# Comparative Analysis of Monocular Depth Estimation Models

Vorrapard Kumthongdee
*New York University*
vk2584@nyu.edu

Nikita Makarov
*New York University*
nom2188@nyu.edu

Haoyang Pei
*New York University*
hp2173@nyu.edu

***Abstract*—Monocular depth estimation predicts depth from a single RGB image, making it useful for tasks like robotics, autonomous driving, and augmented reality. This paper compares five state-of-the-art models: Monodepth2, DenseDepth, DPT, AdaBins, and Depth Anything V2. We evaluate these models on two benchmark datasets, KITTI and NYUv2, using standard metrics such as MAE, RMSE, and threshold accuracy. The results show that Transformer-based and hybrid models outperform traditional CNN-based models. Depth Anything V2 achieves the best qualitative performance, producing smooth and detailed depth maps, but its relative depth predictions limit its quantitative results on older benchmarks. Our findings highlight the need for updated benchmarks to better evaluate modern models and guide future research in monocular depth estimation. The implementation, fine-tuning scripts, and preprocessing tools are available on our GitHub repository: https://github.com/vorrapard/DepthScope.**

***Index Terms*—monocular depth estimation, convolutional neural networks, vision transformers, hybrid models, self-supervised learning, transfer learning.**

## I. Introduction

Monocular depth estimation, the task of predicting the depth of each pixel in a scene from a single RGB image, has become an important challenge in computer vision with applications such as robotics [1], autonomous driving [2], augmented reality (AR) [3] and 3D reconstruction [4]. Unlike stereo depth estimation, which relies on multiple camera viewpoints that requires precise camera calibration and synchronization and leads to a higher cost due to the system complexity [5], monocular depth estimation derives depth information solely from a single image. This makes it a highly ill-posed problem [6], as critical depth cues like parallax and stereopsis are absent.

Early approaches to monocular depth estimation were largely based on traditional computer vision techniques such as structure-from-motion (SfM) [7] or shape-from-shading [8]. However, recent advances in deep learning have greatly enhanced the performance of monocular depth estimation systems. Convolutional neural networks (CNNs), in particular, have shown strong results by learning to predict depth from large datasets of paired RGB-depth images [4], [9], [10]. Models such as Monodepth2 [11] and DenseDepth [12] have demonstrated the ability of CNNs to infer depth in both indoor and outdoor environments with high accuracy by leveraging supervised or self-supervised learning.

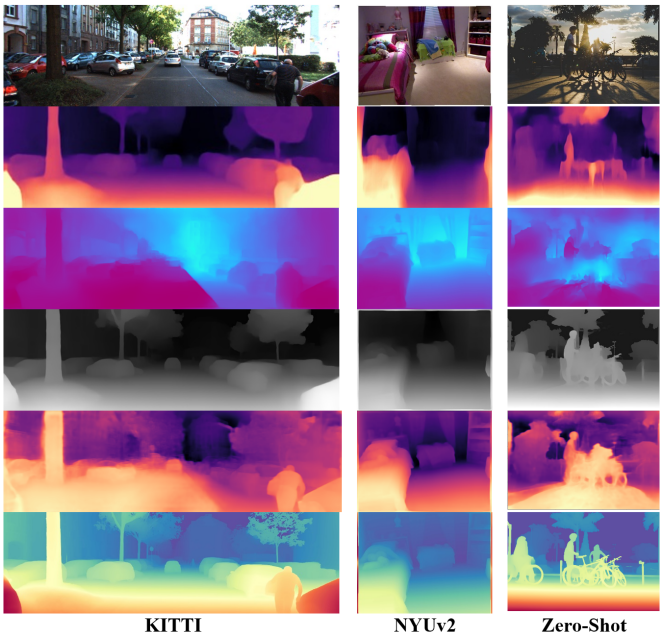


Fig. 1. Qualitative comparison on the KITTI dataset [13], NYUv2 dataset [14], and zero-shot predictions (from top to bottom): original image, Monodepth2 [11], DenseDepth [12], DPT [15], AdaBins [16], and Depth Anything V2 [17]

The introductions of Transformer [18] and Vision Transformers (ViT) [19] has further pushed the boundaries of monocular depth estimation. Transformers can capture long-range dependencies and global context in an image, making them especially effective in complex environments where CNNs might struggle with distant or ambiguous depth cues [20], [21]. While the Dense Prediction Transformer (DPT) [15] is a purely transformer-based model, AdaBins [16] and Depth Anything V2 [17] employ a hybrid architecture integrating both transformers and CNNs. This hybrid approach allows AdaBins and Depth Anything V2 to harness the localized feature extraction capabilities of CNNs alongside the global contextual modeling strengths of transformers. Consequently, all DPT, AdaBins, and Depth Anything V2 are recognized as state-of-the-art models in the field of monocular depth estimation, each contributing uniquely to advancing depth prediction accuracy and robustness.

Despite the progress in monocular depth estimation, there

is still significant interest in evaluating and understanding the trade-offs between different model architectures—such as CNNs, transformers [22], and hybrid approaches [23]—in terms of accuracy, computational efficiency, and ability to generalize across diverse scenes. In this paper, we aim to compare various monocular depth estimation models, including Monodepth2, DenseDepth, DPT, AdaBins, and Depth Anything V2 on KITTI [13] and NYUv2 [14] datasets to provide insights into their strengths and weaknesses under a common evaluation framework.

## II. Related Work

### A. Convolutional Neural Network (CNN) Based Models

**Convolutional Neural Networks (CNNs)** have played a key role in improving monocular depth estimation by effectively capturing and analyzing local spatial features from images. Early CNN-based models introduced encoder-decoder architectures as a foundation for depth estimation. In this setup, the encoder extracts high-level semantic information from the input image, while the decoder generates dense depth maps based on these features.

**Monodepth2** [11] introduced a self-supervised training method that can work with both monocular and stereo video data. The model incorporates innovative techniques such as a novel matching loss, auto-masking, and a multi-scale approach, enabling it to achieve state-of-the-art performance on the KITTI dataset [13] with various models. It uses a U-Net architecture with ResNet18 [24] as the encoder, which has 11 million parameters. This design is smaller and faster compared to earlier models [6].

**DenseDepth** [12] introduced a transfer learning-based approach to monocular depth estimation, leveraging pre-trained CNNs to enhance learning efficiency and accuracy. The model employs a straightforward encoder-decoder architecture with skip connections, resembling a U-Net design. The encoder is based on DenseNet-169 [25], pre-trained on ImageNet [26], which provides a strong foundation for extracting rich semantic features. Proposing a tailored loss function and data augmentation strategies that enable faster training and improved performance, the model demonstrates competitive results on both the NYUv2 [14] and KITTI [13] datasets.

### B. Transformer-Based Models

Vision Transformers (ViT) [19] extend Transformers [18] to image data by dividing images into fixed-size patches and treating each patch as a token. This allows Transformers to capture global context and long-range dependencies, overcoming the limitations of CNN-based models.

**Dense Prediction Transformer (DPT)** [15] uses a ViT as its backbone in an encoder-decoder design created for dense prediction tasks like monocular depth estimation. The encoder divides the input image into small patches, turns them into tokens, and processes them with multi-headed self-attention (MHSA) layers, allowing it to capture global information at every stage. DPT comes in three versions: ViT-Base with 12 transformer layers, ViT-Large with 24 layers, and ViT-Hybrid, which uses ResNet50 [24] for initial feature extraction followed by 12 transformer layers. The decoder produces dense predictions through a three-step process, merging features at different resolutions to balance detailed and global information. DPT can handle images of different sizes by adjusting its position embeddings, making it flexible for various inputs. This combination of global context from Transformers and detailed predictions from the decoder makes DPT suitable for complex depth estimation tasks.

### C. Hybrid CNN-Transformer Models

To harness the strengths of both CNNs and Transformers, hybrid architectures have been developed that integrate convolutional feature extraction with Transformer-based global context modeling.

**AdaBins** [16] introduces an adaptive binning method for monocular depth estimation, improving on fixed-bin approaches by dynamically adjusting bin widths based on the input image features. The architecture has two main parts: a pre-trained EfficientNet B5 [27] encoder-decoder block for extracting high-resolution features and the AdaBins module, which uses a lightweight transformer (mini-ViT) to calculate adaptive bin widths and range-attention maps.

The adaptive binning strategy allows the depth range to be divided flexibly for each image, and final depth values are computed as a linear combination of bin centers. This approach produces smooth depth maps without the sharp transitions seen in discretized outputs. The range-attention maps combine global information from the transformer with local pixel-level features to refine the depth predictions. The model uses a scale-invariant pixel-wise loss and a bin-center density loss to align predicted bins with the ground truth depth distribution, resulting in smooth and accurate depth predictions. This design effectively integrates classification and regression methods to address limitations of earlier approaches.

**Depth Anything v2** [17] builds upon the DPT framework [15] by integrating a hybrid CNN-Transformer architecture for monocular depth estimation. The model employs a ConvNeXt backbone [28] as its encoder to extract local spatial features efficiently, while a lightweight transformer decoder captures global depth relationships through multi-head self-attention (MHSA).

Unlike DPT, which relies solely on supervised learning, Depth Anything V2 leverages pseudo-label supervision and a relative depth loss, enabling it to train on large-scale synthetic and unlabeled datasets. This strategy allows the model to generalize effectively across datasets and perform well in zero-shot scenarios. By combining the local feature extraction capabilities of CNNs with the global reasoning power of Transformers, Depth Anything V2 achieves state-of-the-art performance on the DA-2K benchmark [17], surpassing other current state-of-the-art models such as Marigold [29] and DepthFM [30] in both accuracy and latency.

TABLE I
ARCHITECTURE AND LEARNING PARADIGM COMPARISON OF DEPTH ESTIMATION MODELS

| Model | Architecture | Learning Paradigm |
|---|---|---|
| Monodepth2 [11] | CNNs | Self-supervised learning |
| DenseDepth [12] | CNNs | Supervised learning |
| DPT [15] | Vision Transformers (ViT) | Supervised learning |
| AdaBins [16] | Hybrid model with CNNs and Transformers | Supervised learning |
| Depth Anything v2 [17] | Hybrid model with CNNs and Transformers | Psuedo-label supervised learning |

### *D. Comparative Studies and Evaluation Frameworks*

Benchmarks for monocular depth estimation are limited, primarily due to the challenges of obtaining accurate ground truth depth data. However, two widely recognized datasets in the field are **KITTI** [13] and **NYU Depth V2** [14]. KITTI is commonly used for autonomous driving research, presenting outdoor depth estimation challenges, while NYU Depth V2 focuses on indoor environments with diverse scenes and varying depth complexities. These datasets provide depth mask to enable the evaluation of model accuracy using established metrics below, providing a reliable basis for comparison.

- **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and ground truth depths.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|d_i^{\text{pred}} - d_i^{\text{gt}}\right| \tag{1}$$

- **Root Mean Squared Error (RMSE):** Evaluates the square root of the average squared differences, emphasizing larger errors.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(d_i^{\text{pred}} - d_i^{\text{gt}}\right)^2} \tag{2}$$

- **Threshold Accuracy ($\delta$):** Determines the percentage of predictions within a certain factor of the ground truth depth.

$$\delta = \frac{\text{Number of } d_i^{\text{pred}} \text{ satisfying } \frac{1}{\tau}d_i^{\text{gt}} \le d_i^{\text{pred}} \le \tau d_i^{\text{gt}}}{N} \tag{3}$$

  where $\tau$ is typically set to 1.25, $1.25^2$, and $1.25^3$.

However, each monocular depth estimation model predicts depth in different formats, depending on its pre-training method and the datasets used. For instance, recent models like Depth Anything V2 [17] predict relative depth due to their pseudo-label-based training. Additionally, the NYUv2 and KITTI datasets use different depth units in their ground truth labels. As a result, the predicted depth outputs must be appropriately scaled for each benchmark, or a fine-tuning process is required to achieve optimal performance across the evaluation metrics.

## III. METHOD

This section describes the process of model selection, dataset preparation, and model fine-tuning prior to evaluation, ensuring a fair and objective comparison.

### *A. Model Selection and Implementation*

To provide a comprehensive comparison, we selected five state-of-the-art monocular depth estimation models that represent different architectural paradigms, as seen on Table I. The models were selected on a wide varieties to cover all different architecture and learning paradigm. All models were implemented using the PyTorch framework, leveraging pre-trained weights where applicable to ensure consistency and to expedite the training process.

### *B. Dataset Preparation*

We used two well-known benchmark datasets, KITTI [13] and NYU Depth V2 [14], for our experiments. To ensure consistency and optimize model performance, we applied uniform preprocessing to both datasets, with the exception of image size due to differences in the original resolutions of the datasets. NYU Depth V2 images and depth maps were resized to 480×360 pixels, while KITTI images were resized to 640×192 pixels to enable faster training and reduce computational requirements. All RGB images from both datasets were normalized using the mean and standard deviation of the ImageNet dataset [26], aligning with the pre-trained parameters of the models to support stable and efficient training. Depth values were normalized to a range of [0, 1] by dividing by the maximum depth value in each map, ensuring numerical stability and compatibility with the models' output activations.

For both datasets, we adopted a standardized split of 80% for training, 10% for validation, and 10% for testing. This approach ensured exposure to diverse scenarios during training while reserving sufficient data for unbiased evaluation. The splitting strategy aligns with the practices recommended by the dataset repositories, enabling direct comparisons with existing studies and ensuring that models are tested on entirely unseen data to effectively assess their generalization capabilities.

### *C. Fine-tuning Procedures*

The fine-tuning procedures for each monocular depth estimation model were carefully standardized to ensure a fair and objective comparison. The key components of the training process, including hyperparameters, optimization strategies, and loss functions, were adapted to each model's architecture while maintaining consistency across evaluations.

- **Batch Size:** Set to 8 for both KITTI and NYUv2 datasets, optimizing the use of GPU memory given the image size differences between the datasets.

- **Optimizer:** Adam optimizer with an initial learning rate of $1 \times 10^{-4}$, chosen for its ability to handle sparse gradients and adapt effectively to various architectures.
- **Learning Rate Scheduler:** StepLR with a decay factor of 0.1 every 20 epochs to facilitate gradual convergence.
- **Number of Epochs:** - KITTI: 50 epochs, as the dataset is relatively large and provides sufficient diversity for quicker convergence. - NYUv2: 100 epochs, due to its smaller size and higher depth complexity, requiring additional iterations to fully learn fine-grained details.
- **Loss Functions:** A hybrid loss combining L1 depth regression and the Structural Similarity Index (SSIM) to account for both pixel-level accuracy and structural details:

$$\text{Loss} = 1.0 \cdot \text{SSIM} + 0.1 \cdot \text{L1}.$$

These parameters reflect the unique requirements of each model and the characteristics of the KITTI and NYUv2 datasets, ensuring optimal performance and a consistent evaluation framework.

### D. Evaluation Protocol

Quantitative evaluations were conducted using a suite of standard evaluation metrics commonly employed in monocular depth estimation research. These metrics included Mean Absolute Error (MAE), which measures the average absolute difference between predicted and ground truth depth values; Root Mean Squared Error (RMSE), which emphasizes larger errors by computing the square root of the average squared differences; and Threshold Accuracy ($\tau$ is set to 1.25), which determines the percentage of predictions within a certain factor of the ground truth depth.

## IV. Results and Discussion

### A. Fine-Tuning

All models were fine-tuned on the KITTI [13] and NYUv2 [14] datasets. The training and validation loss curves for Monodepth2 are shown in Figure 2 and Figure 3. These plots show steady convergence, with validation loss slightly higher than training loss in later epochs, indicating minor overfitting. This is typical for CNN-based models, which adapt well to training data but face challenges with generalization.

DenseDepth also showed strong convergence, particularly on NYUv2, but exhibited slightly more overfitting than Monodepth2 due to its deeper architecture. In contrast, transformer-based models like DPT showed smoother convergence and smaller validation loss gaps, demonstrating better generalization. Hybrid models, such as AdaBins and Depth Anything V2, performed robustly, with minimal overfitting.

### B. Model Evaluation

We evaluate all models using three standard metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Threshold Accuracy ($\delta$ at 1.25). Table II summarizes the performance of each model on KITTI and NYUv2 datasets.

CNN-based models, such as Monodepth2 and DenseDepth, perform well but are outperformed by transformer-based and hybrid models like DPT and AdaBins across all evaluation metrics. Surprisingly, Depth Anything V2 achieves notably poor quantitative scores, despite its qualitative results outperforming the other models.

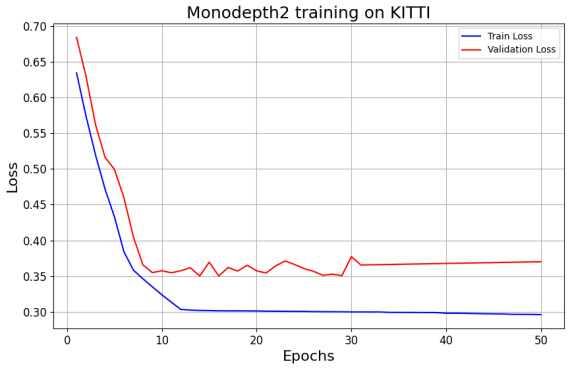


Fig. 2. Training and validation loss curves during fine-tuning for KITTI and NYUv2 datasets.

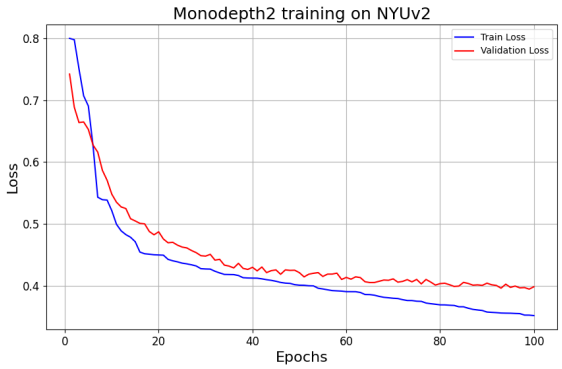


Fig. 3. Training and validation loss curves during fine-tuning for KITTI and NYUv2 datasets.

Figure 1 presents depth predictions generated by the different models. Transformer-based models (DPT and AdaBins) produce smoother depth maps with fewer artifacts and enhanced fine-grained details, particularly in NYUv2 indoor scenes. Notably, the Depth Anything V2 model generates the most visually accurate depth maps, excelling in capturing fine details, transparency, and reflective properties in complex scenes.

We further evaluated the performance and capabilities of the current state-of-the-art model, Depth Anything V2, by comparing its pre-trained model to those of Dense Prediction Transformer (DPT) and AdaBins on a 2D image painting, as shown in Figure 4. Depth Anything V2 demonstrates strong generalization capabilities, producing a smooth and consistent depth map from the painting. DPT also performs well, though it captures fewer fine details. In contrast, AdaBins performs poorly on this task, which reflects the need of fine-tuning of the model. These results suggest that Depth Anything V2 inherits its generalization strengths primarily from its DPT

TABLE II
EVALUATION RESULTS ON KITTI AND NYUV2 DATASETS

| **Model (KITTI)** | **MAE** | **RMSE** | $\delta$ **(1.25)** |
|---|---|---|---|
| Monodepth2 [11] | 0.612 | 2.449 | 69.3% |
| DenseDepth [12] | 0.735 | 2.649 | 65.5% |
| DPT [15] | **0.183** | **0.755** | 71.3% |
| AdaBins [16] | 0.198 | 0.887 | **73.2%** |
| Depth Anything V2 [17] | 2.214 | 2.626 | 12.9% |

| **Model (NYUv2)** | **MAE** | **RMSE** | $\delta$ **(1.25)** |
|---|---|---|---|
| Monodepth2 [11] | 0.415 | 0.946 | 72.1% |
| DenseDepth [12] | 0.394 | 0.715 | 69.1% |
| DPT [15] | 0.171 | 0.720 | 72.4% |
| AdaBins [16] | **0.161** | **0.699** | **74.5%** |
| Depth Anything V2 [17] | 1.815 | 2.179 | 15.4% |

backbone while enhancing fine details through its hybrid CNN architecture and training on a large-scale pseudo-labeled dataset.

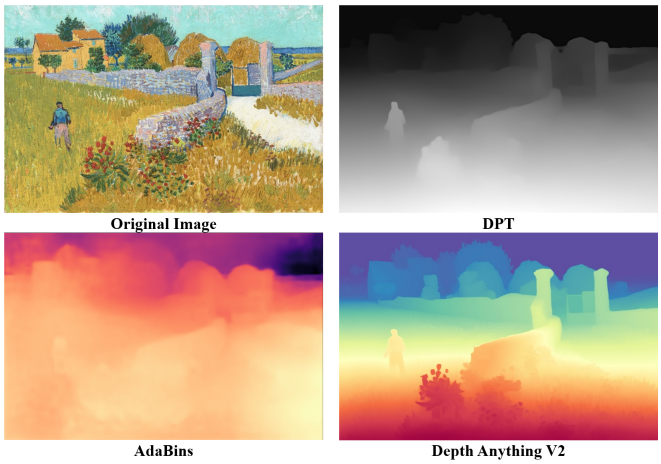


Fig. 4. Depth predictions on a 2D image painting. Depth Anything V2 produces smooth, detailed depth maps, while DPT shows fewer fine details, and AdaBins performs poorly.

### C. Discussion

Transformer-based models, like DPT [15], and hybrid models, such as AdaBins [16] and Depth Anything V2 [17], perform better than traditional CNN-based models. This improvement comes from their ability to combine global depth information with detailed local features of transformers. Hybrid models, in particular, take the advantages of both CNNs and transformers to achieve better results for global and local depth prediction.

Depth Anything V2 performs well on both datasets and zero-shot tasks, showing strong generalization without requiring much fine-tuning. DPT, which uses only transformers, also generalizes well but captures fewer fine details. AdaBins, a hybrid model with adaptive binning, requires fine-tuning for good results. Depth Anything V2 improves on DPT by adding CNNs, which help capture finer details and lighting variations. However, its quantitative performance on KITTI and NYUv2 benchmarks is not as strong. This is because the model is difficult to fine-tune for specific tasks due to its large and complex nature. Additionally, Depth Anything V2 predicts relative depth at the pixel level, which sometimes creates inconsistencies when scaled to real-world metric depths like those in NYUv2 and KITTI. This issue reflects poorly in metrics like MAE, RMSE, and threshold accuracy.

Despite its lower scores on NYUv2 and KITTI, Depth Anything V2 outperforms other models qualitatively. This result raises concerns about the quality of these datasets. Both datasets have outdated low-quality images, small sizes, inconsistent pixel-level depth labels, and lack of diversity. For example, Depth Anything V2 may predict accurate depth for a pixel, but the datasets might mark that pixel as a void or incorrect depth due to measurement limitations (e.g., stereo cameras). These issues increase overall errors and make Depth Anything V2 seem less accurate.

This highlights an important trend: modern models like Depth Anything V2 are now surpassing the limitations of older benchmarks. To address this, a new benchmark called DA-2K [17] was introduced along with Depth Anything V2. This new benchmark offers better evaluation and points toward a new direction for monocular depth estimation research.

## V. CONCLUSION

In this paper, we presented a comparative analysis of five state-of-the-art monocular depth estimation models, including Monodepth2, DenseDepth, DPT, AdaBins, and Depth Anything V2. The models were evaluated on two widely used benchmarks: KITTI (outdoor) and NYUv2 (indoor).

Our results show that Transformer-based and hybrid models outperform traditional CNN-based models. DPT and AdaBins demonstrated strong quantitative performance, while Depth Anything V2 produced the most visually accurate depth maps, especially in challenging scenarios. Depth Anything V2's hybrid design and pseudo-label supervision enable it to generalize well across datasets and perform effectively in zero-shot tasks. However, its relative depth predictions and scaling issues result in lower quantitative scores when evaluated against older datasets like KITTI and NYUv2.

These findings suggest that while existing benchmarks remain valuable, their limitations in label accuracy and depth consistency are becoming clear as models like Depth Anything V2 set new qualitative standards. Moving forward, there is a need for updated benchmarks, such as DA-2K, to better reflect the capabilities of modern monocular depth estimation models and guide future research in this field.

## REFERENCES

[1] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" 2017. [Online]. Available: https://arxiv.org/abs/1703.04977

[2] M. B. Çöteli, O. Olgun, and H. Hacıhabiboğlu, "Multiple sound source localization with steered response power density and hierarchical grid refinement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2215–2229, 2018.

[3] A. Makandar and A. Patrot, "An approach to analysis of malware using supervised learning classification," in *International Conference on Recent Trends in Engineering, Science Technology - (ICRTEST 2016)*, 2016, pp. 1–5.

[4] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," 2016. [Online]. Available: https://arxiv.org/abs/1606.00373

[5] K. Y. Kok and P. Rajendran, "A review on stereo vision algorithm: Challenges and solutions," 11 2019.

[6] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," 2017. [Online]. Available: https://arxiv.org/abs/1609.03677

[7] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," *ACM Trans. Graph.*, vol. 25, no. 3, p. 835–846, Jul. 2006. [Online]. Available: https://doi.org/10.1145/1141911.1141964

[8] B. Kim and P. Burger, "Depth and shape from shading using the photometric stereo method," *CVGIP: Image Understanding*, vol. 54, no. 3, pp. 416–427, 1991. [Online]. Available: https://www.sciencedirect.com/science/article/pii/104996609190040V

[9] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, "Learning like a child: Fast novel visual concept learning from sentence descriptions of images," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2533–2541.

[10] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," *Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2018, pp. 2002–2011, 06 2018.

[11] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," 2019. [Online]. Available: https://arxiv.org/abs/1806.01260

[12] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," 2019. [Online]. Available: https://arxiv.org/abs/1812.11941

[13] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, Aug. 2013. [Online]. Available: https://doi.org/10.1177%2F0278364913491297

[14] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[15] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," 2021. [Online]. Available: https://arxiv.org/abs/2103.13413

[16] S. Farooq Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2021, p. 4008–4017. [Online]. Available: http://dx.doi.org/10.1109/CVPR46437.2021.00400

[17] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv:2406.09414*, 2024.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[20] A. E. Brouwer, "The equivalence of two inequalities for quasisymmetric designs," 2022. [Online]. Available: https://arxiv.org/abs/2203.08910

[21] A. Agarwal and C. Arora, "Depthformer : Multiscale vision transformer for monocular depth estimation with local global information fusion," 2022. [Online]. Available: https://arxiv.org/abs/2207.04535

[22] J. Bae, S. Moon, and S. Im, "Deep digging into the generalization of self-supervised monocular depth estimation," 2023. [Online]. Available: https://arxiv.org/abs/2205.11083

[23] A. Luginov and I. Makarov, "Swiftdepth: An efficient hybrid cnn-transformer model for self-supervised monocular depth estimation on mobile devices," in *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2023, pp. 642–647.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[25] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018. [Online]. Available: https://arxiv.org/abs/1608.06993

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[27] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[28] Z. Liu, H. Lin, Y. Cao, Z. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 976–11 986.

[29] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[30] U. P. P. M. D. K. O. G. S. A. B. V. T. H. B. O. Ming Gui, Johannes S. Fischer, "Depthfm: Fast monocular depth estimation with flow matching," 2024.