Learning to Construct 3D Image from Single 2D Image

Construction of 3D images from single 2D counterpart using Deep Learning

Muhammad Awais

09.06.2019

2019310148

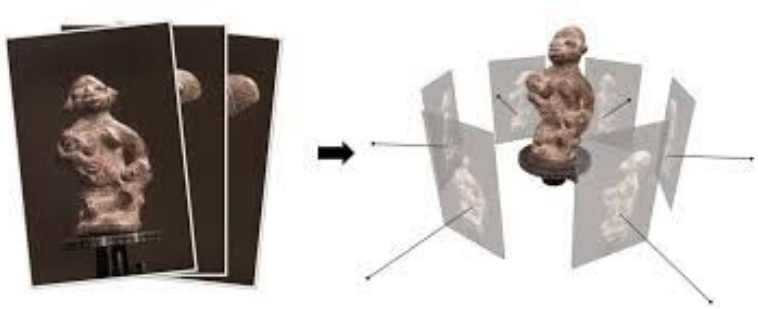
Table of Contents

ABSTRACT

Construction of 3D information from multiple 2D image is well studied problem but similar 3D construction from single image is a difficult problem. With the advent of deep learning and its success in many computer vision tasks, this problem gain some interest. Many new papers have been published recently that tries to solve this problem from different perspectives. In this report, we categorized this problem and tried to summarize a few of these recent methods. We first formulated the problem and discussed a few datasets that are used in this problem. Then we summarize 4 recent attempts to solve this problem. We also proposed a GAN based method to solve this problem.

INTRODUCTION

When we take a photo from a common camera, we essentially project a 3 dimensional image to 2 dimensional space losing much of the information such as depth, relative position etc. Conventionally, multiple images are required to reconstruct 3D representation such as shown in following image.



A simple 3D construction from three images [link]

In these conventional methods, multiple image images, camera position etc. are required and problem is posed as projection of each 2 dimensional point of multiple images to one point on 3 dimensional space. This kind of 3D construction is very useful such as we can use it to construct exact replicas of old archaeological objects for further study in the lab, construct replicas of important buildings, use 3D models in games, in virtual reality or this can be used as a visual aid in education or entertainment etc.

But as this kind of setup requires multiple expensive cameras, calibration and many other technical specialities which makes it difficult to use in broader applications. Also, there are many applications where it is very difficult, even impossible, to take multiple images. For instance, historical image might not have multiple shots. Similarly, in extraterrestrial missions (curiosity rover on Mars for example), it is difficult and redundant to mount multiple cameras to take multiple images of the same spot. But it is also important to have 3D information as this kind of information can help us further our understanding.

To circumvent the above-mentioned problems, single image to 3D construction have been used. Conventionally, many strong priors are assumed on the image to reconstruct its 3D equivalent which prohibits its wide scale applications. Using deep learning to learn the prior can solve this problem. Hence, there is an emerging interest in this problem recently after the successful application of deep learning models on many computer

vision models. After the release of many large scale 3D datasets, it has become relatively convenient to apply deep learning on this problem.

In this report, we will summarise some of the recent deep learning based techniques to solve this problem.

Problem Formulation

Let I be a 3D image and V is it equivalent volumetric representation. Then our goal is to learn a function f such that $\hat{V} = f(I)$ and $||f(I) - V||_p$ is minimized. From deep learning perspective, learned function is deep neural network, either a simple convolutional neural network (CNN) or some generative model such as Generative Adversarial Network or Encoder-Decoder setup. Based on available datasets, we can further divide this problem into two parts i.e. supervised learning where we have single image as well as its volumetric representation. In this setup, neural network can be trained to learn corresponding function. Second option is where we only have some information about the image such as pose etc. In this setup, we can pose problem as weakly supervised problem. Following image visually represents the problem.

A simple 3D construction from single image [link]

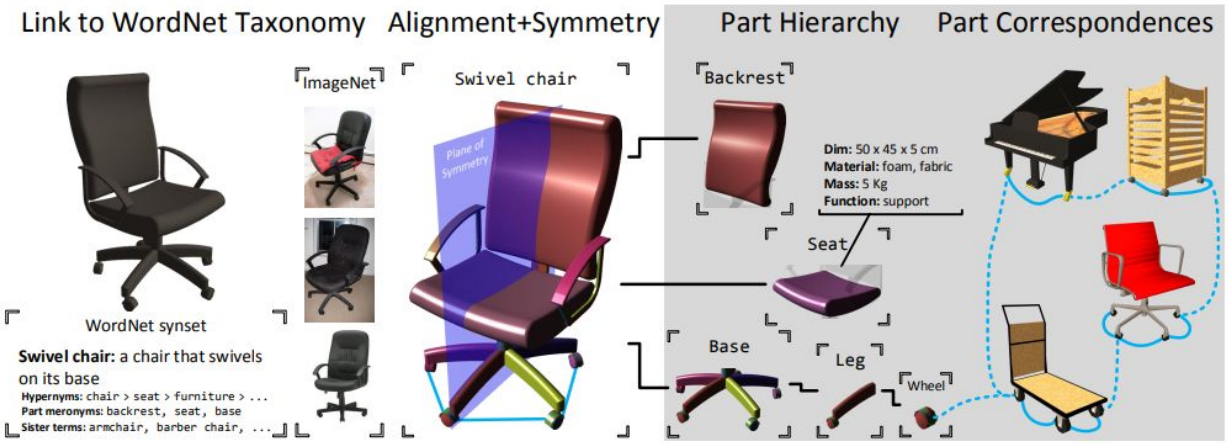
A similar problem is to generate a 3D image of a class by leanrened model without conditioning on any image. This problem is also interesting as it can have many applications in games, virtual reality sets etc. But, here we will only see single shot to 3D construction but the same setup can also be used for this application.

Available Datasets

Deep learning based models heavily rely on large scale datasets contrary to conventional methods. Due to this, it is important to detail datasets that are mostly used in single shot 3D image reconstruction problem.

ShapeNet

ShapeNet is a large scale dataset consisting of 3 million 3D CAD models belonging to 4000+ categories. Each image also has annotations and relation with WordNet. It also has part hierarchy and relationship of the data with other models.



A simple example of ShapeNet dataset illustrating different aspects of it.

A Large Dataset of Object Scans

This dataset consists of 10,000 3D scans of real objects in RGBD.

Other Datasets

Some other datasets that are mostly used are also listed as following

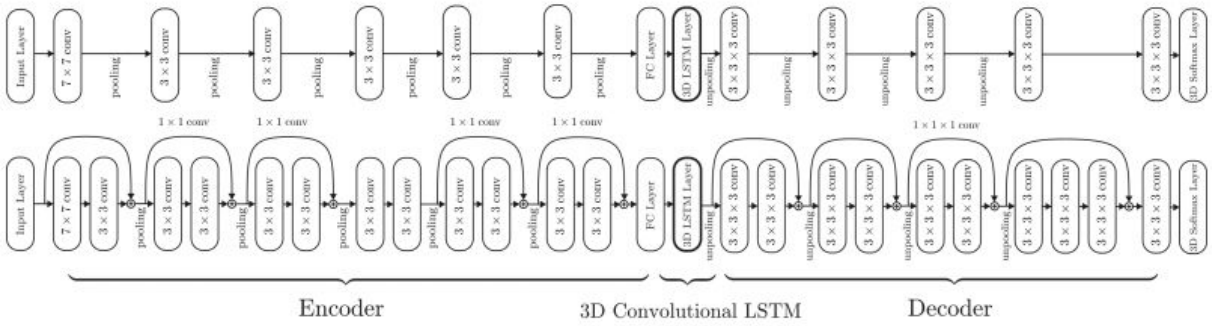
Title	Description	Classes	Number of Images
ModelNet	CAD Models of objects from many different categories	662	127915
IKEA 3D models and aligned images	Furniture models	219	759
Open Surfaces	Annotated surfaces from real world		
Object3D	Data for 3D object recognition	100	90127
Thingi10K	Models for 3D printing of objects		10000
SUNCG	3D models of indoor scenes		45000

Overview of Techniques

We have divided deep learning based single-image to 3D reconstruction methods into two main categories based on the supervision used. Some models use 3D supervision i.e. they require pair of 2D images along with 3D representation of the object while other models only use 2D images along with some other information such as pose etc.

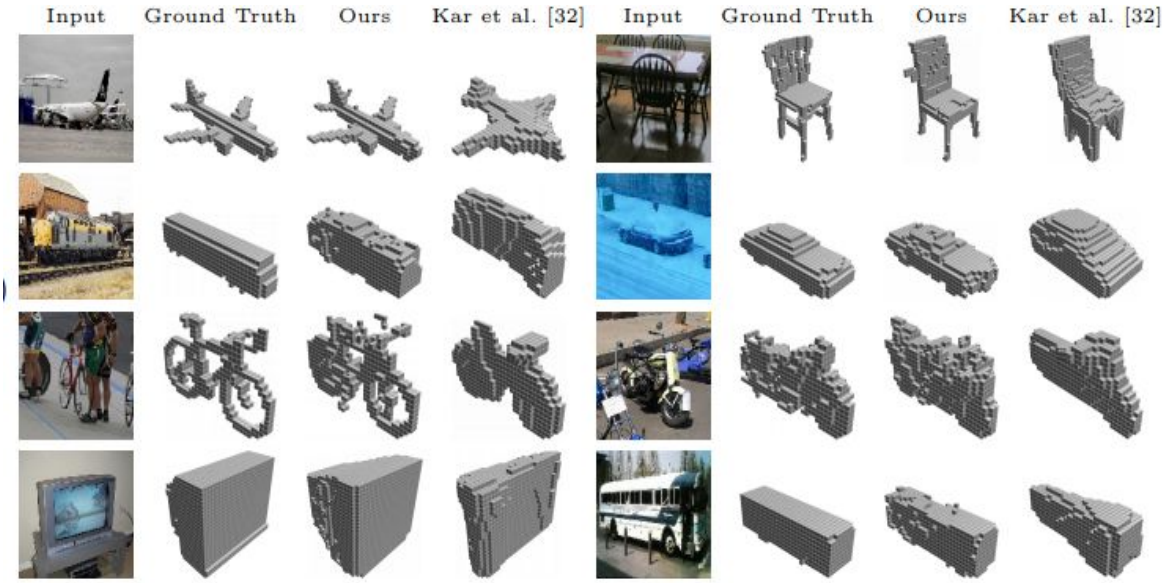
3D Supervised

These methods use a deep neural network along with pair of 2D and 3D images. We first discuss this paper which introduced 3D Recurrent Reconstruction Network (3d-R2N2). Long Short Term Memory (LSTM) is very famous neural network that predicts new output based on the all the previous instances. This paper exploits this property of LSTM along with convolutional layers to construct 3D view from one view point of an object. The network consists of a 2D convnet, 3D Conv LSTM and 3D Deconv CNN as shown in following figure.



Architecture of 3D-R2N2

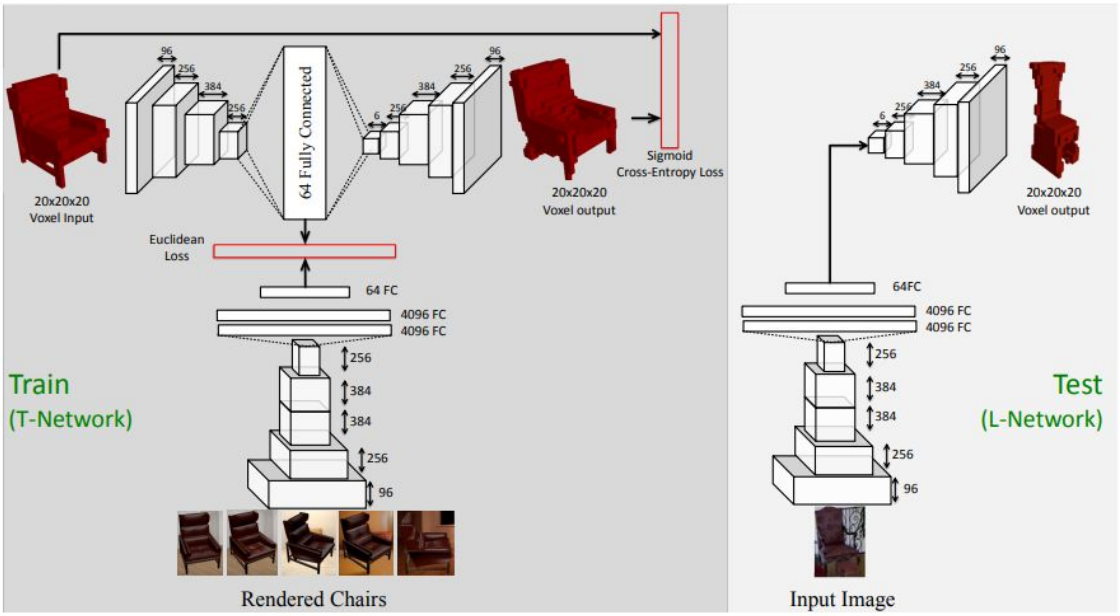
Given an arbitrary view point of an object, 2D-CNN first learns low dimensional features using convolutional layers. Then 3D-LSTM updates it cell states and 3D-DCNN decoder LSTM's states to generate voxel representation. Loss function for the training is cross entropy of voxels. PASCAL VOC and some CAD model based datasets are used for training. The result of this method is shown in the following figure.



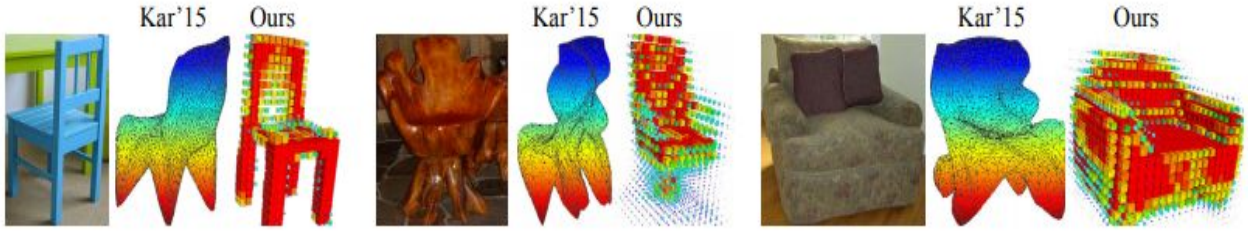
Results of 3D-R2N2

Similarly, this paper first learns a vector representation and then converts them to 3D voxels. In this model, at training time, encoder part of an auto encoder generates latent representation of given 3D voxels and decoder generates similar voxel output from the latent representations. Another CNN is also attached which learns another latent

representation of given images and try to make them as close to voxel based latent representation as possible. At test time though, only one image is fed to this extra CNN which makes a latent representation and fed it to the decoder which then creates 3D voxels of the image. Following image shows this architecture.



Loss function is cross entropy between original and predicted voxels. Some results of this model are shown in the following figure.

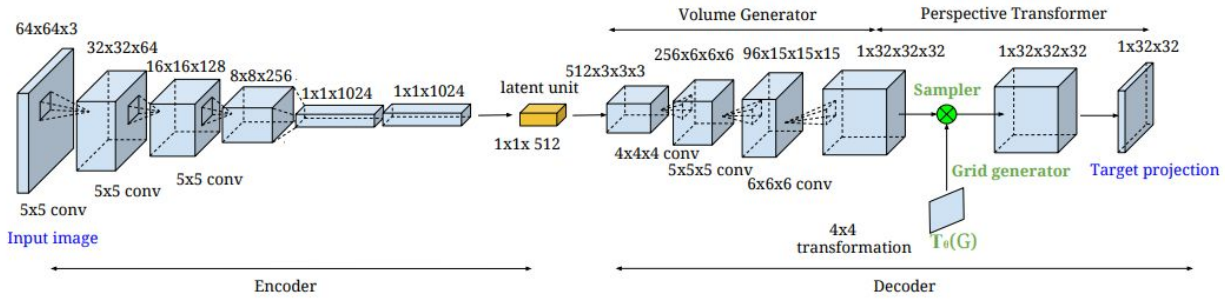


2D Supervised

The 2D supervised methods rely on different information of the 3D model instead of complete 3D information. For instance, the paper we will discuss in the following section relies on the silhouette representation.

The network used by this paper is and encoder decoder setup where an encoder first learn an intermediate representation of given image which is viewpoint invariant. This

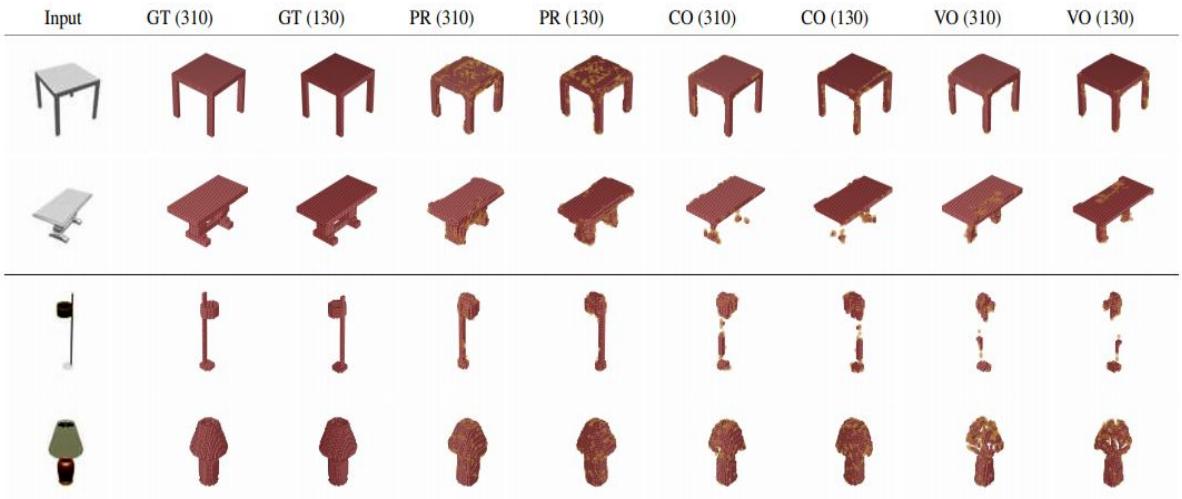
representation is then transformed to a volume. This volume is then used to construct a silhouette which is then compared to the silhouette generated by 2D image. Following figure shows network architecture of this method.



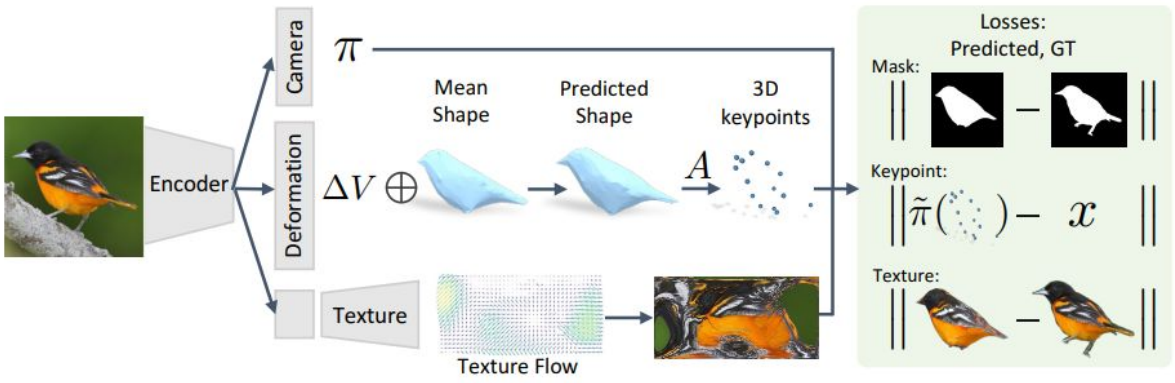
Most interesting part of this paper is their loss function. Given an image I, sillethouse S and camera angle α, loss is defined as;

$$\mathcal{L}_{proj}(I^{(k)}) = \sum_{j=1}^{n} \mathcal{L}_{proj}^{(j)}(I^{(k)}; S^{(j)}, \alpha^{(j)}) = \frac{1}{n} \sum_{j=1}^{n} ||P(f(I^{(k)}); \alpha^{(j)}) - S^{(j)}||_2^2,$$

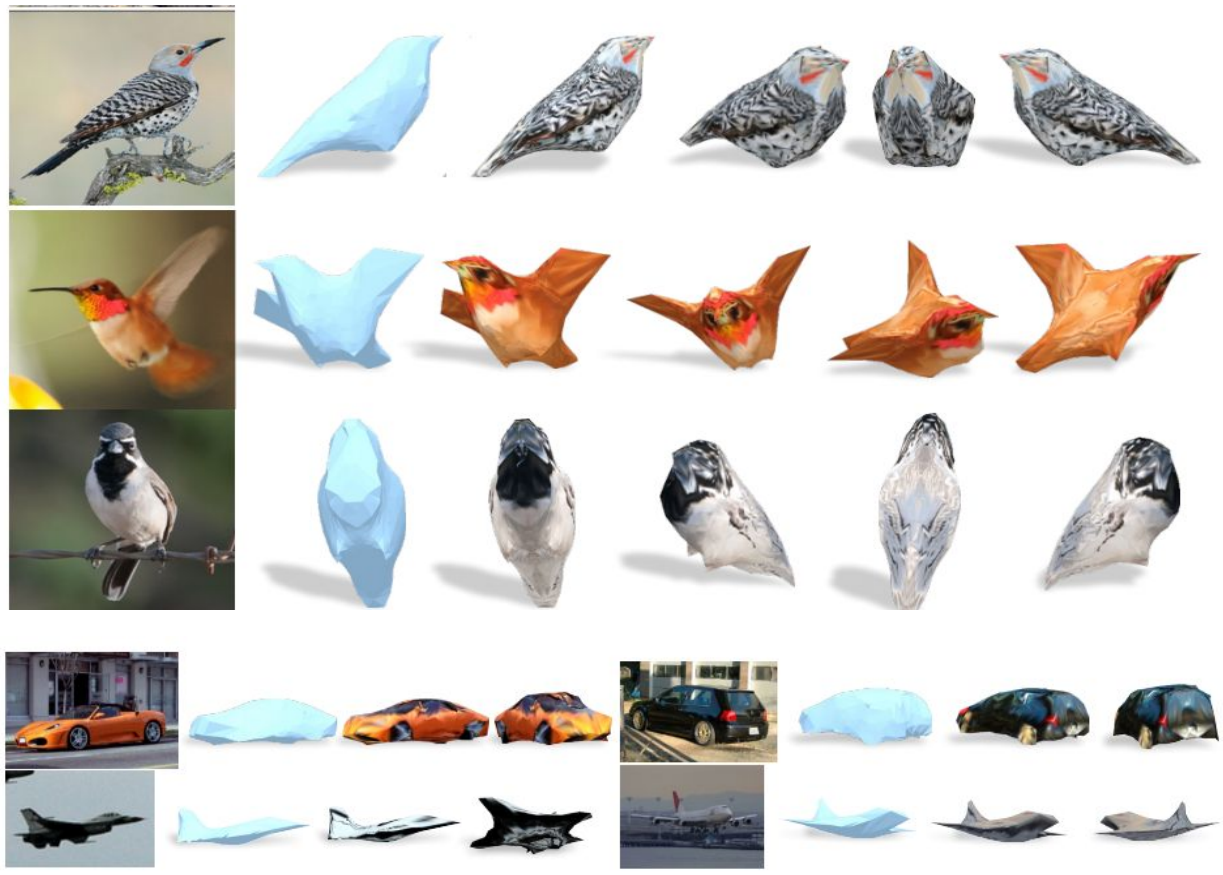
Some results of the paper is shown as following;



This [paper] also uses CNN to learn 3D mesh, camera position and texture. In this approach, many images of same category is used to train a CNN which produces output and a loss of with respect to camera angle, texture and key points from 2D image and from 3D generated mesh is minimized. Following figure shows overall method.



They also used several priors such as smoothness prior, symmetry constraints etc. to enforce consistency. Some results of this paper are shown in the following two figures.



Proposed Improvements

While I have seen many approaches that are based on Auto-Encoders but GAN based approach are very rare. So in this section, I propose a GAN based method. GANs are mostly used to generate natural images and it is a well known fact that GAN generate close to natural images.

In the proposed model, a GAN is first trained to generate 3D voxels image with discriminator of the GAN trained on 3D images. At test time, GAN is reversed i.e. given 3D image, it generates a simple 2D image. Then we start from a random 3D image and fed it to GAN. GAN then generates a 2D representation. This representation is compared with the original image and the random 3D voxel is changed in such a way that 2D image becomes as close to input image as possible.

CONCLUSION

In this report, we first formulated problem of 3D image reconstruction from a single 2D image. We then summarized 4 different deep learning based methods. Two of these methods are 3D supervised i.e. they require both 2D and 3D images to generate 3D image while two of them are 2D supervised i.e. they don't require 3D information even at the training time. Most of these methods use Encoder-Decoder setup to solve this problem. In the last section, we proposed a GAN based different methods to handle the same problem .

REFERENCES

1. Learning single-image 3D reconstruction by generative modelling of shape, pose and shading
2. ShapeNet: An Information-Rich 3D Model Repository
3. A Large Dataset of Object Scans
4. A Curated List of 3D Machine Learning
5. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction
6. Perspective Transformer Nets: Learning Single-View 3D Object Reconstruction without 3D Supervision
7. Learning a Predictable and Generative Vector Representation for Objects
8. Learning Category-Specific Mesh Reconstruction from Image Collections