Active Face Recognition through View Synthesis

Efstratios Kakaletsis, Nikos Nikolaidis
Department of Informatics, AIIA Laboratory, Aristotle University of Thessaloniki
Thessaloniki, Greece, GR-54124
Email: {ekakalets, nnik}@csd.auth.gr

***Abstract*—Active vision exploits the ability of robots to interact with their environment, towards increasing the quantity / quality of information obtained through their sensors and, therefore, improving their performance in perception tasks. Active face recognition is largely understudied in recent literature. In this paper, we propose an active approach that utilizes facial views produced by facial image rendering. The robot that performs face recognition selects the best candidate rotation around the person of interest by simulating the results of such movements through view synthesis. This is achieved by passing to the robot's face recognizer a real world facial image acquired in the current position, generating synthesized views that differ by $\pm\theta^\circ$ from the current view. Then, it decides, on the basis of the confidence of the recognizer, whether to stay in place or move to the position that corresponds to one of the two synthesized views, so as to to acquire a new real image. Experimental results in two datasets verify the superior performance of the proposed method compared to the respective static approach and an approach based on the same face recognizer that involves face frontalization with synthesized views.**

***Index Terms*—active vision, active face recognition, synthesized facial views, photorealistic facial synthesis**

I. Introduction

Recently the robotics and computer vision communities have started researching more thoroughly the field of active vision / perception and exploration [1]. Active perception methods try to obtain more, or better quality, information from the environment by actively choosing from where, when and how to observe it using a camera (or other sensors), in order to accomplish more effectively tasks such as 3D reconstruction [2], [3], [4], [5], [6] or object recognition [7], [8]. This could be achieved, for example, by moving a camera-equipped mobile robot, e.g. a wheeled robot or a UAV, in positions which provide different, hopefully better, views of the object of interest. Although active 3D object reconstruction has attracted considerable interest, mainly towards tackling the "next-best-view" problem (choosing the next viewing position so as to obtain a detailed and complete 3D object model), active approaches for recognition tasks, particularly for face recognition, are much less frequent in the literature. Deep Learning dominates face recognition research due to its superior performance. However the vast majority of recognition approaches adopt a static approach i.e., an approach that is based on an image acquired from a certain viewpoint, even in setups where an active approach could have been used. Indeed, face recognition can be combined with an active approach for directing the movement of a robot towards capturing the face from more informative views and thus obtain more robust results, at the expense of energy consumption and additional decision time. Synthesized views of faces, whose images were captured through a camera, can be used for robot movement guidance in an active face recognition setup. Instead of having the robot move in a physical way for capturing a novel view, one can use a synthesized view as an aid towards choosing a new viewpoint and improving recognition.

In this paper, we propose an active face recognition approach that utilizes facial views synthesized by photorealistic facial image rendering. Essentially, the camera-equipped robot that performs the recognition selects the best among a number of candidate physical movements (rotations) around the face of interest by simulating their results through view synthesis. In other words, once the robot (that is at a certain location with respect to the subject) acquires an image, it provides the face recognizer with this image as well as with synthesized views that differ by $\pm\theta^\circ$ from the current view. Subsequently, it either stays in the current position or moves to the position that corresponds to one of the two synthesized views. The respective decision is based on the confidence of the three recognitions (on the real and the two synthesized views). In the case of a "move" decision, it proceeds to acquire a "real" image from its new location. The procedure repeats in the same manner, for this location, for one or more steps. Using synthesized facial views facilitates the decision-making procedure by providing estimates of what is to be expected in a new robot position. The proposed method involves a face recognizer that is trained to recognize frontal or nearly frontal faces, a fact that facilitates its real-world application. Despite this, it can recognize successfully input facial images obtained from an arbitrary view point, since it utilizes the ability of a robot to move in order to capture more informative views of the subject. This gives it advantage over static approaches. Indeed, although (static) face recognition is a very mature technology, such approaches can operate successfully only if they have been trained to recognize the subject from view angles similar to the one of the input image. This requires that such methods are trained with images captured from a large number of view angles. In contrast, the proposed approach needs to "know" only frontal or nearly frontal views.

The contributions of the paper can be summarized as follows: a) to our knowledge this is one of a very few (2-

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 871449 (OpenDR). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein.

3) works that deal with active face recognition; b) as far as we are aware, this is the first time facial image synthesis is utilized in an active face recognition setup; c) the method requires no training and the involved face recogniser needs to "know" only frontal images of the subjects, and d) the presented results show that the proposed approach performs better than the respective static approach and an approach that involves face frontalization.

II. Related Work

Despite the fact that active object recognition has attracted considerable interest in the computer vision and robotics communities, active face recognition has been scarcely studied. Such a simple method is described in [7] and comprises of a neural network-based face recognizer along with a decision making controller that decides for the viewpoint changes. The authors in [8] propose a deep learning-based active perception method for embedding-based face recognition and examine its behavior on a real multi-view face image dataset. The proposed approach can simultaneously extract discriminative embeddings, as well as predict the action that the robot must take (stay in place, move left or right by a certain amount, on a circle centered at the person) in order to get a more discriminative view.

A significant number of techniques for synthesizing facial images in novel views appeared in the last years since such images can have a number of applications, e.g., in improving face recognition accuracy. For example, since profile faces usually provide inferior recognition results compared to frontal faces, generative adversarial networks (GANs) based methods for the frontalization of profile facial images [9] or generation of other facial views [10] have been proposed for improving face recognition results. A method for the generation of frontal views from any input view that utilizes a novel generative adversarial architecture (ASN) is described in [11]. Towards improving single-sample face recognition by both generating additional samples and eliminating the influence of external factors (illumination, pose), [12] presents an end-to-end network for the estimation of intrinsic properties of a facial image. In [13], a facial image rendering technique is used both in the training and testing stages of a face recognition approach. A method that produces photorealistic facial image views is described in [14]. The basic idea of this approach is that rotating faces in the 3D space and re-rendering them to the 2D plane can serve as a strong self-supervision. A 3D head model (obtained by utilizing the 3D-fitting network 3DDFA [15] accompanied by the projected facial texture of a single view, is being rotated and multi-view images of the face are rendered using the Neural 3D Differential Renderer [16] along with 2D-to-3D style transfer and image-to-image translation with GANs to fill in invisible parts. This last state-of-the-art method was selected due to its robustness and photorealistic quality for the generation of the synthetic facial images required by the method proposed in this paper.

Although facial view synthesis can improve face recognition performance, active perception methods can be expected to provide better results, in cases where acquisition of additional real world facial views is possible due to the existence of e.g. a wheeled robot.

III. Proposed Algorithm

A. Face Recognition

Let us denote as database subset G a set of training facial images for the persons that shall be recognized. Similarly, the facial images to feed the face recognizer are included in the query (test) set T. The face recognition library face.evoLVe [17] which contains many state of the art deep face recognition models, is used. More specifically, an implementation of a certain face recognition approach of face.evoLVe from the OpenDR Toolkit[1] [18] was used. IR-50 (50 layers) [19] trained on MS-CELEB-1M using an ArcFace [20] loss head was used as the 512-dimensional feature extraction backbone. For the database subset G, face detection, facial landmark extraction and face alignment was based on the face.evoLVe module that is based on MTCNN [21], whereas for the query images in T, these processing steps were based on RetinaFace [22]. Face recognition is performed by a nearest-neighbor classifier that uses Euclidean distance in the 512-dimensional feature space to find the database facial image that best matches the query image. Face recognition confidence $FRC \in [0, 1]$, is also evaluated based on the distance between the input query image and the nearest image in the database G. The FRC is given by the following formula:

$$FRC = 1 - \frac{distance}{max\_distance} \tag{1}$$

where $distance$ is the Euclidean distance of query facial image from the nearest neighbor image in the database G and $max\_distance$ is the maximum such distance.

B. Active Face Recognition Through Synthesized Views

The proposed active face recognition algorithm uses the face recognition confidence FRC and facial images synthesized for view angles around the current robot view, in order to select the next robot movement, towards performing a successful recognition. Starting from an initial position, the robot can take one of the following three decisions: stay at the current position, move by θ° to the right or move by θ° to the left, on a circle centered at the person that is to be recognized, in order to acquire a new image. Depending on the achieved recognition confidence, an additional movement, towards the same direction as the first one, might be decided. More specifically, given a facial query image I_r (subscript r stands for real), captured by the robot camera at the robot starting position, the face synthesis algorithm [14] is utilized to estimate the view angle and then render/generate facial views in 2 different view angles i.e. -15° and $+15^\circ$ in pan with respect to the pan of I_r (and the same tilt as I_r). These two images are denoted by I_s^- and I_s^+ respectively (subscript s stands for synthetic). Then, the face recognizer is fed with these three images I_r, I_s^-, I_s^+ (one real, two synthetic ones). Depending on the image that obtained the biggest face recognition confidence FRC, the robot stays in its current position (if FRC was maximum

[1]OpenDR Toolkit: https://github.com/opendr-eu/opendr

in I_r) or physically moves -15° (or $+15^\circ$) (if FRC was maximum in I_s^- (or I_s^+)) and acquires through its camera a new real image I_r^- (or I_r^+). If a "stay" decision was taken, the algorithm outputs the ID of the person it recognized in I_r and terminates. If the robot moved, face recognition is performed again in I_r^- (or I_r^+) and the obtained FRC is compared to an experimentally evaluated threshold t. In case a high enough confidence was observed, the algorithm outputs the ID of the person it recognized in I_r^- (or I_r^+) and terminates. If not, it tries yet another 15° step (movement) in pan, in the same direction as the first step. In more detail, in this second step, it generates/synthesizes a facial view -15° (or $+15^\circ$) in pan from the current pan value (and the same tilt), denoted as I_s^{--} (or I_s^{++}), and evaluates (by calling the face recogniser) FRC on this synthetic image. If $FRC(I_r^-) > FRC(I_s^{--})$ (or $FRC(I_r^+) > FRC(I_s^{++})$) the algorithm decides that the robot shall stay in its current position, outputs the ID of the person it recognized in I_r^- (or I_r^+) and terminates. Otherwise, the robot physically moves -15° ($+15^\circ$) from its current position, acquires a new image I_r^{--} (I_r^{++}) and the algorithm outputs the ID of the person it recognized in this image.

The performance of the proposed procedure obviously depends on whether the synthesis algorithm [14] estimates with sufficient accuracy the view angle of the query image I_r and also on whether the synthesized views are of good quality. In order to limit the possibly negative effect of these factors on the performance of the algorithm (e.g. by leading it to move towards the wrong direction), the algorithm does not actually take a decision based on the last real image it has visited but does so based on the real image where it has obtained the maximum FRC value. In more detail, if the algorithm took one step of -15°, it takes a decision using the real image I given by:

$$I = \underset{x \in \{I_r^-, I_r\}}{argmax} \left(FRC(x)\right) \tag{2}$$

or the equivalent expression that involves I_r^+, I_r, if a step of $+15^\circ$ has been taken. Similarly, if two steps of -15° each have been performed, the algorithm decides on the person ID using the real image I given by:

$$I = \underset{x \in \{I_r^{--}, I_r^-, I_r\}}{argmax} \left(FRC(x)\right) \tag{3}$$

or the equivalent expression that involves I_r^{++}, I_r^+, I_r, if two steps, of $+15^\circ$ each, have been taken. The pseudocode is presented in algorithm 1.

It should be noted that the actual recognition is always performed on a real image, i.e., an image captured by the robot camera. The synthesized views are only used to aid the robot in deciding whether to move in a new position (and acquire a new image there) or stay in the current position. The rationale behind the proposed approach is that in case the initial robot position is far from a frontal or nearly frontal one, the algorithm will hopefully direct it to move towards a position which is closer to a frontal one. Obviously, the procedure can be generalized to include additional steps (movements), i.e., more than the two movements it currently has. It can also work, in the same way, for tilt.

Algorithm 1 Active Face Recognition Algorithm (2 steps) on Pseudocode

Input: I_r, $threshold$, θ°
Result: $Person_{ID}(I_r)$

$\alpha = Estimate\_View\_Angle(I_r)$
$\mathrm{I}_s^- = Render(\alpha - \theta^\circ, I_r)$
$\mathrm{I}_s^+ = Render(\alpha + \theta^\circ, I_r)$
$I = \underset{x \in \{I_r, I_s^-, I_s^+\}}{argmax}(FRC(x))$
if $I = I_r$ **then**
  $I_{ID} = I_r$
  go to 28
else
  if $I = I_s^+$ **then**
    $\theta_{incr} = +\theta^\circ$
  else
    $\theta_{incr} = -\theta^\circ$

$\mathrm{I}_r^{1step} = Move\_and\_Capture(\alpha + \theta_{incr})$
if $FRC(I_r^{1step}) > threshold$ **then**
  $\mathrm{I}_{ID} = \underset{x \in \{I_r, I_r^{1step}\}}{argmax}(FRC(x))$
  go to 28
else
  I_s^{2step}=$Render(\alpha + 2 * \theta_{incr}, I_r^{1step})$
  if $FRC(I_s^{2step}) < FRC(I_r^{1step})$ **then**
    $I_{ID} = \underset{x \in \{I_r, I_r^{1step}\}}{argmax}(FRC(x))$
    go to 28
  else
    $I_r^{2step} = Move\_and\_Capture(\alpha + 2 * \theta_{incr})$
    $I_{ID} = \underset{x \in \{I_r, I_r^{1step}, I_r^{2step}\}}{argmax}(FRC(x))$
    go to 28

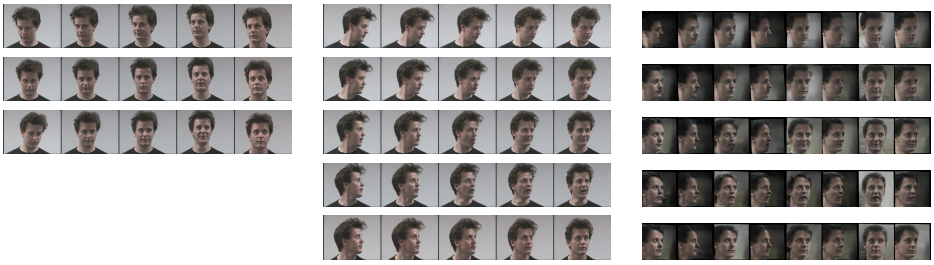
$Person_{ID}(I_r) = Recognize(I_{ID})$

IV. Experimental Evaluation

For the evaluation of the proposed active approach experiments were conducted using the HPID dataset [23] and the Queen Mary University of London Multi-view Face Dataset (QMUL) [24]. In the two datasets, images of all subjects were divided into two non-overlapping subsets: a database subset G (images that the face recognizer uses to decide the ID of the query image through the nearest neighbor classifier) and a query (test) subset T (which includes the images captured by the robot camera in its initial position). Obviously G and T contained images from different pan ranges. This setup was adopted in order to simulate active recognition where the robot is moving only in the pan direction. Concise descriptions of the two datasets are provided below.

A. Datasets

The HPID dataset [23] is a head pose image dataset that consists of 2790 face images of 15 subjects captured by varying the pan and tilt from -90° to $+90^\circ$, in increments of $\theta = 15^\circ$. Two sets of images were captured for



(a) (b) (c)

Fig. 1. (a) Samples from database set G, (b) Samples from test (query) set T, (c) Synthetic images utilised by the algorithm. All images are from the HPID dataset.

each person (93 images in each set). The database subset G (Figure 1.a) contains facial images with tilt in angles $[-30°, -15°, 0°, +15°, +30°]$ and pans $[-15°, 0°]$, i.e., only nearly frontal images. The query subset T (Figure 1.b) contains face images with tilts $[-30°, -15°, 0°, +15°, +30°]$ and pans $[-90°, -75°, -60°, -45°, -30°]$. The selection of the range $[-90°... - 30°]$ in pan, instead of the full (i.e., $[-90°...-30°]$ and $[+30°...+90°]$) semi-circle, in this and the QMUL dataset, was just for simplicity. Similar results were obtained when the experiments involved the entire semi-circle.

Queen Mary University of London Multi-view Face Dataset (QMUL) [24] consists of automatically aligned, cropped and normalised face images of 48 persons. Images of 37 persons are in greyscale (100x100 pixels) whereas those of the remaining 11 persons are in colour and of dimensions 56x56 pixels. For each person 133 facial images exist, populating a viewsphere of $-90°... + 90°$ in pan and $-30°... + 30°$ in tilt in $\theta = 10°$ increments. For the Database split G, images with pan in angles $[-10°, 0°]$ and tilt in the range $[-30°, ..., +30°]$ were used. The Query split T includes images with pan in angles $[-90°, ..., -20°]$ and tilt in the range $[-30°, ..., +30°]$.

B. Experimental Results

The results (in terms of recognition accuracy) are presented in Table I. The line marked "Static" presents the result of the static equivalent of our approach, in which only the initial query facial image is used by the same recogniser involved in the active approach. As can be seen, the proposed active method, implemented to perform up to 4 steps (line "Proposed (Active) (4 steps)") outperforms its static counterpart, increasing the recognition accuracy by 15.61% and 12.97% (absolute increase) in HPID and QMUL datasets, respectively.

The proposed approach was also compared to the frontalization approach that is often used in face recognition when the recognizer is trained only on frontal views. In this case, the facial view synthesis algorithm [14] is used in order to generate a frontal ($0°$ in pan) view from the input (query) image. This image is then provided to the recognizer. The results (line "Frontalization (synthetic frontal views)") show that although frontalization achieves improved performance with respect to the static approach, it is clearly superseded by the proposed active approach.

Statistics regarding the steps taken by the proposed approach were also evaluated and are presented in Table II for HPID dataset. These statistics show that in 26.48% of the cases the robot decided to stay in its initial position whereas in the remaining 73.64% it moved by $\pm 15°, .., \pm 60°$ (one to four steps). It shall be noted however that the decision on the ID of the depicted person is not necessarily obtained from the last position the robot has visited, due to the fact that the image with the maximum recognition confidence (FRC) is used for this purpose (equations (2) and (3)).

TABLE I

FACE RECOGNITION ACCURACY RESULTS AND COMPARISON WITH THE STATIC APPROACH AND OTHER VARIANTS

Method	HPID [23]	QMUL [24]
Static *(non-active, only queries)*	72.49 %	69.88%
Proposed *(Active)* (4 steps)	**88.10%**	**82.85%**
Frontalization *(synthetic frontal views)*	80.75%	75.95%

TABLE II

ACTIVE FACE RECOGNITION STATISTICS (4 STEPS, HPID DATASET): STEPS PERFORMED BY THE ALGORITHM.

Image type	Angle	# Images	Percentage
I_r	0°	397	26.48%
I_r^+	+15°	368	24.54%
I_r^{++}	+30°	84	5.603%
I_r^{+++}	+45°	0	0%
I_r^{++++}	+60°	1	0.066%
I_r^-	-15°	515	34.35%
I_r^{--}	-30°	121	8.07%
I_r^{---}	-45°	9	0.600%
I_r^{----}	-60°	5	0.333%
Total	–	1500	100%

The average number of movements that the algorithm instructs the robot to perform can be easily evaluated from statistics such as the ones presented in Table II. Based on these calculations, the algorithm instructs the robot to make, on average, 0.76 (HPID) or 0.89 (QMUL) movements, a fact that signifies that the time required for active recognition (time for the computations as well as the time for the robot to move) is relatively low. Note that in case the robot decides to performs no movement (stay decision) the number of movements is obviously zero.

V. DISCUSSION AND CONCLUSIONS

An active approach for face recognition that utilizes facial views produced by facial image synthesis was presented in this paper. The robot that performs the recognition selects the best among a number of candidate physical movements around the person of interest by simulating their results through view synthesis. Experimental evaluation showed that the method supersedes both its static version and face recognition that involves frontalization through synthesis of frontal images.

It must be stressed that certain assumptions were adopted in this paper, whereas a number of issues were not fully addressed. First, the actual control of the robot so as to move in

θ° increments on a circle around the person was not dealt with, since it falls outside the scope of the paper. However, a rough estimate of the person position with respect to the robot would suffice to enable robot control. Also, it was assumed that the person being recognized remains relatively static during the recognition process, which can be an acceptable assumption if the process is brief. However, if the person moves, this shall be taken into account by the algorithm. It shall be also noted that the (mild) requirement for a static face is indeed satisfied in certain cases that include sitting or lying persons, as in a healthcare environment, or an elderly care establishment, where a service robot operates in order to aid the inhabitants.

In addition, it was assumed that there are no obstacles in the robot's path. If this is not the case, these obstacles shall be detected (by e.g. depth sensors) and taken into account. Furthermore, obstacles in the space between the robot and the person might occlude the person for certain robot positions. However, since the algorithm decides on the person's identity based on the acquired image where the recognizer obtained the largest recognition confidence, it is rather safe to assume that, in most such cases, the algorithm might not face serious problems, even if it has instructed the robot to move in positions where occlusions occur.

One could also consider, instead of using the synthesized views as proposed in this paper, to estimate the view angle of the robot camera with respect to the person and instruct it to move directly (namely, without intermediate steps) to the position that would allow it to capture a frontal view (0° in pan). However, there are certain issues that make this approach difficult in practice. Indeed, we observed in the experiments that view angle estimates (i.e., those provided by the view synthesis algorithm used) although accurate enough for the purposes of view synthesis, are quite far from the ground truth values, thus rendering this approach problematic. Experiments, which are omitted due to lack of space, verified that such an approach indeed leads to inferior results.

Future plans include evaluation of the algorithm in additional datasets and creation of a realistic simulation so as to investigate some of the issues mentioned above (occlusions, actual robot control, objects that hinder robot motion etc). Employing a more sophisticated face recognizer and comparing it to additional methods, are also planned.

REFERENCES

[1] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, no. 2, pp. 177–196, 2018.

[2] M. Mendoza, J. I. Vasquez-Gomez, H. Taud, L. E. Sucar, and C. Reta, "Supervised learning of the next-best-view for 3D object reconstruction," *Pattern Recognition Letters*, 2020.

[3] J. Delmerico, S. Isler, R. Sabzevari, and D. Scaramuzza, "A comparison of volumetric information gain metrics for active 3D object reconstruction," *Autonomous Robots*, vol. 42, no. 2, pp. 197–208, 2018.

[4] S. Isler, R. Sabzevari, J. Delmerico, and D. Scaramuzza, "An information gain formulation for active volumetric 3D reconstruction," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3477–3484.

[5] C. Forster, M. Pizzoli, and D. Scaramuzza, "Appearance-based active, monocular, dense reconstruction for micro aerial vehicles," *Robotics: Science and Systems Conference, University of California, Berkeley, USA, July 12-16*, 2014.

[6] J. I. Vasquez-Gomez, D. Troncoso, I. Becerra, E. Sucar, and R. Murrieta-Cid, "Next-best-view regression using a 3D convolutional neural network," *Machine Vision and Applications*, vol. 32, no. 2, pp. 1–14, 2021.

[7] M. Nakada, H. Wang, and D. Terzopoulos, "AcFR: Active face recognition using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 35–40.

[8] N. Passalis and A. Tefas, "Leveraging active perception for improving embedding-based deep face recognition," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020, pp. 1–6.

[9] Q. Duan and L. Zhang, "Look More Into Occlusion: Realistic Face Frontalization and Recognition with BoostGAN," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[10] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *IEEE International Conference on Computer Vision*, 2017, pp. 2439–2448.

[11] J. Liao, A. Kot, T. Guha, and V. Sanchez, "Attention selective network for face synthesis and pose-invariant face recognition," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 748–752.

[12] H. Tu, G. Duoji, Q. Zhao, and S. Wu, "Improved single sample per person face recognition via enriching intra-variation and invariant features," *Applied Sciences*, vol. 10, no. 2, p. 601, 2020.

[13] I. Masi, T. Hassner, A. T. Tran, and G. Medioni, "Rapid synthesis of massive face sets for improved face recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 604–611.

[14] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, "Rotate-and-Render: Unsupervised Photorealistic Face Rotation from Single-View Images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5911–5920.

[15] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3D dense face alignment," in *European Conference on Computer Vision (ECCV)*, 2020.

[16] H. Kato, Y. Ushiku, and T. Harada, "Neural 3D mesh renderer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3907–3916.

[17] Q. Wang, P. Zhang, H. Xiong, and J. Zhao, "Face. evoLVe: A high-performance face recognition library," *arXiv preprint arXiv:2107.08621*, 2021.

[18] N. Passalis, S. Pedrazzi, R. Babuska, W. Burgard, D. Dias, F. Ferro, M. Gabbouj, O. Green, A. Iosifidis, E. Kayacan, J. Kober, O. Michel, N. Nikolaidis, P. Nousi, R. Pieters, M. Tzelepi, A. Valada, and A. Tefas, "OpenDR: An Open Toolkit for Enabling High Performance, Low Footprint Deep Learning for Robotics," *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12 479–12 484, 2022.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[20] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *IEEE/CVF conference on Computer Vision and Pattern Recognition, (CVPR)*, 2019, pp. 4690–4699.

[21] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[22] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020, pp. 5203–5212.

[23] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *FG Net workshop on visual observation of deictic gestures*, vol. 6. FGnet (IST–2000–26434) Cambridge, UK, 2004, p. 7.

[24] J. Sherrah and S. Gong, "Fusion of perceptual cues for robust tracking of head pose and position," *Pattern Recognition*, vol. 34, no. 8, pp. 1565–1572, 2001.