End-to-end Learned Multi-View Stereo Reconstruction

Zhisheng Zheng
Technical Univeristy of Munich
finn.zheng@tum.de

Dongyue Lu
Technical University of Munich
dongyue.lu@tum.de

1. Introduction

Multi-view stereopsis is a classic computer vision problem which aims to reconstruct a 3D object given a set of different images with corresponding camera parameters. Many pioneering works have been proposed in this field and a standard pipeline has also been developed [1]. Recent years, convolutional neural network (CNN) has achieved profound impact in this filed. However, it is difficult for CNN to extract long-distance features between global data. Transformer [2] can effectively learn the correlation between long-term features and the self-attention mechanism of transformer can produce more interpretable models, so it has been successfully applied in computer vision. Therefore, in this project, we wish to propose a deep learning framework for end-to-end multi-view stereo reconstruction integrated with transformer module to further improve the performance of current methods.

2. Related Work

SurfaceNet [3] is a pioneering end-to-end learning framework for multi-view stereopsis. Given a set of view pairs of images, it uses a novel space representation structure colored voxel grid and leverages 3D convolutional neural network to predict 3D surface occupancy directly. Jiaming Sun et.al [4] use a feature back-projection approach with a special fragment bounding volume to further extend SurfaceNet. TransformerFusion [5] is a transformer-based end-to-end architecture that can fuse features from different video frames. A transformer module is helpful for attending to the most informative features from each image and generating a decent dense occupancy field. In our work. we leverage transformer on a different perspective by applying attention mechanism in terms of 3D voxel to obtain better feature representation in 3D space.

3. Method

Figure 1 shows the architecture of our network. For a set of images from different viewpoints, we first use an image encoder to extract features, then back-project these features to our generated cubes. To reduce computational cost, a set of 3D CNN module is applied to decrease the size of the cube. Next step, the cube is divided into small cubes and feed into a shared weight transformer module for information exchange. At last we assemble them and use a final 3D CNN module to obtain our final result - the surface occupancy of each voxel.

3.1. Network details and verification steps

Following the method of SurfaceNet and Neuralrecon [4], we need to back-project the feature extracted from images to voxel cubes correctly. After being inputted into CNN, the size of the feature map will be different from the original picture. In order to ensure that the corresponding spatial area and the feature of the image area are consistent, we need to find the accurate receptive field for each feature. In addition, the wrong back-projection will cause sparse or even empty cubes, which leads to meaningless training. For debug and verification we can back-project depth map or RGB values and visualize each step. We plan to integrate the classic transformer module into our network. The structure of DETR [6] will also be a reference. After the back-projection of extracted feature, we divide the feature cube into small cubes, concatenate the features in each small cube and feed them into a shared weight transformer. For position embedding we use a multidim positional encoding method which is a extension of [7]. The output of it will be assemble back to a cube and feed into 3D CNN for final surface occupancy prediction. The final output of our network shall be voxel-grids in which each voxel has a occupancy score value, indicating the confidence of a voxel being on the surface or not.

3.2. Implementation details

We plan to perform the experiments on two possible datasets, ScanNet [8] and Matterport3D [9]. Both datasets contain thousands of indoor scenes with ground-truth camera poses, surface reconstructions and semantic segmentation labels. In oder to supervise the training effectively, voxelization of ground truth data to get an one-to-one matching with our generated cubes is a must. We use binary cross entropy to supervise the training process. For evaluation,

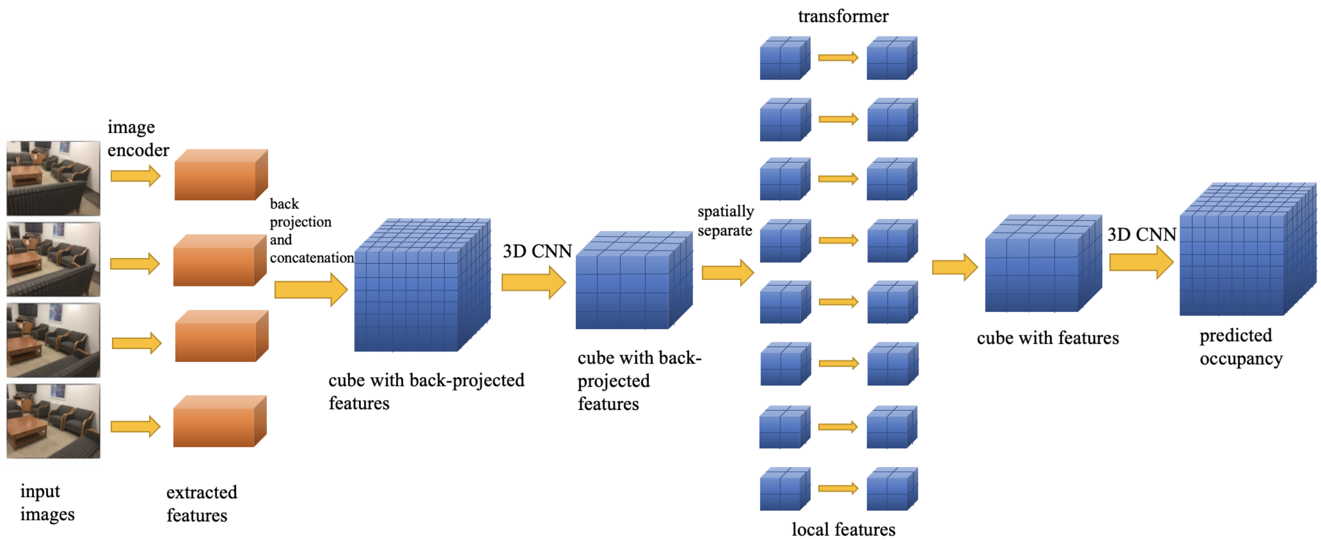

Figure 1. Proposed architecture

we will use Completeness, Accuracy [10] and F-Score [11] as our metrics. To better evaluate our model, we select SurfaceNet, DPSNet [12] and MVSnet [13] as comparison. For ablation study we plan to remove or replace Transformer module and the back-projection step to prove the validity of our model.

4. First Step

Because we follow the pipeline of SurfaceNet, re-implementing the basic architecture of SurfaceNet is a good start for our future work. First we need to back-project RGB value to voxel cube correctly. We shall apply perspective projection to calculate the image coordinate for each voxel in cube, then assign the corresponding RGB value to it. If the obtained coordinate doesn't lie on the image, we set the voxel value to zero. To get a one-to-one matching between GT data and the output of model. Voxelization of GT data is needed. We plan to study the source code of NeuralRecon, because they also use occupancy value as one element of their output. Then we rewrite the basic architecture of SurfaceNet using pytorch and train this model to see if it works well.

5. Time Schedule

Date	Task
27.10 - 10.11	Re-implement of SurfaceNet
10.11 - 17.11	Feature Extraction/Projection
17.11 - 08.12	Transformer
08.12 - 12.01	Model Tuning
12.01 - 26.01	Model Experiments
26.01 - late February	Final Report/Presentation

References

[1] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779. Springer, 2008. 1

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[3] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2307–2315, 2017. 1

[4] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 1

[5] Aljaž Božič, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *arXiv preprint arXiv:2107.02191*, 2021. 1

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 1

[7] Zelun Wang and Jyh-Charn Liu. Translating math formula images to latex sequences using deep neural networks with sequence-level training, 2019. 1

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5828–5839, 2017. 1

[9] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1

[10] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 2

[11] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 2

[12] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019. 2

[13] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 2