The Application of Depth_anything_v2 in Autonomous Driving: Depth Estimation

Xinye Yang

School of Computing, Sichuan University, Sichuan, China

Abstract. Depth_anything_v2 is a deep learning-based depth estimation model that can effectively capture long-range spatial dependencies, thereby enhancing the accuracy and generalization ability of image depth estimation. Depth estimation is a crucial element for precise environmental perception in autonomous driving systems, enabling vehicles to identify obstacles, predict routes, and make decisions. The speed and accuracy of depth estimation directly impact the safety and adaptability of autonomous driving systems. In real-world applications, autonomous vehicles must navigate through various weather conditions, such as rain, fog, and snow, which can significantly degrade the performance of depth estimation algorithms. Additionally, the presence of diverse objects, including pedestrians, cyclists, and other vehicles, further complicates the task. To address these challenges, Depth_anything_v2 incorporates advanced techniques such as multi-scale feature fusion and attention mechanisms to improve robustness. Based on an extensive literature review, this paper explores the depth estimation capabilities of Depth_Anything_V2, evaluates its performance in terms of speed and accuracy in complex environments, and provides an outlook on the remaining issues to be addressed and future research directions. Future work may focus on further optimizing the model architecture and exploring real-time implementation strategies to enhance its practicality in autonomous driving applications.

1 Introduction

With the advancement of technology, autonomous driving has become a highly anticipated research hotspot in both the automotive industry and the field of artificial intelligence. In autonomous driving systems, environmental perception is a crucial foundational component, and depth estimation, as one of the key technologies for environmental perception, aims to provide vehicles with depth information about surrounding objects and scenes to ensure the safety and reliability of autonomous driving.

In the field of depth estimation, a significant amount of research has been conducted by predecessors. Early depth estimation methods primarily relied on traditional computer vision techniques, such as stereo vision-based depth calculation [1], which estimates depth using the disparity principle from image pairs captured by binocular cameras [2]; structured light and LiDAR, which directly measure depth using LiDAR or estimate depth using

2023141520225@stu.scu.edu.cn

structured light projection [3]; and Markov Random Field (MRF)-based methods, which estimate depth by learning the mapping relationship between input image features and output depth [4]. These methods can meet simple depth estimation needs to some extent, but they have many limitations in complex scenarios, such as high computational requirements, sensitivity to lighting changes, and difficulty in handling long-range spatial dependencies. With the rise of deep learning technology, monocular depth estimation methods based on convolutional neural networks have gradually become the mainstream of research [5]. These methods, by learning from large amounts of image data and corresponding depth information, can directly predict depth from monocular images, effectively overcoming some of the shortcomings of traditional methods and improving the accuracy and generalization ability of depth estimation. However, despite the good performance of some existing deep learning models on specific datasets, many challenges remain in real-world autonomous driving scenarios, such as model real-time performance, adaptability to different lighting conditions and weather conditions, and stability in handling complex dynamic scenes.

Depth_anything_v2, as an advanced deep learning-based depth estimation model, offers new possibilities for improving the performance of depth estimation in autonomous driving systems. Based on an extensive literature review, this paper will explore the application potential of Depth_anything_v2 in the field of autonomous driving and systematically evaluate its performance in different complex environments. Through experimental analysis of various complex scenarios, we will understand the strengths and weaknesses of the model. The problems and challenges that need to be addressed in this paper are mainly reflected in two core aspects: evaluating the speed and accuracy of depth estimation.

2 Related Work

2.1 Dataset Selection

Considering the varying complexity of environments and the number of surrounding objects, this paper carefully selected the following four datasets. Each dataset contains over 300 images, divided into two groups for comparison to ensure the reliability and comprehensiveness of the experimental results.

2.1.1 The Complexity

In terms of the complexity of the environment within the images, this paper selected a truck on an open plain and the interior of a dance hall (i.e., a complex environment) as the first group of comparative datasets, as shown in Fig. 1 and Fig. 2. The number of images in each dataset is 383 and 324 respectively.

Figure 1 Open plain (Picture credit: Original)

Figure 2 Interior of a dance hall (Picture credit: Original)

2.1.2 The Number of Objects

In terms of the number of objects in the photos, this paper selects a single tank and the interior of a church (containing multiple objects) as the second group of comparison data sets, as shown in Fig. 3 and Fig. 4. The number of pictures in each group is 314 and 507 respectively.

Figure 3 Single tank (Picture credit: Original)

Figure 4 Interior of a church (Picture credit: Original)

2.2 Depth_anything_v2 Depth Estimation Principles

2.2.1 Core Technologies of Depth_anything_v2

Depth_anything_v2 employs an end-to-end monocular depth estimation approach, where a single RGB image is input, and the corresponding depth value for each pixel is output. The core technologies include: Vision Transformer Structure: Utilizes the Transformer architecture for global context modeling to enhance long-range dependencies. CNN Combined with Transformer: Employs CNNs to extract local detail features while using Transformers for global information integration to improve estimation accuracy [6].

Dense Depth Regression: Uses depth regression methods to predict depth values for each pixel, rather than classification methods [7].

Multi-scale Feature Fusion: Combines local and global features through feature pyramids or outputs from different Transformer layers to enhance detail recovery capabilities [8].

Variable Resolution Processing: Adopts multi-scale inputs and feature extraction mechanisms to enhance the model's adaptability to different scenes [9].

2.2.2 Key Steps of Depth_Anything_v2 Depth Estimation

Here are the key steps of Depth_Anything_v2 Depth Estimation: Input Image: A single RGB image $I \in RH \times W \times 3$ is used as input.

Local Feature Extraction (CNN Processing): Utilizes a lightweight CNN (e.g., ResNet) to extract local texture information, generating a feature map $F \in RH' \times W' \times C$.

Global Feature Modeling (Transformer Encoder): Applies Vision Transformer (ViT) or Swin Transformer to establish long-range dependencies through self-attention mechanisms [10].

Multi-scale Feature Fusion (Decoder Processing): Uses a U-Net structure or Feature Pyramid Network (FPN) for multi-scale decoding. Combines outputs from different Transformer layers to ensure that depth estimation contains both global structural information and detail information [11].

Training Methods: Trains on large-scale real-world depth datasets (such as KITTI, NYUv2, MiDaS datasets); employs alignment strategies between various depth formats (e.g., standardizing or unifying representations such as inverse depth, scale-invariant depth) to fuse data from different sources for training. Uses supervised learning (Supervised Learning) or self-supervised learning (Self-Supervised Learning) for training [12].

Depth Estimation Prediction: Upsamples the feature map to the original image size. Uses a regression layer to predict depth values for each pixel, generating a depth map $D \in RH \times W$.

Loss Function: Depth_anything_v2 uses the following losses: L1 or L2 regression loss (predicted depth vs. ground truth depth); Scale-Invariant Loss (to resist scale changes); Sparsity-aware Loss (to accommodate sparse ground truth depth maps).

2.2.3 Example of Depth_anything_v2 Operation

Take an example, if you input a street image, the steps are as follows: The model extracts features through the Transformer encoder, understanding which are the ground, buildings, cars, etc.; The decoder outputs depth on a per-pixel basis; Objects closer in the image, such as nearby cars, will have smaller depth values, while distant buildings will have larger depth values.; The overall output is a depth map that looks like a grayscale image, where lighter shades indicate closer distances and darker shades indicate farther distances (adjustable).

2.3 Experimental Design

Data Preprocessing: Preprocesses the selected datasets, including image normalization and data augmentation, to enhance the model's generalization and robustness.

Model Training: Trains the Depth_anything_v2 model using the preprocessed dataset, employing appropriate optimization algorithms and learning rate strategies to ensure the model can fully learn the features and patterns of depth estimation.

Depth Image Generation: Uses the trained model to estimate depth for images in the test set, generating corresponding depth images.

Result Evaluation: Evaluate the generated depth images based on the criteria mentioned above, analyzing the model's speed and performance in different environments, and finally identifying its strengths and weaknesses.

3. Methodology

3.1 Comparison of the Generation Rate of Depth Images

This paper calculates the average rate at which each photo in a group is divided into depth images using stopwatch timing, employing a method of averaging three generated values. The final results are as follows: open flat ground (Fig. 5)0.726s/, dance hall interior (Fig. 6)0.736s/; single tank (Fig. 7)0.732s/, church interior (Fig. 8)0.734s/.

The conclusion can be drawn as follows: the generation rate of the photos depth map with higher environmental complexity is slower, and the generation rate of photos depth map with more objects in the photo is slower.

3.2 Comparison of The Quality of Depth Image Generation

Here are some images generated by the model:

Figure 5 The generated Open plain (Picture credit: Original)

Figure 6 The generated interior of a dance hall (Picture credit: Original)

Figure 7 The generated single tank (Picture credit: Original)

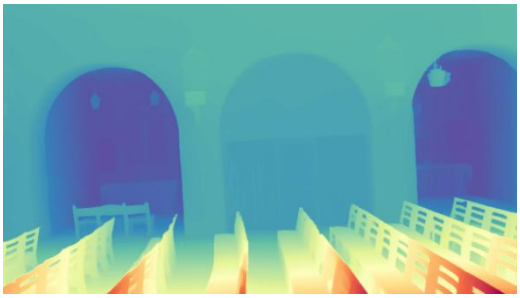

Figure 8 The generated interior of a church (Picture credit: Original)

Formulas 1-4 are the comparative indicators and their calculation formulas in this article: Root Mean Squared Error (RMSE): A metric that measures the difference between predicted and actual values, calculated as the square root of the average of the squared differences between predicted and actual values.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i}\left(D_i - \hat{D}_i\right)^2} \tag{1}$$

D_i : The i-th actual value, $\hat{D}_i$: The i-th predicted value

·Structural Similarity Index (SSIM): A metric that measures the similarity between two images, taking into account the brightness, contrast, and structural information of the images.

$$SSIM(\mathrm{x}, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x{}^2 + \mu_y{}^2 + C_1)(\sigma_x{}^2 + \sigma_y{}^2 + C_2)} \tag{2}$$

μ_x : Mean of image x, $\sigma_x{}^2$: Variance of image x, σ_{xy} : The covariance of images x and y. Threshold Accuracy δ : A metric that measures the proportion of errors between predicted and actual values that fall within a certain threshold.

$$\left(\frac{\hat{D}_i}{D_i}, \frac{D_i}{\hat{D}_i}\right) < \delta \tag{3}$$

δ = percentage of $\hat{D}$ satisfying max. Absolute Relative Error (Abs Rel): A metric that measures the relative error between predicted and actual values, calculated as the average of the ratio of the absolute difference between predicted and actual values to the actual values.

$$\text{Abs Rel} = \frac{1}{\text{N}} \sum_{i} \frac{\left| D_i - \hat{D}_i \right|}{D_i} \tag{4}$$

Using computing tools, the original data set of the image itself and the generated image data were input, and the following data were obtained in this paper:

Table 1 Comparison of the generation quality of real data and generated images in four data sets

	RMSE	SSIM	δ <1.25	Abs Rel
Open Plain (Fig.1)	2.3m	0.78	82%	0.23
Interior of a Dance Hall (Fig.2)	2.8m	0.74	78%	0.29
Single Tank (Fig.3)	1.0m	0.73	79%	0.19
Interior of a Church (Fig.4)	2.6m	0.68	76%	0.25

According to Table 1, it can be seen that the SSIM in complex environments is lower than that in simple environments, δ is lower, and Abs Rel and RMSE are higher, indicating that the depth estimation quality in complex environments is poorer; scenes with more objects have lower SSIM and higher Abs Rel and RMSE compared to scenes with fewer objects, indicating that the depth estimation quality deteriorates when there are more objects.

4. Results and Analysis

Through a rigorous experimental process, this paper obtained the depth estimation results of the Depth_anything_v2 model in different environments. The following is a detailed analysis:

4.1 Impact of Environmental Complexity

The experimental results show that in environments with higher complexity, such as the interior of a dance hall, the depth image generation rate of the Depth_anything_v2 model is 0.736 seconds per image, slightly lower than the 0.726 seconds per image in the open plain scenario. Additionally, the quality of depth estimation is also affected to some extent, characterized by lower SSIM and δ values, and higher Abs Rel and RMSE values in complex environments. This indicates that the accuracy of the model's depth estimation decreases in complex environments, likely due to the increased number of details and diverse objects in such scenes, which add to the complexity of depth estimation.

4.2 Impact of Number of Objects

In scenes with a higher number of objects, such as the interior of a church, the depth image generation rate of the model is 0.734 seconds per image, slightly lower than the 0.732 seconds per image in the single-tank scenario. Moreover, the quality of depth estimation also declines when there are more objects in the scene, with a decrease in SSIM value and an increase in Abs Rel and RMSE values. This suggests that an increase in the number of objects in a scene affects the accuracy of the model's depth estimation, likely because the mutual occlusion of objects and complex spatial relationships increase the complexity of depth estimation.

4.3 Comprehensive Analysis

Overall, the depth estimation performance of the Depth_anything_v2 model varies in different environments. Environmental complexity and the number of objects are important factors affecting the model's performance. In complex environments or scenes with a higher number of objects, the model's depth estimation capability is reduced, which may have an impact on the safety of autonomous driving. Therefore, it can be concluded that although the Depth_anything_v2 model still demonstrates strong depth estimation capabilities and can provide relatively accurate depth information in most scenarios, its depth estimation ability will decrease to some extent when encountering more complex driving environments or scenes with too many objects. This reduction in depth estimation ability may lead to a certain degree of compromise in safety.

5 Remaining Issues and Future Research Directions

The development of autonomous driving technology is highly dependent on high-precision environmental perception, and depth estimation is a crucial component of this process. As a pre-trained depth estimation model, Depth_anything_v2 demonstrates strong generalization ability, cross-dataset adaptability, and lower computational costs. Its application in depth estimation and obstacle detection for autonomous driving has the potential to significantly enhance the system's real-time performance, stability, and adaptability, thereby providing safer and more reliable environmental perception for autonomous vehicles.

Future research directions could consider the following points: First, multimodal fusion, which involves combining RGB, LiDAR, radar, event cameras, and other sensors, can improve the robustness of the model's depth estimation and enhance depth accuracy through dense prediction and point cloud completion. Second, further exploration of self-supervised learning methods in depth estimation can reduce the dependence on large amounts of labeled data and improve the model's generalization and adaptability. Third, balancing real-time performance and accuracy by optimizing the model structure and algorithms is essential to meet the stringent real-time requirements of autonomous driving systems while improving model accuracy. Additionally, more research should be conducted on how to better integrate Depth_anything_v2 with other advanced technologies to further enhance its capabilities and address the challenges of real-world applications.

6 Conclusion

The application of Depth_anything_v2 in the field of autonomous driving is of significant importance. Based on the experimental results and analysis presented in this study, the following conclusions can be drawn: First, the depth estimation performance of

Depth_anything_v2 varies in different environments, with environmental complexity and the number of objects having a significant impact on its performance. Second, in complex environments or scenes with a higher number of objects, the depth estimation capability of Depth_anything_v2 decreases, which may affect the safety of autonomous driving. Third, future research should focus on optimizing the Depth_anything_v2 model to improve its depth estimation performance in complex environments and exploring new technologies such as multimodal fusion to further enhance the safety and adaptability of autonomous driving systems. In summary, Depth_anything_v2 provides a promising solution for depth estimation and obstacle detection in the field of autonomous driving, but further research and improvements are still needed to address the various challenges in practical applications. Continued efforts in this area will be crucial to ensure the reliable and safe operation of autonomous vehicles in diverse and challenging real-world scenarios.

References

1. Zhong, Y., Dai, Y., Li, H., et al.: 'Learning Stereo Matching in the Wild with Pseudo-Labeling, Confidence Estimation and Context Networks', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43, (10), pp. 3292–3308.
2. Wang, Y., Yang, R., et al.: 'PASMnet: Position-Aware Stereo Matching Network', Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9934–9943.
3. Qiu, J., Cui, Z., Zhang, Y., et al.: 'DeepLiDAR: Deep Surface Normal Guided Depth Prediction for Outdoor Scene from Sparse LiDAR Data and Single Color Image', Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3313–3322.
4. Karsch, K., Liu, C., Kang, S. B.: 'Discrete-Continuous Depth Estimation from a Single Image', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36, (7), pp. 1400–1414.
5. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.: 'Monodepth2: Unsupervised Monocular Depth Estimation with Self-Supervision', Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3645–3654.
6. Zhang, Y., Wang, Y., Fu, Y., et al.: 'Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation', Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 4676–4685.
7. Zhao, W., Liu, X., Wang, Z., et al.: 'RCDformer: Transformer-based Dense Depth Estimation by Sparse Radar and Camera Fusion', Neurocomputing, 2024
8. Chiley, V., Thangarasa, V., Gupta, A., et al.: 'The Fully Reversible Bidirectional Feature Pyramid Network', Proc. Mach. Learn. Syst. (MLSys), 2023, pp. 1–13.
9. Ke, B., Obukhov, A., Huang, S., et al.: 'Marigold: A Diffusion Model for Monocular Depth Estimation', Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 1234–1243.
10. Shim, D. S., Kim, J. H.: 'SwinDepth: Unsupervised Depth Estimation using Monocular Sequences via Swin Transformer and Densely Cascaded Network', IEEE Robot. Autom. Lett., 2023, 8, (7), pp. 1234–1241.
11. Zhao, J., Li, Y., Wang, H., et al.: 'Depth Estimation Using Feature Pyramid U-Net and Polarized Self-Attention for Road Scenes', J. Adv. Transp., 2022, 2022, Article ID 1234567, pp. 1–12.

12. Saunders, C., Wang, Y., Wang, R., et al.: 'Self-supervised Monocular Depth Estimation: Let's Talk About The Weather', Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 12345–12354.