

3D Building Reconstruction using 2D GANs

Christopher Siyuan Tao
German Swiss International School
Hong Kong, China
206373@learning.gsis.edu.hk

Abstract—There are many historical sites across the globe that hold a huge importance to mankind as it contains the history and culture of the past. But as buildings wither and lose their beauty over the course of time, we lose a precious piece of the heritage of mankind forever. With the growing emergence of new technologies and the internet, we are instead able to preserve these heritage sites through their reconstruction as 3D models using convolutional neural networks and simple 2D vision. We base our work mainly on a previous implementation of 3D shape reconstruction using 2D image GANs by Xingang Pan et al, where we mine for cues such as albedo, viewpoint, lighting and depth such that we can infer a 3D image from these factors. We believe our work is different from other works in that we use purely an 2D image to reconstruct its 3D form, unlike other related works such as Pixel2Mesh, which uses 3D Mesh Models for ground truths. Our work infers a 3D shape and model from simply one 2D image, which cuts costs and is much easier to implement as it does not require any other input other than the singular 2D image.

Keywords—2D image, 3D Reconstruction, GANs, Deep Learning.

I. INTRODUCTION

There are many historical sites across the globe that hold a huge importance to mankind as it contains the history and culture of the past. But as buildings wither and lose their beauty over the course of time, we lose a precious piece of the heritage of mankind forever. With the growing emergence of new technologies and the internet, we are instead able to preserve these heritage sites through their reconstruction as 3D models using convolutional neural networks and simple 2D vision.

Normally 3D reconstructions are much more complex than using a simple 2D image: after all, the 3D cues that can be mined from a single 2D image are rather limited, and many other methods also use things such as multiple images, 3D ground truths and such. These refer to how multiple references are used for 3D information, such as the use of multiple viewpoints or simply straight up providing the 3D information itself as an input. However, for some historical sites in the world, there exists no physical records of them save for images of them [3-4]. Our project places an emphasis on the digital reconstruction of historical sites which are partially damaged or destroyed through physical records such as images. These images can be used with our project to reconstruct these destroyed historical sites digitally as accurately as possible using limited inputs, unlike other projects which require more inputs that may not exist anymore, hence becoming a limitation.

We base our work mainly on a previous implementation of 3D shape reconstruction using 2D image GANs by Xingang Pan et al [1], where we mine for cues such as albedo,

viewpoint, lighting and depth such that we can infer a 3D image from these factors. We believe our work is different from other works in that we use purely an 2D image to reconstruct its 3D form, unlike other related works such as Pixel2Mesh, which uses 3D Mesh Models for ground truths. Our work infers a 3D shape and model from simply one 2D image, which cuts costs and is much easier to implement as it does not require any other input other than the singular 2D image.

II. RELATED WORKS

Xingang Pan et al. explores the possibility of constructing a 3d model from one single 2D image with great success. They mine the different characteristics such as depth, albedo, viewpoint and lighting from the original image, allowing them to reconstruct an image based on these characteristics and perform reconstruction loss optimization. They also use the method of making pseudo samples using random viewpoints and lighting, and performing GAN inversion on them to obtain projected samples, then using these to redefine and shape the final 3D model.

Our work is mainly branched off from this work as we use the similar approach of mining depth, albedo, viewpoint and lighting characteristics and obtaining pseudo and projected samples to construct and define our 3D model. However, in our work we also tackle some of issues that Xingang Pan et al. had encountered such as the problem of extreme poses which would cause the data mined to be incomplete, and we solve this by adding the option of using more than one input images which contains multiple viewpoints of the same building. However, our model is still based on the reliance of one 2D image to form a complete 3D model and these measures are simply to tackle the extremities that we may face.

Cultural heritage 3D reconstruction from historical materials, Dagmāra Krūmiņa [5]. This paper also shows similarity to our work in that our goals align to protect and preserve potentially damaged historical heritage through 3D reconstruction. They adopts two approaches to tackle the issue of reconstructing buildings: image-based and non image based. A variety of different sources such as photographic products, architectural maps and plans and archive documentations are used in the process of reconstruction.

In our work we aim to derive a complete 3D model from a minimum number of 2D images, and we use a CNN to complete and constantly improve our implementation and reconstruction of 3D models of the heritage buildings, especially the reconstruction of undamaged buildings from simply photographs of the damaged buildings, whilst in this paper no machine learning methods are used but CAD software and image reconstruction software are used instead.

Analyzing and Improving the Image Quality of StyleGAN, Tero Karras et al [2]. Xingang Pan et al uses an implementation of StyleGAN 2 in their work, where they improved upon their network through the removal of common blob-like artifacts and through the transformation of the latent code to a mediate latent code through a mapping network, from which they observed better image qualities and projection of images. We will also be implementing StyleGAN2 in our work as we base our work off Xingang Pan et al's work and use StyleGAN2 to generate better quality 3D models in our work.

Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images, Nanyang Wang et al. Another approach in generating a 3D model from a single 2D image can be traced back to this piece of work by Nanyang Wang et al, who undertook a different approach to Xingang Pan et al by generating 3D Mesh Models using a graph based convolutional network and the VGG16 architecture to extract features from the 2D image as opposed to the generation of pseudo and projected samples and refining an ellipsoid to become a depth mask for the 3D model as shown in Xingang Pan et al's work. Nanyang Wang et al also uses a graph unpooling layer to further improve and train their own network to optimize their 3D mesh model as close to ground truth as possible [6].

Although different from our current approach, their work still holds great importance as they achieve the goal of generating 3D models from a single 3D image at a high accuracy and reliability.

Our method is branched off of Xingang Pan et al's work in using 2D GANs to produce 3D models. We aim to preserve historical heritage through the use of convolutional neural networks and GANs such that we can either a) restore the looks of historical buildings from the past using a single 2D image or b) preserve a current historical building using a single 2D image, which is only achievable by referring to above work's to implement our own methodology.

III. METHODOLOGY

Our work is mainly use the similar approach of mining albedo, depth, viewpoint and lighting characteristics and obtaining pseudo and projected samples to construct and define our 3D model.

A. View Network

The view network (Fig. 1) first takes in an input image of $256 \times 256 \times 3$ pixels. It is an encoding network that compresses the image through multiple convolutional and ReLU layers, which acts as a filter and an activation layer that allows the network to "learn" the 3D information, eventually compressing it to an output of size 4 through a Tanh activation function, which defines how the weights of the network converts the input into an output. As is shown in Fig.1.

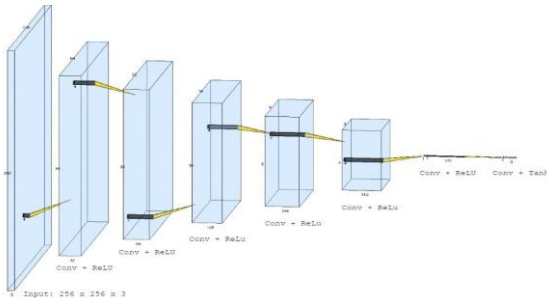


Fig. 1. View Network

B. Depth Network

The depth network (Fig. 2) consists of an encoder and decoder network architecture as the depth is used to construct the projected samples from the pseudo samples. The original input of $256 \times 256 \times 3$ pixels is compressed by many convolutional layers which apply a filter over the original input and mines for features. This is followed by a group normalization layer that divides the channels into groups and normalizes the features within each group. Lastly it is followed by a ReLU layer, a similar activation function to a normal ReLU layer. Then, the output is decoded through many up-sampling convolutional layers which reverses the effect of convolution, and group normalization layers as well as ReLU activation function layers, mining these 3D features to form the projected samples.

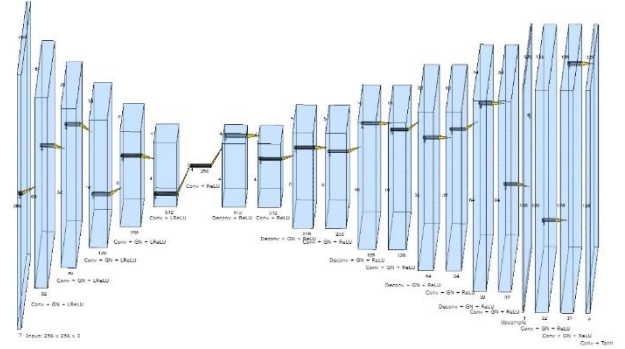


Fig. 2. Depth Network

C. Light Network

The light network (Fig. 3) is similar to the view network in that we do not need to include the decoder structure as the viewpoint and lighting is randomized to form the pseudo samples, which allows the decoder to mine for 3D features and form the better looking projected samples.

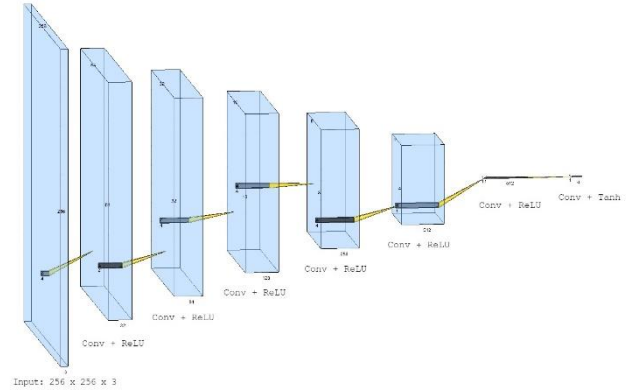
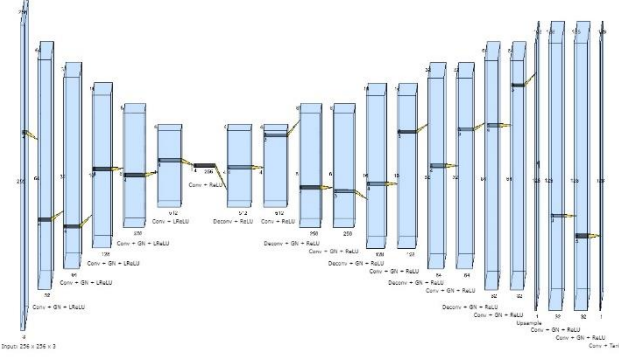
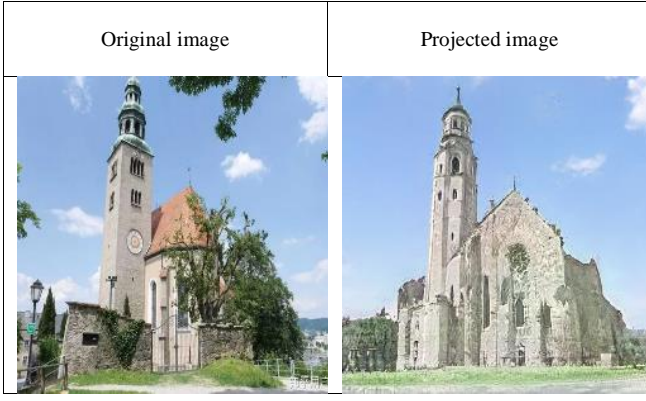


Fig. 3. Light Network

D. Albedo Network

The albedo network (Fig. 4) is also similar to the depth network in that it uses an encoder and decoder network to mine for the albedo features in the pseudo samples using convolutional layers, group normalization and ReLU layers to compress the image into an output of 4, then encodes it using convolutional layers and convolutional layers, as well as group normalization layers and ReLU layers to reconstruct the projected samples using the mined albedo features.





C. Implament details

As mentioned earlier, ReL is used for activation function parameters, and the hyperparameter details are:

Optimizer Adam Learning rate: 1×10^{-4}

Depth map: (0.9, 1.1)

Ellipsoid: (0.91, 1.02)

$\lambda 1$ (in Eq. 2) : 0.01



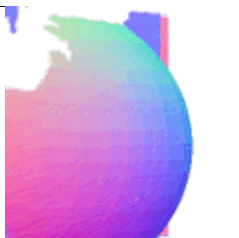
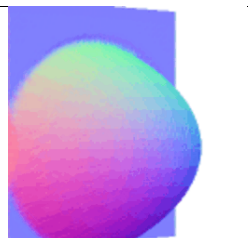
$\lambda 2$ (in Eq. 3): 0.01



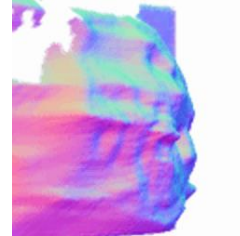
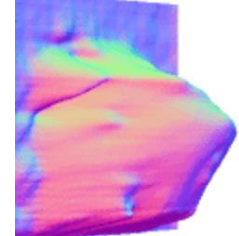
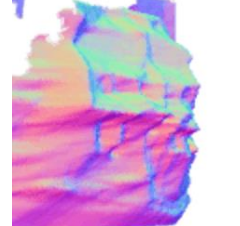
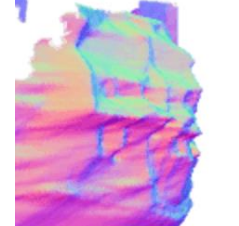
D. Quantitative analysis

The image above was projected using GAN inversion. After that, they go through four stages of relighting and rotating, both masked and unmasked. This is so that the model can mine the different 3D cues from the image from the relighting and rotations. The image on the top left went through a relighting stage whilst the image on the top right went through a rotate stage. Through the Tab. 2, we can see that the results aren't perfect as we can see parts of the building becoming distorted and textures of the building being changed, and the building noticeably becoming wider within the rotated stage. We also see how the initial ellipsoid shape gradually takes on the shape of the building; however with the errors present in the image itself the resulting 3D information does not resemble the original much.

TABLE II.

RESULTS

Items	Demo 1	Demo 2
Relight		
Stage 1		

Items	Demo 1	Demo 2
Relight		
Stage 2		
Stage 3		

Qualitative analysis

These errors (loss curve see Fig. 5) could be potentially caused by the low resolution of the images which makes mining for 3D information much more harder due to the limited size, or it could be caused by some form of interference that covers up a part of the building structure e.g. a tree blocking in front of the building, causing the 3D model to be incomplete and flawed.

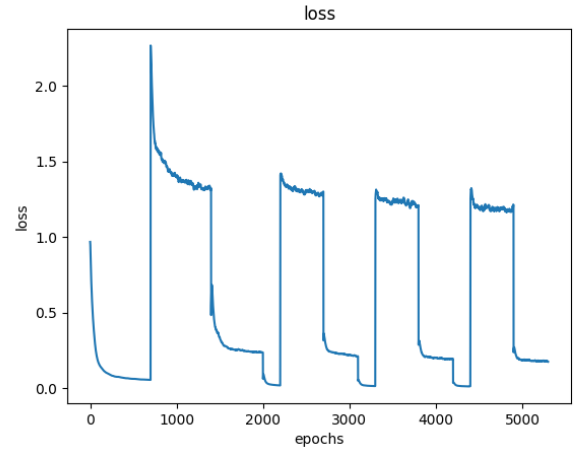


Fig. 5. Loss curve

V. CONCLUSION

This paper utilizes off-the-shelf 2D GANs to recover the shape of 3D objects from images. We find that existing 2D GANs inherently capture enough knowledge to recover the 3D shape of many object categories. Based on weakly convex priors, our method can explore viewpoint and lighting changes

in GAN image manifolds, and exploit these changes to iteratively refine the underlying object shape. We further demonstrate 3D-aware image manipulation on only a single 2D image of a building. The results reveal the potential of 2D GANs for modeling the underlying 3D geometry of 2D image manifolds.

There are two points worth noting in the study:

The influence of prior shape. When the ellipsoid can gradually refine the shape during iterative training. But with flat shapes, the results get worse because these shapes cannot show viewpoint and lighting changes.

Effect of dataset size. The effect of dataset size used to train GANs. Reducing the dataset size from 160k to 3k only slightly degrades performance. When the dataset size is reduced to 1k, the performance deteriorates significantly, as the image quality of GAN starts to deteriorate with insufficient data. Recent GAN data augmentation strategies may address this issue [7].

REFERENCES

- [1] Pan, X. G., & Tsui, S. (2021). (rep.). Do 2D GANs Know 3D Shape? Unsupervised 3D Shape Reconstruction from 2D Image GANs. Retrieved 2022, from <https://github.com/XingangPan/GAN2Shape#do-2d-gans-know-3d-shape-unsupervised-3d-shape-reconstruction-from-2d-image-gans>.
- [2] Karas, T., Laine, S., Aittala, M., Aila, T., Lehtinen, J., & Hellsten, J. (2019). (tech.). Analyzing and Improving the Image Quality of StyleGAN. Retrieved 2022, from <https://arxiv.org/abs/1912.04958?amp=1>.
- [3] Krūmiņa, D. (2019). (tech.). Cultural heritage 3D reconstruction from historical materials. Retrieved 2022, from https://www.clge.eu/wp-content/uploads/2019/04/Winner_GIS_Dagmara_Krumina_LV_3D-reconstruction-historical-materials.pdf.
- [4] Wu, S., 2022. Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the Wild. [online] Elliottwu.com. Available at: <https://elliottwu.com/projects/20_unsup3d/> [Accessed 23 July 2022].
- [5] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In ICCV, pp. 4432–4441, 2019.
- [6] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. ECCV, 2020.
- [7] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In Proc. NeurIPS, 2020a.

[1] Pan, X. G., & Tsui, S. (2021). (rep.). Do 2D GANs Know 3D Shape? Unsupervised 3D Shape Reconstruction from 2D Image GANs. Retrieved 2022, from