Single-view 3D Reconstruction Supported by Classification

Chenyang Li chenyang.li@tum.de
Jiaye Yang jiaye.yang@tum.de
Jiaping Zhang penny.zhang@tum.de
Yihao Wang johanna.wang@tum.de

Abstract

Reconstructing shape from a single-view image has always been an interesting task for both the industrial and academic communities. The work from Wallace in 2019 has shown a paradigm for such tasks: extracting global and local information for the reconstruction. While Wallace's work manages to deliver single-view reconstructions, the training step requires a 3D shape prior, which limits its potential usage in more scenarios. Our contributions are 1) replicating the work from Wallace in PyTorch and 2) designing a new network that can outperform the baseline without the need to train on 3D shapes. Our experiments show that our model can outperform the baseline significantly when training on all 13 classes from the ShapeNet. Codes can be found on GitLab [1].

1. Introduction

Reconstructing 3D models based on a single image is a challenging topic. Wallace et al. [1] use a single-view image to capture local features and refines the average 3D shape of its category to extract global features. With this method, they obtain relatively good 3D reconstruction results. Since global features can also be extracted in a classification task, we can add a classification task to force the network to extract global features from the image to get rid of the 3D shape as prior. Therefore, we can use the same single-view image to extract both global and local features. In a word, we aim to realize 3D reconstruction based on a single-view image supported by a classification network. In this project, we completed the following tasks:

1) Transforming TensorFlow to PyTorch;
2) Replicating the baseline model;
3) Implementing the model we proposed;
4) Comparing part of the results of two models.

2. Baseline

The 3D reconstruction problems are commonly solved based on 3D scans, multi-view RGB images or retrieval methods. In our baseline: Few-Shot Generalization for Single-Image 3D Reconstruction via Priors [1], the authors propose a dual-input method that jointly learns from a single view and a class-specific prior and refines the 3D shape via the input image. The high-resolution single-view images provide detailed local features, while the shape priors entail global information. Their approach is shown in Figure 1.

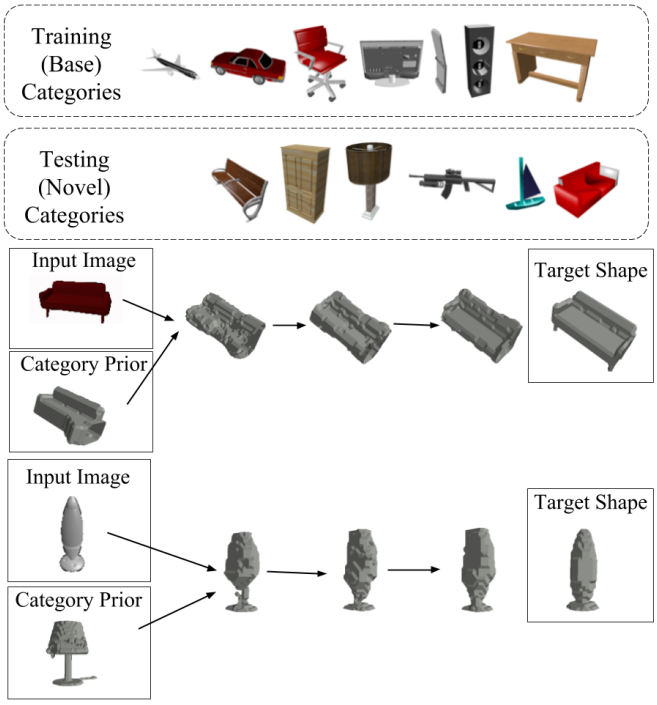

Figure 1. Approach of baseline [1]. It jointly learns the features of a single-view image and a shape prior.

This network of our baseline combines two state-of-the-art methods (see Figure 2). The first input is a single-view image encoded via the category-agnostic encoder of 3D-R2N2 [2]. The second input is the category-specific prior, which is the average voxel grid of the category. Then, the two embeddings are fed into the shape generator to produce the refined 3D shape. The second encoder and shape generator adopts the similar structure of Yang et al. [3].

[1] Codes: https://gitlab.lrz.de/00000000014AE578/3dml-project

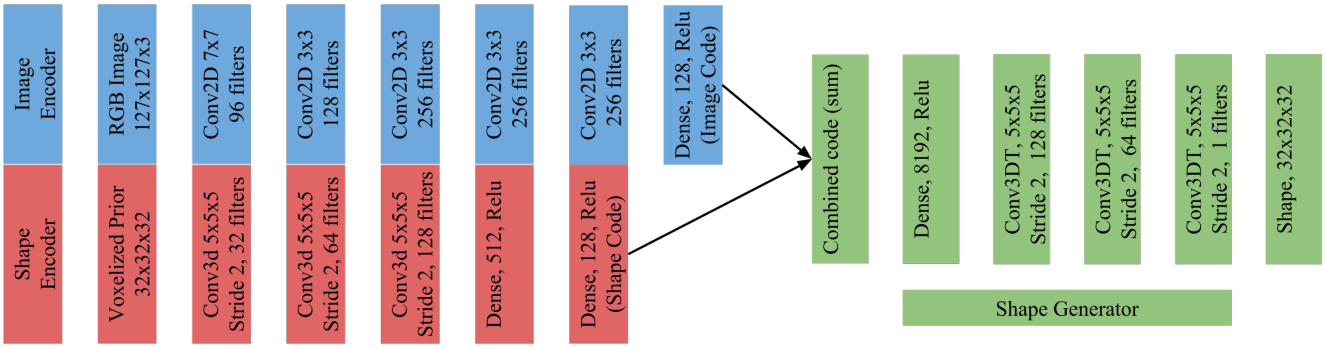

Figure 2. Model of baseline [1]. The image encoder takes a similar structure as [2], while the shape encoder and decoder follow the network of Yang et al. [3].

However, computing the prior is laborious and not intractable in every case because it needs many shapes for every class. To make our model more tractable, we implement the 3D reconstruction with only a few images instead of a shape prior.

3. Main Method

Compared to baseline, we aim to recover 3D shapes based on a single-view image with the help of a classification network instead of average 3D shapes. The principle of the proposed model is using a classification network to extract global features and a single-view image to provide local features so that the 3D network can obtain enough features to accomplish the reconstruction.

Figure 3 illustrates the model structure, which is made of a classification network and a 3D reconstruction network. The input of these two networks is the same image. The classification network uses the image to extract its global features by identifying its class. The reconstruction network builds the 3D shape based on the image and the embedding from the classification network.

Figure 4 shows the model network in detail. We denote the dataset as $D = \{S_i | 1 \leq i \leq N\}$. Each shape S_i is along with the following information: 1) To which class c_i it belongs; 2) A set of n multiple view rendering images $V = \{v_i | 1 \leq i \leq 24\}$. In the training process, we feed the same image v_i to two encoders. The output of the upper encoder is used for the classification task. During the classification process, we could get the embedding containing information about the features of a specific shape class, i.e. global features. It concatenates the classification embedding with the output of the reconstruction encoder and passes the concatenated embedding to the 3D shape decoder. This way, the whole model could get enough features to reconstruct 3D shapes. The loss we choose for 3D reconstruction is binary-cross-entropy (denoted as L_{3d}) and the loss for classification task is cross-entropy (denoted as L_{class}). we optimize the combination

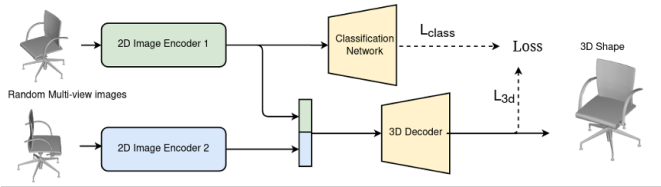

Figure 3. Model Structure.

loss $L = a * L_{class} + b * L_{3d}$. The main parameters we need to decide through experiments are 1. learning rate; 2. loss weights. In the testing phase, we calculate Intersection over Union (IoU) value as performance measurement as the baseline.

4. Experiment

In this section, some experiments on the baseline model and proposed model are introduced in detail. The results are analyzed as well.

4.1. Dataset

In the baseline and our model, we both use the ShapeNet dataset [4]. ShapeNet dataset contains RGB images, their categories, and $32 \times 32 \times 32$ voxelized representations. Each shape has 24 taken pictures from random views. We crop the images to $3 \times 127 \times 127$ and use these images as input. For part of the experiments annotated as 7-class, we use airplanes, cars, chairs, displays, phones, speakers, and tables.

4.2. Experiments on Baseline Model

We split the dataset to 75-5-20 (train-validation-test) following the baseline paper and train the model in order to verify the correctness of the replicated model. Table 1 shows the results by comparing the Intersection over Union (IoU) from the paper and our replication. The results are obtained after 150 epoch training with 0.1 as the learning

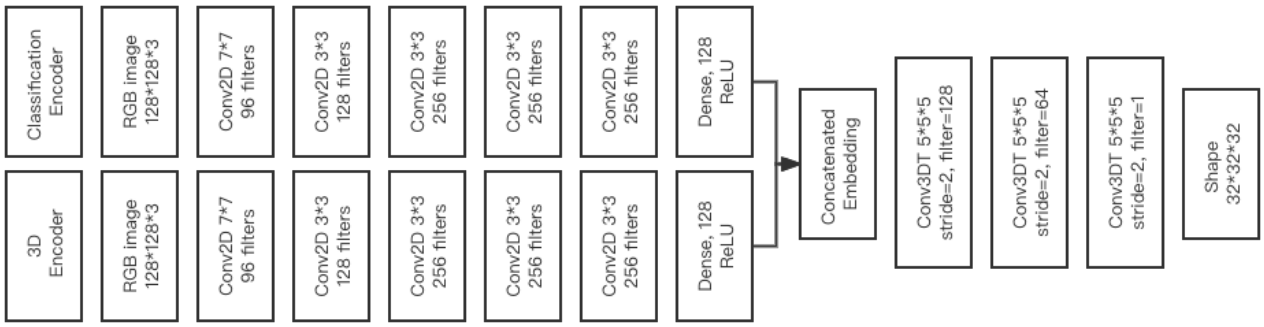

Figure 4. Network Architecture.

rate and using the Adam optimizer. The baseline can also predict unseen classes, so they train the model on only seven classes. We also train the baseline model on 13 classes to compare with our model. The result shows that the replicated model gets comparable results. Moreover, we think the split leads to a too small validation set which may result in instability in the training process, as it suffers from a large variance of the validation loss. Therefore, we split the dataset using 60-20-20 (train-validation-test) and train the model again. The results still keep the same as using 75-5-20.

	7 Classes	**13 Classes**
Paper	0.63	-
Replication	0.56	0.50

Table 1. Comparing the experiment results.

4.3. Experiments on Supposed Model

As the total loss is defined as $L = a * L_{class} + b * L_{3d}$. The first experiment we conduct is to find if the ratios have an influence on the performance. We set different values and find: 1) As the learning rates for the classification task and reconstruction task are different, we need to scale the learning rate using the ratio a and b; 2) If the ratio of the classification task is too large, the model focuses more on classification than reconstruction, which slightly decreases the reconstruction performance. According to a series of experiments, we find the best value of a and b is 0.05 and 1, respectively, with 0.001 as the learning rate using the Adam optimizer.

We first sum the global and local embedding in the length of 128, the same as the baseline. However, by converting the summation to concatenating, the network learns how the features should be combined, i.e., more flexibility can be introduced, which slightly increases its performance. We also enlarge the size of embeddings. After setting the global and local feature size to 256, we only see a very slight improvement. Embeddings with a length of 128 are able to encode sufficient features for reconstruction. Table 2 shows the results.

Embedding size	**128 (sum)**	**128 (cat)**	**256 (cat)**
IoU	0.618	0.620	0.621

Table 2. Results of embedding manipulations.

Figure 5 visualizes some reconstruction results. The model is generally able to reconstruct 3D shapes from pictures. However, it can not reconstruct the fine-grained features, e.g. the armrests of the chair.

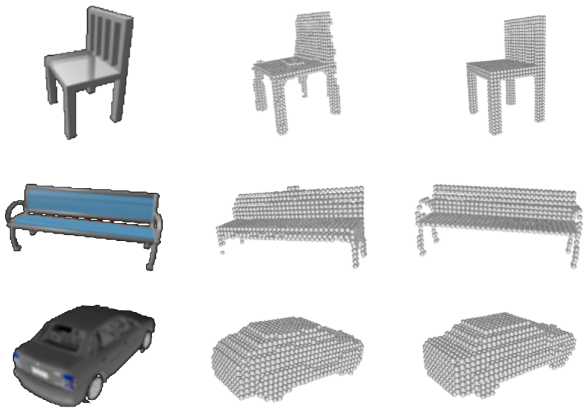

Figure 5. Visualization of reconstruction results. The first column is the input images; The second column shows the reconstruction results; The last column illustrates the ground truths.

4.4. Proposed Model vs Baseline Model

The baseline model encodes the average shape as global features and extracts local features from the single-view image to modify the average shape. Therefore, for each shape, the network extracts the global features from sampled shapes in the same category, which might be incon-

sistent with the input image. We introduce a classification task on the single-view image to force the network to extract global features on the image. In this case, global and local features correspond to the same shape. We expect the consistency of global and local features can improve the performance. Table 3 compares the performance of the baseline and the proposed model. The proposed model outperforms the baseline model on both 7-class and 13-class reconstruction. The proposed model for 7-class reconstruction is trained using an Adam optimizer with a learning rate of 0.001, and the variable a and b is set to 0.001 and 1, respectively. As the classification task is prone to overfitting, we add dropout layers to the classification network. However, this modification does not contribute to better performance.

	7 Classes	**13 Classes**
Baseline	0.56	0.50
Proposed	0.627	0.620

Table 3. Comparing performances.

5. Conclusion

In conclusion, our project consists of replicating Wallace's paper in 2019 and proposing our own network. The replication performs similarly on several experiments conducted on the original paper. Our proposed network achieves similar performance when trained on seven classes from the ShapeNet. Moreover, since our model can potentially generalize to more scenarios(without the need for 3D shapes), we also delivered experiments that train the model on all 13 classes from the ShapeNet datasets to test both models on more diverse scenarios. It is shown that our model can outperform the baseline significantly when the number of training classes increases.

The classification of ShapeNet dataset is not perfect [5], resulting in unstable global embeddings. Future work can use the DeepCluster methods [6] to extract global features more stably.

References

[1] Bram Wallace and Bharath Hariharan. Few-shot generalization for single-image 3d reconstruction via priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3818–3827, 2019.

[2] Christopher B. Choy, Danfei Xu, Jun Young Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of 14th European Conference on Computer Vision*, pages 628–644, 2016.

[3] Guandao Yang, Yin Cui, Serge Belongie, and Bharath Hariharan. Learning single-view 3d reconstruction with limited pose supervision. In *European Conference on Computer Vision*, pages 90–105, 2018.

[4] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository, 2015.

[5] Anny Yuniarti, Nanik Suciati, and Agus Zainal Arifin. Image classification performance evaluation for 3d model reconstruction. In *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, pages 229–233. IEEE, 2020.

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.