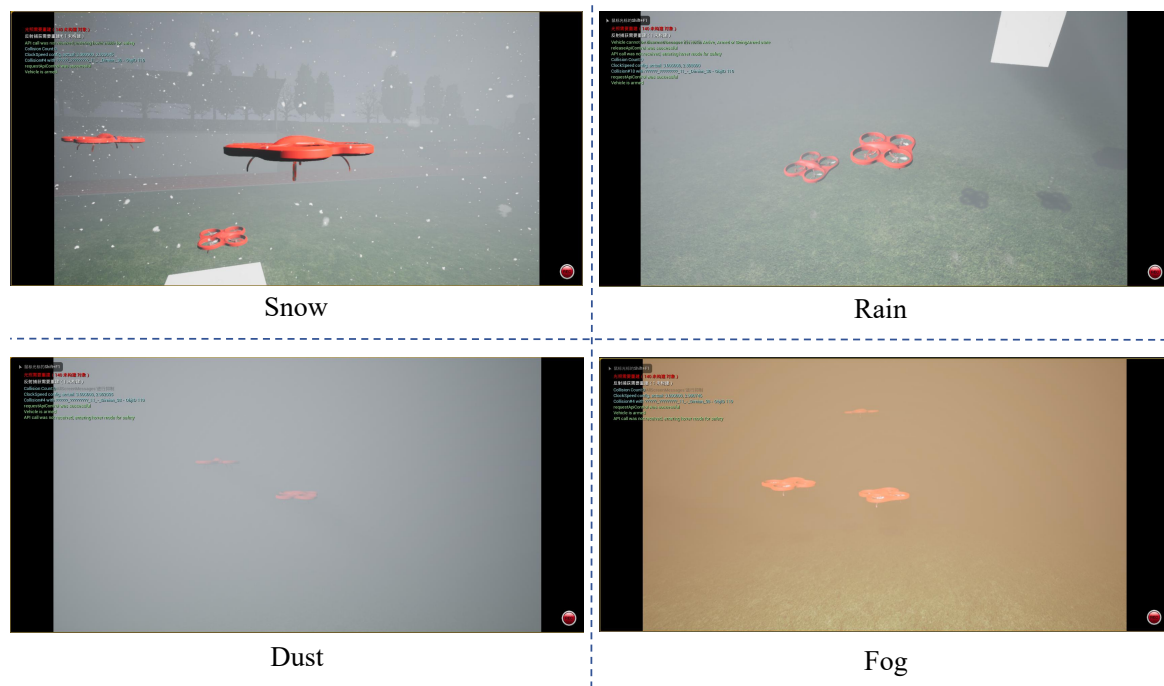


Figure 6. Simulation scenarios under variable weather conditions. The figure displays four different weather scenarios: snow, rain, dust, and fog, used to evaluate the generalization ability of our proposed method.

The results show our method maintains superior average velocity when navigating through static cube obstacles under changing weather, which is critical for mission efficiency. However, similar to previous evaluations, our method exhibits longer extra distances. This suggests UAVs require more frequent path adjustments to compensate for weather-induced sensor noise and visibility changes, ultimately increasing total travel distance.

Table 6. Comparison of the performance (as mean/std) of cube obstacles under different weather conditions.

Weather Type	Unseen/Seen	Method	Success Rate (%)	SPL (%)	Extra Distance (m)	Average Velocity (m/s)
Rain	Seen	SAC + RAE	63.6	51.2	2.653/3.143	0.424/0.137
		Our method	69.5 (↑ 5.9)	54.3 (↑ 3.1)	2.163/3.785	0.374/0.115
Snow	Unseen	SAC + RAE	61.8	51.1	2.699/2.685	0.406/0.131
		Our method	71.2 (↑ 9.4)	56.4 (↑ 5.3)	2.432/3.773	0.385/0.112
Dust	Unseen	SAC + RAE	59.8	51.7	3.124/2.620	0.371/0.131
		Our method	68.1 (↑ 8.3)	58.1 (↑ 6.4)	2.399/4.079	0.381/0.109
Fog	Unseen	SAC + RAE	57.9	49.8	2.965/2.545	0.387/0.134
		Our method	69.5 (↑ 11.6)	59.7 (↑ 9.9)	2.447/4.786	0.379/0.124

4.3.3. More Evaluation Baselines

To thoroughly assess the generalization capability of the CRL approach, we conducted a comparative analysis against several established deep reinforcement learning (DRL) generalization techniques. Specifically, CRL was benchmarked against three prominent baselines: the augmentation-based AutoAugment [32], the regularization-based DrAC [33], and the attention-based SE [34]. All methods were evaluated in a simulated playground environment containing four previously unseen cuboid obstacles. As shown in Table 7,

CRL demonstrates clear advantages over the baselines. It achieves the highest success rate and SPL, outperforming AutoAugment, DrAC, and SE. Furthermore, CRL achieves this high success rate with notable efficiency, while its extra distance is comparable to the other baselines, its average velocity (0.860 m/s) is substantially higher than that achieved by AutoAugment (0.398 m/s), DrAC (0.379 m/s), and SE (0.377 m/s). This suggests that by focusing on invariant causal features, CRL not only improves the reliability of collision avoidance but also enables more decisive and efficient navigation when faced with unfamiliar scenarios. In summary, when evaluated against leading generalization strategies on previously unseen obstacles, the proposed CRL method achieved superior performance—delivering the highest success rate and SPL while maintaining competitive efficiency—validating its effectiveness in enhancing generalization for DRL-based collision avoidance.

Table 7. Performance comparison of various methods in the testing scenario with four cuboid obstacles.

Method	Success Rate (%)	SPL (%)	Extra Distance (m)	Average Velocity (m/s)
SAC + RAE	68.2	59.3	1.635/1.578	0.799/0.131
+ AutoAugment [32]	68.8	61.2	1.505/2.881	0.398/0.131
+ DrAC [33]	69.8	61.5	1.538/2.693	0.379/0.122
+ SE [34]	70.1	61.6	1.564/2.648	0.377/0.122
+ Our method	74.4	62.5	1.542/1.519	0.860/0.125

4.3.4. Ablation Experiment

An ablation study was performed to assess the individual contributions of the invariance loss ($\mathcal{L}_{invariance}$) and the decorrelation loss ($\mathcal{L}_{decorrelation}$) to the model's performance, with results summarized in Table 8. The baseline model, without either of these loss components, achieved a success rate of 68.2%. Introducing only the decorrelation loss ($\mathcal{L}_{decorrelation}$) resulted in a modest improvement, increasing the success rate to 69.6%. When only the invariance loss ($\mathcal{L}_{invariance}$) was utilized, a more significant increase in performance was observed, with the success rate reaching 70.3%. However, the most substantial improvement was achieved when both the invariance and decorrelation losses were applied concurrently, yielding the highest success rate of 74.4%. In addition, under the condition of only $\mathcal{L}_{invariance}$ or $\mathcal{L}_{decorrelation}$, the drone needs to fly more distance and slower speed. This shows that, although $\mathcal{L}_{invariance}$ and $\mathcal{L}_{decorrelation}$ can each bring performance gains to the multi-drone collision avoidance task, the organic combination of the two is crucial to maximizing the ability of the agent to learn effective causal representations and ultimately achieve excellent generalization performance, mission success rate, and navigation efficiency.

Table 8. Ablation study on different loss functions.

$\mathcal{L}_{invariance}$	$\mathcal{L}_{decorrelation}$	Success Rate(%)	SPL(%)	Extra Distance (m)	Average Velocity (m/s)
		68.2	59.3	1.635/1.578	0.799/0.131
✓		69.6	60.1	1.584/1.539	0.814/0.129
	✓	70.3	60.4	1.603/1.554	0.821/0.122
✓	✓	74.4	62.5	1.542/1.519	0.860/0.125

4.3.5. Scalability Analysis

To evaluate the scalability of proposed approach, we set up scenes with different numbers of UAVs and obstacles independently within a playground scene that includes

four unseen cuboid obstacles. As shown in Figure 7, our method consistently exceeds the performance of SAC + RAE, demonstrating its superior robustness and scalability.

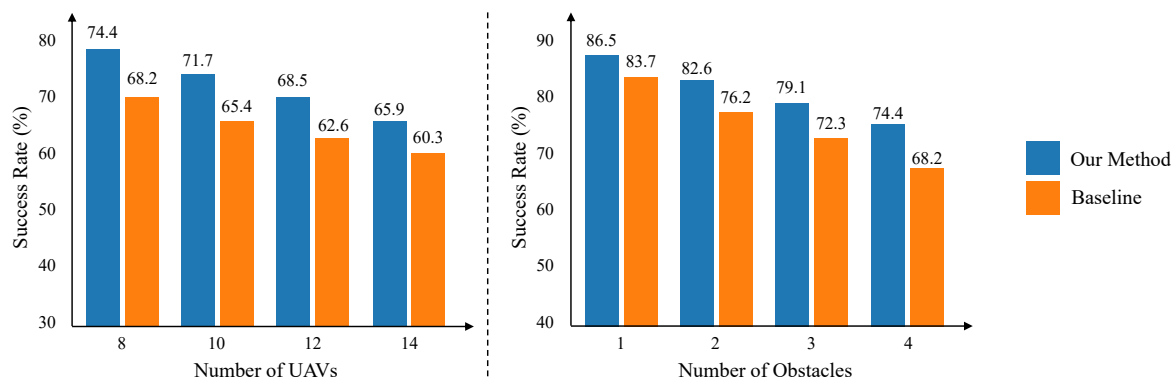


Figure 7. Scalability analysis experiments. Performance comparison between our proposed CRL and the baseline model under conditions with different numbers of UAVs and obstacles. The success rate is used as the evaluation metric.

4.4. Visualization of Trajectory

To visually illustrate the collision evasion efficacy of our methodology, graphical representations of UAV flight paths are presented through perspective and triaxial orthographic projections. As exhibited in Figure 8, the proposed framework manifests enhanced trajectory optimization proficiency and obstacle navigation performance when contrasted with SAC + RAE architecture. This comparative analysis substantiates the operational superiority of our algorithm in multi-agent coordination scenarios.

Figure 8 illustrates the scalability of our approach. As the number of UAVs or obstacles increases, the performance gap between CRL and the baseline widens. For instance, with 14 UAVs, CRL maintains a 65.9% success rate compared to the baseline's 60.3%. This superior scalability indicates that CRL's learned invariant representations are more robust to the increased complexity and potential for spurious correlations inherent in denser multi-agent scenarios. Focusing on fundamental causal factors allows the policy to remain effective even when interactions become more frequent and intricate. In addition, we compare the average extra distance in one round as well as the average speed of baseline and our method. The results are shown in Figure 9, where our method has advantages in average extra distance and average speed.

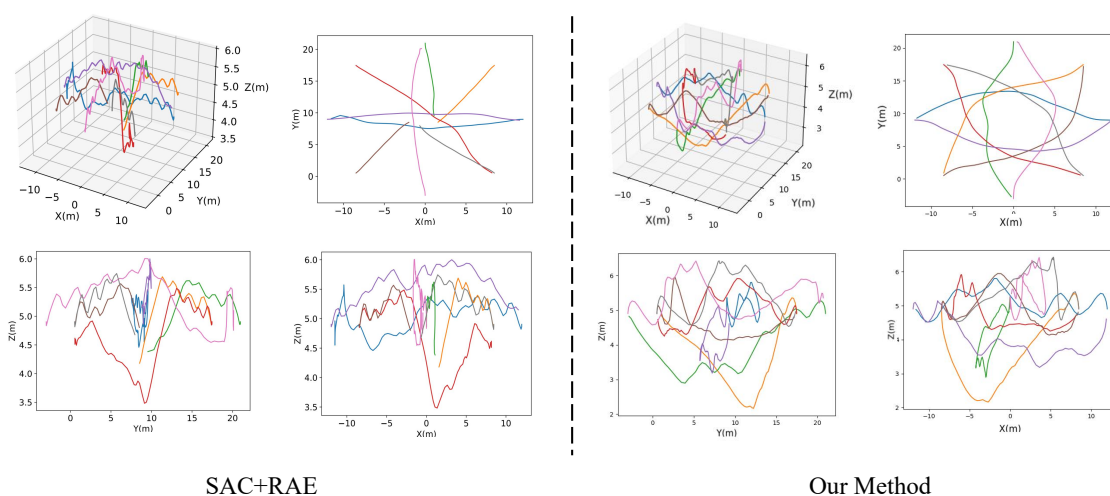


Figure 8. Visualization of UAVs trajectories in both perspective and three-view illustrations. We employ various colors to depict the trajectories of distinct UAVs.

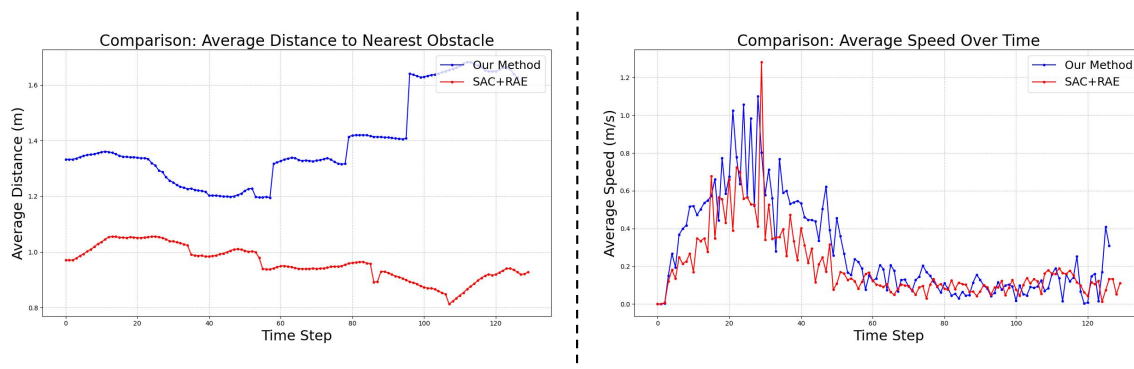


Figure 9. One round average extra distance contrast as well as average speed contrast for baseline and our method.

5. Conclusions

This paper presents a novel causal representation learning (CRL) method to enhance the generalization ability of deep reinforcement learning (DRL) models for multi-UAV collision avoidance. By intervening in the shape of obstacles and maximizing mutual information through the invariance term, our approach effectively identifies invariant causal representations from images. The introduction of the decorrelation term ensures that different dimensions of the representation capture distinct information, further improving the robustness of the learned policy. Extensive experiments demonstrate that our CRL method significantly outperforms state-of-the-art methods in various testing scenarios, including different obstacle shapes and variable weather conditions. The results highlight higher success rates, improved SPL (success weighted by path length), and faster average velocities; while our method shows promise, future work will focus on addressing current limitations, such as enhancing adaptability to background changes and validating the learned strategies in real-world settings. Additionally, we plan to explore more powerful disentanglement methods and advanced causal intervention strategies to further improve the generalization capabilities of this approach.

6. Discussion

The principal aim of our methodological framework was to assess the CRL technique's capacity to mitigate the generalization challenge inherent in DRL frameworks when encountering unforeseen obstacles. Empirical findings substantiate the efficacy and preeminence of the CRL methodology in optimizing generalization capabilities across diverse evaluative contexts, confirming its operational superiority in cross-environmental adaptation tasks. However, several limitations in our current work warrant further investigation. Firstly, our model's adaptability to background changes is limited, and future work should focus on improving this aspect. Secondly, our training and evaluation were conducted exclusively in simulated environments due to hardware constraints; while simulated environments provide a controlled setting for initial testing, real-world validation is crucial for assessing the practical applicability and robustness of the learned strategies. Future work should aim to deploy and evaluate our CRL method in real-world settings to bridge the gap between simulation and real-world deployment. In addition to addressing these limitations, future work will explore more powerful disentanglement methods and advanced causal intervention strategies to further enhance the generalization capability of our CRL approach. We also plan to apply the CRL method to other DRL learning tasks and investigate its potential in various application domains.

Author Contributions: Conceptualization, Z.F.; methodology, G.H. and C.L.; software, C.L., G.H., and Q.W.; validation, C.L. and G.H.; formal analysis, C.L., and G.H.; investigation, C.L. and G.H.; resources, Z.F. and J.Z.; data curation, C.L., and G.H.; writing—original draft preparation, C.L., G.H., and Q.W.; writing—review and editing, C.L., B.W., Q.W., and G.H.; visualization, C.L. and G.H.; supervision, W.L., Z.H., J.Z., G.H., and Z.F.; project administration, J.Z. and Z.F.; funding acquisition, J.Z., W.L., and Z.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Science and Technology Major Project (2021ZD0111502), the National Natural Science Foundation of China (grant numbers 62176147, 62406186, 62476163), the Science and Technology Planning Project of Guangdong Province of China (2022A1515110660), and the STU Scientific Research Foundation for Talents (NTF21001, NTF22030).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	unmanned aerial vehicle
SLAM	simultaneous localization and mapping
DRL	deep reinforcement learning
RAE	regularized autoencoder
CRL	causal representation learning

References

1. Burgués, J.; Marco, S. Environmental chemical sensing using small drones: A review. *Sci. Total Environ.* **2020**, *748*, 141172. [[CrossRef](#)] [[PubMed](#)]
2. Delmerico, J.; Mintchev, S.; Giusti, A.; Gromov, B.; Melo, K.; Horvat, T.; Cadena, C.; Hutter, M.; Ijspeert, A.; Floreano, D.; et al. The current state and future outlook of rescue robotics. *J. Field Robot.* **2019**, *36*, 1171–1191. [[CrossRef](#)]
3. Khan, A.; Gupta, S.; Gupta, S.K. Emerging UAV technology for disaster detection, mitigation, response, and preparedness. *J. Field Robot.* **2022**, *39*, 905–955. [[CrossRef](#)]
4. Betti Sorbelli, F. UAV-based delivery systems: A systematic review, current trends, and research challenges. *J. Auton. Transp. Syst.* **2024**, *1*, 1–40. [[CrossRef](#)]
5. Tsouros, D.C.; Bibi, S.; Sarigiannidis, P.G. A review on UAV-based applications for precision agriculture. *Information* **2019**, *10*, 349. [[CrossRef](#)]
6. Zhuang, L.; Zhong, X.; Xu, L.; Tian, C.; Yu, W. Visual SLAM for unmanned aerial vehicles: Localization and perception. *Sensors* **2024**, *24*, 2980. [[CrossRef](#)] [[PubMed](#)]
7. Sonugür, G. A Review of quadrotor UAV: Control and SLAM methodologies ranging from conventional to innovative approaches. *Robot. Auton. Syst.* **2023**, *161*, 104342. [[CrossRef](#)]
8. Zhang, S.; Wang, H.; Wang, C.; Wang, Y.; Wang, S.; Yang, Z. An improved RANSAC-ICP method for registration of SLAM and UAV-LiDAR point cloud at plot scale. *Forests* **2024**, *15*, 893. [[CrossRef](#)]
9. Junior, F.L.; Moreira, L.A.; Moreira, E.M.; Balddivieso, T.J.; Brunaes, M.S.; Rosa, P.F. UAV path automation using visual waypoints acquired from the ground. In Proceedings of the 2020 IEEE 29th International Symposium on Industrial Electronics (ISIE), Delft, The Netherlands, 17–19 June 2020.
10. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [[CrossRef](#)]
11. Wang, C.; Wang, J.; Shen, Y.; Zhang, X. Autonomous navigation of UAVs in large-scale complex environments: A deep reinforcement learning approach. *IEEE Trans. Veh. Technol.* **2019**, *68*, 2124–2136. [[CrossRef](#)]
12. Hu, K.; Li, H.; Zhuang, J.; Hao, Z.; Fan, Z. Efficient focus autoencoders for fast autonomous flight in intricate wild scenarios. *Drones* **2023**, *7*, 609. [[CrossRef](#)]

13. Huang, H.; Zhu, G.; Fan, Z.; Zhai, H.; Cai, Y.; Shi, Z.; Dong, Z.; Hao, Z. Vision-based distributed multi-UAV collision avoidance via deep reinforcement learning for navigation. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022.
14. Fan, Z.; Xia, Z.; Lin, C.; Han, G.; Li, W.; Wang, D.; Chen, Y.; Hao, Z.; Cai, R.; Zhuang, J. UAV Collision Avoidance in Unknown Scenarios with Causal Representation Disentanglement. *Drones* **2024**, *9*, 10. [\[CrossRef\]](#)
15. Zhuang, J.; Han, G.; Xia, Z.; Wang, B.; Li, W.; Wang, D.; Hao, Z.; Cai, R.; Fan, Z. Robust policy learning for multi-uav collision avoidance with causal feature selection. *arXiv* **2024**, arXiv:2407.04056.
16. Brehmer, J.; De Haan, P.; Lippe, P.; Cohen, T.S. Weakly supervised causal representation learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 38319–38331.
17. Han, G.; Wu, Q.; Wang, B.; Lin, C.; Zhuang, J.; Li, W.; Hao, Z.; Fan, Z. Deep Reinforcement Learning Based Multi-UAV Collision Avoidance with Causal Representation Learning. In Proceedings of the 2024 10th International Conference on Big Data and Information Analytics (BigDIA), Chiang Mai, Thailand, 25–28 October 2024; pp. 833–839. [\[CrossRef\]](#)
18. Çetin, E.; Barrado, C.; Muñoz, G.; Macias, M.; Pastor, E. Drone Navigation and Avoidance of Obstacles Through Deep Reinforcement Learning. In Proceedings of the 2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC), San Diego, CA, USA, 8–12 September 2019; pp. 1–7. [\[CrossRef\]](#)
19. Ouahouah, S.; Bagaa, M.; Prados-Garzon, J.; Taleb, T. Deep-reinforcement-learning-based collision avoidance in UAV environment. *IEEE Internet Things J.* **2021**, *9*, 4015–4030. [\[CrossRef\]](#)
20. Shin, S.Y.; Kang, Y.W.; Kim, Y.G. Obstacle avoidance drone by deep reinforcement learning and its racing with human pilot. *Appl. Sci.* **2019**, *9*, 5571. [\[CrossRef\]](#)
21. Yang, Y.; Zhang, K.; Liu, D.; Song, H. Autonomous UAV Navigation in Dynamic Environments with Double Deep Q-Networks. In Proceedings of the 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 11–15 October 2020, pp. 1–7. [\[CrossRef\]](#)
22. Yu, Y.; Liu, Y.; Wang, J.; Noguchi, N.; He, Y. Obstacle avoidance method based on double DQN for agricultural robots. *Comput. Electron. Agric.* **2023**, *204*, 107546. [\[CrossRef\]](#)
23. Chikhaoui, K.; Ghazzai, H.; Massoud, Y. PPO-based reinforcement learning for UAV navigation in urban environments. In Proceedings of the 2022 IEEE 65th International Midwest Symposium on Circuits and Systems (MWSCAS), Fukuoka, Japan, 7–10 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–4.
24. Liang, C.; Liu, L.; Liu, C. Multi-UAV autonomous collision avoidance based on PPO-GIC algorithm with CNN-LSTM fusion network. *Neural Netw.* **2023**, *162*, 21–33. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Bouhamed, O.; Ghazzai, H.; Besbes, H.; Massoud, Y. Autonomous UAV Navigation: A DDPG-Based Deep Reinforcement Learning Approach. In Proceedings of the 2020 IEEE International Symposium on Circuits and Systems (ISCAS), Sevilla, Spain, 10–12 October 2020; pp. 1–5. [\[CrossRef\]](#)
26. Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; Wang, J. Causalvae: Disentangled representation learning via neural structural causal models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021.
27. Zhang, K.; Xie, S.; Ng, I.; Zheng, Y. Causal representation learning from multiple distributions: A general setting. *arXiv* **2024**, arXiv:2402.05052.
28. Zhao, W.; Tang, D.; Chen, X.; Lv, D.; Ou, D.; Li, B.; Jiang, P.; Gai, K. Disentangled causal embedding with contrastive learning for recommender system. In Proceedings of the Companion Proceedings of the ACM Web Conference 2023, Austin, TX, USA, 30 April–4 May 2023.
29. Shah, S.; Dey, D.; Lovett, C.; Kapoor, A. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In Proceedings of the Field and Service Robotics: Results of the 11th International Conference, Zurich, Switzerland, 12–16 June 2017.
30. Zhang, K.; Gong, M.; Schölkopf, B. Multi-source domain adaptation: A causal view. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
31. Ji, Q.; Li, J.; Hu, J.; Wang, R.; Zheng, C.; Xu, F. Rethinking dimensional rationale in graph contrastive learning from causal perspective. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024.
32. Hansen, N.; Wang, X. Generalization in reinforcement learning by soft data augmentation. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021.
33. Raileanu, R.; Goldstein, M.; Yarats, D.; Kostrikov, I.; Fergus, R. Automatic data augmentation for generalization in reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 5402–5415.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.