RGB 이미지에서 트랜스포머 기반 고밀도 3D 재구성

서가가 [1], 고서 [1], 문명운 [1], 조경은 [1*]
[1] 동국대학교 멀티미디어공학과
2018126716;gaorui; wmy_dongguk@dongguk.edu,
*cke@dongguk.edu (교신저자)

Transformer-based dense 3D reconstruction from RGB images

Jiajia Xu[1], Rui Gao[1], Mingyun Wen[1] and Kyungeun Cho[1*]
[1]Department of Multimedia Engineering, Dongguk University-Seoul

Abstract

Multiview stereo (MVS) 3D reconstruction of a scene from images is a fundamental computer vision problem that has been thoroughly researched in recent times. Traditionally, MVS approaches create dense correspondences by constructing regularizations and hand-crafted similarity metrics. Although these techniques have achieved excellent results in the best Lambertian conditions, traditional MVS algorithms still contain a lot of artifacts. Therefore, in this study, we suggest using a transformer network to accelerate the MVS reconstruction. The network is based on a transformer model and can extract dense features with 3D consistency and global context, which are necessary to provide accurate matching for MVS.

1. Introduction

Recently, multi-view stereo vision [1-3], a key area of computer vision, has been extensively researched. Traditional MVS techniques compute dense correspondences and recover 3D points using engineering regularization and similarity measures [4]. In an ideal Lambertian scenario, these approaches achieve good results but have significant limitations. For example, dense matching becomes difficult in low-textured scenes and in regions containing specular reflections, leading to incomplete reconstructions [1].

2. Related work

The traditional MVS method comprises three processes: camera motion estimation, sparse 3D reconstruction, and dense 3D reconstruction [5]. Popular techniques for dense 3D reconstruction include clustering views for multi-view stereo (CMVS) [6] and patch-based multi-view stereo (PMVS) [7].

Recently, many studies have employed learning-based methods that have achieved significant performance improvements compared to traditional methods. Some studies [8-9] use a coarse-to-fine framework for depth estimation and to reduce computational complexity in using multiple stages. However, most studies have focused on using convolutional neural networks (CNN) as the backbone for feature extraction, but failed to capture various scale features. Because of the recent applications of transformer in the direction of computer vision, many researchers found that a transformer model can effectively capture the global features of images [10].

<Table 1> Limitations of existing methods

Method by the output format	Limitations
Traditional method [6,7]	o Poor performance on non-Lambertian surfaces, areas with little texture, and regions with no texture.
Existing learning-based methods [8-9]	o It is memory expensive. o The bulk of these systems use an inflexible fixed-cost volume representation for the scene.

3. Proposed Method

In contrast to the previous MVS methods, the proposed method extracts the features of the input RGB images using a transformer-based self-attentive hierarchical feature extraction module, which is comparatively more effective.

The proposed method includes three main modules—the RGB image feature extraction module that comprises a transformer network, the cost volume construction module that matches the extracted RGB image feature range to a suitable interval, and the cost regularization module that is used to construct the depth value of each input image and outputs the confidence level of each depth value. Finally, a depth map fusion process generates a single point cloud representation by combining depth maps from multi-view RGB images.

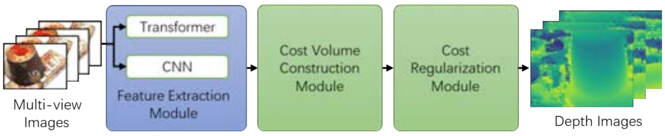
Multi-view Images
Transformer
CNN
Feature Extraction Module
Cost Volume Construction Module
Cost Regularization Module
Depth Images

Figure 1. Overview of the proposed method

4. Experiment

The experiments were conducted using PyTorch, with a Windows 10 and an Nvidia RTX 2070 SUPER GPU. An Intel Core i7-9700 CPU running Python 3.7 was configured to the system.

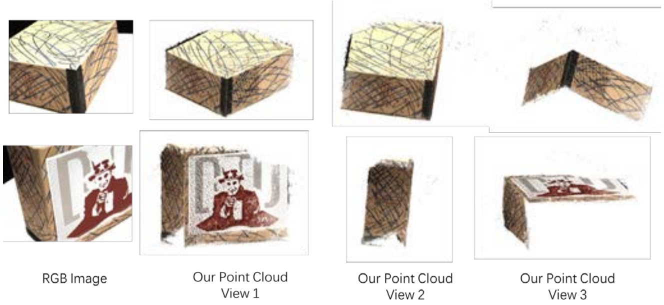

Figure 2. Results of the proposed method

Figure 2 shows the point cloud results in the open dataset DTU [11]. The reconstructed dense point cloud can well represent the appearance of an RGB image, which means the proposed methods can reconstruct a high-quality 3D point cloud.

5. Conclusion

This research proposes using a transformer network for MVS, which can effectively extract features from multiview RGB images. The experiments proved that the proposed method could reconstruct high-quality dense 3D point clouds.

Acknowledgments
This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1A2C200686411).

References

[1] Y. Yao, Z. Luo, S. Li, T. Fang, L. Quan Mvsnet: Depth inference for unstructured multi-view stereo Proceedings of the European Conference on Computer Vision (ECCV) (2018), pp. 767-783.

[2] J. Zbontar, Y. LeCun Computing the stereo matching cost with a convolutional neural network Proceedings of the IEEE conference on computer vision and pattern recognition (2015), pp. 1592-1599

[3] X.X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, F. Fraundorfer Deep learning in remote sensing: A comprehensive review and list of resources IEEE Geosci. Remote Sens. Mag., 5 (4) (2017), pp. 8-36

[4] H. Hirschmüller Stereo processing by semiglobal matching and mutual information IEEE Trans. Pattern Anal. Mach. Intell., 30 (2) (2007), pp. 328-341.

[5] Z. Ma, S. Liu A review of 3d reconstruction techniques in civil engineering and their applications Adv. Eng. Inform., 37 (2018), pp. 163-174.

[6] Y. Furukawa, B. Curless, S.M. Seitz, R. Szeliski Towards internet-scale multi-view stereo Computer Vision and Pattern Recognition, IEEE (2010), pp. 1434-1441

[7] C. Koch, S.G. Paal, A. Rashidi, Z. Zhu, M. Koenig, I. Brilakis Achievements and challenges in machine vision-based inspection of large concrete structures Adv. Struct. Eng., 17 (3) (2014), pp. 303-318.

[8] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14194–14203, 2021.

[9] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. British Machine Vision Conference (BMVC), 2020.

[10] Yu A, Guo W, Liu B, et al. Attention aware cost volume pyramid based multi-view stereo network for 3d reconstruction[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 175: 448-460.

[11] Jensen R, Dahl A, Vogiatzis G, et al. Large scale multi-view stereopsis evaluation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 406-413.