3D Scene Reconstruction from Multi-View Stereo Images Using Machine Learning

Ya-Chu Tsao (1), Pai-Hui Hsu (1)

[1] Nat. Taiwan Univ., No.1, Sec. 4, Roosevelt Road, Taipei, 10617, Taiwan
[2] Inha University, 100 Inha-ro, Michuhol-gu, Incheon 22212, Korea
Email: r07521807@ntu.edu.tw; hsuph@ntu.edu.tw

KEY WORDS: Machine learning, Multi-view stereo, Scene reconstruction

ABSTRACT: 3D scene model is the basic data model in 3D GIS (Geographic Information System) which can be used for 3D geo-visualization and scene analysis. Commonly the 3D scene can be reconstructed by means of LiDAR and photogrammetry technologies, however most of the methods are time-consuming and not fully automatic. How to efficiently and automatically reconstruct the 3D scene models has become an important research issue. This paper proposes a 3D scene reconstruction method from multi-view stereo (MVS) images based on machine learning. Similar to the stereo-pair for 3D vision, the multi-view stereo mimics the human visual system (HVS) to acquire 3D information from multiple overlapping images. Because of the multiple view of an object, the problem of occlusion can be overcome. However, the complex geometric relationship between multiple view stereo images also increase the difficulty of calculation. To make the processing of 3D reconstruction more efficient and automatic, a novel method based on machine learning was introduced. Machine learning is a subset of Artificial Intelligence (AI) that provides the ability to automatically learn from data and improves from experience without too much manual intervention. Therefore, this study intends to use the advantages of machine learning to extract and train the useful features for reconstruction, improving the problems from occlusion. Based on multi-view stereo images and the machine learning model, this study aims to reconstruct the object or even the scene directly. Make the data processing operations simplified and the entire process more efficient or fully automated.

1. INTRODUCTION

In 3D GIS (Geographic Information System), scene analysis and targets distribution are important topics. 3D scene reconstruction is a hot research issue at this stage. Through object and scene reconstruction, the structure or target is visually described and positioned to enhance human's understanding of the scene layout, helping people analyzing and studying the space. Therefore, the reconstruction of 3D scene has its potential and importance to be developed. So far, there are many sources to fulfill 3D scene reconstruction, including aerial imagery, satellite imagery, point cloud data acquired from Light Detection and Ranging (LiDAR), etc. Traditionally, satellite imagery and aerial photography play major data producers. By using stereo pairs with known Azimuth, users artificially measure points and obtain conjugate image points, then proceed the aerial triangulation and produce the Digital Terrain Model (DTM) to present three-dimensional landforms. With the progress of the hardware equipment and software, some manual work, like image correlation or image matching, can be done by computer, which dramatically decrease the time cost and manpower. Make the photogrammetry gradually become automated proceeding techniques.

Multi-view stereo images are usually applied for 3D scene reconstruction. The method acquiring photographs from multiple views mimics the human visual system and fulfills the parallax concept. We can get the relationship between the corresponding points through the images. It is helpful for recover the geometry of the object or the scene. The low cost and accessibility of the equipment become the incentives to users. The way to reconstruct the 3D model do not contact or destroy the scene, keeping the scene from damaged or losing its original shape. It seems that this technique contains lots of advantages. However, there still exist some tasks. While doing massive image processing, users often meet some computation problem, e.g. high computation which cannot be handled, time and manpower-consuming work. Users also have to evaluate the mission success rate, completeness and precision of the model manually. Therefore, the novel technique named machine learning was applied to solve these significant problems.

Over these years, artificial intelligence (AI) has grown and engaged a few amount of researchers. Machine learning is seen as a subset of AI. The algorithms build a mathematical and statistical model based on several sample data, known as "training data", focusing on making predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms depend on the automatic computation, used in a wide variety of applications, such as text categorization (Sebastiani, 2002), sequence prediction (Sutskever *et al.*, 2014), feature detection (Rosten & Drummond, 2006), etc. This kind of algorithms save lots of manpower and the cost of time, promoting the efficiency.

Nowadays, 3D modeling is widely used in various fields, such as Augmented and virtual reality (AR and VR) (Wojciechowski *et al.*, 2004), disaster site reappearance, robot navigation (Ann *et al.*, 2016). With the usage rate raised, peoples seek the work efficiency and precision eagerly. Many imaged-based researches have focused on small-scale objects and scenes recently, showing accomplishments in two-dimensional image-based reconstruction. Most methods use the multiple view images of the target, overcome the problem of occlusion; nevertheless, the objects or scene with complex geometric relationship between stereo images still increase the difficulty of calculation. Also, the similar texture areas are not easily to be matched and reconstructed. To make the processing of 3D reconstruction more efficient and enhance the completeness of the 3D model, we aim to apply machine learning algorithm to the issue.

In this paper, we propose a system for 3D model creation that requires multi-view stereo images to reconstruct the objects and scene, using MVSNet (Yao *et al.*, 2018) as architecture demonstration. The technique is based on machine learning, making the whole workflow more efficient and automatic. In the first part of this architecture, we extract deep visual image features in different levels through each layers. Next step, based on the concept of the differentiable homography, construct the cost volume from the feature maps extracted from the previous step and known camera orientation. Later, generate the depth map. The initial estimated map helps compute the probability contribution along the depth direction and make the depth map refinement completed. After the refinement is done, calculate the loss for both the initial depth map and the refined depth map from the ground truth depth map. In the end, evaluate the precision and completeness of the depth estimation, using the depth map presenting the targets. The way improves the performance evaluation of the processing and result of the architecture.

The training data we used as the benchmark of the model is from the large-scale DTU dataset (Aanæs *et al.*, 2016). The dataset provides different scenes categories and the corresponding camera position and orientation. Also, the dataset is accompanied by exact structured light scans, helping user dominate the environment. The test data are parts of the DTU dataset and some targets from more complex Tanks and Temples dataset (Knapitsch *et al.*, 2017). With these open dataset, we determine to demonstrate the performance of this architecture and evaluate the results generated from the architecture.

2. RELATED WORKS

Recovering 3D geometry from photographs is a typical method and task that has occupied researchers for a long time in Geomatics and Computer Vision. The difference between 2D images and 3D objects is the depth information. Most image-based 3D reconstruction algorithms are aimed to estimate the most similar shape under the assumptions of known materials, viewpoints, and lighting conditions from a series of graphs of an object or a scene (Furukawa *et al.*, 2015). With the known information just mentioned, the geometry of the object or the scene can be described. The amount of viewpoints, the presentation of the target and the computation algorithms will affect the reconstruction performance. The following are the works related to these issues.

2.1 Multi-View Stereo (MVS)

Most work on 3D reconstruction has focused on monocular images (Liu *et al.*, 2016), binocular vision (Scharstein & Szeliski, 2002) and other algorithms that require multiple image, such as structure from motion (Forsyth & Ponce, 2002). Monocular image is captured from single view. In order to get the depth from monocular image or recover the exposure station, it requires to observe color/haze, varied scale, the texture variations and gradients on the photographs (Saxena *et al.*, 2007). The more accurate method is finding the point of view to depict the vanishing lines or position of the ground level (Torralba *et al.*, 2002). The spatial arrangement of the main structures can be roughly described and directly measured. Having said that, there still exist the limitations for modeling the entire shape of the targets, e.g. the occlusion, the unseen side from the single viewpoint. With single camera position and orientation, the photogrammetric network is weak and unstable. Due to these default, the reconstruction method can deal with few specific situations.

Compared to the monocular images, multiple view images have its dominate position on combining the integrated shape from the target. As shown in Fig 1, there exist some cues promoting spatial geometry extraction from a set of images, inclusive of texture, color, shading, contours, depth, and stereo correspondence. Hartley (Hartley & Zisserman, 2003) shows that it is possible to directly and uniformly derive multiple view relations from the intersection properties, especially lights, of back-projected lines and points. That gives assistance to researchers in finding out camera information and rectify the images. Furthermore, the geometry reconstruction can be accelerated to determine the target's 3D location. Make stereo correspondence widely and successfully utilized on the applications.

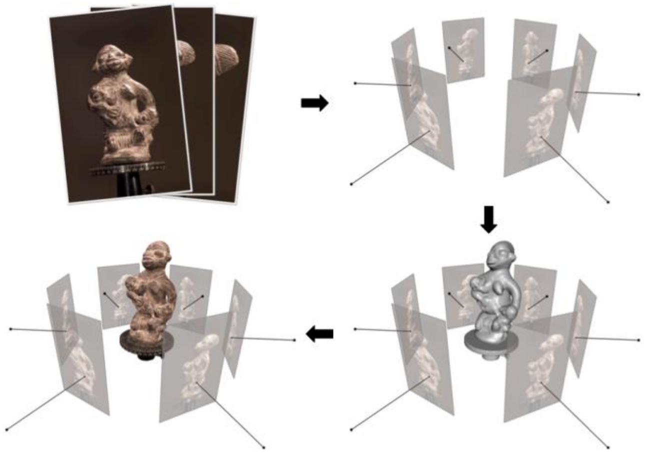

Figure 1 3D reconstruction from multiple images. First step, acquire a series of photographs of an object. Second step, build the corresponding relationship. Then, build the 3D object from extracting the feature in the photographs. Put on the texture on the model. (Furukawa *et al.*, 2015)

Depend on the development in decades, Browns (Brown *et al.*, 2003) organized some literature and made wide reviews on the stereopsis and 3D reconstruction. They put some focuses on image corresponding algorithms, methods on occlusion and real-time applications. Later, Seitz (Seitz *et al.*, 2006) created categories for the multi-view stereo image-based reconstruction algorithms, also mentioned some open 3D model datasets with ground true data. The evaluation methodology to the ground truth model also denoted is to compute the quality indicators, known as accuracy and completeness. Some following papers provide the improved method for calibrated multi-view stereopsis, diminishing the effect of outliers and obstacles in the scene from the photographs (Furukawa & Ponce, 2010). These researches achieve advancement of the reconstruction techniques.

2.2 3D Scene Representation

There are numerous ways to represent the geometry of an object or scene. Voxels, polygon meshes, or depth maps are adopted by the vast majority of multi-view image-based algorithms (Seitz *et al.*, 2006). Pixel is used to record the two dimensional images, while voxel, also called volume pixel, is the used for volumetric representation of the three dimensional objects. The methods divide the 3D target into regular grids or cubes and estimate how much each voxel is adhered to the surface. The smaller the cubes are used, the more carefully the surface is presented, which means there will be the high memory consumption. Plus, while dividing the 3D space, the space discretization error follows. Polygon meshes represent the surface as a set of connected planar polygons, efficient to store and render. It is kind of popular output format for multi-view algorithms. Each presentation, however, has its own defects and flaws making them incompatible with various scenarios or applications. Sometimes the mesh models need to be repaired (Attene *et al.*, 2013). Compared to the representation above, this multi-view depth map is the most flexible 2D representation, convenient especially for smaller datasets. Due to the parallax, we can compute the depth value and generate the depth disparity. As seen in Figure 2, it is plain to display the positional relationship between the objects in the scene through the depth map, getting over the global structure of the original image. The quality of multi-view depth map is more robust and accurate than the one from monocular or binocular images. Focusing on only one reference and a few source images each time makes the multiple image processing divided into relatively small cases of per-view depth map estimation (Yao *et al.*, 2018).

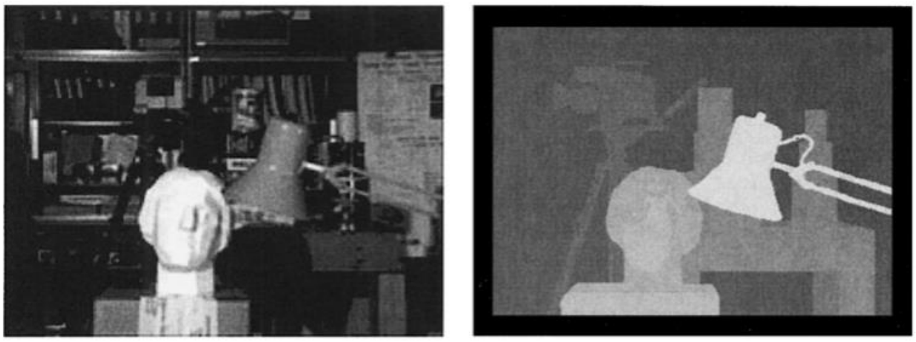

Figure 2 The indoor scene (left) and the corresponding depth map (right) (Scharstein & Szeliski, 2002)

2.3 Machine Learning Algorithm

Machine Learning is the subset of artificial intelligence (AI), born to approximate human thinking and operated for the missions that human expertise cannot accomplish. Counting on the computer programming, it learns from the sample data and experience to optimize a performance and then deal with given problems (Alpaydin, 2009). Among the machine learning models, the artificial neural network (NN) is one of the well-known algorithms. It mimics the operations of biological nervous systems, where simple neurons connect to each other to assemble as multiple layers. In the structure, there is a set of weights or parameters automatically changing through the learning process. After the input data all tackled and statistically calculated by the network, the optimized result is generated as the output data. Deep learning is a new branch of machine learning. The machine processes a large amount of disordered data through multiple processing layers, learning to complete specific tasks and automatically extracting features to represent data characteristics. The processing replaces the runtime consumption from the additional feature extraction in the traditional methods. Not only is the image feature extraction one of the deep learning applications, the algorithm can be applied to dimensional adjustment, non-linear regression, speech recognition (Mikolov *et al.*, 2010), and so on. Make the computer have the intelligence to give birth to sentences, images (Gregor *et al.*, 2015), or even 3D models.

In decades, some deep learning models came out, e.g. convolution neural network (CNN) (Krizhevsky *et al.*, 2012), recurrent neural network (RNN) (Mikolov *et al.*, 2010), Multi-view CNNs (Su *et al.*, 2015), 3D-CNNs (Ji *et al.*, 2012), SurfaceNet (Ji *et al.*, 2017). As seen in Figure 3, most deep learning models focus on Euclidean data, including descriptors, projection, RGB-D data, volumetric data, and multi-view images. According the different data source, users choose the suitable corresponding model. Most of them rely on supervised learning, labeling and defining the feature during the training procedure. On image recognition and classification, CNNs has its outstanding performance. Therefore, lots of image-based models regard CNNs as a clear demonstration.

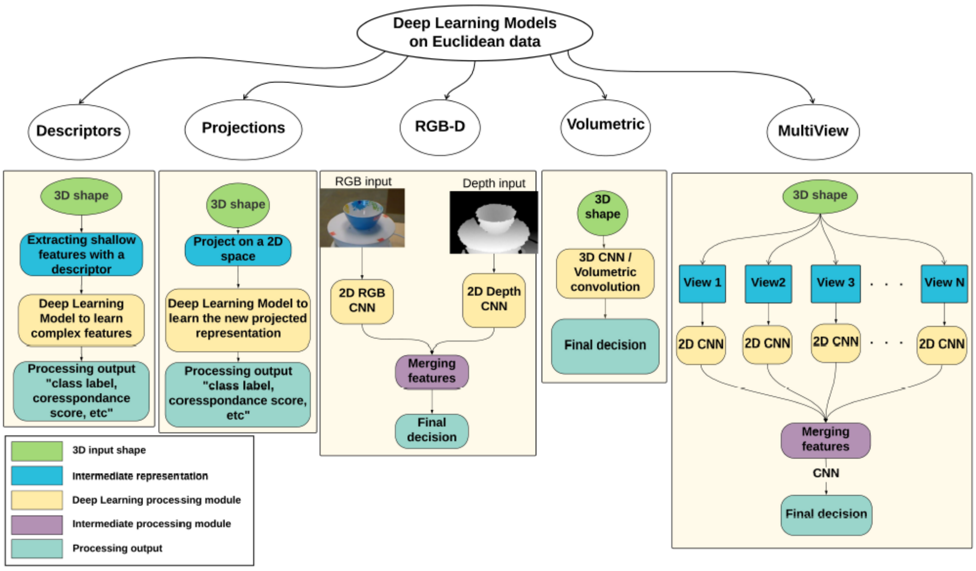
Deep Learning Models on Euclidean data
Descriptors
Projections
RGB-D
Volumetric
MultiView
3D shape
Extracting shallow features with a descriptor
Deep Learning Model to learn complex features
Processing output "class label, coresspondance score, etc"
3D shape
Project on a 2D space
Deep Learning Model to learn the new projected representation
Processing output "class label, coresspondance score, etc"
RGB input
Depth input
2D RGB CNN
2D Depth CNN
Merging features
Final decision
3D shape
3D CNN / Volumetric convolution
Final decision
3D shape
View 1
View2
View 3
View N
2D CNN
2D CNN
2D CNN
2D CNN
Merging features
CNN
Final decision
3D input shape
Intermediate representation
Deep Learning processing module
Intermediate processing module
Processing output

Figure 3 Deep learning models on Euclidean data (Ahmed *et al.*, 2018)

For the architectures based on the multiple view image processing, classical methods use the basic principle of dense matching in images and triangulation to achieve the 3D reconstruction. Multi-view CNNs is the model directly designed for multiple view photographs, first developed on the 3D shape recognition (Su *et al.*, 2015). First, import a collection of images of the 3D polygon object from various views. The algorithm learns and organizes the characteristics and establish the descriptors from the 2D feature. After view-pooling and second CNNs processing, predict the output class in the end. Revised from Multi-view CNNs, some models add more varied types of the photographs, inclusive of resolution, scale, and azimuth, strengthening the recognition ability. Having these various images makes the retrieved geometry more robust. In addition to image recognition, Learnt Stereo Machines (LSM) (Kar *et al.*, 2017) dedicates to using fewer images to reconstruct the target, saving the computation and storage space. Not like some studies predicting object geometry given only images (Choy *et al.*, 2016), this system is also afforded the prior known camera poses, enable to exploit strong geometric cues. LSM showcases the great performance of the object cases, yet leaving the scenes cases for future work. By contrast, the MVSNet faces the problems of reconstruction scale and focuses on producing the depth map for each reference at one time. Moreover, the depth maps are refined by computing the loss from the ground truth depth map, convincing that the depth maps have their credibility and are closed to the reality.

3. WORKFLOW OF THE ARCHITECTURE

In this section, we talk about the materials and the network pipeline. Our framework can be divided into four main parts, feature extraction, homography transformation, cost volume computation and the depth map production. The following are the brief description of each part.

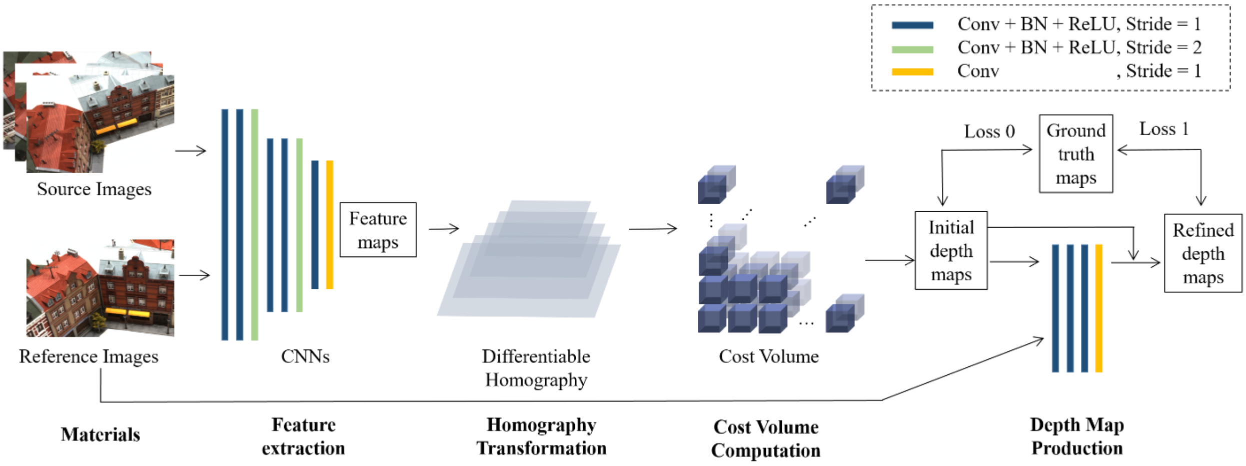

Figure 4 The structure of the architecture

3.1 Materials

We will use the large-scale DTU dataset (Aanæ s *et al.*, 2016) as the training data to set the benchmark in the model. In the freely available dataset, it provides lots of objects and scenes with changing structured lighting conditions, propelling the advance of multi-view stereo methodology and implements. Users can dominate which condition to apply. Given the so many different views of the object and scene categories into the training or testing procedure from the dataset, the model is more flexible to adjust and control the number of viewpoints. The most convenient thing is that the DTU dataset offers the ground truth data of each categories. There is no need for users to do additional computation and find the data in reality. To decrease the runtime, we select some scenes with proper lighting condition to train and test the network. For the testing part, we add the Tanks and Temples dataset (Knapitsch *et al.*, 2017) which supplies the larger scale targets with complete texture and shape. In the future, we would like to add some photographs acquired from the unmanned aerial vehicles (UAVs) as test data, evaluate the possibility of processing.

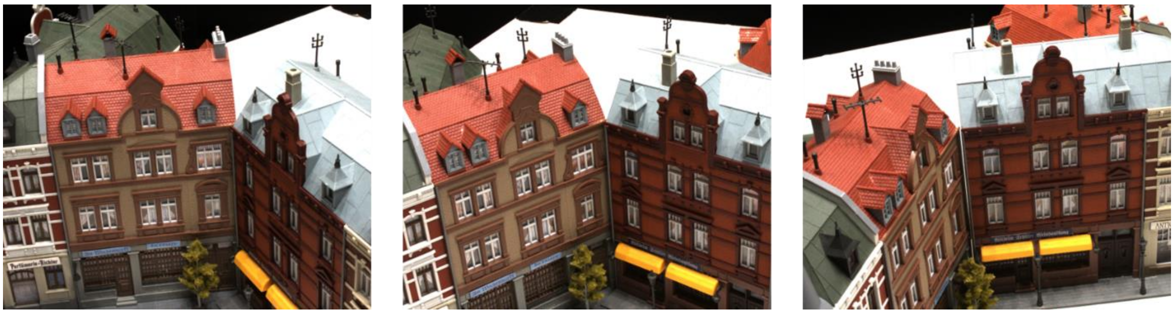

Figure 5 The multi-view images under same lighting condition from DTU dataset (Aanæ s *et al.*, 2016)

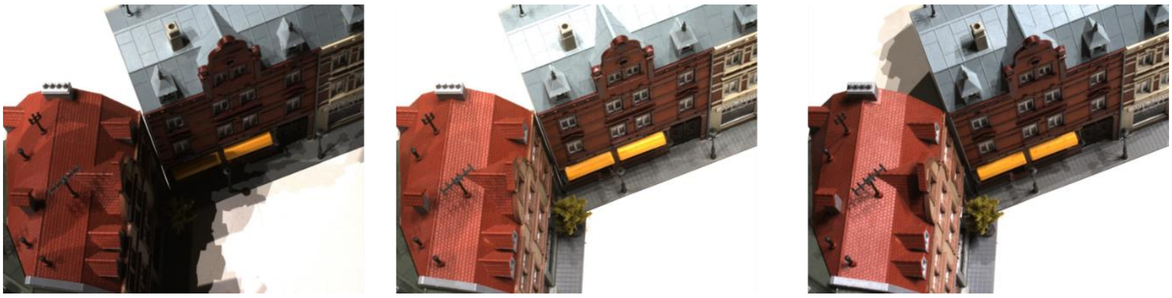

Figure 6 The same-view images under different lighting conditions from DTU dataset (Aanæ s *et al.*, 2016)

3.2 Feature Extraction

In this parts shows the convolution neural network with eight-layers. Basically, a convolution neural network is assembled with an input layer, many hidden layers and the output layer. The input data contain the source data and the reference data. The inside hidden layer can be divided into feature detection layers and fully connected layer. The feature detection layers mainly consist of the convolutional layers and pooling layers. The key component of convolution is the filter, also named feature detector or kernel, which is a two dimensional matrix moving on the image to compute with the pixel value. Each layer has its kernel to do convolution work and then extract the feature. The stride of the third and sixth layer are two, while the others are one, making the scale of the feature map different. The different scale record different level of feature, combined as the feature tower. After convolution computation, there might exist some negative value not allowed. To deal with this problem, we introduce the activation function named rectified linear unit (ReLU) to some layers, rejecting the negative values and only keeping the positive values. This is kind of popular function applied to the neural network, stimulating the neuron action to set up a threshold filtering the signal. Plus, we insert the batch-normalization layer to each layer except the eighth layer for keeping the images in the similar value interval.

3.3 Homography Transformation

Homography matrix is related to the fundamental matrix or essential matrix, using for computing the corresponding relationship between the pixels on the epipolar line of the stereo images. That means homography is an alternative definition of projective transformation, generally a three by three matrix. With given known camera information, projective transformation can project every feature map into a projectively parallel figure, leaving all projective properties invariant (Hartley & Zisserman, 2003). In this architecture, the differentiable homography is operated for the connection between 2D feature extraction and the 3D regularization networks, also enable to proceed the training of depth map inference.

3.4 Cost Volume Computation

After obtaining the epipolar images of the feature maps, we can move on calculating and regularizing the cost volume. Adopting the variance-based cost metric operation of Yao (Yao et al., 2018), each feature map information, such as width (*W*), height (*H*), depth sample number (*D*) and the channel number (*F*), are utilized to define the feature volume (*V*). The *M* function is defined to find the variance of the feature volume (*V*), and the output value is the cost volume (*C*), shown in Formula 2. Then, the probability volume (*P*) can be calculate by the cost volume. Also, the multi-scale 3D CNNs is introduced for doing cost volume regularization, and the softmax operation finally is applied for probability regularization. The softmax function comes from the logistic regression, often used in the last layer of the network. It outputs the value in the range between one and zero, and the outputs usually are seen as the probabilities revelant to each category. These steps remove the noise and the impact from the occlusions and create the initial depth map, making the depth map more smooth and continuous.

$$V = \frac{W}{4} \cdot \frac{H}{4} \cdot D \cdot F \tag{1}$$

$$C = M(V_1, \cdots, V_N) = \frac{\sum_{i=1}^{N} (V_i - \overline{V_i})^2}{N} \tag{2}$$

3.5 Depth Map Production

Traditionally, the depth map is retrieved by the pixel-wise winner-take-all (Yang *et al.*, 2007), well-known approach on the weighted cost volume and a sub-pixel refinement. To achieve the sub-pixel depth map, we adopt the formula 3 (Yao *et al.*, 2018), computing the expectation along the depth direction. Among the hypotheses, the P(*d*) is the probability estimation for all pixels at depth *d*. The computation is along the depth direction and range from d_{max} to d_{min}. After obtaining the depth value and the depth probability, the depth map (*D*) is found through the formula, referred to as the soft argmin operation. Then, the continuous depth distribution is generated to form an initial depth map of the sub-pixel precision. As soon as the initial dep map come out, the reference image and the depth map are applied as the input data during the depth map refinement. The depth refinement consists two parts. One is the depth residual learning network with reference images, and the other is the loss computation from the ground truth map. In the first part, calculate the residuals between the reference image and the initial image in the same size. Due to that the reference image contains natural boundary information, it can be used as a demonstration for refining the depth map. The depth residual learning network has four two dimensional convolution layers to record the residual and output the refined depth map. The residual generated from the network contains positive and negative values. The negative value is not eliminated by the activation function, meaning that all the residuals are held. After merged with the initial depth map, the calculated residual depth map is used as the refined depth map. Then, we import the ground truth map to compute the loss. Both initial depth map and the revised depth map are considered to calculate. As the ground truth depth maps are not fully complete with true value in every pixel, we only think over the valid one actually recording the ground truth value. As seen in the formula 4, p_{valid} is the pixels recorded the valid ground truth value. Loss0 and loss1 are the positive difference respectively generated from the ground truth depth map. The final loss is the combination of the loss0 and loss1, and the parameter is set to 1.0 while working.

$$D = \sum_{d=d_{\max}}^{d_{\max}} d \times P(d) \tag{3}$$

$$Loss = \sum_{p \in p_{valid}} Loss0 + \lambda \times Loss1 \tag{4}$$

4. EXPERIMENTS AND DISCUSSION

Our ultimate goal is to build the 3D model of the large scene, make the implements of our method more flexible and widely used. We first choose many sets of images of the objects from some categories in the DTU dataset as training data, especially the similar targets we would like to reconstruct, e.g. the houses, the streets. After training data, the checkpoint file will be generated and seem as the pre-trained model. With the pre-trained model file, we can skip the training process and save many time and computation. Then we choose one of the objects as the test data, which is the scale-down version of the scene. The object has detailed texture with various color and non-paralleled planes, just like the houses on the street but in the much smaller scale. To apply this architecture to the usual images or the images captured from the unmanned aerial vehicles, we need to test and evaluate the effect or results of the similar targets. There are some settings that need to be clarified and decided before training and testing, inclusive of the image size, the number of the viewpoints, batch size, sample scale, interval size, and so on. Each variable will affect the processing and the results, e.g. the maximum image size when training, down-sample scale for calculating cost volume, the runtime of the training, so it is important to check in advance and adjust after every running step.

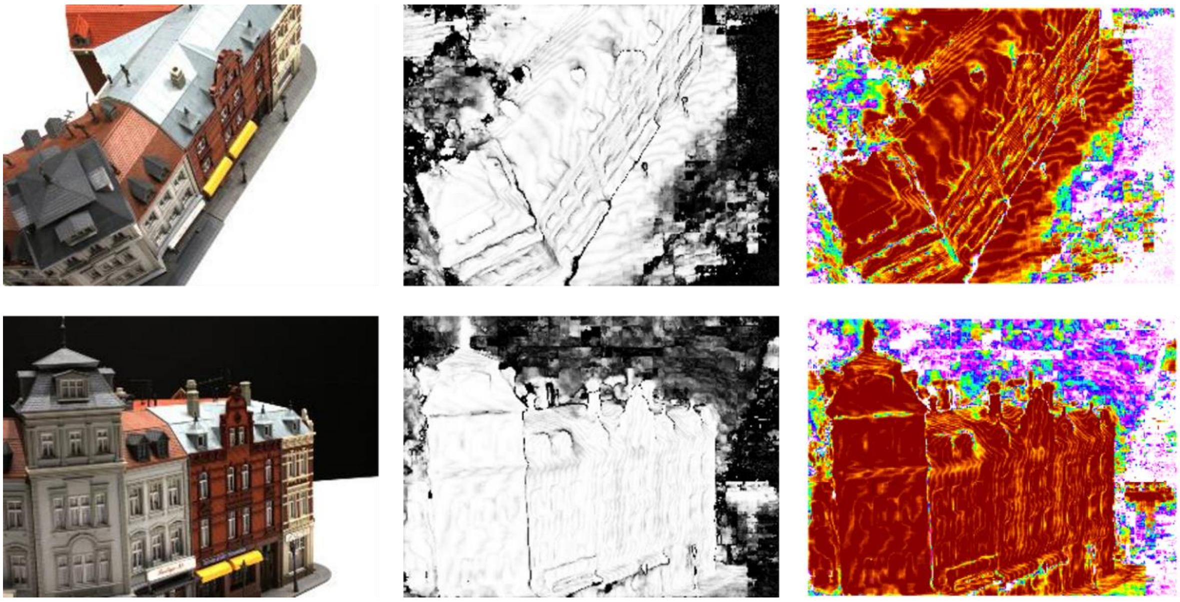

Figure 7 The right row is the original images of the test target from different views, while the middle row presents the probability maps in black and white. The colorful probability maps shown in the right row convert from the probability map in black and white.

Our experiment is now conducted to the generation of probability map. The probability map from test data are shown in figure 7. We decide to use the images from at least three view-points, based on the definition of the multiple view stereo, and we choose three to five viewpoints for the test, not to spending too much time on images from many viewpoints. From figure 7, the middle images are initial probability depth maps presented in the black and white color, which is not good and convenient enough for users to find the probability distribution. To know the spatial information from the probability map, we convert the probability map into rainbow color, making it more visualized in direct way to observe. Observing the colorful probability map, we can find that the architecture obviously marks the area of the target in the brown or dark orange area, but the background is presented in a blurred way. The probability maps reflect the depth estimation quality, showing that the depth value can be easily determined due to the highly concentrated probability distribution. To make the probability map more persuasive, we generate the probability distribution by quantizing the value and present the probability distribution according to the index of the depth value, seen in figure 8. From the target area in the image, it is easy to find the dominate value from the plot. In contract, some noise result in the probability dispersion in the background area, making the true value not obviously defined in a statistical way. There is no dominate probability value in the plot, it needs to use the softmin operation mention before to determine the corresponding index of the depth value.

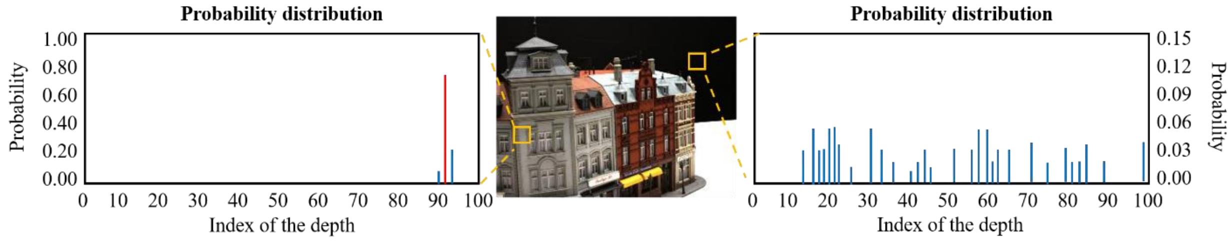

Figure 8 The left-side plot shows the probability distribution of the target itself, while the right-side plot presents the probability scatter of background.

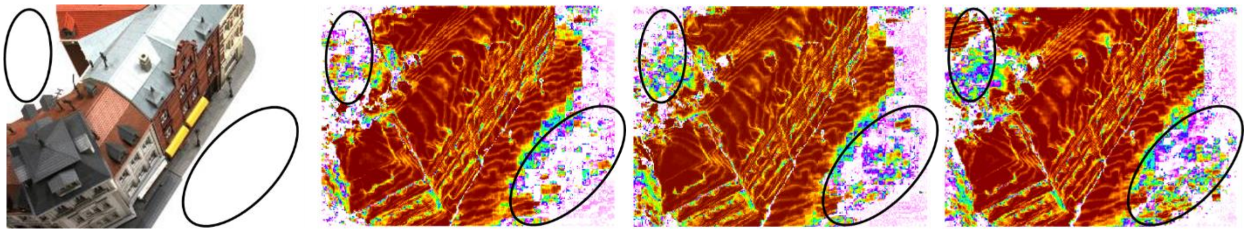

Figure 9 The images from left to right are the original image and the probability maps separately produced from three, four and five viewpoints.

Keeping eye on the images from different input view point, there are some similar feature on the probability maps. The main body of the target on the images look similar to other probability maps, all marked in the brown color. Compared to the target, some noises with unstable probability value exist on the background. In the original images from DTU dataset, the spatial information of the background is empty or ignored. With the more view point data as input data, the probability map has more noises on the background area marked in the figure 9, which means that result unpredictably changes according to the change of the input number of the view point every time. On the other hand, the key point need noticing and mentioning is the probability distribution of the edge of the target. In decades, experts in photogrammetry and computer vision devote much time and energy on the edge detection, matching and model reconstruction, but the results were not usually good enough to present. The discontinuous surface or texture can be seen as parts of the edges. In our architecture, it is not hard to find the edge through the probability map. The probability distribution is not shown with the obvious probability corresponding to the index of the depth value, and it color bar is not as stable as that in the area with similar texture. But we spend few time, for about one half minute, producing the probability map, saving more cost than the traditional methods. The depth estimation of the key area of the target are clearly predicted, which go far towards generating and refining the depth map afterwards.

5. CONCLUSIONS AND FUTURE WORKS

The deep learning has its dominate role on recording the feature of the images. In traditional approaches, many studies met the problem from the occlusion areas. The 3D model of the occlusion area is usually broken and incomplete with few point cloud. Utilizing multiple photographs, it is possible to observe the whole object and reconstruct the more complete model. Besides the problem of occlusion, it has difficulty to conduct the feature extraction and image matching from areas with similar texture in the classic methods. Because it is not easy to capture and recognize the feature, the methods lead the spatial information of that specific areas being ignored and wasted. To deal with the problems, we apply deep learning to the architecture. The convolution neural network has high record ability, without losing features from the imported images. The information of the areas with similar texture can be kept through the feature extraction, not like the procedure from the traditional methods. In our experience, we aim to use the open image dataset to training and testing the architecture, as the above mentioned. We start first on the small scale object, then about to extend to the scene reconstruction. Not only changing the number of viewpoint, we would like to expand the images to different scales, resolutions, the amount of prior data, the scene complexity, and so on. In addition to DTU dataset, we intend to apply the multiple view photographs acquired from the unmanned aerial vehicles as test data, which is much more unstable than DTU dataset or Tanks and Temples. The foreseen problem, on the other hands, are the running time and the GPU memory. It is necessary to make a comparison with the commercial photogrammetry software, e.g. Agisoft Metashape, which comes from Agisoft Photoscan and is made use of by some research groups worldwide, or the novel neural network, such as SurfaceNet, 3D CNNs. The GPU memory is highly related to the input image size, the depth sample number and the batch number. How to adjust the proper number to make the whole work more efficient in speed is the top priority. Therefore, we will keep test and revise the working of the proposed architecture on time, organizing and analyzing the experience results. Then, develop an efficient algorithm suitable for the multiple view image reconstruction.

6. REFERENCE

Aanæ s, H., Jensen, R. R., Vogiatzis, G., Tola, E., & Dahl, A. B. J. I. J. o. C. V., 2016. Large-Scale Data for Multiple-View Stereopsis. International Journal of Computer Vision, 120(2), pp. 153-168.

Ahmed, E., Saint, A., El Rahman Shabayek, A., Cherenkova, K., Das, R., Gusev, G., Ottersten, B, 2018. Deep learning advances on different 3D data representations: A survey. CoRR.

Alpaydin, E., 2009. Introduction to machine learning. MIT press.

Ann, N. Q., Achmad, M. S. H., Bayuaji, L., Daud, M. R., & Pebrianti, D., 2016. Study on 3D scene reconstruction in robot navigation using stereo vision. 2016 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS). pp. 72-77.

Attene, M., Campen, M., & Kobbelt, L. J. A. C. S., 2013. Polygon mesh repairing: An application perspective. 45(2), pp. 15.

Brown, M. Z., Burschka, D., & Hager, G. D., 2003. Advances in computational stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(8), 993-1008.

Choy, C. B., Xu, D., Gwak, J., Chen, K., & Savarese, S., 2016. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. European conference on computer vision. pp. 628-644.

Forsyth, D. A., & Ponce, J., 2002. Computer vision: a modern approach: Prentice Hall Professional Technical Reference. Prentice Hall Professional Technical Reference.

Furukawa, Y., Hernández, C. J. F., Graphics, T. i. C., & Vision., 2015. Multi-view stereo: A tutorial. 9(1-2), pp. 1-148.

Furukawa, Y., & Ponce, J., 2010. Accurate, Dense, and Robust Multiview Stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(8), pp. 1362-1376.

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. J. a. p. a., 2015. Draw: A recurrent neural network for image generation.

Hartley, R., & Zisserman, A., 2003. Multiple view geometry in computer vision. Cambridge university press.

Ji, M., Gall, J., Zheng, H., Liu, Y., & Fang, L., 2017. SurfaceNet: An End-to-end 3D Neural Network for Multiview Stereopsis. Proceedings of the IEEE International Conference on Computer Vision. pp. 2307-2315.

Ji, S., Xu, W., Yang, M., Yu, K. J. I. t. o. p. a., & intelligence, m., 2012. 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence. 35(1), pp. 221-231.

Kar, A., Häne, C., & Malik, J., 2017. Learning a multi-view stereo machine. Advances in neural information processing systems. pp. 365-376.

Knapitsch, A., Park, J., Zhou, Q.-Y., & Koltun, V. J. A. T. o. G., 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG). 36(4), pp. 78.

Krizhevsky, A., Sutskever, I., & Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. pp. 1097-1105.

Liu, F., Shen, C., Lin, G., & Reid, I., 2016. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(10), pp. 2024-2039.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S., 2010. Recurrent neural network based language model. Eleventh annual conference of the international speech communication association.

Rosten, E., & Drummond, T., 2006. Machine learning for high-speed corner detection. European conference on computer vision. Springer. pp. 430-443

Saxena, A., Schulte, J., & Ng, A. Y., 2007. Depth Estimation Using Monocular and Stereo Cues. IJCAI. pp. 2197-2203.

Scharstein, D., & Szeliski, R. J. I. j. o. c. v., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. 47(1-3), pp. 7-42.

Sebastiani, F. J. A. c. s., 2002. Machine learning in automated text categorization. ACM computing surveys (CSUR). 34(1), pp. 1-47.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., & Szeliski, R., 2006. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). 1. pp. 519-528.

Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3d shape recognition. Proceedings of the IEEE international conference on computer vision. pp. 945-953.

Sutskever, I., Vinyals, O., & Le, Q. V., 2014. Sequence to sequence learning with neural networks. Advances in neural information processing systems. pp. 3104-3112.

Torralba, A., Oliva, A. J. I. T. o. p. a., & intelligence, m., 2002. Depth estimation from image structure. 24(9), pp. 1226-1238.

Wojciechowski, R., Walczak, K., White, M., & Cellary, W., 2004. Building virtual and augmented reality museum exhibitions. Proceedings of the ninth international conference on 3D Web technology. pp. 135-144.

Yang, Q., Yang, R., Davis, J., & Nistér, D., 2007. Spatial-depth super resolution for range images. 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1-8.

Yao, Y., Luo, Z., Li, S., Fang, T., & Quan, L., 2018. Mvsnet: Depth inference for unstructured multi-view stereo. Proceedings of the European Conference on Computer Vision (ECCV). pp. 767-783.