# Face depth prediction by the scene depth

Bo Jin[a], Leandro Cruz[a, b], Nuno Gonçalves[a, c]
[a] Institute of Systems and Robotics (ISR), Department of Electrical and Computer Engineering (DEEC), University of Coimbra, Coimbra, Portugal
[b] Siemens Process Systems Engineering (PSE), London, England
[c] The Portuguese Mint and Official Printing Office (INCM), Lisbon, Portugal
jin.bo@isr.uc.pt, l.cruz@psenterprise.com, nunogon@deec.uc.pt

***Abstract*—Depth map, also known as range image, can directly reflect the geometric shape of the objects. Due to several issues such as cost, privacy and accessibility, face depth information is not easy to obtain. However, the spatial information of faces is very important in many aspects of computer vision especially in the biometric identification. In contrast, scene depth information is related easier to obtain with the development of autonomous driving technology in recent years. An idea of face depth estimation inspired is to bridge the gap between the scene depth and the face depth. Previously, face depth estimation and scene depth estimation were treated as two completely separate domains. This paper proposes and explores utilizing scene depth knowledge learned to estimate the depth map of faces from monocular 2D images. Through experiments, we have preliminarily verified the possibility of using scene depth knowledge to predict the depth of faces and its potential in face feature representation.**

***Index Terms*—depth estimation, face depth map, scene depth map**

## I. Introduction

In nature, most creatures with a single eye are extinct. This is because most species in nature are the same as humans, and they need two eyes for three-dimensional positioning. This fact accords with Darwin's theory of evolution [1]. Monocular depth estimation is to estimate the distance between each pixel in the image and the source by using an RGB image from a unique perspective. Humans can easily do monocular depth estimation because of the large amount of prior knowledge. However this goal is still difficult to achieve for machines.

In recent decades, biometrics has attracted the attention of researchers because of its uniqueness, stability, versatility and difficulty to steal and forge. Face is one of the most popular biometric features. Nowadays, face based applications widely exist in security, medical, entertainment and other fields [2] [3]. Because the human vision is three-dimensional, 2D face images are lack of face space stereo information. There is no doubt about the importance of facial spatial information. In recent years, advances and popularity of inexpensive RGB-D sensors enable us utilize three-dimensional information [4] [5] [6]. However, it is still not easy to obtain 3D face data due to privacy issues. So monocular depth estimation inspired us to acquire 3D information from 2D face images.

Depth estimation of a scene from a single photo has a wide range of applications in robotics navigation, augmented reality, three-dimensional reconstruction, and autonomous driving. In the ancient Chinese physiognomy [7], the forehead, nose, cheeks, etc. of the face correspond to various terrains in different positions, which makes us think of using the knowledge of the depth of the scenes learned by existing models.

At present, most scene depth estimations are based on the conversion of two-dimensional RGB images to RGB-D images. They mainly use the Shape-from-X methods obtaining the depth shape of the scene from the image brightness, viewing angles, luminosity, texture information, etc. There are also some methods that combine Structure From Motion (SFM) [8] or Simultaneous Localization And Mapping (SLAM) [9] to estimate the camera pose. Although there are many devices that can directly obtain the scene depth, the equipment is expensive. It is also possible to use binocular stereo vision for depth estimation. Because the binocular stereo vision method requires stereo matching to perform pixel point correspondence and disparity calculation, the calculation complexity is also high, especially for low-texture scenes, the matching effect is not good. Therefore, the monocular depth estimation is relatively cheaper and easier to popularize. There is accumulated research work in the monocular scene depth estimation.

Suppose that there is a 2D image $I$, and we need a function $F$ to calculate its corresponding depth $D$. This process can be written as:

$$D = F(I) \tag{1}$$

There is no doubt that $F$ is a very complex function. Because getting the specific depth from a single image is equivalent to inferring the three-dimensional space from the two-dimensional image. Therefore, traditional depth estimation methods don't work very well in monocular depth estimation, people are more focused on studying stereo vision that is to get depth information from multiple images. We can obtain the change of disparity between two pictures according to the change of viewing angle, so as to achieve the purpose of obtaining the depth.

As early as the end of the last century, researchers began to use machine learning methods to estimate depth from a single picture. Deep learning is currently as the most popular tool for function fitting [10], and researchers hope to use it to infer the corresponding depth of one single image through some inherent properties of the input picture. From 2014 to

This work is supported by the Institute of Systems and Robotics (ISR) and the Portuguese Mint and Official Printing Office (INCM) under Grant FACING: BI-BOLSA1.

978-1-7281-4387-3/19/$31.00 © 2019 IEEE

the present, due to the development of big data and GPU computing [11] [12] [13], a series of results in scene depth estimation have been produced by using deep learning.

In this paper, we propose and explore utilizing scene depth knowledge learned to estimate the depth map of faces from monocular 2D images. In experiments, we designed some case studies using the Bosphorus 3D Face Database [14]. Through experiments, we have preliminarily verified the possibility of using scene depth knowledge to predict the depth of a face and its potential in face feature representation.

## II. Related Work

### A. Face Depth Estimation

Since the 1990s, researchers have started to use machine learning methods to estimate the depth of human faces from monocular images. S. H. Lai et al. used the raw image data in the vicinity of the edge to estimate the depth from defocus [15]. Sun and Lam converted the depth estimation of face images into a independent component analysis (ICA) model problem [16]. Kong et al. estimated the face image depth based on similarity by using Delaunay triangulation [17]. Since 2014, with the development of deep learning, researchers have successively used deep learning methods to perform monocular face depth estimation. Cui et al. used a deep neural network with a cascaded FCN and CNN architecture to estimate depth information of RGB face images [18]. Pini et al. used a conditional Generative Adversarial Network for learning to translate intensity face images into their corresponding depth maps [19]. Arslan and Seke applied a conditional Wasserstein GAN structure to perform face depth estimation [20].

### B. Scene Depth estimation

Eigen and Fergus used a multiscale convolutional network architecture to predict the depth map from a single image [21]. Laina et al. proposed a fully convolutional architecture encompassing residual learning to model the mapping between monocular images and corresponding depth maps [22]. Alhashim and Wonka used a standard encoder-decoder architecture with features extracted using pre-trained networks to get the depth [23]. For the above methods, it is necessary to know in advance the reference standard of the depth value corresponding to a large number of input pictures as training constraints, so as to back-propagate in the deep neural network, and train our neural network to perform depth prediction for scenes. It is referred to as supervised learning. In practical, it is not easy to obtain the depth value corresponding to the scene. At present, the commonly used method is to obtain the depth from the infrared sensor such as kinect [24] or with the help of laser lidar. Though the infrared sensor is relatively cheap, the collected depth range and accuracy are limited. In contrast, the cost of lidar is high. Using unsupervised learning for training, we can get a deep neural network model without knowing the depth before. Godard et al. used unsupervised learning method which is without ground truth to estimate the depth. The basic idea is to match the pixels of the left and right views to get the disparity map so as to calculate and optimize the depth map by Left-Right Consistency [25]. For getting a better performance, Godard et al. used self-supervised learning with a standard, fully convolutional, U-Net to predict the depth [26].

## III. Materials and Methods

In general, 3D scene understanding dataset can be divided into outdoor scene dataset and indoor scene dataset. The representative of outdoor scene and indoor scene datasets are KITTI [27] and NYU Depth V2 [4] respectively. In this project, our flowchart is indicated as Fig. 1. These two 3D scene datasets are utilized to trained by supervised or unsupervised learning by various deep neural network structures.

### A. Supervised Learning

Generally, it is required to know in advance the depth values corresponding to a large number of input pictures as training constraints, so as to back-propagate the deep neural network and train our neural network for depth prediction of similar scenes. The loss function of the depth regression problem is considering the difference between the true value of the depth map and the predicted value of the depth regression network. In Densedepth [23], the loss function can be indicated as:

$$L\left(y,\tilde{y}\right) = \frac{c}{n} * \sum_{i}^{n} \left|y_i - \tilde{y_i}\right| + \frac{1}{n}\sum_{i}^{n}\left|g_x\left(y_i,\tilde{y_i}\right) + g_y\left(y_i,\tilde{y_i}\right)\right| + \frac{1 - SSIM\left(y,\tilde{y}\right)}{2} \tag{2}$$

where $y$ indicates the ground truth depth map, and $\tilde{y_i}$ indicates the generated depth map. $c$ is a constant, empirically set to 0.1. $g_x$ and $g_y$ are functions of computing the differences in components $x$ and $y$ for the depth maps gradients. Structural Similarity Index (SSIM) [28] is a metric to measure the similarity between $y$ and $\tilde{y_i}$.

In this strategy, many well-known multi-layer pre-trained networks with different structures can fully utilize the advantages of deep neural networks as a function simulator.

### B. Self-supervised Learning

*1) Stereo training modality:* In stereo vision, it is supposed that there are two cameras $L$ and $R$, and one point whose coordinates are $(x, D)$. The disparity represents the translation value required for the pixels in the left camera to form the corresponding pixels in the right camera. According to the triangle similarity law, the disparity denoted as $dis$ can be calculated as:

$$dis = x_L - x_R = \frac{f * b}{D} \tag{3}$$

where $f$ is the focal length of the camera, and $b$ is the distance between two cameras. Therefore, a mapping function $F$ for predicting the disparity is expected as:

$$I_L\left(dis + x_L\right) = I_L\left(F\left(x_L\right) + x_L\right) = I_R\left(x_R\right) \tag{4}$$

Thus, $I_L$ is used for the input, and $I_R$ is used for the reference, the model for predicting disparity can be achieved.

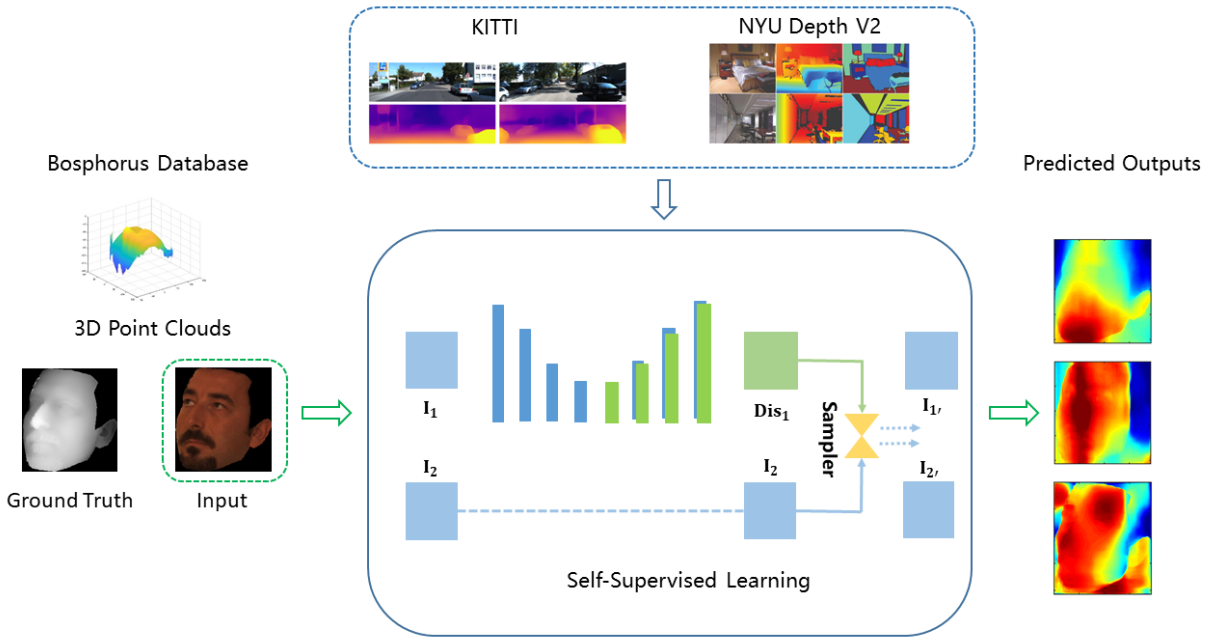


Fig. 1. The schematic diagram of face depth estimation by learning the scene depth knowledge

Finally, the depth map can be obtained with disparity and camera parameters $b$ and $f$. When in the training process, the problem is formulated as the minimization of a photometric reprojection error as:

$$L_p = \min \sum_{\tau} E(I_t, I_{\langle\tau\rangle}) \tag{5}$$

$$L_p = \sum_{\tau} \alpha \frac{1 - SSIM(I_t, I_{\langle\tau\rangle})}{2} + (1-\alpha) \left\| I_t - I_{\langle\tau\rangle} \right\| \tag{6}$$

where $I_t$ represents the target image, $I_\tau$ represents the source image and $I_{\langle\tau\rangle}$ represents the sampled source image. In Monodepth2 [26], the value of $\alpha$ is fixed as 0.85 empirically, and the final loss combining per-pixel smoothness and masked photometric losses is as:

$$L = c_1 L_p + c_2 L_s \tag{7}$$

where

$$L_s = \left| \partial x \frac{d_t}{\overline{d_t}} \right| e^{-|\partial x I_t|} + \left| \partial y \frac{d_t}{\overline{d_t}} \right| e^{-|\partial y I_t|} \tag{8}$$

In the equation above, $\overline{dt}$ represents the mean depth.

In PyDNet [29], in addition to the above losses, the total loss is added to Left-Right Disparity Consistency Loss as:

$$L_c = \sum_{\tau} \left| dis_{\tau}^{l} - dis_{\tau + dis_{\tau}^{l}}^{r} \right| \tag{9}$$

*2) Mono training modality:* Our source image $I_\tau$ could be the second view of $I_t$ in stereo training while $I_\tau$ are the temporally adjacent frames of $I_t$ in mono training, that is, $I_\tau \in \{I_{t-1}, I_{t+1}\}$. Additionally, $I_\tau$ includes both the second view and temporally adjacent frames of $I_t$ in the mix training modality.

### C. Dataset

*1) KITTI dataset:* It is currently the world's largest computer vision algorithm evaluation dataset in autonomous driving scenarios [27]. It contains real image data collected from scenes such as urban areas, rural areas, and highways. Each image can contain up to 15 cars and 30 pedestrians, with various degrees of occlusion and truncation. The entire dataset consists of 389 pairs of stereo images and optical flow diagrams, visual ranging sequences of 39.2 km and more than 200k 3D labeled objects images. The sampled and synchronized frequency is 10 Hz. The scenes of the raw dataset are classified as 'Road', 'City', 'Residential', 'Campus' and 'Person'. There are 8 types annotations in images. They are 'Car', 'Van', 'Truck', 'Pedestrian', 'Person', 'Cyclist', 'Tram' and 'Misc'.

*2) NYU Depth V2 dataset:* It is composed of video sequences of various indoor scenes [4]. The images are recorded by the camera of Microsoft Kinect. The data set contains 1449 densely labeled pairs of RGB and depth images aligned, and contains 464 new scenes in 3 cities and 407024 unlabeled frames.

*3) Bosphorus 3D Face Database:* It contains 105 subjects and 4666 faces in the database [14]. One third of the subjects are professional actors or actresses. There are various expressions (up to 35), head poses (13 yaw and pitch rotations) and varieties of face occlusions for each subject. It can be used for human face processing tasks including but not limited to

facial expression recognition, face recognition under various conditions and 3D face reconstruction.

## IV. Experimental Results and Discussions

### A. Qualitative Results and Analysis

For doing qualitatively evaluation, we present case outputs of face depth maps generated from different advanced models trained with scene depth data (see Fig. 2). The ID of the example case is bs002_CR_RU_0 of Bosphorus 3D Face Database. The ground truth depth image and its corresponding color image are transformed from 3D point cloud file of the Bosphorus database.

Visually, the three sub-figures Fig. 2(c), (d), (g) retain the outline of the human face well. Fig. 2(c) is output from the model PyDNet trained by the KITTI dataset. Fig. 2(c) preserves the contour of the face most intact, but the relative depth displayed is not accurate. The image does not show the correct spatial information for the closest nose tip. Fig. 2(d) is output from the model Monodepth2 trained by the KITTI dataset, and the training modality of it is mono. Fig. 2(d) shows the correct spatial information for the closest nose tip but loses the contour of the ear. Fig. 2(g) is output from the model DenseDepth trained by the NYU Depth V2 dataset. Fig. 2(g) shows the correct spatial information for the closest nose tip, and preserves the contour of the ear. But the depth information shown in the upper left corner and bottom right corner is not correct.

Regarding the texture, Fig. 2(c) seems to be with the largest smoothness among above three depth maps generated visually. Quantitatively, texture is often described by its roughness. We assume the image denoted as $I(x, y)$. Autocorrelation function [30] is usually used as the texture measure as:

$$C\left(\xi, \eta, a, b\right) = \frac{\sum_{x=a-w}^{a+w} \sum_{y=b-w}^{b+w} I(x, y) I(x-\xi, y-\eta)}{\sum_{x=a-w}^{a+w} \sum_{y=b-w}^{b+w} [I(x, y)]^2} \tag{10}$$

where (a,b) is the pixel in the window which size is $(2w + 1) * (2w + 1)$. $\xi, \eta = \pm 0, \pm 1, \pm 2... \pm N$. $\xi$ and $\eta$ are shifting variables on the pixels.

Three autocorrelation function graphs on best three depth maps generated are shown as Fig. 3. In the autocorrelation function graph, a larger downward trend as eta and epsilon increasing means a smaller smoothness of the corresponding image, which accords with our visual feelings. Fig. 3(c) indicated Fig. 2(g) is with a more coarseness because of a larger change in amplitude, which accords with the fact of a face.

### B. Quantitative Results and Analysis

The Structural Similarity Index (SSIM) [28] is the widely used standard for evaluating structural similarity in images that evaluates the quality of a processed image from a ground truth image. We calculate the SSIM for above six models as:

$$SSIM(a, b) = [l(a, b)]^{\alpha} [c(a, b)]^{\beta} [s(a, b)]^{\gamma} \tag{11}$$

where

$$l(a, b) = \frac{2\mu_a \mu_b + C_1}{\mu_a^2 + \mu_b^2 + C_1} \tag{12}$$

$$c(a, b) = \frac{2\sigma_a \sigma_b + C_2}{\sigma_a^2 + \sigma_b^2 + C_2} \tag{13}$$

$$s(a, b) = \frac{\sigma_{ab} + C_3}{\sigma_a \sigma_b + C_3} \tag{14}$$

In the above equations, there are two images denoted as a and b. $\mu_a$ and $\mu_b$ indicate the local mean values of corresponding images, $\sigma_a$ and $\sigma_b$ indicate the standard deviations and $\sigma_{ab}$ indicates the cross-covariance for images.

We test various models trained by scene depth data or face depth data on the Bosphorus database. The results of SSIM values for above six scene models and one face model by the 3D Morphable Model (3DMM) method [31] are summarized as Table I. The 3DMM based method achieves the largest SSIM value, which is not surprising because of face data assistance. Both in visual effects and evaluation indicators, these three generated depth maps can achieve relative good results. Among them, the depth map generated by Densedepth trained by NYU Depth V2 dataset is the best in visual effect and quantitative indicators. We also conclude that models trained by stereo modality for 3D scene depth estimation seem to be not suitable for predicting face depth maps. Fig. 4 shows local SSIM maps of the best three depth maps generated corresponding to the ground truth in the example case. The area with the smaller SSIM value represented by dark pixels corresponds to the area where the generated image is significantly different from the reference image. The area with large local SSIM value represented by bright pixels corresponds to uniform regions of the reference image. Here we find that the depth map generated by Densedepth (NYU Depth V2) predicts well on the nose tip, chin and most of the face area, although its SSIM value is the lowest among above three models.

### C. Face Feature Representation

In this section, we have investigated preliminarily whether the generated depth map can be used as a distinguishable feature to infer its potential in biometric recognition applications. A larger Euclidean distance between features of classes is expected. We assume that in n-dimensional space there are two points $A = (a_1, a_2, ...a_n), B = (b_1, b_2, ...b_n)$. The Euclidean distance is calculated as:

$$d(A, B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{15}$$

The Case 1 is investigated for the same person with different states. In this case, we calculate the average Euclidean distance between images of one subject in the Bosphorus database. The Case 2 is investigated for different persons with different states. In this case, we calculate the Euclidean distance between all images of different subjects. In this section, the model of Monodepth2 with mono training modality is selected

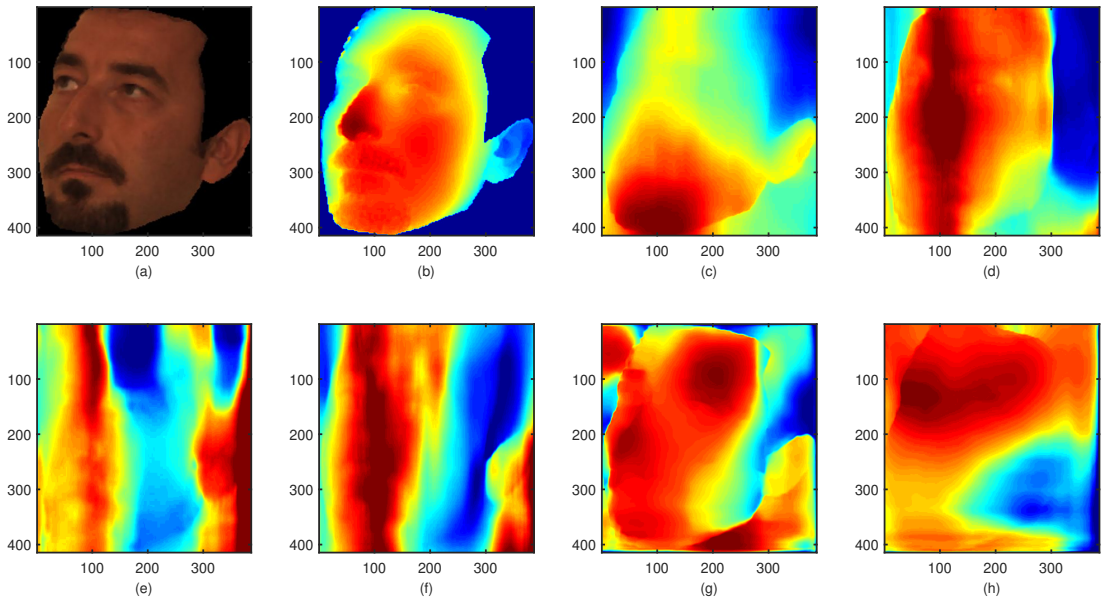


Fig. 2. A case study (ID: bs002_CR_RU_0) of generating depth maps by various models. (a) Ground truth RGB image. (b) Ground truth RGB image. (c) PyDNet pretrained by the KITTI dataset. (d) Monodepth2 trained by the KITTI with mono training modality. (e) Monodepth2 trained by the KITTI with stereo training modality. (f) Monodepth2 trained by the KITTI with mono plus stereo training modality. (g) Densedepth trained by the NYU depth dataset. (h) Densedepth trained by the KITTI dataset.

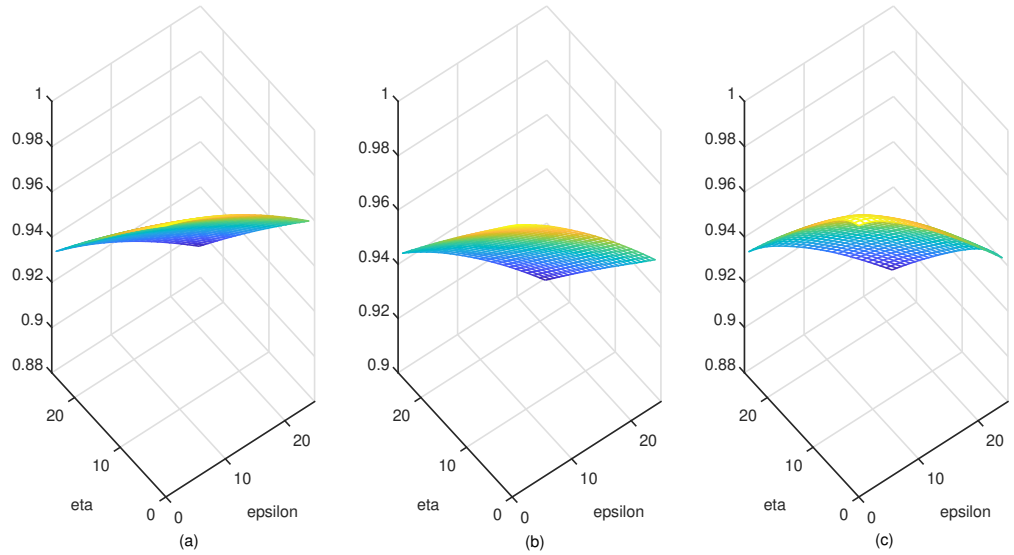


Fig. 3. Autocorrelation function graphs on above best three depth maps generated (a) Fig. 2(c). (b) Fig. 2(d). (c) Fig. 2(g).

TABLE I
**Results measured by the structural similarity (SSIM) index**

| Method | SSIM | Knowledge | Training Modality |
|---|---|---|---|
| PyDNet | 0.552 | Cityscapes + KITTI | Self-Supervised Learning (Mono) |
| Monodepth2 | 0.627 | KITTI | Self-Supervised Learning (Mono) |
| Monodepth2 | 0.510 | KITTI | Self-Supervised Learning (Stereo) |
| Monodepth2 | 0.602 | KITTI | Self-Supervised Learning (Mix) |
| Densedepth | 0.647 | ImageNet + NYU Depth V2 | Supervised Learning |
| Densedepth | 0.609 | ImageNet + KITTI | Supervised Learning |
| 3DMM | 0.745 | Basel Face Model [32] | - |

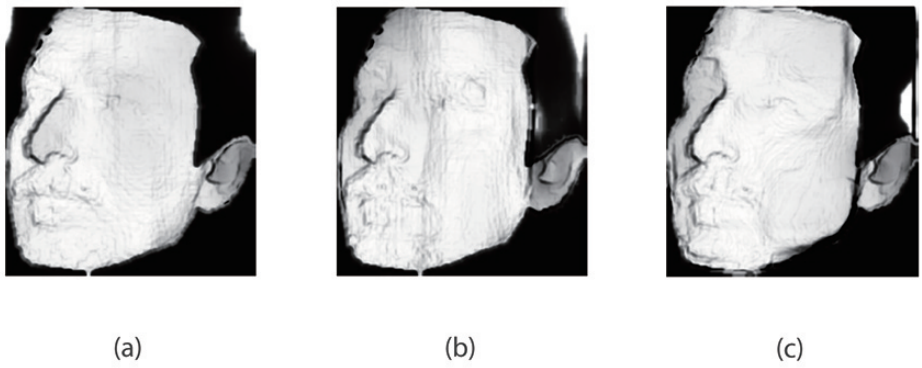


Fig. 4. Local SSIM maps of above best three depth maps generated by (a) PyDNet. (b) Monodepths2 (Mono). (c) Densedepth (NYU Depth V2).

TABLE II
**Case studies of Euclidean Distance in the Bosphrous database**

| Case 1: Intra-Person | RGB (Original) | Depth (Generated) |
|---|---|---|
| Mean Euclidean Distance | $8.64 \times 10^3$ | $2.10 \times 10^4$ |
| Case 2: Inter-Person | RGB (Original) | Depth (Generated) |
| Mean Euclidean Distance | $1.07 \times 10^4$ | $2.22 \times 10^4$ |

to generate the face depth map. In the experiment, all the images are converted to grayscale. The results are shown in Table II, and we find that the mean Euclidean distance between generated depth maps significantly increases to 243% in Case 1 approximately, which are for the same person. In Case 2, the mean Euclidean distance between generated depth maps still significantly increases to 207%. Since that Case 2 is for different persons, the mean Euclidean distance between RGB images increases comparing with Case 1, which is reasonable. Above experiments indicate the generated depth could be a effective feature in biometric recognition applications, because it makes face images more distinguishable.

## V. Conclusion

In this paper, we propose and explore utilizing scene depth knowledge learned to estimate the depth map of faces from monocular 2D images. In the study, we have preliminarily verified the possibility of using scene depth knowledge to predict the depth of faces in the visual effect and quantitative indicators. Through investigating the Euclidean distance changes, we have found that the face depth map predicted by scene models presented could be an effective feature in biometric recognition applications. This paper provides another way to predict the face depth without face related knowledge. In the future, it is expected to be utilized in the face processing tasks.

## References


[1] J. M. Smith and S. J. Maynard, *The theory of evolution*. Cambridge University Press, 1993.

[2] H.-Y. Hu, D. Wu, Y.-Z. You, B. Olshausen, and Y. Chen, "Rg-flow: A hierarchical and explainable flow model based on renormalization group and sparse prior," *arXiv preprint arXiv:2010.00029*, 2020.

[3] B. Jin, L. Cruz, and N. Goncalves, "Deep facial diagnosis: Deep transfer learning from face recognition to facial diagnosis," *IEEE Access*, vol. 8, pp. 123 649–123 661, 2020.

[4] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[5] D. Kim, M. Hernandez, J. Choi, and G. Medioni, "Deep 3d face identification," in *2017 IEEE international joint conference on biometrics (IJCB)*. IEEE, 2017, pp. 133–142.

[6] A. Bud, "Facing the future: The impact of apple faceid," *Biometric Technology Today*, vol. 2018, no. 1, pp. 5–7, 2018.

[7] R. Hassin and Y. Trope, "Facing faces: studies on the cognitive aspects of physiognomy." *Journal of personality and social psychology*, vol. 78, no. 5, p. 837, 2000.

[8] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. M. Reynolds, "'structure-from-motion'photogrammetry: A low-cost, effective tool for geoscience applications," *Geomorphology*, vol. 179, pp. 300–314, 2012.

[9] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 573–580.

[10] Z. Zheng, S. Zhang, B. Yu, Q. Li, and Y. Zhang, "Defect inspection in tire radiographic image using concise semantic segmentation," *IEEE Access*, vol. 8, pp. 112 674–112 687, 2020.

[11] B.-y. Chen, K.-y. Zhang, L.-p. Wang, S. Jiang, and G.-l. Liu, "Generalized extreme value-pareto distribution function and its applications in ocean engineering," *China Ocean Engineering*, vol. 33, no. 2, pp. 127–136, 2019.

[12] H. J. Escalante, V. Ponce-López, J. Wan, M. A. Riegler, B. Chen, A. Clapés, S. Escalera, I. Guyon, X. Baró, P. Halvorsen, H. Müller, and M. Larson, "Chalearn joint contest on multimedia challenges beyond visual analysis: An overview," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 67–73.

[13] M. Zhao, A. Jha, Q. Liu, B. A. Millis, A. Mahadevan-Jansen, L. Lu, B. A. Landman, M. J. Tyskac, and Y. Huo, "Faster mean-shift: Gpu-accelerated embedding-clustering for cell segmentation and tracking," *arXiv preprint arXiv:2007.14283*, 2020.

[14] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *European workshop on biometrics and identity management*. Springer, 2008, pp. 47–56.

[15] S.-H. Lai, C.-W. Fu, and S. Chang, "A generalized depth estimation algorithm with a single image," *IEEE Computer Architecture Letters*, vol. 14, no. 04, pp. 405–411, 1992.

[16] Z.-L. Sun and K.-M. Lam, "Depth estimation of face images based on the constrained ica model," *IEEE transactions on information forensics and security*, vol. 6, no. 2, pp. 360–370, 2011.

[17] D. Kong, Y. Yang, Y.-X. Liu, M. Li, and H. Jia, "Effective 3d face depth estimation from a single 2d face image," in *2016 16th International Symposium on Communications and Information Technologies (ISCIT)*. IEEE, 2016, pp. 221–230.

[18] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2d face recognition via discriminative face depth estimation," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 140–147.

[19] S. Pini, F. Grazioli, G. Borghi, R. Vezzani, and R. Cucchiara, "Learning to generate facial depth maps," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 634–642.

[20] A. T. Arslan and E. Seke, "Face depth estimation with conditional generative adversarial networks," *IEEE Access*, vol. 7, pp. 23 222–23 231, 2019.

[21] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, pp. 2366–2374, 2014.

[22] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[23] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, 2018.

[24] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[25] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[26] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 3828–3838.

[27] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[29] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "Towards real-time unsupervised monocular depth estimation on cpu," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5848–5854.

[30] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: forecasting and control*. John Wiley & Sons, 2011, vol. 734.

[31] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[32] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 2009, pp. 296–301.