Robust Monocular Distance Estimation in Degraded Visual Conditions via a Wide-and-Deep Fusion Framework

Richard Wang, Grant Han, Alexander Liu, Zixuan Ye, Alexander Deng, Lexie Liu, Xing Wei*

Abstract—**Robust visual perception remains a central challenge for autonomous systems operating under degraded environmental conditions such as fog, low light, and artificial glare. In this study, we investigate multiple strategies to enhance the resilience of vision-based models in such scenarios, using monocular distance estimation as a representative task due to its sensitivity to visual quality and availability of ground-truth metrics. A Convolutional Neural Network (CNN) is first trained on clean visual data using a single-camera robotic platform, and its direct inference results serve as the baseline.**

We then explore three enhancement strategies: (1) image preprocessing with a transformer-based large model, (2) model fine-tuning on low-visibility data, and (3) a novel fusion framework that integrates temporal motion history with image features. Inspired by wide-and-deep learning structures, the proposed fusion model combines shallow inputs (e.g., recent speed and displacement) with deep CNN-based image embeddings for improved robustness.

Experimental results across various lighting environments demonstrate that while multiple enhancement strategies improve estimation accuracy, the fusion model achieves the most consistent and efficient performance. The findings provide insight into designing more reliable vision systems for low-cost robotics and general perception tasks under real-world variability.

I. Introduction

Visual perception plays a critical role in autonomous systems, enabling tasks such as navigation, mapping, and environment understanding [1, 2]. Among available sensing approaches, monocular vision remains especially attractive for low-cost and resource-constrained platforms due to its simplicity and minimal hardware requirements [3]. However, vision-based inference often suffers in degraded environments, such as under fog, low lighting, or artificial glare. These conditions are common in real-world deployment but frequently cause deep learning models to underperform, as such models are typically trained under well-lit, clean conditions [4, 5].

While deep neural networks (DNNs) have proven effective in solving a wide range of visual tasks, their performance heavily depends on the quality of the input [6]. When deployed in environments that differ from training conditions, these models often fail to generalize, leading to substantial drops in accuracy [7].

In this work, we use monocular distance estimation as a representative task for evaluating visual robustness strategies. Distance estimation offers a straightforward and measurable performance metric, allowing us to quantify inference accuracy under controlled conditions. Our experiments are conducted on the VEX V5RC platform [8], a low-cost mobile robotics system equipped with a monocular camera. We collect training and evaluation data under both clear and degraded environment. Importantly, we use the platform's built-in odometry to provide ground-truth distance labels for supervised learning. Our experiments cover a range of realistic low-visibility conditions, including dim evening light, nighttime scenes, and artificial lights.

We begin by training a baseline Convolutional Neural Network (CNN) regression model on images captured in clear conditions. We then evaluate three enhancement strategies aimed at improving robustness in degraded visual environments:

1. Image preprocessing using a transformer-based large model,
2. Model fine-tuning with additional data collected in low-visibility settings,
3. A novel fusion model that integrates short-term motion history with visual input.

Our proposed fusion model is inspired by wide-and-deep architectures used in recommendation systems [9]. It incorporates recent speed and displacement measurements into a shallow pathway, which is combined with visual features extracted from a deep CNN encoder. This hybrid structure enables the model to leverage both historical motion context and visual cues, improving resilience when image quality deteriorates.

Through quantitative experiments across varied lighting and weather conditions, we compare these strategies and analyze their effectiveness. Our results show that while multiple approaches provide improvements, the proposed fusion model achieves the highest robustness and efficiency, especially in visually ambiguous environments. These insights can help guide the development of more reliable vision systems in real-world, resource-constrained environments.

The remainder of this paper is organized as follows: Section II reviews related work on monocular distance estimation, visual robustness techniques, and wide-and-deep learning

*All co-authors contributed equally to this paper. Richard Wang (richardbojinwang2008@gmail.com), Grant Han (26ghan@gmail.com), Alexander Liu (alexanderliu089@gmail.com), Zixuan Ye (zixuanye.sh@gmail.com), Alexander Deng (alexanderdeng11@gmail.com), and Lexie Liu (lexieliu689@gmail.com) are students at BASIS Independent Silicon Valley Upper School, San Jose, California, United States and members of Team 1698V in VEX Robotics Competition. Xing Wei (xingweihome @gmail.com) is the AI coach of Team 1698V.

Code repository for this work: https://github.com/richardbwang/Research-Paper-World-Robot-Conference-2025

architectures. Section III details the three enhancement strategies evaluated: transformer-based image preprocessing, model fine-tuning under low-visibility conditions, and the proposed fusion framework. Section IV describes the experimental setup. Section V presents the experimental results and comparative analysis across diverse environmental settings. Finally, Section VI concludes the paper and outlines future directions for deploying robust vision systems in resource-constrained autonomous platforms.

II. Related Work

A. Monocular Distance Estimation under Degraded Conditions

Monocular depth and distance estimation have been widely studied due to their cost-efficiency and hardware simplicity [3, 10]. However, their performance degrades significantly under challenging visual conditions such as fog, low lighting, and artificial glare. To address this, several efforts have focused on training with augmented data, domain adaptation, or self-supervised learning frameworks that reduce dependency on labeled data [5, 6, 7]. Transformer-based restoration models like TransWeather have also shown promise in enhancing image quality before inference [11]. Despite progress, systematic evaluations across strategies remain limited.

B. Visual-Temporal Fusion Models

In robotics, temporal information has been effectively used to improve visual perception. Approaches such as Visual-Inertial Odometry (VIO) combine camera and motion data for improved localization and estimation [12]. Recent works introduce learned fusion architectures that selectively combine image features with temporal priors [13], improving robustness in dynamic or noisy environments. These techniques often rely on multi-modal fusion and temporal attention mechanisms.

C. Wide and Deep Learning Architectures

Wide and deep models, first introduced for recommendation systems [9], combine a wide (memorization) path with a deep (generalization) path. This architecture has since been adapted in various domains to enhance robustness and hybrid reasoning. In visual perception tasks, incorporating motion history as the "wide" component alongside deep visual embeddings enables the model to leverage contextual priors when image quality deteriorates. Our proposed fusion model draws inspiration from this design, integrating shallow features such as recent speed and displacement into a deep CNN framework to improve distance estimation under degraded environments.

Notably, the shallow pathway of the wide side enables efficient integration of contextual priors through online, incremental training. This flexibility presents a significant advantage of the wide-and-deep framework, particularly for adaptive systems operating in dynamic or unpredictable environments.

III. Methodology

This section presents the three strategies evaluated in this work to improve monocular distance estimation under visually degraded environments. The central contribution is a novel Fusion Model based on a Wide-and-Deep Architecture, designed to enhance robustness by integrating temporal context and visual features. We first describe the baseline model and enhancement techniques, then focus in detail on the fusion approach.

We use monocular distance estimation as a representative task to evaluate visual perception robustness under degraded environmental conditions. Distance estimation is particularly suitable for this purpose because it provides a continuous, quantitative metric that can be precisely measured using onboard odometry. While our experiments focus on this task, the proposed methods—especially the fusion model—are broadly applicable to other computer vision problems such as object detection, semantic segmentation, and visual navigation, where robustness to visual degradation is equally critical.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Baseline CNN Model

The baseline model used in this study is a lightweight Convolutional Neural Network (CNN) with a minimal number of layers. It is intentionally not optimized for state-of-the-art performance in monocular distance estimation. Instead, the design prioritizes simplicity and interpretability, serving as a consistent foundation for evaluating the effectiveness of various enhancement strategies under degraded visual conditions.

The model takes as input raw RGB images captured in clear environments and is trained in a supervised manner using distance labels derived from onboard odometry. Its primary role is to establish a reference level of performance in ideal settings, against which we can assess the impact of preprocessing, fine-tuning, and fusion-based approaches when image quality deteriorates.

B. Existing Enhancement Strategies

Strategy I: Transformer-Based Image Preprocessing

To improve image quality under low-visibility conditions, we incorporate a preprocessing step based on generative enhancement using large-scale vision-language models. Specifically, we use a commercial large transformer pretrained model - OpenAI's "create an image" tool via ChatGPT [14], which leverages a transformer-based image generation backbone to enhance visual clarity through semantic reconstruction.

The prompt provided to the model was:

"This image was taken in low light. Could you enhance it to look as if it were taken in bright day light?"

This instruction guides the model to synthesize a high-fidelity, well-lit version of the original image using learned contextual priors. Unlike traditional enhancement techniques, this approach is not constrained to pixel-wise correction but

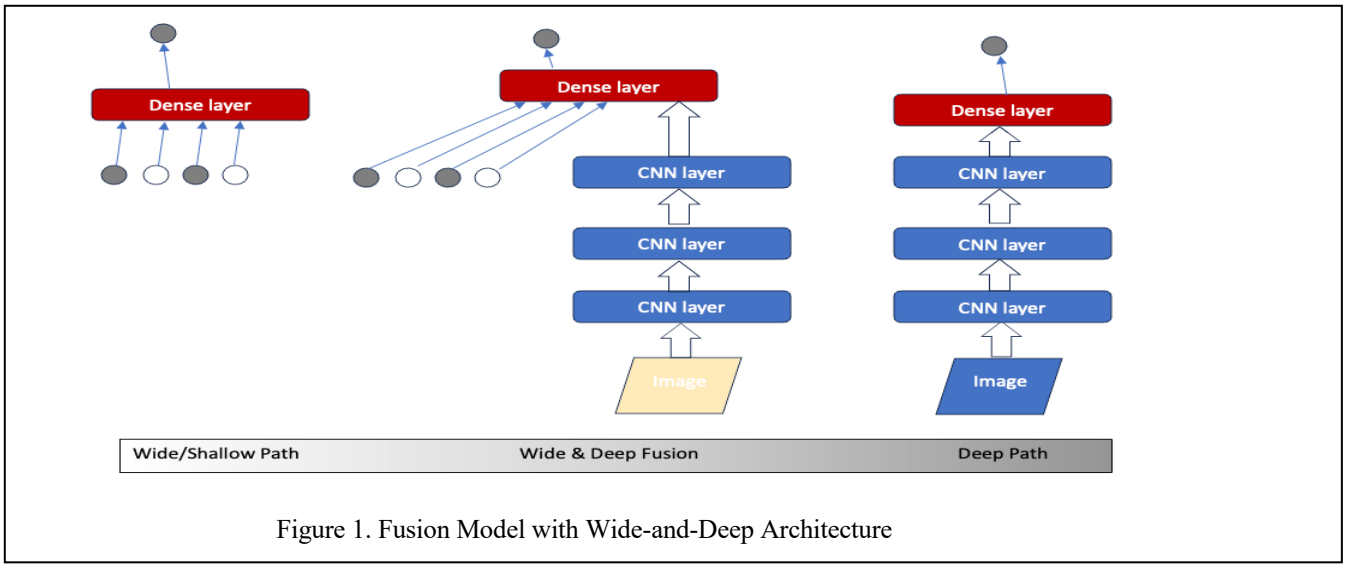

Figure 1. Fusion Model with Wide-and-Deep Architecture

instead generates globally consistent imagery based on scene understanding.

The enhanced images are then passed through the baseline CNN without modifying its structure. This setup isolates the effect of generative enhancement, allowing us to evaluate the contribution of transformer-based preprocessing to distance estimation performance under visually degraded environments.

Strategy II: Fine Tuning

The second strategy improves model robustness via domain adaptation. We collect additional data under adverse conditions and fine-tune the baseline CNN on this extended dataset. Fine-tuning is initialized with pre-trained weights from the clean model and carried out with a lower learning rate to retain prior knowledge while adapting to new domains.

C. *Fusion Model with Wide-and-Deep Architecture*

Our main contribution is a fusion model inspired by the Wide & Deep framework originally proposed for recommender systems [9]. In our context, the "deep" pathway consists of a convolutional encoder that extracts high-level embeddings from monocular images. This pathway is responsible for generalization from visual appearance, and functions similarly to the baseline CNN.

The "wide" pathway models explicit, low-dimensional temporal features from robot odometry, such as recent velocity and displacement. These features are passed through a shallow multi-layer perceptron (MLP), forming a direct and interpretable path that complements the deep visual encoder. This captures patterns based on short-term motion dynamics that correlate with target distance, even when visual information is noisy or absent.

The outputs from both branches—CNN-based visual embeddings and shallow motion features—are concatenated and passed to a final regression layer that predicts distance. This structure allows the model to simultaneously generalize from high-dimensional visual patterns and memorize recent motion trends for context-aware predictions.

The wide-and-deep design offers several advantages. First, it enables learning of complementary representations: the deep path adapts to varied visual inputs, while the wide path provides priors grounded in recent motion. Second, the shallow wide branch can be incrementally updated online, making it efficient for adaptation without full retraining. This is particularly valuable for lightweight, real-time systems deployed in dynamic environments.

The architecture is illustrated in Figure 1. The model's prediction is:

$$\hat{y} = f_{\text{fusion}}([f_{\text{deep}}(x_v), x_t])$$

where

x_v: input image

x_t: temporal features

f_{deep}: deep encoder for visual input

f_{fusion} : fusion function that combines the concatenated features and makes prediction

IV. EXPERIMENTAL SETUP

To evaluate the effectiveness and robustness of the strategies described in section III, particularly our proposed wide-and-deep fusion framework for monocular distance estimation, we implemented and tested the system on a mobile robotic platform based on the VEX V5RC system [8]. This platform is equipped with a forward-facing monocular RGB camera for visual input and a built-in odometry system to provide accurate ground-truth distance labels.

A. *Data Collection*

Experiments were conducted in a controlled indoor environment to ensure consistency and repeatability. During each trial, the robot moved steadily while the forward-facing monocular camera captured frames at a rate of 20 frames per second. Each frame was automatically paired with a ground-truth distance label provided by the robot's onboard odometry system. This setup enabled the collection of temporally dense, labeled data under continuous motion, supporting effective supervised learning and evaluation of the proposed models.

B. Degraded Environments

To assess model robustness under realistic deployment conditions, we captured additional datasets in four degraded environments representing common challenges for monocular vision systems:

- Low Artificial Light (LAL): Images recorded at night with a low-intensity LED light source. This setting induces directional shadows and partial underexposure.
- High Artificial Light (HAL): Similar to LAL, but with a higher-intensity LED light source. This creates overhead glare and uneven brightness across the image.
- Low-Light (LL): Captured at night (approximately 10 PM) with no artificial lighting. The environment was nearly dark, resulting in extreme low visibility and sensor noise.
- Clear Light but Different (CLD): To test sensitivity to temporal lighting shifts, we introduced a subtle domain shift by capturing "clear" images in the morning, in contrast to the baseline dataset captured in the afternoon. This simulates real-world variance due to diurnal lighting changes.

These environments were selected for their simplicity, repeatability, and relevance to real-world robotic operation, particularly in indoor environments with inconsistent or dynamic lighting.

For each environment, a separate dataset was collected using the same camera settings and spatial configuration. Each frame was paired with an odometry-based distance label and included in both the training and evaluation. In total, the dataset encompasses a wide range of visual inputs under varying visibility conditions, supporting a rigorous evaluation of our framework's bias-awareness and generalization capabilities.

V. RESULTS AND ANALYSIS

We evaluate distance estimation performance using Mean Absolute Error (MAE) (measured in inches) and the percentage difference in error between different training conditions. The goal of this study is not to fine-tune a highly accurate distance estimator, but to investigate how well various strategies handle visual degradation.

A. Baseline: Deep-Only CNN Model

The results in Table I reflect the performance of a baseline deep CNN model (without incorporating the shallow/wide path), effectively serving as an ablation study of our wide-and-deep fusion framework. Two evaluation settings are presented:

1. Train and test on the same dataset, under matched lighting conditions.
2. Train on the baseline Clear Light (CL) dataset, then test on other degraded environments.

We observe the following:

- When trained and evaluated on the same condition—even under low or artificial lighting—the CNN performs consistently well, achieving low MAE values across all environments.
- In contrast, training only on CL and testing on degraded lighting conditions leads to significant performance degradation. The most severe case is under Low Light (LL), where the error increases from 1.06 to 58.13.

This gap highlights the deep model's limited generalization ability under visual shifts, motivating the development of more robust solutions such as our proposed fusion framework.

The unit in our experiment is inch. The purpose of this study is not to optimize distance estimation model but to compare the effectiveness of different strategies to handle visual degradation.

TABLE I. DEEP MODEL RESULTS

Environment	Trained & Tested on Same Dataset	Trained on Baseline CL Dataset
CL(Baseline)	1.81	X
LAL	1.79	29.84
HAL	1.47	29.02
LL	1.06	58.13
CLD	1.22	29.86

B. Existing Enhancement Strategies

We evaluated two existing approaches for improving monocular depth estimation in degraded environments:

Strategy I: Transformer-Based Image Preprocessing

Building on recent advances in vision-language models, we adopted a large pre-trained image&text-to-image transformer model, ChatGPT "create an image" tool [14], as described in Section III. The model was prompted to enhance low-light images to appear as if taken under bright daylight conditions.

However, when these enhanced images were directly fed into the baseline CNN model, the resulting Mean Absolute Error (MAE) increased dramatically to 236.50—substantially worse than all other configurations, including applying the baseline model directly to raw low-light data. This outcome suggests that a generic, pre-trained vision transformer, when used without task-specific fine-tuning, may introduce artifacts or distort important visual cues essential for accurate distance estimation.

Although this strategy alone proved ineffective, we observe that fine-tuning the CNN on the preprocessed images, as explored in the next strategy, can partially mitigate the degradation in performance. This implies that image enhancement must be coupled with model adaptation to yield meaningful improvements under visually degraded environments.

TABLE II. FINE TUNING RESULTS

Environment	Baseline MAE	1 Epoch MAE	Δ% (1 Epoch)	5 epoch MAE	Δ% (5 Epoch)	20 epoch MAE	Δ% (20 Epoch)
LAL	29.84	38.37	+28.6%	20.82	-30.2%	6.86	-77.0%
HAL	29.02	25.00	-13.9%	21.14	-27.1	11.51	-60.3%
LL	58.13	45.56	-21.6%	23.05	-60.4	5.2	-91.1%
CLD	29.86	25.42	-14.9%	12.57	-57.9%	4.56	-84.7%
LM	236.5	82.36	-65.2%	23.62	-90.0	21.39	-91.0%

Strategy II: Fine Tuning

In this strategy, we investigated whether fine-tuning the baseline deep CNN model on degraded data could improve performance. Fine tuning helps adapt the model to the new domain, especially when the input distribution has shifted due to environmental degradation or preprocessing (as in Strategy I). This is particularly suitable for real-world robotic deployments where rapid online or incremental updates may be required.

To assess the trade-off between training cost and performance gain, we fine-tuned the model for 1, 5, and 20 epochs. Table II presents the resulting Mean Absolute Error (MAE) and the percentage change compared to the baseline (negative values indicate improvement).

The results demonstrate that fine-tuning significantly improves performance across all environments, with even 5 epochs providing substantial gains. The 20-epoch setting yields the most dramatic reductions, including over 90% error reduction in the low-light (LL) and transformer-preprocessed (LM) cases.

Nevertheless, in practical robotic applications, fine-tuning must remain lightweight to meet real-time constraints. While 20 epochs offer optimal performance in controlled settings, shorter fine-tuning durations (e.g., 5 epochs) offer a strong trade-off between adaptability and deployment efficiency.

C. Strategy III: Wide-and-Deep Fusion

The third and central strategy of this work introduces our proposed Wide-and-Deep Fusion Model, which combines visual features extracted from a deep CNN with structured temporal motion data via a shallow/wide pathway. This architecture is designed to enhance robustness under degraded visual conditions by integrating both high-dimensional image representations and low-dimensional, complementary temporal cues derived from robot odometry.

The wide path incorporates three past distance measurements taken at 50 ms intervals (i.e., over the past 150 ms), reflecting short-term motion dynamics. These temporal features, captured at 20 frames per second, offer predictive information especially useful when visual input is corrupted or ambiguous. They are processed through a shallow Multi-Layer Perceptron (MLP), and fused with the deep visual features. The resulting model is capable of leveraging recent motion context to disambiguate poor-quality visual input.

We evaluate the wide-and-deep fusion model under three configurations:

- *Config 1*: Training on the baseline (clear-light) dataset only, and testing across all environments;
- *Config 2*: Training on baseline + LAL (low artificial light) datasets testing across all environments.
- *Config 3*: Training on baseline, followed by fine-tuning on each target environment.

The results of these configurations are presented in Table III, Table IV and Table V, respectively.

Config 1: Fusion Model Trained on Baseline

TABLE III. PERFORMANCE OF WIDE-AND-DEEP FUSION MODEL (TRAINED ON BASELINE ONLY)

Environment	Baseline MAE	Fusion Model MAE	Δ%
CL(Baseline)	1.81	0.93	-48.6%
LAL	29.84	18.86	-36.8%
HAL	29.02	18.55	-36.1%
LL	58.13	52.48	-9.7%
CLD	29.86	13.53	-54.7%

Even when trained exclusively on clean lighting conditions (CL), the fusion model generalizes significantly better than the deep-only baseline in most degraded environment. Particularly strong gains are observed for CLD (−54.7%), LAL (−36.8%), and HAL (−36.1%), confirming the benefit of integrating past motion cues. While improvement on LL is modest (−9.7%), it nonetheless indicates robustness to the most visually challenging condition.

Config 2: Fusion Model Trained on Baseline + LAL

As shown in Table IV, adding LAL data during training significantly boosts generalization across all conditions—even for those not seen during training (HAL, LL, CLD). This supports the hypothesis that *moderate dataset diversification* with representative degraded samples helps the model learn invariances useful across broader scenarios. Notably, LL error drops by nearly 60%, suggesting the added exposure during training better conditions the model for visual degradation.

TABLE IV. PERFORMANCE OF WIDE-AND-DEEP FUSION MODEL (TRAINED ON BASELINE + LAL)

Environment	Baseline MAE	Fusion Model MAE	Δ%
CL(Baseline)	1.81	1.32	-27.1%
LAL	29.84	12.16	-59.3%
HAL	29.02	13.99	-51.8%
LL	58.13	23.40	-59.7%
CLD	29.86	14.14	-52.6%

Config 3: Fusion Model Trained on Baseline and Fine-Tuned per Environment

Table V demonstrates that fine-tuning the wide-and-deep model for just 5 epochs on each target dataset leads to substantial performance gains—outperforming the deep-only model in both magnitude and efficiency (see Table II). Especially under severe degradation (LL), the error drops dramatically by 74.6% and 79.4%, respectively. The lightweight nature of the wide path makes this approach well-suited for real-world robotics, where fast online adaptation is needed.

TABLE V. PERFORMANCE OF WIDE-AND-DEEP FUSION MODEL (TRAINED ON BASELINE + FINE TUNING)

Environment	Baseline MAE	Fusion Model + Fine-Tuning (5 epochs)	Δ%
CL(Baseline)	1.81	0.93	-48.6%
LAL	29.84	14.99	-49.8%
HAL	29.02	13.80	-52.4%
LL	58.13	14.79	-74.6%
CLD	29.86	6.15	-79.4%

In summary, these three configurations demonstrate that the wide-and-deep fusion strategy provides significant benefits under visibility degradation. Even without additional training, the model generalizes well due to its use of motion priors. Adding a modest amount of degraded training data further enhances robustness, and targeted fine-tuning enables substantial gains with minimal training effort. This makes the proposed fusion approach highly suitable for deployment in dynamic or resource-constrained robotic systems.

E. Results Summary

The experimental results demonstrate several key insights regarding model robustness under degraded visual conditions.

First, the use of a pre-trained large transformer model for image enhancement alone is insufficient and can be detrimental when used without task-specific fine-tuning, as evidenced by a significant increase in estimation error.

Second, fine-tuning the CNN on degraded inputs yields marked performance improvements across all conditions, particularly when extended to 20 epochs, though such approaches may be costly in time-critical deployments.

Third, the proposed wide-and-deep fusion model consistently outperforms the baseline and other strategies at low-cost configurations. When trained solely on clean data, the fusion model exhibits better generalization due to the integration of temporal motion cues. Performance is further enhanced with the addition of limited degraded training data, and maximized when paired with lightweight fine-tuning. Overall, the fusion approach offers the best trade-off between performance, adaptability, and computational cost, making it the most practical and effective solution for real-world robotic applications operating under visual degradation.

VI. CONCLUSION

This study investigated monocular distance estimation as a benchmark for evaluating the resilience of vision-based models under degraded visual conditions, comparing three enhancement strategies to improve inference robustness.

Among them, the proposed fusion model—combining motion history with visual features—consistently achieved the best trade-off between accuracy, adaptability, and computational efficiency in low-quality settings.

These findings provide practical guidance for building vision systems that perform reliably in dynamic, unpredictable environments. Future work will focus on real-time deployment, multimodal integration, and domain adaptation to enhance robustness with minimal retraining.

ACKNOWLEDGMENT

This work was supported in part by the Power Beans Foundation (powerbeans.org), a non-profit organization focusing on robotics research and accessible education.

REFERENCES

[1] M. F. Afshar et al., "An efficient approach to monocular depth estimation for autonomous vehicle perception systems", *Sustainability*, vol. 15, no. 11, 2023.

[2] F. Matos et al., "A survey on sensor failures in autonomous vehicles: Challenges and solutions", *Sensors*, vol. 24, no. 16, 2024.

[3] J. Zhang, "Survey on monocular metric depth estimation", arXiv:2501.11841, submitted for publication.

[4] S. Gasperini et al., "Robust monocular depth estimation under challenging conditions", in *Proc. ICCV*, 2023.

[5] H. Yang et al., "Self-supervised monocular depth estimation in the dark: Towards data distribution compensation", in *Proc. IJCAI*, 2024.

[6] Y. Yao et al., "Improving domain generalization in self-supervised monocular depth estimation via stabilized adversarial training", in *ECCV*, 2024.

[7] J. Lee et al., "Robust monocular depth estimation in adverse weather conditions by unsupervised domain adaptation", in *Proc. ECAI,* 2024.

[8] https://www.vexrobotics.com/v5

[9] H. Cheng et al., "Wide & deep learning for recommender systems", in *Proc. DLRS (Workshop on Deep Learning for Recommender Systems),* 2016.

[10] D. Eigen et al., "Depth map prediction from a single image using a multi-scale deep network", in *Proc. NIPS*, 2014

[11] H. Wang et al., "TransWeather: Transformer-based restoration of images under various adverse weather conditions", in *Proc. CVPR*, 2022.

[12] A. Rosinol et al., "Kimera: An open-source library for real-time metric-semantic localization and mapping", in *Proc. ICRA*, 2020.

[13] C. Chen et al., "Selective sensor fusion for neural visual-inertial odometry", in *Proc. CVPR*, 2019

[14] https://openai.com/index/introducing-4o-image-generation/