Evaluation of Metric Monocular Depth Estimation Models Under Adverse Weather Conditions in Driving Scenarios

Nour Khalefa[1], Roberto Souza[2], Naser El-Sheimy[1]

[1] Dept. of Geomatics Engineering, University of Calgary,
[2] Dept. of Electrical and Software Engineering, University of Calgary,
2500 University Dr. N.W. Calgary, Alberta, Canada T2N 1N4 - (nour.khalefa, roberto.souza2, elsheimy)@ucalgary.ca

Keywords: monocular depth estimation, generalization, autonomous driving

Abstract

Metric monocular depth estimation has become increasingly important and is often used as a redundancy mechanism in autonomous driving, where accurate scene understanding is essential for safe decision-making. In this work, we evaluate three recently proposed models that represent the state-of-the-art (Depth Anything, PackNet-SfM, and UnidDepth) using zero-shot testing on the DrivingStereo dataset across diverse weather conditions, and benchmark their performance. Our analysis considers not only metric depth accuracy metrcis but also each model's ability to generalize under challenging environmental variations. While UniDepth achieves notable improvements over Depth Anything and PackNet-SfM, our results show that substantial progress is still needed for robust real-world deployment. To further assess its practical suitability for autonomous driving applications, we conduct a detailed examination of UniDepth's strengths, limitations, and failure modes.

1. Introduction

Depth estimation is important for understanding 3D structure from 2D representations, essential for navigating and interacting with the environment in applications such as autonomous driving. Monocular depth estimation is a low-cost option that only relies on one camera image stream to estimate depth and is often used as a safety redundancy on autonomous driving scenarios. The challenge in monocular depth estimation lies in accurately inferring depth information, which is inherently absent in a single image. When defining the depth estimation problem, it is important to distinguish between relative depth estimation, which predicts only the ordering of pixel depths, indicating which objects are closer or farther without a fixed scale or offset, and metric depth estimation, which infers real-world distance values from the camera plane. Accurately estimating metric depth is particularly valuable, as it provides precise measurements critical for navigation and planning.

Another key aspect is domain shift, where models trained on specific datasets may not generalize well to others due to variations in environmental conditions, lighting, or scene composition. This highlights the importance of evaluating a model's ability to generalize across diverse environments, ensuring robustness and reliability in real-world scenarios where conditions can vary significantly.

We compared three models in terms of their metric monocular depth estimation capabilities: Depth Anything (Yang et al., 2024), PackNet-SfM (Guizilini et al., 2020), and UniDepth (Piccinelli et al., 2024). Depth Anything is a large transformer model with strong generalization capabilities; although it was primarily trained for relative depth, the authors report that it can still generalize to metric depth. PackNet-SfM is a self-supervised approach designed to generalize across different datasets with metric depth. UniDepth is a state-of-the-art transformer model specifically trained for metric depth estimation. These three models were selected to evaluate a range of approaches, from relative depth generalization to models explicitly trained on metric depth.

2. Related Works

2.1 Supervised depth estimation

In the field of depth estimation, initial efforts treated the task as a regression problem, utilizing ground truth data collected via various sensors. Convolutional Neural Networks (CNNs) were at the forefront of these early approaches (Eigen et al., 2014). Moreover, supervised techniques evolved to formulate the problem as a classification-regression task (Bhat et al., 2021, Fu et al., 2018). These networks rely on the availability and quality of ground truth depth data, which may not always be available.

2.2 Self-supervised stereo depth estimation

In the domain of self-supervised stereo depth estimation, (Garg et al., 2016) used view synthesis as a self-supervised learning approach for stereo depth estimation from stereo pairs. Later, Monodepth (Godard et al., 2017) applied a photometric loss to enforce left-right consistency between stereo reconstructions, enhancing the depth estimation process.

2.3 Self-supervised Monocular depth estimation

Our primary interest lies in self-supervised monocular depth estimation. The first technique to use view synthesis for monocular depth estimation was SfM-learner (Zhou et al., 2017). This method, along with subsequent ones, leveraged a depth network combined with pose estimations to warp images and evaluate them against the original frames by minimizing the photometric loss. Monodepth2 (Godard et al., 2017) introduced per-pixel photometric error minimization and auto-masking to avoid occlusion issues and manage regions lacking texture. However, a notable challenge in monocular depth estimation is the scale ambiguity in pose and depth estimations. To counter this, SC-SfM-Learner (Bian et al., 2019) addressed this by implementing

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-473-2026 | © Author(s) 2026. CC BY 4.0 License.

a differentiable geometric loss and excluding pixels with geometric inconsistencies, thereby enhancing consistency across frames, although at a higher computational cost. PackNet-SfM (Guizilini et al., 2020) introduced a novel velocity loss during training that enabled the pose network, and thus the depth network, to have metrically accurate estimations

2.4 Domain generalization in depth estimation

Domain generalization aims to ensure a model trained on multiple source domains performs well on entirely new, unseen domains, despite potential significant differences between the training and test environments. MiDaS (Ranftl et al., 2020) introduces a novel approach to monocular depth estimation that leverages a mixture of diverse training datasets, including 3D films, to enhance model generalization. It proposes a unique training objective resistant to variations in depth scale and range, and employs multi-objective learning for data integration. The release of MiDaS v3.1 (Birkl et al., 2023) introduces a variety of new models for monocular depth estimation, utilizing diverse encoder backbones including several transformer and convolutional architectures. This update, motivated by the success of transformers in vision tasks, aims to enhance depth estimation quality and efficiency. Zoedepth(Bhat et al., 2023) extends previous work to metric depth by merging relative and metric depth estimation techniques. This model showcases superior generalization by training on 12 datasets for relative depth and fine-tuning on two for metric depth. The approach demonstrates remarkable zero-shot generalization across unseen indoor and outdoor datasets. SQLdepth (Wang et al., 2024) introduces a novel self-supervised method for monocular depth estimation that emphasizes fine-grained scene detail recovery and generalization. By constructing a self-cost volume through a Self Query Layer (SQL) that captures scene geometry, SQLdepth outperforms existing methods when pretrained self-supervised and fine-tuned metrically, highlighting its potential for applications requiring detailed depth perception. Depth Anything (Yang et al., 2024) focuses on creating a versatile and robust foundation model for monocular depth estimation, capable of handling diverse imaging conditions. By utilizing a data engine to gather and auto-annotate a vast dataset ($\sim$62M images), the project significantly broadens data coverage, aiming to minimize generalization errors. The approach integrates strategic data augmentation and auxiliary guidance from pre-trained semantic segmentation models, boosting the model's adaptability. Furthermore, fine-tuning the model with metric depth, inspired by the techniques used in ZoeDepth, resulted in superior performance. UniDepth (Piccinelli et al., 2024) proposed a universal approach to monocular depth estimation capable of delivering accurate metric predictions across a wide range of domains. The model implemented a self-promptable camera module that predicted dense camera representations to condition depth features. Additionally, it leveraged a pseudo-spherical output representation to disentangle camera and depth representations, enhancing the robustness of depth estimation across various environments.

3. Experimental Setup

3.1 Dataset

We chose the DrivingStereo dataset (Yang et al., 2019) for assessment because of its generation of high-quality disparity labels using a model-guided filtering approach applied to multi-frame LiDAR points, enhancing overall dataset quality. This dataset covers a wide range of driving scenarios, including urban, suburban, highway, elevated, and rural roads, as well as various weather conditions like sunny, rainy, cloudy and foggy settings. In comparison to other datasets, DrivingStereo offers numerous advantages including realistic scenes, abundant quantity, diverse scenarios, and superior quality disparity labels (Yang et al., 2019). Furthermore, a subset of the dataset comprises 2000 stereo images taken in four distinct weather conditions: foggy, sunny, cloudy, and rainy, with 500 images for each condition, as shown in the sample in Figure 1. Model evaluation is performed using left camera images from this subset and their corresponding depth maps.

3.2 Depth Anything

Depth Anything is a cutting-edge model in the realm of depth estimation, recently introduced in (Yang et al., 2024). This model falls within the zero-shot depth estimation category, aiming to utilize a varied dataset for training a monocular depth estimation model capable of accurately predicting depth for any given image, as shown in the model's training scheme in Figure 2. Drawing inspiration from a seminal work, MiDaS (Ranftl et al., 2020), which employs an affine-invariant loss to accommodate the varying depth scales and shifts encountered across different datasets, this enabled effective multi-dataset joint training. This approach enables MiDaS to provide relative depth information. The Depth Anything model adopts MiDaS's methodology for relative depth estimation, further reinforced by the findings of ZoeDepth's (Bhat et al., 2023) that a capable relative depth model can be effectively adapted for metric depth estimation with targeted fine-tuning using precise depth measurements. While the core of the model is designed for relative depth estimation, it undergoes fine-tuning to cater to metric depth estimation, particularly using metric depth information derived from the KITTI dataset (Geiger et al., 2012). This fine-tuned version, optimized for metric depth, is the variant evaluated in our work.

The model surpasses both MiDaS and ZoeDepth in estimating relative and metric depth, setting new benchmarks in the field. It also excels in zero-shot depth estimation tests. The paper includes zero-shot test results for the metric depth estimation network, which we will discuss in our results section, complemented by our analysis of the model's performance.

3.3 PackNet-SfM

The second model we evaluate is PackNet-SfM model (Guizilini et al., 2020), shown in Figure 3. It presents a state-of-the-art approach in the field of depth estimation from monocular images, claiming to achieve remarkable results on the KITTI benchmark at the time of its release. The primary purpose of this research is to address the challenges in depth estimation by introducing a novel self-supervised learning architecture that does not rely on labeled data. The paper's major contributions include the innovative PackNet architecture that utilizes symmetrical packing and unpacking blocks alongside 3D convolutions instead of max pooling, bilinear upsampling and 2D convolutions to preserve detailed spatial information, and the introduction of the dense depth for autonomous driving (DDAD) dataset to further challenge depth estimation models with more complex and varied urban driving scenes. Another key contribution is the model's ability to make scale-aware metric depth predictions by incorporating velocity information during training, enhancing the accuracy and applicability of depth estimations.

(a) Sunny (b) Rainy

(c) Foggy (d) Cloudy

Figure 1. Sample images from DrivingStereo under different weather conditions.

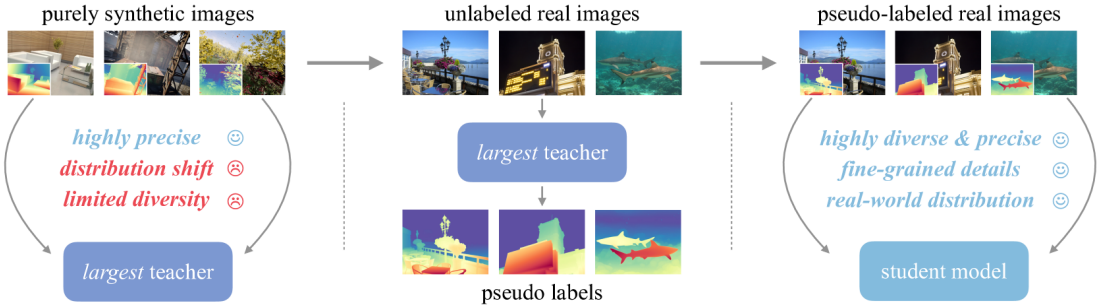


Figure 2. DepthAnything training scheme (Yang et al., 2024). he process begins with supervision from purely synthetic images, then the largest teacher model is then applied to unlabeled real images to generate pseudo labels. The combination of synthetic supervision and pseudo-labeled real data produces a highly diverse and realistic training set, used to train the student model.

Although the paper does not directly address the generalization capability of the model, the authors provide insights into why their approach might excel in this area. The self-supervised learning framework, combined with the detail-preserving PackNet architecture, allows the model to learn robust features that are likely to be applicable across different scenes and conditions. The paper also tests the generalization capabilities of their model, providing evidence of its robustness and adaptability to unseen environments. Their generalization tests along with our own conducted tests will be reported in the results section.

3.4 UniDepth

Most existing models either generalize well across domains but predict only relative depth, or they estimate metric depth accurately but fail to transfer across datasets with different camera setups. UniDepth proposes a semi-supervised, camera-aware framework that addresses both issues simultaneously. It is trained using a mix of labeled and unlabeled data from multiple datasets, with explicit conditioning on camera intrinsics. The goal is to develop a depth estimation model for road scenarios that generalizes to unseen datasets, diverse weather conditions, and varying camera configurations without requiring retraining. We use UniDepthv2 in our testing.

3.4.1 Promptable Camera Module One of UniDepth’s key innovations is its promptable camera module, which transforms camera intrinsics into a dense feature embedding. Rather than using raw intrinsics (e.g., focal length f_x,f_y , principal point c_x,c_y) directly, these values are passed through a lightweight multi-layer perceptron to produce a spatially-distributed, learned camera prompt. This embedding is then fused with the image features to inform the depth decoder about the imaging geometry. This design allows the model to dynamically adapt its predictions to a variety of camera setups without dataset-specific tuning. It plays a critical role in ensuring metric scale alignment across datasets captured with different sensors.

3.4.2 Pseudo-Spherical Representation UniDepth replaces the standard pixel-based output space with a pseudo-spherical output representation. In this formulation, each depth value is predicted along a ray defined by azimuth and elevation angles, rather than the 2D image grid. The predicted quantity corresponds to the radial distance from the camera center. This representation is decoupled from camera intrinsics, meaning that the network is not directly affected by focal length or principal point when generating depth. Instead, at training time, the predicted spherical coordinates are projected into 3D space using the actual camera intrinsics, and the resulting 3D points are supervised through losses defined in metric space. This design provides strong geometric consistency, making UniDepth’s predictions resilient to domain shift and inherently suitable for real-world deployment where camera parameters vary.

3.5 Evaluation metrics

We follow the evaluation metrics employed by (Godard et al., 2019), which consist of error metrics where lower values in-

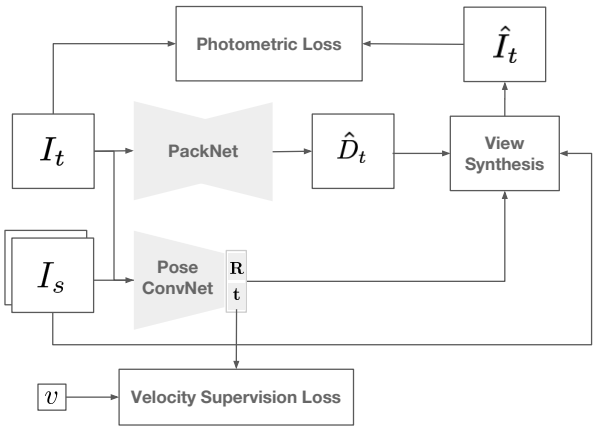


Figure 3. Packnet-sfm scale-aware self-supervised monocular structure-from-motion architecture (Guizilini et al., 2020).

dicate better performance, as well as accuracy metrics where higher values indicate better performance.

3.5.1 Absolute relative error

$$\text{AbsRel} = \frac{1}{n}\sum_{1}^{n}\left|\frac{g_t - p_{\text{red}}}{g_t}\right| \tag{1}$$

Where g_t is the ground truth depth map generated from the Velodyne sensor points included in the KITTI dataset, and p_{red} is the predicted depth map, and n is the number of pixels in the image.

3.5.2 Accuracy metric using threshold

$$\delta = \max\left(\frac{g_t}{p_{\text{red}}}, \frac{p_{\text{red}}}{g_t}\right) \tag{2}$$

$$\delta_1 = \frac{\#\text{pixels where } \delta < 1.25}{n}, \tag{3}$$

where δ_1 measures the proportion of predicted pixels whose depth values are within 25% of the ground truth.

3.5.3 Distance-Based Evaluation To analyze the dependency of depth estimation accuracy on scene distance, we perform a distance-based evaluation by partitioning ground-truth depth values into fixed metric intervals. Let the set of distance bin edges be defined as

$$\mathcal{B} = \{b_0, b_1, \ldots, b_N\}, \quad b_i \in \mathbb{R}^+, \tag{4}$$

where $b_i = 10i$ meters, resulting in bins of width $10\,\text{m}$ over the range $[0, 80)$ meters.

A pixel contributes to bin i if its ground-truth depth g_t satisfies

$$b_i \le g_t < b_{i+1}, \tag{5}$$

and only valid pixels are considered:

$$g_t > 0 \quad \text{and} \quad p_{\text{red}} > 0. \tag{6}$$

For each distance bin i, the Absolute Relative Error is computed as

$$\text{AbsRel}_i = \frac{1}{|\Omega_i|}\sum_{p\in\Omega_i}\left|\frac{g_t - p_{\text{red}}}{g_t}\right|, \tag{7}$$

where Ω_i denotes the set of valid pixels whose ground-truth depth falls within bin i.

The δ_1 accuracy for bin i is defined as

$$\delta_{1,i} = \frac{1}{|\Omega_i|}\sum_{p\in\Omega_i}\mathbf{1}\left(\max\left(\frac{g_t}{p_{\text{red}}}, \frac{p_{\text{red}}}{g_t}\right) < 1.25\right), \tag{8}$$

3.5.4 Image-Region-Based Evaluation To investigate spatial biases in depth estimation, we evaluate performance across different image regions. Each image is partitioned into three concentric regions based on the radial distance of each pixel from the image center: center, mid-periphery, and corners, depicted in Figure 5.

Let (x_p, y_p) denote pixel coordinates and (c_x, c_y) the image center. The radial distance of a pixel is given by

$$r_p = \sqrt{(x_p - c_x)^2 + (y_p - c_y)^2}. \tag{9}$$

Let $r_{\max}$ denote the maximum possible radius in the image. Using two normalized thresholds $0 < \alpha < \beta < 1$, the regions are defined as

$$\text{Center:} \quad r_p \le \alpha\, r_{\max}, \tag{10}$$

$$\text{Mid-periphery:} \quad \alpha\, r_{\max} < r_p \le \beta\, r_{\max}, \tag{11}$$

$$\text{Corners:} \quad r_p > \beta\, r_{\max}. \tag{12}$$

In our experiments, we use $\alpha = 0.33$ and $\beta = 0.66$.

For each image region k, the Absolute Relative Error is computed as

$$\text{AbsRel}_k = \frac{1}{|\Omega_k|}\sum_{p\in\Omega_k}\left|\frac{g_t - p_{\text{red}}}{g_t}\right|, \tag{13}$$

and the δ_1 accuracy as

$$\delta_{1,k} = \frac{1}{|\Omega_k|}\sum_{p\in\Omega_k}\mathbf{1}\left(\max\left(\frac{g_t}{p_{\text{red}}}, \frac{p_{\text{red}}}{g_t}\right) < 1.25\right), \tag{14}$$

where Ω_k denotes the set of valid pixels belonging to region k.

4. Experiments and Discussion

In the following section, we summarize zero-shot results for the three models as reported in their original papers on different datasets. We then present our evaluations on DrivingStereo, enabling comparison with reported results and across all models under a unified setting.

4.1 Depth Anything

We assessed their metric depth model, which is essentially their relative depth model fine-tuned on the KITTI dataset. According to the paper, this general relative depth model can be adapted for metric depth estimation through fine-tuning. However, for effective generalization in a zero-shot context, it re-

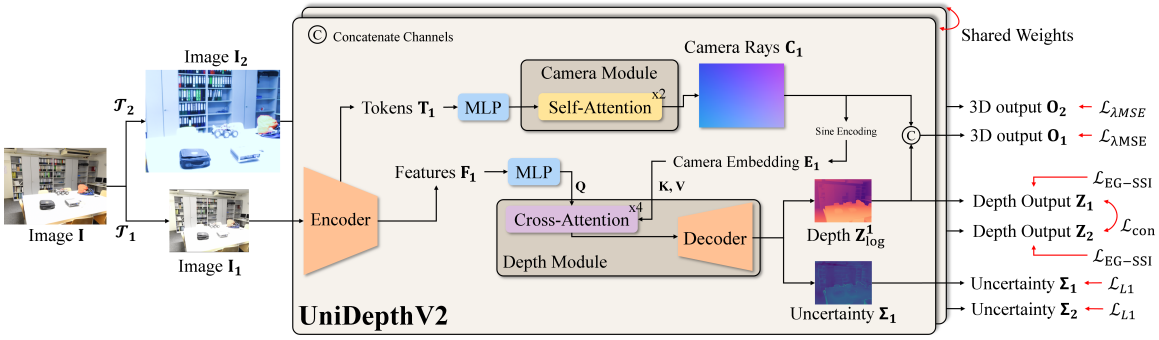
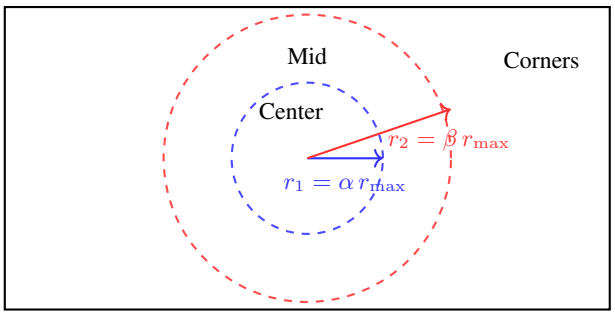


Figure 4. UniDepthv2 architecture (Piccinelli et al., 2024). An input image is augmented twice to produce two transformed views, which are processed by a shared encoder to extract image features and token representations. The camera module refines the token embeddings using multilayer perceptrons and self attention, producing camera ray representations. These camera embeddings serve as key and value inputs to the depth module, where cross attention integrates geometric cues with image features. The decoder then predicts depth maps and corresponding uncertainty estimates for both augmented views. All components operate with shared weights across the two augmented inputs.



$r_{\max}$: maximum image radius (pixels), $\alpha = 0.33$, $\beta = 0.66$

Figure 5. Definition of image regions used in the region-based evaluation. Pixels are grouped into center, mid-periphery, and corner regions based on their radial distance r from the image center, measured in image-plane pixels.

quires fine-tuning on a dataset that shares a similar environment. Thus, they fine-tuned the model on KITTI and conducted zero-shot tests on two other outdoor datasets: Virtual KITTI 2 and DIODE (Dense Indoor and Outdoor Depth Dataset) Outdoor (Vasiljevic et al., 2019), we report the tests from their paper in Table 1. Virtual KITTI 2 is a synthetic dataset designed to mirror the real KITTI environment, including the simulation of camera settings such as position, orientation, and field of view. In contrast, DIODE is a more varied dataset, encompassing scenes from different times of the day and seasons, including summer, fall, and winter. As a result, the model's performance on Virtual KITTI 2 was notably superior. Following this, we report the results of our experiments on DrivingStereo dataset. From the box plots in Figures 6 and 7, which stratify the dataset by weather conditions, we observe a significant variance in the δ_1 metric, particularly under cloud, rain, and fog conditions where less than 10% of the estimated pixels fall within the 25% error threshold. Conversely, performance under sunny conditions was comparable to that on DIODE, which aligns with expectations since the KITTI dataset primarily consists of daytime images. Therefore, the sunny condition only introduces dataset variance without the added complexity of different weather conditions. Figure 8 further illustrates the impact of severe weather on depth prediction accuracy, showcasing images with δ_1 metrics at both extremes. Conversely, Figure 7 reveals that the absolute error metrics are consistently better across all subsets compared to DIODE, possibly due to the varying depth ranges and scales between the datasets. However, a deeper comparison of the datasets is necessary for a definitive conclusion.

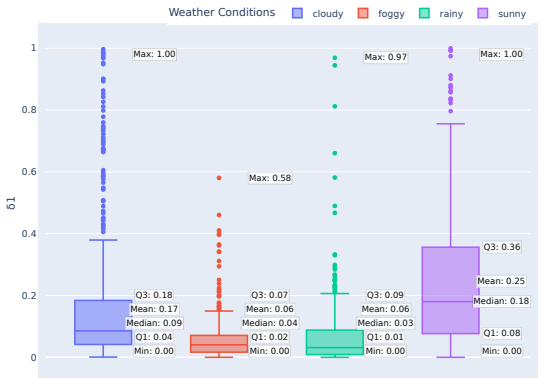


Figure 6. Box plot of δ_1 accuracy for Depth Anything.

Method	Virtual KITTI 2		DIODE Outdoor	
	AbsRel (↓)	δ_1 (↑)	AbsRel (↓)	δ_1 (↑)
ZoeDepth	0.106	0.844	0.814	0.237
Depth Anything	0.085	0.913	0.794	0.288

Table 1. Zero-shot metric depth estimation performance on outdoor scenes reported in (Yang et al., 2024).

4.2 PackNet-SfM

The authors report results for their scale-aware model, trained on KITTI and Cityscapes (Cordts et al., 2016) datasets, with performance on the NuScenes (Caesar et al., 2020) dataset, which was not included in the training data, detailed in Table 2. Given that the NuScenes dataset encompasses a variety of driving conditions and weather scenarios, it is comparable to DrivingStereo. In our experiments on DrivingStereo, analysis of the mean values depicted in Figure 9 reveals that the δ_1 values obtained are on par with those reported by Depth Anything test, with the sunny scenario being the best out of them but slightly worse than the sunny in depth anything test. Despite this, the δ_1

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-473-2026 | © Author(s) 2026. CC BY 4.0 License.

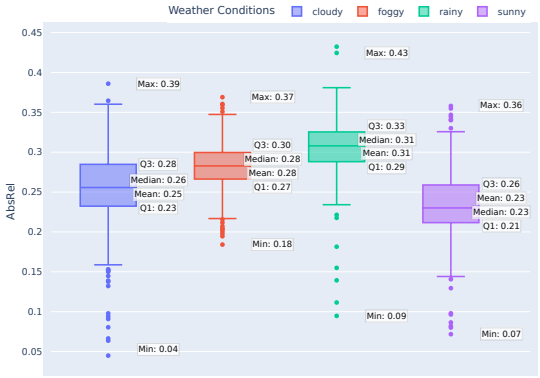


Figure 7. Box plot of absolute relative errors for Depth Anything.

values for the PackNet-SfM model are notably lower than those achieved in the NuScenes evaluation. Conversely, the mean values illustrated in Figure 10 closely align with those observed in Depth Anything test, though they do not surpass the NuScenes test outcomes. It is noteworthy that the PackNet-SfM model exhibits impressive generalization abilities even though it was not trained on large-scale datasets. This performance supports the authors' claims about the effectiveness of the features extracted by their model, which utilizes nearly lossless packing and unpacking blocks to achieve this outcome.

Method	AbsRel (↓)	Sq Rel (↓)	RMSE (↓)	RMSElog (↓)	δ_1 (↑)
ResNet50†	0.210	2.017	8.111	0.328	0.697
PackNet	0.187	1.852	7.636	0.289	0.742

Table 2. Generalization capability of different networks, trained on both KITTI and CityScapes and evaluated on NuScenes.

4.3 UniDepthv2

In our evaluation, we used UniDepth version 2 on the DrivingStereo dataset. Since DrivingStereo was included in the model's training data, we expected UniDepth version 2 to perform very well on this benchmark. The results do show a clear improvement compared to the other two models. Across weather conditions, UniDepth version 2 produces Absolute Relative Error values around ten percent on average. Although this is a solid result, it is still slightly higher than what would typically be expected for data that the model has been exposed to during training. From Figure 12, the cloudy condition in particular shows the widest spread of error values, with very low errors in some samples and much higher errors in others. This indicates a notable amount of variability in how well the model handles cloudy scenes, which often contain soft lighting, low contrast and subtle depth cues that can reduce monocular reliability. A similar pattern appears in the δ_1 accuracy results for the δ_1 model in Figure 11. The overall performance of δ_1 is extremely strong, with values close to perfect in most cases. However, the distribution under cloudy conditions again shows the largest spread. δ_1 achieves very high accuracy in many samples, but some cloudy scenes still lead to significantly lower values. This wide range suggests that even for a model that performs near perfectly on average, certain cloudy scenarios remain challenging. Together, these findings show that UniDepth version 2 achieves meaningful improvements but still yields slightly higher errors than expected for an in distribution dataset. δ_1 reaches near optimal performance on average, yet the variability in cloudy scenes demonstrates that both models can be sensitive to environmental fluctuations within the same weather category. This highlights that even when the model has seen the dataset during training, monocular depth estimation can still exhibit considerable variation under different lighting and scene structures, and cloudy conditions make this particularly evident.

To further understand the behavior of UniDepth, we extended our evaluation by performing two additional analyses. The first examines how the model performance changes with increasing scene distance, and the second explores how accuracy varies across different image regions. Both analyses help reveal the specific conditions under which the model struggles, despite having been trained on the DrivingStereo dataset.

For the distance analysis, we stratified the ground truth depth into intervals of ten meters and evaluated AbsRel and δ_1 independently within each bin. This allows us to distinguish how well the model predicts nearby structures compared to distant ones. The trend is consistent across all weather conditions, as shown in Figures 13 and 14. The model performs very well at close ranges, where strong geometric cues are available, but its accuracy decreases steadily as distance increases. In cloudy weather, for example, AbsRel rises from approximately 0.079 in the zero to ten meter range to nearly 0.145 in the seventy to eighty meter range. At the same time, δ_1 drops from about 0.993 to roughly 0.794. Similar gradual degradation appears in foggy, rainy and sunny conditions, with the most distant bins always showing the lowest accuracy values. These patterns confirm that long range depth estimation remains difficult, even for models that perform strongly at short range. We further examined how prediction accuracy varies across the image. To do this, we divided each frame into three spatial regions: the central region, the mid region surrounding it, and the four corner regions. The results display a consistent structure, as shown in Figures 15 and 16. The center of the image achieves the lowest AbsRel values and the highest δ_1 values. The mid region performs slightly worse but remains close to the center. The corners show the weakest performance in all weather types. For instance, in cloudy conditions, AbsRel increases from about 0.093 in the center to more than 0.103 in the corners, and δ_1 decreases from about 0.966 in the center to about 0.952 in the corners. A similar pattern appears in foggy, rainy and sunny weather. This reflects well known characteristics of camera lenses and dataset distributions. The center of the image typically contains more informative content and less distortion, while the periphery suffers from reduced detail, vignetting and lower texture quality, all of which make depth estimation more difficult. Together, these analyses provide a clearer picture of the model's limitations. UniDepth performs well and shows strong overall accuracy, but its performance declines smoothly as distance increases and as one moves away from the image center. These trends remain visible in all weather conditions. They demonstrate that even when the model has been trained on a dataset, there are predictable geometric and optical factors that constrain performance.

4.4 Additional discussion points

The range and spread depicted in the box plots for all three models, marked by numerous outliers, reflect a significant degree of performance fluctuation. This inconsistency might stem from the inherent diversity within the dataset or potentially from the models inability to perform consistently across the data. Further exploration into the datasets used for the original training

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-473-2026 | © Author(s) 2026. CC BY 4.0 License.

(a) Image (sunny) with best δ_1 (b) Depth map with best δ_1 (c) Image (rainy) with worst δ_1 (d) Depth map with worst δ_1

Figure 8. Depth Anything qualitative test.

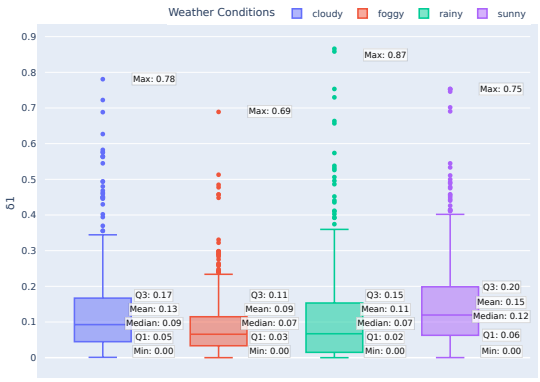


Figure 9. Box plot of δ_1 accuracy for PackNet-SfM.

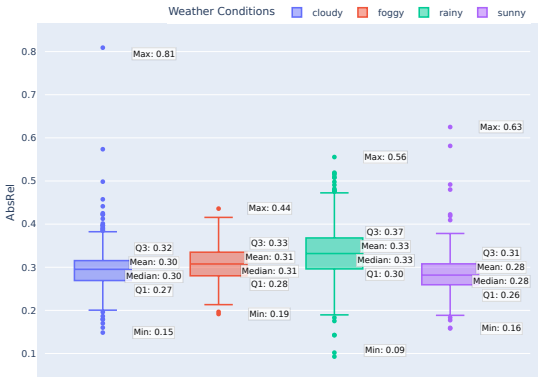


Figure 10. Box plot of absolute relative errors for PackNet-SfM.

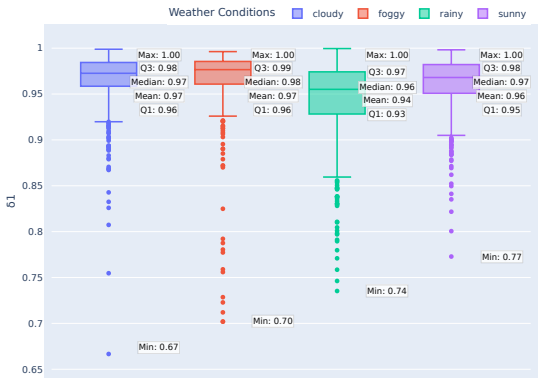


Figure 11. Box plot of δ_1 accuracy for UniDepth

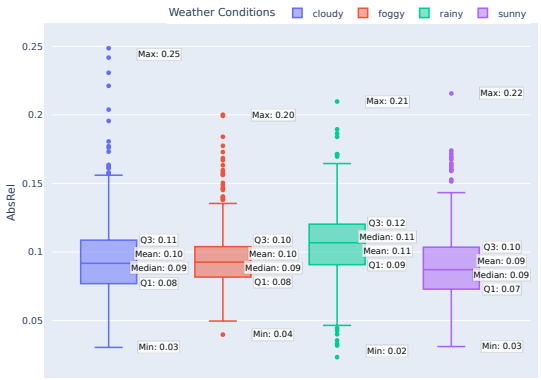


Figure 12. Box plot of absolute relative errors for UniDepth.

and evaluation processes, compared to those employed in our assessments, is essential. Such an analysis would clarify the distinctions that lead to the variability in model performance and the extent of the models’ generalization abilities.

5. Conclusion

This study’s examination of UniDepth, Depth Anything and PackNet-SfM models highlights their ability to adapt to new environments, as shown through zero-shot evaluations of Depth Anything and PackNet-SfM on the DrivingStereo dataset. However, despite the strong potential of these models, the results demonstrate that their generalization across different conditions remains constrained. This underscores the need for further refinement to improve robustness and ensure consistent performance in the diverse and unpredictable scenarios encountered in real-world applications.

For UniDepth, which was evaluated in-distribution rather than zero-shot, we conducted a detailed series of analyses to better understand its behavior and remaining limitations. Distance-based stratification revealed a clear decline in accuracy as scene depth increased. UniDepth performed strongly for nearby objects, yet its accuracy degraded progressively with distance, and the farthest depth bins showed the highest Absolute Relative Error and lowest δ_1 accuracy. This pattern appeared across all weather categories, indicating that long-range depth prediction remains inherently challenging even when the model has been trained on similar data.

Weather-conditioned evaluation further exposed significant variability. In particular, cloudy scenes produced the widest spread in both AbsRel and δ_1, showing that soft lighting and reduced contrast can destabilize the predictions. Additionally, the radial analysis revealed a consistent drop in performance from the image center toward the corners, highlighting the influence

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-473-2026 | © Author(s) 2026. CC BY 4.0 License.

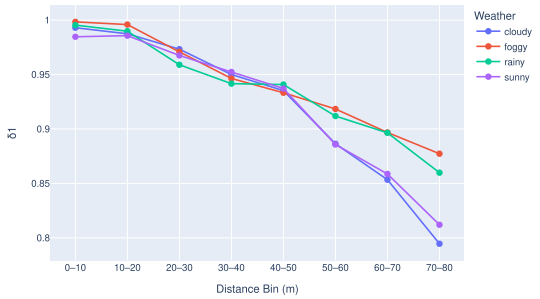


Figure 13. δ_1 accuracy of UniDepth stratified by distance.

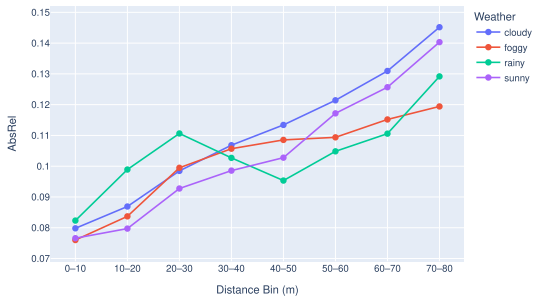


Figure 14. Absolute relative errors of UniDepth stratified by distance.

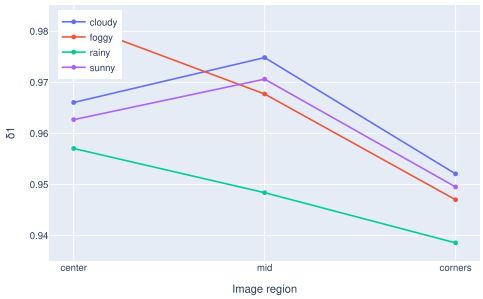


Figure 15. δ_1 accuracy of UniDepth stratified by pixel regions.

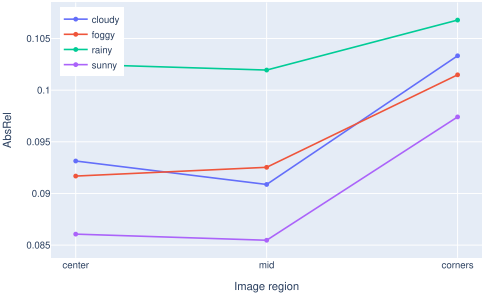


Figure 16. Absolute relative errors of UniDepth stratified by pixel regions.

of lens characteristics, peripheral distortions and dataset biases that reduce texture and geometric cues.

Together, these extended tests show that even when UniDepth is evaluated on data it has previously encountered, its performance is not uniformly reliable. The observed declines across distance ranges, the sensitivity to weather variability and the regional differences across the image all point to systematic limitations that must be addressed. Strengthening model consistency under these factors will be essential for achieving truly robust and dependable depth estimation in practical real-world deployments.

Acknowledgements

This research has been supported by funding from Prof. Naser El-Sheimy from NSERC CREATE and Canada Research Chairs programs. It was also enabled by support provided by the Research Computing Services group at the University of Calgary.

References

Bhat, S. F. et al., 2021. Adabins: Depth estimation using adaptive bins. *IEEE/CVF CVPR*, 4009–4018.

Bhat, S. F. et al., 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.

Bian, J. et al., 2019. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32.

Birkl, R. et al., 2023. MiDaS v3.1 – A Model Zoo for Robust Monocular Relative Depth Estimation. arXiv:2307.14460 [cs].

Caesar, H. et al., 2020. nuscenes: A multimodal dataset for autonomous driving. *IEEE/CVF CVPR*, 11621–11631.

Cordts, M. et al., 2016. The cityscapes dataset for semantic urban scene understanding. *IEEE/CVF CVPR*, 3213–3223.

Eigen, D. et al., 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.

Fu, H. et al., 2018. Deep ordinal regression network for monocular depth estimation. *IEEE/CVF CVPR*, 2002–2011.

Garg, R. et al., 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. *ECCV*, Springer, 740–756.

Geiger, A. et al., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. *IEEE/CVF CVPR*, 3354–3361.

Godard, C. et al., 2017. Unsupervised monocular depth estimation with left-right consistency. *IEEE/CVF CVPR*, 270–279.

Godard, C. et al., 2019. Digging into self-supervised monocular depth estimation. *IEEE/CVF ICCV*, 3828–3838.

Guizilini, V. et al., 2020. 3d packing for self-supervised monocular depth estimation. *IEEE/CVF CVPR*, 2485–2494.

Piccinelli, L. et al., 2024. Unidepth: Universal monocular metric depth estimation. *IEEE/CVF CVPR*, 10106–10116.

Ranftl, R. et al., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3), 1623–1637.

Vasiljevic, I. et al., 2019. DIODE: A dense indoor and outdoor depth dataset. arXiv:1908.00463 [cs].

Wang, Y. et al., 2024. SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation. *Proceedings of AAAI*, 38number 6, 5713–5721.

Yang, G. et al., 2019. DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios. *IEEE/CVF CVPR*, Long Beach, CA, USA, 899–908.

Yang, L. et al., 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. arXiv:2401.10891 [cs].

Zhou, T. et al., 2017. Unsupervised learning of depth and ego-motion from video. *IEEE/CVF CVPR*, 1851–1858.