3D Reconstruction of Simple Objects from A Single View Silhouette Image

Xinhan Di Trinity College Ireland dixi@tcd.ie Pengqian Yu National University of Singapore yupengqian@u.nus.edu

Nov 2016

Abstract

While recent deep neural networks have achieved promising results for 3D reconstruction from a single-view image, these rely on the availability of RGB textures in images and extra information as supervision. In this work, we propose novel stacked hierarchical networks and an end to end training strategy to tackle a more challenging task for the first time, 3D reconstruction from a single-view 2D silhouette image. We demonstrate that our model is able to conduct 3D reconstruction from a single-view silhouette image both qualitatively and quantitatively. Evaluation is performed using Shapenet for the single-view reconstruction and results are presented in comparison with a single network, to highlight the improvements obtained with the proposed stacked networks and the end to end training strategy. Furthermore, 3D reconstruction in forms of IoU is compared with the state of art 3D reconstruction from a single-view RGB image, and the proposed model achieves higher IoU than the state of art of reconstruction from a single view RGB image.

1 Introduction

3D reconstruction techniques rebuild 3D world from 2D information and contribution has been consistently made since the latest two decades. Theoretically, framework are built in this area such as space carving [17], SfM [23], SLAM [5] and CMVS/PMVS [8]. Practically, implementation to reconstruct the 3D world is developed such as VisualSfM [28] and CMPMVS [12].

Among a variety of the techniques aiming at 3D reconstruction, 3D reconstruction from a single view is challenging. In order to overcome this challenge, shape priors are applied in order to compensate for the problem of illposedness. For example, smoothness in the definition of consistency of surfaces is learned and applied for the reconstruction of curved surfaces [19, 20, 21, 26, 32, 11, 22]. Geometric relation is developed as another kind of prior to dissolve ambiguities in the reconstruction process [18, 4, 6, 10, 11]. High lever priors are also used to overcome single-view reconstruction. Semantic relation priors which infer structure information of objects are applied [10, 16, 6, 11]. Also, the learned shape priors are used for the reconstruction of facades [16], human bodies [2], and category specific objects [27, 14] through learning from defined consistency rules.

Thanks to the development of the deep learning techniques, 3D reconstruction from a single view reduces the dependence on classical priors through employing a trained reconstruction model. A 3D recurrent reconstruction neural network(3D-R2N2) is developed to learn a mapping from images of objects to the 3D shapes and is presented as rebuilding 3D shape from a single image [3], 3D Interpreter network is designed for single-view reconstruction with application of two pre-trained models and a projection layer during training a deep learning framework [29], the prediction of both an RGB image and a depth map of the object are made through a convolutional network [25], another joint prediction of a depth map and intrinsic images from single-image input is conducted through training a joint convolutional neu-

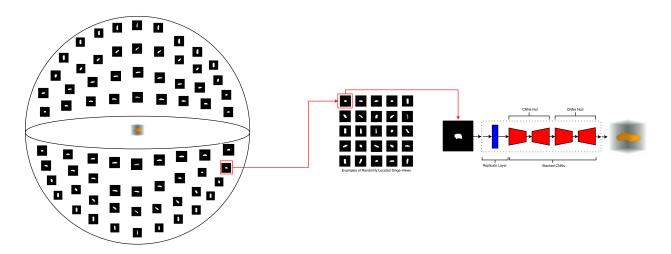


Figure 1: Category Specific 3D Reconstruction from A Single-view Silhouette Image. The Left part represents single-view silhouette images located randomly on a sphere around a object. The right part represents a proposed framework to reconstruct a 3D shape from a single-view silhouette image.

ral field(JCNF) model [15]. Stereo pair images are used to train an auto-encoder without requiring a pre-training stage for single-view depth prediction [9]. Volumetric 3D object reconstruction is achieved from the designed network with extra input of the pre-trained projection transformation code [31].

In this paper, we propose a stacked network for 3D reconstruction from a single-view silhouette image that is demonstrated to be able to reconstruct category-specific objects from a randomly located single-view silhouette image. The overall architecture of the architecture is illustrated in Fig1. The architecture consists of two major components:(1) the replicate layer, which works to produce a single-view 2D segmentation to 3D volume coarse segmentation through simple replication operation. (2) stacked reconstruction networks, which uses the 3D volume segmentation to reconstruct a 3D shape. The proposed network architecture is trained end to end through following a proposed training strategy for stacked hierarchical network without a pre-training stage or separate training for each partial network. The proposed training strategy works as controlling the output of each network during training process of the staked hierarchical networks through gradient calculation. In the stacked hierarchical network architecture, it follows a unique criterion to minimize the error for the final network which is different from combined criterion for each network. As illustrated in Fig1, the proposed architecture achieves good performance for 3D category specific reconstruction from a single-view silhouette image, and the silhouette image is randomly located around a sphere, and previous network framework for 3D reconstruction from single view is not demonstrated to deal with this large view variation .

The main contributions of our work are three-fold. Firstly, we proposed a stacked hierarchical network for 3D reconstruction from a single view and its end-to-end training strategy. And we demonstrate that the design of the stacked networks architecture and its training strategy works much better than a single network. Secondly, we demonstrate that category specific 3D reconstruction from a single-view silhouette image without RGB textures or other extra input is conducted through our proposed model. Thirdly, we demonstrate that for 3D reconstruction from a single-view silhouette image, single views are randomly and widely located on a sphere surface while previous network frameworks work for small view variation, such as single view images located on a circle.

2 Related Work

Network is employed as a data-driven method to provide useful priors for 3D reconstruction from a single view image recently. WarpNet is exploited to align an object in one image with a different object in another which allows single-view reconstructions with quality [13]. Virtual view networks(VNN) [1] are built to produce smooth rotations through the class object collection and points matching for point cloud reconstruction from a single image. The view prior of a single image is estimated through exploiting a deep CNN architecture with a high learning capacity [24]. Two deep network stacks are employed to produce a depth prior for a single image without the need for superpixelation [7].

Furthermore, deep learning networks are applied to build frameworks for 3D reconstructions from a single-view image. Among many frameworks that are applied, the most relevant works are learned NRSfM model with defined consistency and smoothness criterion [14], 3D-R2N2 network with adding 3D convolutional LSTM inside [3], 3D-INN framework with usage of 2 pre-trained models and a projection layer [29], volumetric reconstruction of objects with pre-trained projective transformation code [31] and depth prediction from CNN neural networks [25, 15, 9]. A single network is built for 3D reconstruction from multiple-view silhouette images with multiple-view information [30].

Compared with the above frameworks, our framework is considered as advantages than the exploited networks for 3D reconstruction from a single view. Firstly, despite that a single-view RGB image [3, 29, 31, 25] is used as input, our framework releases RGB dependence to a singleview silhouette image. Secondly, the tested single views [31, 3, 29, 25] are not largely located in space, but our framework works to reconstruct a 3D shape from a single view randomly and widely located on a sphere around the object, while the above frameworks achieves 3D reconstruction from a number of single views or arbitrary views around circles. Thirdly, we reduce the input of singleview silhouette image without key points used in [14]. Finally, we build a stacked hierarchical networks and training the network end to end despite of separate training of each network or usage of pre-trained codes [31, 29].

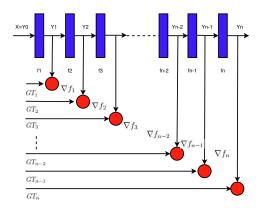


Figure 2: Stacked hierarchical networks.

3 Stacked Hierarchical End-to-end Training Networks

A common choice of completing complex task through networks is a design of stack networks. And train each single network in multiple stages that separate training is conducted for training each network. However, this is not strictly trained end to end. However, we propose a novel design and end to end training strategy for stacked networks and employ it for the challenging task of 3D reconstruction from a single-view silhouette image.

3.1 Networks Architecture and Training Strategy

Stacked hierarchical networks are built for training a model to complete a complex task.

Definition 3.1 Let Y = F(X) be stacked hierarchical networks where X is the input, Y is the output of the stacked networks, F is the overall function of the stack networks, n is the number of the single networks that this stacked networks are comprised of. Consider F can be split as n stacked functions $f_k, k = 1...n$ and f_k is the overall function of kth sub-network in the stacked hierarchical networks. Also consider $Y_k = f_k(Y_k - 1), k = 1...n$, Y_k is the output of the kth sub-network. And $Y_0 = X, Y_n = Y$.

Strategy 3.1 Consider the stacked hierarchical networks Y = F(X), in order to guarantee an effective end-to-end training for stacked hierarchical networks combined with a number of single networks. Both a multiple gradient decent learning strategy and an unique error criterion for the final single network of the stacked networks are employed. This strategy is employed in order to find a global solution for stacked networks to overcome the difficulty of searching global optimization from end-to-end training.

Consider multiple gradient descent learning strategy as the Eq.(1) and Eq.(2).

$$\nabla F(X) = \sum_{k=1}^{n} \lambda_k \nabla f_k(Y_{k-1}) \tag{1}$$

$$\nabla f_k(Y_{k-1}) = |f_k(Y_{k-1}) - GT_k|^{\frac{1}{2}}$$
 (2)

where $\nabla F(X)$ is the multiple gradient for the stacked networks during training, ∇f_k is the gradient for the kth single network. The parameter λ_k follows $\sum_{k=1}^n \lambda_k = 1$. GT_k is the ground truth for the kth single network.

Consider learning the unique error criterion as the Eq.(3).

$$\Theta^* = \arg\min_{\Omega} |F^(X) - GT_n|^{\frac{1}{2}}, \Theta = \theta_1, ..., \theta_n$$
(3)

It works as learning the global optimization parameters set $\Theta^* = \theta_1^*, ..., \theta_n^*$ for each single network in the stacked hierarchical networks. The criterion only focus on the minimization of the error of the final single network in the stacked networks despite of a combination criterion for the minimization of error from all single networks. Both the stacked hierarchical network architecture and it's end-to-end training strategy are illustrated in Fig2.

Application 3.1 Consider a single network Y = f(X), where X is the input, Y is the output, GT is the ground truth, f is the single network overall function, S is the architecture of the single network. Practically, this network is able to achieve the error boundary $E^* = |y - y^*|^{\frac{1}{p}}, p \ge 1$. In the real world situation, in order to increase the learning ability of the single network, that is to reduce the error boundary from E^* to E', E' < E^* , stacked networks are designed as Y = F(X), where $S_{k+1} = S_{k+1}$

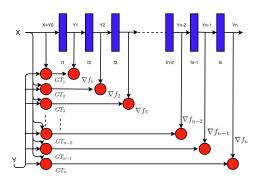


Figure 3: Stacked hierarchical networks for the proposed application.

 $S_k, k=1,..,n-1, GT_k=\eta_k^1\times X+\eta_k^2\times GT, \sum_{i=1}^2\eta_k^i=1$, following the proposed definition and strategy and k is the kth single network, $k\in\{1,2,...,n\}$. The deep architecture for this application is illustrated in Fig3.

3.2 Application for Category Specific 3D Reconstruction from A Single-view silhouette image

Also, before the stacked networks, a replication layer is added to work as replication of single-view 2D segmentation for producing 3D segmentation. In the task of 3D reconstruction from a single-view silhouette image, the single-view silhouette image input is 50×50 , and the replication layer simply works to replicate the 2D segmentation 50 times to produce a 3D segmentation in

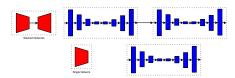


Figure 4: Two network architectures. The top are the concise architecture(top left) of stacked hierarchical networks and the detailed architecture(top right). The bottom are the concise architecture(bottom left) of the single network [30] and the details architecture(bottom left), the replicate layer is not represented.

the 3D grid space. While voxels in the space gets binary value, 1 represents that the voxel is in the segmentation, 0 represents that the voxel is not in the segmentation. The architecture of the stacked deep architecture is illustrated in Fig4.

Furthermore, in order to train the proposed deep architecture from end to end. The proposed training strategy is applied. Here, following the proposed training strategy in Eq.(4), Eq.(5), Eq.(6) and Eq.(7).

$$S_1 = S_2, f_1 = f_2 = f$$
 (4)

$$\nabla f_1 = |Y_1 - 0.5 \times X - 0.5 \times GT|^{\frac{1}{2}} \tag{5}$$

$$\nabla f_2 = |Y_2 - GT|^{\frac{1}{2}} \tag{6}$$

$$\Theta^* = \arg\min_{\Theta} |F(X) - GT|^{\frac{1}{2}}, \Theta = \{\theta_1, \theta_2\} \quad (7)$$

4 Evaluation

Experiments for 3D reconstruction from a single-view silhouette image are conducted to evaluate the performance of the proposed stacked networks and the end-to-end training strategy. To demonstrate the proposed model's ability for class-specific reconstruction from a single-view silhouette image, experiments are conducted for two object categories including the car and the plane. We also conduct 3D reconstruction experiments for both the staked networks and the single network [30] Fig4,

and demonstrate the improvement of the proposed stacked network both qualitatively and quantitatively. Furthermore, results are calculated in voxel IoU in order to make comparison with the state of art single-view reconstruction work.

4.1 Dataset

ShapeNet is a richly-annotated, large-scale repository of shapes represented by 3D CAD models of objects. It contains more than 3,000,000 models, 220,000 models out of which are classified into 3135 categories. We take use of 2 categories of the ShapeNet, planes, and cars for the evaluation of the our reconstruction framework. For each CAD model of the training and testing dataset, we project a CAD model from a large set of single views to get single-view silhouette images of the object.

4.2 Training

We evaluate the proposed single-view reconstruction framework for 2 object categories including cars and planes. For each object category, we randomly pick 100 different CAD models. We train 78 CAD models separately for each object category. As presented, for each CAD model, 180 single-view silhouette images are produced through projection from views widely located on a sphere around the model. Therefore, we train 78×180 single-view silhouette images for each object category. Similarly, for the test, we test 22 CAD models and pick up 180 single-view silhouettes images for each model. Therefore, 22 × 180 single-view silhouette images are test for each object category. In order to measure the quality of the reconstruction shape quantitatively, we measure the voxel overlap between the reconstructed shape and the ground truth shape(voxel IoU). Also, to represent the accuracy improvement of our stacked networks compared with the single network [30], we evaluate rebuilt shapes produced from the two networks for comparison. As illustrated in Fig5, for the reconstruction experiments of cars and planes, we follow the Eq.(8)

$$\alpha_1 = 10l_1, \beta_2 = 20l_2, l_1 \in \{1, 2, ..., 18\}, l_2 \in \{1, 2, ..., 18\}$$
(8)

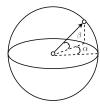


Figure 5: A widely located single view on a sphere.

to get 180 single views on a sphere around each object. And 2D segmentation of the object is projected from each single view. Fig1 demonstrates an example for the choice of single views picked for a car.

4.3 Evaluation on Sphere-surface Located Single-view Silhouette Images (Experiment1)

For this test, we test 22 CAD models for each category and 180 single-view silhouette images are rendered for the test of each model. A single view for each single-view silhouette image is picked following the same rule in the training. That is to say taht the views picked for test is the same views in the training. These views are widely located on the sphere around of the object. Therefore, there are totally 22×180 single-view silhouette images for test for each object category.

As illustrated in Fig6, for 3D reconstruction of cars from a single-view silhouette image, and only a small part of 3D shape is available from a single view, such as the bottom view, the top view, the front view and the back view, it's challenging to reconstruct a good quality of 3D shape from a 2D silhouette image where most part of the 3D shape are unavailable(hard single-view silhouette images). It's reasonable that the single network is only able to reconstruct a very coarse 3D shape in this case. However, for the proposed stacked network and its end-to-end training strategy, a good quality 3D shape is reconstructed even most part of the shape is unavailable from the single-view silhouette images. We test 440 hard single-view silhouette images totally for the reconstruction of cars from a single-view silhouette image.

Similarly, as illustrated in Fig6, for 3D reconstruction of planes from a single-view silhouette image, the single network is able to rebuild a 3D shape but the shape

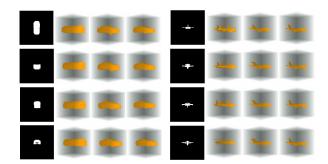


Figure 6: Reconstruction visualization from a hard single-view silhouette image(Experiment1). The left part represents visualization for cars and the right part represents visualization for planesFor each part, the first column represents picked hard single-view silhouette image of an object. The second column represents the rebuilt 3D shape reconstructed from a single network given the corresponding silhouette images. The third column represents the rebuilt 3D shape reconstructed from the proposed stacked networks. The fourth column represents the ground truth shape.

is not reasonable as the silhouette images from these single views do not contain necessary information, such as where to put the airfoil or the tail, what dose the head look like when seen from the back. However, our stacked networks are able to reconstruct good 3D shape for this case, fix the tail airfoil at the right place and reconstruct a good head even these is no information from these views. Also, the number of hard single-view silhouette images of planes for this test is 440.

Also, as illustrated in Fig7, for 3D reconstruction of cars from a single-view silhouette image, and most part of the shape is available in each silhouette image from a single view(easy single-view silhouette image), such as the side views, the left views and the right views, the single network is able to build a reasonable shape but these shape have obvious drawbacks such as big holes, fractures, partial shape damage, partial coarse shape. However, the proposed stacked networks and its end to end training strategy help the reconstruction to overcome these drawbacks, and the 3D reconstructed shapes are very close to the ground truth shapes. We test 3520 easy single-view silhouette images totally for the reconstruction of cars from

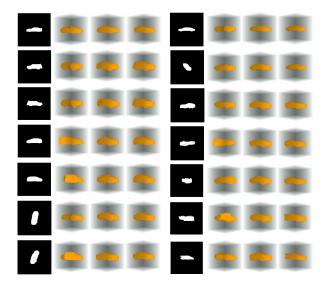


Figure 7: Reconstruction visualization from a easy single-view silhouette image for cars(Experiment1). The first column represents picked easy single-view silhouette images of an object. The second column represents the rebuilt 3D shape reconstructed from a single network given the corresponding silhouette images. The third column represents the rebuilt 3D shape reconstructed from the proposed networks. The fourth column represents the ground truth shape.

a single-view silhouette image.

Also, as illustrated in Fig8, for 3D reconstruction of planes from a single-view silhouette image, and most part of the shape is available in each silhouette image from a single view, the single network is able to rebuild most part of 3D shape of a plane and some parts are missing such as the tail, the wing, the head, the back. Also, it sometimes fix the tail at the wrong position. However, the proposed stacked networks and its training strategy help the reconstruction to solve these issues and rebuild good shapes. Also, the number of easy single-view silhouette images of planes for test is 3520.

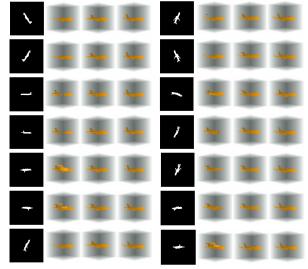


Figure 8: Reconstruction visualization from a easy single-view silhouette image for planes(Experiment1). The first column represents picked easy single-view silhouette images of an object. The second column represents the rebuilt 3D shape reconstructed from a single network given the corresponding silhouette images. The third column represents the rebuilt 3D shape reconstructed from the proposed stacked networks. The fourth column represents the ground truth shape.

4.4 Evaluation on Sphere-surface Randomly Located Single-view Silhouette Images (Experiment2)

For this test, we test 22 CAD models for each category and 180 single-view silhouette images are rendered for the test of each model. A single view for each silhouette image is picked following a different rule Eq.(9) from the training to ensure the all the views are randomly located on a sphere around a object and these views are different with the trained views. Where $l_1 \in \{1, 2, ..., 18\}, l_2 \in \{1, 2, ..., 18\}, \gamma \in (0, 1)$ is a random number.

$$\alpha_1 = 10(l_1 - \gamma), \beta_2 = 20(l_2 - \gamma) \tag{9}$$

This experiment is more challenging than experiment 1 as that the silhouette images are projected from randomly located single views on the sphere, and the views are different from the views used for training. And we test 22×180 single-view silhouette images in total for each object category.

As illustrated in Fig9, for reconstruction of cars from a single-view silhouette images when the views are towards randomly located. and for a single view that small part of shape is available or a large part of shape is available, the single network is able to reconstruct a small part of the car, rebuild a coarse 3D shape that is far from a car structure. Also the single network sometimes fails to produce a reasonable shape and much noise occurs in the grid space. However, the stacked networks are able to reconstruct good quality of shapes without those drawbacks.

Similarly, as illustrated in Fig10, for reconstruction of planes from a single-view reconstruction when the views are randomly located on a sphere surface, and for a single-view silhouette image from views where both a small part of shape or a large part of shape is available, the single network rebuild planes that parts of the plane such as wings, tail are missing or these parts are fixed at the wrong position. Also, it sometimes fails to reconstruct reasonable shapes of planes. However, the proposed stacked networks reconstruct good quality of shapes which overcomes these issues.

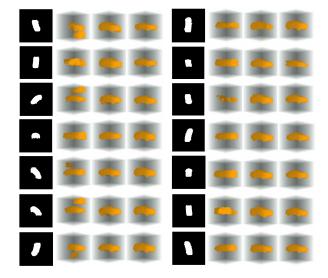


Figure 9: Reconstruction visualization from a randomly located single-view silhouette image for cars(Experiment2). The first column represents picked single-view silhouette images of an object. The second column represents the rebuilt 3D shape reconstructed from a single network given the corresponding silhouette images. The third column represents the rebuilt 3D shape reconstructed from the proposed stacked networks. The fourth column represents the ground truth shape.

	3D-R2N2	S-Hard-E1	S-All-E1	SS-Hard-E1	SS-All-E1	S-All-E2	SS-All-E2
Cars	0.798	0.699	0.765	0.817	0.828	0.601	0.686
Planes	0.513	0.373	0.469	0.473	0.474	0.384	0.430

Table 1: Average IoU for car category and plane category. In the table, S represents the single network, SS represents the proposed stacked networks, E1 and E2 represents experiment1 and experiment2, Hard represents hard single-view silhouette images, All represents silhouette images for all test single views.

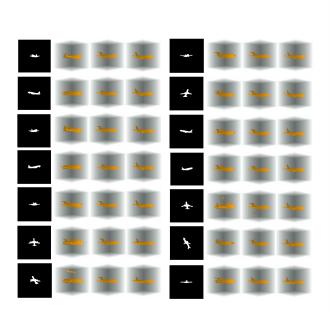


Figure 10: Reconstruction visualization from a randomly located single-view silhouette images for planes(Experiment2). The first column represents picked single-view silhouette images of an object. The second column represents the rebuilt 3D shape reconstructed from a single network given the corresponding silhouette images. The third column represents the rebuilt 3D shape reconstructed from the proposed stacked networks. The fourth column represents the ground truth shape.

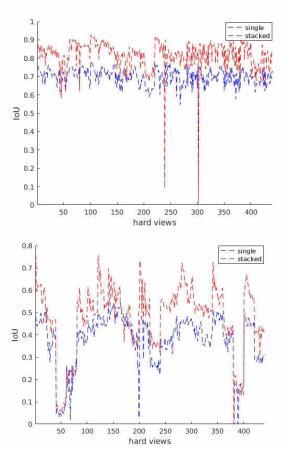


Figure 11: Voxel IoU measure of a single network and stacked networks for experiment1. For cars, voxel IoU for each hard single-view silhouette image are represented(top). Similarly, the IoU for hard single-view silhouette of planes are represented(bottom).

4.5 Comparison between Stacked Networks, 3D-R2N2, Single Networks.

Comparison between the single network [30], proposed stacked networks and the state of art 3D-R2N2 [3] is made. Reconstruction results are calculated in voxel intersection-over-Union(IoU) for the above two experiments. Experiments are conducted with the same configuration for the single network and stacked networks expect 3D-R2N2 which works for the input with RGB textures. As 3D-R2N2 works for RGB images rendered from Shapenet [3] as input in a specific sequence which dose not works for a single-view silhouette image. We compare IoU of the experiments from both the single network and the stacked networks with the same object category test [3] in Shapenet. And also compare the average IoU for the same object category from Shapenet that 3D-R2N2 gets from a single-view RGB image [3].

As illustrated in Tab.1, the average IoU value of the stacked networks is higher than 3D-R2N2 for cars in the experiment1. Although the stacked networks achieves the same IoU lever for planes that is 0.04 lower than [3], but we consider our proposed framework is able to build object category specific reconstruction from a single view silhouette image in a good quality also as represented in Fig6, Fig7, Fig8, Fig9 and Fig10. And the network is demonstrated to be able to do 3D reconstruction from single-view with reduced dependence from RGB texture images to 2D silhouette images. Furthermore, the difficulty of experiment in our work is higher than 3D-R2N2 [3] that the input is reduced to a single-view binary silhouette image. And the views get more variation that the rendered RGB images are projected randomly 5 times on a circle [3], while the single-view silhouette images are rendered 180 times from widely located views on a sphere which is towards random single-view configuration.

Also, the stacked networks achieve higher IoU value than single network Tab.1. Even the increased IoU is about 0.1 and 0.07 higher, but as the Fig6, Fig7, Fig8, Fig9 and Fig10 demonstrate, this improvement demonstrates that the proposed stacked network reconstruct much higher quality shapes than the single network. In detail, we also demonstrate each IoU for the hard single-view silhouette image in experiment1 between the single network and the stacked networks Fig11. The stacked networks achieves distinctly higher IoU than the single net-

work for each single-view silhouette image.

As illustrated in Tab.1, for the experiment of randomly located single-view silhouette images, 74.8737% of 3960 single-view silhouette images of cars achieves high IoU, the percentage for planes is 52.8283% of 3960 randomly located single-view silhouette images of planes. The IoU value is about 0.07 lower than 3D-R2N2 [3]. We consider contribution as the reconstructed shape is in good quality as illustrated in Fig9 and Fig10. Also, the proposed networks work from single-view silhouette images without RGB textures and the views are widely and randomly located around a sphere in spite of a circle [3]. Furthermore, in this experiment, all the views are different in the training.

Also, in order to ensure that the all the average IoU values are credible, we also calculate the standard error, and the standard error is around 10^{-4} which is very low.

5 Conclusion

Firstly, the proposed reconstruction framework is represented to reconstruct good quality 3D category specific shape from a single-view silhouette image, while the state of art of 3D reconstruction from a single view depends on RGB textures and other extra input. To our best knowledge, for object specific category, we firstly demonstrate that networks help the task, 3D reconstruction from single view, reduce the dependence of RGB textures to binary segmentation.

Secondly, a novel network design is proposed for stacked hierarchical networks combined of a number of single networks. Also an end-to-end training strategy for the stacked networks are proposed to help the complex stacked networks find global optimization solution during training. As the application of the 3D reconstruction from a single-view silhouette image demonstrates, this contribution improves the networks' ability in this reconstruction task to overcome many drawbacks and even failures. Further work to the development of complex stacked networks in this end-to-end training direction may be promising.

Thirdly, the proposed networks are demonstrated to build 3D reconstruction from a single view that are both widely and randomly located on a sphere around a object. While to our best knowledge, other network frameworks are demonstrated to build 3D reconstruction from a single view that is randomly located on a circle or partially located on a sphere. Therefore, the proposed network framework is able to reconstruct 3D shape more towards a randomly located single view than the state of art. Further work may include 3D reconstruction for a single view randomly located in space in spite of the surface of a sphere.

Finally, although the stacked hierarchical networks are trained end to end, but the replication layer is not currently in the end to end training strategy. Because we currently reconstruct the ground truth 3D shape to ensure that the 3D shape is fixed without rotation and translation. The future work includes the development of a more completed training strategy for the whole shape reconstruction framework, and some new single networks may be added in the proposed framework for the rotation and translation of a reconstructed shape.

References

- [1] J. Carreira, A. Kar, S. Tulsiani, and J. Malik. Virtual view networks for object reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2937–2946, 2015. 3
- [2] Y. Chen and R. Cipolla. Single and sparse view 3d reconstruction by learning shape priors. *Computer Vision and Image Understanding*, 115(5):586–602, 2011. 1
- [3] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. arXiv preprint arXiv:1604.00449, 2016. 1, 3, 10
- [4] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *International Journal of Computer Vision*, 40(2):123–148, 2000.
- [5] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007.
- [6] E. Delage, H. Lee, and A. Y. Ng. Automatic single-image 3d reconstructions of indoor manhattan world scenes. In *Robotics Research*, pages 305–321. Springer, 2007. 1
- [7] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems* 27, pages 2366–2374. Curran Associates, Inc., 2014. 3

- [8] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010. 1
- [9] R. Garg and I. Reid. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. *arXiv preprint arXiv:1603.04992*, 2016. 2, 3
- [10] F. Han and S.-C. Zhu. Bayesian reconstruction of 3d shapes and scenes from a single image. In *Higher-Level Knowledge in 3D Modeling and Motion Analysis*, 2003. HLK 2003. First IEEE International Workshop on, pages 12–20. IEEE, 2003. 1
- [11] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. ACM transactions on graphics (TOG), 24(3):577– 584, 2005.
- [12] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *Computer Vision* and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 3121–3128. IEEE, 2011. 1
- [13] A. Kanazawa, D. W. Jacobs, and M. Chandraker. WarpNet: Weakly Supervised Matching for Single-view Reconstruction. arXiv preprint arXiv:1604.05592, 2016. 3
- [14] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Category-specific object reconstruction from a single image. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1966–1974. IEEE, 2015. 1, 3
- [15] S. Kim, K. Park, K. Sohn, and S. Lin. Unified Depth Prediction and Intrinsic Image Decomposition from a Single Image via Joint Convolutional Neural Fields. arXiv preprint arXiv:1603.06359, 2016. 2, 3
- [16] P. Koutsourakis, L. Simon, O. Teboul, G. Tziritas, and N. Paragios. Single view reconstruction using shape grammars for urban environments. In 2009 IEEE 12th international conference on computer vision, pages 1795–1802. IEEE, 2009.
- [17] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000. 1
- [18] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. Artificial intelligence, 31(3):355–395, 1987.
- [19] M. R. Oswald, E. Töppe, K. Kolev, and D. Cremers. Non-parametric single view reconstruction of curved objects using convex optimization. In *Joint Pattern Recognition Symposium*, pages 171–180. Springer, 2009. 1
- [20] E. Prados and O. Faugeras. Shape from shading: a well-posed problem? In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2, pages 870–877. IEEE, 2005.

- [21] M. Prasad. Class-Based Single View Reconstruction. PhD thesis, University of Oxford, 2009. 1
- [22] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2009.
- [23] N. Snavely et al. Bundler: Structure from motion (sfm) for unordered image collections. Available online: phototour. cs. washington. edu/bundler/(accessed on 12 July 2013), 2010. 1
- [24] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2686– 2694, 2015. 3
- [25] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3D Models from Single Images with a Convolutional Network. In *European Conference on Computer Vision*, pages 322–337. Springer, 2016. 1, 3
- [26] E. Töppe, M. R. Oswald, D. Cremers, and C. Rother. Image-based 3d modeling via cheeger sets. In *Asian Conference on Computer Vision*, pages 53–64. Springer, 2010.
- [27] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 41–48. IEEE, 2014. 1
- [28] C. Wu. Towards linear-time incremental structure from motion. In 2013 International Conference on 3D Vision-3DV 2013, pages 127–134. IEEE, 2013. 1
- [29] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single Image 3D Interpreter Network. arXiv preprint arXiv:1604.08685, 2016. 1, 3
- [30] R. D. X. Di and M. Prasad. Deep shape from a low number of silhouettes. In *Computer Vision ECCV 2016 Work-shops*. Springer, 2016. 3, 5, 10
- [31] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2016.
- [32] L. Zhang, G. Dugas-Phocion, J.-S. Samson, and S. M. Seitz. Single-view modelling of free-form scenes. The Journal of Visualization and Computer Animation, 13(4):225–235, 2002.