Three-Dimensional Shape Reconstruction from a Single Image by Deep Learning

Kentaro Sakai[1], Yoshiaki Yasumura[2]
Graduate School of Engineering and Science
Shibaura Institute of Technology
Saitama, Japan

Abstract—**Reconstructing a three-dimensional (3D) shape from a single image is one of the main topics in the field of computer vision. Some of the methods for 3D reconstruction adopt machine learning. These methods use machine learning for acquiring the relationship between 3D shape and 2D image, and reconstruct 3D shapes by using the learned relationship. However, since only predefined features (pixels in the image) are used, it is not possible to obtain the desired features of the 2D image for 3D reconstruction. Therefore, this paper presents a method for reconstructing 3D shapes by learning features of 2D images using deep learning. This method uses Convolutional Neural Network (CNN) for feature learning to reconstruct a 3D shape. Pooling layers and convolutional layers of the CNN capture spatial information about images and automatically select valuable image features. This paper presents two types of the reconstruction methods. The first one is to first estimate the normal vector of the object, and then reconstruct the 3D shape from the normal vector by deep learning. The second one is direct reconstruction of the 3D shape from an image by a deep neural network. The experimental results using human face images showed that the proposed method can reconstruct 3D shapes with higher accuracy than the previous methods.**

Keywords—Computer vision; 3D reconstruction; deep learning; convolutional neural network; feature learning; normal vector

I. Introduction

Reconstructing three-dimensional (3D) shapes of objects from two-dimensional (2D) images is one of the most attractive areas of computer vision. Shape-from-shading is a traditional method of 3D reconstruction that uses object shadow images, reflection characteristics, and light source information [1, 2, 3, 4, 5]. Although 3D shapes can be reconstructed in a relatively short time by the methods, it is not practical due to strong constraints such as object texture. Tal Hassner et al. proposed a method of 3D shape reconstruction by the nearest neighbor method using 2D image and 3D shape database [6]. This method searches the database for the most similar image patches of the input image and integrates them to estimate the shape. However, it takes a lot of time to find the most similar patch. Mori et al. proposed a reconstruction method using bagging [7]. Bagging is a kind of ensemble learning method. This method learns the relationship between the pixel values of the patch in the image and the normal vector at the center of the patch. The patch is the window of the k*k size in the image. The learning method is bagging whose weak learner is a regression tree. By using the learned relationship, the method reconstructs 3D shape from an unknown image. From the experimental results, this method could reconstruct 3D shapes more accurately than the previous methods. However, this method uses predefined feature of the image, the pixel value of the patch. If the method can use the desirable features in images, it reconstructs 3D shape more accurately. The desirable features can be acquired by learning. For acquiring the features of the image, deep learning is the most suitable method because the deep learning has ability for feature learning [8, 9, 10].

Therefore, this paper presents a method for 3D reconstruction from a single image by deep learning. This method adopts Convolutional Neural Network (CNN) [11, 12, 13] for feature learning because it is successful in various fields such as object recognition and semantic segmentation. This method outputs the 3D coordinates of an object from the pixel values of 2D image. This method acquires spatial information of an image using the pooling layers and the convolutional layers of the CNN. From the acquired features, this method reconstructs the 3D shape from a 2D image. For 3D reconstruction, two types of the reconstruction methods are proposed in this paper. The first one is first estimating the normal vector of the object, then reconstructing the 3D shape from the normal vector by deep learning. The second one is direct reconstruction of 3D shape from an image by a CNN.

This paper is organized as follows. In Section 2, a method to reconstruct the 3D shape from a single image by using CNN is described. First, a method for 3D reconstruction by estimating normal vector of the object is presented. Second, this paper presents a method for directly reconstructing 3D shape from a single image. In Section 3, the experimental settings such as hyper parameters and experimental results are presented. Finally, conclusion of this paper is described in Section 4.

II. Three Dimensional Reconstruction by Deep Learning

This section presents a method for reconstructing 3D shape from 2D image by the CNN.

A. *Previous Works for 3D Reconstruction*

Fig. 1 shows the overview of the previous method for reconstructing a 3D shape by bagging [7]. This method learns the relationship between 2D images and normal maps of the 3D shapes. First, a normal map is created from the 3D coordinates of the object. Next, the method acquires the relationship between the pixel values in the patch and the normal vector of the center of the patch. The patch is a window of the arbitrary k*k size. The relationship is acquired by bagging with a regression tree as a weak learner. By using the acquired relationship, the normal map of the shape is estimated from a new 2D image. Finally, this method reconstructs the 3D shape from the estimated normal map. Since this method uses predefined features, it cannot reconstruct 3D shape from an image accurately because it does not use the valuable features for 3D shape reconstruction.

B. 3D Reconstruction Method by Estimating Normal Vector

Here, this paper proposes a method for 3D reconstruction by estimating normal vectors of the object. This method basically consists of same procedures of the previous work in Fig. 1. The differences between the proposed method and the previous method are that (1) normal map estimator is created with CNN, and (2) 3D coordinates of the object are estimated from the normal map by deep learning. Since the previous method adopts traditional learning methods for estimating normal vectors, the estimated normal vectors are not sufficiently accurate. This is because the previous method cannot utilize valuable features of the image for 3D reconstruction. On the other hand, Convolutional Neural Network (CNN) enables to acquire valuable features of the image automatically. Therefore, the proposed method can reconstruct 3D shape more accurately by using CNN. The input of the CNN is pixel values of an image and the output is normal vectors of the correspondence pixels.

Next, from the estimated normal vector, the proposed method reconstructs 3D shape of the object. Since 3D coordinate estimation from the normal vectors is hard problem, some previous works tackle this problem. These works uses some constrain of the surface such as smoothness. To solve this problem, the proposed method create a deep neural network to learn the relationship between the normal vectors and 3D coordinates of the object, and the network estimates 3D coordinates from the normal vectors. The merit of this method is that the 3D coordinates can be estimated more accurately by using deep learning in spite of the noisy normal vectors. The input of the deep learning is the normal vectors of the pixel, and the output is the 3D coordinates of the correspondence pixels.

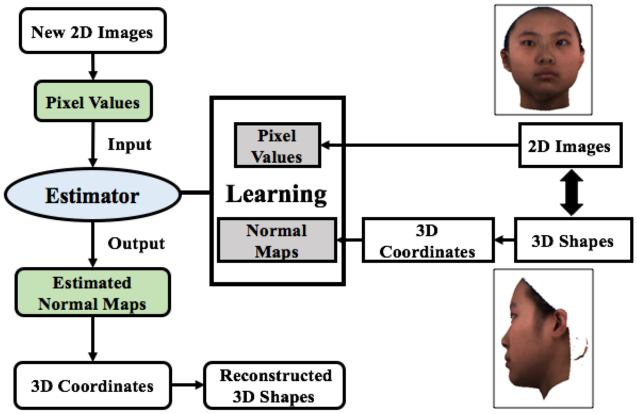


Fig. 1. Overview of the Previous Method.

C. Direct 3D Reconstruction Method

The overview of the direct reconstruction method is shown in Fig. 2. This method directly estimates the 3D coordinates of an object from a 2D image without estimating normal vectors.

The input of this method is pixel values of a 2D image and the output is the corresponding 3D coordinates. CNN learns the relationship between the pixel values and the 3D coordinates of the shape.

Fig. 3 shows the structure of the CNN. The convolution layer contains a set of filters. The filter extracts features in the image. Since some filters detect edge information in the image, they can extract unsmooth shape such as eyes and mouth. Similarly since some filters detect gradation information, they can extract smooth shape such as cheek and forehead. In the learning process of CNN, it can acquire the valuable filters by learning. This characteristic of the convolution layer enables to learn features for 3D shape reconstruction.

The pooling layer summarizes the features in patches, and reduces the spatial dimensions. By using pooling layers, the CNN can be robust to the position shift and rotation in the image. The convolution layers and pooling layers enable the CNN to learn features in the image for 3D reconstruction.

Finally fully connected layers outputs 3D coordinates by combining the features from the convolution layers and pooling layers.

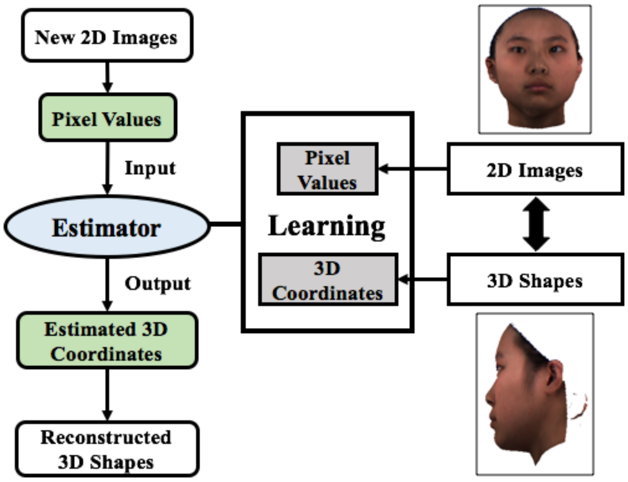


Fig. 2. Overview of the Direct 3D Reconstruction.

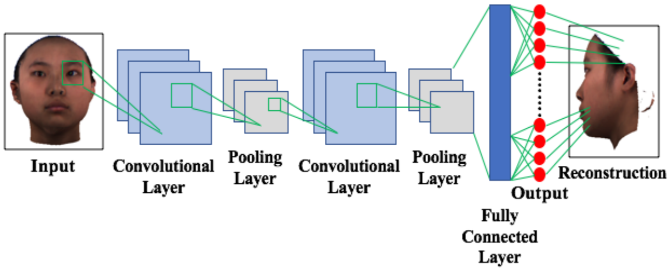


Fig. 3. Convolutional Neural Network for 3D Reconstruction.

III. Experiment

This section presents experimental evaluation of the proposed method. This experiment is conducted with 2D images and 3D coordinates data of actual faces.

A. *Experimental Setting*

The data of this experiments is Beijing University of Technology's "BJUT-3D Face Database" [14]. Fig. 4 shows an example of the data. The dataset consists of 463 face data that are 2D face images and the corresponding 3D coordinates. The dataset is split into training set (313 data), validation set (75 data), and test set (75 data). The training set is used for creating an estimator to reconstruct 3D shape from an image. The validation set is used for preventing overfitting and tuning hyper parameters. The test set is used for accuracy evaluation of 3D shape reconstruction from an unknown image.

To experiment by using the CNN, the proposed method needs to adjust hyper parameters. Epoch is the number of iteration that training data is repeatedly learned. In mini-batch learning, training set is divided into small subsets (mini-batches). The mini-batch is used for reducing the variance of the gradient. The number of filters indicates the number of filters that are used for features extraction in a convolutional layer. Increasing the number of filters enables to obtain various features. Dropout means disabling some of the nodes in the layers with an update process. Dropout is a way to prevent CNN from overfitting.

To evaluate the robustness of the proposed method, this paper experiments with 2D images of five types of light conditions such as light source direction and light color. The light conditions are: (1) light from upper right (2) darker light (3) light from right side (4) blue light (5) light from upper side. The examples of the images of each light condition are presented in Fig. 5.

B. *Experimental Results of Reconstructed Normal Vector*

Table I shows experimental results compared with the previous methods in the various types of the light conditions in Fig. 5. The evaluation of the results uses cosine similarity between the estimated normal vectors and the true normal vectors of the object. If the cosine similarity is higher, the result is better. The proposed method exceeded 0.92 under all condition. On the other hand, the previous method has lowered the value under some conditions. These results showed that the proposed method is more robust for various light conditions than the previous methods. Averagingly the proposed method achieved the best result.

C. *Experimental Results of Reconstructed 3D Shape*

Table II shows experimental results for each hyper parameter by the direct reconstruction method. This experiment adopts Mean Square Error (MSE) for evaluating the results of the proposed method. The MSE means difference between the real shape and the reconstructed shape. From the table, the error is the lowest when the convolution layer has 64 filters. In general, it is assumed that if the number of filters is large, various characteristics can be extracted. However, the error is increasing when the number of filters is the largest. This is because unimportant features are extracted by using too many filters.

The proposed method reduced test error by using dropout. Without dropout, the error in the training set decreases, but the error in the test set increases. From the results, dropout prevents the CNN from overfitting.

Fig. 6 shows the transition of the errors of training and validation data. Since the errors decreases as learning progresses, learning for reconstruction is performed well. Table III shows the errors of the 3D reconstruction in the light condition (1). The table shows that the error by the direct reconstruction method is the lowest. Also the MSE of the reconstruction method by estimating normal vector is less than those of the previous methods. This is due to the reconstruction method from the normal vectors by deep learning. From this result, the proposed two methods can reconstruct more accurately than the previous methods. Table IV shows the MSE by the direct reconstruction method under the different light conditions. The MSEs by the direct reconstruction method in all conditions are higher than the best MSEs of the previous methods as shown in Table III and Table IV. The direct reconstruction method did not lower the MSE values much even under adverse conditions.

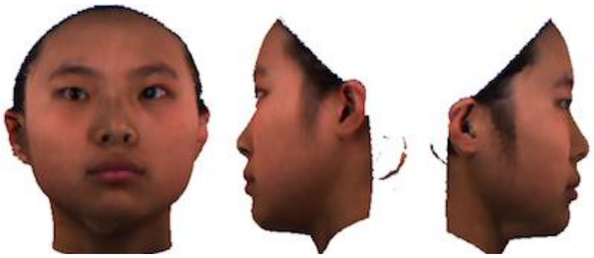


Fig. 4. Face Data.

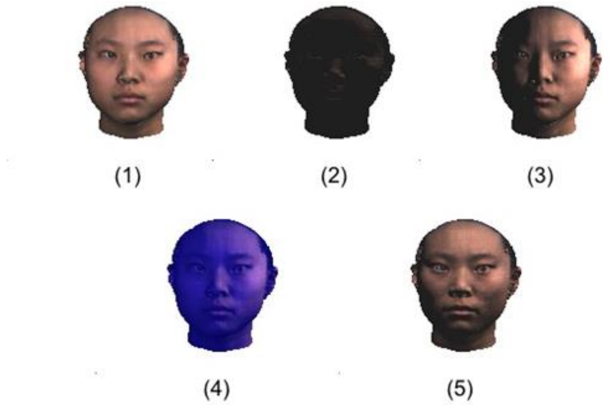


Fig. 5. Examples of Face Images under different Light Condition.

TABLE. I. Results of the Estimated Normal Vectors

	(1)	(2)	(3)	(4)	(5)
Bagging [7]	0.931	0.753	0.882	0.857	0.800
Bagging by optimal learners [7]	0.980	0.884	0.956	0.950	0.908
The proposed method	0.952	0.951	0.950	0.944	0.922

TABLE. II. EXPERIMENTAL RESULTS FOR EACH HYPER PARAMETER

Epoch	50			
No. of mini-batches	8			
No. of filters	32	64	128	64
Dropout	Yes	Yes	Yes	No
Training error	25.51	12.85	17.87	11.07
Test error	16.90	11.39	23.74	14.29

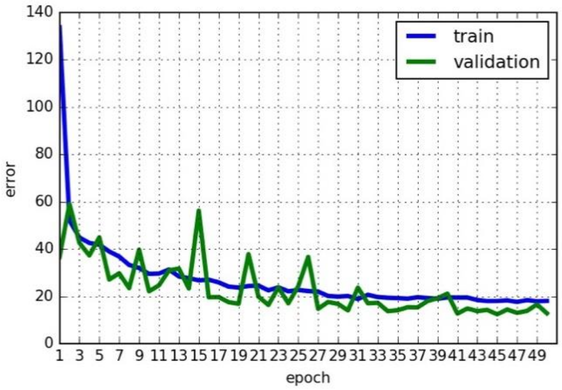


Fig. 6. Sum of the Error in Learning Process.

TABLE. III. THE ACCURACY OF THE RECONSTRUCTED SHAPE

	MSE
Nearest Neighbor method [2]	84.37
Bagging method [3]	80.59
Reconstruction method by estimating normal vector	29.53
Direct reconstruction method	24.87

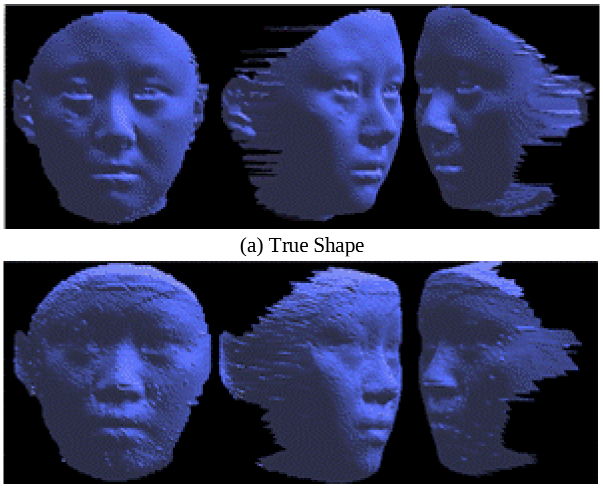
TABLE. IV. MSE BY DIRECT RECONSTRUCTION METHOD UNDER DIFFERENT LIGHT CONDITION

	(1)	(2)	(3)	(4)	(5)
MSE	24.87	30.93	34.25	37.06	27.81

Fig. 7 shows the comparison between the true shape and the reconstructed shape by the proposed method. The true 3D shape is shown in Fig. 7(a), and the reconstructed 3D shapes by the direct reconstruction are shown in Fig. 7(b). From the reconstruct result, the face outline is noisy, but the mouth and nose are reconstructed well. Fig. 8 shows the low accuracy result by direct reconstruction method. Fig. 8(a) shows the true shape of the face, and Fig. 8(b) shows the reconstructed face shape. From the result, the shape is roughly well reconstructed, but there is a lot of noise overall. For more accurate reconstruction, reducing such noise is future task.

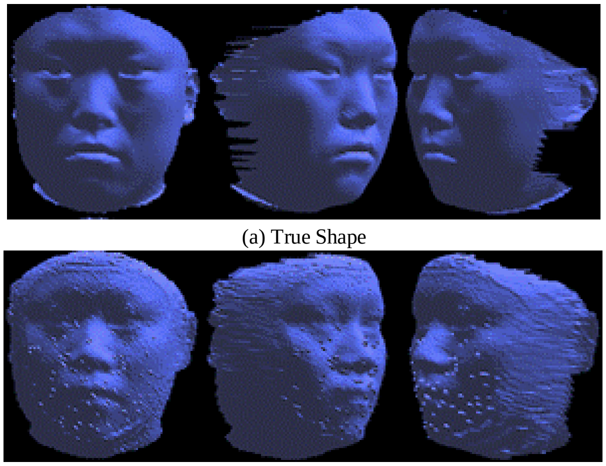
Fig. 9 shows the error area that is painted black. Fig. 9(a) shows higher error area in the shape reconstructed with high accuracy, Fig. 9(b) shows the error area in the shape reconstructed with low accuracy. The black area in the figure indicates that the error of the area is higher than 3.0. From Fig. 9(a), the shape can be well reconstructed even for a part such as the nose and the eye which is difficult to reconstruct. However, the shape in Fig. 9(b) has higher errors throughout the face. This result indicates that the reconstruction result varies in accuracy depending on the image.

The experimental results show that the proposed method can reconstruct better than the previous methods. However, there are the problems that the reconstruction shape contains noise and the reconstruction error depends on the image. To resolve the problem, the number of the layer in the CNN needs to be increased.


(a) True Shape

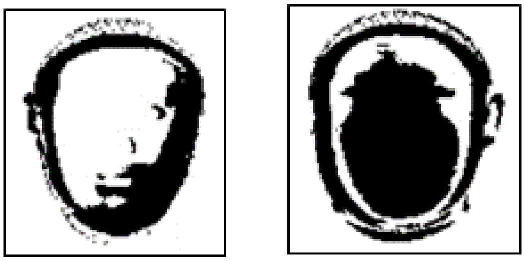
(b) Reconstructed Shape

Fig. 7. Results of High Accuracy Reconstruction.


(a) True Shape

(b) Reconstructed Shape

Fig. 8. Results of Low Accuracy Reconstruction.


(a) High Accuracy

(b) Low Accuracy

Fig. 9. Higher Error Area.

IV. CONCLUSION

This paper presented a method for reconstructing a 3D shape from an image by deep learning. The proposed method outputs 3D coordinates of the shape from the pixel values of an image. The CNN can acquire the valuable features of the image. This paper proposed two types of the reconstruction methods. The first one is first estimating the normal vector of the object, then reconstructing the 3D shape from the normal vector by deep learning. The second one is directly reconstruction of 3D shape from an image by a deep neural network.

From the experimental results, the proposed two methods can reconstruct 3D shape better than the previous method. From the experimental results under the various types of light conditions, the proposed methods are robust to the light conditions. The direct reconstruction method achieved the better results than the previous methods and the reconstruction method by estimating normal vectors by deep learning.

For future work, Generative Adversarial Networks (GAN) [15, 16, 17] is one of the most promising methods for more accurate and smooth shape reconstruction. GAN consists of two neural networks, generative network and discriminative network. For 3D reconstruction, generative network creates a 3D shape from a 2D image, and then discriminative network distinguishes shapes created by the generative network from the true shapes. More accurate shape will be generated by contesting between the two networks.

REFERENCES

[1] R. Zhang, P.S. Tsai, and J.E. Cryer, M. Shah, "Shape-from-Shading: a survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.21, No. 8, pp. 690-706, (1999).

[2] F. Srtori and E.R. Hancock, "Victor transport for shape-from-shading", Pattern Recognition, Vol. 38, No. 8, pp. 1239-1260, (2005).

[3] Abdelrehim H. Ahmed, Aly A. Farag, "A New Formulation for Shape from Shading for Non-Lamberitian Surfaces", 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR06), Volume 2, pp. 1817 - 1824, (2006).

[4] D. Yang, J. Deng, “Shape From Shading Through Shape Evolution”, The IEEE Conference on Computer Vision and Pattern Recognition, pp.3781-3790, (2018).

[5] M. W. Tao, P. P. Srinivasan, S. Hadap, S. Rusinkiewicz, J. Malik, R. Ramamoorthi, “Shape Estimation from Shading, Defocus, and Correspondence Using Light-Field Angular Coherence”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, Issue 3, pp.546-560, (2016).

[6] T. Hassner and R. Basri, “Example Based 3D Reconstruction from Single 2D Images”, IEEE Conference on Computer Vision and Pattern Recognition Workshop, pp. 15-15, (2006).

[7] Y. Mori, Y. Yasumura, and K. Uehara:"3D Face Reconstruction from a Single Image Using Machine Learning Methodology", Proceedings of the 2009 ICMITA, pp.29-32 (2009).

[8] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3415-3424 (2017).

[9] Z. Han et al., "3D2SeqViews: Aggregating Sequential Views for 3D Global Feature Learning by CNN With Hierarchical Attention Aggregation," IEEE Transactions on Image Processing, vol. 28, no. 8, pp. 3986-3999, (2019).

[10] Wen Y., Zhang K., Li Z., Qiao Y. , A Discriminative Feature Learning Approach for Deep Face Recognition, Proc. of ECCV 2016, pp. 499-515 (2016).

[11] Andrew G. Howard, Menglong Zhu, Bo Chen,Dmitry Kalenichenko, Weijun Wang, TobiasWeyand, Marco Andreetto, and Hartwig Adam.Mobilenets: Efficient convolutional neural net-works for mobile vision applications.CoRR,abs/1704.04861, (2017).

[12] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick; Mask R-CNN, The IEEE International Conference on Computer Vision (ICCV), pp. 2961-2969 (2017).

[13] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate nrain lesion segmentation, Medical Image Analysis, Vol;. 36, pp.61-78 (2017).

[14] "BJUT-3D Face Database", Multimedia & Intelligent Software Technology Beijing Municipal Key Laboratory, Beijing University of Technology.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, “Generative Adversarial Nets” , Advances in Neural Information Processing Systems 27, (2014).

[16] M. Mirza, S. Osindero, “Conditional Generative Adversarial Nets”, arXiv preprint arXiv:1411.1784, (2014).

[17] P. Isola, J. Zhu, T. Zhou, A. A. Efros, “Image-To-Image Translation With Conditional Adversarial Networks”, The IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125-1134, (2017).