A Review of Methods for Estimating Depth Based on Various Application Scenarios

Siqi Huang[a]
Engineering College , Shantou University, Shantou, China

Keywords: Depth Estimation, 2D Image, Stereo Image, SPIdepth, HiMODE.

Abstract: In computer vision, depth estimation is a crucial task. The task is to measure the distance between each pixel and the camera. The accuracy and efficiency of estimation tasks have undergone a significant improvement due to the rise of deep learning. There are many distinct application situations for depth estimation, and the task can be broadly classified into two categories: stereo image depth estimation and 2D image depth estimation, depending on the needs of each scenario. In this paper, HR-Depth, HybridDepth, SPIdepth in 2D image depth estimation methods and UniFuse, NLFB, PanoFormer, OmniFusion, HiMODE in stereo image depth estimation methods are introduced and analysed. In addition, NYU-Depth V2, KITTI in 2D image dataset and Stanford3D, and Matterport3D in stereo image dataset are introduced in detail, and the effectiveness of these two kinds of approaches is examined and contrasted using the widely used evaluation indices. It is found that HybridDepth and SPIdepth perform well in 2D image depth estimation, and NLFB and HiMODE perform better in stereo image depth estimation. Future research in depth estimation may focus more on depth estimation studies of stereo images, because stereo images usually contain richer depth information, which can allow for higher accuracy in depth estimation.

1 INTRODUCTION

Depth estimate becomes a crucial element as computer vision advances from basic image identification to intricate scene comprehension. Using a mapping relationship from 2D picture cues to 3D depth, the objective is to extract depth cues from the image and calculate each pixel's distance from the camera. In order to achieve applications in different scenes, image datasets of different scene types such as NYU-Depth, KITTI, Stanford3D, Matterport3D, etc., have been constructed, which have driven the research on depth estimation. As a result, depth estimation has found extensive use in fields like autonomous driving, robotics, VR, AR, 3D reconstruction, and security monitoring.

As the application scenarios become wider and wider, the depth estimation task is gradually divided into two different types: 2D image depth estimation and stereo image depth estimation. Using 2D images without direct depth information, 2D image depth estimation seeks to determine how to create a mapping relationship between 2D image cues and 3D depth in order to estimate each pixel's distance from the camera; Whereas stereo image depth estimation is the study of how to extract effective depth cues from images containing 360-degree scene information in the presence of aberrations or distortions, etc., in order to obtain depth information.

Focusing on the above two different types of depth estimation research, this paper analyses and summarises their research progress and current status, collects different types of research data, and classifies and analyses them. Finally, it provides an outlook on the research development and application of depth estimation.

2 METHOD

2.1 Depth Estimation Methods For 2D Images

The Eigen team was the first to use CNN for 2D image depth estimation in 2014, and since then,

[a] https://orcid.org/0009-0009-1920-1180

Huang, S.
A Review of Methods for Estimating Depth Based on Various Application Scenarios.
DOI: 10.5220/0013681100004670
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, pages 204-211
ISBN: 978-989-758-765-8
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

numerous researchers have improved it in various ways based on this. Meanwhile, unsupervised deep learning methods have emerged, such as training the network by generating new viewpoint maps or using left and right views and video information to solve the depth estimation problem. Supervised means that there is a high dependence on training data with accurate depth labels, whereas unsupervised means that there is no need for depth-labelled data.

2020, HR - Depth was presented in a study by Xiaoyang Lyu et al. This approach, which focuses on optimizing the depth estimation effect through two methodologies, is an enhanced deep network for high-resolution self-supervised monocular depth estimation. On the one hand, the jump connection in DepthNet is redesigned to add intermediate nodes to the original encoder and decoder nodes to aggregate features, so that the decoder can obtain high-resolution features with richer semantic information and thus predict the depth map boundary more accurately; on the other hand, the feature fusion squeeze excitation (fSE) module is proposed, which efficiently fuses features through global average pooling, fully-connected layers and 1×1 convolution to efficiently fuse features and improve network performance while reducing the number of parameters. In addition, a lightweight network Lite - HR - Depth based on MobileNetV3 was constructed and knowledge distillation techniques were applied to further improve its accuracy. The benefits of this approach are substantial： when using Resnet-18 as the encoder, HR-Depth performs better than any prior state-of-the-art techniques with the fewest parameters at both high and low resolutions; Lite - HR - Depth contains only 3.1M parameters, yet achieves comparable or even better performance than Monodepth2 at high resolution; by redesigning the jump connections and introducing the fSE module, the accuracy of depth estimation in large gradient region is effectively improved, and sharper edges can be predicted; the high-resolution depth estimation is deeply analysed, which provides theoretical basis and practical references for subsequent studies.（Lyu、Liu、Wang，2020）

In 2024, HybridDepth was presented in a study by Ashkan Ganj et al. The method aims to address the problems of scale ambiguity, hardware heterogeneity and generalisability in depth estimation by fusing focused stack information and single-image prior for robust metric depth estimation. The process logic is as follows: first, the relative depth map and the metric depth map are generated using the single-image relative depth estimator and the DFF metric depth estimator, respectively; then, the relative depth map is converted into a global-scale depth map by global scale and displacement alignment, and a dense scale map is constructed; finally, the global-scale depth map is corrected using a deep-learning-based refinement layer combined with a scale map and an uncertainty map from the DFF module. depth map with pixel-level scale correction to obtain the final depth map. This method has significant advantages, surpassing existing methods on several datasets, such as on the DDFF12 and NYU Depth V2 datasets, with significant improvement in metrics such as RMSE compared to specific SOTA models; strong zero-sample generalisation capability, with excellent performance on the ARKitScenes and Mobile Depth datasets; a compact model structure, with an inference The model is compact, with inference time of only 20ms and size of 240MB, which is more suitable for mobile device deployment than other models while ensuring accuracy; it effectively solves the scale ambiguity problem of single-image depth estimation, and the depth estimation is more accurate and consistent under different scaling levels. （Ganj、Su、Guo，2024）

In 2024, Mykola Lavreniuk Research introduced SPIdepth for self-supervised monocular depth estimation. The primary flow logic of this technique, which aims to increase depth estimate accuracy by fortifying the pose network, is as follows: Use DepthNet to extract visual characteristics from a single RGB image, then use the encoder-decoder framework and the previously trained ConvNext to generate the depth map; In order to align the depth map and the reference image during view compositing, PoseNet is utilized to estimate the relative poses between the input and the reference image. The Self Query Layer is used to capture the geometric cues for depth estimation, compute the depth intervals and generate the final depth maps through probabilistic linear combinations; both DepthNet and PoseNet are optimised for training, and the interference regions are filtered by an automatic masking strategy, which is combined with various loss functions to improve the model performance. The benefits are significant, demonstrating excellent performance on datasets such as KITTI, Cityscapes and Make3D, with accuracy surpassing many other methods, strong generalisation ability, excellent performance in dealing with dynamic scenes and zero-sample evaluations, and high efficiency in inference using only a single image, with a lightweight model design that is easy to integrate into various types of systems. （Mykola Lavreniuk，2024）

2.2 Depth Estimation Methods For Stereo Image

In 2021, UniFuse was proposed by Hualie Jiang et al. for 360° stereo depth estimation. This method aims at fusing the features of equirect-angular projection (ERP) or cubemap projection (CMP) to enhance the depth estimation. The process is as follows: firstly, the input panoramic image is subjected to ERP and CMP respectively, and the CMP features are reprojected to the ERP mesh through C2E; the U-Net is used as the baseline network, and the processed CMP features are unidirectionally fused to the ERP features in the hopping connection of the decoding stage, which is specifically achieved through the designed CEE fusion module, which firstly performs the residual modulation on the CMP features to reduce their discontinuities, and then uses the SE block to adaptively adjust the importance of the channels; finally, the isometric columnar projection depth map is output from the fused features. This is achieved by the designed CEE fusion module, which firstly modulates the residuals of the CMP features to reduce their discontinuities, and then adaptively adjusts the importance of the channels by using the SE block; finally, the fused features output an isometric bar-projected depth map. This approach has several noteworthy benefits: State-of-the-art performance is attained on the four frequently used datasets, and the benefit is evident on the largest real dataset Matterport3D; the CEE fusion module is made to better fuse the two projected features and reduce the number of parameters; the proposed one-way fusion framework fuses the CMP features to the ERP branch only in the decoding stage, which is more efficient than the two-way fusion; the model complexity is low. Low complexity, UniFuse based on ResNet-18 has a small complexity increase but significant performance increase compared to BiFuse, and can maintain real-time and good performance on MobileNetV2; strong generalization ability, outperforms BiFuse when migrating between different datasets, and is able to generate reasonable depths in regions with no real depth. （Jiang、Sheng、Zhu，2021）

In 2021, NLFB was proposed by Ilwi Yun et al. to improve the 360° monocular depth estimation method. The method consists of three main parts: first, proposing a self-supervised learning method using only gravity-aligned videos, which reduces the dependence on depth data by mining the relationship between the depths of consecutive scenes, and constructing image, depth, and pose consistency loss to train the model; secondly, using a jointly supervised and self-supervised learning approach that uses supervised learning to compensate for the shortcomings of self-supervised learning in areas that reflect light, for instance, and self-supervised learning to improve the supervised learning features and increase the model's ability to adapt to data that is not visible; thirdly, the network can preserve global information when reconstructing depth by creating a non-local fusion block (NLFB), which applies non-local operations to the features entering the fusion block. The advantages of this method are significant: Transformer is successfully applied to 360° depth estimation, which outperforms previous methods on multiple benchmark datasets and reaches the current optimal level; the inaccuracy of supervised learning's predictions because of data scarcity and the erratic performance of self-supervised learning are both successfully improved by the joint learning approach; the non-local fusion block better preserves the global information and improves the accuracy; the self-supervised learning part requires only gravity-aligned videos, which reduces the dependence on depth data and shows advantages in the comparison with other self-supervised learning methods. （Yun、Lee、Rhee，2021）

In 2022, Shen et al. proposed the PanoFormer model, which aims to solve the problem of depth estimation of indoor 360° panoramic images. It starts by designing a hierarchically structured network architecture in which the input stems perform the initial processing of the image, followed by the encoder and decoder progressively extracting and reducing the features through multiple hierarchical stages, where the key lies in the collaborative work of the positional embedding, PST blocks and convolutional layers included in each stage. In terms of feature processing, a pixel-level patch-dividing method is innovatively adopted to finely divide the input features and manually create tokens, which helps the network to capture more detailed features, differentiating it from the traditional Vision Transformer processing. In order to cope with the distortion problem of panoramic images, the relative position embedding is implemented using the Spherical Token Locating Model (STLM), which effectively reduces the effect of distortion on depth estimation by transforming operations between different domains. Meanwhile, to improve the perception of the panoramic geometric structure and provide important information for depth estimation, the enhanced panoramic self-attention mechanism with token flow, which is based on the classic Vision Transformer block, allows the final position of the tokens to be determined by the initialization and learnable flow during the computation of the attention scores, token flow, and resampled features. In the supervised part of model training, the objective function is designed by combining the inverse Huber (Berhu) loss and gradient loss, and the loss value is calculated according to a specific formula, which drives the model to be optimized continuously. Furthermore, two panorama-specific metrics–the left-right consistency error (LRCE) and the polar root mean square error (P-RMSE)–are suggested in light of the features of panoramic images. The P - RMSE focuses on measuring the depth estimation accuracy of the polar regions, while the LRCE is used to assess the depth consistency of the left and right boundaries, and together, they provide a comprehensive and tailored evaluation criterion for panoramic depth estimation to ensure the effectiveness and accuracy of the model in the panoramic image depth estimation task. （Shen、Lin、Liao，2022）

In 2022, Li et al. proposed the OmniFusion method for 360-degree monocular depth estimation. Given the presence of spherical distortion in 360 images affecting depth estimation, the method first converts the ERP input images into a set of tangent images using spherical projection, which can be used to predict the depth map using a conventional CNN architecture due to its lack of distortion. At the same time, considering that the independence of tangent image prediction of depth can cause problems, a geometric embedding network is introduced to combine pixel sphere coordinates with image centre coordinates to provide geometric information, which is fused with image features at an early stage of the encoder to enhance depth consistency. To compensate for the lack of overall information brought by decomposing ERP, Transformer is used to globally aggregate the image block features, and after convolutional dimensionality reduction and spreading to add positional embedding, the features are adjusted by self-attention. In addition, let the network predict the confidence maps, merge the depths in a weighted average manner, and add a regression layer to the decoder and transform the domain. In order to enhance the quality of depth estimation and successfully address the issue of 360-image depth estimation, an iterative depth refinement approach is also created to enhance the geometric embedding based on the iteratively updated depth information.（Li、Guo、Yan，2022）

In 2022, HiMODE was proposed by Masum Shah Junayed et al. as an innovative method for depth estimation of 360-degree panoramic images. The technique, which is based on a CNN+Transformer hybrid architecture, attempts to efficiently address the issues of data loss and distortion in the depth estimation of panoramic images. Architecturally, it employs a deeply separable convolutional CNN backbone network combined with a feature pyramid network that is capable of extracting high-resolution features near the edges, thereby reducing image distortion and artefacts. The Transformer module, on the other hand, plays an important role, with an encoder that enhances the ability to encode depth features by capturing the relationships between pixels and global information in the image through self-attention and cross-attention mechanisms; and a spatial and temporal patch (STP) and a multi-head self-attention (MHSA) layer in the decoder, which can process and recover encoded features to generate accurate depth maps. In addition, the method introduces linear projection and positional coding, where the feature maps extracted by the CNN are appropriately processed to fit the input requirements of the Transformer, and positional coding is utilised to enhance the understanding of the image features. By lowering the number of parameters and computational expenses, the spatial residual block (SRB) setting contributes to the system's increased stability and performance. Conversely, the contextual adjustment layer makes up for the lack of depth data and increases the accuracy of depth estimation by combining the feature maps that the CNN retrieved with the depth maps that the Transformer produced. The advantages of HiMODE include the following: firstly, it achieves excellent performance on multiple datasets and is able to estimate the depth map accurately, especially in recovering surface details and processing complex scenes. Second, the technique is very flexible, robust to the size of the input image, and can be used and trained on datasets of various sizes. In addition, HiMODE's hybrid architecture combines the advantages of CNN and Transformer to provide an effective solution for depth estimation of 360-degree panoramic images with a wide range of applications. （Junayed、Sadeghzadeh、Islam，2022）

3 EXPERIMENTS

3.1 Datasets

NYU-Depth V2 is a classical dataset focusing on indoor scene understanding, with simultaneous acquisition of colour (RGB) images and depth (distance) information by Microsoft Kinect sensors, containing 1449 pairs of densely annotated RGB and depth images (with each object annotated with a category and an individual number), as well as 464 diversified indoor scenes (e.g., home, office) from 3 cities) of more than 400,000 frames of raw, unlabelled video from three cities. The dataset is divided into three parts: Labeled (preprocessed labelled data with complete depth information), Raw (raw sensor data), and a supporting toolkit, which is widely used to train models for depth estimation, semantic segmentation, 3D reconstruction, etc., and is an important research foundation for robotics, AR/VR, and other fields.

KITTI is a reputable dataset in the field of automated driving that was jointly launched by the Toyota Technological Institute (TTI) and the Karlsruhe Institute of Technology (KIT) in Germany. It contains rich data gathered by a range of sensors (such as RGB cameras, stereo cameras, and 3D LIDAR) in real traffic scenarios, covering complex environments like cities and villages.Although its raw data does not provide fine annotations for semantic segmentation (e.g., pixel-level classification), the researchers manually annotated some of the images for different tasks (e.g., road detection, object recognition) involving categories such as roads, vehicles, pedestrians, and sky. This dataset has become a core test benchmark for autonomous driving algorithm development (e.g., environment perception, depth estimation, visual localisation) due to its realistic scenarios and diverse data.

The Stanford3D dataset is an important 3D computer vision dataset, which is mainly collected from indoor environments, covering several different scenes such as offices, bedrooms, and so on. The dataset contains rich data content, including a large number of 3D models in the form of triangular meshes, accompanied by information such as materials and colours, and the corresponding RGB images and depth images. At the same time, the dataset also provides detailed annotation information, including object category annotation, geometric annotation (e.g., 3D coordinates, dimensions, shapes, etc.) and semantic annotation (e.g., object functions, uses, etc.). These annotation information provides important support for tasks such as 3D model reconstruction, object recognition and classification, and scene understanding and interaction, making the Stanford3D dataset widely valuable for research and uses in the domain of computer vision.

A sizable RGB-D dataset for interior scene comprehension is called Matterport3D. 10,800 panoramic views of 90 actual building-scale sceneries made up of 194,400 RGB-D photos are included. A residential building with several rooms and floors that are labeled with surface configurations, camera postures, and semantic segmentation makes up each scene.

3.2 Evaluation Metrics

Indicators of the error or departure of a prediction from the actual value include Absolute Relative Error (Abs Rel), Squared Relative Error (Sq Rel), Root Mean Square Error (RMSE), and Threshold Precision Indicators ($\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$); Abs Rel is the average of the absolute values of the relative errors between the true and predicted values; the lower the value, the better the model performs and the closer the prediction is to the true value; The average of the squares of the relative errors between the true value and the anticipated value is known as Sq Rel. The model performs better and the forecast result is closer to the true value when the value is less. Sq Rel is the average of the squares of the relative errors between the predicted value and the true value, and again the smaller the value, the better the model performance; RMSE calculates the difference between the predicted value and the true value, and seeks for the average of the squares of the difference and then opens the square root, and the smaller the value is, the closer the predicted result is to the true value, and the better the model prediction ability is; and Threshold Precision is a metric that quantifies how much the actual depth value deviates from the projected depth value. The percentage of pixels in a certain threshold range is the ratio of the predicted depth value to the true depth value; the higher the ratio, the more accurately the depth estimate is made under the associated threshold value.

3.3 Experiment Data

Table 1: Experimental data on depth estimation methods for 2D images

Dataset	Method	Abs Rel	RMSE	δ	δ^2	δ^3
NYU-Depth V2	HybridDepth	0.026	0.128	0.988	1.000	1.000
	Metric3Dv2	0.047	0.183	0.989	0.998	1.000
	Marigold	0.055	0.224	0.964	0.991	0.998
	Depth Anything	0.056	0.206	0.984	0.998	1.000
	UniDepth	0.058	0.201	0.984	0.997	0.999
KITTI	Monodepth2	0.115	4.701	0.879	0.961	0.982
	PackNrt-SfM	0.107	4.538	0.889	0.962	0.981
	HR-Depth	0.104	4.410	0.894	0.966	0.984
	GEDepth	0.048	2.044	0.976	0.997	0.999
	EVP	0.048	2.015	0.980	0.998	1.000
	SQLdepth	0.043	1.698	0.983	0.998	0.999
	LightedDepth	0.041	1.748	0.989	0.998	0.999
	SPIDepth	0.029	1.394	0.990	0.999	1.000

As shown in table 1, the minimum values of Abs Rel and RMSE under different datasets are marked in bold, and the maximum value of threshold accuracy index is marked in bold, which shows that the current 2D image depth estimation methods of HybridDepth and SPIDepth have better performance, and the model prediction accuracy is higher.

Table 2: Experimental data on depth estimation methods for stereo images

Dataset	Method	Abs Rel	Sq Rel	RMSE	δ	δ^2	δ^3
Stanford3D	HoHoNet	0.0901	0.0593	0.4132	0.9047	0.9762	0.9933
	Omnidepth	0.1009	0.0522	0.3835	0.9114	0.9855	0.9958
	Bifuse	0.1214	0.1019	0.5396	0.8568	0.9599	0.9880
	SvSyn	0.1003	0.0492	0.3614	0.9096	0.9822	0.9949
	NLFB	0.0649	0.0240	0.2776	0.9665	0.9948	0.9983
	PanoFormer	0.0405	-	0.3083	0.9394	0.9838	0.9941
	UniFuse	0.1114	-	0.3691	0.8711	0.9664	0.9882
	OmniFusion	0.0950	0.0491	0.3474	0.8988	0.9769	0.9924
	HiMODE	0.0532	0.0207	0.2619	0.9711	0.9965	0.9989
Matterport3D	SvSyn	0.1063	0.0599	0.4062	0.8984	0.9773	0.9934
	Omnidepth	0.1136	0.0671	0.4438	0.8795	0.9795	0.9950
	HoHoNet	0.0671	0.0417	0.3416	0.9415	0.9838	0.9942
	Bifuse	0.1330	0.1359	0.6277	0.8381	0.9444	0.9815
	NLFB	0.0700	0.0287	0.3032	0.9599	0.9938	0.9982
	PanoFormer	-	-	0.3635	0.9184	0.9804	0.9916
	UniFuse	0.1063	-	0.4941	0.8897	0.9623	0.9831
	OmniFusion	0.0900	0.0552	0.4261	0.9189	0.9797	0.9931
	HiMODE	0.0658	0.0245	0.3067	0.9608	0.9940	0.9985

As shown in table 2, the minimum values of Abs Rel, Sq Rel, RMSE under different datasets are boldly marked, and the maximum value of the threshold accuracy index is boldly marked, which indicates that HiMODE, NLFB, and PanoFormer have performed better than other stereo image depth estimation techniques in recent years, and that the model prediction accuracy is higher.

4 DISCUSSION

Both 2D image depth estimation and stereo image depth estimation have their own shortcomings. 2D image depth estimation has difficulties in data acquisition and annotation, the model has limited performance in dealing with complex scenes, lighting changes and object occlusion, and feature extraction is insufficient with high model complexity and computational cost. Future research directions include data enhancement, model optimisation, and application scenario expansion, such as synthetic data, innovative neural network structure, and fusion of multimodal information. Stereo image depth estimation, on the other hand, is more complex in data acquisition and processing, and the model has challenges in edge and detail capture, global and local information fusion, and special scene processing, as well as high model complexity and computational cost. Future research can be carried out by optimising the dataset, improving the model structure, fusing multi-scale and multi-modal information, and improving the real-time performance for better application in augmented reality, virtual reality, intelligent robotics and autonomous driving.

5 CONCLUSION

A crucial task in computer vision, depth estimation measures the distance between each pixel in an image and the camera. This information is vital for many applications, including robotics, automatic driving, VR, AR, 3D reconstruction, security monitoring, and more. It also has significant research implications. The accuracy and efficiency of depth estimation have improved significantly with the advent of deep learning, and it is currently primarily separated into two categories: stereo image depth estimation and 2D image depth estimation. In 2D image depth estimation, HR-Depth optimises depth estimation by redesigning jump connections and introducing fSE modules, HybridDepth fuses focused stack information and single-image a priori to solve scale ambiguity, and SPIdepth strengthens the pose network to improve accuracy, which have shown excellent performance on multiple datasets. In stereo image depth estimation, UniFuse fuses different projection features, NLFB combines self-supervised learning with non-local fusion blocks, PanoFormer adopts innovative network architectures and strategies to deal with distortion, OmniFusion solves the problem of spherical distortion by using spherical projection and Transformer, and HiMODE reduces distortion by using a hybrid architecture of CNN+ Transformer. Transformer hybrid architecture to reduce distortion and data loss, each of these approaches has its own advantages in depth estimation of complex scenes. However, at present, 2D image depth estimation suffers from the problems of difficult data acquisition and annotation, limited ability of the model to deal with complex scenes, insufficient feature extraction, and high model complexity and high computational cost, etc. Stereo image depth estimation also faces the challenges of data acquisition and processing, edge detail capture, global and local information fusion, special scene response, and model complexity and computational cost. In the future, depth estimation research can be carried out in various aspects, such as at the data level, solving data-related problems by synthesising data and optimizing datasets; at the model level, innovating neural network structures, simplifying models, and fusing multi-scale and multi-modal information; at the application level, further expanding to augmented reality, virtual reality, intelligent robotics, and automated driving scenarios, and improving real-time and accuracy, in order to promote the depth estimation technology to a wider range of fields. estimation technology to be widely used and developed in more fields. Meanwhile, since stereo images usually contain richer depth information, such methods may become more and more popular in the research of depth estimation.

REFERENCES

Cheng, B., Yu, Y., Zhang, L., et al., 2024. Depth estimation of self-supervised monocular dynamic scene based on deep learning. Journal of Remote Sensing, 28(9).

Ganji, A., Su, H., Guo, T., 2024. HybridDepth: Robust Metric Depth Fusion by Leveraging Depth from Focus and Single-Image Priors. arXiv preprint arXiv:2407.18443.

Godard, C., Mac Aodha, O., Firman, M., et al., 2019. Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.3828-3838.

Guizilini, V., Ambrus, R., Pillai, S., et al., 2019. Packnet-sfm: 3d packing for self-supervised monocular depth estimation. arXiv preprint arXiv:1905.02693.

Hu, M., Yin, W., Zhang, C., et al., 2024. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Jiang, H., Sheng, Z., Zhu, S., et al., 2021. Unifuse: Unidirectional fusion for 360 panorama depth estimation. IEEE Robotics and Automation Letters, 6(2), pp.1519-1526.

Junayed, M.S., Sadeghzadeh, A., Islam, M.B., et al., 2022. HiMODE: A hybrid monocular omnidirectional depth estimation model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.5212-5221.

Ke, B., Obukhov, A., Huang, S., et al., 2024. Repurposing diffusion-based image generators for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9492-9502.

Lavreniuk, M., 2024. SPIdepth: Strengthened Pose Information for Self-supervised Monocular Depth Estimation. arXiv preprint arXiv:2404.12501.

Lavreniuk, M., Bhat, S.F., Müller, M., et al., 2023. Evp: Enhanced visual perception using inverse multi-attentive feature refinement and regularized image-text alignment. arXiv preprint arXiv:2312.08548.

Li, Y., Guo, Y., Yan, Z., et al., 2022. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.2801-2810.

Lyu, X., Liu, L., Wang, M., et al., 2021. Hr-depth: High resolution self-supervised monocular depth estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(3), pp.2294-2301.

Piccinelli, L., Yang, Y.H., Sakaridis, C., et al., 2024. UniDepth: Universal monocular metric depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10106-10116.

Shen, Z., Lin, C., Liao, K., et al., 2022. PanoFormer: Panorama transformer for indoor 360 ∘ depth estimation. In European Conference on Computer Vision, Cham: Springer Nature Switzerland, pp.195-211.

Sun, C., Sun, M., Chen, H.T., 2021. Hohonet: 360 indoor holistic understanding with latent horizontal features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.2573-2582.

Wang, F.E., Yeh, Y.H., Sun, M., et al., 2020. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.462-471.

Wang, Y., Liang, Y., Xu, H., et al., 2024. Sqldepth: Generalizable self-supervised fine-structured monocular depth estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(6), pp.5713-5721.

Yang, L., Kang, B., Huang, Z., et al., 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10371-10381.

Yang, X., Ma, Z., Ji, Z., et al., 2023. Gedepth: Ground embedding for monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.12719-12727.

Yun, I., Lee, H.J., Rhee, C.E., 2022. Improving 360 monocular depth estimation via non-local dense prediction transformer and joint supervised and self-supervised learning. In Proceedings of the AAAI Conference on Artificial Intelligence, 36(3), pp.3224-3233.

Zioulis, N., Karakottas, A., Zarpalas, D., et al., 2019. Spherical view synthesis for self-supervised 360 depth estimation. In 2019 International Conference on 3D Vision (3DV), IEEE, pp.690-699.

Zhu, S., Liu, X., 2023. LightedDepth: Video depth estimation in light of limited inference view angles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.5003-5012.