Face-from-Depth for Head Pose Estimation on Depth Images

Face-from-Depth for Head Pose Estimation on Depth Images

Guido Borghi, Matteo Fabbri, Roberto Vezzani, Simone Calderara and Rita Cucchiara

Abstract—Depth cameras allow to set up reliable solutions for people monitoring and behavior understanding, especially when unstable or poor illumination conditions make unusable common RGB sensors. Therefore, we propose a complete framework for the estimation of the head and shoulder pose based on depth images only. A head detection and localization module is also included, in order to develop a complete end-to-end system. The core element of the framework is a Convolutional Neural Network, called *POSEidon*$^+$, that receives as input three types of images and provides the 3D angles of the pose as output. Moreover, a *Face-from-Depth* component based on a *Deterministic Conditional GAN* model is able to hallucinate a face from the corresponding depth image. We empirically demonstrate that this positively impacts the system performances. We test the proposed framework on two public datasets, namely *Biwi Kinect Head Pose* and *ICT-3DHP*, and on *Pandora*, a new challenging dataset mainly inspired by the automotive setup. Experimental results show that our method overcomes several recent state-of-art works based on both intensity and depth input data, running in real-time at more than 30 frames per second.

Index Terms—Head Pose Estimation, Shoulder Pose Estimation, Automotive, Deterministic Conditional GAN, CNNs.

1 Introduction

Computer vision has been addressing the problem of head pose estimation for several years.
In 2009, Murphy-Chutorian and Trivedi [1] made a first assessment of the proposed techniques. More recently, different approaches have been proposed together with some annotated datasets useful for both training and testing those systems. The interest of the research community is mainly due to a large number of applications that require or are improved by a reliable head pose estimation: face recognition with aliveness detection, human-computer interaction, people behavior understanding are some examples. Moreover, a large effort has been recently devoted to applications in the automotive field, such as monitoring drivers and passengers. Together with the estimation of the upper-body and shoulder pose, the head monitoring is one of the key technologies required to set up (semi)-autonomous driving cars, human-car interaction for entertainment, and driver's attention measurement.

In the automotive field, vision-based systems are required to cooperate or even replace other traditional sensors, due to the increasing presence of cameras inside new car's cockpits and to the ease of capturing images and videos in a completely non-invasive manner.
In the past, encouraging results for driver head pose estimation have been achieved using RGB images [1], [2], [3], [4], [5] as well as different camera types, such as infrared [6], thermal [7], or depth [8], [9], [10]. Among them, the last ones are very promising, since they allow robustness when facing strong illumination variations. Moreover, standard techniques based on RGB images are not always feasible due to poor or absent illumination conditions during the night or to the continuous illumination changes during the day.

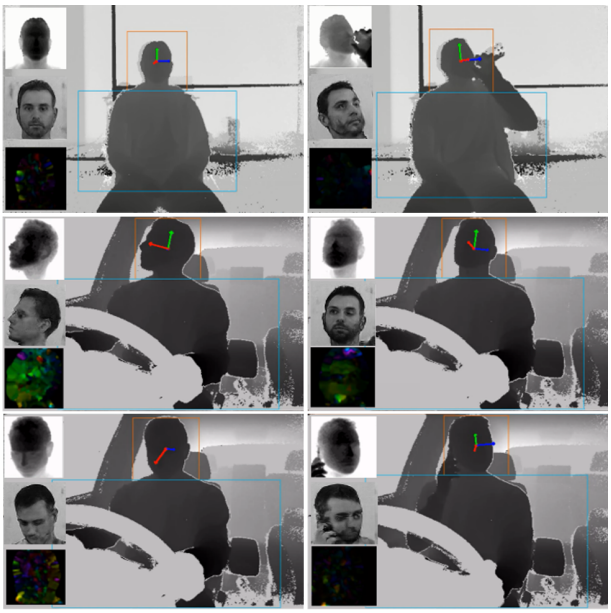


Fig. 1. Visual examples of the proposed framework output in indoor (first row) and automotive (second and third row) settings. Head pose angles are reported as colored arrows. Depth maps, *Face-from-Depth* and Motion Image inputs are depicted on the left of each frame.

Nowadays, the acquisition of depth data is feasible thanks to commercial low-cost, high-quality and small-sized depth sensors, that can be easily placed inside the vehicle.

In this paper, we propose a robust and fast solution for head and shoulder pose estimation, especially devoted to drivers in cars, but that can be easily generalized to any application where depth images are available. The presented framework provides impressive results, reaching

• *The authors are with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Italy.*
E-mail: name.surname@unimore.it

Fig. 2. Example of reliability of the *FfD* network on depth images. Two consecutive frames have been selected from a sequence with an abrupt illumination change (from light to dark). In the first column the auto equalized RGB, then the corresponding depth maps and finally the *FfD* reconstruction output.

an accuracy higher than 73% on the new *Pandora* dataset (see Fig. 3) and a low average error on the *Biwi* dataset, thus overcoming all state-of-art related works.

The core of the framework is a Convolutional Neural Network (CNN), called *POSEidon*$^+$, that combines depth, appearance and Motion Images as input to estimate the 3D pose angles in regression. An overview of the model is depicted in Figure 4. The model is enhanced with a *Face-from-Depth (FfD)* component.
This is motivated by recent literature results [11], [12] that testifies the importance of intensity images for the task. The *FfD* component is able to reconstruct the gray-level appearance of a face directly from the corresponding depth image. Thanks to the insensitivity of the depth image to the external illumination conditions, the provided reconstruction is more stable and reliable than gray or color images captured from the same RGB-D sensor. Moreover, the reconstruction can be applied in situations where the depth sensor is exploited alone without the color stream for computational or implementation constraints.
As an example, in Figure 2, we have reported two frames captured from an RGB-D sensor in correspondence of an abrupt illumination change (from light to dark). The depth images are not affected by the illumination change and thus the corresponding *FfD* reconstructions are identical. The provided output highlights the reliability of the developed network as well as the quality of the results.

The overall system is split into two components: the *Face-from-Depth* architecture followed by the pose estimation module, that takes as input the reconstructed gray level images. From a first glance, this approach could be improper since we are somehow forcing the *FfD* model to output a human understandable intermediate representation, *i.e.*, the gray level image. Training an end-to-end system enables the network to find the best internal/intermediate representation. However, in addition to a performance improvement as reported in Section 6, the introduction of the *Face-from-Depth* component allows the second part of the system to be trained on wider datasets since more annotated datasets on gray-level images are usually available rather than on depth ones. More generally, *FfD* moves input depth images on a domain where more experience is available in order to understand and process them.

This paper is an improved and extended version of our preliminary work, that has been described in [10], where the body pose estimation task was carried on through a baseline version of the *POSEidon*$^+$ framework, here referred as *POSEidon*. In this paper we present the overall framework, introducing a new *Face-from-Depth* architecture, which exploits the recent *Deterministic Conditional GAN* models [13] to reconstruct gray-level face images. To the best of our knowledge, this is one of the first proposal to generate intensity images from depth data for the head pose estimation task with an *adversarial* approach. Moreover, we evaluate and check the overall quality of the computed face images and results confirm their high quality and accuracy.
Extensive experiments have been carried out and results show that the *POSEidon*$^+$, equipped with the improved version of the *Face-from-Depth* architecture, achieves significant improvements in the head pose estimation task. Besides, we show that is possible to obtain competitive results exploiting a CNN trained on gray-level faces and tested on generated ones.

2 Related Work

The complete framework proposed in this paper merges together several modern aspects of computer vision. Among the others, the detection, localization, and pose estimation of the head and the shoulders on depth images have been included. In the following, we describe the state of the art of each mentioned topic, including the *Domain Translation* research area related to the *Face-from-Depth* module.

Head Pose. Head pose estimation approaches can rely on different input types: RGB images, depth maps, or both. For this reason, in order to discuss related works, we adopt a classification based on the input data types leveraged by each method.

RGB methods take monocular or stereo intensity images as input. In [14] a discriminative approach to frame-by-frame tracking the head pose is presented, based on the detection of the centers of both eyes, the tip of the nose and the center of the mouth. Also, [15], [16], [17] leverage well visible facial features on RGB input images, and [18] on 3D data. [19] proposed to predict pose parameters from high-dimensional feature vectors, embedding a Gaussian mixture of linear inverse-regression model into a dynamic Bayesian model. However, these methods need facial (*e.g.* nose and eyes) or pose-dependent features, that should be always visible: consequently, these methods fail when such features are not detected.
A different approach for head pose estimation involves 3D model registration techniques. Firstly, Blanz and Vetter [20] propose a technique for modeling textured 3D faces automatically generated from one or more photographs. Cao *et al*. [21] exploited a 3D regression algorithm that learns an accurate, user-specific face alignment model from an easily acquired set of training data, generated from images of the user performing a sequence of predefined facial poses and expressions. Furthermore, [22] proposed a hybrid approach, which exploits the flexibility of a generative 3D facial model in a combination with a fitting algorithm. However, those

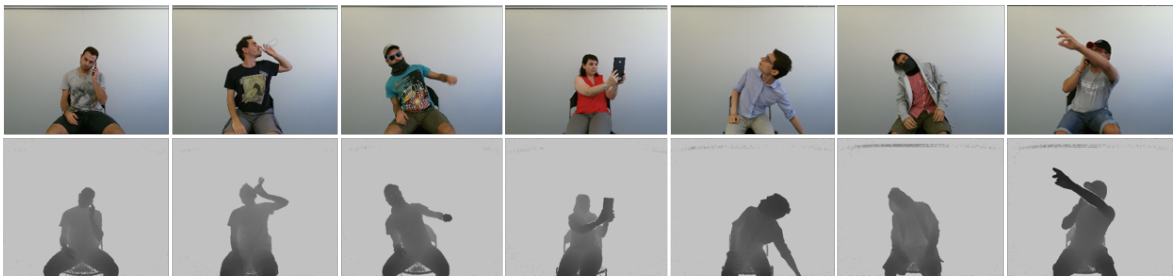


Fig. 3. Sample frames from the *Pandora* dataset. As depicted, extreme poses and challenging camouflage can be present.

techniques often need a manual initialization which is indeed critical for the effectiveness of the method.
A first attempt to use deep learning techniques combined with the regression task in head the pose estimation problem has been performed by Ahn *et al*. [11], through a CNN trained on RGB input images. Also, [23] exploits a CNN by mapping images of faces on a low dimensional manifold parameterized by pose. In [24] a framework to jointly estimate the head pose and the face alignment using global and local CNN features has been presented while a hybrid approach based on CNN and Gaussian mixture was proposed in [25] and [26]. With deep learning-based approaches, synthetic datasets were often used to train CNNs, that generally require a huge amount of data [27].
Additionally, a bunch of methods regard head pose estimation as an *optimization* problem: in [28] a multi-template, *Iterative Closest Point* (ICP) [29] based gaze tracking system is introduced. Besides, other works use linear or nonlinear regression with extremely low-resolution images [30]. HOG features and a Gaussian locally-linear mapping model were used in [12] and, finally, recent works produce head pose estimations performing a face alignment task [31] using CNNs.
In general, RGB based methods are highly sensitive to illumination, partial occlusions and bad image quality [1].

Depth methods, on the other hand, exploit only range data to perform the pose estimation task. A first attempt to localize accurate nose locations from depth maps in order to perform head tracking and pose estimation was done in [9]. Consequently, [32] used geometric features to identify nose candidates to produce the final pose estimation. A more robust approach was done in [33], [34], [35], where a Random Regression Forest [36] algorithm is exploited for both head detection and pose estimation purposes. Furthermore, in [37] facial point clouds were matched with pose candidates, through a novel triangular surface patch descriptor.
As previously stated for RGB methods, those techniques require facial attributes, thus are prone to errors when such features are not detected.
Remaining depth methods regard the head pose estimation task as an *optimization* problem: [38] used the *Particle Swarm Optimization* (PSO) [39] while [8] perform pose estimation by registering a morphable face model to the measured depth data combining PSO and ICP techniques. Furthermore, [40] used a least-square technique to minimize the difference between the input depth change rate and the prediction rate, to perform 3D head tracking. Finally, in [41] a generative model is proposed, that unifies pose tracking and face model adaptation on-the-fly.
However, no previous method that uses depth maps as the only input exploits CNNs in an effective way. In this work we propose a method based on [10] which uses depth maps to produce accurate head pose predictions by leveraging CNNs.

RGB-D methods combine together RGB images and depth maps. A first effort to leverage both data was done in [42], where a Neural Network is exploited to perform head pose predictions. HOG features [43] were extracted from RGB and depth images in [44], [45], then a *Multi Layer Perceptron* and a linear SVM [46] were used for feature classification, respectively. In [47] Random Forests and tensor regression algorithms are exploited while [48] used a cascade of tree classifiers to tackle extreme head pose estimation task. Recently, in [49] a multimodal CNN was proposed to estimate gaze direction: a regression approach was only approximated through a classifier with a granularity of 1° and with 360 classes. As for RGB and depth methods, these appearance-based techniques are not robust enough: they still strongly depend on the detection of visible facial features.
Following 3D model registration techniques, [50] leverage intensity and depth data to build a 3D constrained local method for robust facial feature tracking. Furthermore, in [51], [52], [53], [54] a 3D morphable model is fitted, using both RGB and depth data to predict head pose. Finally, [55], based on a particle filter formalism, presents a new method for 3D face pose tracking in color images and depth data acquired by RGB-D cameras.
Several works based on head pose estimation, however, do not take in consideration the head localization task.

Head Detection and Localization. To propose a complete head pose estimation framework, head detection is firstly required to find the complete head or a particular point, for example the head center [56]. With RGB or intensity images Viola and Jones [57] face detector is often exploited, *e.g.* in [42], [50], [51], [52], [55]. A different approach demands the head location to a classifier, *e.g.*, [48]. As reported in [8], these approaches suffer due to the lack of generalization capabilities of exploited models, with different acquisition devices and scene contexts.
Recently, deep learning approaches trained on huge face datasets allowed to reach impressive results [58], [59]. However, very few works in literature propose methods for head detection or localization using *only* depth images as input. A method based on a novel head descriptor

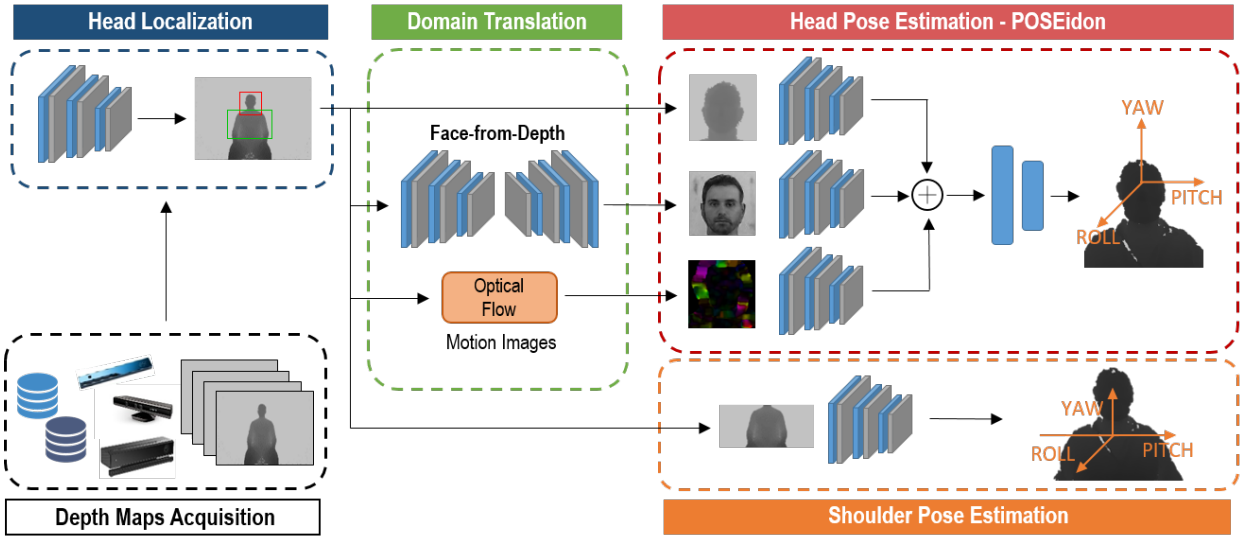


Fig. 4. Overview of the whole *POSEidon*$^+$ framework. Depth input images are acquired by depth sensors (black) and provided to a head localization CNN (blue) to suitably crop the images around the upper-body or head regions. The head crop is used to produce the three inputs for the following networks (green), that are then merged to output the head pose (red). In particular, the *Face-from-Depth* architecture reconstructs gray-level face images from the corresponding depth maps, while the Motion Images are obtained by applying the *Farneback* algorithm. Finally, the upper-body crop is used for the shoulder pose estimation (orange). [best in color]

and an LDA classifier is described in [60]. Every single pixel is classified as head or non-head, and all pixels are clustered for final head detection. In [61] a fall detection system is proposed, in which is included a module for head detection. Heads are detected only on moving objects through a background suppression. In [33] patches extracted from depth images are used to both compute the location and the pose of the head, through a regression forest algorithm.

Driver Body Pose. Only a limited number of works in literature tackle the problem of driver body pose estimation, focusing only on upper-body parts and taking into account automotive contexts. Ito *et al.* [62], adopting an intrusive approach, placed six marker points on the driver body to predict some typical driving operations. A 2D driver body tracking system was proposed in [63], but a manual initialization of the tracking model is strictly required. In [7] a thermal long-wavelength infrared video camera was used to analyze occupant position and posture. In [64] an approach for upper body tracking system using 3D head and hands movements was developed.

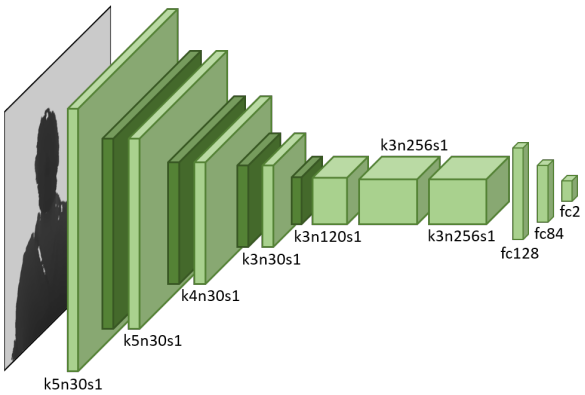


Fig. 5. Architecture of the Head Localization network with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

Domain Translation. Domain translation is the task of learning a parametric translation function between two domains. Recent works have addressed this problem by exploiting Conditional Generative Adversarial Networks (cGAN) [65] in order to learn a mapping from input to output images. Isola *et al.* [13] demonstrated that their model, namely *pix2pix*, is effective at synthesizing photos from label maps, reconstructing objects from edge maps and colorizing images. Moreover, Wang *et al.* [66] proposed a method that acts as a rendering engine: given a synthetic scene, their *Style GAN* is able to render a realistic image. In [67] a cGAN is capable of translating an RGB face image to depth data. Recently, a coupled generative adversarial networks framework has been proposed [68], to generate pairs of corresponding images in two different domains. In our preliminary work [10], we proposed one of the first approach, based on a traditional CNN with common aspects with respect to autoencoders [69] and Fully Convolutional Networks [70], that was trained to compute the appearance of a face using the corresponding depth information.

3 The POSEidon$^+$ framework

An overview of the *POSEidon*$^+$ framework is depicted in Figure 4. The final goal is the estimation of the pose of the driver's head and shoulders, defined as the mass center position and the corresponding orientation relative to the reference frame of the acquisition device [1]. The orientation is provided using three rotation angles *pitch*, *roll* and *yaw*. *POSEidon*$^+$ directly processes the stream of depth frames

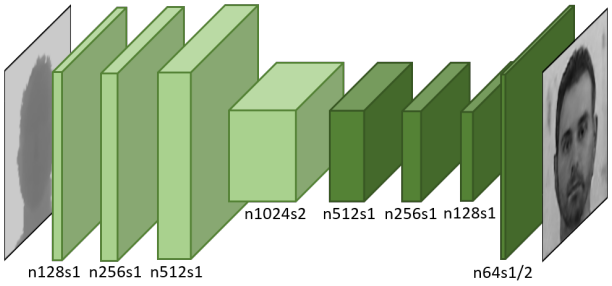


Fig. 6. Architecture of the *Face-from-Depth* network.

captured in real-time by a commercial sensor. Position and size of the foremost head are estimated by a head localization module based on a regressive CNN (Sect. 3.1). The output provided is used to crop the input frames around the head and the shoulder bounding boxes, depending on the further pipeline type. Frames cropped around the head are fed to the head pose estimation block, while the others are exploited to estimate the shoulders pose.
The core components of the system are the *Face-from-Depth* network (Sect. 4), and *POSEidon*$^+$ (Sect. 5), the network which gives the name to the whole framework. Its trident shape is due to the three included CNNs, each working on a specific source: depth, gray level (the output of *FfD*) and *Motion Images* data. The first one – *i.e.*, the CNN directly connected to the input depth data – plays the main role on the pose estimation, while the others cooperate to reduce the estimation error.

3.1 Head Localization

In this step, we defined and trained a network for head localization, relying on the main assumption that a single person is in the foreground. The image coordinates (x_H, y_H) of the head center are the network outputs, or rather, the center mass position of all head points in the frame [56].
Details on the deep architecture adopted are reported in Figure 5. A limited depth and small-sized filters have been chosen to meet real-time constraints while keeping satisfactory performance. For this reason, input images are firstly resized to 160×132 pixels. A max-pooling layer (2×2) is run after each of the first four convolutional layers, while a dropout regularization ($\sigma = 0.5$) is exploited in fully connected layers. The hyperbolic tangent activation (*tanh*) function is adopted, in order to map continuous output values to a predefined range $[-\infty, +\infty] \rightarrow [-1, +1]$. The network has been trained by *Stochastic Gradient Descent* (SGD) [71] and the L_2 loss function.
Given the head position (x_H, y_H) in the frame, a dynamic size algorithm provides the head bounding box with center (x_H, y_H), width w_H and height h_H, around which the frames are cropped:

$$w_H = \frac{f_x \cdot R_x}{D}, \quad h_H = \frac{f_y \cdot R_y}{D} \tag{1}$$

where f_x, f_y are the horizontal and the vertical focal lengths in pixels of the acquisition device, respectively. R_x, R_y are the average width and height of a face (for head pose task $R_x = R_y = 320$) and D is the distance between the head center and the acquisition device, computed averaging the depth values around the head center.
Some examples of bounding boxes estimated by the network are superimposed in the frames of Figure 1.

4 Face-from-Depth

Due to illumination issues, the appearance of the face is not always available if acquired with a RGB camera, *e.g.* inside a vehicle at night. On the contrary, depth maps are generally invariant to illumination conditions but lack of texture details.
We aim to investigate if it is possible to imagine the appearance of a face given the corresponding depth data. Metaphorically, we ask the model to mimic the behavior of a blind person when he tries to figure out the appearance of a friend through the touch.

4.1 Deterministic Conditional GAN

The *Face-From-Depth* network exploits the *Deterministic Conditional GAN* (det-cGAN) paradigm [13] and it is obtained as a generative network G capable of estimating a gray-level image I^E of a face from the corresponding depth representation I^D. The generator G is trained to produce outputs as much indistinguishable as possible from *real* images I by an adversarially trained discriminator D, which is expressly trained to distinguish the *real* images from the *fake* ones produced by the generator. Differently from a traditional GAN [74], [75], the Generator network of a det-cGAN takes an image as input (to be *Conditional*) and not a random noise vector (to be *Deterministic*). As a result, a det-cGAN learns a mapping from observed images x to output images y: $G : x \rightarrow y$.

The objective of a det-cGAN can be expressed as follows:

$$\begin{aligned} L_{det-cGAN}(G, D) = \mathbb{E}_{I \sim p_{data}(I)}[\log D(I)] \\ + \mathbb{E}_{I^D \sim p_{data}(I^D)}[\log(1 - D(G(I^D)))] \end{aligned} \tag{2}$$

where $\log D(I)$ represents the log probability that I is *real* rather than *fake* while $\log(1 - D(G(I^D)))$ is the log probability that $G(I^D)$ is *fake* rather than *real*. G tries to minimize the term $L_{det-cGAN}(G, D)$ of Equation 2, against D that tries to maximize it. The optimal solution is:

$$G^* = \arg \min_G \max_D L_{det-cGAN}(G, D) \tag{3}$$

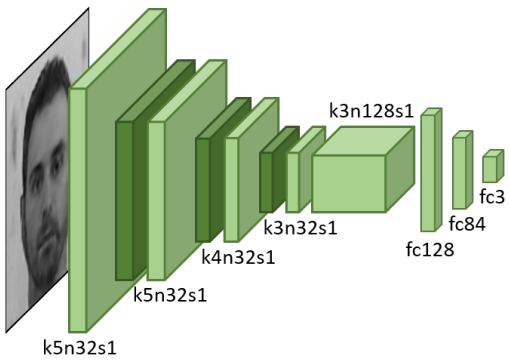


Fig. 7. Architecture of the Head and Shoulder Pose Estimation networks.

TABLE 1
Head Pose Estimation Results on *Biwi*. To allow fair comparisons with state of the art methods, **POSEidon**$^+$ has been evaluated using different evaluation protocols.

Validation Procedure	Year	Data		Head			Avg
		Depth	RGB	Pitch	Roll	Yaw	
ALL SEQUENCES USED AS TEST SET							
Padeleris [38]	2012	√		6.6	6.7	11.1	8.1
Rekik [55]	2013	√	√	4.3	5.2	5.1	4.9
Martin [72]	2014	√		2.5	2.6	3.6	2.9
Papazov [37]	2015	√		2.5 ± 7.4	3.8 ± 16.0	3.0 ± 9.6	4.0 ± 11.0
Meyer [8]	2015	√		2.4	2.1	2.1	2.2
Li [54]	2016	√	√	1.7	3.2	2.2	2.4
Sheng [41]	2017	√		2.0	1.9	2.3	2.1
LEAVE ONE OUT (LOO)							
Drouard [12]	2015		√	5.9 ± 4.8	4.7 ± 4.6	4.9 ± 4.1	5.2 ± 4.5
Drouard [19]	2017		√	10.0 ± 8.7	8.4 ± 8.0	8.6 ± 7.2	9.0 ± 7.9
POSEidon$^+$	2017	√		**2.4 ± 1.3**	**2.6 ± 1.5**	**2.9 ± 1.5**	**2.6 ± 1.4**
K4-FOLD SUBJECT CROSS VALIDATION							
Fanelli [35]	2011	√		3.5 ± 5.8	5.4 ± 6.0	3.8 ± 6.5	- ± -
POSEidon$^+$	2017	√		**2.8 ± 1.7**	**2.9 ± 2.1**	**3.6 ± 2.5**	**3.1 ± 2.1**
K5-FOLD SUBJECT CROSS VALIDATION							
Fanelli [34]	2011	√		8.5 ± 9.9	7.9 ± 8.3	8.9 ± 13.0	8.43 ± 10.4
POSEidon$^+$	2017	√		**2.8 ± 1.8**	**2.8 ± 2.2**	**3.6 ± 2.2**	**3.0 ± 2.1**
K8-FOLD SUBJECT CROSS VALIDATION							
Lathuiliere [25]	2017		√	4.7	3.1	**3.1**	3.6
POSEidon$^+$	2017	√		**2.8 ± 1.9**	**2.8 ± 1.8**	3.3 ± 2.0	**3.0 ± 1.9**
FIXED TRAIN AND TEST SPLITS							
Yang [44]	2012	√	√	9.1 ± 7.4	7.4 ± 4.9	8.9 ± 8.3	8.5 ± 6.9
Baltrusaitis [50]	2012	√	√	5.1	11.3	6.3	7.6
Kaymak [47]	2013	√	√	7.4	6.6	5.0	6.3
Wang [73]	2013	√	√	8.5 ± 14.3	7.4 ± 10.8	8.8 ± 14.3	8.2 ± 12.0
Ahn [11]	2014		√	3.4 ± 2.9	2.6 ± 2.5	2.8 ± 2.4	2.9 ± 2.6
Saeed [45]	2015	√	√	5.0 ± 5.8	4.3 ± 4.6	3.9 ± 4.2	4.4 ± 4.9
Liu [27]	2016		√	6.0 ± 5.8	5.7 ± 7.3	6.1 ± 5.2	5.9 ± 6.1
POSEidon [10]	2017	√		1.6 ± 1.7	1.8 ± 1.8	1.7 ± 1.5	1.7 ± 1.7
POSEidon$^+$	2017	√		**1.6 ± 1.3**	**1.7 ± 1.7**	**1.7 ± 1.3**	**1.6 ± 1.4**

As a possible drawback, the images generated by G are forced to be realistic thanks to D, but they can be unrelated with the original input. For instance, the output could be a nice image of a head with a very different pose with respect to the input depth. Thus, is fundamental mixing the GAN objective with a more traditional loss, such as SSE distance [76]. While discriminators job remains unchanged, the generator, in addition to fooling the discriminator, tries to emulate the ground truth output in an SSE sense. The pixel-wise SSE is calculated between downsized versions of the generated and target images, first applying an averaged pooling layer. We formulate the final objective as the weighted sum of a content loss and an adversarial loss as:

$$G^* = \arg\min_G \max_D L_{det-cGAN}(G, D) + \lambda L_{SSE}(G) \quad (4)$$

where λ is the weight controlling the content loss impact.

4.2 Network Architecture

We propose to modify the classic hourglass generator architecture, performing a limited number of upsampling and downsampling operations. As shown in the following experimental section, the *U-Net* architecture [79] can be adopted in order to shuttle low-level information between input and output directly across the network [13], but it is less convenient in our case.

Following the main architecture guidelines for stable Deep Convolutional GANs by Radford *et al.* [75], we instead adopt the architecture illustrated in Figure 6 for the Generator. Specifically, in the encoder part, we use three convolutional layers followed by a strided convolutional layer (with stride 2) to halve the image resolution.

The decoding stack uses three convolutional layers followed by a transposed convolutional layer (also referred as fractionally strided convolutional layers) with stride $1/2$ to double the resolution, and a final convolution. The number of filters follows a power of 2 pattern, from 128 to 1024 in the encoder and from 512 to 64 in the decoder. *Leaky ReLU* is used as activation function in the encoding phase while *ReLU* is used in the decoding phase.

We adopt *batch normalization* before each activation (except for the last layer) and a kernel size 5×5 for each convolution. The discriminator architecture complies with the generators encoder in terms of activation and number of filters, but contains only strided convolutional layers (with stride 2) to halve the image resolution each time the number of filters is

TABLE 2
Evaluation metrics computed on the reconstructed gray-level face images with *Biwi* and *Pandora* datasets. Starting from the left, L_1 and L_2 distances are reported, then the absolute and the squared differences, the root-mean-square error and, finally, the percentage of pixels under a certain threshold. Further details about metrics are reported in [77].

Dataset	Method	Norm ↓		Difference ↓		RMSE ↓			Threshold ↑		
		L_1	L_2	Abs	Squared	linear	log	scale-inv	1.25	2.5	3.75
Biwi	FfD [10]	33.35	2586	0.454	24.07	40.55	**0.489**	**0.445**	0.507	**0.806**	**0.878**
	FfD	**24.44**	**2230**	**0.388**	**19.81**	**35.50**	0.653	0.610	**0.615**	0.764	0.840
Pandora	FfD [10]	41.36	3226	0.705	46.00	50.77	0.603	**0.485**	0.263	0.725	0.819
	pix2pix [13]	19.37	1909	**0.468**	24.07	30.80	0.568	0.539	0.583	0.722	0.813
	AVSS [78]	23.93	2226	0.629	34.49	35.46	0.658	0.579	0.541	0.675	0.764
	FfD + U-Net	23.75	2123	0.653	34.96	33.89	0.639	0.553	0.555	0.689	0.775
	FfD	**18.21**	**1808**	0.469	**22.90**	**28.90**	**0.556**	0.501	**0.605**	**0.743**	**0.828**

doubled. The network then outputs one *sigmoid* activation. In the discriminator, we use batch normalization before every Leaky ReLU activation, except for the first layer.

4.3 Training details

We trained the det-cGAN with depth images and simultaneously providing the network with the original gray-level images associated with the depth data in order to compute the L_{SSE}. To optimize the network we adopted the standard approach from Goodfellow *et al.* [74] and alternate the gradient descent updates between the generator and the discriminator with $K = 1$. We used mini-batch SGD applying the *Adam* solver [80] with $\beta_1 = 0.5$ and batch size of 64. We set $\lambda = 10^{-1}$ in Equation 4 for the experiments. Moreover, to encourage the discriminator to estimate soft probabilities rather than to extrapolate extremely confident classifications, we used a technique called *one-sided label smoothing* [81] where the target for the real examples are replaced with a value slightly less than 1, such as 0.9. This solution prevents the discriminator to produce extremely confident predictions that could unbalance the adversarial learning.

5 Pose Estimation from Depth

5.1 POSEidon$^+$ network

The *POSEidon*$^+$ network is a fusion of three CNNs and has been developed to perform a regression on the 3D pose angles. As a result, continuous Euler values – corresponding to the *yaw*, *pitch* and *roll* angles – are estimated (right part of Fig. 4). The three *POSEidon*$^+$ components have the same shallow architecture based on 5 convolutional layers with kernel size of 5×5, 4×4 and 3×3 and a 2×2 max-pooling is conducted only on the first three layers due to the limited size of the input (64×64). The first four convolutional layers have 32 filters each, the last one has 128 filters. *tanh* is exploited as activation function; we are aware that *ReLU* [82] converges faster, but better performance in term of accuracy prediction are achieved.

The three networks are fed with different input data types: the first one, directly takes as input the head-cropped depth images; the second one is connected to the *Face-from-Depth* output and the last one operates on Motion Images, obtained applying the standard *Farneback* algorithm [83] on pairs of consecutive depth frames. The presence of depth discontinuities around the nose and the eyes generates specific motion patterns which are related to the head pose. Motion Images, thus, provide useful information for the estimation of the pose of a moving head. Frames with motionless heads are very rare in real videos. However, in those cases the common image compression creates artifacts around the face landmarks which allow the estimation of the head pose.

A fusion step combines the contributions of the three above described networks. The last fully connected layer of each component is removed in order to provide the following layers with more data and not only the estimated angles. As a results, the output of the whole *POSEidon*$^+$ network is not only a weighted mean of the three component outputs, but a more complex combination. Different fusion approaches that have been proposed by Park *et al.* [84] are investigated. Given two feature maps x^a, x^b with a certain width w and height h, for every feature channel d_a^x, d_b^x and $y \in R^{w \times h \times d}$:

- **Multiplication**: computes the element-wise product of two feature maps, as $y^{mul} = x^a \circ x^b, d^y = d_a^x = d_b^x$
- **Concatenation**: stacks two features maps, without any blend $y^{cat} = [x^a | x^b], d^y = d_a^x + d_b^x$
- **Convolution**: stacks and convolves feature maps with a filter k of size $1 \times 1 \times (d_a^x + d_b^x)/2$ and β as bias term, $y^{conv} = y^{cat} * k + \beta, \quad d^y = (d_a^x + d_b^x)/2$

The final *POSEidon*$^+$ framework exploits a combination of two fusing methods, in particular, a convolution followed by a concatenation. After the fusion step, three fully connected layers composed of 128, 84 and 3 activations respectively and two dropout regularization ($\sigma = 0.5$) complete the architecture. *POSEidon*$^+$ is trained with a double-step procedure. First, each individual network described above is trained with the following L_2^w weighted loss:

$$L_2^w = \sum_{i=1}^{3} \| w_i \cdot (y_i - f(x_i)) \|_2 \tag{5}$$

where $w_i \in [0.2, 0.35, 0.45]$. This weight distribution gives more importance to the yaw angle, which is preponderant in the selected automotive context. During the individual training step, the last fully connected layer of each network is preserved, then is removed to perform the second training phase. Holding the weights learned for the trident components, the new training phase is carried out on the last three fully connected layers of *POSEidon*$^+$ only, with the loss function L_2^w reported in Equation 5. In all training

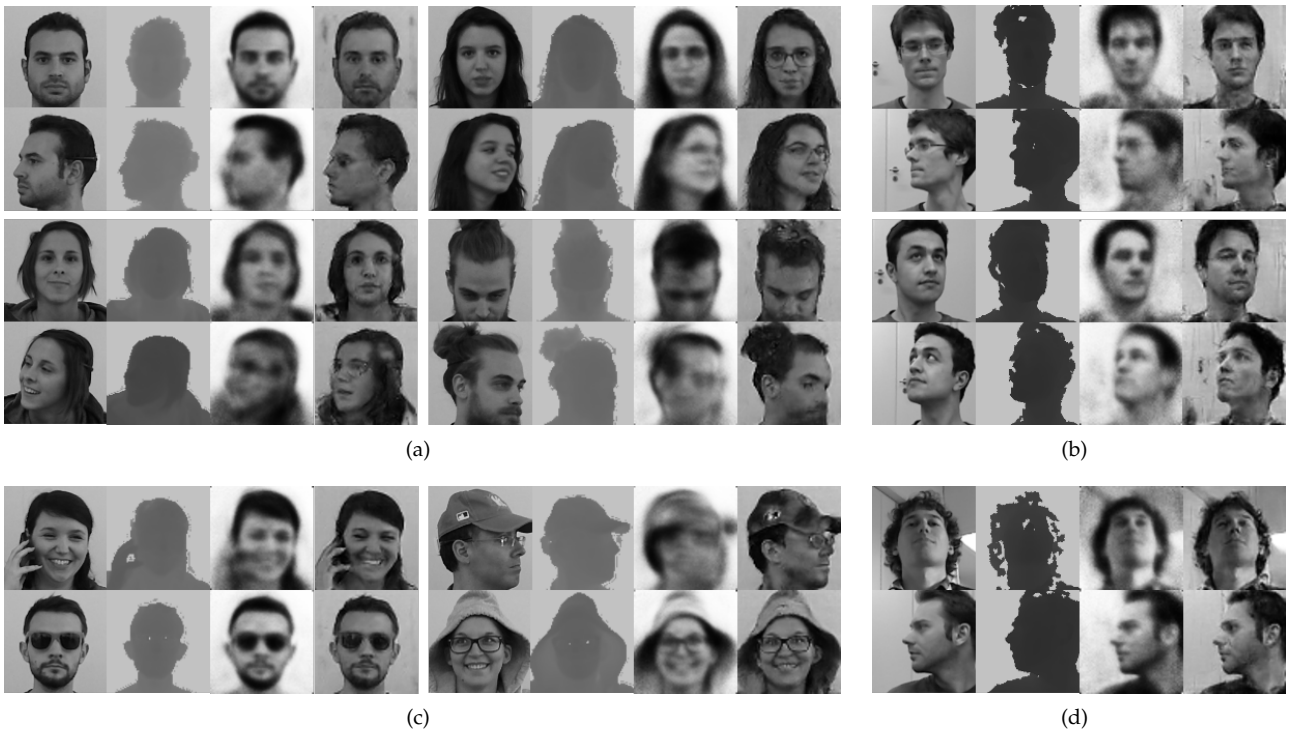


Fig. 8. Test (a) and train (c) images on *Pandora* dataset, test (b) and train (d) images on *Biwi* dataset. For each block, gray-level images and then the corresponding depth faces are depicted in the first columns; face images taken from the method described in [10] are reported in the third column; finally, the output of the *Face-from-Depth* network proposed in this paper is depicted in the last column.

steps, the SGD optimizer [71] is exploited, the learning rate is set initially to 10^{-1} and then is reduced by a factor 2 every 15 epochs.

5.2 Shoulder Pose Estimation

The framework is completed with an additional network for the estimation of the shoulder pose. We employ the same architecture adopted for the head (Sect. 5.1), performing regression on the three pose angles.
Starting from the head center position (Sect. 3.1), the depth input images are cropped around the driver neck, using a bounding box $\{x_S, y_S, w_S, h_S\}$ with center $(x_S = x_H, y_S = y_H - (h_H/4))$, and width and height obtained as described in Equation 1, but with different values of R_x, R_y to produce a rectangular crop: these values are tested and discussed in Section 6. The network is trained with SGD optimizer [71], using the weighted L_2^w loss function described above (see Eq. 5). As usual, hyperbolic tangent is exploited as activation function.

6 Experimental Results

6.1 Datasets

Network training and testing phases have been done exploiting two publicly available datasets, namely *Biwi Kinect Head Pose* and *ICT-3DHP*. In addition, we collected a new dataset, called *Pandora*, which also contains shoulder pose annotations. Data augmentation techniques are employed to enlarge the training set, in order to achieve space invariance and avoid overfitting [71].
Random translations on vertical, horizontal and diagonal directions, jittering, zoom-in and zoom-out transformation of the original images have been exploited. Percentile-based contrast stretching, normalization and scaling of the input images are also applied to produce zero mean and unit variance data.
Other datasets for head pose estimation and related tasks have been collected in last decades [85], [86], [87], [88], [89], but in most cases there are some not desirable features, for instance, no depth or 3D data, no continuous ground truth annotations and not enough data for deep learning techniques.
Follows a detailed description of the three adopted datasets.

6.1.1 Biwi Kinect Head Pose dataset

Fanelli *et al.* [33] introduced this dataset in 2013. It is acquired with the *Microsoft Kinect* sensor, i.e., a structured IR light device. It contains about 15k frames, with RGB (640×480) and depth maps (640×480). Twenty subjects have been involved in the recordings: four of them were recorded twice, for a total of 24 sequences. The ground truth of yaw, pitch and roll angles is reported together with the head center and the calibration matrix. The original paper does not report the adopted split between training and testing sets; fair comparisons are thus not guaranteed. To avoid this, we clearly report the adopted split in the following.

6.1.2 ICT-3DHP dataset

ICT-3DHP dataset has been introduced by Baltrusaitis *et al.* in 2012 [50]. It has been collected using a *Microsoft Kinect* sensor and contains RGB images and depth maps of about

14k frames, divided into 10 sequences. The image resolution is 640×480 pixels. An additional hardware sensor (*Polhemus Fastrack*) is exploited to generate the ground truth annotation. The device is placed on a white cap worn by each subject, visible in both RGB and depth frames. The presence of a few subjects and the limited number of frames make this dataset unsuitable for training deep learning approaches.

6.1.3 Pandora dataset

In addition to publicly available datasets, we have also collected and used a new challenging dataset, called *Pandora*. It has been specifically created for head center localization, head pose and shoulder pose estimation in automotive contexts (See Fig.3). A frontal and fixed device acquires the upper body part of the subjects, simulating the point of view of a camera placed inside the dashboard. The subjects mainly perform driving-like actions, such as holding the steering wheel, looking to the rear-view or lateral mirrors, shifting gears and so on. *Pandora* contains 110 annotated sequences of 10 male and 12 female actors. Each subject has been recorded five times. *Pandora* is the first publicly available dataset which combines the following features:

- **Shoulder angles**: in addition to the head pose annotation, *Pandora* contains the ground truth data of the shoulder pose expressed as yaw, pitch, and roll.
- **Wide angle ranges**: subjects perform wide head ($\pm 70°$ roll, $\pm 100°$ pitch and $\pm 125°$ yaw) and shoulder ($\pm 70°$ roll, $\pm 60°$ pitch and $\pm 60°$ yaw) movements. For each subject, two sequences are performed with constrained movements, changing the yaw, pitch and roll angles separately, while three additional sequences are completely unconstrained.
- **Challenging camouflage**: garments, as well as various objects are worn or used by the subjects to create head and/or shoulder occlusions. For example, people wear prescription glasses, sunglasses, scarves, caps, and manipulate smart-phones, tablets or plastic bottles.
- **Deep-learning oriented**: the dataset contains more than 250k full resolution RGB (1920×1080) and depth images (512×424) with the corresponding annotation.
- **Time-of-Flight (ToF) data**: a *Microsoft Kinect One* device is used to acquire depth data, with a better quality than other datasets created with the first *Kinect* version [90].

Each frame of the dataset is composed of an RGB appearance image, the corresponding depth map, and the 3D coordinates of the skeleton joints corresponding to the upper body part, including the head center and the shoulder positions. For convenience's sake, the 2D coordinates of the joints on both color and depth frames are provided as well as the head and shoulder pose angles with respect to the camera reference frame. Shoulder angles are obtained through the conversion to Euler angles of a corresponding rotation matrix, obtained from a user-centered system [91] and defined by the following unit vectors (N_1, N_2, N_3):

$$N_1 = \frac{p_{RS} - p_{LS}}{\|p_{RS} - p_{LS}\|} \quad U = \frac{p_{RS} - p_{SB}}{\|p_{RS} - p_{SB}\|}$$
$$N_3 = \frac{N_1 \times U}{\|N_1 \times U\|} \quad N_2 = N_1 \times N_3 \tag{6}$$

where p_{LS}, p_{RS} and p_{SB} are the 3D coordinates of the left shoulder, right shoulder and spine base joints. The annotation of the head pose angles has been collected using a wearable *Inertial Measurement Unit* (IMU) sensor. To avoid distracting artifacts on both color and depth images, the sensor has been placed in a non-visible position, *i.e.*, on the rear of the subject's head. The IMU sensor has been calibrated and aligned at the beginning of each sequence, assuring the reliability of the provided angles. The dataset is publicly available (http://aimagelab.ing.unimore.it/pandora/).

6.2 Quantitative Results

The proposed framework has been deeply tested using the datasets described in Section 6.1.3. For evaluation with the *Pandora* dataset, sequences of subjects 10, 14, 16 and 20 have been used for testing, the remaining for training and validation. With *Biwi* dataset, test subjects are determined by the validation procedure adopted. Finally, we tested the system on all the sequences contained in *ICT-3DHP* dataset.

Domain Translation. First, we check the capabilities of the *Face-from-Depth* network alone. Some visual examples of input, output, and ground-truth images are reported in Figure 8.

With this aim, we propose two types of evaluation. The first is based on metrics related to the reconstruction accuracy. Following the work of Eigen etal [77], Table 2 reports some results. The system is evaluated both on *Biwi* and on *Pandora* datasets. *FfD* network is compared with other Image-to-Image methods taken from the recent literature. In particular, we trained from scratch the deep models proposed in [13], [78] (referred here as *pix2pix* and *AVSS*, respectively), following procedures reported in the corresponding papers. Moreover, in order to investigate how architectural choices impact the reconstruction quality of *FfD*, we tested a different design. We modified the network adding the *U-Net* [79] skip connections between mirrored layers (cf. Sect. 4.2).

We also compared the presented approach with our preliminary version of *Face-from-Depth* network [10], that fuses the key aspects of *encoder-decoder* [69] and *fully convolutional* [70] neural networks.

For the sake of comparison, we report here key details about the preliminary *FfD* version [10]. It has been trained in a single step, with input head images resized to 64×64 pixels. The activation function is the hyperbolic tangent and best training performance are reached through the self adaptive *Adadelta* optimizer [92]. A particular loss function is

TABLE 3
Results obtained on *Pandora* dataset with head pose network trained on gray level images and tested with the original gray-level and reconstructed ones.

Testing input	**Head** Pitch	Roll	Yaw	**Acc.**
gray-level	**7.1 ± 5.6**	**5.6 ± 5.8**	**9.0 ± 10.9**	**0.613**
pix2pix [13]	7.9 ± 8.0	5.9 ± 6.3	12.8 ± 21.4	0.581
AVSS [78]	8.9 ± 8.5	6.2 ± 6.4	13.4 ± 20.4	0.543
FfD [10]	8.5 ± 8.9	6.1 ± 6.2	12.4 ± 17.3	0.559
FfD + U-Net	8.7 ± 8.4	6.4 ± 6.6	13.5 ± 19.9	0.552
FfD	**7.6 ± 6.9**	**5.8 ± 6.0**	**10.1 ± 12.6**	**0.613**

TABLE 4
Results of the head pose estimation on *Pandora* comparing different system architectures. The baseline is a single CNN working on the source depth map (Row 1). The accuracy is the percentage of correct estimations (err <15°). FfD: *Face-from-Depth*, MI: *Motion Images*.

HEAD POSE ESTIMATION ERROR [EULER ANGLES]

#	Input				Crop	Fusion	Head			Accuracy
	Depth	FfD	MI	Gray			Pitch	Roll	Yaw	
1	√					-	8.1 ± 7.1	6.2 ± 6.3	11.7 ± 12.2	0.553
2	√				√	-	6.5 ± 6.6	5.4 ± 5.1	10.4 ± 11.8	0.646
3		√			√	-	6.8 ± 6.1	5.8 ± 5.0	10.1 ± 12.6	0.658
4			√		√	-	7.7 ± 7.5	5.3 ± 5.7	10.0 ± 12.5	0.609
5				√	√	-	7.1 ± 6.6	5.6 ± 5.8	9.0 ± 10.9	0.639
6	√	√			√	concat	5.6 ± 5.0	4.9 ± 5.0	9.7 ± 12.1	0.698
7	√		√		√	concat	6.0 ± 6.1	4.5 ± 4.8	9.2 ±11.5	0.690
8	√	√	√		√	conv+concat	**5.6 ± 5.2**	**4.8 ± 5.0**	**8.2 ± 9.8**	**0.736**

exploited in order to highlight the central area of the image, where the face is supposed to be after the cropping step, and takes in account the distance between the reconstructed image and the corresponding gray-level ground truth:

$$L = \frac{1}{R \cdot C} \sum_{i}^{R} \sum_{j}^{C} \left(||y_{ij} - \bar{y}_{ij}||_2^2 \cdot w_{ij}^{\mathcal{N}} \right) \quad (7)$$

where R, C are the number of rows and columns of the input images, respectively. $y_{ij}, \bar{y}_{ij} \in \mathcal{R}^{ch}$ are the intensity values from ground truth ($ch = 1$) and predicted appearance images. Finally, the term $w_{ij}^{\mathcal{N}}$ introduces a bivariate Gaussian prior mask. Best results have been obtained using $\mu = [\frac{R}{2}, \frac{C}{2}]^T$ and $\Sigma = \mathbb{I} \cdot [(R/\alpha)^2, (C/\beta)^2]^T$ with α and β empirically set to $3.5, 2.5$ for squared images of $R = C = 64$. Other details about network architecture and training are reported in [10].

The second set of tests is specific to the head pose estimation task. The head pose network described in Section 5, trained with gray-level images taken from the *Pandora* dataset, is tested on the reconstructed face images. Since the network has been trained on real gray-level images to output the angles of the head pose, we can suppose that the more generated images are similar to the corresponding gray-level ones, the better the results are. The comparison is presented in Table 3. In the first row, results obtained using gray-level images as testing input are reported, this is the best case and should be used as a reference baseline. Results present in the following rows confirm that our *FfD* is able to generate high-quality faces, very similar to gray-level faces. Moreover, we note that the head pose network has the ability to generalize well on cross-dataset evaluations since we generally obtain a good accuracy even with different types of face images as input. The *Face-from-Depth* network has been created to this goal, even if the output is not always realistic and visually pleasant: however, the promising results confirm their positive contribution in the estimation of the head pose.

Head Pose Estimation. An ablation study of *POSEidon*$^+$ framework on *Pandora* is conducted and results are reported in Table 4, providing mean and standard deviation of the estimation errors obtained on each angle and for each system configuration. Similar to Fanelli *et al.* [35], we also report the mean accuracy as percentage of good estimations (*i.e.*, angle error below 15°).

The first row of Table 4 shows the performance of a baseline system, obtained using the head pose estimation network only, and input depth frames are directly fed to the network without processing and cropping the input around the head. As expected, results are reasonable proving the ability of the deep network to extract useful features for head pose estimation from whole images.
The cropping step is included instead in the other rows, using the ground truth head position as the center and the cropping method described in Section 3.1. All three branches (i.e., depth, *FfD*, and Motion Images) of *POSEidon*$^+$ framework are individually evaluated. In particular, the fifth row includes an indirect evaluation of the reconstruction capabilities of the *Face-from-Depth* network. The same network trained and tested on the original gray level images performs similarly to the one trained and tested on *FfD* outputs (Row 3). The similar results confirm that the image reconstruction quality is sufficiently accurate, at least for the pose estimation task.
Results obtained using couples of networks are shown in rows 6 and 7, exploiting concatenation to merge the final layers of each component. Finally, the last row reports the performance of the complete framework. To merge layers, we use a *conv* fusion of couples of input types, followed by the *concat* step. We found that it is the best combination, as described in [10]. Even if the choice of the fusion method has a limited effect (as deeply investigated in [84], [93]), the most significant improvement of the system is reached by combining and exploiting the three input types together.

Figure 9 shows a comparison of the performance provided by each trident component: each graph plots the error distribution of a specific network with respect to the ground

TABLE 5
Results for head pose estimation task on *Pandora* dataset. In particular, here we compare our preliminary work [10] with the proposed one. In addition, we include a comparison with *POSEidon*$^+$ framework, in which we replace the head pose estimation network trained on reconstructed face images with the same network trained on gray-level images, here referred as *POSEidon**.

Method	Head			Acc.
	Pitch	Roll	Yaw	
POSEidon [10]	5.7 ± 5.6	4.9 ± 5.1	9.0 ± 11.9	0.715
POSEidon*	5.6 ± 5.8	4.8 ± 5.0	8.8 ± 10.9	0.720
POSEidon$^+$	**5.6 ± 5.2**	**4.8 ± 5.0**	**8.2 ± 9.8**	**0.736**

TABLE 6
Estimation errors and mean accuracy of the shoulder pose estimation on *Pandora*

Parameters		Shoulders			Accuracy
R_x	R_y	Pitch	Roll	Yaw	
No crop		2.5 ± 2.3	3.0 ±2.6	3.7 ± 3.4	0.877
700	250	2.9 ± 2.6	2.6 ±2.5	4.0 ± 4.0	0.845
850	250	2.4 ± 2.2	2.5 ±2.2	3.1 ± 3.1	0.911
850	500	**2.2 ± 2.1**	**2.3 ±2.1**	**2.9 ± 2.9**	**0.924**

TABLE 7
Results on *Biwi*, *ICT-3DHP* and *Pandora* dataset of the complete *POSEidon*$^+$ pipeline (*i.e.*, head localization, cropping and pose estimation).

Dataset	Local.	Head		
		Pitch	Roll	Yaw
Biwi	3.27±2.19	1.5±1.4	1.6±1.6	2.2±2.0
ICT-3DHP	-	4.9±4.2	3.5±3.4	6.8±6.0
Pandora	4.27±3.25	7.3±8.2	4.6±4.5	10.3±11.4

truth value. Depth data allows reaching the lowest error rates for frontal heads, while the other input data types are better in presence of rotated poses. The graphs highlight the averaging capabilities of *POSEidon*$^+$ too.
Furthermore, in Table 5 we compare best performance of *POSEidon*$^+$ on *Pandora* dataset, obtained exploiting the *FfD* network proposed in this paper and the previous one described in [10]. We also evaluate *POSEidon*$^+$ replacing the central CNN (see Fig. 4) trained on reconstructed face images with the same CNN but trained on gray-level images (this experiment is here referred as *POSEidon**). Results confirm that the proposed *POSEidon*$^+$ overcomes our preliminary work. The overall quality of reconstructed face images is confirmed and also the feasibility to train and test the pose network on different dataset without a significant drop in performance.

Finally, we compare the results of *POSEidon*$^+$ with the state-of-art on the *Biwi* dataset. Due to the lack of a common validation and test protocol, Table 1 is split accordingly to the evaluation procedures adopted, in order to allow fair comparisons. For each validation procedure, we report results of *POSEidon*$^+$. In particular, we implement a 2-folds (half subjects in train and half in test), 4-folds, 5-folds (as adopted in the original works [34], [35], respectively) and 8-folds subject independent cross evaluations. We also conduct the *Leave-One-Out* (LOO) validation protocol. We dedicate the last section of Table 1 also for those methods that do not follow a standard evaluation procedure since they create a fixed or random [11] sets with a limited number of subjects (or sequences) to test their systems. Besides, we note that a fair comparison with methods reported in the top part of Table 1 is not possible since they exploit all sequences of *Biwi* dataset for test, while deep learning approaches need a certain amount of training data. Results confirm the excellent performance of *POSEidon*$^+$ and the generalization ability across different training and testing subsets with different validation protocol. The system overcomes all the reported methods, included our previous proposal [10]. The average error is lower than other approaches, even those are not using all the frames available on *Biwi* dataset (some works exclude the frames on which the face detection fails [34], [35]).

Shoulder Pose Estimation. The network performing the shoulder pose estimation has been tested on *Pandora* only, due to the lack of the corresponding annotation in the other datasets. Results are reported in Table 6.
In particular, we conduct evaluation on different input types, varying the values R_x and R_y (cf. Section 5.2) that affect head and shoulder crops. We test also the shoulder pose network using the whole input depth frame, without any crop. The reported results are very promising, reaching an accuracy of over 92%.

Complete pipeline. In order to have a fair comparison, results reported in Tables 1 and 4 are obtained using the ground truth head position as input to the crop procedure. We finally test the whole pipeline, including the head localization network described in section 3.1, using also *ICT-3DHP* dataset. The mean error of the head localization (in pixels) and the pose estimation errors are summarized in Table 7. Sometimes, the estimated position generates a more effective crop of the head and, as a result, the whole pipeline performs better on the head pose estimation over the *Biwi* dataset. *POSEidon*$^+$ reaches valuable results also on the *ICT-3DHP* dataset and it provides comparable results with respect to state-of-the-art methods working on both depth and RGB data (4.9±5.3, 4.4±4.6, 5.1±5.4 [45], 7.06, 10.48, 6.90 [50], for pitch, roll and yaw respectively). We note that *ICT-3DHP* does not include the head center annotation, but the position of the device used to acquire pose data placed on the back of the head, and this partially compromises the performance of our method. Besides, we can not suppose a coherency between the annotations obtained with different IMU devices, in particular regarding the definition of the null position (*i.e.*, when the head angles are equal to zero).
The complete framework – except for the *FfD* module – has been implemented and tested on a desktop computer equipped with a *NVidia Quadro k2200* GPU board and on a laptop with a *NVidia GTX 860M*. Real-time performance has been obtained in both cases, with a processing rate of

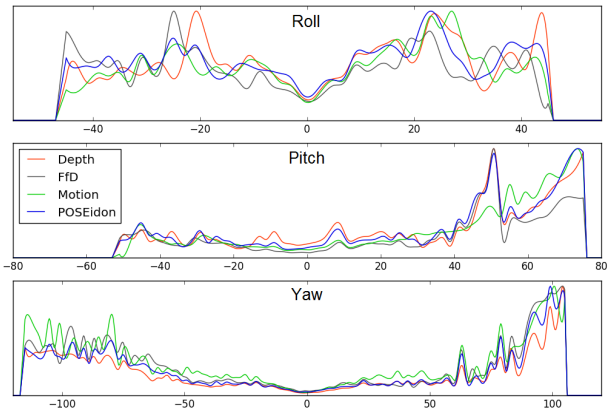


Fig. 9. Error distribution of each *POSEidon*$^+$ components on *Pandora* dataset. On x-axis are reported the ground truth angles, on y-axis the distribution of error for each input type.

more than 30 frames per second. The whole system has been tested instead on a *Nvidia GTX 1080* and is able to run at more than 50 frames per second. Some examples of the system output are reported in Figure 1. In addition, the original depth map, the *Face-from-Depth* reconstruction and the motion data given in input to *POSEidon*$^+$ are placed on the left of each frame.

7 Conclusions

An end-to-end framework to monitor the driver's body pose called *POSEidon*$^+$ is presented. In particular, a new *Face-from-Depth* architecture is proposed, based on a Deterministic Conditional GAN approach, to convert depth faces in gray-level images and supporting head pose prediction.
The system is based only on depth images, no previous computation of specific facial features is required and has shown real-time and impressive results with two public datasets. All these aspects make the proposed framework suitable to particular challenging contexts, such as automotive. Since the system has been developed with a modular architecture, each module can be used as single or in combination, reaching worst but still satisfactory performances. This work provides a comprehensive review and a comparison of recent state-of-art works and can be used as a brief review to understanding the current state of the 3D head pose estimation task.

Acknowledgments

This work has been carried out within the projects Citta educante (CTN01-00034-393801) of the National Technological Cluster on Smart Communities funded by MIUR and *FAR2015 - Monitoring the car drivers attention with multisensory systems, computer vision and machine learning* funded by the University of Modena and Reggio Emilia. We also acknowledge *Ferrari* SpA and CINECA for the availability of real car equipments and high performance computing resources, respectively.

References

[1] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009. 1, 3, 4

[2] C. Tran and M. M. Trivedi, "Vision for driver assistance: Looking at people in a vehicle," in *Visual Analysis of Humans*. Springer, 2011, pp. 597–614. 1

[3] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez, "Real-time system for monitoring driver vigilance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 63–77, 2006. 1

[4] A. Doshi and M. M. Trivedi, "Head and eye gaze dynamics during visual attention shifts in complex environments," *Journal of vision*, vol. 12, no. 2, pp. 9–9, 2012. 1

[5] B. Czupryński and A. Strupczewski, "High accuracy head pose tracking survey," in *International Conference on Active Media Technology*. Springer, 2014, pp. 407–420. 1

[6] Q. Ji, Z. Zhu, and P. Lan, "Real-time nonintrusive monitoring and prediction of driver fatigue," *IEEE transactions on vehicular technology*, vol. 53, no. 4, pp. 1052–1068, 2004. 1

[7] M. M. Trivedi, S. Y. Cheng, E. M. Childers, and S. J. Krotosky, "Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 6, pp. 1698–1712, 2004. 1, 4

[8] G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz, "Robust model-based 3d head pose estimation," in *Proc. of IEEE International Conference on Computer Vision*, 2015, pp. 3649–3657. 1, 3, 6

[9] S. Malassiotis and M. G. Strintzis, "Robust real-time 3d head pose estimation from range data," *Pattern Recognition*, vol. 38, no. 8, pp. 1153–1165, 2005. 1, 3

[10] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 4, 6, 7, 8, 9, 10, 11

[11] B. Ahn, J. Park, and I. S. Kweon, "Real-time head orientation from a monocular camera using deep neural network," in *Proc. of Asian Conference on Computer Vision*, 2014, pp. 82–96. 2, 3, 6, 11

[12] V. Drouard, S. Ba, G. Evangelidis, A. Deleforge, and R. Horaud, "Head pose estimation via probabilistic high-dimensional regression," in *Proc. of IEEE International Conference on Image Processing*, 2015, pp. 4624–4628. 2, 3, 6

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016. 2, 4, 5, 6, 7, 9

[14] J. Whitehill and J. R. Movellan, "A discriminative approach to frame-by-frame head pose tracking," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–7. 2

[15] T. Vatahska, M. Bennewitz, and S. Behnke, "Feature-based head pose estimation from images," in *Proc. of 7th IEEE-RAS International Conference on Humanoid Robots*, 2007, pp. 330–335. 2

[16] R. Yang and Z. Zhang, "Model-based head pose tracking with stereovision," in *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, 2002, pp. 255–260. 2

[17] Y. Matsumoto and A. Zelinsky, "An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement," in *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 499–504. 2

[18] Y. Sun and L. Yin, "Automatic pose estimation of 3d facial models," in *Proc. of International Conference on Pattern Recognition*, 2008, pp. 1–4. 2

[19] V. Drouard, S. Ba, and R. Horaud, "Switching linear inverse-regression model for tracking head pose," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 1232–1240. 2, 6

[20] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proc. of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194. 2

[21] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3d shape regression for real-time facial animation," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 41, 2013. 2

[22] M. Storer, M. Urschler, and H. Bischof, "3d-mam: 3d morphable appearance model for efficient fine head pose estimation from still images," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 192–199. 2

[23] M. Osadchy, Y. L. Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *Journal of Machine Learning Research*, vol. 8, no. May, pp. 1197–1215, 2007. 3

[24] X. Xu and I. A. Kakadiaris, "Joint head pose estimation and face alignment framework using global and local cnn features," in *Proc. 12th IEEE Conference on Automatic Face and Gesture Recognition, Washington, DC*, vol. 2, 2017. 3

[25] S. Lathuilière, R. Juge, P. Mesejo, R. Munoz-Salinas, and R. Horaud, "Deep mixture of linear inverse regressions applied to head-pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3, 6

[26] K. Khan, M. Mauro, P. Migliorati, and R. Leonardi, "Head pose estimation through multi-class face segmentation," in *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 2017, pp. 175–180. 3

[27] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei, "3d head pose estimation with convolutional neural network trained on synthetic images," in *Proc. of IEEE International Conference on Image Processing*, 2016, pp. 1289–1293. 3, 6

[28] T. Bär, J. F. Reuter, and J. M. Zöllner, "Driver head pose and gaze estimation based on multi-template icp 3-d point cloud alignment," in *2012 15th International IEEE Conference on Intelligent Transportation Systems*, 2012, pp. 1797–1802. 3

[29] K.-L. Low, "Linear least-squares optimization for point-to-plane icp surface registration," *Techrep - Chapel Hill, University of North Carolina*, vol. 4, 2004. 3

[30] J. Chen, J. Wu, K. Richter, J. Konrad, and P. Ishwar, "Estimating head pose orientation using extremely low resolution images," in *2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, 2016, pp. 65–68. 3

[31] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face Alignment Across Large Poses: A 3D Solution," *ArXiv e-prints*, Nov. 2015. 3

[32] M. D. Breitenstein, D. Kuettel, T. Weise, L. Van Gool, and H. Pfister, "Real-time face pose estimation from single range images," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8. 3

[33] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 617–624. 3, 4, 8

[34] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real time head pose estimation from consumer depth cameras," in *Joint Pattern Recognition Symposium*, 2011, pp. 101–110. 3, 6, 11

[35] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3d face analysis," *Int. J. Comput. Vision*, vol. 101, no. 3, pp. 437–458, 2013. 3, 6, 10, 11

[36] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002. 3

[37] C. Papazov, T. K. Marks, and M. Jones, "Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4722–4730. 3, 6

[38] P. Padeleris, X. Zabulis, and A. A. Argyros, "Head pose estimation on depth data based on particle swarm optimization," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 42–49. 3, 6

[39] J. Kennedy, "Particle swarm optimization," in *Encyclopedia of machine learning*. Springer, 2011, pp. 760–766. 3

[40] F. A. Kondori, S. Yousefi, H. Li, S. Sonning, and S. Sonning, "3d head pose estimation using the kinect," in *Proc. of International Conference on Wireless Communications and Signal Processing (WCSP)*, 2011, pp. 1–4. 3

[41] L. Sheng, J. Cai, T.-J. Cham, V. Pavlovic, and K. Ngi Ngan, "A generative model for depth-based robust 3d facial pose tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4488–4497. 3, 6

[42] E. Seemann, K. Nickel, and R. Stiefelhagen, "Head pose estimation using stereo vision for human-robot interaction." in *Proc. of Sixth International Conference on Face and Gesture Recognition*. IEEE Computer Society, 2004, pp. 626–631. 3

[43] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893. 3

[44] J. Yang, W. Liang, and Y. Jia, "Face pose estimation with combined 2d and 3d hog features," in *Proc. of International Conference on Pattern Recognition*, 2012, pp. 2492–2495. 3, 6

[45] A. Saeed and A. Al-Hamadi, "Boosted human head pose estimation using kinect camera," in *Proc. of IEEE International Conference on Image Processing*, 2015, pp. 1752–1756. 3, 6, 11

[46] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998. 3

[47] S. Kaymak and I. Patras, "Exploiting depth and intensity information for head pose estimation with random forests and tensor models," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 160–170. 3, 6

[48] S. Tulyakov, R.-L. Vieriu, S. Semeniuta, and N. Sebe, "Robust real-time extreme head pose estimation." in *Proc. of International Conference on Pattern Recognition*, 2014, pp. 2263–2268. 3

[49] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2094–2107, 2015. 3

[50] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2610–2617. 3, 6, 8, 11

[51] R. S. Ghiass, O. Arandjelović, and D. Laurendeau, "Highly accurate and fully automatic head pose estimation from a low quality consumer-level rgb-d sensor," in *Proc. of the 2nd Workshop on Computational Models of Social Interactions: Human-Computer-Media Communication*, 2015, pp. 25–34. 3

[52] Q. Cai, D. Gallup, C. Zhang, and Z. Zhang, "3d deformable face tracking with a commodity depth camera," in *Proc. of European Conference on Computer Vision*, 2010, pp. 229–242. 3

[53] A. Bleiweiss and M. Werman, "Robust head pose estimation by fusing time-of-flight depth and color," in *Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2010, pp. 116–121. 3

[54] S. Li, K. N. Ngan, R. Paramesran, and L. Sheng, "Real-time head pose tracking with online face template reconstruction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1922–1928, 2016. 3, 6

[55] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "3d face pose tracking using low quality depth cameras." in *VISAPP (2)*, 2013, pp. 223–228. 3, 6

[56] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013. 3, 5

[57] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, 2004. 3

[58] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Faceness-net: Face detection through deep facial part responses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017. 3

[59] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017. 3

[60] S. Chen, F. Bremond, H. Nguyen, and H. Thomas, "Exploring depth information for head detection with depth images," in *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*. IEEE, 2016, pp. 228–234. 3

[61] A. T. Nghiem, E. Auvinet, and J. Meunier, "Head detection using kinect camera and its application to fall detection," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*. IEEE, 2012, pp. 164–169. 4

[62] T. Ito and T. Kanade, "Predicting driver operations inside vehicles," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6. 4

[63] A. Datta, Y. Sheikh, and T. Kanade, "Linear motion estimation for systems of articulated planes," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8. 4

[64] C. Tran and M. M. Trivedi, "Introducing xmob: Extremity movement observation framework for upper body pose tracking in 3d," in *Proc. of IEEE International Symposium on Multimedia*. IEEE, 2009, pp. 446–447. 4

[65] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014. 4

[66] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 318–335. 4

[67] W. Zhang, Z. Shu, D. Samaras, and L. Chen, "Improving heterogeneous face recognition with conditional adversarial networks," *arXiv preprint arXiv:1709.02848*, 2017. 4

[68] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in neural information processing systems*, 2016, pp. 469–477. 4

[69] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International Conference on Artificial Neural Networks*. Springer, 2011, pp. 52–59. 4, 9

[70] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440. 4, 9

[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 5, 8

[72] M. Martin, F. v. d. Camp, and R. Stiefelhagen, "Real time head model creation and head pose estimation on consumer depth cameras," in *Proceedings of the 2014 2Nd International Conference on 3D Vision - Volume 01*, ser. 3DV '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 641–648. [Online]. Available: http://dx.doi.org/10.1109/3DV.2014.54 6

[73] B. Wang, W. Liang, Y. Wang, and Y. Liang, "Head pose estimation with combined 2d sift and 3d hog features," in *Image and Graphics (ICIG), 2013 Seventh International Conference on*. IEEE, 2013, pp. 650–655. 6

[74] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680. 5, 7

[75] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015. 5, 6

[76] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544. 6

[77] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2366–2374. 7, 9

[78] M. Fabbri, S. Calderara, and R. Cucchiara, "Generative adversarial models for people attribute recognition in surveillance," in *14th IEEE International Conference on Advanced Video and Signal based Surveillance*, 2017. 7, 9

[79] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241. 6, 9

[80] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 7

[81] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242. 7

[82] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814. 7

[83] G. Farneback, "Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field," in *Proc. of IEEE International Conference on Computer Vision*, vol. 1. IEEE, 2001, pp. 171–177. 7

[84] E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition," in *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–8. 7, 10

[85] A. D. Bagdanov, I. Masi, and A. Del Bimbo, "The florence 2d/3d hybrid face datset," in *Proc. of ACM Multimedia Int.l Workshop on Multimedia access to 3D Human Objects (MA3HO11)*. ACM Press, December 2011. 8

[86] 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, 2004. 8

[87] J. Nuevo, L. M. Bergasa, and P. Jiménez, "Rsmat: Robust simultaneous modeling and tracking," *Pattern Recognition Letters*, vol. 31, pp. 2455–2463, December 2010. 8

[88] K. Yuen, S. Martin, and M. M. Trivedi, "On looking at faces in an automobile: Issues, algorithms and evaluation on naturalistic driving dataset," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 2777–2782. 8

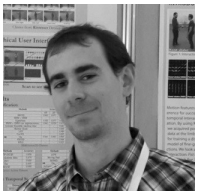
[89] S. Martin, K. Yuen, and M. M. Trivedi, "Vision for intelligent vehicles & applications (viva): Face detection and head pose challenge," in *Intelligent Vehicles Symposium (IV), 2016 IEEE*. IEEE, 2016, pp. 1010–1014. 8

[90] H. Sarbolandi, D. Lefloch, and A. Kolb, "Kinect range sensing: Structured-light versus time-of-flight kinect," *Comput. Vis. Image Und.*, vol. 139, pp. 1–20, 2015. 9

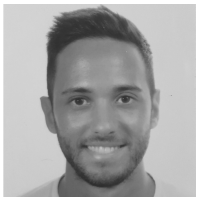
[91] G. T. Papadopoulos, A. Axenopoulos, and P. Daras, "Real-time skeleton-tracking-based human action recognition using kinect data," in *International Conference on Multimedia Modeling*, 2014, pp. 473–483. 9

[92] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012. 10

[93] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," *arXiv preprint arXiv:1604.06573*, 2016. 10



Guido Borghi received the master's degree in computer engineering from the University of Modena and Reggio Emilia in 2015. He is currently a Ph.D. candidate within the AImageLab group in Modena. His research topics are about Computer Vision and Deep Learning oriented to Human-Computer Interaction applications with 3D data.



Matteo Fabbri is currently a Ph.D. student at the International Doctorate School in ICT of the University of Modena e Reggio Emilia. He works under the supervision of Prof. Rita Cucchiara and Ing. Simone Calderara, on computer vision and deep learning for people behavior understanding. His research interests include generative models and multiple object tracking.

Roberto Vezzani Roberto Vezzani graduated in Computer Engineering in 2002 and received his Ph.D. course in Information Engineering in 2007 at the University of Modena and Reggio Emilia, Italy. Since 2016 is Associate Professor. His research interests mainly belong to computer vision systems for human computer interaction, video surveillance, with a particular focus on motion detection, people tracking and re-identification. He is a member of ACM, IEEE, and IAPR.

Simome Calderara received a computer engineering masters degree in 2005 and the Ph.D. degree in 2009 from the University of Modena and Reggio Emilia, where he is currently an assistant professor within the AImageLab group. His current research interests include computer vision and machine learning applied to human behavior analysis, visual tracking in crowded scenarios, and time series analysis for forensic applications. He is a member of the IEEE.

Rita Cucchiara received the masters degree in Electronic Engineering and the Ph.D. degree in Computer Engineering from the University of Bologna, Italy, in 1989 and 1992, respectively. Since 2005, she is a full professor at the University of Modena and Reggio Emilia, Italy, where she heads the AImageLab group and is Director of the SOFTECH-ICT research center. She is currently President of the Italian Association of Pattern Recognition, (GIRPR), affiliated with IAPR. She published more than 300 papers on pattern recognition computer vision and multimedia, and in particular in human analysis, HBU, and egocentric-vision. The research carried out spans on different application fields, such as video surveillance, automotive and multimedia big data annotation. Currently, she is AE of IEEE Transactions on Multimedia and serves in the Governing Board of IAPR and in the Advisory Board of the CVF.