

XAI-Guided Fourier Domain Generalisation for Pure-Simulation Sim-to-Real Deep Reinforcement Learning

Wei Yao, Chuyao Wang*, Ziwei Wang, and Nabil Aouf

City St George's, University of London

London, United Kingdom

Email: wei.yao, chuyao.wang.2, ziwei.wang.3, nabil.aouf@city.ac.uk,

Abstract—Deep reinforcement learning has shown strong potential for vision-based autonomous driving, but policies trained in simulation often suffer severe performance degradation when deployed in the real world. Existing sim-to-real methods commonly address this gap through image-transfer-based domain adaptation or policy-agnostic domain randomisation. However, the former requires laborious acquisition of target-domain data and is often limited to the deployment domain, whereas the latter introduces unguided perturbations that may not reflect the underlying sim-to-real discrepancy, leading to suboptimal transfer performance. To overcome these limitations, we propose XFD, an explainable artificial intelligence (XAI)-guided Fourier domain generalisation framework for pure-simulation sim-to-real reinforcement learning. The proposed framework extracts explainability attribution maps from simulated observations, and converts these attributions into policy-conditioned Fourier-domain interventions. By perturbing appearance-sensitive spectral components while preserving task-defining spatial structures, XFD constructs robustness-enhancing visual domains entirely within simulation. Subsequently, the proposed robust policy optimisation guides the policy towards attending to task-critical features while suppressing sensitivity to irrelevant visual information. Extensive experiments in simulation and zero-shot real-world deployment demonstrate that XFD consistently improves traveled distance and reward over standard policy and randomly domain-randomised baseline, especially on unseen maps with cluttered backgrounds. Attribution-based analysis further proves that the proposed policy preserves more consistent task-relevant visual evidence across simulation and real-world observations.

Index Terms—Autonomous driving, deep reinforcement learning, explainable artificial intelligence, sim-to-real transfer, visual domain generalisation.

I. INTRODUCTION

AUTONOMOUS driving is a closed-loop decision-making problem in which perception and control must remain reliable under uncertain and changing environments. Visual scene understanding and deep reinforcement learning have shown strong potential for autonomous driving, particularly when high-dimensional visual observations must be interpreted and mapped to low-level control actions [1]–[4]. However, training such policies directly on physical robots [5] is often unsafe, sample-inefficient, and difficult to reproduce. Simulation therefore provides a scalable and controllable platform for policy learning. Nevertheless, policies optimised in simulation often suffer performance degradation after deployment, because real-world visual appearance, camera response, vehicle

dynamics, and environmental statistics may differ from those observed during training [6]. This sim-to-real gap remains a central obstacle to robust learning-based autonomous navigation [7]–[9].

Sim-to-real transfer requires sufficient resemblance between the training and deployment environments in visual appearance, scene statistics, sensor response, and robot dynamics [10]. Yet constructing a target-faithful simulator is rarely feasible, since real deployment involves uncontrolled illumination, texture variation, sensing noise, and hardware-dependent dynamics [11]. Rather than requiring an exact reconstruction of the target environment, domain randomisation relaxes this requirement by training the policy over a distribution of randomised simulation domains. Existing randomisation strategies can be broadly categorised into target-informed and target-free settings. Target-informed methods exploit target-domain evidence, including real images [12], real-world rollouts [13], or domain adaptation [14], to calibrate the randomised distribution toward the deployment domain. Target-free methods instead assume no access to target-domain data and enlarge visual, physical, or dynamical variability in simulation, aiming to include the real domain within the induced training support [9], [15]. Both settings seek to improve the overlap between training and deployment distributions, but differ in whether this overlap is guided by target-domain evidence or approximated through source-domain diversity.

The above resemblance-driven and randomisation-based strategies are commonly implemented at the image or pixel level, either by aligning visual appearance with the target domain or by perturbing simulator-induced visual factors. Although such pixel-level interventions can reduce apparent domain discrepancy, they entangle style-dependent appearance with task-defining spatial structure and provide limited control over which visual changes affect closed-loop behavior. Frequency-domain adaptation provides a more structured way to control visual shifts. Spectral amplitude is closely related to style-dependent factors such as texture, illumination, and color statistics, whereas phase information better preserves geometry and spatial structure [16], [17]. However, existing Fourier-based methods usually borrow target-domain spectra or mix spectra across observed domains [18], and the generated shifts are not explicitly linked to closed-loop policy behavior. To make spectral interventions policy-sensitive, it is necessary to identify which visual evidence contributes to the action decision. Attribution methods such as Grad-CAM

*Corresponding author

and DeepLIFT provide such a mechanism by revealing the visual evidence behind model outputs [19], [20]. Nevertheless, these explanations are still mainly used for post-hoc analysis [21], and explanation-guided learning has rarely been exploited to construct simulation interventions for sim-to-real reinforcement learning [22]. As a result, existing methods remain limited in connecting structured visual shifts with the policy-relevant evidence that determines closed-loop behavior.

This paper proposes XFD, an XAI-guided Fourier domain generalisation framework for target-free sim-to-real reinforcement learning. The framework replaces policy-agnostic randomisation with policy-sensitive domain intervention: a frozen teacher policy first provides attribution maps that identify action-relevant visual evidence, and these maps are then used to modulate Fourier-domain perturbations generated entirely within simulation. Since training under target-free visual domain shifts can destabilise policy optimisation, the teacher is also used as an auxiliary stabilising reference [23], rather than as an imitation target or performance upper bound. The student policy remains optimised by the reinforcement learning objective under these policy-conditioned domains, enabling robust policy improvement while preserving task behavior. Soft Actor-Critic (SAC) [8] is adopted as the baseline reinforcement learning algorithm.

The main contributions of this work are summarised as follows:

- We propose a DeepLIFT-based XAI-guided Fourier domain intervention mechanism for visual generalisation. The proposed approach translates policy attributions into structured spectral perturbations that modify deployment-sensitive appearance statistics while preserving task-defining spatial structures, enabling robustness-enhancing visual domain generalisation entirely within simulation.
- We introduce a target-domain-free robust policy optimisation strategy within a teacher-student framework. The teacher provides attribution guidance and training stabilisation rather than imitation targets, enabling the student policy to surpass the teacher in task performance while learning more focused and task-relevant representations across visual domains.
- We conduct extensive evaluations in both simulation and zero-shot real-world deployment using the Duckietown platform. Results on lane-following and obstacle-avoidance tasks demonstrate that XFD consistently improves sim-to-real transfer performance on unseen maps with substantial background clutter, achieving robust generalisation through pure simulation training without requiring target-domain data acquisition or real-world fine-tuning.

II. RELATED WORK

A. Sim-to-Real Visual Domain Generalisation

Sim-to-real reinforcement learning has been widely studied to reduce the cost and risk of training visual control policies directly on physical robots. A common line of work narrows the sim-to-real gap by incorporating target-domain information. Bousmalis et al. [14] combined randomised simulation with

domain adaptation for robotic grasping from monocular RGB images, while Chebotar et al. [13] proposed SimOpt to adapt simulator parameter distributions using a small number of real-world rollouts. These methods improve transfer by aligning simulation with the deployment domain, but they still rely on real observations, real interactions, or target-domain statistics. Such dependence increases data-collection cost and may restrict transferability to the deployment conditions covered by the collected target data.

Visual appearance transfer addresses sim-to-real transfer from the perspective of image-domain alignment. Pixel-domain methods directly transform synthetic images toward a target visual style. Shrivastava et al. [24] proposed SimGAN to refine simulator outputs using adversarial learning, Bousmalis et al. [25] studied unsupervised pixel-level domain adaptation, and Hoffman et al. [26] introduced CyCADA by combining image-level and feature-level adaptation with semantic consistency. Similar photorealistic translation has also been explored for synthetic driving scenes, such as CARLA2Real [27]. These methods reduce visual discrepancy between synthetic and real data, but they usually assume access to target-domain images and optimise image realism, feature alignment, or perception accuracy rather than closed-loop policy robustness. More importantly, unconstrained pixel-level translation may alter action-critical structures such as lane geometry, road boundaries, and obstacle layout.

Fourier-domain methods provide a more structured alternative by manipulating image spectra rather than pixels directly. Yang and Soatto [16] proposed Fourier Domain Adaptation, which replaces the low-frequency amplitude spectrum of source images with that of target images. Xu et al. [17] further showed that phase information tends to preserve semantic and geometric structure, whereas amplitude statistics are more closely related to domain-dependent appearance such as color, texture, and illumination. Compared with pixel-domain translation, Fourier-domain manipulation offers a clearer separation between appearance variation and structural preservation. However, existing Fourier-based methods either rely on target-domain spectra or perform generic spectrum mixing, and their perturbations remain disconnected from the visual evidence used by a closed-loop policy. This motivates a policy-aware use of Fourier-domain generalisation, where spectral-domain interventions are regulated by action-relevant visual evidence to modify deployment-sensitive appearance statistics while preserving task-defining spatial structure.

B. Policy-Relevant Domain Construction

Domain construction concerns how to generate source-domain variations that are not only diverse, but also effective for improving transfer robustness. Early domain randomisation methods broaden the simulation distribution by manually varying visual or physical factors. Tobin et al. [9] introduced visual domain randomisation by training models with randomised textures, distractors, camera parameters, lighting conditions, and image noise. Peng et al. [28] extended this idea to control by randomising dynamics parameters during simulation training. These methods established the effectiveness of

distributional broadening, but the induced domains are largely determined by predefined simulator parameters and sampling ranges.

Recent studies have moved toward objective-guided domain construction, where the randomised distribution is adapted according to a task- or transfer-related criterion. Bayesian domain randomisation searches source-domain distribution parameters to improve real-world performance using sparse target-domain data [29]. Active domain randomisation selects informative environment instances according to trajectory-level discrepancy or task difficulty [30], while automatic and entropy-based randomisation progressively expands the domain distribution according to policy performance or success constraints [31], [32]. These methods show that domain generation should be guided by learning objectives rather than by uniform sampling alone. However, their objectives are typically defined over simulator parameters, distributional spread, trajectory-level discrepancy, real-world rollouts, or episode-level outcomes.

For visual closed-loop policies, such objective signals remain indirect: they indicate whether a domain is useful at the task level, but do not reveal which visual evidence drives the policy's action selection. As a result, existing objective-guided randomisation may improve domain coverage or task-level robustness without explicitly controlling whether the generated visual shifts affect policy-relevant perception. This motivates a policy-relevant domain construction strategy, where action-related visual evidence is used to guide robustness-enhancing visual shifts generated entirely within simulation.

C. Explanation-Guided Training

Attribution-based explanation methods have been widely used to analyse the visual evidence behind model predictions and XAI has become an important direction for safe and trustworthy autonomous driving [33]. Model-agnostic methods such as LIME [34] approximate local decision behavior by perturbing input samples, while SHAP and its deep variants, such as DeepSHAP, estimate feature contributions from the perspective of Shapley values [35]. Gradient-based methods provide a more direct way to attribute model outputs to input features. Sundararajan et al. [36] proposed Integrated Gradients, which accumulates gradients along a path from a reference input to the actual input. Selvaraju et al. [19] proposed Grad-CAM, which produces class-discriminative localisation maps from convolutional feature maps. Shrikumar et al. [20] proposed DeepLIFT, which propagates contribution scores relative to a reference activation and assigns dense input-level relevance to model outputs.

These methods differ in their suitability for closed-loop visual reinforcement learning. Perturbation-based and Shapley-value methods are general but often require repeated input evaluations or background sampling, making them less suitable for frequent attribution extraction during policy training. Grad-CAM is efficient and intuitive, but its localisation is usually coarse because it is derived from high-level feature maps. Integrated Gradients provides input-level attribution, but its path integration can be computationally more expensive. In

contrast, DeepLIFT provides dense input-level contributions through reference-based propagation and is less affected by gradient saturation. This makes it suitable for identifying fine-grained visual evidence that influences continuous action outputs.

Beyond post-hoc interpretation, explanation-guided learning uses explanations as regularisation, supervision, or intervention signals to influence model training [22]. In reinforcement learning, Zuo et al. [37] studied contrastive explanation learning for explaining agent decisions. Existing works show that explanations can reveal decision-relevant evidence in visual policies, but they are mainly used for interpretation, debugging, or explanation-level supervision. In contrast, this work uses DeepLIFT-based attribution as a training-time signal for domain construction, linking action-relevant visual evidence with policy-conditioned visual domain shifts for target-free sim-to-real transfer.

III. METHODOLOGY

A. Overview

The proposed framework aims to construct policy-relevant visual domain shifts entirely within simulation for target-free sim-to-real reinforcement learning. As illustrated in Fig. 1, the framework follows a teacher-student structure. The teacher is a trained domain-randomised SAC policy and is kept frozen throughout training. The student is trained from scratch. During training, the student receives the XAI-guided Fourier-perturbed observations and is optimised by the proposed target-free robust policy objective. This student policy is indicated as the XFD model used for simulation evaluation and zero-shot Duckiebot deployment.

Based on the teacher attribution, the input image for the student model is perturbed in the Fourier domain. Following Fourier-domain adaptation [16] and Fourier-based domain generalisation [17], the amplitude component is modified to introduce appearance variations related to illumination, texture, contrast, and color statistics, while the phase component is preserved to maintain lane geometry, road boundaries, and obstacle layout. The attribution map regulates the spatial expression of the perturbation, so that task-critical regions are protected and less relevant regions can absorb stronger appearance shifts. This converts generic spectral augmentation into a policy-conditioned domain intervention.

To maintain stable policy optimisation, the intervention is activated through a reward-gated curriculum. The student is first trained with standard SAC until it acquires basic driving behavior, after which the XAI-guided Fourier perturbation is gradually introduced. The student policy is then optimised under these policy-conditioned and task-consistent visual domains using the SAC objective with auxiliary teacher-guided regularisation terms. The auxiliary teacher-guided terms are used only to stabilise the student response under generated visual domains, while the policy improvement remains driven by the SAC objective.

B. SAC-Based Visual Driving Policy

The Duckiebot navigation task is modeled as a continuous-control Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$.

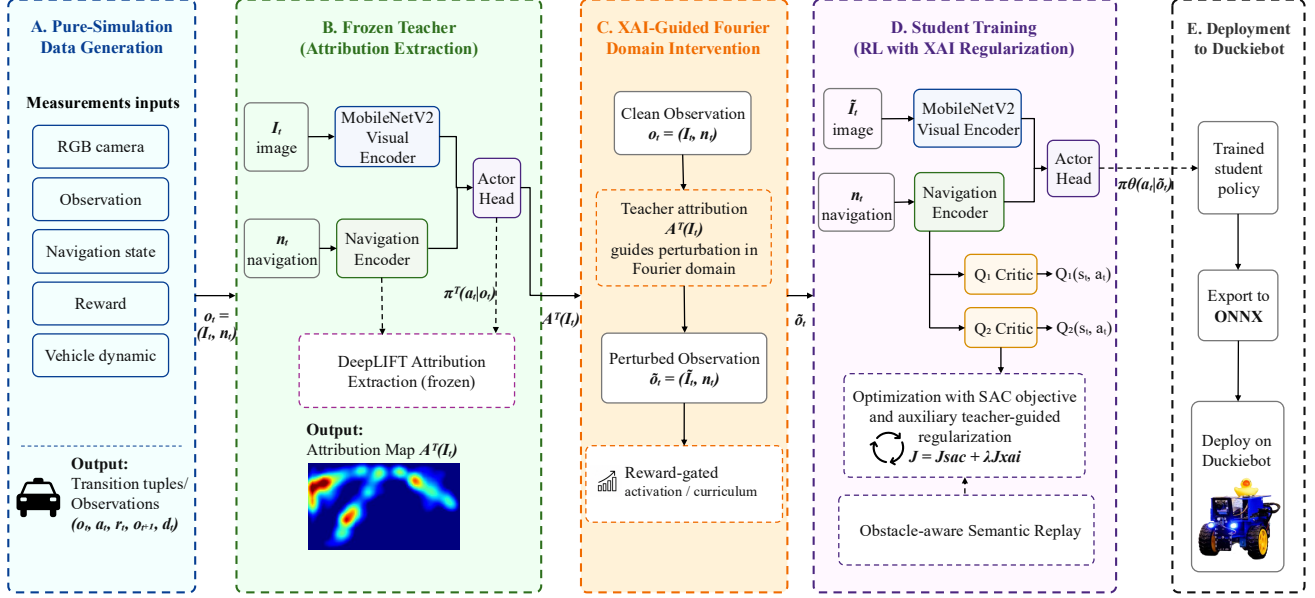


Fig. 1. Overall framework of the proposed target-free XAI-guided Fourier domain generalisation method. A single obstacle-aware policy is trained in simulation with randomised obstacle placements, where teacher attributions guide Fourier-domain visual intervention and semantic replay improves sampling of sparse obstacle-interaction transitions. The trained student policy is exported to ONNX for zero-shot Duckiebot deployment.

At each time step t , the agent receives an observation $o_t = (I_t, n_t)$, where I_t is the front-view RGB image and $n_t \in \mathbb{R}^5$ denotes the low-dimensional navigation state. The upper part of the image is cropped before resizing and normalisation to reduce task-irrelevant background information. The action $a_t \in \mathbb{R}^2$ represents the left and right wheel commands.

We adopt soft actor-critic (SAC) [8] as the base visual control backbone. SAC-style actor-critic formulations have also been used in autonomous-driving decision-making tasks to improve control efficiency and safety under dynamic traffic conditions [38]. The actor $\pi_\theta(a_t|o_t)$ predicts a stochastic wheel-control policy, while two critics Q_{ψ_1} and Q_{ψ_2} estimate the corresponding soft action values using the clipped double-Q design. In our implementation, the actor and critic employ separate visual encoders. This separation allows policy and value learning to form independent visual representations, and prevents the actor-side regularisation introduced later from directly constraining the critic.

This subsection only defines the visual control backbone. The complete student optimisation objective under the constructed target-free visual domains is given in Section III-E.

C. Attribution Extraction

To identify the visual evidence that contributes to the control decision, we extract DeepLIFT [20] attribution maps from a frozen teacher actor. The teacher is instantiated from a domain-randomised SAC policy pretrained in simulation, providing an attribution reference that has already been exposed to visual appearance variations. During XFD training, the teacher network is kept fixed and is used only for attribution extraction and local regularisation.

Given an observation $o_t = (I_t, n_t)$, the deterministic output of the teacher actor is denoted by $\mu_T(o_t) \in \mathbb{R}^{d_a}$, where $d_a = 2$ corresponds to the left and right wheel commands. DeepLIFT propagates contribution scores from the teacher action output back to the image input relative to a reference activation. For the RGB image I_t , the resulting attribution tensor is written as

$$C_T(I_t) \in \mathbb{R}^{H \times W \times 3 \times d_a} \quad (1)$$

Since the policy output is continuous and multi-dimensional, the contributions from both action dimensions are aggregated into a single spatial attribution map:

$$\mathcal{A}_T(u, v) = \frac{1}{Z} \sum_{c=1}^3 \sum_{k=1}^{d_a} |C_T(u, v, c, k)| \quad (2)$$

where Z normalizes $\mathcal{A}_T$ into $[0, 1]$. A larger value of $\mathcal{A}_T(u, v)$ indicates that the corresponding image location contributes more strongly to the teacher's wheel-control output.

The obtained DeepLIFT attribution map is used as a training-time signal for domain construction, which is consistent with interpretable driving studies that introduce intermediate visual evidence to improve policy transparency [39]. Regions with high attribution are treated as control-relevant visual evidence, whereas regions with low attribution provide larger degrees of freedom for appearance perturbation. This attribution-guided separation is passed to the Fourier intervention module, allowing the subsequent domain construction to modify visual appearance while preserving the image structures most relevant to closed-loop control.

D. XAI-Guided Fourier Domain Intervention

The proposed intervention constructs policy-conditioned and task-consistent visual domains by perturbing simulation

images in the Fourier domain. The key idea is to alter appearance-dependent statistics while preserving the geometric structures required for closed-loop control. Unlike policy-agnostic domain randomisation [9] or dynamics randomisation [28], the perturbation is conditioned on the DeepLIFT attribution map extracted in Section III-C, so that the spatial strength of the intervention is regulated according to control relevance, as illustrated by the teacher attribution responses in Fig. 2(b) and Fig. 2(g).

Directly applying strong visual perturbations at the beginning of training may destabilise exploration, since the student policy has not yet acquired a reliable driving prior. Therefore, the intervention is activated by a reward-gated curriculum, following the curriculum-learning principle of introducing stronger interventions after basic task competence has been acquired [40]. Let R_e denote the return of episode e , and let $\bar{R}_e$ be the moving average over the most recent K episodes. The intervention gate is defined as

$$g_t = \mathbb{I}(\bar{R}_e \geq \rho) \quad (3)$$

where ρ is a task-dependent activation threshold. Once the gate is enabled at step t_g , the perturbation strength is gradually increased:

$$\eta_t = g_t \cdot \min \left(\eta_{\max}, \eta_{\min} + \frac{t - t_g}{T_{\text{ramp}}} (\eta_{\max} - \eta_{\min}) \right) \quad (4)$$

where T_{ramp} controls the ramp-up duration, and $\eta_{\min}$ and $\eta_{\max}$ denote the minimum and maximum perturbation strengths. This schedule delays the increase of intervention strength until the policy has learned basic control behavior, thereby avoiding abrupt distribution shifts during early optimisation.

Given a preprocessed RGB image I_t , as exemplified by the lane-following and obstacle-avoidance inputs in Fig. 2(a) and Fig. 2(f), the Fourier transform is computed independently for each channel:

$$\mathcal{F}(I_t^c) = A_t^c \exp(jP_t^c), \quad c \in \{1, 2, 3\} \quad (5)$$

where A_t^c and P_t^c denote the amplitude and phase components, respectively. The amplitude component is perturbed to modify appearance statistics, including illumination, contrast, texture, and color distribution, while the phase component is preserved to maintain lane geometry, road boundaries, and obstacle layout. The perturbed amplitude is defined as

$$\tilde{A}_t^c = A_t^c \odot (1 + \eta_t \Delta_t^c) \quad (6)$$

where Δ_t^c is a bounded spectral perturbation and $\odot$ denotes element-wise multiplication. The intermediate Fourier-perturbed image is reconstructed by

$$I_t^F = \mathcal{F}^{-1} \left(\tilde{A}_t \exp(jP_t) \right) \quad (7)$$

The top-attribution regions in Fig. 2(c) and Fig. 2(h) visualise the most action-relevant areas selected from the teacher attribution maps. These binary regions are used only for interpretation, whereas the continuous attribution map is used in the actual intervention. Let $A_T(I_t) \in [0, 1]^{H \times W}$ denote the normalised teacher attribution map. The attribution-weighted Fourier shifts shown in Fig. 2(d) and Fig. 2(i) indicate where

the reconstructed appearance variation is allowed to be expressed after being modulated by the teacher attribution. A complementary intervention mask is defined as

$$M_t = 1 - A_T(I_t) \quad (8)$$

The final perturbed image is obtained by blending the original image and the Fourier-perturbed image:

$$\tilde{I}_t = I_t + M_t \odot (I_t^F - I_t) \quad (9)$$

This formulation assigns lower perturbation freedom to high-attribution regions and higher perturbation freedom to low-attribution regions. The final perturbed observations in Fig. 2(e) and Fig. 2(j) show that the generated images change domain-dependent appearance while retaining the lane, boundary, and obstacle structures that are most relevant to control. The intervention therefore converts generic Fourier augmentation into a policy-conditioned visual-domain intervention. The student policy is subsequently trained on $\tilde{o}_t = (\tilde{I}_t, n_t)$, allowing reinforcement learning to optimise control behavior under visually shifted but structurally consistent simulation domains.

E. Target-Free Robust Policy Optimisation

The student policy is optimised without target-domain images, real-world rollouts, or demonstrations. To write the optimisation compactly, we define the training observation as

$$x_t = \begin{cases} o_t, & g_t = 0 \\ \tilde{o}_t, & g_t = 1 \end{cases} \quad (10)$$

where g_t is the intervention gate defined in Section III-D. Before the intervention is activated, the student is trained on the original simulation observation o_t . After activation, it is trained on the attribution-guided Fourier observation $\tilde{o}_t$.

For compact notation, let $Q_{\min}^{\psi}(x, a) = \min_{i=1,2} Q_{\psi_i}(x, a)$ and $\bar{Q}_{\min}^{\psi}(x, a) = \min_{i=1,2} \bar{Q}_{\psi_i}(x, a)$. For a transition $(x_t, a_t, r_t, x_{t+1}, d_t)$, the SAC target is

$$V_{\bar{\psi}}(x_{t+1}) = \bar{Q}_{\min}^{\psi}(x_{t+1}, a_{t+1}) - \alpha \log \pi_{\theta}(a_{t+1}|x_{t+1}) \quad (11)$$

where $a_{t+1} \sim \pi_{\theta}(\cdot|x_{t+1})$. The SAC target is then written as

$$y_t = r_t + \gamma(1 - d_t)V_{\bar{\psi}}(x_{t+1}) \quad (12)$$

The critic is optimised by

$$\mathcal{L}_Q = \sum_{i=1}^2 \mathbb{E}_{\mathcal{D}} \left[(Q_{\psi_i}(x_t, a_t) - y_t)^2 \right] \quad (13)$$

The teacher is not involved in the critic target or critic loss. Thus, value learning is driven only by the reinforcement learning objective under the constructed domains.

The SAC actor loss is

$$\mathcal{L}_{\text{SAC}} = \mathbb{E}_{\mathcal{D}} \left[\alpha \log \pi_{\theta}(a_t|x_t) - Q_{\min}^{\psi}(x_t, a_t) \right] \quad (14)$$

This term allows the student to improve its control behavior through reinforcement learning, rather than merely imitating the teacher.

To stabilise policy learning under appearance-shifted observations, a behavioral consistency term is introduced between

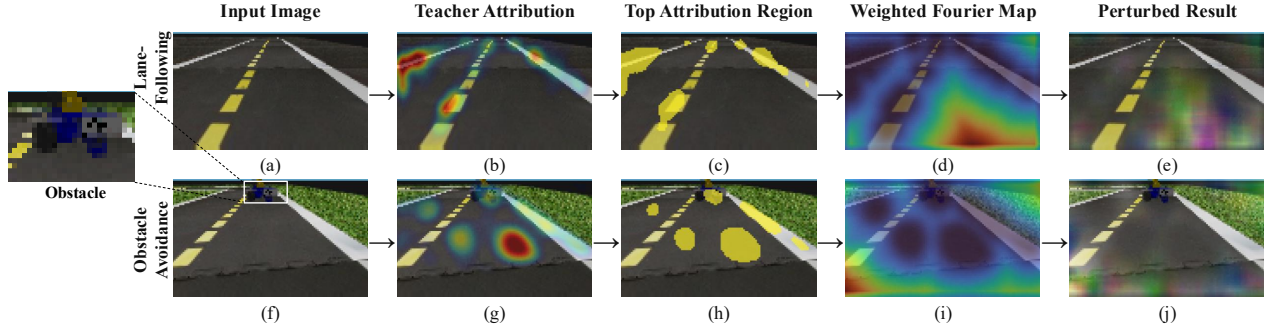


Fig. 2. Visualisation of the XAI-guided Fourier intervention pipeline under lane-following and obstacle-avoidance observations. The columns show the input image, teacher attribution, top-attribution region, attribution-weighted Fourier map, and final perturbed result.

the student response to the perturbed observation and the teacher response to the corresponding clean observation. Let $\mu_S(\tilde{o}_t)$ and $\mu_T(o_t)$ denote the deterministic mean actions of the student and teacher actors, respectively. The teacher-student consistency loss [23] is defined as

$$\mathcal{L}_{\text{cons}} = \mathbb{E}_{\mathcal{D}} \left[\|\mu_S(\tilde{o}_t) - \text{sg}(\mu_T(o_t))\|_2^2 \right] \quad (15)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. This term uses the frozen teacher only as a local behavioral reference and does not update the teacher parameters.

In addition to action-level consistency, attribution-level constraints [21] are applied to regulate the visual evidence used by the student actor. Let $\mathcal{A}_T(I_t)$ and $\mathcal{A}_S(\tilde{I}_t)$ denote the normalised DeepLIFT attribution maps of the teacher and student actors, respectively. The outside-attribution loss penalizes student attribution assigned to regions that are weakly supported by the teacher attribution:

$$\mathcal{L}_{\text{out}} = \mathbb{E}_{\mathcal{D}} \left[\frac{\|M_t \odot \mathcal{A}_S(\tilde{I}_t)\|_1}{\|M_t\|_1 + \epsilon} \right] \quad (16)$$

where $M_t = 1 - \mathcal{A}_T(I_t)$ is the complementary attribution mask defined in Section III-D. This discourages the student from relying on visually perturbed regions that are not strongly supported by the teacher attribution.

A complementary attribution-alignment term preserves the global structure of the teacher attribution under the perturbed observation: Let $\mathbf{a}_T = \text{vec}(\mathcal{A}_T(I_t))$ and $\mathbf{a}_S = \text{vec}(\mathcal{A}_S(\tilde{I}_t))$. The attribution-alignment loss is defined as

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{\mathcal{D}} \left[1 - \frac{\langle \mathbf{a}_T, \mathbf{a}_S \rangle}{\|\mathbf{a}_T\|_2 \|\mathbf{a}_S\|_2 + \epsilon} \right] \quad (17)$$

While $\mathcal{L}_{\text{out}}$ suppresses attribution leakage into low-teacher-attribution regions, $\mathcal{L}_{\text{align}}$ encourages the student to preserve the overall spatial attribution structure of the teacher.

The final student actor objective combines reinforcement learning, behavioral consistency, and attribution regularisation:

$$\mathcal{L}_{\pi} = \mathcal{L}_{\text{SAC}} + g_t \left(\lambda_{\text{cons}} \mathcal{L}_{\text{cons}} + \lambda_{\text{xai}} (\mathcal{L}_{\text{out}} \mathcal{L}_{\text{out}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}) \right) \quad (18)$$

where g_t is the intervention gate defined in Section III-D. When $g_t = 0$, the student is optimised by the standard SAC

objective. When $g_t = 1$, the auxiliary terms regularise the actor toward stable action-level and attribution-level behavior under target-free visual domain shifts.

F. Obstacle-Aware Semantic Replay

Since a single obstacle-aware policy is trained in this work, experience replay is configured for the obstacle-aware training environment rather than for two separately trained task policies. During simulation training, the Duckiebot encounters randomised obstacle placements while still being required to maintain lane-following behavior. Therefore, the replay buffer contains both lane-tracking transitions and sparse obstacle-interaction transitions. The resulting replay organisation is illustrated in Fig. 3.

Each collected transition is assigned to one semantic replay category:

$$\mathcal{B} = \{\mathcal{B}_{\text{normal}}, \mathcal{B}_{\text{near}}, \mathcal{B}_{\text{success}}, \mathcal{B}_{\text{fail}}\}, \quad (19)$$

where the four categories correspond to normal lane-following transitions, near-obstacle interactions, successful avoidance maneuvers, and failure-related transitions, respectively. The normal category preserves lane-tracking competence, while the remaining categories increase the exposure of sparse safety-critical obstacle-avoidance cases.

Mini-batches are sampled from a progress-dependent category distribution

$$\omega(p_t) = [\omega_{\text{normal}}, \omega_{\text{near}}, \omega_{\text{success}}, \omega_{\text{fail}}], \quad (20)$$

which satisfies

$$\sum_{b \in \mathcal{B}} \omega_b(p_t) = 1, \quad (21)$$

where $p_t = t/T$ denotes the normalised training progress. In the early stage, the sampler increases the probability of near-obstacle and failure-related transitions to expose the policy to decision boundaries and unsafe outcomes. As training proceeds, the sampling distribution shifts toward successful avoidance transitions while retaining normal lane-following samples. This design allows the same policy to learn obstacle-aware control without sacrificing lane-following behavior.

This replay mechanism is independent of the proposed Fourier intervention.

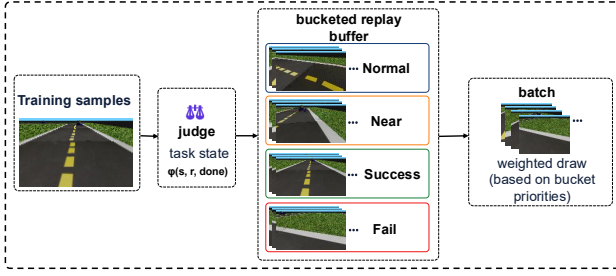


Fig. 3. Obstacle-aware semantic replay sampling. Transitions collected from randomised obstacle-placement training are assigned to normal, near-obstacle, success, and failure categories, and mini-batches are sampled by weighted category selection.

It acts as a task-level sampling strategy within the same SAC optimisation framework, correcting the long-tailed transition distribution induced by randomised obstacle placement while preserving standard lane-tracking competence.

G. Reward Formulation

The reward function is used to induce stable lane-following behavior and to suppress degenerate policies such as stopping, excessive heading deviation, and boundary-hugging. Since reward design is not the main contribution of this work, both the baseline and the proposed method are trained with the same task reward.

At each timestep, the simulator first checks whether the vehicle can be assigned to a valid lane pose. If the lane pose is invalid, a fixed penalty $r_t = -5$ is returned. For valid lane poses, the reward is defined as

$$r_t = r_t^{\text{prog}} + r_t^{\text{move}} - p_t^{\text{stop}} - p_t^{\text{angle}} - p_t^{\text{edge}} \quad (22)$$

The progress reward encourages forward motion along the lane direction:

$$r_t^{\text{prog}} = 1.5 s_t \max(c_t, 0) \quad (23)$$

where $s_t = \text{clip}(v_t, 0, 1)$ is the normalised speed and c_t denotes the projection of the vehicle heading onto the lane direction.

A small motion reward is added to discourage stationary behavior:

$$r_t^{\text{move}} = 0.02 \min\left(1, \frac{s_t}{s_{\min}}\right) \quad (24)$$

where $s_{\min} = 0.10$.

The stopping penalty is imposed when the normalised speed is lower than $s_{\min}$:

$$p_t^{\text{stop}} = \mathbb{I}(s_t < s_{\min}) k_{\text{stop}} \left(\frac{s_{\min} - s_t}{s_{\min}} \right)^2 \quad (25)$$

The heading penalty is applied outside an angular dead zone:

$$p_t^{\text{angle}} = k_{\text{angle}} \max(0, |\theta_t| - \theta_{\text{dead}}) \quad (26)$$

where θ_t denotes the heading error.

The edge penalty discourages driving close to the lane boundary:

$$p_t^{\text{edge}} = k_{\text{edge}} \max(0, |e_t| - e_{\text{thr}}) \quad (27)$$

where e_t is the lateral deviation from the lane reference. The constants are set to $k_{\text{stop}} = 3.0$, $k_{\text{angle}} = 0.2$, $\theta_{\text{dead}} = 30^\circ$, $k_{\text{edge}} = 4.0$, and $e_{\text{thr}} = 0.10$.

This reward only defines the task objective. The proposed Fourier-domain intervention and attribution-guided policy regularisation are independent of the reward formulation; therefore, improvements over the baseline are attributed to the proposed target-free robust training strategy rather than to reward redesign.

IV. EXPERIMENTS

This section describes the experimental protocol used to train and evaluate the proposed XAI-guided Fourier domain generalisation framework, including the task and map configurations, baseline methods, training protocols, and evaluation. The evaluation metrics are designed to assess closed-loop robustness, cross-map generalisation, and zero-shot sim-to-real transfer under visual domain shifts. Quantitative comparisons, real-world deployment results, and attribution-based qualitative analyses are presented in Section V.

A. Environment Configurations

The experiments are conducted in the Duckietown environment using an obstacle-aware training task. During training, the policy is optimised on Map 2, shown in Fig. 4, with randomised obstacle placements. The policy is required to maintain lane-following behaviour while responding appropriately to obstacle interactions.

The experimental map configuration is organised as follows:

- **Map 2:** the simulation training map, where the single obstacle-aware policy is trained with randomised obstacle placements.
- **Map 1 and Map 3:** unseen simulation evaluation maps used to assess cross-map generalisation.
- **Map 4 and Map 5:** physical Duckietown layouts used for zero-shot real-world deployment, corresponding to the evaluation layouts represented by Map 1 and Map 3, respectively.

The same trained policies are evaluated on all simulated and real-world maps without target-domain fine-tuning or target domain data acquisition. All the map layouts used in the experiments are demonstrated in Fig. 4.

B. Baseline Methods

All compared methods use the same SAC-based visual control backbone and are trained under the same obstacle-aware simulation setting on Map 2. The baseline is a standard SAC policy trained without the proposed sim-to-real components. A randomly domain-randomised SAC policy is included to represent the conventional visual randomisation. The proposed XFD policy uses the same backbone and training environment but introduces DeepLIFT-guided Fourier domain intervention and target-free robust policy optimisation during training.

For each method, the same trained policy is evaluated under both lane-following and obstacle-avoidance tasks. This setting keeps the comparison focused on whether the proposed

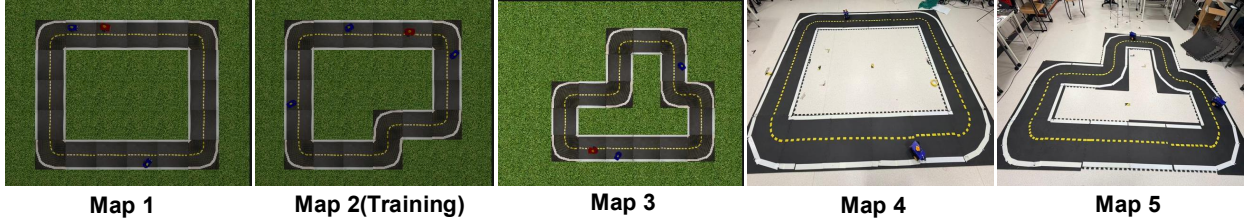


Fig. 4. Duckietown map layouts for simulation training, simulation evaluation, and zero-shot real-world deployment.

visual-domain intervention improves the robustness of a single obstacle-aware policy, rather than relying on task-specific policy training. For compact notation, the compared methods are denoted as follows:

- **Base**: standard SAC without any sim-to-real techniques.
- **DR**: conventional domain-randomised SAC with target-free visual randomisation. This model also serves as the teacher policy during XFD training.
- **XFD**: the full proposed method.

C. Training Protocols

All policies are trained entirely in simulation using Map 2 with randomised obstacle placements. No target-domain images, real-world rollouts, demonstrations, or real-world fine-tuning are used during training. The input observation consists of a cropped front-view RGB image and a five-dimensional navigation state. The action is a two-dimensional continuous wheel command corresponding to the left and right wheel velocities.

The Base and DR policies are trained using the standard SAC framework. For the DR baseline, the visual randomisation parameters are empirically tuned to maximise training performance in simulation. Both policies are trained until the episode reward reaches a stable plateau, indicating policy convergence. XFD is subsequently trained using the converged DR policy as a frozen teacher. The student policy is initially optimised on clean simulation observations using the standard SAC objective. Once the reward-gated criterion is satisfied, the attribution-guided Fourier intervention is activated and the perturbation strength is progressively increased. During this phase, the teacher serves only as an attribution source, while only the XFD student policy is updated.

After training, the same policy checkpoint is evaluated without further adaptation on simulation Map 1, Map 2, and Map 3, as well as on the corresponding real-world Map 4 and Map 5. The evaluation includes both lane-following and obstacle-avoidance tasks.

D. Evaluation Metrics

The trained policies are evaluated in both simulation and zero-shot real-world deployment. Each trained checkpoint is

evaluated under both lane-following and obstacle-avoidance tasks using the same set of metrics.

For the lane-following task, each method is evaluated over ten runs on each map in the absence of obstacles. For the obstacle-avoidance task, each map contains two obstacle-placement configurations to increase evaluation diversity. Each configuration is evaluated over five runs, resulting in ten trajectories per map, which are pooled to compute the obstacle-avoidance metrics. The same evaluation protocol is applied in both simulation and real-world experiments. For both tasks, successfully completing one closed-loop circuit of the map is considered a success. A run is terminated if the vehicle leaves the drivable lane or collides with an obstacle.

During real-world evaluation, the Duckiebot pose is recorded using an OptiTrack motion-capture system. The recorded pose sequence is projected onto the calibrated map coordinate frame and resampled at 10 Hz. The resulting map-plane trajectory is used to compute travelled distance, lateral deviation, reconstructed reward, and yaw-based smoothness.

After the trajectories are collected, four complementary metrics are used for quantitative evaluation:

- **Travelled distance**: the cumulative map-plane trajectory length after 10 Hz resampling, reflecting how far the vehicle progresses before termination.
- **Reward**: the cumulative simulator-like return reconstructed from forward speed, heading alignment, and centerline deviation. A terminal failure penalty is added when the vehicle does not complete one lap.
- **Lateral deviation**: the mean absolute distance between the vehicle trajectory and the fitted right-lane centerline, measuring lane-tracking accuracy.
- **Smoothness**: a yaw-based regularity score computed from the measured vehicle yaw acceleration:

$$S_{\text{smooth}} = \frac{100}{1 + \text{RMS}(\ddot{\psi})/r_{\text{ref}}}, \quad (28)$$

where $\text{RMS}(\cdot)$ denotes the root mean square operator, $\ddot{\psi}$ denotes the yaw acceleration, and r_{ref} is a reference normalisation factor. Higher smoothness indicates less abrupt yaw variation. The reported results are presented as the mean and standard deviation across these ten runs.

All methods are evaluated under the same task settings, map layouts, obstacle placements, and reward reconstruction procedure. Therefore, differences in these metrics reflect the

TABLE I
SIMULATION EVALUATION RESULTS OF THE OBSTACLE-AWARE POLICIES UNDER LANE-FOLLOWING AND OBSTACLE-AVOIDANCE TASKS. RESULTS ARE REPORTED AS MEAN $\pm$ STANDARD DEVIATION OVER TEN EVALUATION TRAJECTORIES.

Task	Map	Method	Distance	Reward	Deviation	Smoothness
Lane following	Map 1	Base	10.95 $\pm$ 0.01	834.40 $\pm$ 0.78	0.0333 $\pm$ 0.0005	91.44 $\pm$ 0.32%
		DR (Teacher)	10.54 $\pm$ 0.00	798.46 $\pm$ 0.36	0.0372 $\pm$ 0.0003	93.86 $\pm$ 0.26%
		XFD (Student)	11.39 $\pm$ 0.01	851.39 $\pm$ 0.21	0.0796 $\pm$ 0.0003	93.23 $\pm$ 0.12%
	Map 2	Base	5.88 $\pm$ 0.02	338.73 $\pm$ 1.21	0.0460 $\pm$ 0.0023	88.92 $\pm$ 0.41%
		DR (Teacher)	10.22 $\pm$ 0.01	734.70 $\pm$ 0.77	0.0500 $\pm$ 0.0003	89.65 $\pm$ 0.43%
		XFD (Student)	11.00 $\pm$ 0.00	819.31 $\pm$ 0.36	0.0766 $\pm$ 0.0001	90.59 $\pm$ 0.11%
	Map 3	Base	4.04 $\pm$ 0.02	198.38 $\pm$ 0.88	0.0504 $\pm$ 0.0021	85.88 $\pm$ 0.51%
		DR (Teacher)	9.09 $\pm$ 2.27	587.94 $\pm$ 195.99	0.0739 $\pm$ 0.0063	86.53 $\pm$ 3.40%
		XFD (Student)	10.72 $\pm$ 0.01	799.75 $\pm$ 0.46	0.0725 $\pm$ 0.0003	89.29 $\pm$ 0.12%
Obstacle avoidance	Map 1	Base	11.07 $\pm$ 0.04	780.54 $\pm$ 4.19	0.0564 $\pm$ 0.0005	85.64 $\pm$ 0.64%
		DR (Teacher)	10.42 $\pm$ 0.01	720.42 $\pm$ 0.67	0.0677 $\pm$ 0.0004	90.98 $\pm$ 0.20%
		XFD (Student)	11.36 $\pm$ 0.01	802.13 $\pm$ 0.90	0.0848 $\pm$ 0.0019	86.86 $\pm$ 0.32%
	Map 2	Base	6.10 $\pm$ 0.06	277.63 $\pm$ 7.86	0.0762 $\pm$ 0.0015	82.62 $\pm$ 0.61%
		DR (Teacher)	10.01 $\pm$ 0.04	689.37 $\pm$ 7.16	0.0641 $\pm$ 0.0015	89.93 $\pm$ 0.69%
		XFD (Student)	10.86 $\pm$ 0.02	763.65 $\pm$ 1.01	0.0796 $\pm$ 0.0016	87.96 $\pm$ 0.18%
	Map 3	Base	4.13 $\pm$ 0.01	188.48 $\pm$ 0.96	0.0733 $\pm$ 0.0008	81.89 $\pm$ 0.50%
		DR (Teacher)	8.30 $\pm$ 2.21	447.62 $\pm$ 201.06	0.1056 $\pm$ 0.0094	84.68 $\pm$ 0.81%
		XFD (Student)	10.96 $\pm$ 0.01	759.08 $\pm$ 3.37	0.0832 $\pm$ 0.0004	86.08 $\pm$ 0.25%

effect of the proposed method rather than changes in the evaluation protocol or reward definition. Quantitative results are reported in Section V.

E. Implementation Details

All policies are implemented in PyTorch using the same SAC-based visual actor-critic architecture. The actor and critic do not share the visual backbone; instead, they use separate MobileNetV2 visual encoders to avoid actor-side regularisation directly constraining critic value estimation. Each RGB observation is encoded by a MobileNetV2-based visual encoder, while the five-dimensional navigation state is processed by a lightweight fully connected navigation encoder. The extracted visual and navigation features are concatenated and passed to the corresponding policy or value head. The actor head outputs the mean and log-standard deviation of a two-dimensional Gaussian wheel-command policy, whereas the critic consists of two independent Q-value heads following the clipped double-Q design. The obstacle-aware student policy is trained on Map 2 with randomised obstacle placements. The main hyper-parameter settings are summarised in Table II.

The attribution-guided Fourier intervention is activated only after the reward-gated condition is satisfied, and the perturbation strength is increased progressively during the ramp-up stage. Attribution regularisation is applied periodically on a small mini-batch during training. During deployment, no Fourier perturbation or DeepLIFT computation is required; only the trained student actor is exported in Open Neural Network Exchange (ONNX) format and deployed on the Duckiebot. Therefore, the proposed method does not introduce additional inference cost during real-world execution.

TABLE II
HYPER-PARAMETER SETTINGS USED FOR POLICY TRAINING.

Item	Setting
Training map	Map 2 with randomised obstacles
Training steps	5.0×10^5 environment steps
Replay buffer size	8.0×10^4
Batch size	256
Action repeat	2
Discount factor γ	0.99
Target smoothing τ	0.005
Actor learning rate	1.0×10^{-4}
Critic learning rate	3.0×10^{-4}
Temperature learning rate	3.0×10^{-4}
Fourier activation step	5.0×10^4 steps
Fourier ramp-up duration	1.5×10^5 steps
Attribution method	DeepLIFT
Deployment format	ONNX
Deployment platform	Jetson Nano

V. RESULTS

A. Simulation Evaluation

Table I reports the simulation results of the three obstacle-aware policies in lane-following and obstacle-avoidance tasks. All policies are trained only on Map 2 with randomised obstacle placements and are evaluated on Map 1, Map 2, and Map 3 without further adaptation. Map 1 and Map 3 are unseen during training, and therefore mainly reflect cross-map generalisation. The corresponding simulation trajectories are shown in Fig. 5.

In the lane-following task, XFD achieves the highest distance and reward across all three maps. On the training map, Map 2, XFD increases the travelled distance from 5.88 for the base SAC policy and 10.22 for DR to 11.00, while improving the reward from 338.73 and 734.70 to 819.31, respectively.

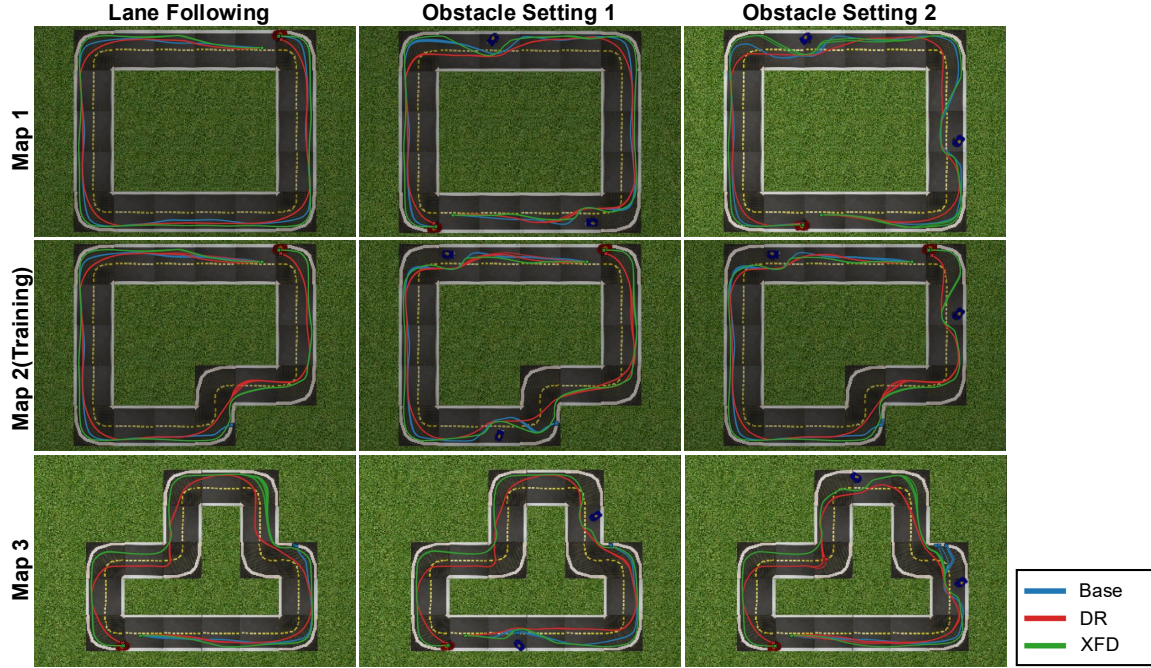


Fig. 5. Simulation trajectory comparisons on Map 1, Map 2, and Map 3 under lane-following and obstacle-avoidance tasks. Each row corresponds to a simulation map, and each column corresponds to lane following, obstacle setting 1, and obstacle setting 2.

The improvement is more pronounced on the unseen Map 3, where the base policy terminates early with a distance of 4.04 and DR reaches 9.09 with large variance, whereas XFD maintains a distance of 10.72 and a reward of 799.75. This indicates that the proposed visual-domain intervention improves lane-following generalisation even though no separate lane-following policy is trained.

In the obstacle-avoidance task, XFD also obtains the highest distance and reward across all maps. On Map 2, the proposed method improves the distance from 6.10 for the base policy and 10.01 for DR to 10.86. On the unseen Map 3, XFD further increases the distance from 4.13 and 8.30 to 10.96, and improves the reward from 188.48 and 447.62 to 759.08, respectively. These results show that the proposed method improves both lane-tracking behavior and obstacle-aware closed-loop control within the same trained policy.

Averaged over the three simulation maps, XFD improves the lane-following distance by approximately 58.6% over the base policy and 10.9% over DR. For obstacle avoidance, the average distance improvement is approximately 55.8% over the base policy and 15.5% over DR. The corresponding reward gains are also consistent across both tasks, confirming that the proposed method improves task progress rather than only changing low-level trajectory statistics.

These simulation results also highlight the effectiveness of the teacher-student framework in XFD. Although the student receives guidance from the teacher through action and attribution information, it is not trained as an imitation policy. Instead, the teacher provides local stabilisation signals under the generated visual domains, while the student continues to

optimise its actor and critic networks through the objective function. Consequently, the student is able to improve beyond the teacher's performance rather than being constrained by it. This is reflected in Table I: XFD outperforms DR in travelled distance and reward across all simulated maps and both evaluation tasks. Averaged over the three maps, XFD improves the lane-following distance and reward over DR by approximately 10.9% and 16.5%, respectively. For obstacle avoidance, the corresponding gains over DR are approximately 15.5% in distance and 25.2% in reward. These results show that the proposed method already improves policy performance within simulation, before considering the additional sim-to-real transfer gains.

The deviation and smoothness metrics should therefore be interpreted together with distance and reward. In several cases, the base or DR policy gives lower deviation or higher smoothness, but these values are often obtained from shorter trajectories, conservative motion, or early termination. In contrast, XFD consistently completes longer trajectories and receives higher rewards, which exposes the policy to more curved segments and obstacle-interaction regions. In conclusion, the simulation results indicate that XFD improves closed-loop completion and cross-map robustness while maintaining acceptable trajectory regularity.

B. Real-World Deployment

Table III presents the zero-shot real-world deployment results on physical Map 4 and Map 5. The term target-free refers to the training and adaptation process: the policies are trained entirely in simulation and are deployed without

TABLE III

ZERO-SHOT REAL-WORLD EVALUATION RESULTS OF THE SINGLE OBSTACLE-AWARE POLICY UNDER LANE-FOLLOWING AND OBSTACLE-AVOIDANCE TASKS. RESULTS ARE REPORTED AS MEAN $\pm$ STANDARD DEVIATION OVER TEN EVALUATION TRAJECTORIES.

Task	Map	Method	Distance	Reward	Deviation	Smoothness
Lane following	Map 4	Base	1.37 ± 0.31	-92.16 ± 15.83	0.2286 ± 0.0249	$22.54 \pm 3.83\%$
		DR (Teacher)	4.06 ± 3.17	133.01 ± 210.06	0.0975 ± 0.0347	$47.73 \pm 6.14\%$
		XFD (Student)	10.26 ± 0.12	576.48 ± 16.55	0.1494 ± 0.0054	$42.18 \pm 5.17\%$
	Map 5	Base	0.77 ± 0.08	-82.53 ± 4.69	0.1748 ± 0.0193	$28.03 \pm 1.60\%$
		DR (Teacher)	2.91 ± 0.87	60.82 ± 52.65	0.1270 ± 0.0264	$55.52 \pm 2.42\%$
		XFD (Student)	9.63 ± 0.49	471.17 ± 99.60	0.1537 ± 0.0121	$42.27 \pm 4.50\%$
Obstacle avoidance	Map 4	Base	1.42 ± 0.34	-83.87 ± 28.42	0.1318 ± 0.0668	$24.10 \pm 7.11\%$
		DR (Teacher)	2.77 ± 1.17	61.10 ± 81.26	0.0887 ± 0.0283	$50.01 \pm 6.40\%$
		XFD (Student)	9.95 ± 0.97	482.25 ± 73.69	0.1822 ± 0.0109	$47.32 \pm 5.12\%$
	Map 5	Base	0.98 ± 0.40	-92.30 ± 37.12	0.1140 ± 0.0246	$32.59 \pm 7.72\%$
		DR (Teacher)	2.85 ± 1.39	-3.06 ± 65.58	0.1892 ± 0.0574	$49.35 \pm 15.71\%$
		XFD (Student)	9.77 ± 0.14	411.28 ± 106.57	0.1873 ± 0.0174	$44.91 \pm 2.42\%$

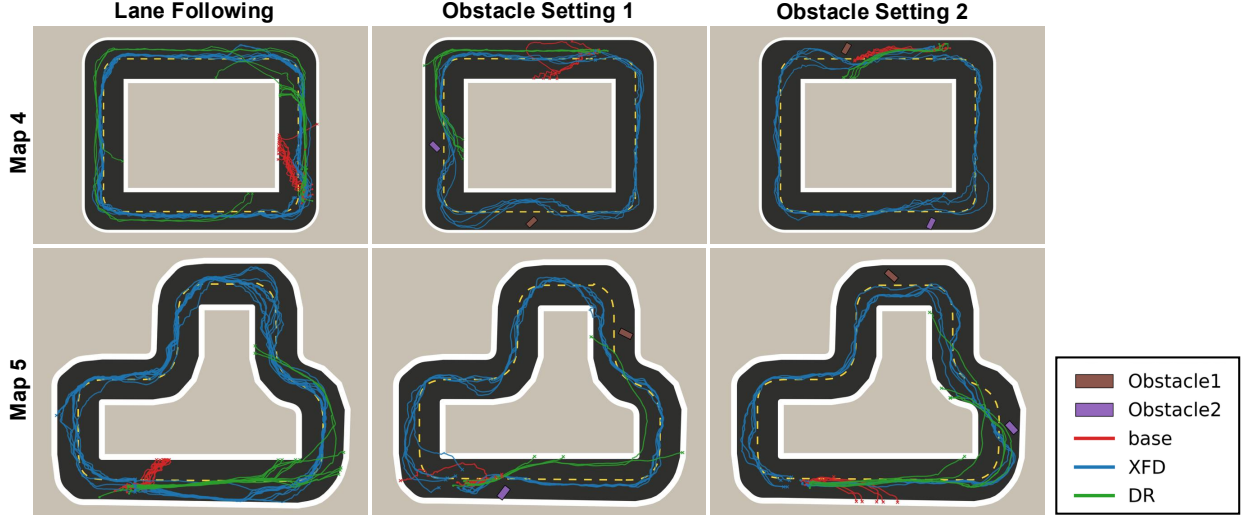


Fig. 6. Real-world trajectory comparisons on two physical Duckietown maps. The columns correspond to lane following, obstacle setting 1, and obstacle setting 2, respectively.

using target-domain images, real-world rollouts, demonstrations, or real-world fine-tuning. The physical maps differ from the simulation training map in layout scale, background appearance, illumination conditions, and robot-environment interaction. Therefore, the real-world evaluation directly tests whether the robustness acquired from the proposed policy-conditioned Fourier intervention transfers to unseen physical visual observations and Duckiebot dynamics.

In the lane-following task, XFD substantially outperforms both the base SAC and DR policies in travelled distance and reward on the two physical maps. Map 4 represents a relatively regular closed-loop layout, where the main challenge comes from real-world appearance changes, camera response, and actuation mismatch. In this setting, XFD reaches a distance of 10.26, compared with 1.37 for the base policy and 4.06 for DR, indicating that the proposed policy can maintain stable lane tracking after direct deployment. Map 5 is more challenging because its compact layout contains consecutive

turns and sharper geometric transitions. The base and DR policies terminate early on this map, reaching only 0.77 and 2.91 in distance, respectively, whereas XFD maintains a distance of 9.63. This result shows that the proposed method does not merely adapt to an easy physical layout, but also preserves closed-loop control robustness when the vehicle must continuously adjust its heading through successive turning segments. The reward follows the same trend: the base policy receives negative returns on both maps, DR provides limited improvement, and XFD achieves the highest positive reward. These results demonstrate that the proposed method enables the obstacle-aware policy to maintain robust driving behaviour, both with and without obstacles, despite substantial sim-to-real discrepancies arising from real-world visual shifts, lighting variations, texture discrepancies, background clutter, and differing physical map geometries. This highlights the ability of the proposed framework to learn task-relevant representations that generalise across diverse visual conditions and deploy-

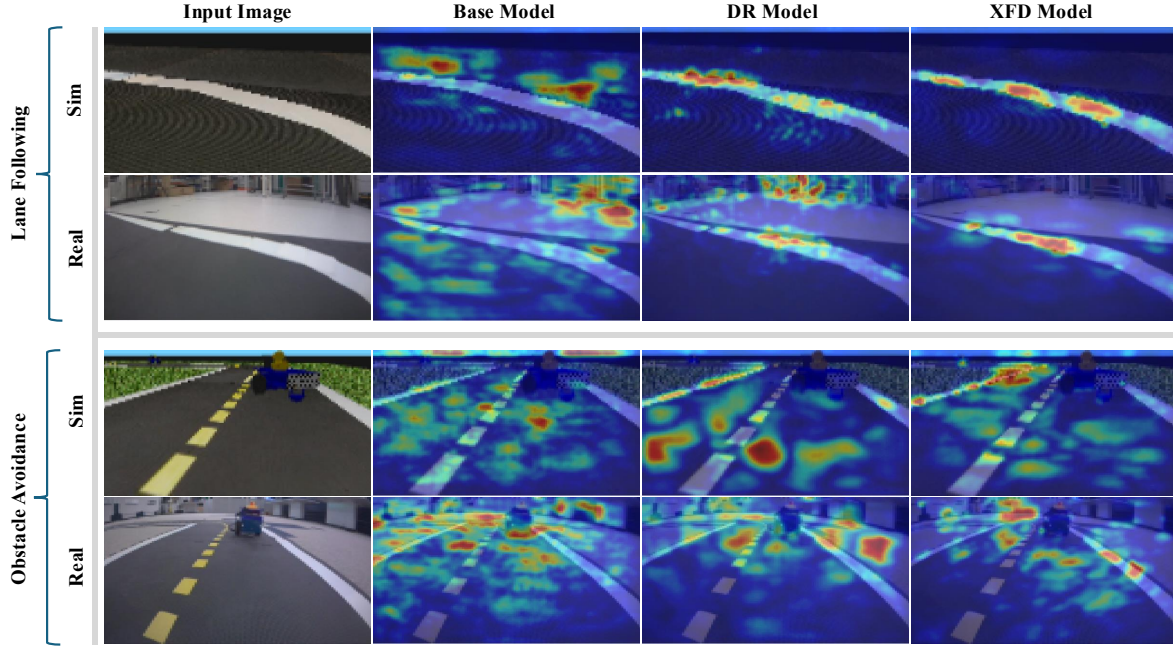


Fig. 7. Attribution map comparisons among the base model, DR model, and XFD model under two evaluation tasks.

ment environments, supporting robust zero-shot transfer from simulation to the physical Duckiebot platform. Furthermore, despite being trained under the guidance of the DR teacher policy, XFD consistently outperforms its teacher across both simulation and real-world evaluations, demonstrating that the proposed framework facilitates policy improvement beyond the capabilities of the teacher rather than merely replicating its behaviour.

The advantage of XFD is also evident in the obstacle-avoidance task. On Map 4, XFD improves the travelled distance from 1.42 for the base policy and 2.77 for DR to 9.95. On the most difficult Map 5, the distance increases from 0.98 and 2.85 to 9.77. The reward is also markedly improved, rising from negative or near-zero values for the baselines to 482.25 on Map 4 and 411.28 on Map 5. These results demonstrate that XFD enables robust transfer of both lane-following and obstacle-avoidance behaviours to the physical Duckiebot, maintaining reliable performance even in previously unseen environments with consecutive turns and increased navigational complexity. The trajectory visualisation in Fig. 6 further reveals a common failure pattern of the base and DR policies. In both physical maps, the baseline policies tend to terminate early around obstacle-interaction or turning regions, where the vehicle must simultaneously maintain lane alignment, avoid the obstacle, and recover to the lane. DR improves visual diversity during training, but its real-world trajectories still show unstable recovery and limited progress after obstacle encounters. In contrast, XFD produces longer and more continuous trajectories across both obstacle settings, suggesting that the proposed policy-conditioned Fourier intervention helps the student preserve task-relevant lane and obstacle cues under real appearance changes.

Averaged over the two real-world maps, XFD improves the

lane-following distance from 1.07 for the base policy and 3.49 for DR to 9.95. In obstacle avoidance, the average distance increases from 1.20 and 2.81 to 9.86. Although DR sometimes achieves lower deviation or higher smoothness, its travelled distance and reward remain much lower than those of XFD, indicating that these smoother or more centred trajectories are often associated with conservative motion or early termination rather than successful task completion. Together with Fig. 6, these results demonstrate that the proposed XFD largely improves zero-shot sim-to-real transfer by enabling the trained policy to maintain progress, recover from obstacle interactions, and remain robust across different physical map geometries.

C. XAI Attribution Analysis

To further investigate the decision-making behavior of the learned visual policies, we visualise the attribution responses of the baseline SAC model, the domain-randomised model, and the XFD model under both simulation and real-world observations. The visualisation is shown in Fig. 7. The first column presents the input observations, while the remaining columns show the attribution maps generated by different policies. Brighter regions indicate image areas that contribute more strongly to the corresponding control output.

As shown in Fig. 7, the baseline model produces less consistent attribution responses after the transition from simulation to real-world observations. In lane following, its highlighted regions are partially shifted away from the lane structure under real observations, indicating that the policy may rely on visual cues that are sensitive to appearance changes. The domain-randomised model improves visual coverage to some extent, but its attribution is still relatively scattered and includes responses to background or texture-dependent regions. In

contrast, the proposed model maintains more concentrated attribution around lane boundaries and road geometry in both simulation and real-world cases. This suggests that the XAI-guided Fourier intervention encourages the student policy to preserve control-relevant visual evidence under visual domain shifts.

A similar tendency can be observed in the obstacle-avoidance task. Compared with lane following, obstacle avoidance requires the policy to jointly process lane geometry, road boundaries, and obstacle-related cues. The baseline and DR models show broader and less stable attribution patterns across simulation and real-world observations, indicating that their action decisions are more easily affected by background texture, appearance changes, or local visual distractors. In contrast, XFD maintains stronger responses around task-relevant regions, including lane markings, road boundaries, and obstacle-related visual cues. This provides a visual explanation for the simulation results in Section V-A: even before real-world deployment, the student policy improves travelled distance and reward because it learns to rely on more control-relevant evidence rather than merely following the teacher or fitting simulator-specific appearance.

The same attribution trend also supports the real-world deployment results in Section V-B. When the observation changes from simulation to physical Duckietown maps, the baseline and DR policies exhibit less consistent evidence selection, which is consistent with their early termination and limited travelled distance in real experiments. By contrast, XFD preserves a more stable attribution structure between simulated and real observations. This suggests that the proposed Fourier intervention improves transfer not only by increasing visual diversity during training, but also by constraining the student policy to keep its action selection grounded in task-critical structures. Therefore, the attribution analysis explains both levels of performance gain: in simulation, XFD improves policy learning beyond the teacher-guided reference; in real-world deployment, it preserves the same task-relevant visual evidence under unseen appearance, lighting, and map-layout changes.

It should be noted that attribution quality is not a direct substitute for closed-loop performance, and a visually plausible attribution map does not by itself guarantee successful control. Nevertheless, prior XAI and explanation-guided learning studies suggest that attribution analysis can reveal whether a policy relies on task-relevant or spurious visual evidence [21], [22], [33]. In this work, the quantitative results and attribution maps are consistent: the model with more stable and task-focused visual evidence also achieves longer travelled distance and higher reward in both simulation and real-world deployment. Therefore, the analysis does not claim a universal causal relation between visual explanation and performance, but provides empirical evidence that preserving task-relevant visual evidence is an important factor behind the improved closed-loop robustness of XFD.

VI. CONCLUSION

This paper presented XFD, an XAI-guided Fourier domain generalisation framework for target-domain-free sim-to-real

reinforcement learning. By leveraging DeepLIFT attributions to regulate Fourier-domain perturbations, the proposed approach generates robustness-enhancing visual domains entirely within simulation, enabling policy learning under structured visual shifts while preserving task-relevant spatial information. To support robust learning, a teacher-student training strategy was introduced, where the teacher provides attribution guidance and training stabilisation while the student remains optimised through the reinforcement learning objective.

Extensive evaluations in simulation and zero-shot real-world deployment demonstrated that XFD consistently outperforms both standard SAC and random domain randomisation baselines across lane-following and obstacle-avoidance tasks. The results demonstrate the complementary benefits of the proposed XAI-guided Fourier intervention and robust policy optimisation strategy. While the Fourier intervention generates policy-relevant visual domains that improve visual generalisation beyond random domain randomisation, the robust optimisation framework enables effective learning under these challenging visual shifts. In combination, XFD improved significantly in closed-loop robustness and sim-to-real transfer across previously unseen environments. Notably, XFD consistently surpasses the DR teacher policy from which it learns in simulation, demonstrating that the teacher serves as a source of guidance and stabilisation rather than a performance upper bound.

Attribution analyses further revealed that XFD promotes more stable and task-focused visual evidence across simulation and real-world observations, supporting the hypothesis that effective sim-to-real transfer should regulate visual shifts according to policy-relevant features rather than relying on indiscriminate visual perturbations. Future work will extend the proposed framework to more complex dynamic environments, larger-scale deployment scenarios, and online attribution-guided adaptation for continual robust policy learning.

REFERENCES

- [1] B. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. Sallab, S. Yogamani, and P. Perez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, Jun. 2022.
- [2] S. Aradi, "Survey of deep reinforcement learning for motion planning of autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, p. 740–759, Feb. 2022. [Online]. Available: <https://doi.org/10.1109/TITS.2020.3024655>
- [3] J. Chen, S. E. Li, and M. Tomizuka, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5068–5078, 2022.
- [4] C. Wang and N. Aouf, "Fusion attention network for autonomous cars semantic segmentation," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, 2022, pp. 1525–1530.
- [5] A. Amini, I. Gilitschenski, J. Phillips, J. Moseyko, R. Banerjee, S. Karaman, and D. Rus, "Learning robust control policies for end-to-end autonomous driving from data-driven simulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1143–1150, 2020.
- [6] H. Zhang, Y. Liu, Y. Jiang, W. Lin, and C. Ye, "Depth-pc: Sim-to-real transfer for zero-shot visual servoing via cross-modal fusion," *IEEE Robotics and Automation Letters*, vol. 10, no. 11, pp. 11976–11983, 2025.
- [7] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=IQdXeXDoWtI>

- [8] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," 2018. [Online]. Available: <https://arxiv.org/abs/1801.01290>
- [9] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," 2017. [Online]. Available: <https://arxiv.org/abs/1703.06907>
- [10] G. D. I, I. G. R, M. LM, B. O, and F. L. D, "Sim-to-real transfer and reality gap modeling in model predictive control for autonomous driving," *APPLIED INTELLIGENCE*, vol. 53, no. 10, p. 12719–12735, 2023. [Online]. Available: <https://link.springer.com/article/10.1007/s10489-022-04148-1>
- [11] F. Muratore, F. Ramos, G. Turk, W. Yu, M. Gienger, and J. Peters, "Robot learning from randomized simulations: A review," *CoRR*, vol. abs/2111.00956, 2021. [Online]. Available: <https://arxiv.org/abs/2111.00956>
- [12] A. Carlson, K. A. Skinner, R. Vasudevan, and M. Johnson-Roberson, "Sensor transfer: Learning optimal sensor effect image augmentation for sim-to-real domain adaptation," *CoRR*, vol. abs/1809.06256, 2018. [Online]. Available: <http://arxiv.org/abs/1809.06256>
- [13] Y. Chebotar, A. Handa, V. Makovychuk, M. Macklin, J. Issac, N. Ratliff, and D. Fox, "Closing the sim-to-real loop: Adapting simulation randomization with real world experience," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8973–8979.
- [14] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," 2017. [Online]. Available: <https://arxiv.org/abs/1709.07857>
- [15] O. M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020. [Online]. Available: <https://doi.org/10.1177/0278364919887447>
- [16] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," 2020. [Online]. Available: <https://arxiv.org/abs/2004.05498>
- [17] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A fourier-based framework for domain generalization," 2021. [Online]. Available: <https://arxiv.org/abs/2105.11120>
- [18] D. Peng, Y. Lei, L. Liu, P. Zhang, and J. Liu, "Global and local texture randomization for synthetic-to-real semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 6594–6608, 2021.
- [19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct. 2019. [Online]. Available: <http://dx.doi.org/10.1007/s11263-019-01228-7>
- [20] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," 2019. [Online]. Available: <https://arxiv.org/abs/1704.02685>
- [21] C. Wang and N. Aouf, "Explainable deep adversarial reinforcement learning approach for robust autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 10, no. 4, pp. 2551–2563, 2025.
- [22] Y. Gao, S. Gu, J. Jiang, S. R. Hong, D. Yu, and L. Zhao, "Going beyond xai: A systematic survey for explanation-guided learning," *ACM Comput. Surv.*, vol. 56, no. 7, Apr. 2024. [Online]. Available: <https://doi.org/10.1145/3644073>
- [23] H. Wang, H. Luo, W. Zhang, and H. Chen, "Cts: Concurrent teacher-student reinforcement learning for legged locomotion," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9191–9198, 2024.
- [24] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," 2017. [Online]. Available: <https://arxiv.org/abs/1612.07828>
- [25] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," 2017. [Online]. Available: <https://arxiv.org/abs/1612.05424>
- [26] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," 2017. [Online]. Available: <https://arxiv.org/abs/1711.03213>
- [27] S. Pasios and N. Nikolaidis, "Carla2real: A tool for reducing the sim2real appearance gap in carla simulator," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 11, p. 18747–18761, Nov. 2025. [Online]. Available: <http://dx.doi.org/10.1109/TITS.2025.3597010>
- [28] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2018, p. 3803–3810. [Online]. Available: <http://dx.doi.org/10.1109/ICRA.2018.8460528>
- [29] F. Muratore, C. Eilers, M. Gienger, and J. Peters, "Data-efficient domain randomization with bayesian optimization," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 911–918, 2021.
- [30] B. Mehta, M. Diaz, F. Golemo, C. J. Pal, and L. Paull, "Active domain randomization," *CoRR*, vol. abs/1904.04762, 2019. [Online]. Available: <http://arxiv.org/abs/1904.04762>
- [31] OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang, "Solving rubik's cube with a robot hand," 2019. [Online]. Available: <https://arxiv.org/abs/1910.07113>
- [32] G. Tiboni, P. Klink, J. Peters, T. Tommasi, C. D'Eramo, and G. Chalkatzaki, "Domain randomization via entropy maximization," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=GXTmviVrOM>
- [33] A. Kuznetsov, B. Geyvna, C. Wang, S. Peters, and S. V. Albrecht, "Explainable ai for safe and trustworthy autonomous driving: A systematic review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 19342–19364, 2024.
- [34] M. T. Ribeiro, S. Singh, and C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016. [Online]. Available: <http://arxiv.org/abs/1602.04938>
- [35] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.07874, 2017. [Online]. Available: <http://arxiv.org/abs/1705.07874>
- [36] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *CoRR*, vol. abs/1703.01365, 2017. [Online]. Available: <http://arxiv.org/abs/1703.01365>
- [37] R. Zuo, S. Khan, Z. Wang, G. E. Katz, and Q. Qiu, "Why the agent made that decision: Contrastive explanation learning for reinforcement learning," 2025. [Online]. Available: <https://arxiv.org/abs/2411.16120>
- [38] J. Guan, G. Chen, J. Huang, Z. Li, L. Xiong, J. Hou, and A. Knoll, "A discrete soft actor-critic decision-making strategy with sample filter for freeway autonomous driving," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 2, pp. 2593–2598, 2023.
- [39] S. Teng, L. Chen, Y. Ai, Y. Zhou, Z. Xuanyuan, and X. Hu, "Hierarchical interpretable imitation learning for end-to-end autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 673–683, 2023.
- [40] P. Morere, L. Ott, and F. Ramos, "Learning to plan hierarchically from curriculum," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2815–2822, 2019.