Cyclops! A Survey of Monocular Geometry

Jose Cuaran and Al Smith
University of Illinois at Urbana-Champaign
(jrc9, ad10)@illinois.edu

Abstract

Monocular geometry is a very well-researched and nearly solved field in 3D computer vision. The concept is simple: what 3D understanding can we obtain from a single-camera image? Many methods have explored this concept as far back as the 1970's, and recently, advances in machine learning have begun to show great promise in resolving substantial detail in depth information from single-view images. In this work, we aim to provide a scoping review on works focused on monocular geometry and depth estimation, briefly summarizing explorations in classical works, and focusing primarily on two key areas that have been extensively explored recently: discriminative models and generative models for 3D depth modeling from single-view images.

1. Introduction

Monocular geometry is the process of predicting the geometry of a scene from a single image, a task that is crucial for understanding the three-dimensional structure of the environment. It has a wide range of applications including autonomous driving, robotics, 3D reconstruction and augmented reality, where understanding spatial relationships is essential for navigation, interaction, and scene interpretation. By enabling geometry perception from a single image, monocular geometry reduces the need for complex hardware setups, such stereo cameras or depth sensors, making it more accessible and cost-effective.

Unlike stereo vision, which utilizes two or more images to infer depth through disparity, monocular geometry relies only on the information contained within a single frame. Therefore, this is inherently an ill-posed problem, as many different 3D structures can project to the same 2D image. Factors such as occlusions, varying lighting conditions, and texture gradients further complicate the task.

Several approaches have been developed to tackle the challenges of monocular geometry estimation. Traditional methods often relied on handcrafted features and geometric constraints, such as texture gradients and shading. However, recent advancements in deep learning have revolutionized the field, leading to the emergence of data-driven techniques that leverage convolutional neural networks (CNNs), transformers and large-scale datasets. In this survey we present a compilation of some of the main approaches for monocular geometry estimation, including classical methods and deep learning-based methods. We additionally present a summary of metrics for evaluation, the performance of some methods and the limitations and open research questions in this field.

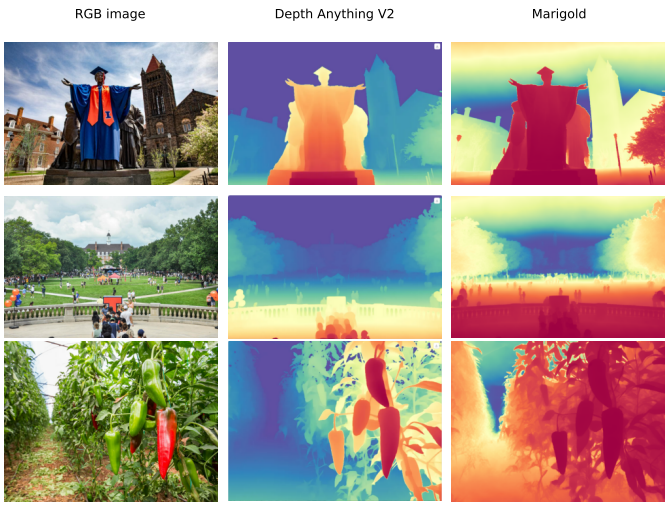

Figure 1. Depth estimated by two state-of-the-art methods: Depth Anything V2 [22](a discriminative model) and Marigold [8](a generative model). Note that Depth Anything V2 capture fine-grained details from the objects and predicts more consistent depth values (e.g. the sky).

2. Taxonomy

2.1. Classical methods

Extracting 3D information from singular cameras has long been a challenge in classical computer vision and are highly dependent on camera sensor hardware, imaging characteristics, and lenses. Berthold Horn wrote the seminal work in this field, focused on the challenge of identifying the shape

of a smooth opaque object from a single view by utilizing knowledge about the surface, position of the lightsource, and environment features [7]. Dozens of paradigms since have been explored and proposed to solve monocular geometry challenges, some of which include the seminal work in this field in shape-from-shading (SFS) [14], photometric surface orientation from albedo [20], simple image segmentation [6], and probabalistic approaches to pixel-wise depth [15], In each of these methods, we rely on physical attributes of the components that constrain the environment and measured objects by reflectance properties, camera properties, color, orientation, and depth.

2.2. Deep Learning-based methods

Deep learning-based methods for monocular geometry estimation have gained significant attention in the last decade due to their ability to capture semantic and spatial information from single images, without the need for handcrafted features. Leveraging convolutional neural networks (CNNs) or vision transformers, these methods learn complex mappings from images to depth maps through large-scale datasets. Additionally, advancements in self-supervised training techniques have reduced the dependency on costly ground truth depth data, allowing models to be trained on monocular videos or stereo pairs.

Deep learning-based methods for monocular geometry estimation can be classified into two main categories: discriminative methods and generative methods. Discriminative methods focus on learning a direct mapping from input data (a single image) to a desired output (e.g., a depth map). Generative methods, in contrast, focus on modeling the underlying distribution of the data, allowing to sample multiple solutions for a given input image. Table 1 shows a summary of deep learning-based methods that we describe below.

2.2.1 Discriminative methods

These methods can be classified into supervised methods that require labeled data, self-supervised methods that leverage stereo images or monocular video sequences to estimate geometric information, and semi-supervised methods that leverage both unlabeled and labeled data.

Supervised methods: An early work on supervised models for monocular depth estimation is introduced by Liu et al.[10], who integrate Conditional Random Fields (CRF) with deep learning. While CRFs effectively capture spatial relationships between superpixels in an image, the use of deep convolutional neural networks (CNNs) eliminates the need for hand-crafted features. During training, the negative conditional log-likelihood of the data is minimized, incorporating a regression term to ensure depth accuracy and a smoothness term to promote consistent and smooth depth maps. For inference, they employ Maximum a Posteriori (MAP) estimation, which is performed in closed form. This approach combined with the use of fully convolutional networks and a superpixel pooling strategy outperformed previous works in accuracy and runtime at the time of this paper.

Laina et al. [9] propose an end-to-end architecture for monocular depth estimation that is both less complex and more efficient than previous approaches. Their method leverages fully convolutional networks to significantly reduce the number of parameters while enabling dense prediction. In addition, residual blocks are used to train deeper layers while increasing the receptive field, capturing global context for depth estimation. Additionally, the authors introduce up-convolution blocks, which facilitate the prediction of high-resolution depth maps. Trained in a supervised manner using a reverse Huber loss, this approach achieves real-time inference and outperforms prior methods in terms of depth accuracy.

Typical learning-based depth estimation methods predict affine-invariant depth, meaning depth is estimated up to an unknown scale and shift from a single monocular image. To achieve consistent 3D reconstruction, these parameters must be accurately estimated afterward. To address this challenge, Yin et al. [23] propose a network that, given a distorted point cloud, leverages learned geometric priors to estimate both the focal length and the depth shift, resulting in realistic reconstructions. While this approach performs well in structured environments, its performance degrades in scenes lacking geometric cues.

Self-supervised methods: As labeled data is costly and not always reliable (e.g. sensor depth data is noisy and incomplete), several approaches have emerged for depth estimation without supervision. Overall, they leverage geometric constraints imposed by binocular stereo images or monocular video sequences to predict disparity or depth maps. This problem is usually formulated as a reconstruction problem, where source color images are predicted based on the relative pose and predicted depth maps, and a reconstruction loss is minimized between the ground truth source images and the generated images.

For instance, Garg et al. [2] propose a simple approach to train a convolutional neural network (CNN) for depth estimation without supervision. They utilize pairs of images with known camera motion (e.g., stereo images or images from SLAM systems). Specifically, a fully convolutional neural network takes the left image as input and predicts a disparity map, which is then used to warp the right image onto the left image plane. The model is trained using a reconstruction loss (photometric error) in conjunction with a disparity smoothness term. This method achieves results competitive with those of supervised approaches. However, its ability to generalize to other cameras is limited due to its

Method	Year	Learning approach	Architecture	Number of parameters	Training dataset (Number of samples)
DCNF-FCSP [10]	2015	Discriminative, Supervised	Fully convolutional	5.8M	NYU v2 dataset (1.5k), Make3D (534)
[9]	2016	Discriminative, Supervised	Fully convolutional	62M	NYU Depth v2 (12k)
[23]	2020	Discriminative, Supervised	ResNet50 + CNN decoder	¯25M	Taskonomy (114k), 3D Ken Burns (51k) DIML (121k), Holopix50K (48k), HRWSI (20k)
Depth Anything [21, 22]	2024	Discriminative, Semi-supervised	Transformer	Teacher (1.3B), Student (25M)	63.5M samples
[2]	2016	Discriminative, Self-supervised	Fully convolutional	-	KITTI dataset (23.5k)
[4]	2017	Discriminative, Self-supervised	Fully convolutional	31M	KITTI dataset (30.2k)
[24]	2018	Discriminative, Self-supervised	ResNet50-1by2 + CNN decoder	¯7M	KITTI dataset (23.5k)
Monodepth [5]	2019	Discriminative, Self-supervised	Fully convolutional U-Net	¯15M	KITTI dataset (23.5k)
HR-Depth[11]	2020	Discriminative, Self-supervised	U-Net	HR-Depth (14.6M), Lite-HR-Depth (3.1M)	KITTI dataset (¯40k)
MonoVit [25]	2022	Discriminative, Self-supervised	Transformer + CNN	-	KITTI dataset (¯40k)
Marigold [8]	2024	Generative-based, Diffusion model	U-Net	¯1.5B	Hypersim (54k), Virtual KITTI (20k)
DDVM [17]	2023	Generative-based, Diffusion model	Efficient U-Net	¯300M	ScanNet (2.5M), SceneNet RGBD (5M) Waymo Open dataset (200k)

Table 1. Comparison of different deep learning-based depth estimation models

strong dependence on the camera baseline and focal length.

Godard et al. [4] developed a fully convolutional network for monocular depth estimation without the need for ground truth depth data during training. To achieve this, they utilize binocular stereo images, which inherently depend on the depth of the scene. The model takes the left image as input and is designed to predict both the left and right disparity maps, as well as the corresponding color images. The proposed loss function comprises (i) a reconstruction term based on the photometric error between the ground truth color images and the generated images, (ii) a disparity smoothness term to ensure locally consistent depth values, and (iii) a left-right disparity consistency loss to enhance depth prediction accuracy. During inference, the depth map is recovered using the known baseline and focal length of the camera. This approach yields accurate and smooth depth estimates for monocular images captured with the same camera used for data collection. However, its accuracy diminishes when tested on other datasets, though it still produces visually plausible depth maps.

Similar to previous works, Zhan et al. [24] utilize stereo images for single-view depth estimation. However, unlike earlier methods, their model predicts not only disparity maps but also the relative camera pose (visual odometry). This is accomplished by leveraging both the spatial constraints from the left and right images and the temporal constraints from consecutive stereo pairs. They train two convolutional networks: one for depth prediction and the other for motion estimation. These models are jointly trained using an image reconstruction loss and a depth smoothness loss, similar to prior approaches. Additionally, to address issues such as specular reflections and photometric limitations, they introduce a feature reconstruction loss to enhance robustness and accuracy. This approach performs comparably to supervised methods for depth estimation and achieves results close to visual odometry and SLAM methods for camera motion estimation.

Godard et al. [5] introduced further advancements in self-supervised depth estimation with their Monodepth2 model, which leverages stereo or monocular image sequences to predict depth maps and generate warped images for supervision. A key contribution of this work is a novel approach to handle occluded regions between source and generated images, as these regions can significantly impact the photometric reconstruction loss. Additionally, they propose an auto-masking loss to mitigate the effects of a stationary camera, addressing the assumption of camera motion inherent in these methods. In the case of monocular self-supervision, they jointly train a network for pose estimation. Monodepth2 demonstrates substantial improvements over previous methods, particularly when using a combination of both stereo and monocular data for training.

Previous works have struggled to capture fine details, even when using high-resolution input images. To address this, Lyu et al. [11] introduce HR-Depth, a self-supervised model for depth estimation that predicts high-resolution depth maps. The authors adopt a U-Net architecture and employ dense skip connections to provide the decoder with enhanced semantic and spatial information. This design, combined with an efficient feature fusion strategy, significantly improves the quality and accuracy of the predicted depth maps. Furthermore, the authors develop a lightweight model trained using a teacher-student distillation approach. This model is five times smaller than Monodepth in terms of parameters while maintaining comparable performance.

Previous works primarily rely on convolutional neural networks (CNNs) for depth estimation. However, the limited receptive fields of these networks constrain their ability to capture global relationships within the image, which are crucial for improving depth estimation. To address this issue, Zhao et al. [25] propose Monovit, a vision transformer model for monocular depth estimation. Specifically, they combine convolutional layers with transformer blocks to capture both local and global context for disparity map prediction. The model is trained using the conventional image reconstruction approach, employing a loss function that includes both a reconstruction term and a smoothness term. The training process is self-supervised, utilizing monocular video sequences, with an additional network jointly trained for pose estimation. This approach effectively captures fine-grained details and demonstrates superior generalization to unseen datasets compared to previous methods.

Semi-supervised methods While labeled data provides accurate depth for supervision, it is limited in both quantity and diversity of domains. To address this limitation, semi-supervised methods aim to leverage existing labeled datasets along with large-scale and multi-domain unlabeled data. For instance, motivated by the strong performance of foundation models trained on large-scale datasets, Yang et al. [21] developed Depth Anything V1, a foundation model for monocular depth estimation. Addressing the challenge of limited labeled datasets for this task, the authors propose an approach that leverages both labeled and unlabeled data. They first train a teacher model using 1.5 million labeled samples, which is then employed to generate pseudo-labels for 62 million unlabeled images, thus creating a large-scale dataset. Subsequently, a student model is trained using both the labeled and pseudo-labeled datasets. To enhance depth estimation, strong perturbations are applied to the unlabeled images during training, encouraging the student model to extract additional visual cues. The architecture employs DinoV2 as the encoder and a vision transformer as the decoder. This combination of strategies, along with a carefully designed loss function that accounts for feature alignment, enables the model to surpass state-of-the-art performance, demonstrating improved accuracy and generalization.

Yang et al. [22] report further improvements on Depth Anything by training the teacher model exclusively on synthetic labeled data. This approach is motivated by the observation that real depth images tend to be noisy and incomplete, whereas synthetic images provide accurate ground truth, even for small objects, allowing the model to capture fine details in the scenes. Additionally, they train the student model on pseudo-labeled real images to enhance its generalization capabilities by mitigating the distribution shift between synthetic and real data.

2.2.2 Generative methods

Generative methods are used to identify the underlying distribution of depth, which can be used to generate multiple possible depth maps from a given input. The primary approach to this involves diffusion models, of which significant amounts of work has been done in the past two years.

Diffusion models: Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation [8], also known as Marigold, explores how pre-trained diffusion models, originally designed for generating realistic images, can be repurposed for the task of monocular depth estimation. This work builds on the intuition that diffusion models learn visual representations that can capture spatial and structural relationships in the input images. The authors propose a method to fine-tune pre-trained image generators using a task-specific objective for depth estimation. By adapting the learned diffusion process to generate depth maps instead of RGB images, the model achieves competitive results without the need for extensive depth-specific data. This approach demonstrates how the generalization capabilities of diffusion-based generators can be harnessed for 3D reconstruction tasks such as depth estimation.

The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation [17] investigates the use of diffusion models for both optical flow and monocular depth estimation, demonstrating their potential in addressing multiple vision tasks with a unified framework. The authors propose a novel conditioning mechanism that enables diffusion models to handle depth estimation tasks by progressively refining depth predictions over several denoising steps. The primary insight is that diffusion models are great for capturing uncertainty in the prediction space, which is valuable for depth estimation. Through their experiments, the authors show that diffusion-based methods outperform several existing approaches, highlighting the adaptability of this approach to various tasks beyond their original domain of image generation.

ECoDepth: Effective Conditioning of Diffusion Models for Monocular Depth Estimation [13] introduces a novel approach to condition diffusion models for monocular depth estimation by incorporating task-specific priors and constraints. The method, ECoDepth, focuses on optimizing the conditioning process of the diffusion model, allowing it to more effectively generate accurate depth maps. A key contribution of this work is the use of a multi-scale conditioning strategy, which enables the model to capture both global scene context and fine-grained details. By carefully integrating geometric priors, such as depth smoothness and boundary alignment, into the diffusion process, ECoDepth achieves state-of-the-art results, significantly improving depth prediction quality over previous methods. Additionally, this approach demonstrates the flexibility of diffusion models when conditioned with appropriate domain-specific knowledge.

DiffusionDepth: Diffusion Denoising Approach for Monocular Depth Estimation [1] leverages the power of diffusion denoising models and transfer learning for predicting depth maps from monocular images. Taking inspiration from approaches that use CLIP to improve zero-shot transfer in other domains, DiffusionDepth formulates depth estimation as a conditional image generation task where the diffusion model gradually refines the predicted depth map through multiple noise-removal steps. Using a frozen ViT model contained within DiffusionDepth's CIDE module, the overall pipeline can be further conditioned to more accurately generate depth maps for a given input image. DiffusionDepth also emphasizes the robustness of diffusion models in handling various lighting and environmental conditions, making it a versatile solution for monocular depth estimation.

Fine-Tuning Image-Conditional Diffusion Models is Easier than You Think [12] is a very recent paper that presents a simplified approach for tuning pre-trained diffusion models for monocular depth estimation tasks, and extends this to geometry estimation as well. This work demonstrates that by carefully adjusting a small number of parameters in an already trained diffusion model, high-quality depth maps can be generated with low overhead. The fine-tuning process leverages pre-existing visual knowledge from large-scale image-conditioned diffusion models like Marigold, enabling them to be adapted for monocular depth estimation with minimal labeled depth data. The authors show that this fine-tuning approach leads to strong performance across various datasets and tasks, without requiring extensive retraining from scratch. The simplicity and effectiveness of this method make it a practical choice for depth estimation in real-world applications, where labeled data might be scarce, also allowing users to infer geometry from detph maps.

These works show the advances and potential of diffusion models as generative frameworks for monocular depth estimation. Their ability to model complex distributions and handle noisy, multimodal predictions gives them an advantage in generating relative depth maps from 2D images, especially in scenarios where traditional methods might struggle.

3. Summary of evaluations

Metrics. There are different metrics to evaluate the performance of depth estimation models, including:

- Average relative error (AbsRel)

$$AbsRel = \frac{1}{T}\sum_{i \in T}\frac{|d_i^{gt} - d_i|}{d_i^{gt}}$$

- Squared relative error

$$SqRel = \frac{1}{T}\sum_{i \in T}\frac{|d_i^{gt} - d_i|^2}{d_i^{gt}}$$

- Root mean squared error (RMS)

$$RMS = \sqrt{\frac{1}{T}\sum_{i \in T}(d_i^{gt} - d_i)^2}$$

- Accuracy with threshold: Percentage of pixels such that:

$$\max(\frac{d_i^{gt}}{d_i}, \frac{d_i}{d_i^{gt}}) = \delta < threlshold$$

Where d_i^{gt} is the ground-truth depth for pixel i, d_i is the estimated depth, and T is the total number of pixels in all the evaluated images. For affine-invariant depth estimation methods, these metrics are computed after aligning the depth predictions to the ground truth depth maps (e.g. in [8] the authors solve a least squares problem).

Datasets. Several datasets have been developed for depth estimation tasks. Among the most widely used is the KITTI dataset [3], which provides stereo images and LiDAR measurements from urban environments. Other notable datasets include the NYU-v2 dataset [19], offering RGB-D data from indoor environments; the Make3D dataset [16], which includes image and range data from outdoor scenes; and the ETH3D dataset [18], comprising stereo images and laser scans collected from both indoor and outdoor settings.

Table 2 shows the performance of some of the methods presented in this survey on the KITTI dataset [3]. We found that out of the discriminative methods, Depth Anything V2 (supervised) and MonoVit (self-supervised) exhibit the best performance. Both rely on the transformer architecture. Regarding the generative methods, DDVM outperforms other methods. Figure 1 presents the depth maps generated by Depth Anything V2 and Marigold (a generative method). While both methods produce smooth depth maps, Depth Anything V2 demonstrates superior detail capture and greater consistency in its estimations. This difference can likely be attributed to the difference in training data: Depth Anything V2 is trained on 65 million samples, whereas Marigold is finetuned on only 74,000 depth samples, highlighting the critical role of data volume in performance.

4. Discussion

Significant advances in monocular geometry estimation methods have been made over the last decade. State-of-the-art approaches are now capable of estimating high-resolution depth maps and surface normals in a zero-shot manner from a single image. These methods exhibit strong generalization across diverse domains, thanks to large-scale datasets, self-supervised techniques, and mature neural network architectures such as vision transformers. However, several limitations and open research questions remain, as discussed below.

Real-Time Processing Current high-performance models for depth estimation rely on neural networks with millions of parameters, requiring powerful hardware to run in real-time. This poses a limitation for applications in robotics or mobile devices with constrained computational resources. Developing efficient algorithms that can operate under such constraints while maintaining accuracy is a crucial area for future research.

Data Limitations A major challenge in monocular depth estimation is the lack of large-scale, diverse datasets with high-quality ground truth depth information. Most existing datasets focus on specific scene types, limiting the generalization of models trained on them. Additionally, real-world datasets often contain noise and incomplete informa-

Method	Error↓			Accuracy↑		
	AbsRel	SQ Relative	RMS	δ <1.25	δ <1.25^2	δ <1.25^3
DCNF-FCSP [10]	0.236	-	7.421	0.613	0.858	0.949
[23]	0.149	-	-	0.784	-	-
Depth Anything V1 [21]	0.076	-	-	0.947	-	-
Depth Anything V2 [22]	0.075	-	-	0.948	-	-
[2]	0.169	1.08	5.104	0.74	0.904	0.962
[4]	0.114	0.898	4.935	0.861	0.949	0.976
[24]	0.135	1.132	5.585	0.82	0.933	0.971
Monodepth2 [5]	0.106	0.806	4.63	0.876	0.958	0.98
HR-Depth [11]	0.101	0.716	4.395	0.899	0.966	0.983
MonoVit [25]	0.093	0.671	4.202	0.912	**0.969**	**0.985**
Marigold [8]	0.099	-	-	0.916	-	-
DDVM [17]	**0.055**	-	**2.613**	**0.965**	-	-

Table 2. Error and Accuracy Metrics for Different depth estimation methods on the KITTI dataset

tion, which negatively impacts the performance of learning-based models. While synthetic datasets have become more realistic and accessible, they remain limited to certain environments. Thus, the need for extensive training data encompassing a wide variety of scenes continues to be an open challenge.

Scale Ambiguity Monocular depth estimation inherently suffers from scale ambiguity, meaning that the estimated depth can be scaled without affecting the perceived structure of the scene. Although some methods claim to predict metric depth, this typically refers to affine-invariant depth, which is accurate only up to an unknown scale and depth shift. This presents difficulties for applications requiring precise absolute depth measurements, such as 3D reconstruction and navigation.

References

[1] Yiqun Duan, Xianda Guo, and Zheng Zhu. Diffusiondepth: Diffusion denoising approach for monocular depth estimation, 2023.

[2] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 740–756. Springer, 2016.

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[4] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[5] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019.

[6] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75:151–172, 2007.

[7] Berthold K.P. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. *PhD Thesis, Massachusetts Institute of Technology (MIT)*, 1970.

[8] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[9] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[10] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2015.

[11] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2294–2301, 2021.

[12] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. *arXiv preprint arXiv:2409.11355*, 2024.

[13] Suraj Patni, Aradhye Agarwal, and Chentan Arora. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2024.

[14] Zhang Ruo, Ping-Sing Tsai, James E. Cryer, and Mubarak Shah. Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999.

[15] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. Learning depth from single monocular images. *Neural Information Processing Systems*, 2005.

[16] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008.

[17] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017.

[19] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012.

[20] Robert J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19:139–144, 1980.

[21] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

[22] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.

[23] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021.

[24] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 340–349, 2018.

[25] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 international conference on 3D vision (3DV)*, pages 668–678. IEEE, 2022.