Final Report - 3D Point Cloud Completion Using Improved MSN with Novel SoftPool++ Architecture

Jiongyan Zhang Zichen Zhang Chengzhi Shen
Technical University of Munich

Abstract

The point cloud is a common data structure when dealing with 3D computer vision and graphics, and deep learning-based methods to tackle point cloud-related problems are popular choices these days. This paper aims at solving the task of 3D point cloud completion. The architecture proposed in this paper is an improved version of the Morphing and Sampling Network(MSN) using the SoftPool-truncated features, local convolution, and global feature transform. The result shows improvement both in loss criteria and convergence speed, which proves the effectiveness of the SoftPool operator with the truncation strategy, local convolution, and the global feature transform.

1. Introduction

There are several data structures for 3D shapes, and there is no doubt that the point cloud is an important one among them. The paper focuses on point cloud completion, which is one of the popular domains regarding point cloud processing that is crucial for many applications, such as the perception of autonomous vehicles. The main target of the point cloud completion is to complete the geometry of the incomplete point cloud obtained by a partial scan. The completion process includes both global structure reconstruction and local fine detail preservation. So this requires the point cloud completion network to be able to learn both global and local features.

The network proposed in this paper is an end-to-end deep learning framework, and it is an improved version of the Morphing and Sampling Network(MSN) [1] - one of the state-of-the-art frameworks dealing with the point cloud completion task. The input is a set of incomplete point clouds representing the partial scan of an object, whereas the output is the complete point clouds of the object. Since our network is an improvement of the MSN, we do not expect it to have significantly more parameters than the original MSN to compare their performance. Therefore when designing the network, we strictly restrict the new network's number of parameters.

In this paper our main contributions are listed as below.

- We reorganize the MSN codebase as the baseline model. Additionally we add a transformation network that is similar to the Spatial Transformer Network [2] to the encoder part of the MSN.
- We use the novel SoftPool-truncated features [3] instead of the original PointNet [4]-generated features, and local convolution designed by us to modify the encoder of the MSN.
- We add a global feature transform as a skip connection to the first part of the MSN to preserve and fuse information, this idea is inspired by the architecture of SoftPool++ [3].

The improved architecture proposed in this paper shows better performance and faster convergence within 10 epochs of training, which proves the effectiveness of the improvements. Our code is released here: https://github.com/hinczhang/Machine-Learning-for-3D-Geometry

2. Related Work

There have been a lot of exceptional deep learning architectures dealing with point cloud processing for the past several years. PointNet [4] is a classical architecture that can tackle several point cloud processing tasks. It can create a permutation invariant feature for point clouds. FoldingNet [5] deforms a 2D grid with multi-layer perceptrons to perform object completion. AtlasNet [6] and PCN [7] are more complicated networks for reconstruction. MSN [1] improves the point cloud completion using the novel morphing and sampling idea. PointNet++ [8] adds the farthest point sampling and grouping to the network for hierarchical processing. PMP-Net [9] performs the object completion from the observed regions to the nearest occluded regions. GRNet [10] further leverage both the point cloud and the voxel grid. This architecture voxelizes the point cloud and processes it. Finally, it converts the output back to the point

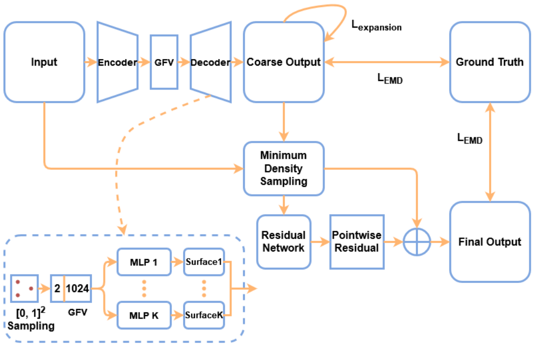

Figure 1. The MSN architecture

cloud representation. VE-PCN [11] achieves an improvement in the completion by adding a supplement to the features of the decoder in the volumetric completion.

3. Method

Our architecture is an improved version of the MSN [1]. In this section, details of all parts of the project will be illustrated in the following subsections.

3.1. Morphing and Sampling Network

The MSN architecture in Fig. 1 takes the input incomplete point cloud and passes it through an encoder-decoder architecture to output a coarse completion of the point cloud.

The decoder part consists of a unit square sampling and K multi-layer perceptrons (MLPs). This decoder learns a mapping that morphs the 2D points in a 2D unit square to the 3D points on a 3D surface using these MLPs. N points are sampled in a unit square and concatenated with the encoded feature vector to pass through the MLPs. Finally, each sampled 2D point is mapped to K 3D points on the K different surfaces.

The output of the decoder is the coarse completion of the point cloud that will be merged with the input incomplete point cloud. The authors of MSN apply the minimum density sampling (MDS) to this merged result. And this strategy will obtain an evenly distributed subset point cloud. A residual network will take this point cloud for point-wise residual prediction, whose output will be added to the previous point cloud to get the final complete point cloud.

3.2. Encoder

The encoder of the MSN is a simplified version of the PointNet [4]. The feature transform in the PointNet is removed in the MSN paper, but we decide to add it back. The architecture is as Fig. 2.

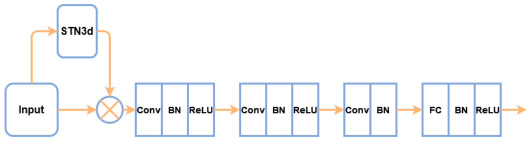

Figure 2. The encoder

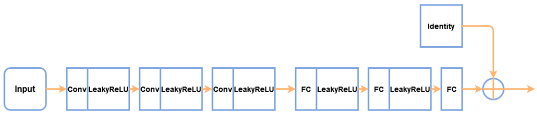

Figure 3. The STN3d architecture

The feature transform here is named STN3d, which consists of several convolutional layers and fully-connected layers. Its structure is as Fig. 3. This network takes the input point cloud and outputs a 3×3 matrix representing a rotation matrix for the point cloud.

Apart from the above modifications, the major contribution to the encoder is the SoftPool-truncated features and the local convolution.

3.3. SoftPool-Truncated Features

To replace the original pooling methods to reduce the information loss while generating the feature map, researchers have proposed a novel pooling operator - SoftPool [12]. It is an exponentially weighted activation downsampling to preserve the information in the reduced activation maps. Furthermore, we try to use the SoftPool-truncated features [3] to deliver the fine-grained geometric details to obtain a better consequence as Fig. 4. This method mainly consists of three operations: SoftPool with truncation strategy, local convolution, and reshaping. The SoftPool with truncation strategy will sort the rows of the feature maps and select the rows with the highest values in the feature maps. Then these selected feature rows will be truncated and merged together. This operation reduces the computation cost without the loss of key information.

3.4. Local Convolution

Because each point from the point cloud is fed into the network independently, the rows of the feature maps remain independent. Besides, since the current feature maps are formed by merging the truncated pieces from the original feature maps, we propose the local convolution method using different convolutional kernels for sliced feature rows from different previous feature maps.

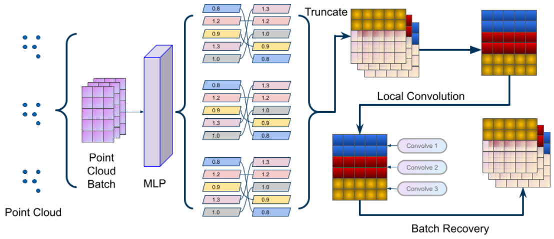

Figure 4. The SoftPool-truncated features and local convolution

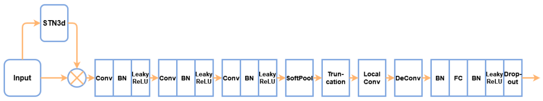

Figure 5. The improved encoder

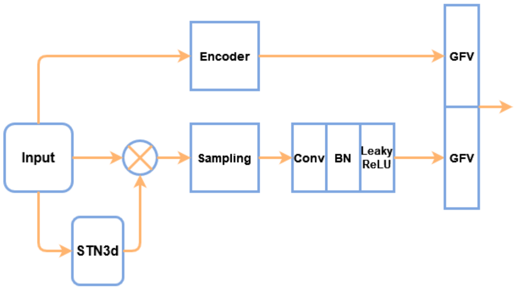

Figure 6. The global feature transform

3.5. The Improved Encoder

The architecture of the improved encoder with the mentioned modifications is shown in Fig. 5.

3.6. Global Feature Transform

The global feature transform shown in Fig. 6 acts as a skip connection to the encoder part. The encoder solely encodes the feature vector in the original MSN architecture. In our new architecture, the output from the encoder and the global feature transform will be fused together to form the encoded feature vector.

3.7. Loss Functions

The loss function we use in this paper is the same as the one in the original MSN paper. It is a linear combination of two Earth Mover's Distance (EMD) and one expansion loss. The EMD loss is a similarity metric to measure the similarity between two point clouds. And the expansion loss acts as a regularizer to ensure that the surface elements are compact.

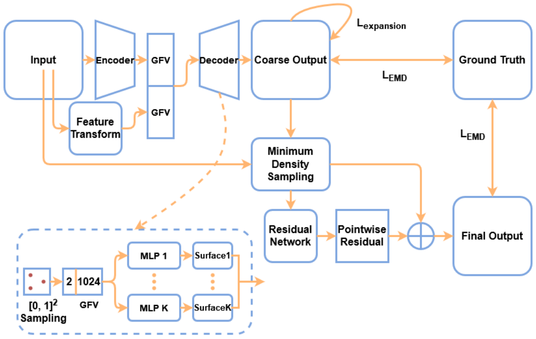

Figure 7. The entire architecture

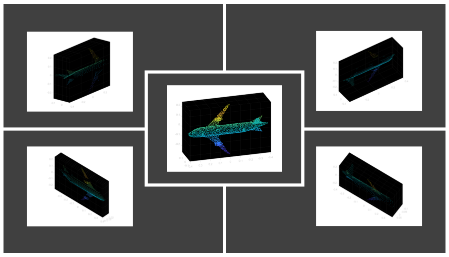

Figure 8. Data generation

$$L = L_{EMD}(S_1, S_{gt}) + \alpha L_{expansion} + \beta L_{EMD}(S_2, S_{gt}) \quad (1)$$

S_1 denotes the coarse point cloud completion, S_2 denotes the complete point cloud, and S_{gt} denotes the ground truth point cloud. α is 0.1 and β is 1.0.

3.8. Entire Architecture

The entire architecture is shown in Fig. 7.

4. Experiment

4.1. Dataset Configuration

We use the ShapeNet dataset (Core.v1) [13] as our original 3D model source. This dataset contains diverse 3D objects that provide rich cases for training. To prepare our experiment data, we use the blender to take 50 random poses for each 3D object and record their depth images in the OpenEXR format. After that, we regenerate the partial 3D models corresponding to the original complete models as the data that we will use for the project. An example is shown in Fig. 8. The dataset sizes for training, validation, and testing are 7000, 700, and 50 respectively.

4.2. Experiment Result

For the experiment, the hyper-parameters are as in Tab. 1. After training the baseline model (MSN) and our model, we obtain the loss curves shown in Fig. 9.

Hyper-Parameters	Values
Batch Size	64
Points Number	4096
Primitive Number	16
Epoch	10
Learning Rate	0.001
Optimizer	Adam

Table 1. The hyper-parameters

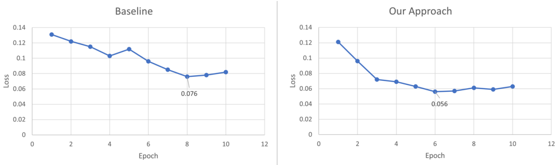

Figure 9. The loss curves

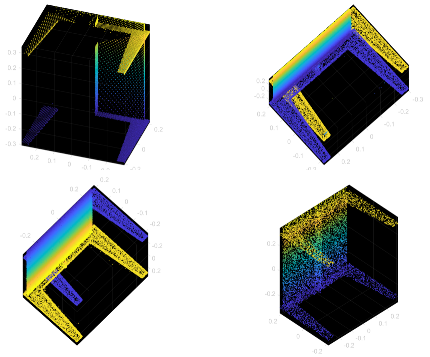

Figure 10. The visualization of the point clouds, the upper left is the input, the upper right is the result of the MSN, the lower left is the result of our approach, the lower right is the ground truth

The parameter number of MSN is 30.2M, while our approach is 33.7M. It can be concluded from the loss curves that our approach has faster convergence and better validation loss within the first ten epochs. Due to the limitation of hardware resources, the limit performance of the model is not tested. However, our approach gets a quick overfit as well. The test result shows that the loss is 0.083 for MSN and 0.074 for our approach.

4.3. Result Visualization

The visualization is shown in Fig. 10. It shows that the MSN and our model can reach approximately the same visual effect.

5. Conclusion

We perform some improvements to the original MSN architecture. We apply the STN3d, SoftPool-truncated features, and local convolution to modify the encoder. Besides, we add a global feature transform to the first part of the MSN for information fusion. The result of the experiment shows that our model achieves faster convergence and better validation loss within the first ten epochs. Regarding future works, improvements in loss functions, sampling strategy, and ablation studies can be performed.

References

[1] Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11596–11603, 2020. 1, 2

[2] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 1

[3] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Softpool++: An encoder–decoder network for point cloud completion. *International Journal of Computer Vision*, 130(5):1145–1164, 2022. 1, 2

[4] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1, 2

[5] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018. 1

[6] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 1

[7] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pages 728–737. IEEE, 2018. 1

[8] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1

[9] Xin Wen, Peng Xiang, Zhizhong Han, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Yu-Shen Liu. Pmp-net: Point cloud completion by learning multi-step point moving paths. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7443–7452, 2021. 1

[10] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun. Grnet: Gridding residual network for dense point cloud completion. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020. 1

[11] Xiaogang Wang, Marcelo H Ang, and Gim Hee Lee. Voxel-based network for shape completion by leveraging edge generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13189–13198, 2021. 2

[12] Alexandros Stergiou, Ronald Poppe, and Grigorios Kalliatakis. Refining activation downsampling with softpool. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10357–10366, 2021. 2

[13] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3