High-quality Voxel Reconstruction from Stereoscopic Images

Arturo Navarro
Department of Computer Science
Universidad Católica San Pablo, Peru

Manuel Loaiza
Department of Computer Science
Universidad Católica San Pablo, Peru

***Abstract*—Volumetric reconstruction from one or multiple RGB images has shown significant advances in recent years, but the approaches used so far do not take advantage of stereoscopic features such as distance blur, perspective disparity, textures, etc. that are useful to shape the object volumes. Our study is to evaluate a convolutional neural network architecture for reconstruction of 128^3 voxel models from 960 pairs of stereoscopic images. The preliminary results show an 80% of coincidence with the original models in 2 categories using the Intersection over Union metric. These results indicate that good reconstructions can be made from a small dataset. This will reduce the time and memory usage for this task.**

***Keywords*—Voxel reconstruction; stereoscopy; convolutional neural networks; disparity maps**

I. Introduction

Voxel Reconstruction is a wide field of research within computer vision. It is currently getting more attention thanks to the appearance of high-resolution RGB-D and stereo cameras. This field has benefitted from the development of machine learning neural networks, which have allowed volumetric reconstruction to obtain promising advances both for single-image and multiple-image cases. This reconstruction takes pictures from real or synthetic 3D objects and obtains a model made of voxels. Voxels are the 3D equivalent of the 2D pixels; this means that voxels are mainly indivisible blocks of position and color data, that can be put together and broken down. Such properties are important for applications like robotic vision, particle simulation, or volume comparison.

Computer vision itself offers multiple opportunities for further development since it is still a budding topic to deal with due to the gap between the information provided by the pixels and the interpretation that can be given to that information. Current volumetric reconstruction methods based on Convolutional Neural Networks (CNN) have achieved significant progress in the quality of the results, however reconstructions are made usually under good conditions that are rarely replicated in reality: multiple views, good lighting, and definition, or extensive datasets. They also don't make any use of depth information provided by stereoscopic vision. Stereoscopy is a vital function for human beings for depth perception. It gives good information about volume and depth mainly from depth cues like accommodation, focus, occlusion, linear and aerial perspective, relative size, density, and motion parallax [1]. Much of this information is present in only one of the ocular perspectives through the so-called monocular depth, but it is the complementation that allows a better perception, through the binocular depth.

Current approaches for 3D reconstruction tasks have been obtained mainly under optimal conditions and multiple viewpoints, which is often not possible in real situations. On the other hand, while single-image reconstruction methods have also made improvements, they rely on intensive pretraining and large datasets to estimate the hidden parts and fill in the volume.These datasets are usually formed by images with simple and unique objects; this is good for techniques testing but it has limited practical applications. Reconstructed voxel meshes also tend to be low-resolution (32^3 voxels), losing important details that could be significant to some applications.

This study aims to validate a deep learning model for stereo 3D reconstruction, to get a good resolution model of 128^3 voxels from 960 couples of stereoscopic images. For that purpose, The architecture proposed by [7] was implemented. This generated more detailed voxel models for the applications previously mentioned. The dataset used for this study will be kept small due to limited computational resources. We aimed to prove that a 3D reconstruction is feasible from a small dataset of detailed images and limited computational resources that could be use for practical applications, where optimal image capture is not ideal.

The rest of this document is as follows: chapter two will include the related work to this study. Methodology and dataset is presented in chapter three. The results will be shown in chapter four. Chapter five will review the discussion and the final chapter will include the conclusions.

II. Stereoscopy

It is a technique of depth artificial representation from two stereographic images, where each of them represents the vision from one of the human eyes, thus imitating the real vision. Depth is reconstructed in the brain from the 2D images captured by the retina, and from certain signals or cues present in those images [2]. Much of this information is present in only one of the ocular perspectives through the so-called monocular depth, but it is the vision complementation that allows a better perception, through the binocular depth.

Main cues:

- Occlusion: Visual obstruction of a near object on a more distant one.
- Linear Perspective: Phenomenon by which parallel lines seem to get closer the further they are apart.

- Aerial perspective: Perception of distance through the absorption of light. Closer objects appear more colorful and brighter than distant objects in general.
- Relative size: Related to perspective. Nearby objects appear larger.
- Density: The denser an object's texture appears, the farther away it is assumed to be. It is also known as texture gradient.
- Adequacy: It is the adjustment of the muscle and the ocular lens to focus an image. The further away the object is, the more relaxed the muscle and the more circular the lens.
- Focus: It is the rotation of the eyes, to focus together on an object at a certain distance. The eyes converge on near objects and diverge on distant ones.
- Motion Parallax: When the viewer (human, camera, etc.) is in motion, nearby objects appear to move faster than distant objects.

Fig. 1 shows the main binocular cues for depth perception.

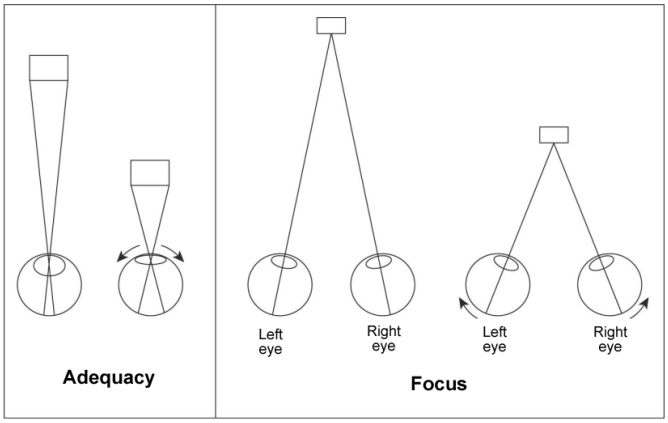

Fig. 1. Binocular Cues. [1]

III. Convolutional Neural Networks

They are supervised learning neural networks, based on multilayer perceptrons, where the hidden layers are specialized to detect certain shapes, starting from simple lines and curves, up to more specialized layers capable for complex figures detection. [3]. These networks are especially used for artificial vision since they are biologically inspired by the functioning of the human visual cortex.

The input for the convolutional network is a normalized set of image pixels, where each input neuron corresponds to a pixel, separated by channels (red, blue, green). A group of nearby pixels is then convolved with a filter matrix called kernel. These kernels are applied one by one to each neuron input, which are previously trained. The result of that convolution is a matrix that feeds the next network layer. For this first layer, the output of each neuron is calculated as:

$$Y_j = f\left(b_j + \sum_{i=1} K_{ij} * Y_i\right) \quad (1)$$

Where Y_j is the output of each neuron, Y_i are its inputs, K_{ij} is the kernel , b_j is the bias , and $f()$ is an activation function, commonly the Rectified Linear Function or ReLu; which is defined as $f(x) = max(0, x)$.

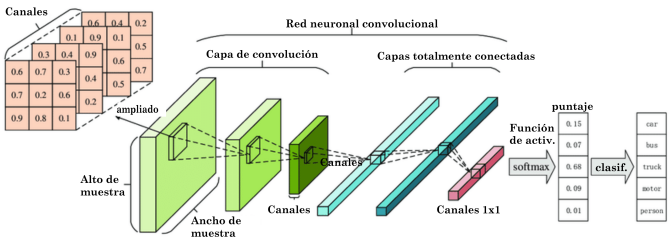

Fig. 2. Example of Convolutional Neural Network [4].

The next step is sampling, which is reducing the size of the output matrix of the previous layer, thus reducing the computational cost of the convolutions. The most common reduction method is Max Pooling which selects the max value from $m \times n$ group of neighbour cells. This also helps to reduce noise from the extracted images. On the reduced outputs a new set of kernels is applied to perform a finer detail extraction followed by another sampling process. This is repeated as necessary.

Finally, as shown in Fig. 2, the outputs of the last convolution-sampling process are sent to a fully connected supervised learning network that is responsible for classifying the details to a previously established category [5].

IV. Related Work

In [6], the authors made a classification for 3D reconstruction approaches, focusing on shape representations, the network architectures, and the training mechanisms, stating specific issues such as limited datasets, fine-scale reconstruction, and unseen objects. Recently, Xie et al. [7] proposed to reconstruct synthetic 3D models into a low-quality voxel mesh and a point cloud using a dataset of couples of RGB images and their correspondent disparity maps to feed a 3 CNN pipeline (Fig. 3), this work is the first to use stereo images for reconstruction based on deep learning. It obtained similar results to reconstruction from multiple-image-based reconstructions.

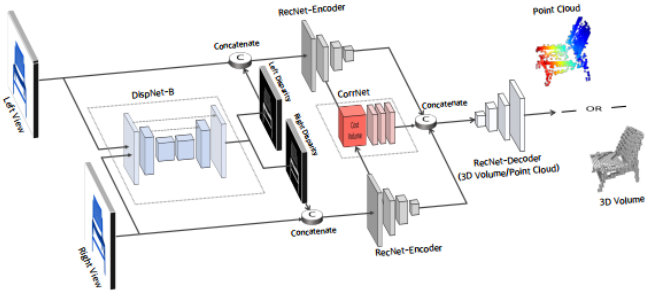

Fig. 3. CNN Architecture for 3d Mesh and Point Cloud Reconstruction from RGB and Disparity Sterteoscopy Images. [7]

Similarly, Xie worked also on a point view and voxel reconstruction approach for single and multiple views named

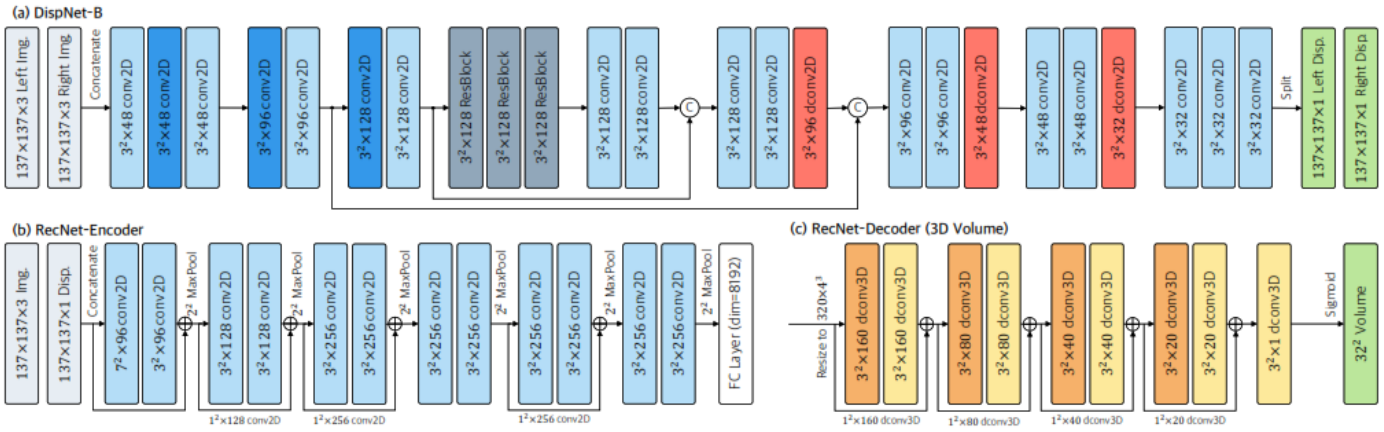

Fig. 4. DispNet and RecNet Architecture. [7]

Pix2Vox [8] [9]. For the voxel mesh upscaling task, Wang et al. [10] presented a long-term recurrent CNN and Generative Adversarial Network (GAN) combination to complete and improve resolution for voxel volumes. These works generate low-resolution 32^3 meshes and get small and simple images as inputs, our work starts from complex and detailed images from a small dataset to get similar results in completion.

A relatively recent study to improve recognition by CNNs is that of Qi et. al [11] which improves the results on both volumetric and multiview-based reconstructions by using two CNNs with specialized jobs. These architectures produce also low-quality voxel meshes. Dryanovski et al [12] using mobile devices and the *Google Tango* platform to achieve a low computational cost and low memory consumption reconstruction of large scenes by using a hash table to store the volumetric information. The method also includes a fusion process to improve the quality and completeness of reconstructions. Kar, Häne, and Malik [13], propose a stereo reconstruction system based on projective geometry to transfer from 2D to 3D elements from multiple images joined by a CNN to a volumetric mesh of 32^3 and a depth point map. Riegler et al [14] present a CNN fusion architecture for multiview reconstruction, based on truncated signal distance function (TSDF) and octree data structures for simple objects. The efficiency of representation is addressed by Liu et al. [15], where the authors propose a point-voxel convolution to reduce the memory cost of voxel models. Multiple-view approaches rely on large datsets, camera poses and costly 3D fusion processes. Moreover, getting many views for a scene is not always feasible. Stereo-based reconstruction uses only two images; therefore, partial 3D mesh fusion is less costly.

In [16], Firman et al. start from a trained set of decision trees in a random forest to complete a scene making predictions about the reconstructed geometry in *voxlets* or groups of voxels, for a low-resolution mesh, using only a single input depth image. Yang et al. [17], seek a higher resolution also starting from a single image, but applying Generative Adversarial Networks (GANs) to identify a specific set of objects (benches, chairs, armchairs, and tables) from a single image, but with meshes of 256^3 voxels. In contrast, Häne et al. used Convolutional Neural Networks (CNN), [18] to achieve a hierarchical prediction by subdividing an initial 16^3 mesh into a 256^3 voxel to refine the result. For that purpose, they use also a single image, but the framework is prepared to work with more inputs. More recently, Tatarchenko et al. [19] proposed two new methods in the recognition task that is based on the Intersection over Union (IoU) metric: Clustering and Recovery, although also pointing to low-resolution meshes. TSDF volumes and regression instead of depth maps are used by Choe [20] and Murez [21], making their reconstructions based on camera poses. Jadhav [22], worked on a homography multiview image correspondence to complete voxel meshes. Addressing the limitation of single image perspective, Watson et al. [23] introduces a process for obtaining stereo images from a single RGB input, adding depth on key information for reconstruction. Volumen completion from single image input is made by Varley et al. [24], which work consists of a fast voxel mesh completion method applied to robot grasping. One-image-based reconstructions require semantic labeled datasets and they are restricted to specific domains, since they complete occluded parts from similar preproccesed scenes. Stereo reconstruction obtains their depth information from the very input, and our work relies only in the ground truth model for refinement. Other approaches like TSDF volumes and homography need also camera poses and multiple images which are not always available.

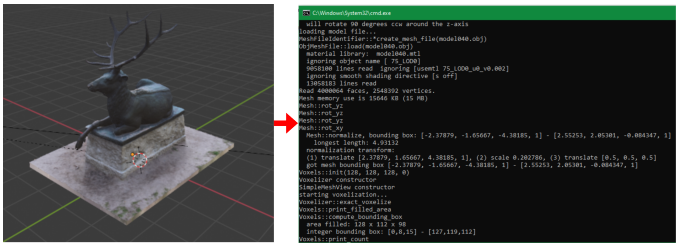

Fig. 5. Voxelization Process on Binvox.

Stereo image inputs carry enough depth information to obtain a suitable 3D voxel mesh model that can be used in practical applications. Our work starts from a small dataset of detailed images, so the computational cost is small both in image processing and 3d mesh fusion. To overcome domain-restricted limitations, we tested two different categories of objects: complex low-res scenes and single high-detailed objects.

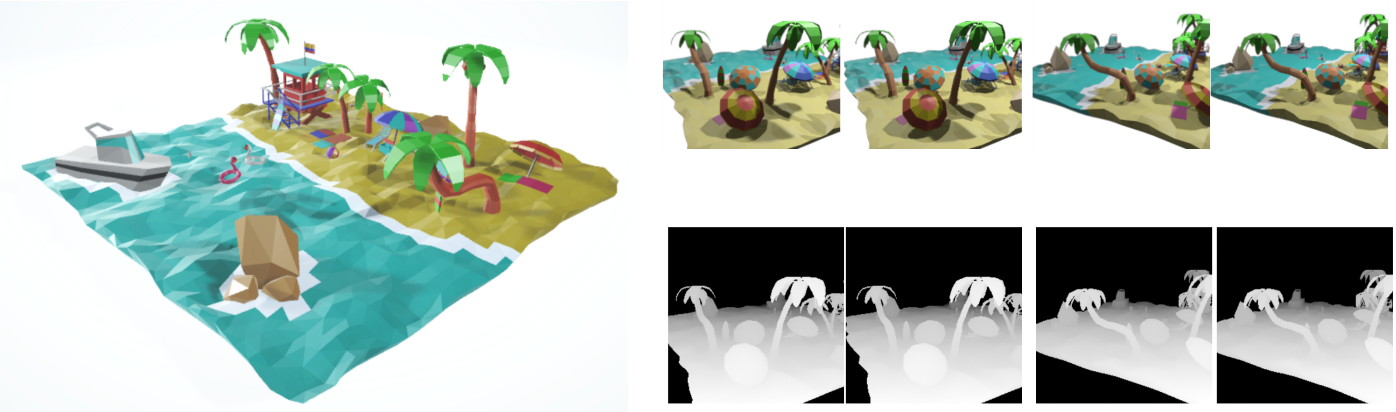

Fig. 6. Left: Original Model. Right up: Pairs of Left-Right RGB Images. Right Down: Pairs of Left-Right Disparity Maps.

V. Methodology

We use the DispNet, and RecNet architecture from [7] for high-quality volumetric reconstruction from a dataset of pairs of stereoscopic images with details that recreate real conditions, such as a Gaussian blur or low lighting. This dataset was generated from 3D models that could be found on https://sketchfab.com. The Blender tool was used to create the RGB and depth images according to the procedure described in [9]. The dataset is extended by adding the same images with a depth of field of 3m.

The tested architecture includes interconnected convolutional and residual neural networks of the encoder-decoder type, as can be seen in Fig. 4. DispNet includes 5 convolution and 4 deconvolution layers to deliver the two disparity images. RecNet includes 6 convolution layers, one fully connected layer, and 7 deconvolution layers to finally build the volumetric mesh. Two modifications were made to the original RecNet decoder architecture to produce 128^3 voxel grids. Several residual blocks are included in both networks. CorrNet .. is a three-dimensional convolutional network with nine convolutional layers and a final fully connected layer; Although it was implemented, due to limitations of the development platform it could not be tested with more than 500 images.

A comparison of the refinement of the volumetric mesh was obtained by training the current convolutional network to obtain higher resolution meshes; specifically 128^3 voxels. This will be useful for models belonging to taxonomies with more complex shapes. For the validation of the results, a voxelization process will be used from the tested models to obtain ground truth meshes with Blender and Binvox (Fig. 5). The results were measured using the IoU metric against the previously obtained *ground truth* reconstructions.

A. Implementation Details

The architecture was implemented with Kaggle Notebook, using a NVIDIA Tesla P100 GPU with 13 GB of GDDR5 memory, an Intel(R) Xeon(R) 2.30GHz CPU, 16 GB of RAM. and 359 GB of disk space. The implementation has been carried out with the Python programming language and the Keras *deep learning* library, which is already equipped with all the types of convolutional layers required to build the proposed models. Due to limitations in space and use time, the number of epochs for both networks are limited to 100, which we found was enough to achieve good results.

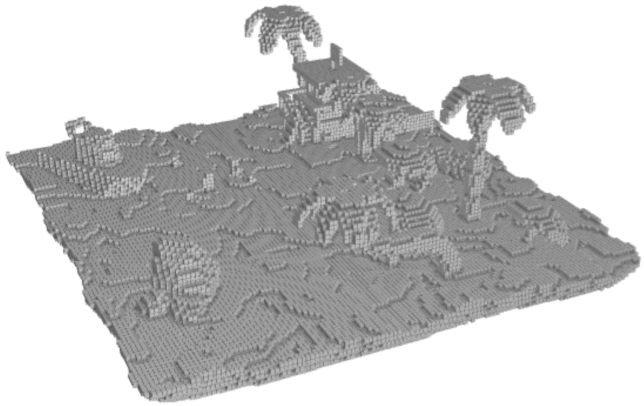

Fig. 7. Binvox Mesh Sample.

B. Data set

Dataset has been built with pairs of stereoscopic images in RGB and their correspondent disparity maps captured from 40 3d objects selected from 2 categories of free models obtained from sketchfab.com: complex scenes with low polygon count and high poly count statues. A production pipeline for these images has been created in the **Blender** (https://www.blender.org/) tool, where a stereoscopic camera with azimuth angle $\theta_{az} \in [0.360)$ and elevation angle $\theta_{el} = 30$, with a 35mm focal length, 32mm sensor size, and a 130mm [7] stereo baseline. The camera is positioned at 12 different angles with an angular increment of 30º. The RGB images obtained are 256×256 pixels. In the case of disparity images, these are obtained by applying a depth filter which is then color inverted to generate the disparity details. These images also have 256×256 pixels. Fig. 6 shows some images in RGB and disparity. Each model has been voxelized with the **binvox** tool (https://www.patrickmin.com/binvox/), since this format

is more compatible with Keras (Fig. 7). A three-dimensional mesh of occupancy (binary values) is obtained, loaded, and stored for use. The size of the dataset is then increased by applying a 6-pixel neighborhood Gaussian blur effect to the RGB images.

VI. Results

A. DispNet Results

The DispNet network received two RGB stereoscopic images concatenated as a 6-channel array for a set size of 960 pairs. As output, it generated two stereoscopic grayscale disparity images (1 channel), which are also concatenated in a 2-channel array with a set size of 960 pairs. The network is configured as shown in Fig. 4. The Dispnet class includes a minimum square error *loss* function, Adam optimizer, 100 training epochs, *batch* of size 6, and validation set equal to 30%. After training, the following metrics are obtained:

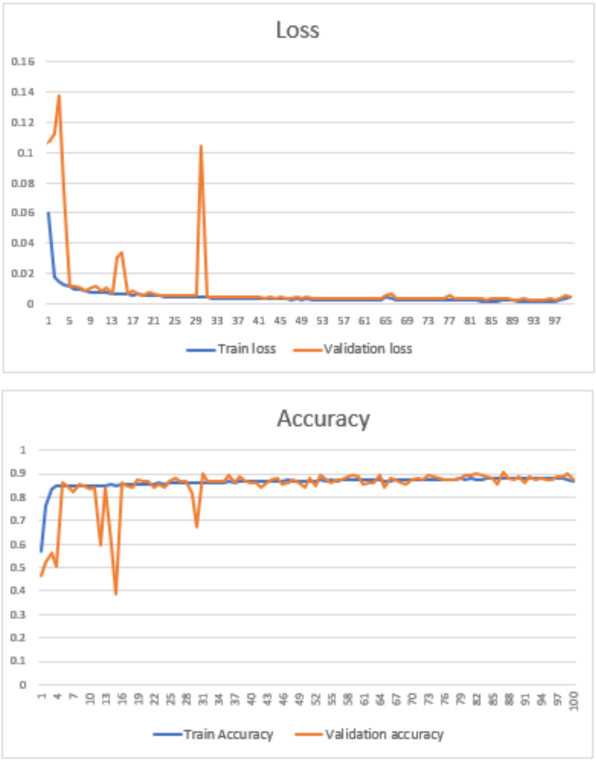

Fig. 8. Dispnet Loss and Accuracy during 100 Epochs of Training.

TABLE I. Training Results for DispNet After 100 Epochs

Metric	Train	Validation
loss	0.0016	0.0027
accuracy	0.8634	0.8431

Table I shows an error value below of 0.01% for both sets and an accuracy of about 86% for the training set and 84% for the validation set. The training progress can be seen in Fig. 8. These results are quite acceptable, even though it is a small dataset. Visually the disparity images produced by Dispnet are very similar to the original in shape and color. This can be seen in Fig. 9. These generated images are feasible to use as input for RecNet.

B. RecNet Results

Two inputs are used for RecNet: the left RGB image and disparity pair merged into 4 channels, and the corresponding right side RGB-disparity merge, as output; the occupancy maps are given by *binvox* with each of the 40 models multiplied 12 times. Unlike DispNet, RecNet is implemented with a binary entropy *loss* function. Adam optimizer, 100 epochs and *batch* of 6 are also used. The metrics shown in Table II and Fig. 10 indicate a very low precision that could not be corrected either. with CorrNet aggregation [9]. The generated models are acceptable in both categories as seen in Fig. 11. There is usually a loss of fine detail such as thin lines or small extremities.

TABLE II. RecNet Training Metrics for 100 Epochs

Metric	Train	Validation
loss	0.007	0.0128
accuracy	0.1256	0.1139

C. Intersection over Union Metric

The IoU metric calculates the total number of matches between two 3D meshes produced by an intersection or coincidences in both arrays over the total number of occupied cells of both meshes. This is represented in the following equation

$$IoU = \frac{\sum_{i,j,k} I(\hat{V}^{(i,j,k)} > t) I(V^{(i,j,k)})}{\sum_{i,j,k} I[I(\hat{V}^{(i,j,k)} > t) + I(V^{(i,j,k)})]} \quad (2)$$

Where $\hat{V}(i,j,k)$ is the generated mesh, $V(i,j,k)$ the original mesh; t is a limit (threshold) applied to the values of the mesh produced to approximate the values to 0 or 1 respectively. The I function is a limit function to obtain only binary values before performing the summations. The IoU result is a percentage value indicating the matches between arrays.

For both categories with normal and blurred variants on each one; the resulting IoU metric is shown in Table III. Fig. 12 indicates the comparison of metric values for each category, as well as the IOU frequencies for grouping entries with an interval of 0.05.

TABLE III. Averages IoU Values for Every Category Trained

Category	Value
Scenes	0.896
Statues	0.815
Blurred scenes	0.895
Blurred statues	0.816
Average	**0.856**

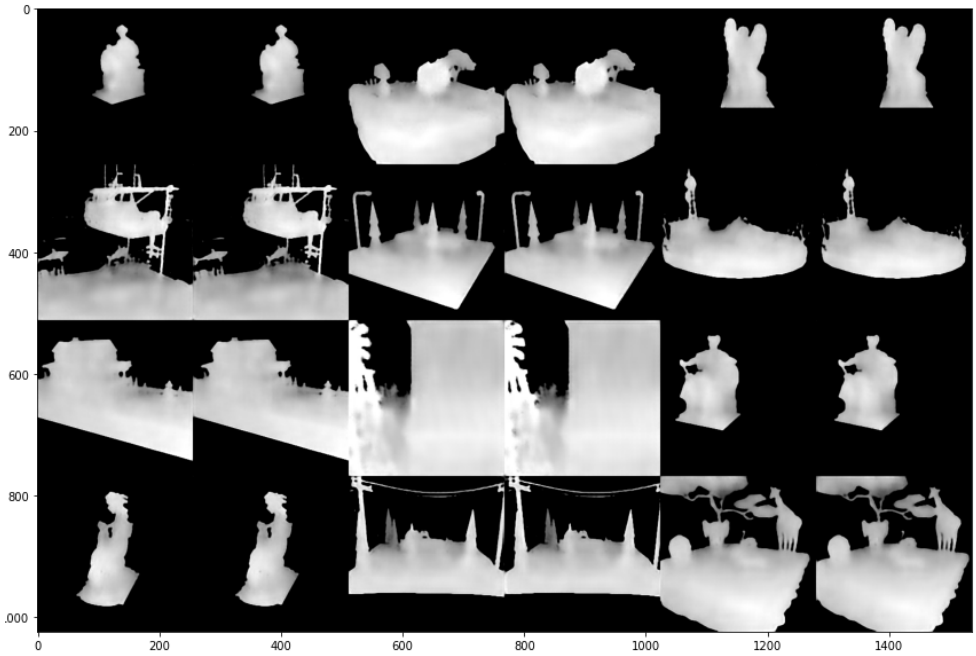

Fig. 9. Disparity Maps Generated by DispNet.

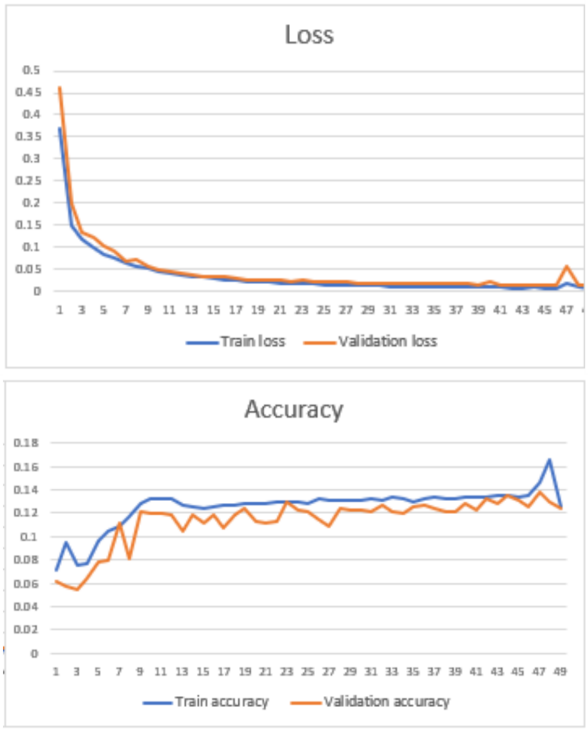

Fig. 10. RecNet Loss and Accuracy during 100 Epochs of Training.

VII. Discussion

Our research shows that a 3D volumetric reconstruction can be achieved from a small set of pairs of stereoscopic images. The metrics get values of 80% for a small dataset compared to [7], where the author uses more than a million pairs of images, with values between 65% and 70% for model precision. The additional information given by the depth maps helps to get more accurate reconstructions. A smaller set of better quality images requires less computational power and less time consumption, which is ideal for applications on free GPU developing platforms like Google Colab or Kaggle Notebook. It is possible that the balance of quality of inputs and the size of the set leads to good reconstructions with fewer resources. This is related to the main limitations faced during this study: the time and size constraints of GPU use and the access to bigger datasets. Our CNN architecture wasn't specially modified for a 128^3 voxel model, apart from a stride modification on the final layer of the pipeline. Therefore, a more precise model could be achieved with additional configurations on the architecture.

The quality of inputs has been established as an important variable to reconstruction accuracy. The author in [7] used 137×137 RGB pictures, relying more on the number of inputs given. 256×256 or even bigger RGB and depth pictures have shown similar results for this CNN architecture as smaller pictures. This could be extended for reconstruction research on single and multiple RGB images, which usually use low-quality pictures. These results could be extended for real good-quality pictures taken with stereo and depth cameras. Better tuned CNNs based on this architecture can be made for these real pictures, the model resolution could also be extended with extra steps in the pipeline.

VIII. Conclusion

In this work, we have developed a two-step convolutional network architecture for 3d reconstruction from a small dataset of pairs of stereoscopic images with complex scenes and detailed objects. Our work is influenced by depth perception cues noticed as disparity maps which are generated in the first step and then used to feed the reconstruction step. Re-

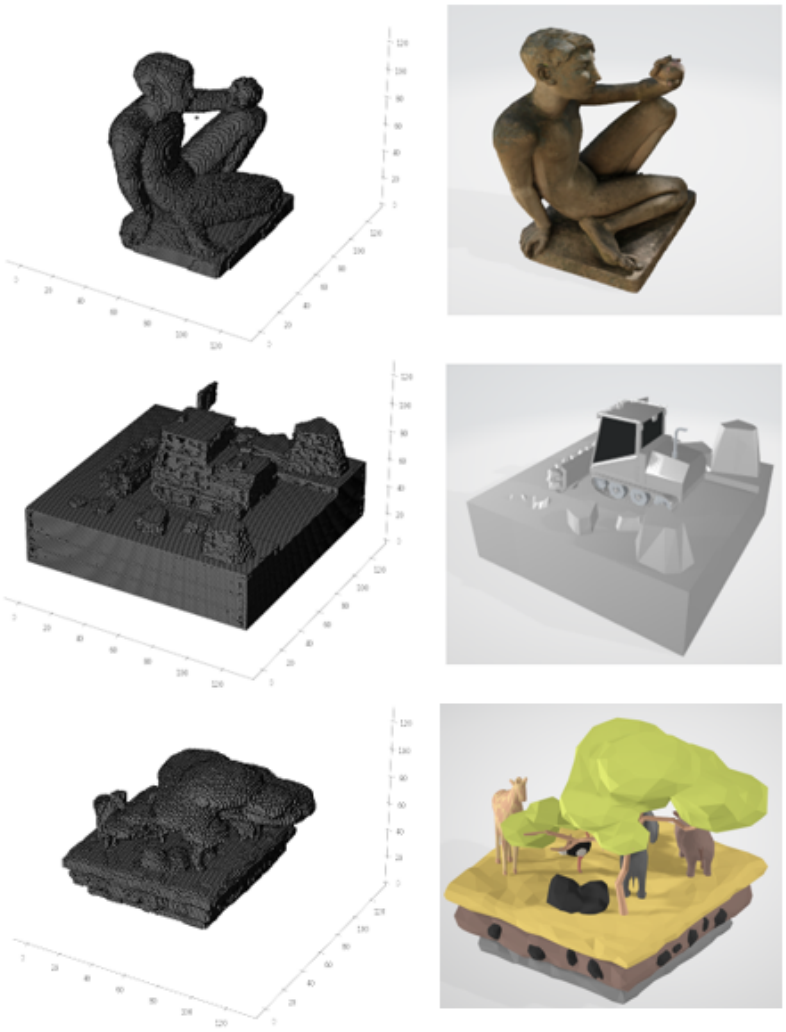

Fig. 11. RecNet Result Meshes (Left) Compared to Ground Truth Meshes (Right).

constructed models have on average 80% on the intersection over union metric. This percentage is similar to state-of-the-art proposals with bigger datasets. The models also have better resolution (128^3) than previous works. This suggests that a set of hundreds of rich and detailed stereoscopic images could be enough for the reconstruction of these volumetric models. The dataset was enlarged with a blurred version of the RGB images as input. Future work will focus on refining the final result with bigger datasets, applying another transformations to images like rotation and contrast; and tuned networks settled for this task that could overcome low accuracy levels on network training and validation.

ACKNOWLEDGMENT

We acknowledge the financial support of the "Proyecto Concytec - Banco Mundial", through its executing unit "Fondo Nacional de Desarrollo Científico, Tecnológico y de Innovación Tecnológica (Fondecyt)", for their research work entitled "Reconstrucción y modelado 3D de las superficies de componentes y piezas de maquinaria pesada usada en Minería, con nivel de precisión milimétrica, para su aplicación en un nuevo proceso optimizado de manutención especializada".

The authors would like to thank EdwinRC (https://sketchfab.com/Edwin3D), noe-3d.at (https://sketchfab.com/www.noe-3d.at), and VIMUNE (https://sketchfab.com/vimune), for the fantastic models used in this study.

REFERENCES

[1] J. LaViola Jr., E. Kruijff, R. McMahan, D. Bowman, I. Poupyrev, *3D User interfaces, theory and practice,* 2017.

[2] E. Abbott, *Chapter 2 Principles of Stereoscopic Depth Perception and Reproduction,* 2004

[3] J.I. Bagnato, *¿Cómo funcionan las Convolutional Neural Networks? Visión por Ordenador,* 2018, *https://www.aprendemachinelearning.com/como-funcionan-las-convolutional-neural-networks-vision-por-ordenador/* , retreived on 17/02/2022.

[4] X. Kang, B. Song and F. Sun, *A deep similarity metric method based on incomplete data for traffic anomaly detection in IoT,* Applied Sciences, 01, 2019.

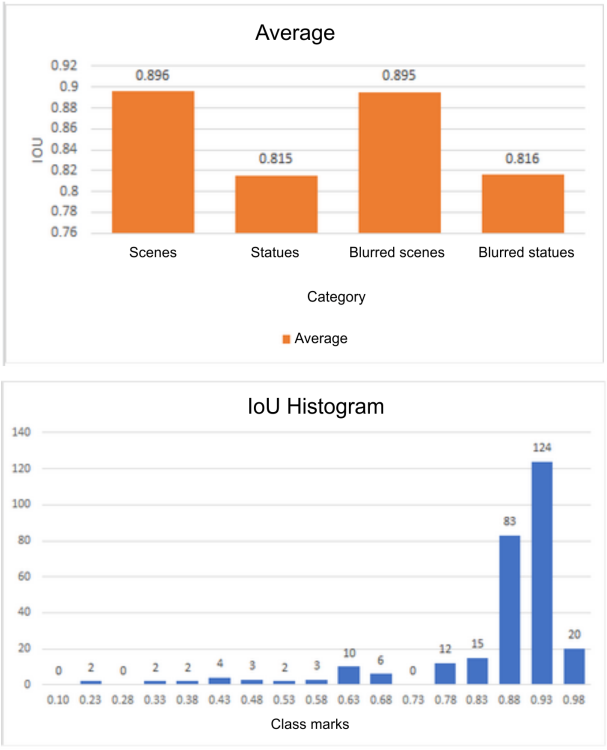

Fig. 12. Left: IOU Metric Average for Each Category. Derecha: IoU Histogram with 0.05 Interval Width.)

[5] S. Saha, *A comprehensive guide to Convolutional Neural Networks — the ELI5 way,* 2018, *https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53?gi=c6d122c7b4c*, retreived on 17/02/2022.

[6] X.F. Han, H. Laga and M. Bennamoun, *Image-based 3D Object reconstruction: state-of-the-Art and trends in the Deep Learning era,* 11 2019.

[7] H. Xie, H.Yao, S. Zhou, S. Zhang, X. Sun, and W. Sun, *Toward 3d object reconstruction from stereo images,* 10 2019.

[8] H. Xie, H. Yao, X. Sun, S. Zhou and S. Zhang, *Pix2Vox: Context-aware 3D reconstruction from single and multi-view images,* 01 2019.

[9] H. Xie, H. Yao, S. Zhang, S. Zhou, and W. Sun, *Pix2vox++:Multi-scale context-aware 3d object reconstruction from single and multiple images*,International Journal of Computer Vision, 2020.

[10] W. Wang, Q. Huang, S. You, C. Yang and U. Neumann, *Shape inpainting using 3D Generative Adversarial Network and Recurrent Convolutional Networks,* 11 2017.

[11] C.R. Qi, H. Su, M. Niessner, A. Dai, M. Yan, and L. Guibas,*Volumetric and multi-view CNNs for object classification on 3d data,*04 2016.

[12] I. Dryanovski, M. Klingensmith, S. Srinivasa, and J. Xiao, *Large-scale,real-time 3d scene reconstruction on a mobile device,*AutonomousRobots, vol. 41, 02 2017.

[13] A. Kar, C. Häne, and J. Malik, *Learning a multi-view stereo machine,*08 2017.

[14] G. Riegler, A.O. Ulusoy, H. Bischof and A. Geiger, *OctNetFusion: learning depth fusion from data,* 10 2017.

[15] Z. Liu, H. Tang, Y. Lin and S. Han, *Point-Voxel CNN for efficient 3D Deep Learning,* 07 2019.

[16] M. Firman, O. Aodha, S. Julier, and J. Gabriel, *Structured prediction of unobserved voxels from a single depth image,* pp. 5431–5440, 062016.

[17] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen, *3d object dense reconstruction from a single depth view,*IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PP, 02 2018.

[18] C. Häne, S. Tulsiani, and J. Malik, *Hierarchical surface prediction for 3d object reconstruction,* pp. 412–420, 10 2017.

[19] M. Tatarchenko, S. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox, *What do single-view 3d reconstruction networks learn?,* pp. 3400–3409, 06 2019.

[20] J. Choe, S. In, F. Rameau, M. Kang and I. S. Keon, *VolumeFusion: deep depth fusion for 3D scene reconstruction,* 08 2021.

[21] Z. Murez et al., *Atlas: end-to-end 3D scene reconstruction from posed images,* ECCV 2020: Computer Vision – ECCV 2020 pp 414-431, 11 2020.

[22] T. Jadhav, K. Singh and A. Abhyankar, *Volumetric 3D reconstruction of real objects using voxel mapping approach in a multiple-camera environment,* Turk J Elec Eng & Comp Sci (2018) 26: 755 – 767, 03 2018.

[23] J. Watson, O. Mac Aodha, D. Turmukhambetov, G.J. Brostow and M. Firman. *Learning stereo from single images,* 08 2020.

[24] J. Varley, C. DeChant, A. Richardson, J. Ruales and Peter Allen, *Shape completion enabled robot grasping,* 03 2017.