Multi-View Stereo Method

Subjects: Mathematics, Applied

Contributor: Zhen Liu, Guangzheng Wu, Tao Xie, Shilong Li, Chao Wu, Zhiming Zhang, Jiali Zhou

As a 3D reconstruction method, multi-view stereoscopic (MVS) plays a vital role in 3D computer vision, and has a wide range of applications in the fields of virtual reality, augmented reality, and autonomous driving. With the rapid development of deep learning technology in the field of computer vision, the learning-based multi-view stereo method has produced advanced results.

Keywords: multi-view stereo ; cost volume

1. Introduction

As a 3D reconstruction method, multi-view stereoscopic (MVS) plays a vital role in 3D computer vision, and has a wide range of applications in the fields of virtual reality, augmented reality, and autonomous driving. With a series of images from different viewpoints and the corresponding camera parameters as inputs, multi-view stereo can estimate the depth information of each pixel and generate a corresponding 3D representation of the observed scene. As a key problem in 3D computer vision, multi-view stereo has received extensive research attention [1,2,3,4].

In recent years, with the rapid development of deep learning technology in the field of computer vision, learning-based multi-view stereoscopic methods have achieved advanced results [4,5,6].]。 Learning-based multi-view stereo algorithms usually consist of multiple components, including feature extraction, deep sampling, cost quantity construction, cost quantity regularization, and deep regression. However, the large GPU memory requirements not only limit image processing to low resolutions, but also hinder the adoption of multi-view stereo on various edge computing devices. In the real-world application of 3D vision, the deployed devices often have limited computing resources. For example, in autonomous driving scenarios, lidar data is often processed using 3D point cloud compression technology to reduce storage and transmission costs [7]. Unlike LiDAR data processing, the main computational challenge for multi-view stereo is to generate a point cloud from a 2D image and camera parameters from a given input source. Therefore, reducing the memory consumption of the algorithm can greatly improve the practicability of the technology. Recently, many researchers have proposed improved methods to deal with the high computational problem of learning-based multi-view stereo methods. In particular, coarse-to-fine architectures have been widely used to design efficient multi-view stereo networks [6,8,9,10,11,12]. Typically, in these methods, the initial cost volume is typically built at a low resolution rather than a fixed resolution, and then a new cost volume is iteratively built at a higher resolution based on the final stage results, and finally a depth map is obtained. The assumption that the depth plane is progressively reduced at different stages [6,8,9,10,11,12] is also a key strategy to reduce the amount of computation. Although the coarse-stage output is of great significance to the final result as an input to the construction of fine-stage cost-quantities, these existing methods need to pay more attention to the characteristic information of the coarse stage. If the feature extraction phase of the coarse stage is insufficient, the poor initial results may adversely affect the final results and final output of the subsequent stages. However, intensive feature extraction steps always increase computational load and GPU consumption, and there is still a need to balance accuracy and computational efficiency.

In addition, another existing challenge for cascade-based multi-view stereoscopic is the adaptation to the depth hypothetical range. In the initial phase, planar scanning covers the entire imaginable depth range. At the same time, in many cascade-based algorithms, the estimated depth value of the previous stage is used as the center of the sampling interval during the generation of depth assumptions at a more granular stage [6,8,10,12], and each pixel has a fixed sampling distance within its respective stage. However, setting a uniform sampling distance for each pixel is not an ideal approach because the optimization of the depth refinement stage will vary between different pixels in the same depth map, where some pixels may have a stable depth while others may exhibit significant variations. With this challenge in mind, Cheng [11] used the probability distribution of each pixel to set the sampling distance; However, this approach does not perform well in terms of GPU memory usage and runtime, while its training time is also significant.

2. Multi-view stereo mode

2.1. Traditional multi-view stereo method

Multi-view stereoscopic (MVS), as a fundamental problem in the field of 3D reconstruction in computer vision, solves the problem of recovering the spatial geometry of a scene from a photo. Before the advent of deep learning, it had already attracted a great deal of attention and made substantial progress. Traditional multi-view stereo methods can be broadly divided into the following four categories: voxel-based methods [25,26,27,28,29] and grid-based methods [30,31]. Surface-based methods [19,32,33] and depth map-based methods [1,20,21,34,35]. The mesh-based approach is less reliable because its final reconstruction performance depends on its initialization. At the same time, the surfel-based method represents the surface as a set of surfels, which is simple but efficient. However, surfel-based methods require additional cumbersome post-processing steps to generate the final 3D model. The depth map-based approach calculates the depth value of each pixel in each image, reprojects the pixels into 3D space, and then fuses the points to generate a point cloud model. Among the four methods, the depthmap-based method is the most flexible and the most widely used in this field. In recent years, depth map-based methods have achieved remarkable success, and good algorithmic frameworks are in use, such as Furu [19], Gipuma [21], Tola [20], and COLMAP [1]. While the performance of traditional multi-view stereoscopic is commendable, there are still shortcomings that need to be improved: high computational requirements, slow processing speed, and poor handling of scenes with weak textures or weak reflective surface blocks.

2.2. Learning-based multi-view stereoscopic approach

In recent years, with the convergence of deep learning, the learning-based multi-view stereo method has experienced rapid development and achieved outstanding performance. Yao [4] launched MVSNet, the first multi-view stereoscopic network based on end-to-end learning, laying the foundation for rapid growth in the coming years. MVSNet [4] uses a shared-weight 2D-CNN network to extract feature maps from the input images. Differential monotonic transformations [36] are then applied to distort these feature maps into reference perspectives. The method utilizes a series of depth assumption planes to construct a cost volume that represents the correlation between the source and reference images. Subsequently, the 3D-CNN network was used for cost regularization. Finally, the estimated depth map of the output as a reference image is obtained by depth regression. In the DTU benchmark dataset [17], MVSNet [4] not only outperforms the previous traditional MVS methods [1,19,20], but also runs much faster. However, due to the high GPU memory consumption, only low-resolution images can be used as input images in MVSNet. A number of learning-based MVS approaches have been proposed to deal with GPU memory consumption. Yao [22] proposed an improved method, R-MVSNet [22], which replaces the deeply refined 3D-CNN network with a series of GRU convolutions. This improvement reduces GPU memory consumption and enables 3D reconstruction at high resolution. Gu [6] proposed the CasMVSNet model, which is based on the Feature Pyramid Network (FPN) [13] to construct cascading costs. Thanks to its novel coarse-to-fine architecture, CasMVSNet can process input images from DTU datasets at native resolution [17]. Similar to CasMVSNet [6], CVP-MVSNet [8] and Fast-MVS [23] also contain coarse-to-fine frameworks, and both exhibit excellent performance on benchmark datasets [17,18]. Based on the coarse-thickness cascade framework, UCS-Net [11] further introduces a depth sampling strategy that uses uncertainty estimation to adaptively generate spatially varying depth assumptions. Vis-MVSNet [9] also uses uncertainty to explicitly infer and integrate pixel occlusion information in multi-view cost volume binning. PatchMatch [2], as a classical and traditional stereo matching algorithm, has also been integrated into the learning-based MVS framework, and the resulting model is named PatchmatchNet [2]. Recently, Effi-MVS [10] has been proposed, demonstrating a new method for constructing dynamic cost quantities in deep refinement. In addition, TransMVSNet [37] is the first learning-based MVS approach that leverages Transformer [38] to enable powerful, remote global context aggregation within and between images.