Robust Policy Learning for Multi-UAV Collision Avoidance with Causal Feature Selection

Jiafan Zhuang
Shantou University
Shantou, China
jfzhuang@stu.edu.cn

Gaofei Han
Shantou University
Shantou, China
22gfhan@stu.edu.cn

Zihao Xia
Shantou University
Shantou, China
22zhxia@stu.edu.cn

Che Lin
Shantou University
Shantou, China
19clin1@stu.edu.cn

Boxi Wang
Shantou University
Shantou, China
19bxwang@stu.edu.cn

Wenji Li
Shantou University
Shantou, China
liwj@stu.edu.cn

Dongliang Wang
Shantou University
Shantou, China
dlwang@stu.edu.cn

Zhifeng Hao
Shantou University
Shantou, China
haozhifeng@stu.edu.cn

Ruichu Cai
Guangdong University of Technology
Guangzhou, China
cairuichu@gmail.com

Zhun Fan ✉
University of Electronic Science and
Technology of China
Shenzhen, China
fanzhun@uestc.edu.cn

ABSTRACT

Collision avoidance navigation for unmanned aerial vehicle (UAV) swarms in complex and unseen outdoor environments presents a significant challenge, as UAVs are required navigate through various obstacles and intricate backgrounds. While existing deep reinforcement learning (DRL)-based collision avoidance methods have shown promising performance, they often suffer from poor generalization, leading to degraded performance in unseen environments. To address this limitation, we investigate the root causes of weak generalization in DRL models and propose a novel causal feature selection module. This module can be integrated into the policy network to effectively filter out non-causal factors in representations, thereby minimizing the impact of spurious correlations between non-causal elements and action predictions. Experimental results demonstrate that the proposed method achieves robust navigation performance and effective collision avoidance, particularly in scenarios with unseen backgrounds and obstacles, which significantly outperforms state-of-the-art (SOTA) algorithms.

KEYWORDS

Reinforcement Learning; Representation Learning; Vision-Based Navigation; Multi-Robot Systems

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. NowÃľ (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

ACM Reference Format:
Jiafan Zhuang, Gaofei Han, Zihao Xia, Che Lin, Boxi Wang, Wenji Li, Dongliang Wang, Zhifeng Hao, Ruichu Cai, and Zhun Fan ✉. 2025. Robust Policy Learning for Multi-UAV Collision Avoidance with Causal Feature Selection. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

1 INTRODUCTION

In recent years, unmanned aerial vehicle systems [5, 33, 60] have made significant progress and been applied in domains, like agriculture [26, 50], search and rescue operations [47, 48], mining industry [43], and patrol inspections [29]. To enable effective collaboration among UAVs, it is crucial to identify an optimal path to target position while avoiding collisions, especially in large swarms [24]. As a result, multi-UAV collision avoidance has emerged as a fundamental and critical task, drawing increasing attention from researchers. Traditional approaches of multi-UAV collision avoidance [17, 38, 49] predominantly rely on real-time simultaneous localization and mapping (SLAM) [3]. These methods employ sensors, such as LiDAR, to perceive the surrounding environment and generate trajectories through path planning algorithms[1]. Additionally, prior maps [27, 31, 51] are often integrated to enhance SLAM system performance. However, these traditional methods typically require substantial computational resources and are constrained by availability of prior maps, which makes it less adaptable to complex and dynamic environments.

To address the limitations of traditional methods, deep reinforcement learning (DRL) [2, 7] has been extensively studied and applied in robotics. DRL facilitates end-to-end collision avoidance navigation [20] using only sensor data as input, thereby eliminating the

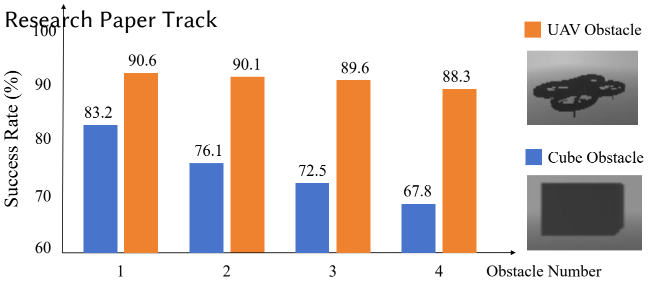


Figure 1: The influence of obstacle shape on success rate. For the DRL model, UAV-shaped obstacles have been seen during training while cube-shaped obstacles are unseen. During testing, the former has little influence on performance while the latter would significantly reduce the individual success rate of navigation.

dependence on prior environmental maps. Specifically, DRL learns visual representations from sensor inputs and maps these representations to optimal flight strategies by designing appropriate policy networks and reward functions. However, DRL is fundamentally a data-driven approach that typically assumes the training and testing data are sampled from independent and identically distributed (IID) environments [16]. In real-world applications, this IID assumption is often violated, as the deployment environment typically differs from training environment to meet practical requirements [55]. Consequently, this leads to a generalization issue [62].

To investigate the generalization issue in DRL, we revisit the pioneering work on DRL-based multi-UAV collision avoidance, specifically SAC+RAE [23]. An intriguing analysis is provided in Figure 1. In SAC+RAE, eight UAVs are simulated during the training phase for collision avoidance experiments. During this stage, each UAV treats the other UAVs as primary obstacles, enabling the model to learn a promising collision avoidance strategy. In the testing phase, we introduce both UAV-shaped obstacles and previously unseen cube-shaped obstacles into the environment. Interestingly, while the UAV-shaped obstacles have little effect on the UAVs' navigation, the cube-shaped obstacles significantly reduce the success rate of navigation. This experiment reveals that the DRL model mistakenly associates obstacle shapes with its learned strategies. As a result, when faced with unseen obstacles (*e.g.*, cube-shaped obstacles), the model is unable to execute an effective strategy, leading to a marked reduction in the generalization ability of DRL models.

Enhancing the generalization capability of DRL models to adapt to unseen scenarios is critical for practical UAV applications. Existing approaches can be broadly categorized into two groups: augmentation-based and regularization-based methods. Augmentation based methods [9, 15, 41] aim to enhance the training environment's observations through data augmentation techniques. However, recent studies [35, 56] indicate that certain data augmentation techniques may reduce sample efficiency, in some cases, lead to model divergence. Regularization-based methods [8, 44, 52] introduce constraints during model training, such as encouraging weight sparsity or smoothness, to enhance generalization and mitigate overfitting. However, both augmentation-based and regularization-based methods rely on manually designed strategies, which are inherently limited in their ability to cover the vast diversity of potential deployment scenarios. As a result, they often yield suboptimal performance when tested in novel or unseen environments.

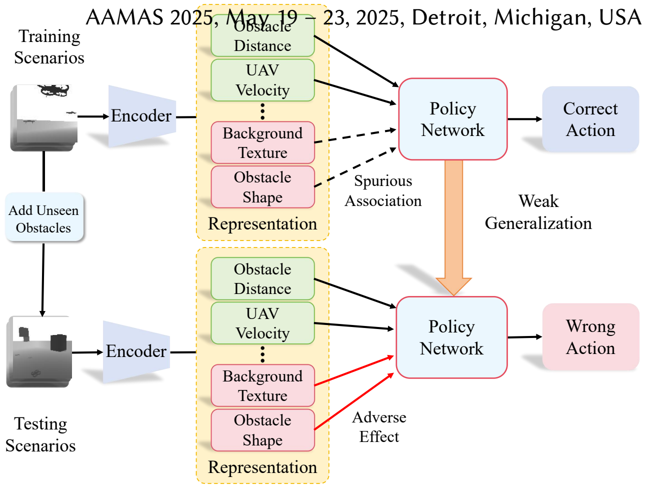


Figure 2: Illustration on the influence of non-causal representation factors. During policy learning, non-causal factors would construct spurious correlations with action prediction. When testing scenarios are different from the training scenarios, these non-causal factors would bring adverse effect and result in wrong actions.

To address the generalization issue in DRL, we analyze the structure of SAC+RAE and identify the root cause of its weak generalization ability as stemming from unstable and error-prone visual representations. As shown in Figure 2, SAC+RAE utilizes a regularized auto-encoder (RAE)[11] to encode input depth images into compact visual representations. The regularized auto-encoder, trained with reconstruction supervision, tends to encode all visual information from the input data, including obstacle distance, obstacle shape, and background texture. Among these encoded features, some are crucial for the collision avoidance task, such as obstacle distance and UAV velocity, which we refer to as *causal factors*. In contrast, other features, such as obstacle shape and background texture, are specific to certain environments and irrelevant to the task, referred to as *non-causal factors*. Without distinguishing between these representation components, passing all of them to the policy network can lead to spurious correlations [10, 53] between non-causal factors and action predictions, ultimately undermining the model's generalization capability. Consequently, when the testing environment changes, such as through the introduction of unseen obstacles, the non-causal factors may negatively impact policy predictions due to spurious correlations, ultimately leading to collisions. This is the primary reason for the weak generalization ability of DRL models in unknown scenarios.

Therefore, effectively filtering out non-causal components from visual representations is essential for improving the generalization ability of DRL. To achieve this, causal representation learning (CRL) [4, 32, 42] offers a promising solution, which has become a key research area in artificial intelligence. It aims to identify causal representation factors by constructing underlying causal structures, thus effectively addressing generalization challenges in out-of-distribution scenes [18, 37]. To address this issue, we first analyze the decision-making process in DRL from perspective of CRL, which describes the relationships among causal features, non-causal features, and predicted actions. As illustrated by causal assumptions in Figure 3, non-causal features function as confounding factors,

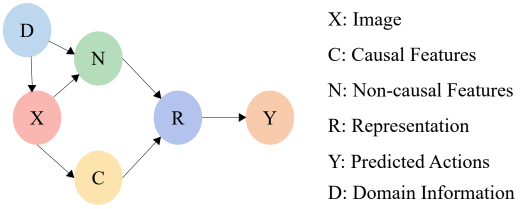


Figure 3: Structural causal model (SCM) for representation learning. Image X is composed of both causal factors S and non-causal factors U, but only the causal factors S have a direct causal impact on the representation learning process.

opening a backdoor path that introduces spurious correlations between non-causal features and predicted actions.

In this work, we design a plug-and-play **C**ausal **F**eature **S**election (CFS) module that can be integrated into the policy network. Based on the common assumption that different channels capture distinct visual factors [10, 25, 59], the CFS module incorporates a differentiable mask to effectively filter out non-causal feature channels. Additionally, we introduce a reward-based guidance and hierarchical consistency constraint to facilitate the identification of causal and non-causal feature components. As a result, only the causal representation factors are passed on for policy learning, explicitly minimizing the influence of non-causal factors and thereby enhancing the model's generalization ability.

As this is the first work to investigate the generalization issue in multi-UAV collision avoidance, we establish a benchmark comprising unseen backgrounds and obstacles to evaluate effectiveness of our method in improving generalization ability. Experimental results show that our method significantly increases the success rate of collision avoidance compared to previous SOTA methods, demonstrating the superiority and effectiveness of our method.

2 RELATED WORK

2.1 DRL-based Collision Avoidance Navigation

Research on DRL-based collision avoidance navigation has gained significant attention. To address the lack of drone simulators that closely resemble real environments, Cetin *et al.* [6] introduce a simulation platform named Airsim [40]. They utilize RGB images as observation signals and propose a discrete space-based deep deterministic policy gradient algorithm for navigation. However, their model is trained and tested exclusively within the same urban environment, limiting its ability to generalize and adapt to unseen environment. Huang *et al.* [22] replaced RGB images with depth images as observation states to better bridge the gap between simulation and reality. They trained a value network using the deep Q-network algorithm in a discrete action space. This method successfully transitioned from simulation to real-world environments by training the network with a pillar matrix. To resolve the stuttering issue caused by discrete action spaces in drone motion, Xue *et al.* [57] propose a vision-based collision avoidance approach using soft actor-critic (SAC) [14] in a continuous action space. By combining SAC with a variational auto-encoder (VAE), the drone can perform collision avoidance tasks in a simulated environment with multiple wall obstacles.

Unlike previous works, our research focuses on addressing the generalization issue in DRL-based collision avoidance navigation. We investigate the relationship between visual representations and weak generalization ability, and propose a novel CFS module to mitigate spurious correlations and enhance generalization.

2.2 Causal Representation Learning

Traditional representation learning methods often rely on the assumption of IID data. However, in real-world applications, it is challenging to ensure that deployment scenarios match the training environment. When the data deviates from the IID assumption, the performance of machine learning algorithms typically degrades significantly. CRL seeks to discover causal factors by constructing underlying causal structures, effectively addressing the generalization issue in out-of-distribution (OOD) scenarios. Lin *et al.* [34] tackled unsupervised video anomaly detection from the perspective of causal inference, using a causal framework to reduce the impact of noisy pseudo-labels on detection outcomes. Liu *et al.* [36] decoupled physical laws, mixed styles, and non-causal features in pedestrian motion patterns to ensure the robustness and reusability in motion trajectory prediction. Huang *et al.* [21] proposed an adaptive reinforcement learning algorithm with a graph model to minimize state representations, capturing domain-specific variations while maintaining shared representations across common domains. This approach achieved effective policy transfer with minimal samples from the target domain. Yang *et al.* [58] enhanced the traditional VAE model with causally structured layers to influence representations, though it required true causal variables as labels.

In this work, we are the first to apply CRL within a DRL-based collision avoidance navigation framework and introduce a CFS module. This module effectively addresses the generalization challenge when UAVs are deployed in unseen environments.

3 APPROACH

3.1 Problem Formulation and DRL Setting

This study aims to equip UAVs with the ability to adapt to unseen environments. The UAVs receive depth images from a front-facing camera and pose information from an inertial measurement unit, resulting in limited environmental observation during the interaction process. Consequently, it is framed as a partially observable Markov decision process (POMDP).

Observation space. At each timestep t, the observation information obtained is $o^t = [o_z^t, o_g^t, o_v^t]$. Here, o_z^t refers to the depth images, which contain distance information. o_g^t denotes the target position information, and o_v^t represents the velocity of the UAV.

Action space. To enhance the diversity and controllability of the UAVs, we employ a continuous action space. At each timestep t, each UAV generates an action command with three degrees of control: $a = [v_x^{cmd}, v_z^{cmd}, v_y^{cmd}]$. Here, v_x^{cmd} represents the forward velocity, v_z^{cmd} denotes the climb velocity, and v_y^{cmd} refers to the steering velocity.

Reward function. Sparse rewards can significantly increase the difficulty of reinforcement learning (RL). To address the UAV collision avoidance problem, this study introduces a non-sparse reward

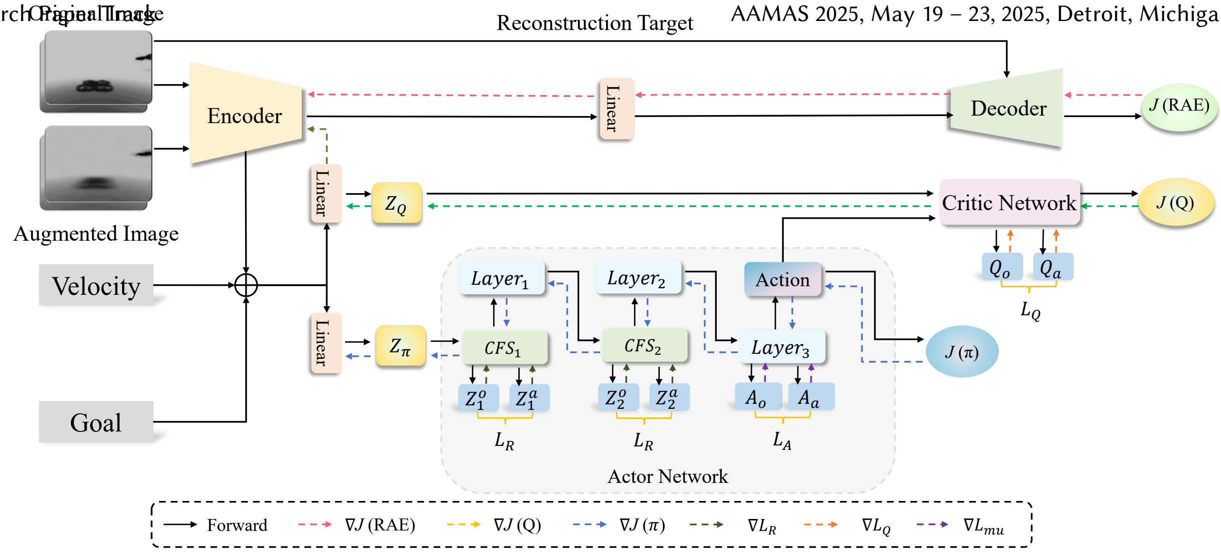


Figure 4: The architecture of our framework for Multi-UAV collision-avoidance. The framework follows the SAC paradigm and uses an regularized auto-encoder for visual representation extraction, which takes depth images, velocity, and relative goal position as input and outputs flight control actions. In the actor network, we insert our design CFS module for feature selection.

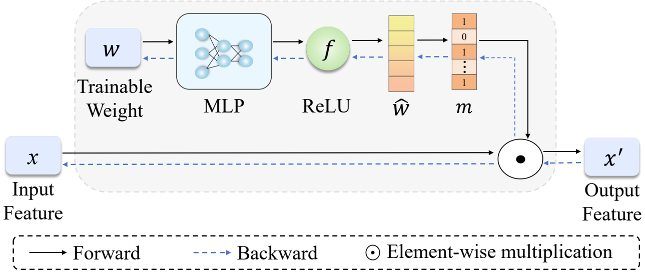


Figure 5: Causal Feature Selection Module. This module transforms a trainable weight into a binary mask for feature selection.

function. The reward function consists of two components: obstacle avoidance reward and arrival reward. Detailed explanations are provided below:

$$r^t = r^t_{goal} + r^t_{avoid} \tag{1}$$

The reward r^t_{goal} is designed to guide UAVs toward their target. If a UAV is within 0.5 meters of its target position at timestep t, it will receive an arrival reward $r_{arrival}$. For the i^{th} UAV, if the distance between its goal g_i and current position p^t_i is smaller than the distance between the goal and previous position p^{t-1}_i, a reward r^t_{goal} will be added, which is scaled by a weighting factor ω_{goal} based on the difference in distances.

$$r^t_{goal} = \begin{cases} r_{arrival} & if \ \|p^t_i - g_i\| < 0.5 \\ \omega_{goal} \cdot (\|p^{t-1}_i - g_i\| - \|p^t_i - g_i\|) & otherwise \end{cases} \tag{2}$$

The reward r^t_{avoid} is designed to encourage each UAV to avoid collisions. A UAV incurs a penalty $r_{collision}$ if it collides with other UAVs or obstacles in the environment. Additionally, a penalty is applied if the minimum distance d^t_{min} in the depth image, falls below the safe distance d_{safe}. This penalty is scaled by a weight factor ω_{avoid} to appropriately adjust the severity of penalty:

$$r^t_{avoid} = \begin{cases} r_{collision} & if \ \ collision \ \ occurs \\ \omega_{avoid} \cdot max(d_{safe} - d^t_{min}, 0) & otherwise \end{cases} \tag{3}$$

3.2 Overview

The framework of our method is illustrated in Figure 4. It follows the SAC paradigm and consists of two main modules: the visual module and policy module (*i.e.*, the actor and critic networks). The visual module is responsible for extracting visual representations from depth images, containing essential visual information for collision avoidance. These visual representations, along with the current velocity and relative goal position, are then fed into the policy module for strategy learning. At each timestep, the policy module generates a set of actions to guide UAVs in collision avoidance.

For visual information extraction, we employ a regularized auto-encoder [11], consisting of an encoder for feature extraction and a decoder for reconstruction. To ensure that the visual representation retains all necessary information, a reconstruction supervision loss $J(RAE)$ is applied to update both the encoder p_ϕ and decoder g_φ:

$$J(RAE) = \mathbb{E}_x \left[\log p_\phi(x|z) + \lambda_z \|z\|^2 + \lambda_\phi \|\phi\|^2\right] \tag{4}$$

For policy learning, the policy module is divided into an actor network and a critic network, following the SAC paradigm. The actor network is designed to generate actions, while the critic network is responsible for evaluating the quality of these actions. During the training phase, the parameters of the actor network are updated using the loss function $J(\pi)$, which can be expressed as follows:

$$J(\pi) = \mathbb{E}_{o\sim\mathcal{B}} \left[D_{KL}(\pi(\cdot|o) \| \mathcal{Q}(o, \cdot))\right] \tag{5}$$

where $\mathcal{Q}(o, \cdot) \propto \exp\{\frac{1}{\alpha}Q(o, \cdot)\}$. The parameters of critic network is updated through loss function $J(Q)$, which can be expressed as

$$J(Q) = \mathbb{E}_{(o,a,r,o')\sim\mathcal{B}} \left[(Q(o, a) - r - \gamma\bar{V}(o'))^2\right] \tag{6}$$

Since the visual representation extracted from visual module tends to encode all information of scene, including domain-specific but task-irrelevant details, spurious associations between visual representation and predicted actions can arise, leading to poor generalization capability. To address this issue, as shown in Figure 4, we propose a novel CFS module. This module can be embedded into the actor network to filter out task-irrelevant information from the representation, improving generalization performance.

3.3 Causal Identifiability Analysis

In this work, we design a novel CFS module to discover causal feature components while filtering out non-causal ones, motivated by the established structural causal model (SCM) shown in Figure 3. In this section, we provide a proof of causal identifiability analysis to offer a solid theoretical foundation for our proposed method.

The framework of autoencoders facilitates efficient learning in deep latent-variable models. However, a key challenge remains: these models often lack guarantees of identifiability. Khemakhem et al. [28] demonstrate that any model relying on an unconditional latent distribution is inherently unidentifiable. Similarly, in the context of this paper, the absence of additional domain information results in both causal and non-causal features being equally represented in the observed depth image data, making it difficult for autoencoders to differentiate between them. To address this issue, Tian et al. [46] show that leveraging multiple source domains with varying data distributions helps identify causal structures, as causal knowledge can be inferred from the changes in these distributions. Therefore, constructing multi-domain data and integrating additional domain information is crucial for ensuring that the model accurately identifies non-causal features.

We adhere to the SCM depicted in Figure 3 and make the following standard assumptions, as discussed in [28, 30, 32]:

(i) Smooth and Positive Density: The probability density function associated with the latent variables is smooth and strictly positive throughout the domains of Z and D. In other words, for every $z \in Z$ and $d \in D$, it is true that $p_{z|d}(z|d) > 0$.

(ii) Conditional Independence: Given the variables d, each latent variable z_i is independent of the remaining latent variables z_j for all $i, j \in \{1, \dots, n\}$ where $i \neq j$. Formally, the logarithm of the conditional density can be expressed as:

$$\log p_{z|d}(z|d) = \sum_{i=1}^{n} \log p(z_i|d). \tag{7}$$

(iii) Linear Independence: For any $z \in Z$, there exist $n_1 + 1$ distinct values of d such that the corresponding n_1 vectors defined by $v(z, d_j) - v(z, d_0)$ are linearly independent. The vector $v(z, d)$ is defined as:

$$v(z,d) = \left(\frac{\partial q_1(z_1, d)}{\partial z_1}, \dots, \frac{\partial q_n(z_n, d)}{\partial z_n} \right). \tag{8}$$

Under these assumptions, the causal and non-causal features are identifiable within their respective subspaces.

Proof Sketch: Initially, we define an invertible transformation function h that maps the true latent variables z to the estimated variables $\hat{z}$. Subsequently, by utilizing the different components of the z domain, which possess domain invariance and domain-specific characteristics, we construct a system of linear equations with full rank, ensuring a unique solution where $\frac{\partial z_i}{\partial \hat{z}_j} = 0$. Given that the Jacobian matrix of the transformation h is invertible, it follows that for each latent variable z_i (where $i \in \{1, \dots, n\}$), there exists a corresponding function h_i such that $z_i = h_i(\hat{z}_i)$.

3.4 Causal Feature Selection

In this work, we design a plug-and-play and lightweight CFS module, which can be embedded into the actor network. The CFS module generates a differentiable binary mask for channel selection, explicitly suppressing the influence of non-causal channels. As shown in Figure 5, to perform causal feature selection, we multiply the input feature $x \in \mathbb{R}^C$ by the generated binary mask m as follows:

$$x^{'} = x \odot m, \tag{9}$$

where C is the number of feature channels. In this way, the binary mask m explicitly activates causal feature channels and deactivates non-causal ones. The modulated feature $x^{'}$ then eliminates the influence of non-causal factors, improving the generalization ability.

The core of the CFS module lies in generating a differentiable binary mask m, which can be integrated into the actor network and trained in an end-to-end manner. Specifically, we design a differentiable mask generation process. Given an intermediate vector $x \in \mathbb{R}^C$ within the actor network, we assign a trainable weight $w \in \mathbb{R}^C$, and the corresponding mask m is determined as follows:

$$\hat{w} = ReLU(MLP(w)), \tag{10}$$

$$m = \frac{\hat{w}^2}{\hat{w}^2 + \epsilon}, \tag{11}$$

where ϵ is an infinitesimally small positive number. We first generate an intermediate variable $\hat{w}$ by transforming the input weight w through a small MLP, followed by a ReLU activation. For each channel, the mask m contains a value of 0 if $\hat{w}$ is 0; otherwise, it contains 1, as ϵ is extremely small. This process transforms the trainable weight w into a differentiable binary mask m without the need for manual threshold design.

3.5 Hierarchical Consistency Constraint

After embedding the CFS module into actor network, a crucial step is to guide the CFS module to function as intended *i.e.*, filtering out non-causal channels while retaining the causal ones.

To identify the causal feature channels by constructing multi-domain data as demonstrated in Sec. 3.3, we design a simple yet effective strategy. Specifically, we apply basic data augmentation operations (*e.g.*, random noise, motion blurring) to the original input images I^o, then feed the augmented images I^a into the pipeline for feature extraction and action generation. Since these augmentations affect only low-level statistical information while preserving task-relevant data, such as obstacle distance. We assume that the causal feature channels should remain consistent between the two images, while non-causal channels may exhibit significant changes. Based on this assumption, we design a consistency constraint on the visual representations processed by the CFS module to guide it in filtering out non-causal channels. To further aid feature selection, we apply consistency constraints to the subsequent action and Q-value predictions.

As shown in Figure 4, we implement a hierarchical consistency constraint to guide the CFS module, applied at the representation, action, and Q-value levels. Specifically, the first constraint is the representation consistency loss L_R, which aims to ensure non-causal

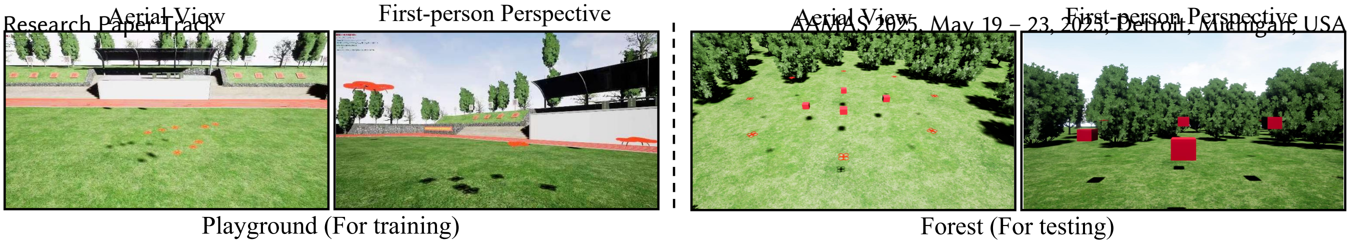


Figure 6: Simulation scenarios for model training and testing. Specifically, playground scenario is used for model training, while canyon, snow mountain, temple and forest scenarios are used for testing.

TABLE 1: Performance comparison with existing methods.

Method	SSR (%)	ISR (%)	SPL (%)	Extra Distance (m)	Average Speed (m/s)
SAC+RAE	29.6	87.7	68.7	10.882/2.771	0.448/0.093
+ AutoAugment [15]	53.4	90.6	85.0	**1.587/0.567**	**1.143/0.081**
+ DrAC [41]	50.0	90.9	74.9	5.180/1.384	1.073/0.144
+ $L1$ Norm [8, 39]	69.6	92.9	84.1	2.527/0.769	1.011/0.102
+ $L2$ Norm [44, 45]	58.4	93.8	79.1	4.564/1.807	0.573/0.092
+ SE [19]	29.8	85.6	65.1	10.498/4.641	0.5944/0.112
+ CBAM [54]	49.6	88.3	78.3	4.406/3.835	0.499/0.091
+ MaDi [13]	52.8	92.2	**85.3**	2.638/1.345	0.720/0.062
+ Our CFS	**96.0**	**98.6**	72.8	11.773/1.029	0.684/0.030

factors that do not affect the representation after augmentation, processed by the CFS modules CFS_1 and CFS_2:

$$L_R = \frac{1}{N}\sum_{i=1}^{N} \|Z_j^o - Z_j^a\|^2, (j = 1, 2) \tag{12}$$

where Z_j^o and Z_j^a are representations extracted from the original and augmented images by the jth CFS module. The second constraint is the Q-value consistency loss, which identifies non-causal factors that do not influence the Q-value change after augmentation:

$$L_Q = \frac{1}{N}\sum_{i=1}^{N} \|Q_o - Q_a\|^2 \tag{13}$$

where Q_o and Q_a represent the predicted Q-values from the original and augmented images. The third constraint is the action consistency loss, which aims to ensure that non-causal factors do not affect the predicted action after augmentation:

$$L_A = \frac{1}{N}\sum_{i=1}^{N} \|A_o - A_a\|^2 \tag{14}$$

where A_o and A_a denote the mean of predicted action distributions from the original and augmented images.

To prevent trivial solutions, such as filtering out nearly all feature channels, we provide additional supervision to ensure that the processed representation remains effective for collision avoidance. Specifically, we use the actor loss $J(\pi)$ to update the trainable weights in the CFS module when generating binary masks (*i.e.*, the pre-set weights and the small MLP). If the CFS module mistakenly filters out essential feature channels, the actor loss $J(\pi)$ will increase, encouraging the CFS module to retain the causal channels.

4 EXPERIMENT AND RESULTS

4.1 Experiment Metrics and Scenarios

4.1.1 Performance Metrics. Following the previous work [23], we use the following performance metrics for evaluation:

- Swarm Success Rate (SSR): The ratio of the number of times all UAVs reach the target point to the total number of trials is a critical metric, particularly for practical applications involving UAV cluster collaboration. In such scenarios, a mission is typically considered complete only when all drones successfully reach their targets.
- Individual Success Rate (ISR): The percentage of UAVs that successfully reach target positions without collisions in limited time.
- Success weighted by Path Length (SPL): The proportion of UAVs that successfully reach target positions, accounting for the path length, and calculated across the test scenario with N UAVs and M episodes as follows:

$$SPL = \frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M} S_{i,j}\frac{l_{i,j}}{max(p_{i,j}, l_{i,j})} \tag{15}$$

where $l_{i,j}$ denotes the shortest-path distance from the i^{th} UAV's initial position to the target position in episode j, $p_{i,j}$ represents the actual path length traversed by the i^{th} UAV, and $S_{i,j}$ is a binary indicator denoting success or failure in that episode.

- Extra Distance: The average additional distance traveled by UAVs relative to the straight-line distance between the initial and goal positions.
- Average Speed: The average speed of all UAVs during testing, calculated as the mean speed across all episodes.

TABLE 2: Performance comparison under different backgrounds.

Scene	Seen/Unseen	Method	SSR (%)	ISR (%)	SPL (%)	Extra Distance (m)	Average Speed (m/s)
Playground	Seen	SAC+RAE	87.5	96.4	**76.4**	**8.573/2.839**	0.521/0.072
		Our method	**97.1** (↑ 9.6)	**99.6** (↑ 3.2)	73.7	11.370/0.756	**0.789/0.035**
Forest	Unseen	SAC+RAE	29.6	87.7	68.7	**10.882/4.771**	0.448/0.093
		Our method	**96.0** (↑ 66.4)	**98.6** (↑ 10.9)	**72.8**	11.773/1.029	**0.684/0.030**
Snow Mountain	Unseen	SAC+RAE	65.1	92.7	**72.4**	**9.113/2.785**	0.503/0.064
		Our method	**94.4** (↑ 29.3)	**99.3** (↑ 6.6)	70.5	13.069/1.010	**0.665/0.045**
Canyon	Unseen	SAC+RAE	52.3	91.2	67.8	**11.657/5.383**	0.454/0.095
		Our method	**92.8** (↑ 40.5)	**99.1** (↑ 7.9)	**70.4**	13.068/1.029	**0.664/0.042**

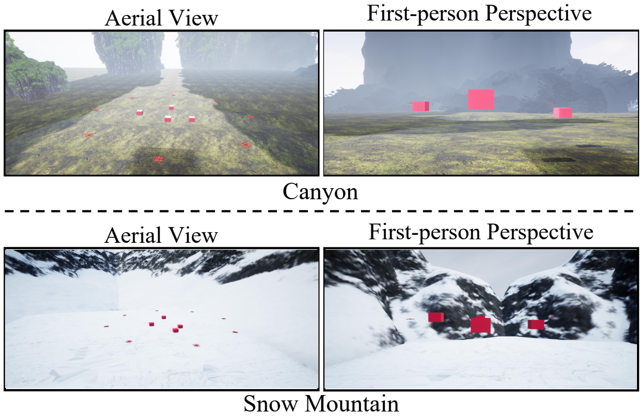


Figure 7: More Evaluation Scenes. We design two additional typical scenes (*i.e.*, snow mountain and canyon) for evaluation. Best viewed in zoom and color.

TABLE 3: Scalability analysis experiments. ISR and SSR are adopted as the evaluation metric.

	Num	SAC+RAE		+ Our CFS	
		SSR (%)	ISR (%)	SSR (%)	ISR (%)
Obstacle	4	29.6	87.7	**96.0** (↑ 66.4)	**98.6** (↑ 10.9)
	6	22.4	82.6	**91.8** (↑ 69.4)	**93.1** (↑ 10.7)
	8	18.4	80.6	**90.5 (**↑ 72.1**)**	**92.9** (↑ 12.3)
	10	16.6	78.9	**87.6** (↑ 71.0)	**89.7** (↑ 10.8)
UAV	8	29.6	87.6	**96.0** (↑ 66.4)	**98.6** (↑ 11.0)
	10	20.0	84.0	**91.2** (↑ 71.2)	**93.4** (↑ 9.4)
	12	11.8	82.4	**88.4** (↑ 76.6)	**90.6** (↑ 8.2)
	14	2.5	80.6	**85.9** (↑ 83.4)	**88.5** (↑ 7.9)

4.1.2 Scenarios. To investigate the generalization ability of DRL-based multi-UAV collision avoidance systems, we design several typical scenarios by varying backgrounds and introducing unseen obstacles, as illustrated in Figure 6.

For training, we use a playground scenario with no obstacles. At each training round, the initial and target positions of UAVs are randomly generated. The UAVs' flight range is limited to a $(16 \times 16 \times 4)$ space, allowing them to fully explore the high-dimensional observation space and improving the robustness of learned strategies.

For testing, we modify the scenarios in three key aspects to thoroughly evaluate the generalization capability of DRL models. Specifically, we change the background from a playground to a completely different forest scene and introduce several unseen obstacles, creating a typical out-of-distribution (OOD) scenario for evaluating generalization ability. Additionally, we change the initialization pattern of the UAVs from a random to a circular configuration, a more challenging setting. In this case, the UAVs' initial positions are uniformly distributed within a circular area at the same height, and their target positions are placed on the opposite side of the circular area from their starting points.

4.2 Performance Comparison

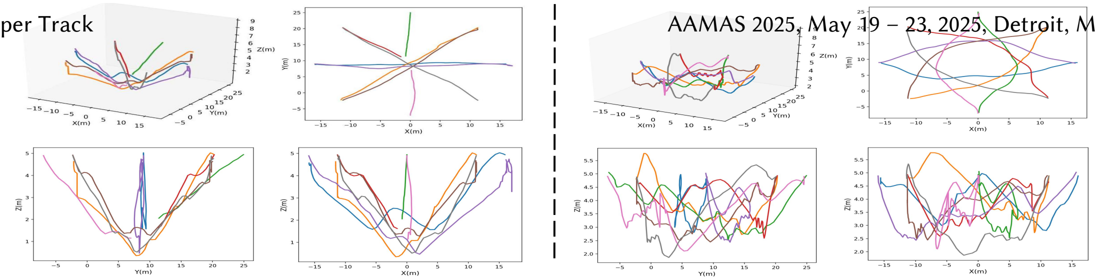
To clearly evaluate the generalization ability of our method, we build upon previous SOTA method, SAC+RAE [23], and compare our proposed CFS method with other existing approaches that address the generalization issue in DRL models. Specifically, we compare our method with two groups: augmentation-based methods (*i.e.*, AutoAugment [15] and DrAC [41]) and regularization-based methods ($L1$ Norm [8, 39] and $L2$ Norm [44, 45]). Additionally, since our CFS module can be interpreted as a novel attention mechanism, we also compare it with several popular attention-based methods (*i.e.*, SE [19], CBAM [54], and MaDi [13]). For fairness, our proposed CFS module and other methods are built on the same baseline model (SAC+RAE) with identical training and testing configurations.

As shown in TABLE 1, our CFS significantly improves the navigation success rate compared to other methods, demonstrating its effectiveness in enhancing the generalization capability of DRL models. While CFS results in a slightly longer planned flight path and slower speed, this is due to UAVs performing more collision avoidance maneuvers, ultimately leading to a higher success rate.

4.3 Ablation Study

In this subsection, we conduct extensive ablation experiments to further demonstrate the effectiveness of our proposed module.

4.3.1 More Evaluation Scenes. To further evaluate the adaptability of our method to unseen backgrounds, we design two additional typical scenes (*i.e.*, snow mountain and canyon) for evaluation, as shown in Figure 7. These scene designs are inspired by practical UAV applications, such as wildlife monitoring[12] and reconnaissance and surveillance [61]. As indicated in TABLE 2, our CFS consistently outperforms the baseline model across different scenes, including both seen and unseen environments.



SAC+RAE Our Method

Figure 8: Visualization of UAV trajectories in perspective and three-view drawings. The trajectories of different UAVs are represented in distinct colors. Best viewed in color.

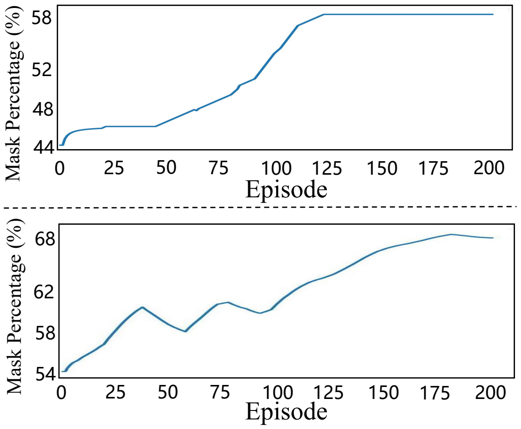


Figure 9: Visualization on masked percentage in our CFS module during training. During training, the percentage of zero value in the binary mask consistently increases and finally converges in two CFS modules.

4.3.2 *Scalability.* To assess the scalability of our method, we adjust the testing scenario by varying the number of UAVs and obstacles independently. As shown in TABLE 3, our method consistently brings significant performance improvements over the baseline model, demonstrating its superior robustness and scalability.

4.3.3 *Casual Feature Selection.* The key to our method's improvement in generalization ability lies in the causal feature selection mechanism, which utilizes a differentiable binary mask. To clearly illustrate the feature selection process, we analyze the two embedded CFS modules and visualize the percentage of zero values in the generated binary mask during training, as shown in Figure 9. Throughout the training, the percentage of zero values consistently increases and eventually stabilizes at around 58% and 68%, indicating that the CFS module effectively identifies and eliminates a substantial portion of feature channels.

4.3.4 *Hierarchical Consistency Constraint.* One of the key components of the CFS module is the hierarchical consistency constraint (*i.e.*, L_R, L_Q, and L_A), which effectively guides the CFS module in identifying causal feature channels while filtering out non-causal ones. To assess the effectiveness and necessity of each consistency constraint, we conduct an ablation study. As shown in TABLE 4, all constraint components contribute to improving the generalization capability of DRL models, with their combination yielding the best performance. Notably, the representation consistency loss L_R introduced in CFS demonstrates the most significant improvement compared to the other constraints.

TABLE 4: Ablation study on the hierarchical consistency constraints.

L_Q	L_A	L_R	SSR (%)	ISR (%)
✓			0	2.1
	✓		5.6	64.8
		✓	15.2	71.3
✓	✓		46.7	81.4
	✓	✓	60.2	86.3
✓		✓	65.1	87.5
✓	✓	✓	**96.0**	**98.6**

4.3.5 *Trajectory Visualization.* To validate the quality of the planned paths, we provide a visualization of UAV trajectories. As shown in Figure 8, our method achieves smoother and more complete flight trajectories, whereas SAC+RAE results in the collision of 4 UAVs in unseen scenarios. These results demonstrate that our method generates a more robust and effective flight strategy.

5 CONCLUSIONS

In this paper, we propose a robust policy learning approach by introducing a CFS module to enhance the generalization ability of DRL techniques in unseen scenarios. The CFS module can be integrated into the policy network, effectively filtering out non-causal components during representation learning and thereby reducing the influence of spurious correlations. To demonstrate its generalization capability, we conduct extensive experiments across various testing scenarios, including altered backgrounds and the introduction of unseen obstacles. The experimental results show that our method significantly outperforms the previous state-of-the-art, demonstrating excellent generalization ability and robustness.

ACKNOWLEDGMENTS

This work is supported in part by the National Science and Technology Major Project (2021ZD0111502), the National Natural Science Foundation of China (62176147, 62406186, 62476163), the Science and Technology Planning Project of Guangdong Province of China (2022A1515110660, 2022A1515110566, 2023A1515011574), and the STU Scientific Research Foundation for Talents (NTF22030, NTF21001, NTF21052).

REFERENCES

[1] Shubhani Aggarwal and Neeraj Kumar. 2020. Path planning techniques for unmanned aerial vehicles: A review, solutions, and challenges. *Computer Communications* 149 (2020), 270–299.

[2] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine* 34, 6 (2017), 26–38.

[3] Mitch Bryson and Salah Sukkarieh. 2007. Building a Robust Implementation of Bearing-only Inertial SLAM for a UAV. *Journal of Field Robotics* 24, 1-2 (2007), 113–143.

[4] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. 2019. Learning disentangled semantic representation for domain adaptation. In *IJCAI*, Vol. 2019. NIH Public Access, Macau, China, 2060.

[5] Yuwei Cai, Guijie Zhu, Huaxing Huang, Zhaojun Wang, Zhun Fan, Wenji Li, Ze Shi, and Weibo Ning. 2021. The behavior design of swarm robots based on a simplified gene regulatory network in communication-free environments. In *Proc. Int. Workshop Adv. Comput. Intell. Intell. Inform.* IEEE, Beijing, China, 48–55.

[6] Ender Cetin, Cristina Barrado, Guillem Muñoz, Miquel Macias, and Enric Pastor. 2019. Drone navigation and avoidance of obstacles through deep reinforcement learning. In *DASC*. IEEE, San Diego, CA, USA, 1–7.

[7] Yuhui Chen, Haoran Li, and Dongbin Zhao. 2024. Boosting Continuous Control with Consistency Policy. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. IEEE, Auckland, New Zealand, 335–344.

[8] Yikun Cheng, Pan Zhao, Fanxin Wang, Daniel J Block, and Naira Hovakimyan. 2022. Improving the Robustness of Reinforcement Learning Policies With ℓ_1 Adaptive Control. *IEEE Robotics and Automation Letters* 7, 3 (2022), 6574–6581.

[9] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, Long Beach, CA, USA, 113–123.

[10] Pim De Haan, Dinesh Jayaraman, and Sergey Levine. 2019. Causal confusion in imitation learning. *NeurIPS* 32 (2019), 11698–11709.

[11] Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Scholkopf. 2020. From Variational to Deterministic Autoencoders. In *International Conference on Learning Representations*. OpenReview. net, Millennium Hall, Addis Ababa, 107877.

[12] Luis F Gonzalez, Glen A Montes, Eduard Puig, Sandra Johnson, Kerrie Mengersen, and Kevin J Gaston. 2016. Unmanned aerial vehicles (UAVs) and artificial intelligence revolutionizing wildlife monitoring and conservation. *Sensors* 16, 1 (2016), 97.

[13] Bram Grooten, Tristan Tomilin, Gautham Vasan, Matthew E Taylor, A Rupam Mahmood, Meng Fang, Mykola Pechenizkiy, and Decebal Constantin Mocanu. 2024. MaDi: Learning to Mask Distractions for Generalization in Visual Deep Reinforcement Learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Auckland, New Zealand, 733–742.

[14] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR, Stockholm, Sweden, 1861–1870.

[15] Nicklas Hansen and Xiaolong Wang. 2021. Generalization in reinforcement learning by soft data augmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Xi'an, China, 13611–13617.

[16] Bruce Hoadley. 1971. Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of mathematical statistics* 42, 6 (1971), 1977–1991.

[17] Stefan Hrabar. 2011. Reactive obstacle avoidance for rotorcraft uavs. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, San Francisco, USA, 4967–4974.

[18] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. 2020. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Seattle, WA, USA, 10951–10960.

[19] Jie Hu, Li Shen, and Gang Sun. 2020. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Vol. 42. IEEE Computer Society, Seattle, USA, 2011–2023.

[20] Kaiyu Hu, Huanlin Li, Jiafan Zhuang, Zhifeng Hao, and Zhun Fan. 2023. Efficient Focus Autoencoders for Fast Autonomous Flight in Intricate Wild Scenarios. *Drones* 7, 10 (2023), 609.

[21] Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, and Kun Zhang. 2021. AdaRL: What, Where, and How to Adapt in Transfer Reinforcement Learning. In *International Conference on Learning Representations*. OpenReview. net, Virtual, arXiv–2107.

[22] Haitao Huang, Jingjing Gu, Qiuhong Wang, and Yi Zhuang. 2019. An autonomous UAV navigation system for unknown flight environment. In *International Conference on Mobile Ad-Hoc and Sensor Networks*. IEEE, Shenzhen, China, 63–68.

[23] Huaxing Huang, Guijie Zhu, Zhun Fan, Hao Zhai, Yuwei Cai, Ze Shi, Zhaohui Dong, and Zhifeng Hao. 2022. Vision-based Distributed Multi-UAV Collision Avoidance via Deep Reinforcement Learning for Navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Kyoto, Japan, 13745–13752.

[24] Sunan Huang, Rodney Swee Huat Teo, and Kok Kiong Tan. 2019. Collision avoidance of multi unmanned aerial vehicles: A review. *Annual Reviews in Control* 48 (2019), 147–164.

[25] Qirui Ji, Jiangmeng Li, Jie Hu, Rui Wang, Changwen Zheng, and Fanjiang Xu. 2024. Rethinking dimensional rationale in graph contrastive learning from causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. AAAI Press, Vancouver,Canada, 12810–12820.

[26] Chanyoung Ju and Hyoung Il Son. 2019. Modeling and control of heterogeneous agricultural field robots based on Ramadge–Wonham theory. *IEEE Robotics and Automation Letters* 5, 1 (2019), 48–55.

[27] Alif Ridzuan Khairuddin, Mohamad Shukor Talib, and Habibollah Haron. 2015. Review on simultaneous localization and mapping (SLAM). In *ICCSCE*. IEEE, Penang, Malaysia, 85–90.

[28] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*. PMLR, San Diego, USA, 2207–2217.

[29] San Kim, Donggeun Kim, Siheon Jeong, Ji-Wan Ham, Jae-Kyung Lee, and Ki-Yong Oh. 2020. Fault diagnosis of power transmission lines using a UAV-mounted smart inspection system. *IEEE access* 8 (2020), 149999–150009.

[30] Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. 2022. Partial disentanglement for domain adaptation. In *International conference on machine learning*. PMLR, Virtual, 11455–11472.

[31] Rainer Kümmerle, Bastian Steder, Christian Dornhege, Alexander Kleiner, Giorgio Grisetti, and Wolfram Burgard. 2011. Large scale graph-based SLAM using aerial images as prior information. *Autonomous Robots* 30 (2011), 25–39.

[32] Zijian Li, Ruichu Cai, Guangyi Chen, Boyang Sun, Zhifeng Hao, and Kun Zhang. 2024. Subspace Identification for Multi-Source Domain Adaptation. *NeurIPS* 36 (2024), 34504–34518.

[33] Zhenhua Li, Shuo Zhang, Xinye Cai, Qingfu Zhang, Xiaomin Zhu, Zhun Fan, and Xiuyi Jia. 2021. Noisy optimization by evolution strategies with online population size learning. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 52 (2021), 5816–5828.

[34] Xiangru Lin, Yuyang Chen, Guanbin Li, and Yizhou Yu. 2022. A causal inference look at unsupervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. Association for the Advancement of Artificial Intelligence (AAAI), Vancouver, Canada, 1620–1629.

[35] Siao Liu, Zhaoyu Chen, Yang Liu, Yuzheng Wang, Dingkang Yang, Zhile Zhao, Ziqing Zhou, Xie Yi, Wei Li, Wenqiang Zhang, et al. 2023. Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Paris, France, 23436–23446.

[36] Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. 2022. Towards robust and adaptive motion forecasting: A causal representation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, New Orleans, LA, USA, 17060–17071.

[37] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. 2021. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*. OpenReview. net, Virtual, 2060.

[38] Yuncheng Lu, Zhucun Xue, Gui-Song Xia, and Liangpei Zhang. 2018. A survey on vision-based UAV navigation. *Geo-spatial information science* 21, 1 (2018), 21–32.

[39] Jun Ma, Liming Yang, and Qun Sun. 2020. Capped L1-norm distance metric-based fast robust twin bounded support vector machine. *Neurocomputing* 412 (2020), 295–311.

[40] Ratnesh Madaan, Nicholas Gyde, Sai Vemprala, Matthew Brown, Keiko Nagami, Tim Taubner, Eric Cristofalo, Davide Scaramuzza, Mac Schwager, and Ashish Kapoor. 2020. Airsim drone racing lab. In *Neurips 2019 competition and demonstration track*. PMLR, Vancouver, Canada, 177–191.

[41] Roberta Raileanu, Maxwell Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2021. Automatic data augmentation for generalization in reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 5402–5415.

[42] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proc. IEEE* 109, 5 (2021), 612–634.

[43] Javad Shahmoradi, Elaheh Talebi, Pedram Roghanchi, and Mostafa Hassanalian. 2020. A comprehensive review of applications of drone technology in the mining industry. *Drones* 4, 3 (2020), 34.

[44] Guang Shi, Jiangshe Zhang, Huirong Li, and Changpeng Wang. 2019. Enhance the performance of deep neural networks via L2 regularization on the input of activations. *Neural Processing Letters* 50 (2019), 57–75.

[45] Sahil Singla, Surbhi Singla, and Soheil Feizi. 2021. Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100. *arXiv preprint arXiv:2108.04062* 35 (2021), arXiv–2108.

[46] J Tian and J Pearl. 2001. *Causal discovery from changes: a bayesian approach, ucla cognitive systems laboratory.* Technical Report. Technical Report. 512–521 pages.

[47] Yulun Tian, Katherine Liu, Kyel Ok, Loc Tran, Danette Allen, Nicholas Roy, and Jonathan P How. 2020. Search and rescue under the forest canopy using multiple UAVs. *The International Journal of Robotics Research* 39, 10-11 (2020), 1201–1221.

[48] Teodor Tomic, Korbinian Schmid, Philipp Lutz, Andreas Domel, Michael Kassecker, Elmar Mair, Iris Lynne Grixa, Felix Ruess, Michael Suppa, and Darius Burschka. 2012. Toward a fully autonomous UAV: Research platform for indoor and outdoor urban search and rescue. *IEEE robotics & automation magazine* 19, 3 (2012), 46–56.

[49] Juan-Carlos Trujillo, Rodrigo Munguia, Edmundo Guerra, and Antoni Grau. 2018. Cooperative monocular-based SLAM for multi-UAV systems in GPS-denied environments. *Sensors* 18, 5 (2018), 1351.

[50] Dimosthenis C Tsouros, Stamatia Bibi, and Panagiotis G Sarigiannidis. 2019. A review on UAV-based applications for precision agriculture. *Information* 10, 11 (2019), 349.

[51] Olga Vysotska and Cyrill Stachniss. 2017. Improving SLAM by exploiting building information from publicly available maps and localization priors. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science* 85 (2017), 53–65.

[52] Kaixin Wang, Bingyi Kang, Jie Shao, and Jiashi Feng. 2020. Improving generalization in reinforcement learning with mixture regularization. *Advances in Neural Information Processing Systems* 33 (2020), 7968–7978.

[53] Andrew Ward. 2013. Spurious correlations and causal inferences. *Erkenntnis* 78 (2013), 699–712.

[54] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. European Conference on Computer Vision, Munich, Germany, 3–19.

[55] Huici Wu, Xiaofeng Tao, Ning Zhang, and Xuemin Shen. 2018. Cooperative UAV cluster-assisted terrestrial cellular networks for ubiquitous coverage. *IEEE Journal on Selected Areas in Communications* 36, 9 (2018), 2045–2058.

[56] Zhenghua Xu, Shengxin Wang, Gang Xu, Yunxin Liu, Miao Yu, Hongwei Zhang, Thomas Lukasiewicz, and Junhua Gu. 2024. Automatic data augmentation for medical image segmentation using Adaptive Sequence-length based Deep Reinforcement Learning. *Computers in Biology and Medicine* 169 (2024), 107877.

[57] Zhihan Xue and Tad Gonsalves. 2021. Vision based drone obstacle avoidance by deep reinforcement learning. *AI* 2, 3 (2021), 366–380.

[58] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. 2021. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, Nashville, TN, USA, 9588–9597.

[59] Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. 2023. Multi-View Causal Representation Learning with Partial Observability. In *The Twelfth International Conference on Learning Representations*. OpenReview. net, Vienna, Austria, arXiv–2311.

[60] Yutong Yuan, Xiaomin Zhu, Zhun Fan, Li Ma, Ji Ouyang, and Ji Wang. 2021. Online Planning-based Gene Regulatory Network for Swarm in Constrained Environment. In *2021 7th International Conference on Big Data and Information Analytics (BigDIA)*. IEEE, Chongqing, China, 464–471.

[61] Jian Zhang and Yang Zhang. 2020. A method for UAV reconnaissance and surveillance in complex environments. In *2020 6th International Conference on Control, Automation and Robotics (ICCAR)*. IEEE, Singapore, 482–485.

[62] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4396–4415.