IEICE TRANS. INF. & SYST., VOL.E105–D, NO.2 FEBRUARY 2022

LETTER

Learning Pyramidal Feature Hierarchy for 3D Reconstruction

Fairuz Safwan MAHAD[†a)], ***Nonmember***, **Masakazu IWAMURA**[†b)], ***Senior Member***, ***and*** **Koichi KISE**[†c)], ***Fellow***

SUMMARY Neural network-based three-dimensional (3D) reconstruction methods have produced promising results. However, they do not pay particular attention to reconstructing detailed parts of objects. This occurs because the network is not designed to capture the fine details of objects. In this paper, we propose a network designed to capture both the coarse and fine details of objects to improve the reconstruction of the fine parts of objects.

key words: *computer vision, 3D reconstruction, deep learning, multi-view*

1. Introduction

Three-dimensional (3D) reconstruction using RGB images has been widely researched in the computer vision community. 3D reconstruction has a wide range of applications in various fields, such as the medical sector [1], archaeological sector [2], and civil engineering sector [3]–[5]. However, reconstructing a highly accurate 3D model remains a challenging task, even with a state-of-the-art 3D reconstruction method. Soltani et al. proposed a method to reconstruct a high-resolution 3D model that inputs multiple depth maps or silhouettes [6], which has been shown to outperform other 3D reconstruction methods in terms of multi-viewpoint cloud-based methods [7]. However, the method ignores an important aspect, which is the reconstruction of detailed parts of the 3D model. In most cases, the method fails to reconstruct the fine parts of the 3D model, as shown in Fig. 1. The fine parts are either slightly reconstructed or not reconstructed at all. Slightly reconstructed means that the reconstruction is incomplete or lacks density. The method fails particularly in the reconstruction of fine parts, such as the legs of a chair, and the grip and tip of a rifle. Our research aims to improve the quality of 3D models, specifically focusing on reconstructing detailed parts of 3D models, which have been ignored by most 3D reconstruction methods.

In this paper, we propose a simple but effective approach to improve the reconstruction of detailed parts of 3D models using multiple viewpoints*. Our proposed method is based on a state-of-the-art method [6] and introduces the pyramidal hierarchical-based network concept from [9], which was originally designed for object detection. In [9], a multi-scale feature map was built, where each feature map consisted of high-level semantic features with different spatial resolutions. By leveraging semantically strong features extracted at different levels, the reconstruction quality of detailed parts of the 3D model was improved. We demonstrate that the mechanism of the pyramidal hierarchical-based network is effective in enhancing the quality of the reconstructed 3D model, as shown in Fig. 1.

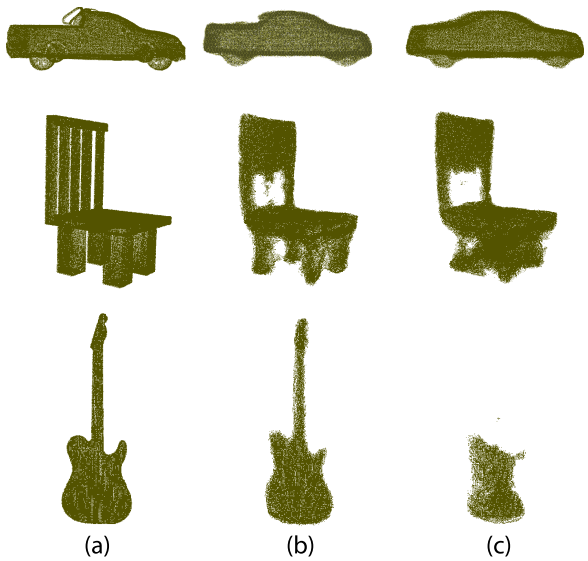

Fig. 1 Comparison of reconstructed 3D models using our proposed method and that of Soltani et al. [6] (a) Ground truth. (b) Proposed method. (c) Method of Soltani et al. [6]. Our method improved the reconstruction of detailed parts using both local and generic features.

2. Approach

2.1 Soltani et al.'s Method [6]

Soltani et al. proposed a network that can run in three settings. We used the simplest setting, in which the network takes in 20 depth map images, and produces a set of 20 depth map images and 20 silhouette images. The output from the network, that is, a total of 40 images, is used to render the final 3D model. The 20 depth map images are projected back to 3D space to create an initial 3D model. The 3D model is further refined using the silhouette images to filter outliers.

*This approach is based on the method in [10].

Manuscript received October 7, 2021.
Manuscript publicized November 16, 2021.

†The authors are with Dept. of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai-shi, 599–8531 Japan.
a) E-mail: fsafwan88@gmail.com
b) E-mail: masa@cs.osakafu-u.ac.jp
c) E-mail: kise@cs.osakafu-u.ac.jp

DOI: 10.1587/transinf.2020ZDL0001

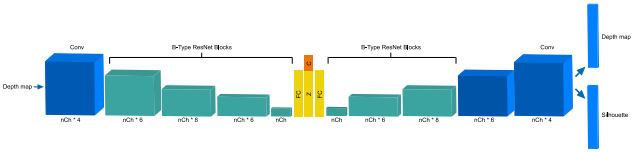

Fig. 2 Pipeline structure of the method proposed by Soltani et al. [6] nCh denotes the number of channels (74). FC denotes the fully connected layer. Z and C are the latent variable and a class label, respectively. C is not used in this paper.

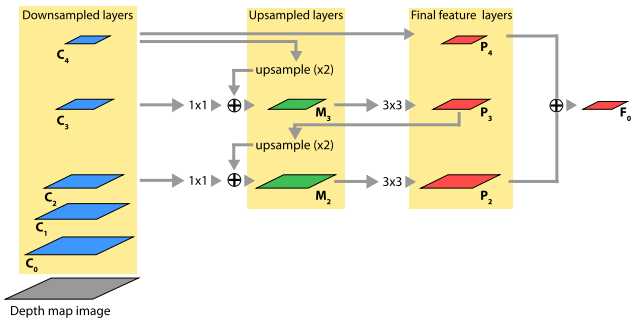

Fig. 3 Our proposed network architecture.

An overview of the pipeline proposed by Soltani et al. is shown in Fig. 2. The core of its network structure is a deep generative network and variational autoencoder (VAE) for both its encoder and decoder, which results in a high-resolution 3D model.

2.2 Proposed Method

Our proposed method is based on that of Soltani et al. Although the method of Soltani et al. can reconstruct a high-resolution 3D model, the network was designed to learn only single-level features. Thus, most detailed parts are ignored during reconstruction. To further enhance the quality of the reconstructed 3D model, our strategy is to focus on capturing features at tight spots, such as small and thin parts of the object. This will improve the quality of the reconstructed 3D model. Thus, we implement a multi-scale layered network with skip connections inspired by the method in [9] in the encoder part of the VAE network structure used in [6].

According to Lin et al. [9], multi-scale upsampled layers consist of semantically stronger features than downsampled layers, which is the main advantage of using pyramidal-based networks. Our proposed network architecture is shown in Fig. 3. It features bottom-up and top-down pathways with skip connections. The network starts with a bottom-up pathway by scaling down the input image to feature maps of sizes $\{110^2, 53^2, 25^2, 11^2, 4^2\}$, which are denoted by $\{C_0, C_1, C_2, C_3, C_4\}$, respectively. The layers M_3 and M_2 are obtained by concatenating features from their respective previous layer and the layer from the current level. The final feature layers are denoted by $\{P_2, P_3, P_4\}$. $\{P_2, P_4\}$ are concatenated to produce a final merged feature map denoted by $\{F_0\}$. $\{P_3\}$ is not included to reduce the number of parameters. The network in [6] can be trained either in an unsupervised manner or conditionally. For our purpose, we train our network using the unsupervised method. Therefore, we use the same loss function as Eq. (1) in [6]. Our proposed method adopts a similar concept to that in [9] with two distinct differences:

1. Lin et al. [9] considered every output level $\{P_1, P_2, P_3, P_4\}$ independently, whereas in our proposed network, we concatenate the final feature maps $\{P_2, P_4\}$ to produce a final merged feature map followed by a fully connected layer.
2. We implement our proposed network in the encoder of the VAE structure.

3. Experiments

In this section, we evaluate our method against that of Soltani et al. [6]. We obtained our results in both qualitative and quantitative evaluations.

3.1 Experimental Settings

We trained our model on the ShapeNet dataset [8], which consists of 57 object categories, with a total of 56,652 3D models. To use the dataset for training, we rendered depth maps from all 3D models in the dataset using fixed camera angles. The rendered size was 224×224 pixels. To evaluate our method against that in [6], we used the same dataset distribution as [6], which was 92.5% for training and the remaining 7.5% for validation. We used the pre-trained dataset, which was the same model used to produce the results in [6], along with the authors' original source code. We trained our model on an NVIDIA GeForce GTX TITAN V.

3.2 Qualitative Evaluation

Figure 4 shows our results against those of the method of Soltani et al. [6]. We reconstructed the 3D models using the 7.5% test dataset that was set aside. We observed two main points in Fig. 4. First, our method captured the overall shape of the object better than the method of Soltani et al. This is because of the fusion of features across multi-scale layers. Second, our method was better at preserving the thin parts of the object. The method of Soltani et al. could not reconstruct the thin parts of the object in most cases. We designed our network architecture to use features across multiple levels. The features in upsampled layers are semantically stronger than those in downsampled layers. Because of this, there are many more meaningful features for both global and local features. However, the features in upsampled layers are weak in terms of localization. By contrast, the features in downsampled layers are not as semantically strong as those in upsampled layers, but they have better localization. By associating semantically stronger features with weak localization in upsampled layers with semantically weaker features with strong localization in downsampled layers at each level, the network can simultaneously learn both the overall

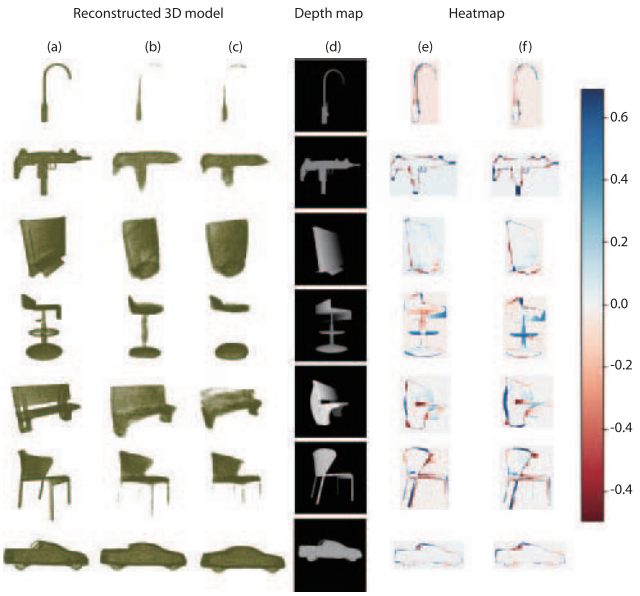

Fig. 4 Qualitative results for successful cases. Comparison of 3D reconstructed models and heatmaps between our proposed method and that of Soltani et al. [6]. (a) Ground truth 3D model. (b) Reconstructed 3D model using the proposed method. (c) Reconstructed 3D model using the method of Soltani et al. [6]. (d) Depth map ground truth. (e) Heatmaps of the proposed method. (f) Heatmaps of the method of Soltani et al. [6].

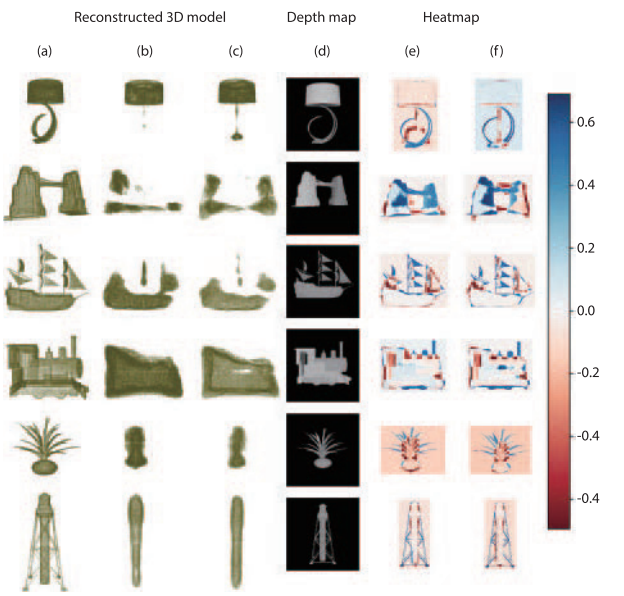

Fig. 5 Qualitative results for failed cases. Comparison of 3D reconstructed models and heatmaps between our proposed method and that of Soltani et al. [6]. (a) Ground truth 3D model. (b) Reconstructed 3D model using the proposed method. (c) Reconstructed 3D model using the method of Soltani et al. [6]. (d) Depth map ground truth. (e) Heatmaps of the proposed method. (f) Heatmaps of the method of Soltani et al. [6].

shape and detailed parts of an object. These factors lead to a much more complete reconstructed 3D model compared with that resulting from the method of Soltani et al. By contrast, our method is not perfect. It failed to reconstruct uncommon and complex shapes, as shown in Fig. 5. This may have been caused by a lack of training data. For example, the topmost lamp in the figure had a particular shape that was not included in the training data.

3.3 Quantitative Evaluation

In this section, we discuss our results quantitatively using the mean intersection-over-union (IoU). We reconstructed all the 3D models in the test dataset in point cloud form and computed the IoU by converting the point clouds into 3D voxels. We trained a neural network model on all 57 object classes five times. In Table 1, we tabulate our results for each class and compare them with those of the method of Soltani et al. [6]. We also present the standard deviation for each category over the five runs to indicate that the results obtained were stable and not coincidental. Hence, the table shows the mean IoU and its standard deviation. Note that the average IoU was reported as 84.0 in [6]. However, after training and running it for five times using the original source code provided by the authors, we achieved an average IoU of 81.1. Therefore, we used this value as our benchmark for the method of Soltani et al. [6]. Table 1 shows the best results for each method. The table shows that our method outperformed the method of Soltani et al. in 45 out of 57 categories. Our method performed better in categories such as lamp, guitar, table, pistol, microphone, and chair because most of the 3D models in those categories contained thin structures, which gave a clear advantage to our method. In particular, in the categories such as lamp, bookshelf, camera, and vase, despite their complex shapes, our method coped with the shape complexity better than the method of Soltani et al. [6].

Additionally, to measure the significance of our proposed method against that of the method of Soltani et al. [6], we ran a statistical test: the Student's T-test. The test produces a P-value; if the P-value is less than 0.05, this indicates that the experiment is statistically significant. Our experiment achieved a P-value of 0.0301, which proves that our proposed method was statistically significant.

3.4 Limitation

Although our method improved the 3D reconstruction of the overall shape, and also detailed parts, to a certain extent, it did not perform well in categories such as knife, rocket, and cap. These categories had simple-shaped objects, and most of them looked similar to each other. Additionally, a weakness of depth map-based 3D reconstruction is that it does not capture thin parts of objects well; this claim was supported in [7]. Despite the improvements made in the reconstruction of certain thin and detailed parts, the limitation of depth map-based 3D reconstruction undermined the effectiveness and capabilities of our method.

4. Conclusion

In this paper, we proposed a simple yet effective method

Table 1 Comparison of the method of Soltani et al. [6] and our proposed method. "#M" represents the number of models. "Accuracy ± SD" represents the accuracy in terms of IoU followed by its standard deviation value over five runs. "Gain" represents the improved accuracy of our proposed method against that of Soltani et al.

Category	#M	Accuracy ± SD		Gain
		Soltani [6]	Ours	
aeroplane	304	71.8 ±0.002	**72.7** ±0.001	0.9
bag	8	78.8 ±0.006	**80.6** ±0.008	1.8
basket	11	83.6 ±0.002	**84.5** ±0.003	0.9
bathtub	62	87.3 ±0.006	**88.3** ±0.002	1.0
bed	10	76.3 ±0.007	**77.6** ±0.006	1.3
bench	132	78.1 ±0.007	**78.6** ±0.001	0.5
bicycle	6	43.1 ±0.005	**43.5** ±0.001	0.4
birdhouse	2	82.8 ±0.009	**84.4** ±0.002	1.6
bookshelf	39	73.3 ±0.007	**74.2** ±0.002	0.9
bottle	35	92.5 ±0.006	**93.0** ±0.0008	0.5
bowl	13	93.5 ±0.005	**93.9** ±0.002	0.4
bus	82	91.1 ±0.005	**91.4** ±0.002	0.3
cabinet	115	89.2 ±0.011	**91.1** ±0.001	1.9
camera	8	67.7 ±0.009	**69.2** ±0.005	1.5
can	8	93.9 ±0.007	**94.8** ±0.004	0.9
cap	3	**81.1** ±0.023	77.6 ±0.003	−3.5
car	554	82.2 ±0.003	**82.7** ±0.000	0.5
cellphone	45	91.8 ±0.005	**92.3** ±0.005	0.5
chair	491	77.0 ±0.003	**78.4** ±0.005	1.4
clock	40	**79.8** ±0.009	79.6 ±0.006	−0.2
dishwasher	8	93.9 ±0.002	**94.1** ±0.009	0.2
display	81	86.5 ±0.001	**86.6** ±0.001	0.1
faucet	47	**66.1** ±0.007	65.6 ±0.002	−0.5
filecabinet	18	**91.9** ±0.007	91.8 ±0.008	−0.1
flowerpot	42	65.3 ±0.005	**65.7** ±0.001	0.4
guitar	65	78.5 ±0.004	**81.5** ±0.006	3.0
headphone	5	55.7 ±0.010	**60.7** ±0.006	5.0
helmet	16	75.2 ±0.005	**75.6** ±0.0008	0.4
keyboard	5	87.6 ±0.005	**88.0** ±0.001	0.4
knife	42	**80.4** ±0.025	78.0 ±0.006	−2.4
lamp	181	68.2 ±0.008	**68.3** ±0.004	0.1
laptop	34	97.0 ±0.003	**97.1** ±0.001	0.1
letterbox	7	**71.2** ±0.008	70.0 ±0.017	−1.2
microphone	6	62.9 ±0.003	**63.6** ±0.001	0.7
microwave	11	**93.7** ±0.002	93.2 ±0.005	−0.5
motorcycle	28	**75.7** ±0.006	75.4 ±0.002	−0.3
mug	17	**84.3** ±0.002	**84.3** ±0.003	0.0
piano	13	79.4 ±0.004	**80.9** ±0.005	1.5
pillow	6	**86.7** ±0.001	86.6 ±0.007	−0.1
pistol	19	84.9 ±0.004	**85.4** ±0.001	0.5
printer	18	79.4 ±0.002	**80.9** ±0.009	1.5
remote control	4	89.4 ±0.007	**89.7** ±0.001	0.3
rifle	171	77.5 ±0.005	**77.8** ±0.005	0.3
rocket	7	**73.2** ±0.002	71.3 ±0.006	−1.9
ship	147	79.6 ±0.002	**79.7** ±0.001	0.1
skateboard	18	80.7 ±0.005	**81.4** ±0.001	0.7
sofa	242	87.1 ±0.003	**87.5** ±0.001	0.4
speaker	121	84.2 ±0.002	**84.4** ±0.0008	0.2
stove	8	88.5 ±0.002	**91.1** ±0.004	2.6
table	652	84.8 ±0.002	**85.2** ±0.0009	0.4
telephone	92	92.6 ±0.002	**92.7** ±0.001	0.1
tower	12	**76.6** ±0.004	76.2 ±0.001	−0.4
train	25	84.7 ±0.007	**85.1** ±0.002	0.4
trashcan	28	85.3 ±0.002	**85.4** ±0.0005	0.1
vase	38	82.4 ±0.003	**83.2** ±0.004	0.8
vessel	85	**81.5** ±0.01	80.6 ±0.004	−0.9
washing machine	17	92.4 ±0.003	**92.7** ±0.002	0.3
Average		**81.1**	**81.5**	0.4

to improve the reconstruction of detailed parts of an object, particularly thin parts. To enable the network to focus on learning detailed parts in addition to the overall shape of the object, we designed a network that uses multi-scale layers to learn and merge features of different scales. We compared our results with those of a state-of-the-art method (Soltani et al. [6]) in both qualitative and quantitative evaluations. Our results demonstrated that our method outperformed the state-of-the-art method [6] in most cases and that they are statistically significant. Our method improved the reconstruction of thin parts of objects that, in most cases, could not be reconstructed by the state-of-the-art method [6]. This led to the improvement of accuracy. Our qualitative results also demonstrated that the 3D models reconstructed by our method were more complete and resembled the ground truth more than those reconstructed by the state-of-the-art method [6].

References

[1] L. Humbert, J.A. De Guise, B. Aubert, B. Godbout, and W. Skalli, "3D reconstruction of the spine from biplanar X-rays using parametric models based on transversal and longitudinal inferences," Medical Engineering & Physics, vol.31, no.6, pp.681–687, 2009.

[2] F. Bruno, S. Bruno, G. De Sensi, M.-L. Luchi, S. Mancuso, and M. Muzzupappa, "From 3D reconstruction to virtual reality: A complete methodology for digital archaeological exhibition," Journal of Cultural Heritage, vol.11, no.1, pp.42–49, 2010.

[3] E. Kwak, I. Detchev, A. Habib, M. El-Badry, and C. Hughes, "Precise photogrammetric reconstruction using model-based image fitting for 3D beam deformation monitoring," Journal of Surveying Engineering, vol.139, no.3, pp.143–155, 2013.

[4] X. Brunetaud, L. De Luca, S. Janvier-Badosa, K. Beck, and M. Al-Mukhtar, "Application of digital techniques in monument preservation," European Journal of Environmental and Civil Engineering, vol.16, no.5, pp.543–556, 2012.

[5] Y. Ham and M. Golparvar-Fard, "Three-dimensional thermography-based method for cost-benefit analysis of energy efficiency building envelope retrofits," Journal of Computing in Civil Engineering, vol.29, no.4, B4014009, 2014.

[6] A.A. Soltani, H. Huang, J. Wu, T.D. Kulkarni, and J.B. Tenenbaum, "Synthesizing 3D shapes via modeling multi-view depth maps and silhouettes with deep generative networks," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1511–1519, 2017.

[7] X.-F. Han, H. Laga, and M. Bennamoun, "Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era," IEEE Trans. Pattern Anal. Mach. Intell., vol.43, no.5, pp.1578–1604, 2021.

[8] A.X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An information-rich 3D model repository," arXiv preprint arXiv:1512.03012, 2015.

[9] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.936–944, 2017.

[10] F.S. Mahad, M. Iwamura, and K. Kise, "Leveraging pyramidal feature hierarchy for 3D reconstruction," International Workshop on Frontiers of Computer Vision, pp.347–362, Springer, 2020.