Face Attribute Analysis from Structured Light: An End-to-End Approach

Vikas Thamizharasan, Abhijit Das, Daniele Battaglino, Francois F Bremond, Antitza Dantcheva

Noname manuscript No.
(will be inserted by the editor)

Face Attribute Analysis from Structured Light: An End-to-End Approach

Vikas Thamizharasan · Abhijit Das · Daniele Battaglino · Francois Bremond · Antitza Dantcheva

Received: date / Accepted: date

Abstract In this work we explore the use of structured-light imaging for face analysis. Towards this and due to lack of a publicly available structured-light face dataset, we (a) firstly generate a synthetic structured-light face dataset constructed based on the RGB-dataset London Face and the RGB-D dataset Bosphorus 3D Face. We then (b) propose a conditional adversarial network for depth map estimation from generated synthetic data. Associated quantitative and qualitative results suggest the efficiency of the proposed depth estimation technique. Further, we (c) study the estimation of gender and age directly from (i) structured-light, (ii) binarized structured-light, as well as (iii) estimated depth maps from structured-light. In this context we (d) study the impact of different subject-to-camera distances, as well as pose-variations. Finally, we (e) validate the proposed gender and age models that we train on synthetic data on a small set of real data, which we acquire. While these are early results, our findings clearly indicate the suitability of structured-light based approaches in facial analysis.

Keywords Soft biometrics, age estimation, gender estimation, depth imagery, structured light, IRDP, C-GAN.

1 Introduction

Face analysis has witnessed significant advances in the past decades, aiming to extract discriminative features towards determining subject's identity, emotions, as well as face attributes (age, gender, race, hair style). Despite recent progress in face attribute prediction [7, 10, 11, 17, 48] such work is predominantly focused on extracting features in the RGB-color space. More recently, depth information has been additionally

V. Thamizharasan, A. Das, F. Bremond · Antitza Dantcheva
Inria Sophia Antipolis - Méditerranée, France
E-mail: abhijit.das@inria.fr

D. Battaglino
Blu Manta, Sophia Antipolis, France

considered (i.e., RGBD), seeking to achieve higher robustness, stemming from the additional information on geometric relation between objects [24–27, 59]. This has been exploited in 3D reconstruction, robotics, face analysis, pose estimation, segmentation, as well as virtual reality. In the context of face analysis, RGBD has intuitively brought to the fore an increased robustness, in particular in the presence of different poses, illumination variations, and occlusions [10, 11, 17, 31]. We note that such depth data has been acquired with consumer depth cameras (e.g., Kinect V1 [3], Asus Xtion Pro) with underlining technology related to structured-light (SL); time-of-flight (ToF) cameras, as well as passive stereo cameras. Among them, SL-based sensors have been predominantly utilized in face analysis for associated higher resolution of depth information, seamless operation in the presence of low illumination and noise. Further SL-acquired data avoids multi-path errors [40] and artifacts or texture-less regions [5], which are able to alter pertinent face characteristics and therefore might significantly degrade the accuracy of face analysis tasks. Further benefits include the incorporated absolute size, as well as robustness to some presentation attacks. Moreover, SL-images that have been captured in night condition allow for better analysis and e.g., recognition accuracy than RGB-images captured in comparable conditions.

While SL-based technology incorporates named benefits, methods which compute depth maps from raw sensor data still require significant computational and memory resources. In addition, resulting depth estimation is often noisy, prone to errors due to missing zones or occlusions [40]. Such limitations motivate the here pursued end-to-end approach, omitting the explicit step of depth estimation. The benefits of such an end-to-end approach include i) the fact that utilizing raw data without intermediate representations allows for a single optimization of an end-to-end model from infra-red dot pattern (IRDP) data to the target task, ii) improvement of accuracy brought to the fore by implicit task-tailored depth representation, iii) reduction in computational costs, and iv) reduction/elimination of the burden of manual tuning of the parameters to compute the depth-map.

1.1 Contributions

- Due to lack of publicly available datasets, which contain IRDP-images of human faces, we firstly propose a *synthetic data generation framework*, referred to as *SynthIRDP*, in which we construct 3D faces based on single RGB images, then render the three modalities (RGB, depth, IRDP) by simulating structured light sensor physics in the 3D rendering engine Blender[1]. *SynthIRDP* allows for simulation of constrained and unconstrained settings.
- We then proceed to *reconstruct depth maps from IRDP* with a proposed conditional generative adversarial network (C-GAN), aimed at domain mapping between distortion in the projected pattern and absolute depth. We note that proposed depth estimation approach does not require additional manual tuning of parameters, as for other standard methods as block-matching [61].
- Capitalizing on earlier mentioned benefits, we propose an *end-to-end approach for gender and age estimation based on IRDP-data*. We leverage the capabili-

[1] https://www.blender.org/

ties of an end-to-end approach, i.e., by training a network targeted for age and gender classification based on raw SL-data, omitting the classical step of depth estimation. To the best of our knowledge, this is the first end-to-end face analysis approach and related study.
- We then estimate gender and age based on (a) IRDP, (b) binarized IRDP, as well as (c) reconstructed depth map.
- While in all our experiments we utilize synthetic data, due to the lack of availability of IRDP real-life data with corresponding high quality depth maps, we finally *validate the accuracy* of our proposed end-to-end approach on a *real-life dataset*, comprising of 22 subjects.

The remaining paper is organized as follows. Section 2 revisits related work, Section 3 introduces the proposed approach for synthetic IRDP generation. Section 4 presents the depth estimation method. Section 5 demonstrates our end-to-end gender and age estimation framework. We present and discuss quantitative and qualitative experimental results in Section 6. Section 7 concludes the paper.

2 Related Work

In this section we briefly review literature related to SL-sensors, depth estimation, gender and age estimation, face analysis based on RGBD, 3D imagery, as well as GANs.

2.1 SL sensors

Structured-light (SL) sensors incorporate an infrared projector and two cameras. The projector emits a determined and fixed reference-pattern, represented by emitted infrared dots. While one camera captures the projected and distorted IR, the other captures visible light. In addition to one camera that captures the projected and distorted IR, there may be another camera foreseen, that captures visible light, as illustrated in Figure 1. The sensed pattern undergoes distortions, as well as illumination variations that are instrumental in representing the scenery. Therefore, by establishing the correspondence between reference- and sensed-pattern, the scenery can be deduced. Similar to depth estimation performed in stereo cameras, depth is derived for each dot through triangulation, where the projector acts as a second camera in a typical passive stereo camera setup.

The *PrimeSense* algorithm[2], as utilized by Kinect, performs local stereo matching, in order to estimate disparity between the difference in horizontal coordinates of corresponding image pixels (**d**), from which depth (**Z**) is estimated via $Z = \frac{bf}{d}$, where b denotes the baseline of the sensor and f refers to the focal length. We note that this approach is computationally expensive and necessitates the projector-camera system to be perfectly calibrated since the disparity is computed along the horizontal

[2] http://www.primesense.com/

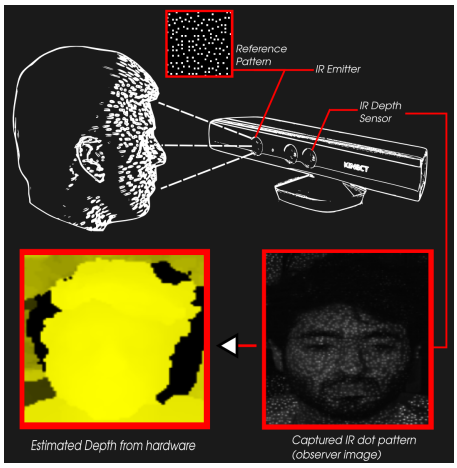

Fig. 1 IRDP sensor. A projector emits a determined and fixed reference-pattern, represented by emitted pseudo-random infrared dots. While one camera captures the projected and distorted IR, the other captures visible light (Kinect V1).

lines. More recently stereo matching for disparity estimation has been tackled using deep learning techniques, operating in a supervised and unsupervised setting on RGB and structured light data in a monocular or stereo setup [39]. While these methods show promising results, they rely on large datasets.

2.2 Depth Estimation

With the advancement of deep learning techniques, depth estimation has witnessed significant improvement over the last years. In this context a vast literature has explored a monocular setting through supervised methods [19], [34, 35, 64], as well as through unsupervised methods [20, 22] and in a stereo setting [39, 46] on RGB data. Associated performances have been evaluated on publicly available datasets such as NYU Depth V2 [44] and KITTI [21]. While RGB methods remain more economical, RGB cameras and respective datasets are ubiquitous, performing poorly in textureless regions, being less robust to noise (illumination variance) and being ill-posed for obtaining absolute depth as compared to techniques that use SL-data.

Towards tackling the computationally-expensive stereo matching problem faced by traditional algorithms using structured light data, HyperDepth [55] employed a learning-based method involving an ensemble of cascaded random forests. Despite improvement in computation time, similar methods require, as reference ground truth depth maps, usage of the PatchMatch algorithm [4], thus placing a limitation on the quality of disparity estimation. In contrast, ActiveStereoNet [65] incorporated an end-to-end network for depth estimation using active stereo systems. Due to lack of ground truth, the authors proposed a self-supervised method to predict precise depth using a two stage *Siamese* network. While one branch built a low resolution cost volume to infer an initial disparity estimate, followed by a bilinear upsampling

and a residual network to predict the final disparity map, the other branch, referred to as *Invalidation Network* was trained end-to-end to predict a confidence map. ActiveStereoNet achieved promising results, capable of handling occlusions and reducing artefacts. Ren *et al.* [49] proposed face video deblurring using 3D facial priors. Hu *et al.* [28] investigated face super-resolution guided by 3D facial priors.

2.3 Gender and Age estimation

Gender and age are demographic facial attributes (often referred to as soft biometrics [13]), which are beneficial for the associated (a) semantic meaning to humans, offering interpretation beyond that achieved by classical face recognition, (b) complementary information they offer a biometric system, (c) omitting of enrollment, i.e., a previously unseen face can be classified without it being present in the training set. Recent deep learning models [1, 12, 14, 52, 53, 58, 63] have boosted estimation accuracy significantly. Such approaches either (i) extract a generic face representation across large amounts of face data and then train shallower classifiers for attribute prediction [37] or (ii) optimize over attributes directly [54].

2.4 Face analysis based on RGBD and 3D imagery

Face analysis based on RGBD and 3D is inherently more robust to variations in pose, illumination, as well as occlusion [6, 10, 11, 17, 31]. Former three works have showcased the significant improvement of accuracy, based on deep learning models and classical learning techniques employing shape (depth) and texture (RGB) in the contexts of face recognition and classification tasks such as gender and age.

2.5 Generative Adversarial Networks (GANs)

Firstly proposed by Goodfellow [23], GANs are composed of two minmax adversaries, a *generator and discriminator*, where the $generator$ attempts to reconstruct distribution of the training data by generating realistic images, while the $discriminator$ estimates the probability that a generated image stems from the training data or is synthetic (i.e., real or fake). Conditional GANs (C-GANs) are an extension of GANs, where the model is trained to learn a conditional distribution by having both adversaries conditioned on additional information. GANs have received increased interest for various computer vision tasks such as image-image translation [66], image editing [67], representation learning [47], image inpainting [45], future prediction [41], video [62]. Pertaining to image to image translation, a nonparametric texture model was employed in the seminal work of Efros and Leung [18]. Long *et al.* [38] firstly presented the concept of image translation based on CNNs. Isola *et al.* [29] proposed flexible learning of different mappings between input and output images with a single loss function. We here note that all aforementioned works were designed for pairwise image generation/synthesis. Rosales *et al.* [51] firstly proposed unpaired image translation. Co-GAN [36] and cross-modal scene networks [2, 60] employed input

and output to share certain ”content“ features. STAR-GAN [9] represented a scalable approach for image-to-image translations for multiple domains using only a single model. Cycle-GAN [68] dealt with image-to-image translation for domain-to-target translation in the absence of paired examples. C-GAN [42] incorporated unpaired image translation, based on provided conditions.

The above discussion on the state-of-the art showcases that there is no exiting work related to IRDP to depth map estimation in the context of face analysis, which we tackle in this work. In the following section we proceed to describe proposed framework for generation of synthetic IRDP data based on RGB images.

3 Synthetic IRDP data generation

The size and quality of annotated data used to train deep learning models are critical to the associated accuracy and reliability of such models, as generally thousands to millions of parameters are trained in the process. Insufficient data can bring to the fore generalization problems. Towards overcoming cumbersome large data acquisition, recent works have proposed to train data-hungry deep learning models with synthetic data [32, 43, 50] and have proceeded to validate the reliability of such models on real-life data.

Given the lack of a face dataset based on IRDP and motivated by the above works, we firstly propose to synthetically generate IRDP data, which we depict in Figure 2 (a), wherein we emulate the physics of the structured light hardware in the Blender 3D rendering engine. Specifically, we load an existing 3D-image of a face into a Blender-scene, which contains a virtual IR camera, as well as a virtual IR projector, projecting the pseudo-random dot patterns into the scene. The rendering operator of Blender is used to render IRDP along with depth (see Figure 2 (a)-(i)). This synthetic setup allows for flexible rendering of objects at varying distance, pose and illumination. The overall pipeline of our synthetic *IRDP* data generation for RGB and 3D face dataset is illustrated in Figure 2 (j).

Existing 3D face datasets such as the Bosphorus [56] and the Face Warehouse [8] datasets contain few hundreds of subjects. Given that such size is unsuitable for training deep learning algorithms, we propose to reconstruct 3D faces from existing 2D RGB face datasets [15], as inspired by Sela *et al.* [57]. In particular we aim to reconstruct geometric structure of faces from single images.

Hereby a pixel-based geometric representation i.e., a depth and correspondence map, is learned, followed by performing geometric deformation and refinement steps to obtain a detailed textured mesh. We illustrate results of example reconstructed faces in Figure 2. Using the above method, we create a dataset of 10940 IRDP-face samples based on RGB and depth images.

We then proceed to elaborate on depth estimation based on IRDP data.

4 Depth estimation from IRDP

We tackle depth estimation as a domain-to-domain translation task, from the raw IR-domain to data-to-absolute depth-domain. To do so we train a C-GAN to learn an

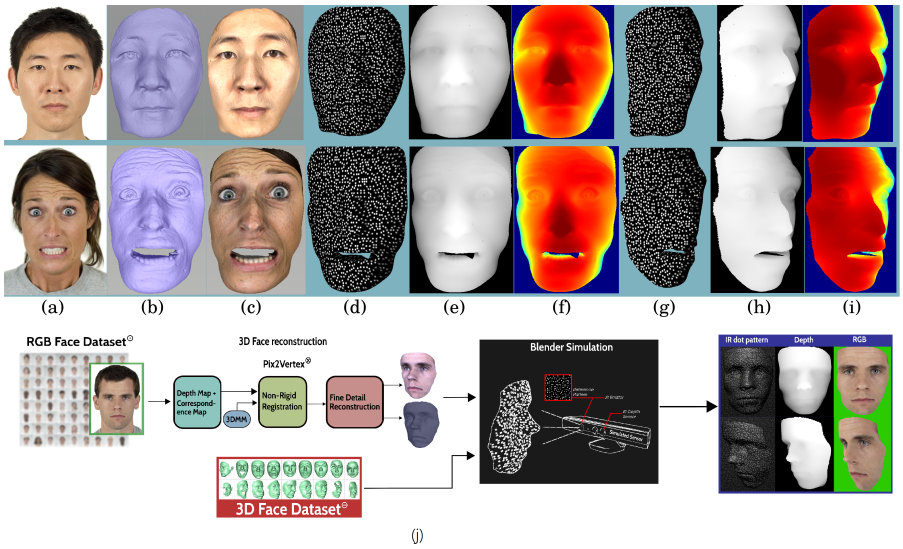

Fig. 2 Example images of *SynthIRDP*. Column (a) represents RGB images from a 2D face dataset, (b) 3D faces reconstructed by [57], (c) corresponding texture mapped faces. Column (d) to (f) show the *IRDP*, depth and grey to RGB depth respectively rendered using our *SynthIRDP* pipeline. Column (g) to (i) contain the same faces rendered as in d) to (f) for different pose and (j) Pipeline of our synthetic *IRDP* data generation for RGB and 3D face dataset. We load an existing 3D-image of a face into a Blender-scene, which contains a virtual IR camera, as well as a virtual IR projector, projecting the pseudo-random dot patterns into the scene. The rendering operator of Blender is used to render IRDP along with depth.

adaptation to change/distortions in shape of the projected pattern on the face conditioned on the additional information of depth. The goal is to learn a conditional generative model from an observed image x and a random noise vector z to an output depth y. We note that deviating from previous standard methods, the network is only provided an observed image as input, withno manual tuning.

Our primary aim is to utilize the learned adversarial loss, which bares two benefits: (i) the need for a hand-crafted loss to tackle this problem is removed and (ii) any possible structural difference between the output and target is penalized, as opposed to using conditional random forests (CRFs) or perceptual losses, as pointed out by Zhu *et al.* [68]. In addition to generator loss, we add a traditional loss, L1, in the GAN objective, in order to encourage depth estimation to be close to the ground-truth depth, as well as to assure that depth relationship between pixels is independent. The weight assigned to this L1 term depends on the λ parameter. As we are performing a one-to-one deterministic mapping, we do not provide the random noise vector z to our GAN-model.

Our $generator$ consists of a U-Net architecture, representing an encoder-decoder structure with skip connections. The skip connections are added between layer i and layer $n - i$, where n is the total number of layers. The rationale is to transfer low level information from the $encoder$ to the $decoder$, which otherwise would be lost

via the series of convolution layers up to the bottleneck layer. The $discriminator$ is a convolutional PatchGAN classifier [68].

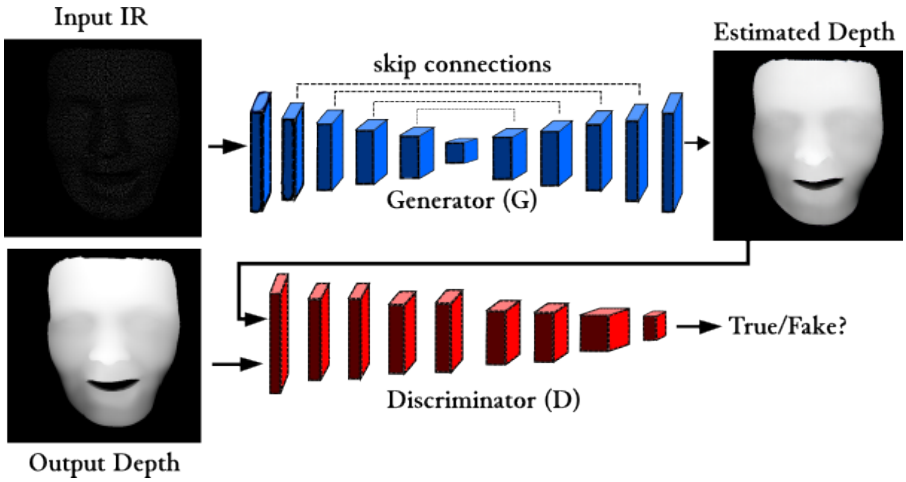

Fig. 3 (a) Depth estimation. Proposed framework for depth estimation from IRDP imagery based on C-GAN.

To optimization we train the generator G to maximize $\log D\left(x, G\left(x, z\right)\right)$ rather than minimizing $\log\left(1 - D\left(x, G\left(x, z\right)\right)\right)$. The objective is divided by 2 while optimizing D to slow down the learning rate of D relative to G. The proposed framework for depth estimation is in Figure 3(a).

5 Face Attribute Analysis

We adopt the architecture proposed by Levi and Hassner [33] for both age and gender classification. A detailed diagram of the entire network is provided in Figure 4.

The network contains three convolutional layers, each followed by a rectified linear operation and pooling layer. The first two layers also follow normalization using local response normalization. The first convolutional Layer contains 96 filters of 7×7 pixels, the second convolutional Layer contains 256 filters of 5×5 pixels, The third and final convolutional Layer contains 384 filters of 3×3 pixels. Then, two fully-connected layers are added, each containing 512 neurons. Finally, the output of the last fully connected layer is fed to a soft-max layer that assigns a probability for each class. The prediction itself corresponds to the class with maximal probability for the given test image.

We use the same architecture for both, gender and age classification. The network is trained separately on three different modalities: IRDP, depth and RGB. We tackle the binary gender classification (*2 class*: male and female), as well as age classification (*4 class*: teen, young, adult and senior) to study the ability of the network to

learn, when facing three different modalities (RGB, RGBD and IRDP). We note that to the best of our knowledge, IRDP data has not been explored before in age and gender estimation.

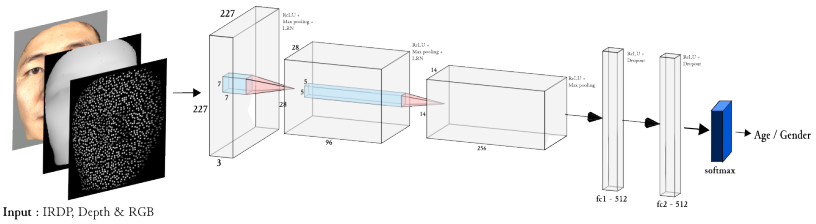

Fig. 4 Gender and Age estimation CNN architecture. The network is trained separately on three different modalities: IRDP, depth and RGB.

6 Experimental Results

In this section we proceed to describe experimental results and related discussion on (i) proposed depth estimation strategy from IRDP imagery based on C-GAN (ii) proposed face attribute analysis based on IRDP and depth imagery, as well as (iii) merging the performance gap between real and synthetic reconstructed depth data.

6.1 Depth estimation and IRDP image synthesis

6.1.1 Training procedure

As proposed by Zhu et al. [68], we use minibatch stochastic gradient descent (SGD) and apply ADAM optimization [30], with a learning rate of 0.0002, and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. We initialize $\lambda = 100$. Towards gasping the ability of GANs to learn depth estimation from a single dot pattern image, we train and test the network over a diverse dataset in pose, distance and subjects. We do so with two modalities, (i) dot pattern image which contains both the distortion and illumination variation information of the projected patterns and (ii) binarized dot-pattern image, which contains only the distortion variation of the projected patterns. We do so in order to assess the importance of each information provided to the GAN for the task of depth estimation and analyze related robustness to unseen pose, illumination and distance of the subject from the sensor. Loss functions pertaining to C-GAN for depth estimation employing (a) IRDP images and (b) binarized-IRDP images, referred to as B-IRDP, are shown in Figure 5. Specifically, we observe the loss cure of C-GAN, while learning the depth estimation from IRDP versus binary IRDP. We conclude from the pattern in the loss curve that while employing IRDP, the GAN model was able to learn better.

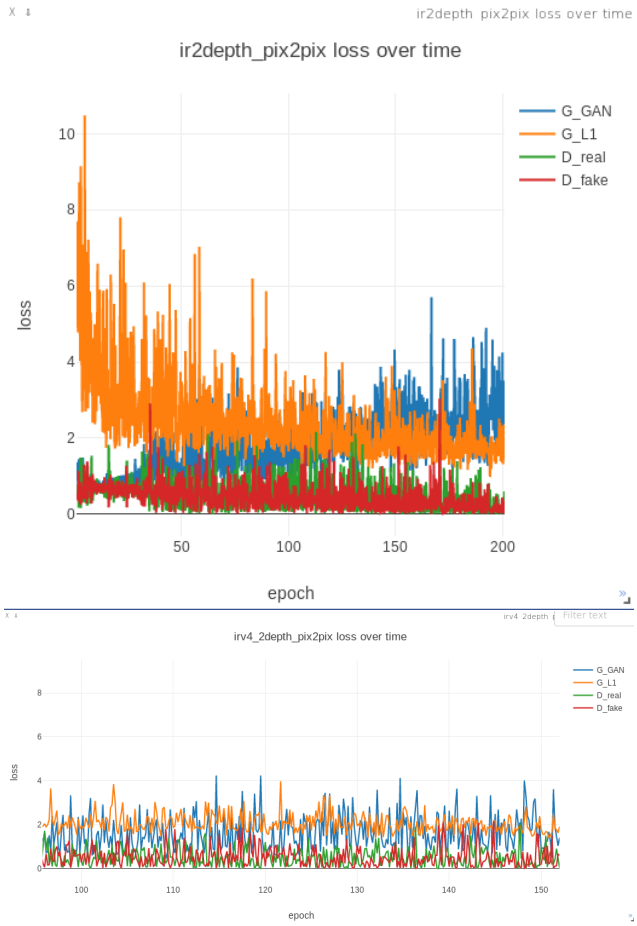

Fig. 5 Depth estimation. Loss curve of C-GAN in estimating depth employing (a) IRDP images and (b) binarized-IRDP images.

The size of the training set varies from 1000 - 10,000 images and the size of the test set ranges from 300 - 2,000 images for 500 subjects. The model is trained with input image resolution of 512 $\times$ 512 on *4* $\times$ *GTX 1080 Ti* GPU for 100 epochs.

6.1.2 Qualitative results

Figure 6 demonstrates network's ability to perform depth estimation from a single IRDP image. While when trained on only frontal poses, surprisingly the model is able to estimate depth of unseen poses (Figure 6 (c)), it is unable to estimate finer depth details at high image gradients. We note that this is expected from supervised learning techniques but can be solved by introducing random poses into the training set (see row 3 of (c) and row 3 of (e) in Figure 6).

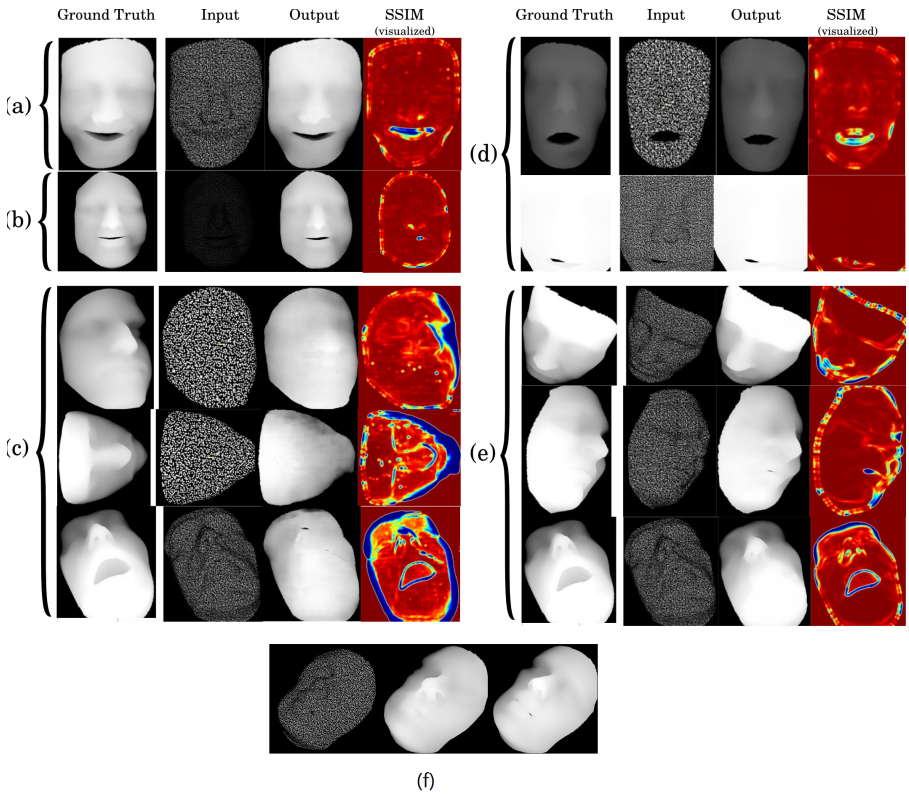

Fig. 6 Qualitative results of the experiments of depth estimation and generated hallucination. **(a)** $\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_F;test:\mathcal{D}_F}$, **(b)** $\mathcal{E}$:IRDP$_{train:\mathcal{D}_F;test:\mathcal{D}_F}$, **(c)** $\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_F;test:\mathcal{D}_P}$, **(d)** $\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_{PD};test:\mathcal{D}_{PD}}$ **(e)** $\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_P;test:\mathcal{D}_P}$. **(f)** Generator hallucinating facial features during training to fool discriminator.

Towards further analysis, in Figure 6(f) we show the hallucinating or unknown artifact generated by the model. Possible reason might be that, in order to minimize the loss function, in some cases the generator generates depth structures of the noise that is not in accordance with the ground truth depth. This might be due to the generator not merely learning to perform mapping of IRDP distortions to depth, but also learning some facial features, where it has found a loophole to fool the discriminator by hallucinating noises. This only occurs in early stages of the training.

The used metrics are enlisted hereafter.

– **Average Relative Error (rel)**. The relative absolute error is the absolute error divided by the magnitude of the exact value. The percent error is the relative error expressed in terms of per 100.

$$rel = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{G(x)_i - y_i}{y}\right|,$$

where $G(x)_i$ is a pixel in the generated depth image $G(x)$, y_i is a pixel in the ground truth depth image y, n is the total number of pixels for each depth image.

– **Root Mean Square Error (RMSE)** represents the standard deviation of the residuals prediction errors. Residuals are a measure of how far data points are from the regression line. Hence RMSE is a measure of how spread out these residuals are.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(G(x)_i - y_i\right)^2}$$

– **Average log_{10} error**. To introduce the percentual difference an MSE logarithm scale is used. It makes MSE to only focus on the relative difference between the true and the predicted value.

$$\log_{10}\text{ error} = \frac{1}{n}\sum_{i=1}^{n}\left|\log_{10}\left(G(x)_i\right) - \log_{10}\left(y_i\right)\right|$$

– **Threshold Accuracy (δ_n)** captures the percentage of match in a image with respect to a reference image for a given threshold following following equation.

$$\delta_n = \%\ \text{ of }\ \mathrm{y}_i \ni \max\left(\frac{y_i}{G(x)_i}, \frac{G(x)_i}{y_i}\right) = \delta_n$$
$$\text{, where } \delta_n < 1.25^n \text{ for n=1,2,3}$$

– **Structural Similarity Index (SSIM)** indicates the image quality based on an initial image as reference.

$$SSIM = \frac{1}{N}\sum_{t=1}^{N} SSIM\left(G(x)_i, y_i,\right),$$

where N is the total number of images in the test set.

To explore the robustness of our model w.r.t. pose variation, orientation and distance, we test the performance in following settings.

1. **Frontal IRDP** *[$\mathcal{E}$:IRDP$_{train:\mathcal{D}_F;test:\mathcal{D}_F}$]*: Trained and tested on a dataset ($\mathcal{D}_F$) of IRDP containing only frontal pose of faces with a fixed distance from the camera and projector.
2. **Frontal B-IRDP** *[$\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_F;test:\mathcal{D}_F}$]*: Trained and tested on a dataset ($\mathcal{D}_F$) of B-IRDP containing only frontal pose of faces with a fixed distance from the camera and projector.
3. **Frontal and varied pose IRDP** *[$\mathcal{E}$:IRDP$_{train:\mathcal{D}_F;test:\mathcal{D}_P}$]*: Trained on a dataset ($\mathcal{D}_F$) of IRDP containing only frontal pose of faces with a fixed distance from the camera and projector while tested on a dataset ($\mathcal{D}_P$) of IRDP containing varied pose and orientation of faces.
4. **Frontal and varied pose B-IRDP** *[$\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_F;test:\mathcal{D}_P}$]*: Trained on a dataset ($\mathcal{D}_F$) of B-IRDP containing only frontal pose of faces with a fixed distance from the camera and projector while tested on a dataset ($\mathcal{D}_P$) of B-IRDP containing varied pose and orientation of faces.

5. **Varied pose IRDP** *[$\mathcal{E}$:IRDP$_{train:\mathcal{D}_P;test:\mathcal{D}_P}$]*: Trained and tested on a dataset ($\mathcal{D}_P$) of IRDP containing varied pose and orientation of faces with a fixed distance from the camera and projector.
6. **Varied pose B-IRDP** *[$\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_P;test:\mathcal{D}_P}$]*: Trained and tested on a dataset ($\mathcal{D}_P$) of B-IRDP containing varied pose and orientation of faces with a fixed distance from the camera and projector.
7. **Varied pose and distance IRDP** *[$\mathcal{E}$:IRDP$_{train:\mathcal{D}_{PD};test:\mathcal{D}_{PD}}$]*: Trained and tested on a dataset ($\mathcal{D}_{PD}$) of IRDP containing faces with varied pose, orientation and distance from the camera and projector.
8. **Varied pose and distance B-IRDP** *[$\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_{PD};test:\mathcal{D}_{PD}}$]*: Trained and tested on a dataset ($\mathcal{D}_{PD}$) of B-IRDP containing faces with varied pose, orientation and distance from the camera and projector.

Table 1 summarizes the quantitative results of the experiments, where depth estimation from IRDP images using the evaluation metrics are reported. We conclude based on all metrics that the IRDP outperforms the other modalities.

Table 1 Quantitative results of the depth estimation.

Experiment	Error metrics ↓			Accuracy metrics ↑			
	rel	rms	$\log_{10}$	δ_1	δ_2	δ_3	SSIM
$\mathcal{E}$:IRDP$_{train:\mathcal{D}_F;test:\mathcal{D}_F}$	**0.0827**	**0.0303**	**0.0177**	**0.9808**	**0.9846**	0.9871	**0.9821**
$\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_F;test:\mathcal{D}_F}$	0.0844	0.0323	0.0180	0.9780	0.9824	0.9866	0.9792
$\mathcal{E}$:IRDP$_{train:\mathcal{D}_F;test:\mathcal{D}_P}$	1.2834	0.1816	0.1683	0.8647	0.8939	0.9015	0.7961
$\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_F;test:\mathcal{D}_P}$	1.4811	0.2177	0.1884	0.7351	0.7500	0.7581	0.7705
$\mathcal{E}$:IRDP$_{train:\mathcal{D}_P;test:\mathcal{D}_P}$	0.1311	0.03422	0.0183	0.9806	0.9870	**0.9894**	0.9758
$\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_P;test:\mathcal{D}_P}$	0.1403	0.03612	0.0201	0.9710	0.9789	0.9807	0.9666
$\mathcal{E}$:IRDP$_{train:\mathcal{D}_{PD};test:\mathcal{D}_{PD}}$	0.6314	0.0602	0.0338	0.9661	0.9765	0.9797	0.9470
$\mathcal{E}$:B-IRDP$_{train:\mathcal{D}_{PD};test:\mathcal{D}_{PD}}$	0.6631	0.0896	0.0492	0.9366	0.9409	0.9473	0.9157

In addition, we demonstrate the robustness of IRDP for depth estimation as opposed to RGB data in Table 2. With exception of the extreme lighting scenario, the performance of IRDP is better for depth estimation as compared to RGB. The reason for such poor performance of IRDP in extreme lighting condition has to do with the physics underlying its working principal, where the heavy exposition to light affects the projected pattern sensing. Further we proceed to illustrate the result with varying pose and height. Cross pose i.e., training with frontal pose and testing with other poses results in a decrease for most evaluation metrics. However, the result is significantly better, when trained and tested with varying poses. In addition, performance decreases for varying distance. One possible reason might be that the synthetic simulations, which do not mimic exactly the behaviour of real conditions. More work should be done in future to fill the gap between simulation and real sensors.

6.2 Face attribute analysis

6.2.1 Experimental details

We train the above network on our synthetic dataset created using *SynthIRDP*; train set of size 7440, validation set of size 1500 and test set of size 2000 images for each

Table 2 RGB and IRDP for depth estimation in varying illumination condition. These experiments involve altering the external environment lighting in the scene to test the robustness of each modality. Results indicate that IRDP is more robust as opposed to RGB.

Experiment	Error metrics ↓			Accuracy metrics ↑			
	rel	rms	$\log_{10}$	δ_1	δ_2	δ_3	SSIM
Ideal Conditions : IRDP	**0.0827**	**0.0303**	**0.0177**	**0.9808**	**0.9846**	**0.9871**	**0.9821**
Ideal Conditions : RGB	0.2030	0.04317	0.0238	0.9754	0.9840	0.9855	0.9772
Poor lighting : IRDP	**0.4284**	**0.1818**	**0.1044**	**0.9613**	**0.9775**	**0.9840**	**0.9698**
Poor lighting : RGB	5.0828	1.525	1.118	0.8100	0.8266	0.8437	0.8588
Extreme lighting : IRDP	6.0844	2.1323	1.5180	0.8080	0.8124	0.8166	0.8098
Extreme lighting : RGB	**5.1403**	**1.7612**	**1.1201**	**0.8310**	**0.8389**	**0.8407**	**0.8366**

modality: IRDP, RGB, depth. The train, validation and test set includes 870 subjects, 450 of whom are female and 420 male across four age groups: teen, young, adult and senior. We have utilized same splits employed for age and gender classification, in train, validation and test set pertaining to different poses and illuminations for all three modalities. Pose and illumination distributions are random with no duplicates present in either train, validation or test set. We also train the model on a binarized IRDP, which we refer to as B-IRDP, in order to visualize how the network learns, when provided with only distortion variation in structured light dot pattern. We note that B-IRDP only contains shape variations unlike IRDP, which also contains illumination variations.

Input images are firstly re-scaled and aligned to 256 $\times$ 256 and then fed to the network, as illustrated in Figure 4. We set *eta* as 0.001, *batch size* as 32 and *dropout* as 0.5 and train the network for 20,000 *iterations*. The same models are used later to evaluate the captured real-life images.

6.2.2 Evaluation on synthetic dataset

We estimate the performance of proposed gender and age estimation approaches with respect to *precision p*, *recall r*, and $F1$ score. We note that for each class (male, female; age-categories), precision is defined as the number of correctly predicted cases divided by the number of all predictions of this class. Recall denotes the number of correctly predicted cases divided by the number of all cases of this class. F1 is the harmonic mean of precision and recall

$$F1 = \frac{2pr}{p+r}.$$

Table 3 reports the F1 scores of gender and age classification on all modalities. The network tends to perform similary on depth data and IRDP for both, age and gender classification with and without pre-training. One possible reason to why using a pre-trained network, trained on ImageNet [16] does not favour an improvement in IRDP as opposed to depth, is due to the structural (dis-)similarity present in the imagery. In comparison to RGB the network tends to perform similar on depth data and on IRDP for gender classification with and without pre-training. Whereas, while

the network is trained on ImageNet [16], the result is better for RGB for gender classification. A possible reason is due to the pre-training network being trained on RGB images. For age classification the network tends to perform better for RGB in comparison to depth data and on IRDP with and without pre-training. B-IRDP is not as effective in this scenarios, as expected.

Table 3 Results on synthetic dataset *SynthIRDP*. * indicates that the network was initialized with a pre-trained inception v3 checkpoint trained on ImageNet [16].

Input	F1 Score (Gender)	F1 Score (Age)
RGB	0.956	**0.875**
reconstructed depth	**0.957**	0.822
IRDP	0.954	0.804
B-IRDP	0.912	0.761
RGB^*	**0.986**	**0.880**
$reconstructeddepth^*$	0.961	0.848
$IRDP^*$	0.936	0.834
$B-IRDP^*$	0.919	0.780

Figure 7 portrays the loss on training the network on our synthetic dataset generated using our *SynthIRDP* pipeline for age and gender classification. We can observe that the objective function converges slower for IRDP than for depth i.e., the network takes more iterations and thus longer training time. For these reasons, as well as the better performance on depth than IRDP stated in Table 3, we conclude that the network proposed in Figure 4 learns better with depth data than IRDP.

6.2.3 Evaluation on real-life dataset

To evaluate the model performance on a real-life dataset, we acquire around 1500 images of 22 subjects, 8 female and 14 male, capturing a number of poses and distances between subject and sensor. The sensor used was the Asus Xtion Pro Live, which shares the same pseudo-random dot pattern structure as Kinect V1. Figure 8 shows example images acquired from the sensor. While the sensors also provide an estimated depth map, such maps are of low resolution and lack subpixel precision. Thus they fail to capture detailed structures like faces.

We note that in all our experiments, where we test with real data, we train the models on only synthetic data due to the lack of availability of real-life data with high quality depth maps. Related results are summarized in Table 4.

These are preliminary results, which would need validation on a larger real-life dataset containing both depth and IRDP and using other deep learning architectures like the ones proposed by Cui et al. [11]. A deeper analysis suggests that IRDP has promising properties, especially for embedded sensors (such as Internet-of-Things devices or smartphones), where computational cost of reconstructing explicitly the depth-map is not negligible.

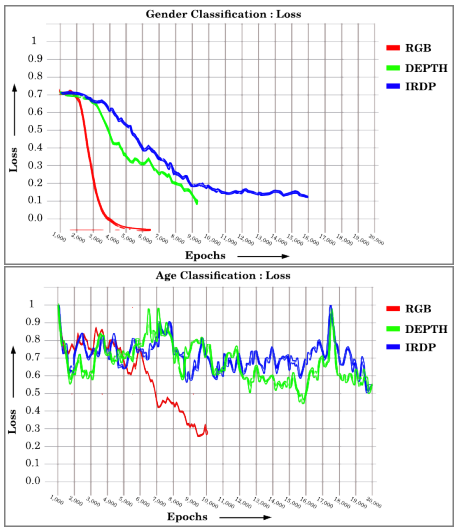

Fig. 7 Gender and age estimation. Loss curve of the network on the synthetic dataset for the task of gender and age classification.

Fig. 8 Real-life data. Example images of the real-life IRDP-dataset, which we acquired using Asus Xtion Pro Live sensor.

7 Conclusions and future work

In this work we explored the use of structured light in face attribute analysis. Towards this, we firstly generated a synthetic face dataset based on RGB and RGB-D datasets. We then proposed a C-GAN for depth map reconstruction architecture. We then compared facial images pertaining to RGB, reconstructed depth maps, infra-red dot pattern (IRDP), as well as binarized IRDP-imagery in estimating gender and age.

Table 4 Results pertaining to the real-life dataset, * indicates that the network was initialized with a pre-trained inception v3 checkpoint trained on ImageNet [16].

Input	F1 Score (Gender)	F1 Score (Age)
RGB	0.831	0.712
Depth	**0.855**	**0.731**
IRDP	0.847	0.713
B-IRDP	0.814	0.691
RGB^*	**0.870**	**0.761**
$Depth^*$	0.863	0.741
$IRDP^*$	0.851	0.723
$B-IRDP^*$	0.822	0.715

Presented experimental results revealed the ability of gender and age to be gleaned directly from IRDP-images, as well as from reconstructed depth maps.

We note that these are early results, however our findings clearly indicate the feasibility of IRDP-based approaches for facial analysis tasks.

Future works involve mitigating the gap between synthetic and real data, to better generalize in unseen realistic conditions. In addition, an end-to-end approach has the potential to temper the impact of the calibration, which can be learned during training. Other possible future work directions include the analysis of proposed methods under computational and memory constraints.

Declarations

Vikas Thamizharasan has received a research funding from the Company Blu Manta for this work.

References

1. Abate, A.F., Barra, P., Barra, S., Molinari, C., Nappi, M., Narducci, F.: Clustering facial attributes: Narrowing the path from soft to hard biometrics. IEEE Access (2019)
2. Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., Torralba, A.: Cross-modal scene networks. IEEE transactions on pattern analysis and machine intelligence (2017)
3. B. Freedman A. Shpunt, M.M., Arieli., Y.: Depth mapping using projected patterns. US Patent 8,150,142 (2012)
4. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**, 24:1–24:11 (2009)
5. Bleyer, M., Breiteneder., C.: Stereo matching - state-of-the-art and research challenges. In Advanced Topics in Computer Vision, Springer pp. 143–179 (2013)
6. Boutellaa, E.: Contribution to face analysis from RGB images and depth maps (2017)
7. Cai, Y., Lei, Y., Yang, M., You, Z., Shan, S.: A fast and robust 3d face recognition approach based on deeply learned face representation. Neurocomputing **363**, 375–397 (2019)
8. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics **20**, 413–425 (2014)
9. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. arXiv preprint arXiv:1711.09020 (2017)

10. Chowdhury, A., Ghosh, S., Singh, R., Vatsa, M.: RGB-D face recognition via learning-based reconstruction. 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS) pp. 1–7 (2016)
11. Cui, J., Zhang, H., Han, H., Shan, S., Chen, X.: Improving 2d face recognition via discriminative face depth estimation. 2018 International Conference on Biometrics (ICB) pp. 140–147 (2018)
12. Dantcheva, A., Brémond, F.: Gender estimation based on smile-dynamics. IEEE Transactions on Information Forensics and Security **12**(3), 719–729 (2017)
13. Dantcheva, A., Elia, P., Ross, A.: What else does your biometric data reveal? a survey on soft biometrics. IEEE Transactions on Information Forensics and Security **11**(3), 441–467 (2016)
14. Das, A., Dantcheva, A., Bremond, F.: Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In: ECCVW 2018-European Conference of Computer Vision Workshops (2018)
15. DeBruine, L., Jones, B.: Face Research Lab London Set (2017). DOI 10.6084/m9.figshare.5047666.v3. URL `https://figshare.com/articles/Face_Research_Lab_London_Set/5047666`
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
17. Drira, H., Amor, B.B., Srivastava, A., Daoudi, M., Slama, R.: 3D face recognition under expressions, occlusions, and pose variations. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**, 2270–2283 (2013)
18. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 2, pp. 1033–1038. IEEE (1999)
19. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS (2014)
20. Garg, R., Kumar, B.V., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European Conference on Computer Vision, pp. 740–756. Springer (2016)
21. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. 2012 IEEE Conference on Computer Vision and Pattern Recognition pp. 3354–3361 (2012)
22. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
23. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680 (2014)
24. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. ECCV (2014)
25. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Aligning 3D models to RGB-D images of cluttered scenes. CVPR (2015)
26. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture. ACCV (2016)
27. He, Y., Chiu, W., Keuper, M., Fritz, M.: RGB-D semantic segmentation using spatio-temporal data-driven pooling. CVPR (2017)
28. Hu, X., Ren, W., LaMaster, J., Cao, X., Li, X., Li, Z., Menze, B., Liu, W.: Face super-resolution guided by 3d facial priors. In: European Conference on Computer Vision, pp. 763–780. Springer (2020)
29. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint (2017)
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
31. Kittler, J., Koppen, P., Kopp, P., Huber, P., Rätsch, M.: Conformal mapping of a 3d face representation onto a 2d image for cnn based face recognition. 2018 International Conference on Biometrics (ICB) pp. 124–131 (2018)
32. Kortylewski, A., Schneider, A., Gerig, T., Egger, B., Morel-Forster, A., Vetter, T.: Training deep face recognition systems with synthetic data. CoRR **abs/1802.05891** (2018)
33. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops (2015)
34. Li, J., Klein, R., Yao, A.: A two-streamed network for estimating fine-scaled depth maps from single RGB images. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 3392–3400 (2017)

35. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5162–5170 (2015)
36. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: Advances in neural information processing systems, pp. 469–477 (2016)
37. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
38. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440 (2015)
39. Luo, W., Schwing, A.G., Urtasun, R.: Efficient deep learning for stereo matching. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5695–5703 (2016)
40. M. Hansard S. Lee, O.C., Horaud., R.P.: Time-of-flight cameras: principles, methods and applications. Springer Science Business Media (2012)
41. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440 (2015)
42. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
43. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: Ganerated hands for real-time 3d hand tracking from monocular RGB. In: Proceedings of Computer Vision and Pattern Recognition (CVPR) (2018). URL `https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/`
44. Nathan Silberman Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV (2012)
45. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)
46. Pilzer, A., Xu, D., Puscas, M.M., Ricci, E., Sebe, N.: Unsupervised adversarial depth estimation using cycled generative networks. 2018 International Conference on 3D Vision (3DV) pp. 587–595 (2018)
47. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
48. Ratyal, N., Taj, I.A., Sajid, M., Mahmood, A., Razzaq, S., Dar, S.H., Ali, N., Usman, M., Baig, M.J.A., Mussadiq, U.: Deeply learned pose invariant image analysis with applications in 3d face recognition. Mathematical Problems in Engineering **2019** (2019)
49. Ren, W., Yang, J., Deng, S., Wipf, D., Cao, X., Tong, X.: Face video deblurring using 3d facial priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9388–9397 (2019)
50. Richardson, E., Sela, M., Kimmel, R.: 3d face reconstruction by learning from synthetic data. 2016 Fourth International Conference on 3D Vision (3DV) pp. 460–469 (2016)
51. Rosales, R., Achan, K., Frey, B.J.: Unsupervised image translation. In: iccv, pp. 472–478 (2003)
52. Rose, J., Bourlai, T.: Deep learning based estimation of facial attributes on challenging mobile phone face datasets. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1120–1127 (2019)
53. Rozsa, A., Günther, M., Rudd, E.M., Boult, T.E.: Facial attributes: Accuracy and adversarial robustness. Pattern Recognition Letters **124**, 100–108 (2019)
54. Rudd, E.M., Gunther, M., Boult, T.E.: Moon: A mixed objective optimization network for the recognition of facial attributes. In: Proceedings of the European Conference on Computer Vision (2016)
55. Ryan Fanello, S., Rhemann, C., Tankovich, V., Kowdle, A., Orts Escolano, S., Kim, D., Izadi, S.: Hyperdepth: Learning depth from structured light without matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5441–5450 (2016)
56. Savran, A., Alyüz, N., Dibeklioglu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3d face analysis. In: BIOID (2008)
57. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. arxiv (2017)
58. Shekhawat, H.S., Rathor, H.S.: Impacts of change in facial features on age estimation and face identification: A review. In: A.K. Somani, R.S. Shekhawat, A. Mundra, S. Srivastava, V.K. Verma (eds.) Smart Systems and IoT: Innovations in Computing, pp. 801–812. Springer Singapore, Singapore (2020)
59. Socher, R., Huval, B., Bhat, B., Manning, C.D., Ng, A.Y.: Convolutional-recursive deep learning for 3d object classification. NIPS (2012)

60. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. arXiv preprint arXiv:1611.02200 (2016)
61. Tornow, M., Grasshoff, M., Nguyen, N., Al-Hamadi, A., Michaelis, B.: Fast computation of dense and reliable depth maps from stereo images. In: F. Solari, M. Chessa, S.P. Sabatini (eds.) Machine Vision, chap. 3. IntechOpen, Rijeka (2012). DOI 10.5772/34976. URL https://doi.org/10.5772/34976
62. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Advances In Neural Information Processing Systems, pp. 613–621 (2016)
63. Xie, J.C., Pun, C.M.: Chronological age estimation under the guidance of age-related facial attributes. IEEE Transactions on Information Forensics and Security **14**(9), 2500–2511 (2019)
64. Xu, D., Wang, W., Tang, H., Liu, H.W., Sebe, N., Ricci, E.: Structured attention guided convolutional neural fields for monocular depth estimation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 3917–3925 (2018)
65. Zhang, Y., Khamis, S., Rhemann, C., Valentin, J., Kowdle, A., Tankovich, V., Schoenberg, M., Izadi, S., Funkhouser, T., Fanello, S.: Activestereonet: End-to-end self-supervised learning for active stereo systems. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 784–801 (2018)
66. Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial network. arXiv preprint arXiv:1609.03126 (2016)
67. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: European Conference on Computer Vision, pp. 597–613. Springer (2016)
68. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks