Artículo original

Redes Generativas Adversariales (GANs) para Imágenes de Profundidad: Reducción de la Brecha Simulación-Realidad en la Percepción de UAVs

Generative Adversarial Networks (GANs) for Depth Image Transformation: Reduction of Simulation-to-Reality Gap in UAV Perception

Pablo José Salazar Villacis: Loughborough University, PhD in Computer Science,
https://orcid.org/0000-0001-7137-477X
Autor de correspondencia: p.salazar-villacis@lboro.ac.uk

DOI: *https://doi.org/10.64424/rcu41202560*

Recibido: 26 febrero 2025
Publicado: 12 marzo 2025

Resumen:
Los vehículos aéreos no tripulados (UAV) dependen de la percepción de profundidad para la navegación autónoma y la evasión de obstáculos. Sin embargo, los modelos entrenados en simulación tienen dificultades para generalizar debido a la brecha entre imágenes de profundidad sintéticas y reales, causada por diferencias en el ruido del sensor, la variabilidad del entorno y las texturas de los objetos, lo que reduce su eficacia en aplicaciones reales. Este estudio aborda la adaptación de dominio mediante redes generativas adversariales (GAN) para transformar imágenes de profundidad simuladas en representaciones más realistas. Se implementan dos enfoques: Pix2Pix, un modelo supervisado que requiere datos emparejados, y CycleGAN, un método no supervisado que adapta imágenes sin correspondencias directas. Para una evaluación rigurosa, se construye un conjunto de datos alineado con imágenes sintéticas y reales.Los resultados muestran que Pix2Pix supera a CycleGAN en la replicación de características de profundidad del mundo real al minimizar errores de intensidad, mientras que CycleGAN, aunque conserva la geometría, tiene dificultades para modelar el ruido del sensor. La adaptación adversarial reduce significativamente la brecha simulación-realidad, mejorando la precisión de la imagen de profundidad para la percepción de UAV. Para validar su aplicabilidad, las imágenes adaptadas se integran en el Sistema Operativo de Robots (ROS), permitiendo la percepción en tiempo real. Los hallazgos demuestran que la adaptación de dominio basada en GAN mejora la visión robótica basada en profundidad, facilitando una navegación más fiable de los UAV en entornos complejos.

Palabras claves: Adaptación de dominio, Imágenes de profundidad, Redes Generativas Adversariales, Brecha de simulación a realidad.

Abstract:
Unmanned Aerial Vehicles (UAVs) rely on depth perception for autonomous navigation and obstacle avoidance. However, models trained in simulation often struggle to generalize due to the domain gap between synthetic and real depth images. This gap results from differences in sensor noise, environmental variability, and object textures, reducing the effectiveness of simulation-trained models in real-world applications. This paper explores domain adaptation using Generative Adversarial Networks (GANs) to transform simulated depth images into more realistic counterparts. Two approaches are implemented: Pix2Pix, a supervised model requiring paired datasets, and CycleGAN, an unsupervised method that adapts images without paired samples. A dataset of aligned synthetic and real-world depth images is constructed to enable robust evaluation. Results show that Pix2Pix outperforms CycleGAN in replicating real-world depth characteristics by minimizing depth intensity errors, while CycleGAN, despite preserving object geometry, struggles to model sensor noise. The adversarial adaptation method significantly reduces the simulation-to-reality gap, improving depth image accuracy for UAV perception.To validate real-world applicability, the adapted depth images are integrated into the Robot Operating System (ROS), enabling real-time UAV perception. The findings demonstrate that GAN-based domain adaptation enhances depth-based robotic vision, facilitating more reliable UAV navigation in complex environments.

Keywords: Domain Adaptation, Depth Images, Generative Adversarial Networks, Simulation-to-Reality Gap.

Como citar la obra: Salazar Villacis, P. J. (2025). Adaptación de Dominio Adversarial para Imágenes de Profundidad: Reduciendo la Brecha Simulación-Realidad en la Percepción de UAVs. *Revista Científica Unanchay, 4*(1), 1-20, doi: https://doi.org/10.64424/rcu41202560

Introduction

Unmanned Aerial Vehicles (UAVs) have emerged as a crucial technology in robotics, enabling applications in exploration, indoor navigation, object transportation, and environmental mapping (Al Radi et al., 2024). Their ability to operate in three-dimensional (3D) environments makes them particularly well-suited for tasks that require autonomous perception and navigation in unstructured scenarios. However, robust UAV navigation in real-world environments remains challenging due to sensor limitations, perception inaccuracies, and the need for reliable adaptation from simulated training environments to real-world deployment (Xu et al., 2024).

Deep learning techniques have significantly advanced robotic perception, facilitating improved object detection, scene understanding, and autonomous navigation (Le et al., 2024; Wu et al., 2019). Many of these methods rely on training in simulated environments, where large datasets can be generated efficiently and safely. However, a persistent issue in transferring learned models to real-world applications is the domain gap between synthetic and real data. This gap arises from differences in sensor noise, lighting conditions, object textures, and environmental variability, leading to poor generalization of simulation-trained models when deployed in reality (Sadeghi and Levine, 2017; Sampedro et al., 2018). Addressing this issue is essential to enable the deployment of robust robotic systems capable of operating reliably across different environments. This study hypothesises that GAN-based domain adaptation methods, particularly Pix2Pix, can significantly reduce perceptual errors in depth images, thereby improving UAV navigation in real environments.

The domain adaptation problem has been widely studied in the context of robotic vision and deep learning. UAV navigation systems heavily rely on perception mechanisms that integrate depth images to understand spatial constraints and detect obstacles (Jing Chen et al., 2016; Sikang Liu et al., 2016). While various works have explored domain adaptation for image classification and semantic segmentation, the problem of adapting depth images remains underexplored. Prior research has attempted to mitigate the reality gap through physics-based simulations, domain randomization, and adversarial learning techniques (Westerski and Teck, 2023). Among these, Generative Adversarial Networks (GANs) have demonstrated promising capabilities in translating images from one domain to another while preserving structural consistency (Goodfellow et al., 2014). Early works on domain adaptation applied GANs to generate realistic textures from synthetic images, yielding improvements in object detection and scene reconstruction (Xu et al., 2023). However, the adaptation of depth images, particularly for UAV perception, has not been fully addressed.

Depth cameras play a fundamental role in UAV perception, providing crucial spatial awareness for obstacle avoidance and trajectory planning. However, simulated depth images often fail to capture the full complexity of real-world depth sensors, including noise artifacts and non-uniform depth distribution (Tzeng et al., 2020). These discrepancies contribute to the sim-to-real gap, leading to degraded model performance when transitioning from simulation to real-world applications. Bridging this gap is crucial for ensuring the reliable deployment of UAVs in dynamic and unknown environments (James et al., 2019).

One promising approach to mitigating this challenge is domain adaptation, which enables models trained on synthetic data to generalize more effectively to real-world conditions. In this work, we investigate domain adaptation techniques based on Generative

ISSN: 2953-6707

Adversarial Networks (GANs) to transform simulated depth images into their real-world counterparts (James and Johns, 2016). The goal is to reduce the perceptual discrepancy between synthetic and real depth images, thereby improving the accuracy and robustness of depth-based perception models for UAV navigation.

This paper presents a comparative study of two domain adaptation approaches for depth image transformation: Pix2Pix, a supervised method requiring paired depth images for training (Isola et al., 2017), and CycleGAN, which enables adaptation without the need for paired datasets(Zhu et al., 2017). To ensure accurate ground truth for training and evaluation, a dataset of aligned synthetic and real-world depth images is constructed. A key aspect of this study is the quantitative and qualitative assessment of the adapted depth images, focusing on how effectively they replicate real-world depth characteristics. Results indicate that the adversarial-based Pix2Pix model significantly reduces the adaptation error compared to the reconstruction-based CycleGAN approach, which, while preserving object geometry, struggles to replicate the noise characteristics inherent to real depth sensors.

Furthermore, to demonstrate the real-world applicability of these adaptation techniques, the generated depth images are integrated into the Robot Operating System (ROS), allowing real-time perception for UAVs (Quigley et al., 2019). The performance of the adapted images is analysed through error distribution comparisons, revealing that adversarial-based adaptation provides a notable improvement in depth perception accuracy. By improving the fidelity of depth perception, the proposed methods directly contribute to more reliable obstacle detection and trajectory planning, which are critical for safe and autonomous UAV navigation in unstructured environments. This ultimately leads to a measurable reduction in perception-related navigation errors during real-world deployment.

Páginas: 1-20

ISSN: 2953-6707

Methodology

This study follows a structured methodology to ensure a controlled and reproducible evaluation of domain adaptation techniques for depth image refinement. The approach involves constructing a high-fidelity simulation environment, carefully modelling real-world objects, and integrating real-time pose synchronisation to maintain consistency between simulated and physical setups. By implementing precise depth camera calibration and leveraging adversarial learning techniques, the study provides a robust framework for improving UAV perception. The evaluation focuses on the effectiveness of Pix2Pix and CycleGAN, two generative models with distinct adaptation strategies, in transforming simulated depth images into realistic counterparts. The methodological design allows for a direct comparison of their strengths and limitations, offering valuable insights into the trade-offs between paired and unpaired domain adaptation and their implications for real-world robotic applications.

Virtual Environment Construction and Depth Image Generation

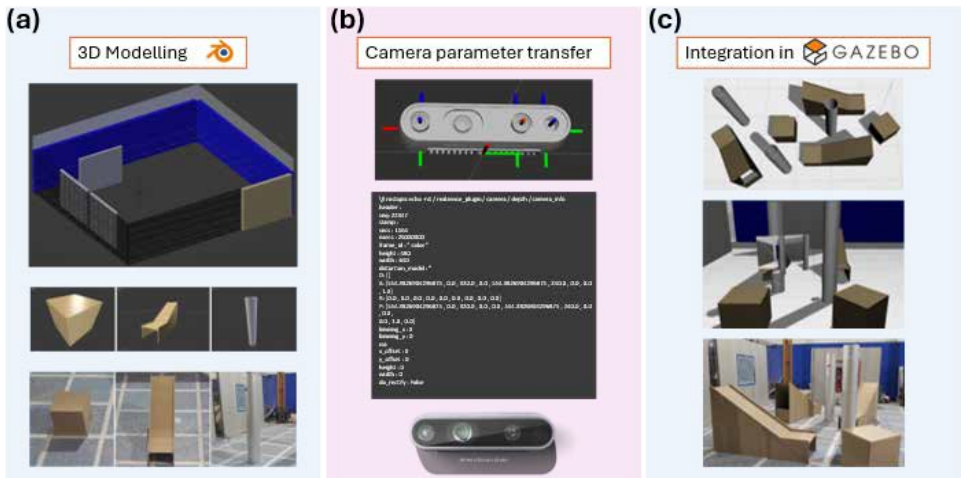
To facilitate domain adaptation for depth image refinement, a controlled virtual environment was developed to closely replicate real-world conditions. The simulation setup was designed to ensure that the generated depth images remained structurally consistent with those captured in real-world experiments. This required precise 3D modelling, camera calibration, and integration into a physics-based simulation environment, as illustrated in Figure 1.

Three key objects were selected for depth perception evaluation: a wooden cube, a cylindrical obstacle, and a ramp-like structure. These objects introduce diverse spatial patterns that enhance model generalisation, particularly for UAV navigation tasks. The wooden cube, measuring 60 cm, functions as a UAV takeoff and landing station. The cylindrical obstacle, standing 2 metres tall with a 30.5 cm diameter, mimics structural elements commonly found in urban and industrial environments. The ramp, composed of segmented components, introduces inclined surfaces that simulate uneven terrains. These objects were precisely modelled in Blender using Boolean operations where necessary to create hollow structures and merge components (Figure 1a). To enhance realism, UV mapping was used to apply textures that closely match real-world materials, ensuring that the visual and depth properties remained consistent across domains. The wooden cube and ramp were assigned wood-textured surfaces, while the cylindrical obstacle was colour-matched to its real-world counterpart. The final models were exported in .dae format for seamless integration into the simulation.

ISSN: 2953-6707

Figure 1

Virtual environment setup for depth image generation.



Source: Author (2025). (a) 3D modelling of the Robotics Arena and key objects in Blender. (b) Camera parameter transfer for accurate RealSense D435 simulation. (c) Integration of objects into Gazebo for realistic depth image acquisition.

The virtual testing environment, referred to as the Robotics Arena, was developed within Gazebo, a widely used physics-based simulation platform (Figure 1c). The spatial configuration of this environment was carefully aligned with the real-world experimental setup, ensuring one-to-one correspondence between the placement of physical and virtual objects. Structural elements such as safety nets and walls were incorporated to maintain environmental consistency between the real and simulated setups.

A critical component of the simulation environment was the integration of a depth sensor to generate synthetic depth images that accurately reflect real-world depth perception. The Intel RealSense D435 was selected due to its compact design and maximum sensing range of 10 metres, making it ideal for UAV applications. Since Gazebo only provides a default RealSense R200 model, a customised RealSense D435 sensor was implemented to ensure an accurate simulation of its optical properties. A Gazebo plugin was developed to replicate the real camera's intrinsic parameters, including focal length, baseline distance, and depth sensing characteristics (Figure 1b).

To maintain consistency between real and simulated depth images, the horizontal field of view of the virtual camera was calculated based on the intrinsic parameters of the RealSense D435. This calibration process ensured that the depth images captured in the simulation accurately reflected those obtained in real-world experiments. By integrating a properly calibrated depth sensor, the study enabled direct comparisons between generated and real-world depth images, forming the foundation for effective domain adaptation.

Pose Synchronisation and Depth Image Acquisition for Domain Adaptation

The alignment of real and simulated objects was critical for ensuring consistency in depth image acquisition for domain adaptation. The OptiTrack motion capture system was employed to track the real-world positions of objects using retroreflective markers, enabling precise pose synchronisation between physical and virtual environments. The Motive software processed these marker-based detections, transmitting rigid body pose data to the Robot Operating System (ROS) through the vrpn_client_ros package. This ensured continuous real-time synchronisation, allowing objects in simulation to be accurately positioned to match their real-world counterparts.

Páginas: 1-20 ISSN: 2953-6707

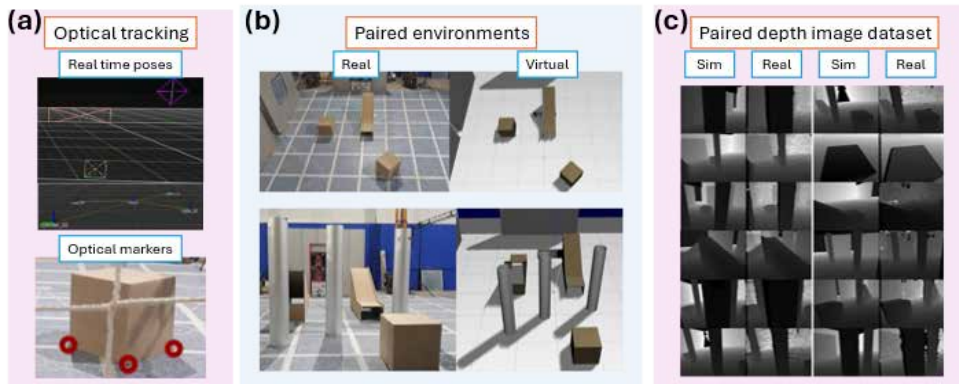
A ROS-based service was implemented to automatically spawn 3D models in Gazebo using the real-world pose data. By subscribing to OptiTrack updates, this service ensured that the placement of objects in the virtual environment remained consistent with the physical setup. This synchronisation was essential for acquiring paired depth images, which formed the foundation for supervised domain adaptation with Pix2Pix. Without this level of environmental fidelity, inconsistencies between simulated and real-world conditions would limit the adaptation capability of the trained models.

The process of depth image acquisition was tailored to support both paired and unpaired domain adaptation techniques. A dataset of depth images was collected, ensuring that each simulated image had a directly corresponding real-world counterpart. This was achieved by continuously tracking the Intel RealSense D435 camera's pose using OptiTrack and replicating its position in simulation through the /gazebo/set_model_state service. Depth images were synchronised and stored using a ROS bag file, capturing a range of experimental conditions.

The dataset was structured to support the different requirements of the domain adaptation models. Pix2Pix, which relies on supervised learning, was trained using paired depth images to learn direct pixel-wise transformations between simulated and real depth domains. In contrast, CycleGAN was trained on unpaired depth images, learning to perform synthetic-to-real depth translation without explicit one-to-one correspondences. This distinction in training strategies provided a comparative framework for evaluating the advantages and limitations of paired versus unpaired adaptation methods.

Figure 2

Pose synchronisation and depth image acquisition.



Source: Author (2025). (a) OptiTrack enables real-time pose tracking. (b) Paired environments ensure alignment between real and simulated setups. (c) Paired depth images form the basis for supervised domain adaptation.

The dataset was structured to support the different requirements of the domain adaptation models. Pix2Pix, which relies on supervised learning, was trained using paired depth images to learn direct pixel-wise transformations between simulated and real depth domains. In contrast, CycleGAN was trained on unpaired depth images, learning to perform synthetic-to-real depth translation without explicit one-to-one correspondences. This distinction in training strategies provided a comparative framework for evaluating the advantages and limitations of paired versus unpaired adaptation methods.

ISSN: 2953-6707

The alignment process, depth acquisition setup, and structured dataset are illustrated in Figure 2a shows the OptiTrack optical tracking system used for real-time pose synchronisation. Figure 2b presents the paired real and simulated environments, demonstrating the accuracy of object placement between domains. Figure 2c displays a sample of the paired depth image dataset, highlighting the correspondence between simulated and real-world depth images, which was essential for evaluating the performance of the adaptation models.

Generative Models for Depth Image Adaptation

To bridge the simulation-to-reality gap, this study implemented two domain adaptation techniques: Pix2Pix and CycleGAN, both of which leverage Generative Adversarial Networks (GANs). These models were trained to refine synthetic depth images, making them more representative of real-world sensor outputs. They were selected due to their effectiveness in image-to-image translation tasks and their contrasting training paradigms: one supervised and the other unsupervised. Adversarial domain adaptation methods such as CoGAN (Liu and Tuzel, 2016), SimGAN (Shrivastava et al., 2017), and PixelDA (Bousmalis et al., 2017) have introduced strategies like coupled discriminators, self-regularisation, and noise conditioning to enhance the realism of generated images. However, these models often require more complex architectural configurations and may struggle to preserve structural features that are critical for depth-based robotic perception. Pix2Pix and CycleGAN were therefore selected to provide a balanced evaluation of paired versus unpaired training regimes while maintaining architectural simplicity and structural fidelity.

Pix2Pix is a supervised GAN-based model that learns a direct mapping from synthetic to real depth images using paired training data. The model architecture consists of a U-Net-based generator, which transforms simulated depth images into their real-world counterparts, and a PatchGAN discriminator, which evaluates the realism of the generated images. The training process optimises adversarial loss, encouraging the generator to produce more realistic images, and L1 loss, ensuring that translated images remain structurally consistent with ground truth. This combination allows Pix2Pix to retain fine structural details while improving the realism of synthetic images.

In contrast, CycleGAN is an unsupervised GAN-based model that learns bidirectional mappings between synthetic and real domains without requiring paired training data. Instead of relying on direct pixel-wise correspondences, CycleGAN introduces cycle consistency loss, ensuring that an image translated to the target domain and then mapped back retains its original structure. The model consists of two generators and two discriminators, learning transformations between synthetic and real domains in both directions. This flexibility allows CycleGAN to adapt synthetic images to real-world characteristics even when paired datasets are unavailable. However, due to the lack of direct supervision, CycleGAN may introduce structural inconsistencies and fail to replicate certain depth features.

While both models are based on adversarial learning and aim to reduce the perceptual gap between synthetic and real domains, they differ significantly in their underlying mechanisms and training requirements. Pix2Pix relies on supervised learning

ISSN: 2953-6707

with paired data and uses a combination of adversarial and pixel-wise losses to ensure both realism and structural fidelity. In contrast, CycleGAN adopts an unsupervised approach, introducing cycle-consistency loss to preserve content in the absence of paired samples. This distinction not only affects training strategies but also influences the models' ability to replicate sensor-specific noise and geometric accuracy. Theoretical correlation between the two lies in their shared generative framework, yet their architectural configurations and learning objectives reflect different trade-offs between data alignment, realism, and generalisation.

After training, both models were integrated into ROS, enabling real-time transformation of depth images. The adapted depth images were published as ROS topics and visualised using rqt_image_view, allowing UAVs operating in simulation to process depth images that closely resemble real-world sensor outputs.

Experimental Setup and Model Evaluation

A structured experimental framework was designed to evaluate the performance of Pix2Pix and CycleGAN. The evaluation focused on depth accuracy, noise replication, and geometric structure preservation. The models' ability to reduce discrepancies between simulated and real-world depth images was assessed through quantitative error analysis and qualitative structural comparisons.

A dataset of 2378 depth images was used to evaluate Pix2Pix, while CycleGAN was tested on 603 images, with the difference in dataset size attributed to the higher computational cost of CycleGAN inference. Each dataset consisted of simulated, generated, and real depth images, with depth profiles extracted along the centre row of each image (640 pixels). The absolute depth errors were computed to quantify the improvements achieved by each model.

For statistical validation, a Wilcoxon signed-rank test was conducted to determine whether the observed error reductions were statistically significant. Additionally, Cliff's Delta was used to assess the practical significance of these reductions. The evaluation process combined histogram and boxplot visualisations with qualitative depth image assessments, providing a comprehensive analysis of each model's effectiveness in depth adaptation.

Páginas: 1-20

ISSN: 2953-6707

Results

This section evaluates Pix2Pix and CycleGAN in refining simulated depth images by analyzing error reduction and visual realism. Error distributions are quantified using histograms, boxplots, and statistical tests, including the Wilcoxon signed-rank test and Cliff's Delta. Representative images—corresponding to minimum, median, and maximum MedAE—are assessed through radar charts and depth comparisons for structural accuracy. A final comparison highlights Pix2Pix's advantage in paired supervision versus CycleGAN's reliance on unpaired training, determining which method better refines depth maps for applications in robotics and autonomous systems.

Performance of the Pix2Pix Model

The results of this study are structured into two primary analyses: the evaluation of error distributions across the dataset and the qualitative assessment of depth image reconstruction. The first analysis examines the extent to which the generative model reduces the discrepancies between simulated and real depth values, while the second analysis focuses on the visual realism and structural integrity of the generated depth images. To quantify the performance of the Pix2Pix model, we employ the Median Absolute Error (MedAE) as the primary metric, which provides a robust measure of error by computing the median of the absolute differences between the predicted and real depth values. MedAE is particularly useful in this context as it mitigates the influence of outliers, ensuring a more stable evaluation of the generative model's performance.

The boxplots of error distributions, presented in Figure 3d, illustrate the variations in absolute error across the three selected images corresponding to the minimum, median, and maximum MedAE cases. A significant reduction in error magnitude is evident when comparing the Real-Gen errors with the Sim-Real errors, indicating that the Pix2Pix model effectively refines depth information and brings the generated outputs closer to real-world depth representations.

Across the analysed dataset, the Real-Gen errors remain consistently lower than the Sim-Real errors, confirming that the generative model significantly reduces the gap between the original simulated depth maps and the corresponding real depth images. The reduction in error is most pronounced in the images associated with the minimum and median MedAE cases, where the generated depth maps closely approximate the real depth values. In contrast, the maximum MedAE case demonstrates a less effective correction, with larger residual errors remaining even after generation.

A statistical summary of the absolute errors further supports these observations. In the case of the minimum MedAE image, the Real-Gen error exhibits a median of 0.0392 meters, whereas the corresponding Sim-Real error has a median of 1.7255 meters. The reduction is also evident in the median MedAE image, where the Real-Gen median error is 0.4314 meters, compared to 1.0588 meters for Sim-Real. In the case of the maximum MedAE image, while the Pix2Pix model still reduces the error, the Real-Gen median error is 4.9804 meters, which remains substantially high. The interquartile range for the Real-Gen errors in this case is notably larger, indicating a greater degree of variance and confirming that the generative model exhibits reduced effectiveness in highly complex scenes.

For the minimum MedAE case, the generated depth profiles closely match real data, effectively correcting distortions, especially in smooth surfaces and well-defined boundaries, making Pix2Pix effective in structured scenes. In the median MedAE case, while depth estimates remain realistic, discrepancies appear at occlusion boundaries and

ISSN: 2953-6707

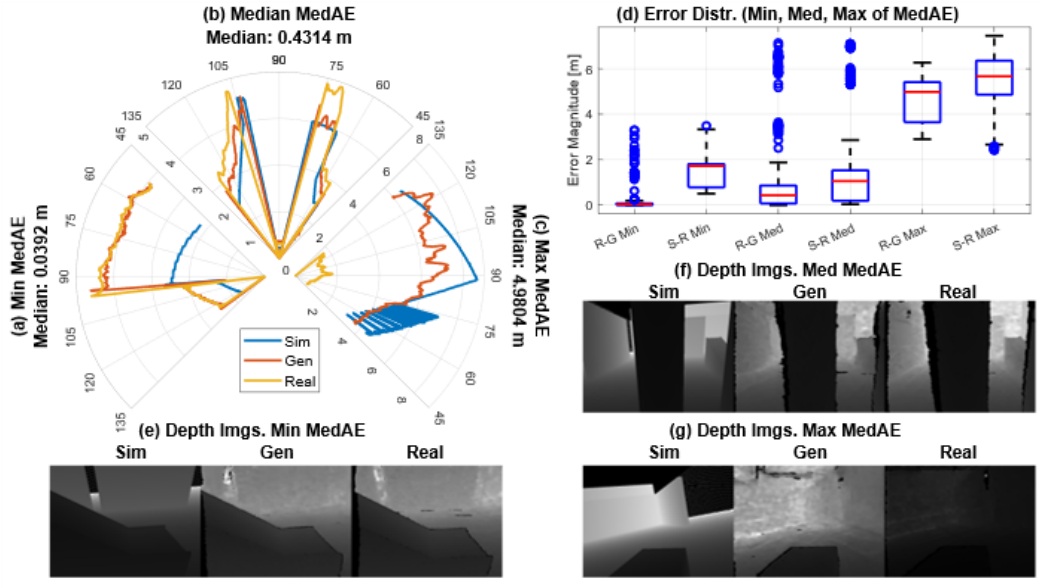
high-gradient regions, where the model struggles with depth ambiguities. In the maximum MedAE case, deviations become pronounced, with over-smoothing in high-gradient transitions, leading to a loss of fine geometric details and highlighting the model's limitations in handling complex structural changes.

The qualitative assessment of generated depth images provides further insights into the strengths and weaknesses of the Pix2Pix model. Figure 3e, figure 3f and figure 3g present the simulated, generated, and real depth images corresponding to the minimum, median, and maximum MedAE cases, respectively.

For the minimum MedAE case, the generated depth profiles closely match real data, effectively correcting systematic distortions in smooth surfaces and well-defined object boundaries, suggesting Pix2Pix performs well in structured scenes. In the median MedAE case, while the generated depth profiles align with real data, discrepancies emerge at occlusion boundaries and high-gradient regions, with errors increasing near depth discontinuities, indicating challenges in resolving occlusions and textured surfaces. In the maximum MedAE case, deviations become more pronounced, with noticeable distortions and over-smoothing in high-gradient transitions, suggesting over-regularization that leads to a loss of fine geometric details, particularly in areas with depth discontinuities.

Figure 3

Consolidated error analysis for Pix2Pix.



Source: Author (2025). Radar charts in (a), (b), and (c) compare simulated (Sim), generated (Gen), and real (Real) depth profiles for the minimum, median, and maximum MedAE cases, showing how closely Gen aligns with Real. (d) Boxplots illustrate error distributions, highlighting the reduction in Real-Gen (R-G) error. (e), (f), and (g) display the corresponding depth images, demonstrating Pix2Pix's effectiveness in refining depth structure and noise characteristics.

This highlights a fundamental limitation of the Pix2Pix model in handling highly complex depth variations, where the generative approach fails to fully reconstruct the fine geometric features of the scene.

Performance of the CycleGAN Model

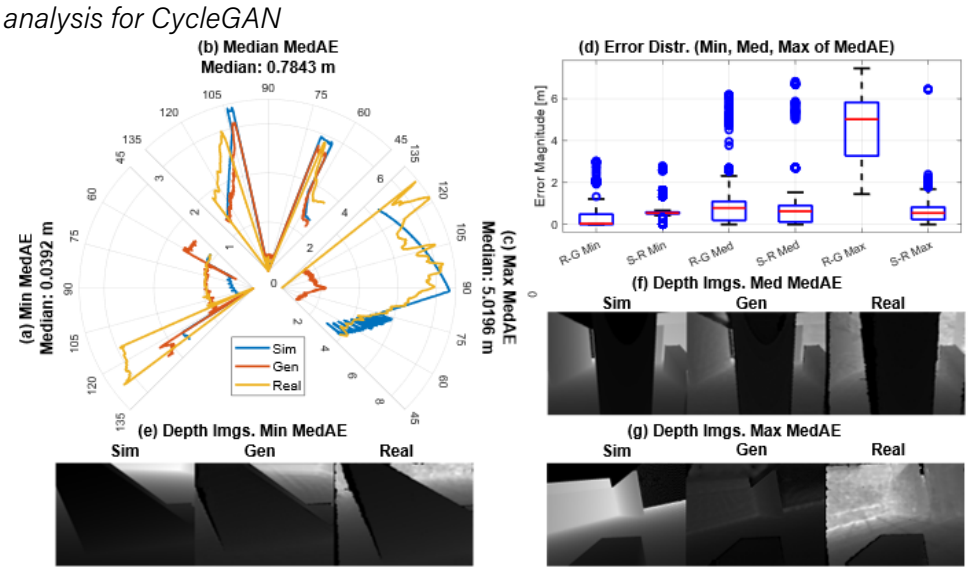
The assessment of the CycleGAN model's performance is structured around two key analyses: the evaluation of error distributions across the dataset and the qualitative examination of generated depth images. The first analysis aims to quantify the extent to which the CycleGAN model reduces discrepancies between simulated and real depth values. The second analysis focuses on the visual realism and structural consistency of the

ISSN: 2953-6707

generated depth maps, providing insights into the model's ability to replicate real-world depth distributions.

Figure 4

Consolidated error analysis for CycleGAN



Source: Prepared by the author (2025). . Radar charts in (a), (b), and (c) compare simulated (Sim), generated (Gen), and real (Real) depth profiles for the minimum, median, and maximum MedAE cases, showing limited improvement. (d) Boxplots indicate minimal reduction in Real-Gen error. (e), (f), and (g) present depth images, highlighting challenges in realistic depth reconstruction.

Performance of the CycleGAN Model

The assessment of the CycleGAN model's performance is structured around two key analyses: the evaluation of error distributions across the dataset and the qualitative examination of generated depth images. The first analysis aims to quantify the extent to which the CycleGAN model reduces discrepancies between simulated and real depth values. The second analysis focuses on the visual realism and structural consistency of the generated depth maps, providing insights into the model's ability to replicate real-world depth distributions.

The statistical evaluation of the model's error distributions is illustrated in figure 4 (d), where the boxplots of absolute errors corresponding to the minimum, median, and maximum MedAE cases are presented. The analysis reveals that CycleGAN effectively reduces error magnitudes, as indicated by the significant decrease in Real-Gen errors when compared to Sim-Real errors. This suggests that the model successfully refines the depth distributions, aligning the generated outputs more closely with real-world depth representations.

Despite the improvement in depth estimation, CycleGAN demonstrates varying levels of effectiveness across different cases. For the image associated with the minimum MedAE, the Real-Gen median error is 0.0392 meters, contrasting with a Sim-Real median error of 0.5490 meters. This reduction confirms that the model achieves a meaningful refinement in relatively simple scenarios. In the median MedAE case, the Real-Gen median error is 0.7843 meters, while the corresponding Sim-Real median error is 0.6275 meters. Unlike the Pix2Pix model, where a clear reduction in error is evident, CycleGAN does not consistently outperform the simulated depth maps in all cases. In the maximum MedAE case, the Real-Gen median error reaches 5.0196 meters, showing that while the model does reduce error to some extent, it struggles significantly in complex scenarios, often failing to produce substantial refinements in regions with intricate depth structures.

Páginas: 1-20 ISSN: 2953-6707

The statistical summary of absolute errors confirms that while CycleGAN can reduce discrepancies between simulated and real depth distributions, it does not consistently outperform the direct simulated-to-real comparison. Notably, the interquartile ranges for Real-Gen errors in all cases remain relatively large, indicating considerable variance in model performance. This suggests that while the generative process introduces some corrections, the degree of refinement is highly scene-dependent, particularly in regions characterized by complex geometries and occlusions.

The radar charts presented in Figure 4a, Figure 4b, and Figure 4c provide an in-depth comparative analysis of depth profiles extracted from the simulated, generated, and real images. The results indicate that for cases with lower MedAE, the generated depth maps exhibit a general alignment with real depth data, particularly in regions with gradual depth transitions. However, in cases with higher MedAE, the generated depth maps deviate significantly from the real data, revealing the limitations of CycleGAN in handling abrupt depth discontinuities.

For the minimum MedAE case, the generated depth profiles closely resemble real depth, correcting distortions and accurately reconstructing smooth depth variations, making CycleGAN effective in simple scenes. In the median MedAE case, artifacts appear at occlusion boundaries and high-gradient transitions, as the model struggles with sharp depth variations, leading to localized distortions. In the maximum MedAE case, deviations are significant, with over-smoothing in complex geometries, loss of fine details, and difficulty preserving intricate depth transitions. These findings highlight a key limitation of CycleGAN in its ability to generalize across highly complex scenes, where real-world depth distributions exhibit significant variance.

The qualitative assessment of the generated depth images provides additional insights into the strengths and weaknesses of CycleGAN.

Figure 4e, Figure 4f, and Figure 4g present the simulated, generated, and real depth images for the minimum, median, and maximum MedAE cases, respectively. A critical observation from these results is that while the model does produce depth maps that are statistically closer to real data, the visual quality of the generated images is noticeably inferior compared to Pix2Pix.

For the minimum MedAE case, the generated depth image shows structural consistency with the real depth map but appears blurred, refining edges while losing fine details. In the median MedAE case, artifacts become more prominent, especially in occluded areas and high-gradient transitions, as CycleGAN struggles with noise characterization and fails to preserve finer geometric structures. In the maximum MedAE case, discrepancies are most pronounced, with over-smoothing eliminating sharp depth transitions and failing to reconstruct high-frequency variations, highlighting CycleGAN's difficulty in replicating complex depth distributions.This qualitative analysis reveals that CycleGAN, while effective in reducing numerical errors, struggles to produce visually convincing depth reconstructions. Unlike Pix2Pix, which maintains structural coherence in most cases, CycleGAN introduces distortions that make the generated depth maps appear unrealistic. The model appears to prioritize statistical alignment over perceptual accuracy, leading to results that, while numerically valid, fail to capture the true characteristics of real-world depth distributions.

Páginas: 1-20

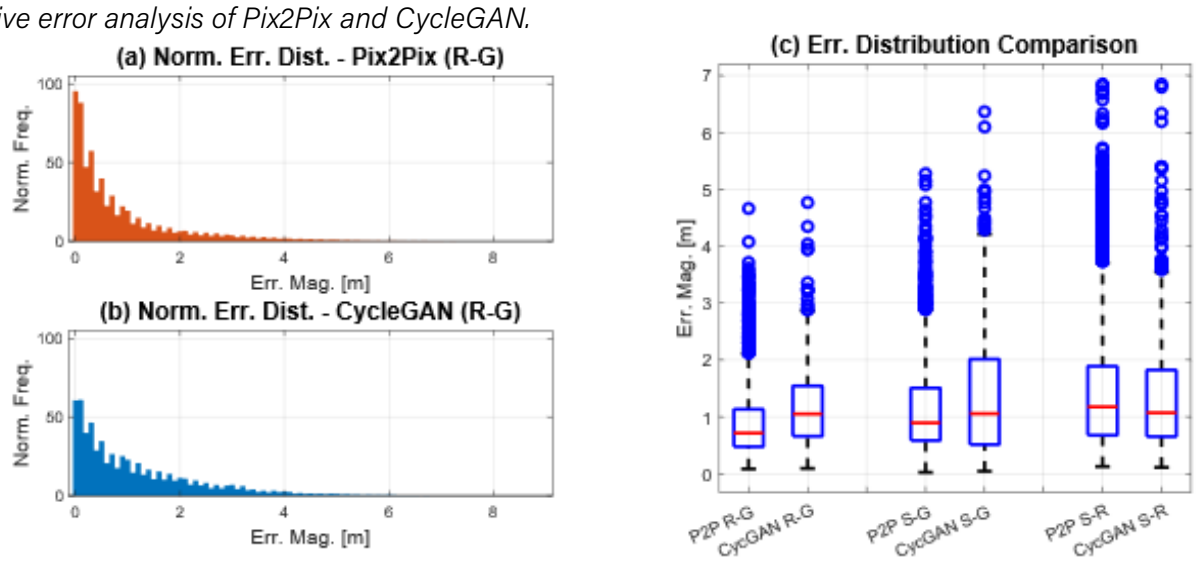
ISSN: 2953-6707

Comparative Performance of Pix2Pix and CycleGAN

The comparative analysis between Pix2Pix and CycleGAN provides insight into the relative strengths and limitations of each model. The error histograms shown in Figure 5a and Figure 5b illustrate the normalized distributions of Real-Gen errors for both models. Pix2Pix demonstrates a higher frequency of lower error values, suggesting that its generated depth images are generally more accurate when compared to real depth data. CycleGAN, on the other hand, exhibits a broader error distribution, with a noticeable shift toward higher error magnitudes, indicating that its generated depth images retain greater discrepancies from real-world measurements.

Figure 5

Comparative error analysis of Pix2Pix and CycleGAN.



Source: Prepared by the author (2025). (a) and (b) show normalized Real-Gen error distributions, with Pix2Pix exhibiting lower errors. (c) Boxplots compare error distributions, highlighting Pix2Pix's significant Sim-Real error reduction, while CycleGAN shows minimal improvement.

The statistical analysis presented in Table 1 further confirms these observations. The median Real-Gen error for Pix2Pix is 0.7143 meters, whereas CycleGAN has a higher median Real-Gen error of 1.0509 meters. Similarly, while the Sim-Gen errors are comparable for both models, the Sim-Real error for Pix2Pix is significantly larger than that of CycleGAN. This suggests that CycleGAN operates within a more constrained transformation space, achieving only marginal improvements over the original simulated depth maps. In contrast, Pix2Pix applies more substantial corrections, leading to a stronger reduction in the Sim-Real gap.

Table 1

Median errors and statistical significance for Pix2Pix and CycleGAN.

Model	Median Error (Real-Gen) [m]	Median Error (Sim-Gen) [m]	Median Error (Sim-Real) [m]	Wilcoxon p-value	Cliff's Delta
Pix2Pix	0.7143	0.8945	1.1744	4.0893e-189	0.5172
CycleGAN	1.0509	1.0562	1.0706	0.00914	-0.0116

Source: Author (2025).

ISSN: 2953-6707

The statistical significance of these differences is supported by the Wilcoxon signed-rank test, which evaluates whether the observed reduction in error is consistent across the dataset. Pix2Pix achieves an exceptionally low p-value (4.0893e-189), providing very strong evidence that the reduction in Real-Gen error is statistically significant. The effect size, measured using Cliff's Delta, is 0.5172, indicating a large practical effect and confirming that the improvement is not only statistically significant but also meaningful in real-world applications. In contrast, CycleGAN's p-value is 0.00914, which, while still statistically significant, suggests a weaker improvement. The corresponding effect size of -0.0116 is classified as negligible, further reinforcing the conclusion that CycleGAN does not meaningfully reduce the Sim-Real error.

Table 2 highlights the practical implications of these differences. Pix2Pix achieves an average Real-Gen error reduction of 0.46 meters, demonstrating a strong improvement in depth accuracy. CycleGAN, by contrast, achieves a significantly lower improvement of only 0.20 meters. The practical impact of these findings is clear: Pix2Pix is far more effective at refining simulated depth maps and reducing the Sim-Real gap, while CycleGAN's improvements are relatively minor and statistically weak.

Table 2.

Comparative analysis of Pix2Pix and CycleGAN in reducing Sim-to-Real error, including mean error reduction, statistical significance, and practical impact.

Comparison	Pix2Pix	CycleGAN
Mean Real-Gen Error Reduction	0.46m	0.20m
Wilcoxon p-value	4.0893e-189 (very strong evidence)	0.0091 (weak evidence)
Effect Size (Cliff's Delta)	0.5172 (large effect)	-0.0116 (negligible effect)
Practical Improvement	Strong reduction in Sim-to-Real gap	Very weak or negligible improvement

Source: Author (2025).

These findings are consistent with prior research in domain adaptation and sim-to-real transfer for robotic perception. For instance, Sadeghi and Levine (2017) demonstrated the effectiveness of domain randomisation for UAV control, although their approach lacked fine structural preservation compared to adversarial learning models. James et al. (2019) employed GAN-based adaptation for robotic grasping and reported statistically significant performance gains when using paired supervision. More recently, Westerski and Fong (2023) surveyed state-of-the-art synthetic data generation techniques and emphasised that methods which preserve semantic and spatial structure, such as those based on photorealistic simulation or structured domain adaptation, tend to outperform approaches relying solely on randomisation. The results of the present study reinforce this trend, showing that Pix2Pix, which leverages paired data, offers superior performance in refining depth images. This supports the idea that supervised adaptation strategies provide measurable advantages when high-fidelity domain alignment is required for UAV navigation.

ISSN: 2953-6707

Discussion

This study evaluated the performance of Pix2Pix and CycleGAN for domain adaptation in depth image refinement, focusing on their ability to bridge the simulation-to-reality gap. The results demonstrate that both models reduce depth estimation errors, but Pix2Pix consistently outperforms CycleGAN in terms of structural accuracy, numerical error reduction, and overall realism. The comparison between these two approaches highlights the trade-offs between paired and unpaired domain adaptation methods, offering insights into their applicability for real-world UAV perception and navigation.

A crucial aspect of this study was the meticulous replication of the real-world environment in simulation, achieved through precise 3D modelling and real-time pose synchronisation. Objects such as the wooden cube, cylindrical obstacle, and ramp structure were carefully reconstructed in Blender and integrated into the Gazebo simulation platform, ensuring that the geometric features encountered in simulation closely matched those in the real world. This alignment was further reinforced by the OptiTrack motion capture system, which enabled real-time tracking and pose synchronisation of physical objects with their virtual counterparts. The ability to track object positions with millimetre accuracy and replicate their exact placement in simulation was essential for acquiring paired depth images, forming the foundation for supervised domain adaptation with Pix2Pix. Without this level of environmental fidelity, the domain adaptation process would lack consistency, reducing the effectiveness of the trained models.

The error analysis confirms that Pix2Pix achieves a stronger reduction in Sim-Real error compared to CycleGAN. The boxplot analysis demonstrates that Real-Gen errors are substantially lower than Sim-Real errors across all three representative cases—minimum, median, and maximum MedAE—confirming that Pix2Pix successfully transforms simulated depth images into realistic depth distributions. The model exhibits its strongest performance in structured environments, where depth variations are gradual, and occlusion boundaries are well-defined. However, in highly complex scenes, Pix2Pix struggles with fine-scale geometric details, occasionally over-regularising depth transitions and introducing smoothing artefacts.

CycleGAN, in contrast, presents a different set of strengths and weaknesses. While the model achieves a measurable reduction in numerical error, its ability to produce visually coherent depth images is significantly weaker than that of Pix2Pix. The unpaired nature of CycleGAN training, which relies on cycle consistency loss rather than direct supervision, leads to inconsistent structural corrections. As a result, the generated depth maps often contain artefacts and distortions, failing to achieve the fine-grained depth refinement required for high-precision applications. Notably, CycleGAN's effect size is negligible, indicating that its improvements over the simulated depth images are minor in practical terms.

A key finding of this study is that Pix2Pix benefits significantly from paired supervision, allowing it to learn direct mappings between synthetic and real depth images. The effectiveness of this approach is reflected in the Wilcoxon signed-rank test results, which show overwhelmingly strong statistical significance (p-value = 4.0893e-189) and a large effect size (Cliff's Delta = 0.5172). This confirms that the model's refinement process is not only statistically significant but also practically meaningful, making Pix2Pix a highly suitable approach for depth adaptation in UAV perception tasks.

Páginas: 1-20

ISSN: 2953-6707

CycleGAN, despite demonstrating some level of numerical improvement, exhibits limited effectiveness in depth refinement. The statistical analysis supports this conclusion, as its Wilcoxon p-value (0.00914) suggests only weak statistical significance, and its effect size (-0.0116) is classified as negligible. These findings suggest that CycleGAN struggles to meaningfully close the Sim-Real gap, reinforcing the notion that unpaired domain adaptation alone may not be sufficient for high-precision depth transformations.

Beyond numerical error reduction, the qualitative assessment of depth images further underscores the advantages of Pix2Pix. The radar charts reveal that Pix2Pix-generated depth profiles closely follow real data, especially in scenes with low to moderate complexity. In contrast, CycleGAN fails to reconstruct fine-grained structures, often producing depth maps that exhibit unnatural distortions and spatial inconsistencies. These observations highlight the importance of paired supervision in generative depth refinement, as Pix2Pix consistently generates more accurate and visually coherent depth maps than CycleGAN.

Despite Pix2Pix's strong performance, challenges remain. The model occasionally struggles with high-frequency depth variations, particularly in occlusion regions where depth discontinuities are abrupt. One promising direction is to explore hybrid approaches that integrate the structural consistency of CycleGAN with the supervised learning advantages of Pix2Pix, potentially yielding a more balanced and robust depth refinement framework.

Another avenue would be to investigate the integration of alternative refinement strategies with Pix2Pix to address its possible limitations in handling high-frequency depth variations and occlusion boundaries. Techniques such as spatially-adaptive normalisation (Park et al., 2019) and multi-scale feature alignment (Xu et al., 2021) have shown promise in preserving fine structural details during image-to-image translation. Godard et al., (2019) on monocular depth estimation demonstrated that incorporating edge-aware smoothness and temporal consistency constraints can improve depth accuracy in dynamic scenes. Combining such architectural enhancements or post-processing methods with Pix2Pix may further improve its ability to generalise across complex geometries. Additionally, attention-based mechanisms (Wang et al., 2024) offer a way to prioritise high-gradient regions and could help mitigate over-smoothing in fine-grained depth structures.

Páginas: 1-20

ISSN: 2953-6707

Conclusion

This study investigated the application of Pix2Pix and CycleGAN for domain adaptation in depth image refinement, assessing their ability to transform simulated depth images into realistic counterparts. The findings confirm that Pix2Pix significantly outperforms CycleGAN, achieving greater reductions in depth estimation error and generating more visually coherent depth maps. The paired supervision approach used in Pix2Pix proves highly effective, enabling the model to learn direct mappings that optimise depth accuracy.

A major strength of this study lies in the high-fidelity replication of the real-world environment in simulation. The use of precisely modelled objects, OptiTrack motion capture for real-time pose synchronisation, and Gazebo-based virtual scene construction ensured that the training and evaluation process remained as consistent as possible across real and simulated domains. This structured experimental framework facilitated the acquisition of paired depth images, which were crucial for Pix2Pix's successful adaptation. The study highlights that maintaining a high level of environmental consistency is fundamental to achieving reliable depth refinement through domain adaptation.

The statistical analysis confirms the superiority of Pix2Pix, with a highly significant p-value and a large effect size, indicating that the reduction in Sim-Real error is both statistically and practically meaningful. CycleGAN, despite offering some numerical improvements, exhibits weak structural accuracy and negligible practical impact. The unpaired nature of its training process limits its ability to generate depth maps that convincingly resemble real-world data.

These findings underscore the importance of selecting the appropriate domain adaptation strategy based on specific application requirements. In tasks where high-precision depth estimation is crucial, Pix2Pix emerges as the preferred model, providing substantial reductions in error and strong structural consistency. However, CycleGAN's ability to learn without paired data may still be useful in scenarios where labelled datasets are unavailable, though its effectiveness remains limited.

Beyond its immediate contributions to UAV perception and navigation, this research has broader implications for robotic vision, computer graphics, and sensor simulation. The ability to generate realistic depth images from synthetic environments can benefit applications in autonomous systems, mixed reality, and robotic training simulations. Additionally, this tool holds potential for educational applications, particularly in settings where students and researchers lack access to real-world depth cameras. By providing a cost-effective, simulation-based approach to depth sensing, this framework can serve as a valuable resource for teaching and research in robotics, computer vision, and AI-driven perception.

Future research should explore hybrid approaches that combine the structural consistency of CycleGAN with the depth accuracy of Pix2Pix, potentially leading to a more flexible and robust depth refinement framework. Additionally, investigating alternative architectures or post-processing techniques could further enhance depth consistency and mitigate over-smoothing in complex scenes. Expanding this study to other depth sensing technologies and simulation environments may further improve its generalisation and practical impact.

Páginas: 1-20

ISSN: 2953-6707

Ultimately, this study confirms that adversarial domain adaptation can significantly bridge the simulation-to-reality gap in depth perception, enabling autonomous systems to operate more effectively in real-world environments while also offering new opportunities for education, research, and real-time robotic applications.

Acknowledgements

The author acknowledges the support of the Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) of Ecuador.

Páginas: 1-20

ISSN: 2953-6707

References

Bousmalis K, Silberman N, Dohan D, Erhan D, Krishnan D. Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks. 2017 IEEE Conf. Comput. Vis. Pattern Recognit., vol. 2017- Janua, IEEE; 2017, p. 95–104.

Godard C, Aodha O Mac, Firman M, Brostow G. Digging Into Self-Supervised Monocular Depth Estimation. 2019 IEEE/CVF Int. Conf. Comput. Vis., vol. 2019- Octob, IEEE; 2019, p. 3827–37.

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Adv. Neural Inf. Process. Syst., vol. 3, Wiesbaden: Springer Fachmedien Wiesbaden; 2014, p. 2672–80.

Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-Image Translation with Conditional Adversarial Networks. CVPR 2017.

James S, Johns E. 3D Simulation for Robot Arm Control with Deep Q-Learning 2016.

James S, Wohlhart P, Kalakrishnan M, Kalashnikov D, Irpan A, Ibarz J, et al. Sim-To-Real via Sim-To-Sim: Data-Efficient Robotic Grasping via Randomized-To-Canonical Adaptation Networks 2019:12627–37.

Jing Chen, Tianbo Liu, Shaojie Shen. Online generation of collision-free trajectories for quadrotor flight in unknown cluttered environments. 2016 IEEE Int. Conf. Robot. Autom., vol. 2016- June, IEEE; 2016, p. 1476–83.

Le H, Saeedvand S, Hsu CC. A Comprehensive Review of Mobile Robot Navigation Using Deep Reinforcement Learning Algorithms in Crowded Environments. J Intell Robot Syst Theory Appl 2024;110:1–22.

Liu MY, Tuzel O. Coupled generative adversarial networks. Adv. Neural Inf. Process. Syst., vol. 29, 2016, p. 469–77.

Park T, Liu MY, Wang TC, Zhu JY. Semantic image synthesis with spatially-adaptive normalization. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019- June, IEEE; 2019, p. 2332–41.

Quigley M, Gerkey B, Conley K, Faust J, Foote T, Leibs J, et al. ROS: an open-source Robot Operating System 2019.

Al Radi M, AlMallahi MN, Al-Sumaiti AS, Semeraro C, Abdelkareem MA, Olabi AG. Progress in artificial intelligence-based visual servoing of autonomous unmanned aerial vehicles (UAVs). Int J Thermofluids 2024;21:100590.

Sadeghi F, Levine S. CAD2RL: Real Single-Image Flight Without a Single Real Image. Robot. Sci. Syst. XIII, vol. 13, Robotics: Science and Systems Foundation; 2017.

Sampedro C, Bavle H, Rodriguez-Ramos A, de la Puente P, Campoy P. Laser-Based Reactive Navigation for Multirotor Aerial Robots using Deep Reinforcement Learning. 2018 IEEE/RSJ Int. Conf. Intell. Robot. Syst., IEEE; 2018, p. 1024–31.

Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R. Learning From Simulated and Unsupervised Images Through Adversarial Training 2017:2107–16.

Sikang Liu, Watterson M, Tang S, Kumar V. High speed navigation for quadrotors with limited onboard sensing. 2016 IEEE Int. Conf. Robot. Autom., vol. 2016- June, IEEE; 2016, p. 1484–91.

Tzeng E, Devin C, Hoffman J, Finn C, Abbeel P, Levine S, et al. Adapting Deep Visuomotor Representations with Weak Pairwise Constraints. Springer Proc. Adv. Robot., vol. 13, Springer, Cham; 2020, p. 688–703.

Páginas: 1-20

ISSN: 2953-6707

Wang F, Zhang Q, Zhao Q, Wang M, Sun F. Unsupervised image-to-image translation with multiscale attention generative adversarial network. Appl Intell 2024;54:6558–78.

Westerski A, Teck FW. Synthetic Data for Object Detection with Neural Networks: State of the Art Survey of Domain Randomisation Techniques. ACM Trans Multimed Comput Commun Appl 2023;21.

Wu K, Esfahani MA, Yuan S, Wang H. Depth-based Obstacle Avoidance through Deep Reinforcement Learning. Proc. 5th Int. Conf. Mechatronics Robot. Eng., vol. Part F1476, New York, NY, USA: ACM; 2019, p. 102–6.

Xu C, Zhou M, Ge T, Jiang Y, Xu W. Unsupervised Domain Adaption With Pixel-Level Discriminator for Image-Aware Layout Generation 2023:10114–23.

Xu X, Chen Z, Yin F. Multi-Scale Spatial Attention-Guided Monocular Depth Estimation With Semantic Enhancement. IEEE Trans Image Process 2021;30:8811–22.

Xu Y, Cao H, Xie L, Li X-L, Chen Z, Yang J. Video Unsupervised Domain Adaptation with Deep Learning: A Comprehensive Survey. ACM Comput Surv 2024;56:36.

Zhu J-Y, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2017 IEEE Int. Conf. Comput. Vis., vol. 2017-Octob, IEEE; 2017, p. 2242–51.

Páginas: 1-20

ISSN: 2953-6707