Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Active Object Reconstruction Using a Guided View Planner

Xin Yang[*1 2], **Yuanbo Wang**[*1], **Yaru Wang**[1]
Baocai Yin[1], **Qiang Zhang**[1], **Xiaopeng Wei**[1], **Hongbo Fu**[2]
[1] Dalian University of Technology
[2] City University of Hong Kong
xinyang@dlut.edu.cn, yuanbodlut@gmail.com, wangyaru@mail.dlut.edu.cn
{ybc, zhangq, xpwei}@dlut.edu.cn, hongbofu@cityu.edu.hk

Abstract

Inspired by the recent advance of image-based object reconstruction using deep learning, we present an active reconstruction model using a guided view planner. We aim to reconstruct a 3D model using images observed from a planned sequence of informative and discriminative views. But where are such informative and discriminative views around an object? To address this we propose a unified model for view planning and object reconstruction, which is utilized to learn a guided information acquisition model and to aggregate information from a sequence of images for reconstruction. Experiments show that our model (1) increases our reconstruction accuracy with an increasing number of views (2) and generally predicts a more informative sequence of views for object reconstruction compared to other alternative methods.

1 Introduction

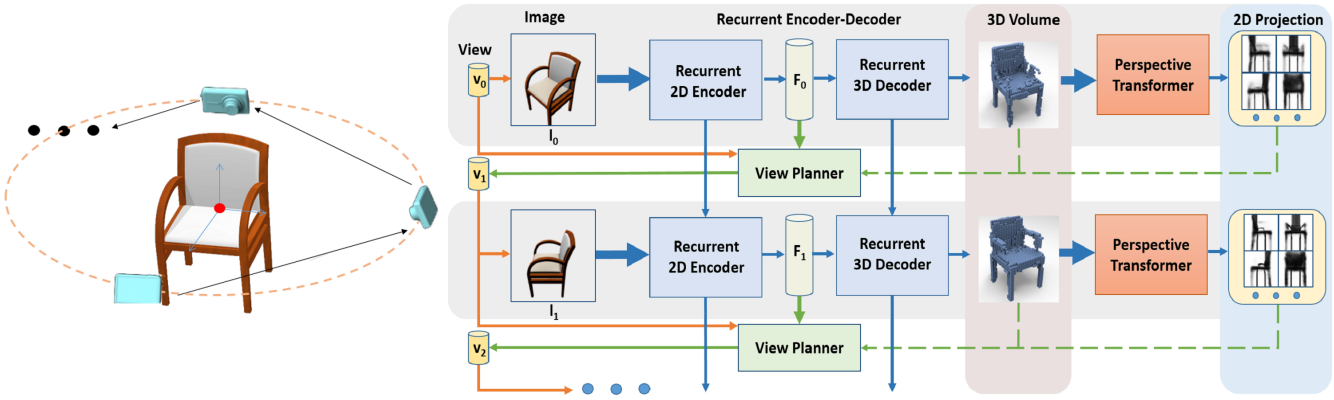
With the growing application demand of robot-object manipulation and 3D printing, automatic and efficient 3D model reconstruction from 2D images has recently been a hot topic in the research field of computer vision. Classic 3D reconstruction methods, based on the Structure-from-Motion technology [Halber and Funkhouser, 2017; Snavely *et al.*, 2006], are usually limited to the illumination condition, surface textures and dense views. On the contrary, benefited from prior knowledge, learning-based methods [Liu *et al.*, 2017; Yan *et al.*, 2016] can utilize a small number of images to reconstruct a plausible result without the assumptions on the object reflection and surface textures.

Regarding object reconstruction as a predictive and generative issue from a single image, learning based methods usually utilize CNN-based encoder and decoder to predict a 3D volume [Girdhar *et al.*, 2016; Dai *et al.*, 2017; Wu *et al.*, 2016b; Kar *et al.*, 2015], 3D structure [Wu *et al.*, 2016a] or point set [Fan *et al.*, 2017] trained by 3D supervision. Recent work attempts to use only 2D images for 2D supervision to train image-based object reconstruction models. For example, with a differentiable style, Yan et al. [2016] propose Perspective Transformer Nets with a novel projection loss that enables the 3D learning using 2D projection without 3D supervision.

A crucial assumption in the above-mentioned models, however, is that the input images contain most information of a 3D object. As a result, these models fail to make a reasonable prediction when the observation has severe self-occlusion as they lack the information from other views. An effective solution is to utilize more views to make up the information. Choy et al. [2016] propose a 3D Recurrent Neural Networks (3D-R2N2) to map multiple random views to their underlying 3D shapes. In contrast to 3D-R2N2, we focus on how much information is sufficient and how to aggregate these information for 3D object reconstruction. In other words, how many views and which views of image can capture the most informative feature and maximize the quality of reconstruction? This is a problem about dynamical view prediction when reconstructing an object. It means an active process of capturing new information for 3D reconstruction.

There are many methods receiving the maximal information gain such as the decrease of uncertainty [Xu *et al.*, 2015], Monte Carlo sampling [Denzler and Brown, 2002] and Gaussian Process Regression [Huber *et al.*, 2012]. Some methods regard this information gain task as a sequential prediction of the next best view (NBV) by reducing scanning effort [Wu *et al.*, 2014] or uncertainty of object [Wu *et al.*, 2015] with the least observations. All these attempt to receive the maximal information gain with the minimal number of views. Intuitively, it is not absolute that the best performance comes from the maximal information. The reason is perhaps that learning based methods attempt to exploit the spatial and temporal structure of the sequential observations [Xu *et al.*, 2016] and learn to predict an approximate views sequence to optimize the deviation between the prediction and the ground truth. The attention model [Mnih *et al.*, 2014] is a good solution for the problem of sequential locations prediction as it utilizes the recurrent neural network to extract the information from an image sequence, and adaptively selects a sequence of discriminative regions or locations. There are many successful attention-based applications such as image classification

[*]Equal Contribution.



(a) Viewing Space (b) Active Object Reconstruction Model

Figure 1: We present an active object reconstruction model. Targeting a 3D object, we utilize the guidance from both 3D volume and 2D projection to train a View Planner, which continuously predicts the next informative and discriminative views parameterized as camera azimuth angles on a viewing circle † around the object (a). Based on the predicted view sequence, our Recurrent Encoder-Decoder takes the feature sequence encoded by a Recurrent 2D Encoder as input and recurrently decodes it to a 3D volume by using a Recurrent 3D Decoder (b), which gradually improves the accuracy with an increasing number of views.

[Xiao *et al.*, 2015] and shape recognition [Xu *et al.*, 2016].

However, different from image identification building the inconsistency between predicted category labels and ground truth, object reconstruction requires a dense prediction for each voxel, and it thus needs to explore a deeper relation between 3D volume and 2D images and to use this relation to guide the aggregation of multi-view information and the planning of sequential views. To achieve this, our method differs from the other attention-based models in two major aspects. First, to constrain the consistency between 3D volume and 2D images, we combine the volumetric and projective supervision in the process of view aggregation. Second, for guided view planning, our reward is set upon the performance of reconstruction and volume-projection consistency, facilitating the view planner to capture more information.

Our experiments show that our model aggregates more discriminative information from multi-view images and apparently increases the accuracy with an increasing number of views. For the view planning task, we demonstrate our sequences can give better prediction than other strategies. The main contributions of this paper are as follows: (1) We build a Recurrent Encoder-Decoder based on multiple Conv-RNN layers and a volume-projection supervision, leading to a better reconstruction performance. (2) We combine 3D volume prediction and 2D projection to design the reward for view planning policy learning. Under the control of the combined reward, we can implicitly learn the deep relationship between 3D reconstruction and 2D images, and optimize the planning policy. (3) We propose an active framework that learns a view planner for 3D object reconstruction. Our model can dynamically determine the views based on information gain and discrimination, which makes the reconstruction more accurate.

†Note that our model can apply to a viewing sphere as well but we found viewing parameters in a circle are enough for the training of the synthesis data.

2 Methodology

2.1 Overview

Figure 1 illustrates our model, which is summarized as follows. In the training stage, starting at a random view in the viewing circle (Figure 1 (a)), we select the pre-rendered color image with its associated view close to the random view and feed it into a **Recurrent 2D Encoder**, which encodes the image to a latent unit F_0 and propagates the information when absorbing new views. The encoded unit is then fed into two branch pathways. One is a **Recurrent 3D Decoder**, which maps the latent unit extracted from all past views to a predicted 3D volume. The other one called **View Planner** serves as a dynamical view prediction module that continually receives and integrates the encoded units from current and all past views (Figure 1 (b)), and regresses view parameters for next observation (Section 2.2). To combine view planning and object reconstruction into a unified and correlative model, we propose a volume-projection guidance (Section 2.3) for the supervised learning of view-based volume mapping and reinforcement learning of continuous view prediction. The volume comes from the Recurrent Encoder-Decoder and the projection is generated by a differentiable **Perspective Transformer** [Yan *et al.*, 2016]. In the test stage, our model keeps acquiring the images observed from a target object under the guidance from the View Planner and reconstructs the 3D model with the Recurrent Encoder-Decoder.

2.2 Network Architecture

We consider multi-view volumetric reconstruction as a dense prediction from a sequence of views and develop a unified framework for both volume reconstruction and view planning in an active scheme. The procedure can be divided into multiple time steps and we plot one step of data flow in Figure 2. Next we discuss the detailed architecture.

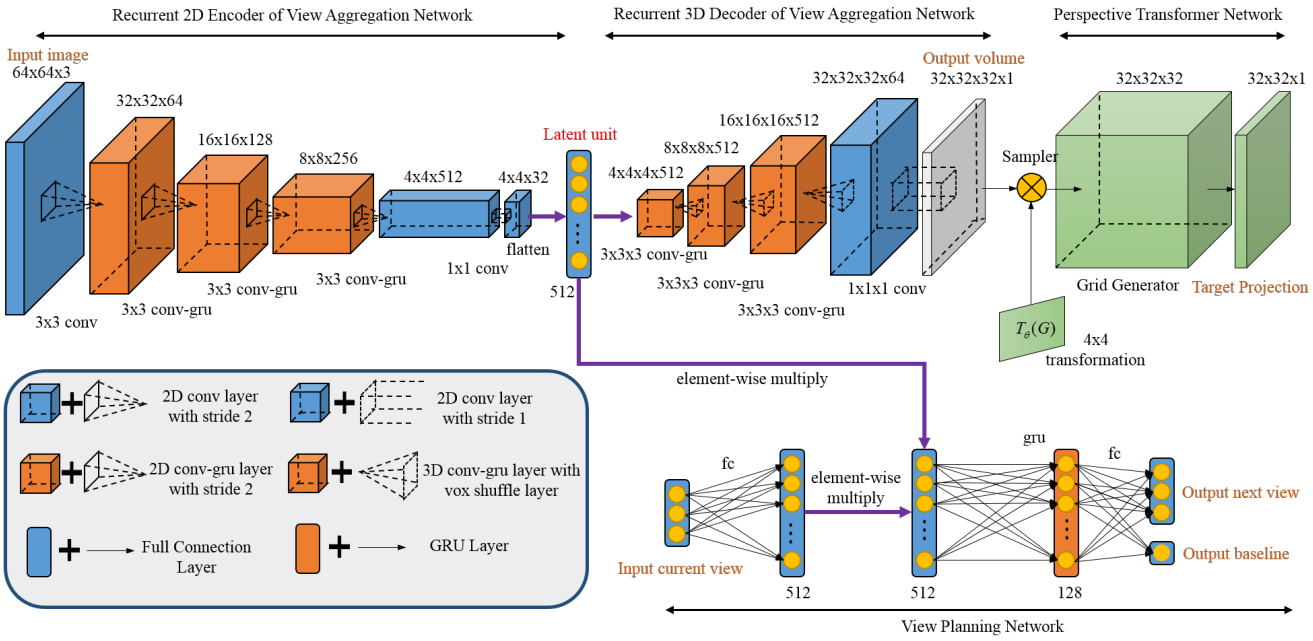


Figure 2: Illustration of network architecture. Our entire network consists of four components: a Recurrent 2D Encoder, a Recurrent 3D Decoder, Perspective Transformer and a View Planner. Taking current view and a rendered RGB image as input, the network predicts occupancy probability for each voxel in a $32 \times 32 \times 32$ volume and the next view.

Recurrent 2D Encoder. This network is utilized to extract features from input image I_t and aggregate them with the past views to a latent unit: $F_t = f_{enc}(I_t, S^{enc}_{t-1})$, where S^{enc}_{t-1} refers to all past states of hidden layers in the encoder network. In our implementation, we build a Recurrent 2D Encoder upon multiple 2D Conv-GRU layers to extract the spatial features from images and integrate the sequential past states. 2D Conv-GRU layers update their hidden states under the control of three convolution-operator gates with the hidden states arranged in 2D space. Compared to convolution based auto-encoder networks with a single 3D-GRU, our network is better at feature extraction on image sequence, since our reconstruction performance improves a large margin (demonstrated in Section 3.4).

Recurrent 3D Decoder. Taking features F_t extracted by the encoder network as input, Recurrent 3D Decoder utilizes multiple 3D Conv-GRU layers, which are similar to 2D Conv-GRU layers but arranged in 3D space, to decode F_t to a 3D volume where each voxel grid retains the probability of occupancy: $V_t = f_{dec}(F_t, S^{dec}_{t-1})$, where S^{dec}_{t-1} refers to all past states of hidden layers in the decoder network. To increase the resolution of feature maps, we add a voxel shuffle layer after each 3D Conv-GRU layer, which allocates the depth dimension of feature vector to 3D space.

View Planner. The task of view planning is to actively regress a sequence of views parameterized as camera azimuth angles on a viewing circle around the object. Taking a random angle as initial view, we sequencely feed the rendered image under the current view into the Recurrent 2D Encoder to extract and aggregate features. We then merge the features with the current view parameters by element-wise multiplication to get a viewing "glimpse" [Mnih *et al.*, 2014], which fuses the information of both the image sequence and current view: $g_t = f_{enc}(I_t, S^{enc}_{t-1}) * f_{view}(v_{t-1})$. With a GRU layer, our model retains all past glimpse information and continuously absorbs new views. The glimpse information can be formulated as $h^{gru}_t = f_{gru}(g_t, h^{gru}_{t-1})$, which discriminatively describes the relation of images sequence, 3D volume and parameters of a sequence of views. Feeding the states to an extra full connection layer after the GRU layer, we finally predict view parameters of the next view to get the next image input.

Perspective Transformer. We use the Perspective Transformer Network proposed by Yan et al. [2016] to obtain a 2D projection from the 3D volume. Utilizing this 3D differentiable transformation, we project the 3D voxel prediction to a 2D grid, which looks like a projection silhouette. Combining this differentiable 2D projection with the predicted volume, we build a projective guidance on the training of volume prediction and view planning (see Section 2.3 for details).

2.3 Volume-Projection Guidance

We combine the procedure of object reconstruction and view planning into a unified framework. For view planning, it optimizes a view prediction policy under the control of feedback signals based on the evaluation of reconstruction performance (as shown in Figure 1 (b)). For object reconstruction, the Recurrent Encoder-Decoder receives input images from the informative views predicted by the View Planner, which ensures a sufficient information gain and boosts the improvement of reconstruction performance. We jointly train two modules by both volumetric and projective patterns but use different strategies: a reinforcement learning under the con-

trol of volume-projection reward and a supervised learning using volume-projection supervision.

Volume-Projection Reward. At each time step, the guidance of View Planner comes from the performance of volume predicted by the reconstruction module. In other words, the View Planner receives a reward signal which is built upon the reconstruction feedback from the Recurrent Encoder-Decoder. We only calculate the accumulative reward during a whole episode to update a view planning policy, which maps the image observation to the camera view. To accommodate the dense prediction on the whole 3D volume and ensure the reconstruction improvement with new views fed in, the increment of voxel Intersection-over-Union (IoU) is utilized to measure the reconstruction reward. Mathematically, the reward at step t can be formulated as follows:

$$r^t_{cons} = IoU(\widehat{V}_t, V) - IoU(\widehat{V}_{t-1}, V), \tag{1}$$

where $\widehat{V}_t$ is the 3D volume predicted by the Recurrent Encoder-Decoder, and V is the corresponding ground truth in the dataset.

To implicitly learn the relation between 3D volume and 2D projection, we design a projection reward to encourage the consistence between 3D construction and 2D projection from different viewpoints. The projection reward is defined as the increment of pixel IoU value on 2D silhouettes sampled by the Perspective Transformer from multiple different views. The reward at step t can be formulated as follows:

$$\begin{aligned} r^t_{proj} = \frac{1}{n}\sum_{i=1}^{n} & IoU(f_{ptn}(\widehat{V}_t, v_i), f_{ptn}(V, v_i)) \\ & - IoU(f_{ptn}(\widehat{V}_{t-1}, v_i), f_{ptn}(V, v_i)), \end{aligned} \tag{2}$$

where n is the number of projection views and v_i is the $i-th$ view.

In addition, to punish for selecting similar views, we add an additional movement cost defined as the minimum value of the circle distance between the current location and past views. Integrating the reconstruction reward, projection reward, and movement cost, the final reward is defined as follows:

$$r = \lambda_v r_{cons} + \lambda_p r_{proj} - \lambda_m C_{move}, \tag{3}$$

where λ_v, λ_p and λ_m are the weights of the reconstruction reward, the projection reward, and the movement cost, respectively.

Using this reward, we can control the update of policy, which corresponds to the gradient policy algorithm. We sample the views predicted by the View Planner according to a normal distribution with a predefined standard deviation at each time step, and minimize the following loss function to optimize the view planning policy:

$$L_{rl} = \sum_{t=1}^{t=T} -log(p(v^p_t \mid I_t, \theta_{vp})) * (R_t - b_t), \tag{4}$$

where v^p_t is a sampled view at time step t, R_t is the reward at time t, and b is a predicted value as a various baseline which is utilized to center the reward ([Mnih *et al.*, 2014]). The log probability can derivative by back propagation of network and reward R is the signal received from the reconstruction feedback of Recurrent Encoder-Decoder.

Volume-Projection Supervision. The loss function of Recurrent Encoder-Decoder is defined as the mean value of voxel-wise square error (MSE):

$$L_{vox} = \|V_{pre} - V\|^2, \tag{5}$$

where V_{pre} is the final output of the Recurrent Encoder-Decoder.

Besides the 3D volumetric loss, we add a 2D projective loss to implicitly learn the effect of 2D projection on 3D prediction, which improves the multi-view reconstruction performance. The 2D supervision loss is formulated as:

$$L_{proj} = \frac{1}{n}\sum_{j=1}^{n} \|f_{ptn}(V_{pre}, T^j) - M^j\|^2, \tag{6}$$

where f_{ptn} is the Perspective Transformer Network, T^j is the parameters of $j-th$ view with a 4-by-4 transformation matrix, and M^j is the projection of the ground truth voxel. We combine the 3D supervision (Equation 5) and 2D supervision (Equation 6) using a weighted sum as:

$$L = \lambda_{vox} L_{vox} + \lambda_{proj} L_{proj}, \tag{7}$$

where λ_{vox} and λ_{proj} are the weights of volumetric loss and perspective loss, respectively.

3 Evaluation

In this section, we discuss the following three questions: (1) Can our model improve the accuracy of reconstruction with an increasing number of views? (Section 3.2) (2) Can our View Planner obtain more informative and discriminative views to boost the reconstruction performance compared to the other alternative methods? (Section 3.3) (3) Do our network structures learn better than other settings? (Section 3.4)

3.1 Implementation Details

Our model was trained and tested under the Pytorch framework, accelerated by a GPU (NVIDIA GTX 1080Ti). We used the dataset from [Yan *et al.*, 2016], which is based on the ShapeNetCore [Wu *et al.*, 2015]. Each model is represented as a 3D volume of $32 \times 32 \times 32$ from its canonical orientation, and images are rendered from 24 azimuth angles with 30° elevation angle. For each rendered image, we cropped and resized the centered region to 64×64 pixels with 3 channels (RGB). We updated the weights by using ADAM solver with batchsize 16, epoch 200, $\lambda_{vox} = \lambda_{proj} = 0.5$.

3.2 Evaluation on Reconstruction Performance

We compare our method with PTN (Perspective Transformer Network [Yan *et al.*, 2016]), OGN (Octree Generating Networks [Tatarchenko *et al.*, 2017]) and 3D-R2N2 (proposed by Choy et al. [2016]). PTN uses an encoder-decoder model to make a 3D volume prediction trained with a combined loss of both projection supervision and volume supervision. OGN generates volumetric 3D outputs in a compute- and memory-efficient manner by using an octree representation. To evaluate the multi-view performance, we also compare to 3D-R2N2, which performs both single- and multi-view 3D reconstruction using a 3D recurrent network. We trained and tested

methods	PTN	OGN	3D-R2N2			Ours		
# views	1	1	1	3	5	1	3	5
plane	0.553	0.587	0.513	0.549	0.561	0.605	0.657	**0.679**
bench	0.482	0.481	0.421	0.502	0.527	0.498	0.569	**0.597**
cabinet	0.711	0.729	0.716	0.763	0.772	0.715	0.769	**0.789**
car	0.712	0.816	0.798	0.829	0.836	0.757	0.805	**0.838**
chair	0.458	0.483	0.466	0.533	0.550	0.532	0.590	**0.617**
monitor	0.535	0.502	0.468	0.545	0.565	0.524	0.596	**0.624**
lamp	0.354	0.398	0.381	0.415	0.421	0.415	0.445	**0.461**
speaker	0.586	0.637	0.662	0.708	0.717	0.623	0.685	**0.723**
firearm	0.582	0.593	0.544	0.593	0.600	0.618	0.664	**0.693**
couch	0.643	0.646	0.628	0.690	0.706	0.679	0.723	**0.749**
table	0.471	0.536	0.513	0.564	0.580	0.547	0.590	**0.617**
cellphone	0.728	0.702	0.661	0.732	0.754	0.738	0.793	**0.822**
watercraft	0.536	**0.632**	0.513	0.596	0.610	0.552	0.606	0.626
mean	0.565	0.596	0.560	0.617	0.631	0.600	0.653	**0.680**

Table 1: The per-category multi-view reconstruction comparison by 3D-R2N2, ours and reference values of the sing-view based model PTN and OGN. Except for the watercraft, our method performs consistently the best in each category.

our network using 13 categories with train/test data split used by 3D-R2N2's authors, which was adopted by OGN's author as well. For a fair comparison, we followed 3D-R2N2's setting and used 5 random views along the viewing circle to evaluate our Recurrent Encoder-Decoder model. For PTN, We re-trained the model for multi-category reconstruction using the code released by the authors, since they originally trained their model only on chair category. In the test stage, we compute voxel IoU (1 with threshold 0.4 as the evaluation metric.

We evaluate the reconstruction performance of the compared methods on 13 categories as shown in the table of Table 1. It can be seen that our method performs better than the baseline volumetric reconstruction methods (PTN and OGN) when using only single view and outperform it a large margin with an increasing number of views. It proves the ability of our model predicting a reasonable reconstruction result by using only a single image. Compared to 3D-R2N2, our model gets higher IoUs when using the same number of views. The reason is perhaps that our Recurrent Encoder-Decoder extracts more discriminative features and aggregates the information from different views at a deeper level. The examples of reconstruction results shown in Figure 4 qualitatively show that our model can generally make a reasonable prediction on a global shape even from a single view and succeed to optimize the local details that 3D-R2N2 fails (pointed out by the red circles) when using more views.

3.3 Evaluation on Information Gain

To evaluate the performance of view prediction, we compared our View planner against two baselines and an alternative method. The two baselines consist of a random planners that selects random view as the next view and a farthest planners that selects the view which is the farthest away from previous views in the viewing circle around the targeting object. The alternative methods is the NBV technique proposed in ShapeNet [Wu *et al.*, 2015] which estimates the information gain of a view from 3D volume. We train and test our active reconstruction model using the chair models in ShapeNet and rendering images under the train/test data split used by PTN's authors [Yan *et al.*, 2016]. We set $\lambda_v = 10$, $\lambda_p = 10$, $\lambda_m = 0.04$. For comparison, we respectively feed the image sequence predicted by these strategies into the pre-trained Recurrent Encoder-Decoder to reconstruct the 3D volume.

We plot the IoU values and the decrease of Shannon Entropy over the number of views in Figure 3. Compared to other methods, our model not only attains more information but also gets a more accurate results, showing that our model is able to predict a both informative and discriminative view sequence for more accurate reconstruction results.

Structure	2E-R-3D	R2E-3D	2E-R3D	R2E-R3D
Encoder	2D Enc	R-2D	2D Enc	R-2D
Decoder	3D Dec	3D Dec	R-3D	R-3D
Loss	0.0241	0.021	0.014	**0.012**
IoU	0.6285	0.704	0.785	**0.798**
IoU(test)	0.519	0.542	0.571	**0.605**

Table 2: The comparison of quantitative results under different variations on network structures (evaluated using 5 views).

3.4 Network Structures Comparison

To demonstrate that our Recurrent Encoder-Decoder extracts a more discriminative feature from an image sequence and gets a better reconstruction performance, we compare four kinds of network architectures under different combinations: fully convolutional Encoder-Decoder with a 3D RNN (2E-R-3D), Recurrent 2D Encoder with a 3D CNN-based decoder (R2E-3D), 2D CNN-based Encoder with a Recurrent 3D Decoder (2E-R3D), and our Recurrent 2D Encoder with Recurrent 3D Decoder (R2E-R3D). We trained all these four models on the chair category using the rendered images from random views and ground truth 3D volume under the train/test data split of ShapeNet database used by PTN's authors. For comparison, we utilize 5 random views to reconstruct the 3D volume and show in Table 2 the results of MSE loss and IoU

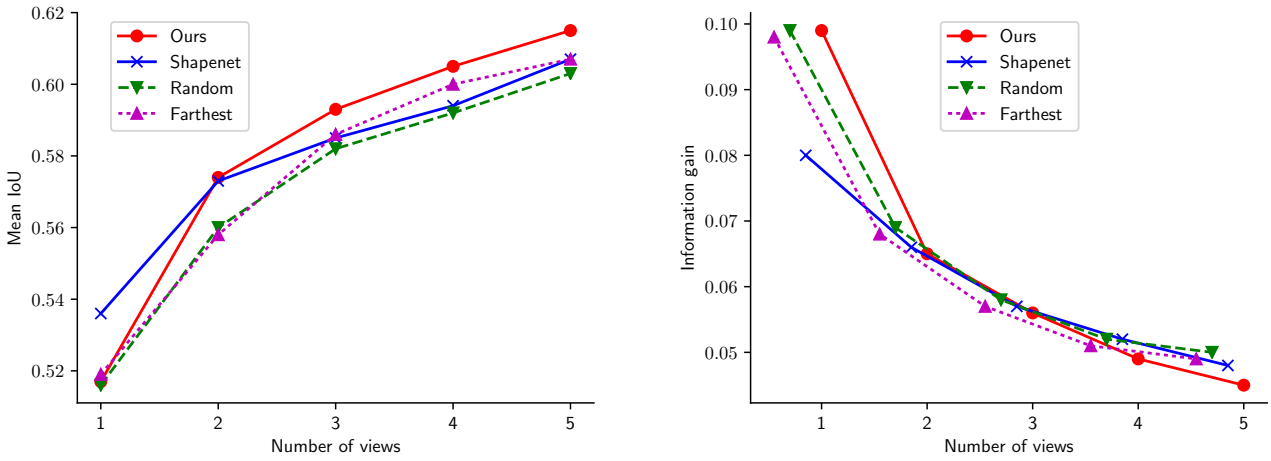


Figure 3: The view prediction comparison against the baselines (Random and Farthest) and ShapeNet. (a): IoU values (the higher and the better) over the number of views. (b): Information gain (the lower and the better) as the decrease of Shannon Entropy.

values. The results show that our R2E-R3D architecture performs the best on both training losses and testing IoU values. Using R2E-R3D model, we can achieve the best reconstruction performance against the other settings, which validates our model superior in view-based reconstruction task.

4 Conclusion

In this paper, we have presented an learning-based model with active perception which unifies the guided information acquisition and multi-view object reconstruction. Under the guidance from both volume and projection, we jointly train the Recurrent Encoder-Decoder and View Planner. Experiments demonstrate that our model obtains more information and increases the reconstruction performance with an increasing number of views. Our model only extracts the semantic features but ignores the correspondence of geometrical features from different camera viewpoints, leading to a slow growth when feeding in more than 5 views. In the future, we would utilize multi-modal features to optimize or jointly learn the object reconstruction and utilize more efficient data representations to increase the output resolution. Besides, it is interesting to extend our approach to multi-object reconstruction by predicting the transformation of camera view from one object to another one.

Acknowledgments

We thank the anonymous reviewers for the insightful and constructive comments. The work was partially funded by the Research Grants Council of HKSAR, China (Project No. CityU 11237116 and CityU 11300615), ACIM-SCM, the Hong Kong Scholars Program, and by NSFC grant from National Natural Science Foundation of China (NO. 91748104, 61632006, 61425002, U1708263).

References

[Choy *et al.*, 2016] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, 2016.

[Dai *et al.*, 2017] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Computer Vision and Pattern Regognition*, 2017.

[Denzler and Brown, 2002] J Denzler and C Brown. An information theoretic approach to optimal sensor data selection for state estimation. *Pattern Analysis and Machine Intelligence*, 2002.

[Fan *et al.*, 2017] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image. In *Computer Vision and Pattern Regognition*, 2017.

[Girdhar *et al.*, 2016] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*, 2016.

[Halber and Funkhouser, 2017] Maciej Halber and Thomas Funkhouser. Fine-to-coarse global registration of rgb-d scans. In *Computer Vision and Pattern Recognition*, 2017.

[Huber *et al.*, 2012] Marco F Huber, Tobias Dencker, Masoud Roschani, and Jürgen Beyerer. Bayesian active object recognition via gaussian process regression. In *Information Fusion*, 2012.

[Kar *et al.*, 2015] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *Computer Vision and Pattern Recognition*, 2015.

Figure 4: Qualitative results of reconstruction samples for example view sequences. 3D-R2N2 generally fails in the categories with much higer variation (eg. the lamp in the bottom right corner) while our model does better in feature extraction and view aggregation, leading to a more accurate reconstruction.

[Liu *et al.*, 2017] Xiaobai Liu, Yadong Mu, and Liang Lin. A stochastic image grammar for fine-grained 3d scene reconstruction. In *International Joint Conferences on Artifical Intelligence*, 2017.

[Mnih *et al.*, 2014] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, 2014.

[Snavely *et al.*, 2006] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Transactions on Graphics*, 2006.

[Tatarchenko *et al.*, 2017] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *International Conference on Computer Vision*, 2017.

[Wu *et al.*, 2014] Shihao Wu, Wei Sun, Pinxin Long, Hui Huang, Daniel Cohen-Or, Minglun Gong, Oliver Deussen, and Baoquan Chen. Quality-driven poisson-guided autoscanning. *ACM Transactions on Graphics*, 2014.

[Wu *et al.*, 2015] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Computer Vision and Pattern Recognition*, 2015.

[Wu *et al.*, 2016a] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3d interpreter network. In *European Conference on Computer Vision*, 2016.

[Wu *et al.*, 2016b] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, 2016.

[Xiao *et al.*, 2015] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Computer Vision and Pattern Recognition*, 2015.

[Xu *et al.*, 2015] Kai Xu, Hui Huang, Yifei Shi, Hao Li, Pinxin Long, Jianong Caichen, Wei Sun, and Baoquan Chen. Autoscanning for coupled scene reconstruction and proactive object analysis. *ACM Transactions on Graphics*, 2015.

[Xu *et al.*, 2016] Kai Xu, Yifei Shi, Lintao Zheng, Junyu Zhang, Min Liu, Hui Huang, Hao Su, Daniel Cohen-Or, and Baoquan Chen. 3d attention-driven depth acquisition for object identification. *ACM Transactions on Graphics*, 2016.

[Yan *et al.*, 2016] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, 2016.