Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research on 3D Reconstruction Methods Based on Deep Learning

Lu Ge

School of Electrical and Computer Engineering Xiamen University Malaysia, 43900 Sepang, Selangor, Malaysia

DMT2109239@xmu.edu.my

Abstract. Deep learning applications have been applied extensively and have made tremendous strides in the 3D reconstruction field in recent years. This paper offers a methodical review of deep learning-based techniques for single-view images, multi-view images, and video-based sequence approaches. For single-view methods, we focus on depth estimation using Convolutional Neural Networks (CNNs) and image-to-depth mapping using Generative Adversarial Networks (GANs). For multi-view methods, we explore 3D reconstruction based on multi-view stereo matching method, 3D points cloud reconstruction method and stereo flow estimation method. For video-based methods, we introduce depth estimation method based on optical flow and video sequence modeling using Recurrent Neural Networks (RNNs). Generally, the multi-view image method is more accurate and sophisticated than the single-view image method, while the method based on video sequences is more challenging and complex. Different 3D reconstruction methods depend on specific application scenarios and requirements. This review provides a considerable insight in research for 3D reconstruction and also make the conclusion as well as future prospect for this field.

Keywords: 3D Reconstruction; Methods; Deep Learning.

1. Introduction

Deducing the three-dimensional geometry and structure of objects and situations from one or more two-dimensional photographs is the aim of image-based 3D reconstruction. Large amounts of applications, such as industrial control, medical diagnosis, 3D modeling and animation, object recognition, scene understanding, robot navigation, and industrial control, depend on this technology. To be more specific, in autonomous driving, accurate 3D reconstruction of the environment is important and essential for obstacle detection and path planning. In manufacturing, 3D reconstruction is used for quality control and inspection of products. In healthcare, it can be used in medical imaging and surgical planning.

Traditional methods of 3D reconstruction often facing difficulties of dealing with complex scenes, occlusions, and textureless surfaces. Handcrafted features and geometric constraints can be computationally expensive and may not generalize well to diverse environments. Deep learning-based approaches offer a promising solution by automatically learning features and representations from data, potentially leading to more accurate and efficient 3D reconstructions.

Convolutional neural networks (CNNs) and generative adversarial networks (GANs), two examples of deep learning models, perform exceptionally well in a variety of computer vision applications, including object identification, semantic segmentation, and picture classification. These models can capture the substructure of 3D scenes and objects by learning complex patterns and representations from large-scale datasets, which makes deep learning particularly well-suited for 3D reconstruction, where the relationship between 2D images and 3D shapes is inherently complex and non-linear.

By using the capabilities of deep learning, researchers have developed new methods for 3D reconstruction that which is better than traditional approaches in terms of accuracy and robustness.

This review paper focus on 3D reconstruction methods grounded in deep learning, which have already shown the progress in reconstructing 3D shapes from 2D images with many details. We can divide these methods into three main groups: single-view images, multi-view images, and video sequences.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

We discuss basic principles, advantages, and disadvantages of different approach, and put forward some future research directions. By analyzing these, researcher can better apply the deep learning techniques for 3D reconstruction, and also better develop the application ssuch as robotics, autonomous driving, and virtual reality.

2. Mainbody

An overview of the 3D reconstruction technique based on single-view, multi-view, and video sequences is given in this section. It covers methods like employing CNNs to estimate depth from single-view photos., single-scale and multi-scale methods, as well as GANs. In Addition, it discusses the methods based on multi-view images, such as Structure from Motion (SfM), Multi-View Stereo (MVS), and Depth from Stereo (DFS), compared their similarities and differences in reconstructing scenes from multiple views. Moreover, it covers methods based on video sequences, which focus on video depth estimation techniques that utilize optical flow estimation to predict depth information. These are methods that all combine deep learning with optical flow estimation, which could finally improve accuracy and robustness of depth estimation from videos.

2.1. Single-view image-based methods

Single-view image-based methods use CNNs for depth estimation. These methods can be further categorized into single-scale depth estimation and multi-scale depth estimation, each with its strengths and limitations.Additionally, we review approaches that use GANs to generate depth maps from images. These include unidirectional GANs, bidirectional GANs, conditional GANs, and self-supervised GANs, each offering unique advantages in capturing fine details and handling occlusions.

2.1.1. CNN-based Depth Estimation Methods

Monocular Depth Estimation is the process of estimating depth from a single RGB image [1]. This task is relatively easy for humans because we can leverage visual cues such as perspective, object size, lighting, and occlusion. For computational models, it is a difficult problem nevertheless, as one 2D image can equate to a plethora of different 3D situations. A growing amount of models for deep learning have been used for single-scale depth estimate in recent years as deep learning has advanced. These models are mostly built on deep CNN and no longer depend on manually created features.

Based on deep learning, monocular depth estimation methods learn depth features through multi-layer neural networks, exhibiting higher accuracy in processing single-eye images. Even in cases of image occlusion or missing ground truth depth, this method can estimate scene depth well with lower error rates. When faced with extensive occlusion or lack of actual ground depth information, this method can improve depth estimation accuracy and robustness by introducing network constraints to learn scene depth. However, one disadvantage is their high computational complexity, which can make real-time applications challenging, especially in resource-constrained environments. Despite this limitation, the significant advancements in accuracy and robustness make deep learning-based monocular depth estimation a valuable tool in various computer vision applications.

As shown in Figure 1, the single-view image-based monocular depth estimation method employs 2D and 3D CNNs combined with contextual information. In order to determine the ultimate depth map using contextual data. During training, feature maps are extracted from the left and right images using two 2D CNNs that have shared weights. These feature maps are then concatenated into the cost volume module of a 3D convolutional network. PSMNet is a well-liked model that was trained using a top-down/bottom-up approach for unsupervised monocular depth estimation. This technology aggregates semi-global environment data using a spatial pyramid pooling module as a matching cost volume. A stack of hourglass-based 3D CNNs under intermediate supervision adjusts the matching cost volume to enhance the robustness and accuracy of depth estimation [2].

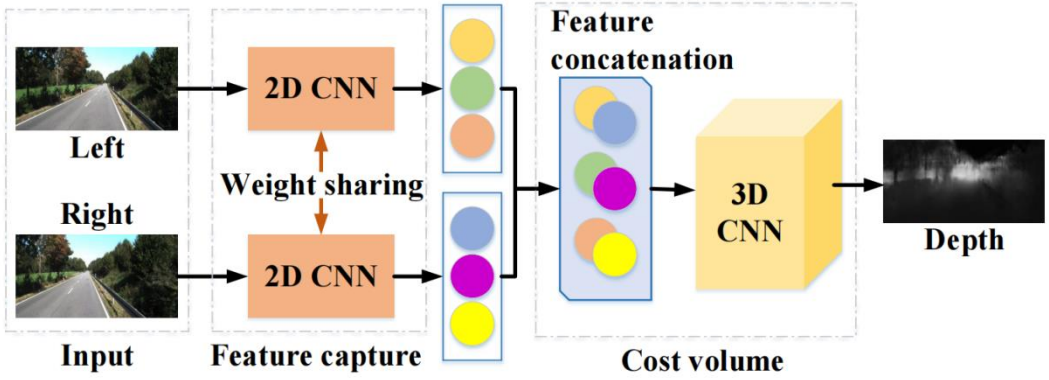

Fig. 1. The overall monocular depth estimation approach is built on unsupervised 2D and 3D CNNs. [2]

2.1.2. Single-scale Depth Estimation

In monocular depth estimation, features are extracted using deep learning models like CNNs. It is particularly in the context of Single-scale Depth Estimation from Single-view pictures. These features, including edges, textures, and colors, which can be helpful to understand the depth and geometric structure of a scene. Compared with multi-scale depth estimation, Single-scale Depth Estimation is more simpler and direct, as it do not need to process the information at multiple scales, it would lead to higher computational efficiency. However, since Single-scale Depth Estimation considers only the single scale information, it may perform poorly when dealing with multiple scales objects or complex scenes, this would make depth estimate less reliable and accurate.

One advantage of Single-scale Depth Estimation is its simplicity and typically higher computational efficiency. As Single-scale methods only need to process images at a single scale, they perform well in scenarios with limited computational resources or high real-time requirements. However, Single-scale Depth Estimation also has limitations, particularly in handling scene details and complex structures. Because Single-scale methods only consider information from a single scale, they may not perform as well as multi-scale methods when dealing with complex scenes or objects of multiple scales.

2.1.3. Multi-scale Depth Estimation

Multi-scale depth estimation could improve depth estimation accuracy by combining data from several scales. This approach is to processimages at different resolutions or scales. It can capture both details overall scene structure, creating more comprehensive results. The process of multi-scale depth estimation typically involves several of main steps.First of all, the input image is sampled to create multiple scales or resolutions of the image.Then, each scale of the image is processed independently through a convolutional neural network or a similar model to extract relevant features.These features capture information at different levels of detail, allowing the model to analyze the scene from various perspectives. Next, these features are combined to create a comprehensive representation of the scene. This process integrates information from different scales, enhancing the model's understanding of the scene's geometry and depth distribution. The final depth map, which depicts the scene's depth at each pixel, is estimated using the fused features.

One advantage of multi-scale depth estimation is it can process the scenes with different levels of complexity and object scales. By processing images at different scales, the model can adapt to different scene characteristics. As a result, it can have a more complete understanding of the scene's structure. However, multi-scale depth estimation also faces some challenges. On the one hand, it is determining the optimal scales or resolutions to use for processing. Choosing the right scales can significantly influence the accuracy and efficiency of this estimation process. In Addition, integrating information from multiple scales requires careful design which could ensure that the model can effectively combine features from different scales, without introducing artifacts or facing some

inconsistent situations in the final depth map.In conclusion, multi-scale depth estimation is a approach which is worth for developing, it is useful to improve depth estimation accuracy and robustness in computer vision applications. By utilizing information from multiple scales, this approach can enhance the model's understanding of scene geometry and depth distribution, leading to more accurate and reliable depth estimation results.

2.2. The method of generating depth maps from images using GANs

Adversarial Generative Networks. Because GANs can mimic computer vision challenges like rivalry between two networks, they are growing in popularity. Impressive results have been obtained by this technology in tasks including editing, representation learning, and image generalization. Translations from text to image and image to image have also been used recently [3]. This approach employs GANs to generate more vivid and clear depth maps than traditional models. The two main modules in the process are the discriminator, it assesses if the input depth map is generated or real, and the generator, which predicts the depth map using a depth estimation network. The generator is incentivized to generate depth maps that are identical to ground truth depth maps due to this adversarial setting. Many GAN designs have been used in depth estimation. These architectures take advantage of the adversarial training framework to produce more realistic and high-quality depth maps. They also offer useful restrictions that help to increase the resilience and accuracy of depth estimation.

The deep learning-based monocular depth estimation framework uses the encoder-decoder network topology, as shown in Figure 2. It produces depth maps after receiving RGB photos as input. The decoder part uses deconvolutional layers to transform features into pixel-level depth maps the same size as the input, while the encoder section uses convolutional and pooling layers to extract depth information. Skip connections are also used to pair matching layers of the encoder and decoder so as to maintain features at each scale. To produce the desired depth maps, the complete network progressively converges under the constraints and training of the depth loss function.

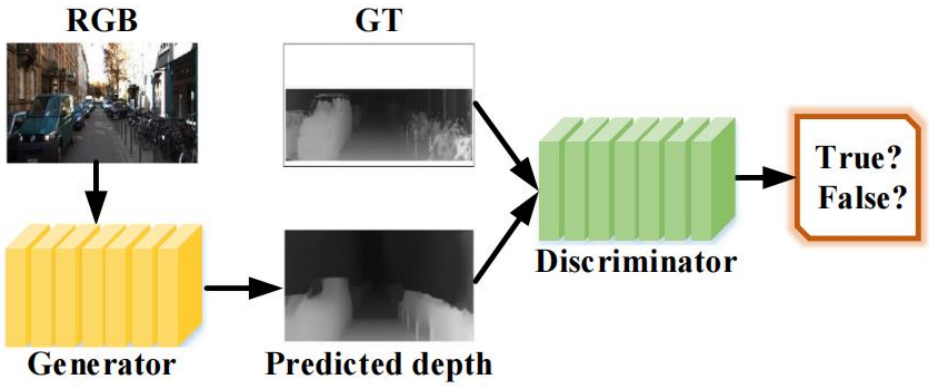

Fig. 2. The general GAN-based framework for supervisedmonocular depth estimation [2]

2.3. Methods Based on Multi-view Images

Methods based on multi-view images are an essential part of 3D reconstruction, using information from multiple images captured which is from different viewpoints to create detailed three-dimensional models of objects or scenes. With the success of deep learning in various applications, deep multi-view learning methods have also gained increasing attention, showing promising results [4].

Compared to single-view image-based methods, there are advantage of multi-view image-based methods. They can provide more detailed and accurate reconstructions, especially in complex scenes with occlusions and texture less regions. Additionally, they can handle dynamic scenes more effectively, as they can incorporate information from multiple frames captured over time.

2.3.1. Three-dimensional point cloud reconstruction methods

The SfM photogrammetric technique is based on the concepts of binocular vision and shifting vision of an object that is either moving or being observed from a moving point.[5].It is used to reconstruct the structure of a 3D scene and the motion of cameras from multiple images. By using feature points in multiple views and observe the correspondence relationship of them between different views, it can infer the 3D structure of the scene and also camera trajectory. The core principle of SfM is triangulation, which involves observing feature points in multiple views and then using their projection relationships in different views to calculate their positions in 3D space. By collecting a sufficient number of feature points and views, the entire 3D structure of the scene can be reconstructed. SfM can be useful when utilizing the geometric relationships between multiple views to reconstruct the 3D scene, without getting know the camera's internal parameters or the scene's structure beforehand. This is the widen use of SfM in various fields, such as drone aerial photography, 3D modeling, VR and so on.

Multi-View Stereo (MVS) is a technique used to reconstruct a 3D scene using images from multiple views. Through matching feature points in multiple views and take use of their relationship, MVS could estimate the 3D positions of points in the scene. MVS use the overlapping regions between multiple view to calculate the depth of feature points. By matching feature points between different views and using their projection positions in the camera coordinate system and parameters, positions of these points can be inferred. Finally, result of the entire scene can be achieved by merging the 3D positions of feature points from all views. MVS can use the geometric information and overlapping regions between multiple views that can improve the accuracy and density of 3D reconstruction. Compared with single-view methods, MVS can better handling situations such as occlusion and texture loss, that performing better in complex scenes.

Depth from Stereo is a method estimating scene depth by utilizing the disparity information between pair of images which is captured by a stereo camera. The basic concept is to use two slightly offset cameras (left and right eyes) to capture images of the same scene. Analyzing the difference between these two images, it could infer the depth information of points in the scene. Focusing on triangulation based on the geometric relationship between disparity and camera parameters, the technique can calculate the depth of each point in the scene. Specifically, the camera parameters are calibrated, and images are rectified to align at corresponding points in the left and right images. Next, by integrating the camera's internal parameters with the baseline length between the cameras, and computing the difference between matching points, it can achieve the depth information of each point. It is a useful technique to infer the three-dimensional structure of the scene which is based on the disparity information between image pairs captured by a stereo camera that do not use other sensors or complex equipment.

SfM (Structure from Motion), MVS (Multi-View Stereo), and DFS (Depth from Stereo) are three common methods for multi-view 3D reconstruction, contain both similarities and differences in reconstructing scenes from multiple views. They all infer the 3D structure and camera motion of a scene using images from multiple views. These methods utilize the principle of triangulation, that could infer the 3D position of points in the scene by observing the projection positions and features difference in different views. However, they differ in some of their focus and methods during the reconstruction process [6]. To be specific, SfM mainly focus on estimating camera motion and scene structure, while MVS is more concerned about the precise position of each point in the scene through matching and triangulation. DFS, on the other hand, is special that uses the disparity information from a stereo camera to estimate the depth of the scene without the need for camera motion estimation.

The advantage of 3D point cloud reconstruction lies in its utilization of UAV images for dense matching, enabling the recovery of the three-dimensional geometric shape and structure of buildings and terrain from point cloud data. This method can provide high-precision 3D models of buildings, which is highly effective for applications such as architectural reconstruction and terrain modeling. However, this method requires a large amount of image data and complex computation, placing high demands on computational resources and processing time.

2.4. Methods Based on Video Sequences

The computer science field has study reconstructing 3D scenes from image sequences methods for many years. In recent years, with the continuous progress of high-quality reconstruction systems, scentists have made great advances in achieving more precise capture and producing denser reconstructions [7].

Methods based on video sequences is important in deep learning. Taking use the space relationships between frames in videos, these methods can better analyze video content. In video processing, preprocessing of videos is the first step of extract motion information from the video, such as inter-frame differencing and interpolation. Then, using extracted features such as optical flow and color histograms, models can capture the dynamic changes in video sequences. These methods are more than important in different applications. It provide strong support for the understanding and processing of video content.

2.4.1. Video Depth Estimation Method Based on Optical Flow Estimation

Two core issues in computer vision are optical flow analysis and depth estimation. By combining these two tasks, dense 3D scene flow may be computed, a useful tool in areas like video analysis, autonomous driving, and robot navigation [8]. The optical flow information between the input frame and future frame is predicted using the optical flow estimation-based video prediction technique. Pixels from the input frame are sampled to construct the final forecasted frame [9]. Optical flow is a representation of the movement of pixels between consecutive frames in an image sequence. It reveals the temporal changes in pixel positions across the image. To determine the motion information of each pixel in the image, we first compute the optical flow field between the present frame and the following frame in video prediction. We then sample pixels from the current frame to create the anticipated next frame image based on this motion information.

DeepV2D predicts depth from a calibrated video sequence [10]. It combines deep learning together with optical flow estimation, to estimate scene depth information from a video sequence. Firstly, DeepV2D uses Convolutional Neural Network to extract features from video frames that contains rich spatial information. Then by taking use the result of CNN and within optical flow estimation, it can produce a depth map through the network. It taking good use of the advantages of deep learning in feature learning and representation learning, also with the advantages of optical flow estimation in capturing motion information, it finally get a significant progress in the video depth estimation.

The focus of DORN is to enhance the accuracy and resilience of depth estimation by utilizing camera motion and optical flow. In order to improve the accuracy of the depth map, it extends the RAF algorithm for optical flow estimate by estimating optical flow between numerous viewpoints. To be specific, DORN firstly use Convolutional Neural Networks to get features from video frame and then combines the optical flow information between multiple views to achieve precise estimation of the depth map. This approach take good use of optical flow estimation in motion information estimation and also deep learning in advantage of extraction and learning, that achieving good results in video depth estimation.

RAFT-3D is a method that improve and extend the existing optical flow estimation algorithms which would better adapt to the need of 3D reconstruction. The basic idea is using optical flow information between multiple views in order to increase the quality of the depth map. RAFT-3D could get more accurate and dense optical flow fields by estimating optical flow between multiple views, that would be more accurate to infer depth information. It is significant that this method focus on the improvement and extension of optical flow estimation algorithms, for the purpose of meeting specific requirements of 3D reconstruction, which is a great progress in video depth estimation field.

By combining deep learning with optical flow estimation approaches, all three of these methods enhance the precision and resilience of video depth estimation. They all integrate the outcomes of optical flow estimation to produce depth maps after using convolutional neural networks to extract characteristics from video frames.

3. Conclusion

This paper provides a detailed discussion and analysis of methods based on single-view images, multi-view images, and video sequences. The research on these methods is of great significance for advancing the field of computer vision and 3D scene understanding.

First of all, the development of intelligent interaction interfaces, virtual reality, and augmented reality technologies is fundamentally supported by the ongoing advancement of depth estimation techniques in single-view and multi-view scenarios.By calculating depth information more precisely, the virtual worlds can be more real and interactive.Furthermore, multi-view image-based techniques offer powerful technological support for a variety of applications, including virtual navigation, environmental monitoring, and map creation. Three-dimensional scenes can be more precisely rebuilt according to merge data from many points of view, offering a more real environment simulation for applications, such as intelligent navigation and simulation training.

These techniques will continue to be refined and extended in their fields of application as long as technology like generative adversarial networks, deep learning, and multi-view photography continue to advance. Applying depth estimation and picture creation technologies, for instance, to smartphones and virtual reality glasses, there will be a more realistic and immersive user experience in the intelligent interaction interfaces field. Combining deep learning technology with multi-view imaging can make the medical image diagnosis more precisely. Furthermore, video sequence-based techniques, like optical flow estimation for video prediction, which is very important for understanding and evaluating video information. These techniques use the spatiotemporal information found in movies in order to make improvement in behavior analysis, object tracking, and action.

Reference

[1] Bhoi, A. Monocular depth estimation: A survey. arXiv preprint arXiv:1901.09402. 2019

[2] Ming, Y., Meng, X., Fan, C., & Yu, H. (2021). Deep learning for monocular depth estimation: A review. Neurocomputing, 438, 14-33.

[3] Aleotti, F., Tosi, F., Poggi, M., & Mattoccia, S. (2018). Generative adversarial networks for unsupervised monocular depth prediction. ECCV workshops. 2018.

[4] Roca-Pardinas, J., Lorenzo, H., Arias, P., & Armesto, J. From laser point clouds to surfaces: Statistical nonparametric methods for three-dimensional reconstruction. Computer-Aided Design, 40(5), 646-652, 2008

[5] Iglhaut, J., Cabo, C., Puliti, S., Piermattei, L., O'Connor, J., & Rosette, J. Structure from motion photogrammetry in forestry: A review. Current Forestry Reports, 5, 155-168, 2019.

[6] Zhou, Y., Zhang, L., Xing, C., Xie, P., & Cao, Y. Target three-dimensional reconstruction from the multi-view radar image sequence. IEEE Access, 7, 36722-36735, 2019

[7] Luo, X., Huang, J. B., Szeliski, R., Matzen, K., & Kopf, J. Consistent video depth estimation. ACM ToG, 39(4), 71-1, 2020

[8] Wang, Y., Wang, P., Yang, Z., Luo, C., Yang, Y., & Xu, W. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In CVPR (pp. 8071-8081, 2019

[9] Lu, W., Cui, J., Chang, Y., & Zhang, L. A video prediction method based on optical flow estimation and pixel generation. IEEE Access, 9, 100395-100406, 2021

[10] Teed, Z., & Deng, J. Deepv2d: Video to depth with differentiable structure from motion. arXiv preprint arXiv:1812.04605, 2018.