Domain Translation with Conditional GANs: from Depth to RGB Face-to-Face

Matteo Fabbri Guido Borghi Fabio Lanzi Roberto Vezzani Simone Calderara Rita Cucchiara

Department of Engineering "Enzo Ferrari"
University of Modena and Reggio Emilia
via Vivarelli 10 Modena 41125, Italy
{name.surname}@unimore.it

***Abstract*—Can faces acquired by low-cost depth sensors be useful to catch some characteristic details of the face? Typically the answer is no. However, new deep architectures can generate RGB images from data acquired in a different modality, such as depth data. In this paper, we propose a new *Deterministic Conditional GAN*, trained on annotated RGB-D face datasets, effective for a face-to-face translation from depth to RGB. Although the network cannot reconstruct the exact somatic features for unknown individual faces, it is capable to reconstruct plausible faces; their appearance is accurate enough to be used in many pattern recognition tasks. In fact, we test the network capability to hallucinate with some *Perceptual Probes*, as for instance face aspect classification or landmark detection. Depth face can be used in spite of the correspondent RGB images, that often are not available due to difficult luminance conditions. Experimental results are very promising and are as far as better than previously proposed approaches: this domain translation can constitute a new way to exploit depth data in new future applications.**

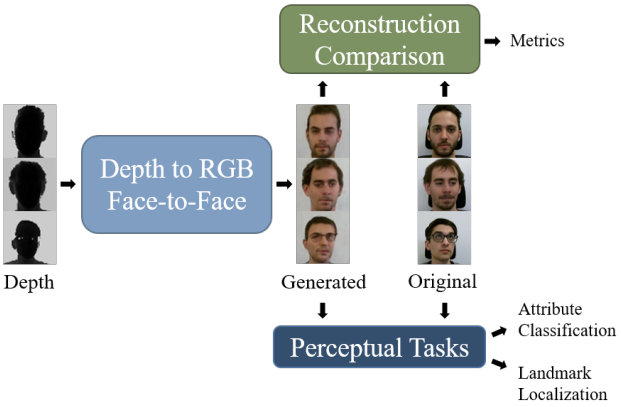

Fig. 1. Overview of Reconstruction Comparison and Probe Perceptual Tasks for performance evaluation.

I. Introduction

Generative Adversarial Networks (GANs) have been adopted as a viable and efficient solution for the Image-to-Image translation task, or rather the ability to transform images into other images across domains, according to a specific training set. Initially, Autoencoders, and in particular Convolutional Autoencoders [1], have been investigated and designed for several image processing tasks, such as image restoration [2], deblurring [3], and for image transformations such as image inpainting [4] or image style transformation. They have been used also as transfer learning mechanism for the Domain Transfer task: for sensor to image transformation [5] or from depth to gray-level images of faces [6]. As mentioned above, the Goodfellow's proposal of GANs [7] became the reference architecture for unsupervised generative modeling and for sampling new images from the underlying distribution of an unlabeled dataset by exploiting the joint capabilities of a Generative and a Discriminative Networks [8]. Furthermore, Conditional GANs [9], [10] provided conditional generative models by conditioning the sampling process with a partially observed input image. Several experiments show the power and effectiveness of conditional GANs, as for instance to improve resolution or to provide de-occlusion of images of people [10].

In this work, we explore the capability of face-to-face domain translation exploiting conditional GANs. The ability of a network to hallucinate and define a face aspect (in color or gray level), starting from a range map, could be a useful basic step for many computer vision and pattern recognition tasks, from biometric to expression recognition, from head pose estimation to interaction, especially in those contexts where intensity or color images cannot be recorded, for instance when shadows, light variations or darkness make the luminance and color acquisition not feasible enough. Our contribution is the definition of a *Conditional Generative Adversarial Network* that, starting from an annotated dataset with coupled depth and RGB faces (acquired by RGB-D sensors), learns to generate a plausible RGB face from solely the depth data. The network learns a proper transformation across the color and depth domains. Nevertheless, in generative settings, the result is likely a plausible face which could be qualitatively satisfactory (*e.g.*, the Discriminator network is fooled by it) but it is objectively difficult to properly measure the adherence to the conditioned input.
Therefore, another important contribution of our proposal is the adoption of some vision tasks as *Perceptual Probes* for performance evaluation, under the assumption that the domain translation task is viewed as an initial step of a more com-

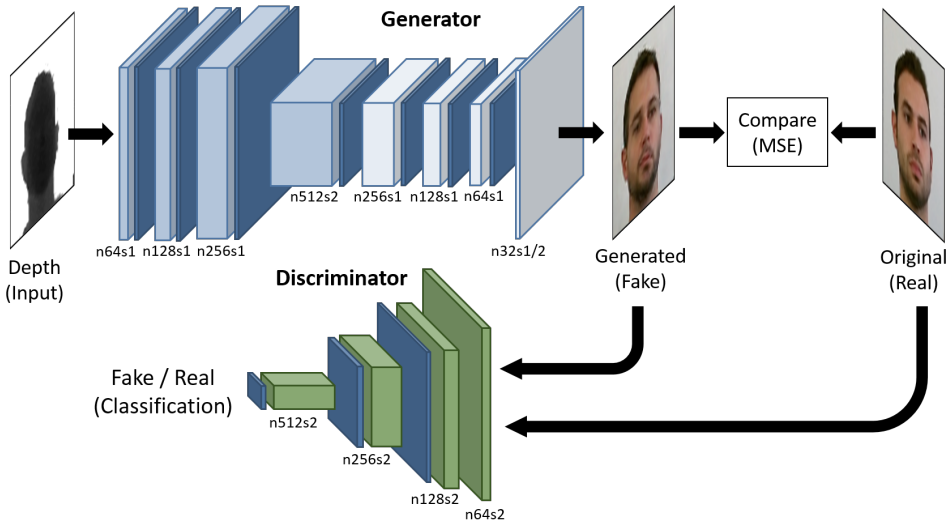

Fig. 2. Training schedule for Conditional GANs. The Discriminator learns to classify between generated fake images and real images while the Generator learns to fool the Discriminator. For each layer, the image provide information about number of filters (n) and stride (s).

plex visual recognition task. We assess that the face-to-face translation is acceptable if the new generated RGB face (from depth input) exhibits similar proprieties of other RGB-native faces in the selected probe perceptual tasks (*i.e.*, categorical attributes are maintained across domains). In accordance with this assumption, we will provide several experiments to test the proposed solution: we will use two different perceptual probes – namely, a network for face attribute classification and a method for landmark extraction – and we will evaluate how these tasks perform on generated faces. The overview of our Probe Perceptual Task is depicted in Figure 1. Results are really encouraging so that this approach could be a first attempt to "see and recognize faces in the dark", in analogy to how blind people captures the appearance only by touching a face and sensing the depth shape.

II. Related works

GAN for Image-to-Image translation. GANs have been defined very recently and tested in several contexts. Our work is inspired by the first idea of Goodfellow *et al.*, of Generative Adversarial Networks [7] with some variant in terms of conditional and discriminative GANs. GANs have been successfully used for Image-to-Image translation; they have been initially presented in [9] and then applied to some contexts as unpaired image to image translation [11]. A previous work starting from the same depth images has designed an Autoencoder to create gray-level faces from depth, with the final goal of head estimation [6]. An extension of this work is performed in [12] where a GAN is trained for the same final goal. In this paper, we compare our architecture with [9], using a similar dataset, other datasets, and with some probe perceptual tasks.

Depth Maps and Deep Learning. The latest spread of high-quality, cheap and accurate commercial depth sensors has encouraged the researchers of the computer vision community. Depth data are a useful source of information especially for systems that have to work in presence of darkness or dramatic light changes. Besides, recent depth sensors usually exploit infrared lights instead of lasers, so their use is safer for humans.

In the literature, the potentiality of depth images used as input for deep learning approaches has not been fully investigated yet. Only recently, Convolutional Neural Networks (CNNs) and depth maps have been exploited for various tasks, like head pose estimation [6], [14], facial landmark detection [13], head detection and obstacle detection [15]. Various type of deep architecture have been investigated, like LSTM [16] or Siamese networks [17]. The importance of this source of information is proved by the presence of works that aim to retrieve depth starting from monocular RGB images [18], [19].

III. Proposed Method

A general view of the proposed method is depicted in Figure 2. It consists of a GAN trained and tested on two different datasets, detailed in the following section.

GANs are generative models that learn a mapping from random noise vector z to output image y: $G : z \rightarrow y$ [7]. Conditional GANs instead are generative models introduced by [9] that learn a mapping from an observed image x and random noise z to an output image y: $G : \{x, z\} \rightarrow y$. Like GANs, they are composed of two components: a Generator G and a Discriminator D. The Generator G is trained to generate outputs that are indistinguishable from "real" by the adversarially trained Discriminator D which is trained to recognize the Generator's "fake" images from the "real" ones.

Using random noise as input, the generator G creates completely new samples, drawn from a probability distribution that approximates the distribution of the training data. This procedure leads to a non-deterministic behavior, that

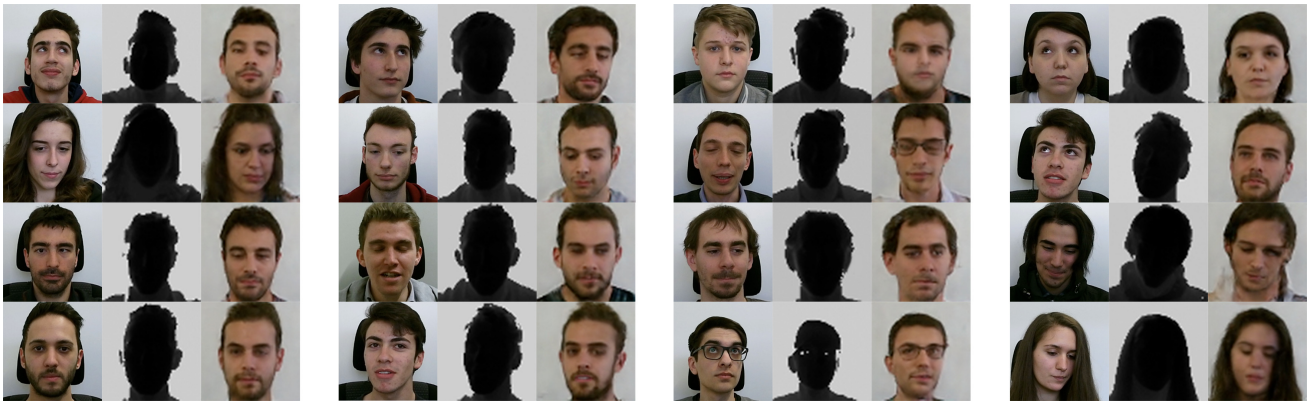

Fig. 3. Best results on the *MotorMark* [13] dataset. For each triplets of images: (Leftmost) the original image; (Middle) the input depth map; (Rightmost) the Generated face image.

is undesired for our goal. By removing the noise z, the probability distribution approximated by the model becomes a delta function with the property of preserving a deterministic behavior. Deterministic Conditional GANs (det-cGAN) thus learn a mapping from observed image x to output image y: $G : x \rightarrow y$.

A. Framework

The main goal is to train a generative function G capable of estimating the RGB face appearance I^{gen} from the corresponding depth input map I^{dpt} with the objective of reproducing the original image I^{rgb} associated with the depth map. To this aim, we train a Generator Network as a feed-forward CNN G_{θ_g} with parameters θ_g. For N training pairs images (I^{dpt}, I^{rgb}) we solve:

$$\hat{\theta}_g = \arg\min_{\theta_g} \frac{1}{N} \sum_{n=1}^{N} Loss_G \left(G_{\theta_g}(I_n^{dpt}), I_n^{rgb} \right). \quad (1)$$

We obtained $\hat{\theta}_g$ by minimizing the loss function defined at the end of this subsection. Following the det-cGAN paradigm we further define a Discriminator Network D_{θ_d} with parameters θ_d that we train alongside G_{θ_g} with the aim of solving the adversarial min-max problem:

$$\min_{\theta_g} \max_{\theta_d} \mathbb{E}_{I^{rgb} \sim p_{data}(I^{rgb})} [\log D(I^{rgb})] + \mathbb{E}_{I^{gen} \sim p_{gen}(I^{gen})} [\log 1 - D(G(I^{dpt}))] \quad (2)$$

where $D(I^{rgb})$ is the probability of I^{rgb} being a "real" image while $1 - D(G(I^{dpt}))$ is the probability of $G(I^{dpt})$ being a "fake" image. The main idea behind this min-max formulation is that it gives the possibility to train a generative model G with the target of fooling the discriminator D, which is adversarially trained to distinguish between generated "fake" images and "real" ones. With this approach, we achieve a generative model capable of learning solutions that are highly similar to "real" images, thus indistinguishable from the Discriminator D.

As a possible drawback, those solutions could be highly realistic thanks to D but unrelated to the input. A generated output could be a realistic face image with very different visual attributes and different pose with respect to the original image. This setup does not guarantee, for example, that a depth map of a girl with wavy hair looking to the right will generate an RGB image preserving those features. In order to tackle this problem, we mixed the Generator loss function $Loss_G$ with a more canonical loss such as MSE. Borrowing the idea from [10], we propose a Generator loss that is a weighted combination of two components:

$$Loss_G = \lambda Loss_{MSE} + Loss_{adv} \quad (3)$$

where $Loss_{MSE}$ is calculated using the mean squared errors of prediction (MSE) which measure the discrepancy between the generated image I^{gen} and the ground truth image I^{rgb} associated with the corresponding input depth map I^{dpt}. The MSE component is subject to a multiplication factor λ which controls its impact during training. The $Loss_{adv}$ component is the actual adversarial loss of the framework which encourages the Generator to produce perceptually good solutions that reside in the manifold of face images. The loss is defined as follows:

$$Loss_{adv} = \sum_{n=1}^{N} -\log(D(G(I^{dpt}))) \quad (4)$$

where $D(G(I^{dpt})$ is the probability of the Discriminator labeling the generated image $G(I^{dpt})$ as being a "real" image. Rather than training the Generator to minimize $\log(1 - D(G(I^{dpt})))$ we train G to minimize $\log(D(G(I^{dpt})))$. This objective provides strongest gradients early in training [7]. The combination of those two component grants the required behavior: the Generator has not only to fool the Discriminator but has to be near the ground truth output in an MSE sense.

B. Architecture

The task of Image-to-Image translation can be expressed as finding a mapping between two images. In particular, for the specific problem we are considering, the two images share the same underlying structure despite differing in surface appearance. Therefore, the structure in the input depth image is

TABLE I

EVALUATION METRICS COMPUTED ON THE GENERATED RGB FACE IMAGES WITH *MotorMark* DATASET. STARTING FROM LEFT ARE REPORTED $L1$ AND $L2$ DISTANCES, ABSOLUTE AND SQUARED ERROR DIFFERENCES, ROOT-MEAN-SQUARED ERROR AND FINALLY THE PERCENTAGE OF PIXELS UNDER A DEFINED THRESHOLD. FURTHER DETAILS ARE REPORTED IN [20]

Method	**Norm** ↓		**Difference** ↓		**RMSE** ↓			**Threshold** ↑		
	L_1	L_2	Abs	Squared	linear	log	scale-inv	1.25	2.5	3.75
Autoencoder	39.80	6327	2.21	273.33	58.74	1.248	1.791	1.389	1.878	2.120
pix2pix [9]	37.77	6150	2.06	253.11	56.01	1.240	1.846	1.400	1.882	2.157
Our	**37.12**	**6021**	**2.05**	**245.88**	**54.86**	**1.222**	**1.749**	**1.423**	**1.914**	**2.188**
Our (Binary Maps)	43.58	6868	2.45	320.93	62.53	1.320	1.830	1.319	1.778	2.047

roughly aligned with the structure in the output RGB image. In fact, both images are representing the same subject in the same pose thus details like mouth, eyes, and nose share the same location through the two images. The generator architecture was designed following those considerations.

A recent solution [9] to this task adopted the "U-Net" [21] architecture with skip connections between mirrored layers in the encoder and decoder segments in order to shuttle low-level information between input and output directly across the network. We found this solution less profitable because the strictly underlying structural coherence between input and output makes the network use the skip connections to jump at easier but not optimal solutions and ignoring the main network flow.

Consequently, our architecture implementation follows the *FfD* implementation in [12]. We relaxed the structure of the classical hourglass architecture performing less upsampling and downsampling operations in order to preserve the structural coherence between input and output. We found that using the half of feature maps described in [12] at each layer in both Generator and Discriminator networks sped up the training without a significant reduction of qualitative performance.

We propose the Generator's architecture depicted in Figure 2. Specifically, in the encoder, we used three convolutions followed by a strided convolution (with stride 2, in order to reduce the image resolution). The decoder uses three convolutions followed by a fractionally strided convolution (also known in literature as transposed convolutions) with stride $1/2$ to increase the resolution, and a final convolution. Leaky ReLU is adopted as activation function in the encoding stack while ReLU is preferred in the decoding stack. Batch normalization layers are adopted before each activation, except for the last convolutional layer which uses the Tanh activation. The number of filters follows a power of 2 pattern: from 64 to 512 in the encoder and from 256 to 32 in the decoder. All convolutions use a kernel of size 5×5. The Discriminator architecture is similar to the Generator's encoder in terms of number of filters and activations functions but uses only strided convolutional layers with stride 2 to halve the image resolution each time the number of filters is doubled. The resulting 512 feature maps are followed by one sigmoid activation to obtain a probability useful for the classification problem.

C. Training Details

We trained our det-cGAN with 64×64 resized depth maps as input and simultaneously providing the original RGB images associated with the depth data in order to compute the MSE loss. We adopted the standard approach in [7] to optimize the network alternating gradient descent updates between the generator and the discriminator with K = 1. We used mini-batch SGD applying the *Adam* solver with momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. In our experiments we chose a λ value of 10^1 in Equation (3) and a batch size of 64. Some best results are presented in Figure 3.

IV. EXPERIMENTS

Generally, evaluating the quality of reconstructed images is still an open problem, as reported in [9]. Traditional metrics such as $L1$ distance are not sufficient to assess joint statistic on the produced images, and therefore do not extrapolate the full structure of the result. In order to more holistically investigate

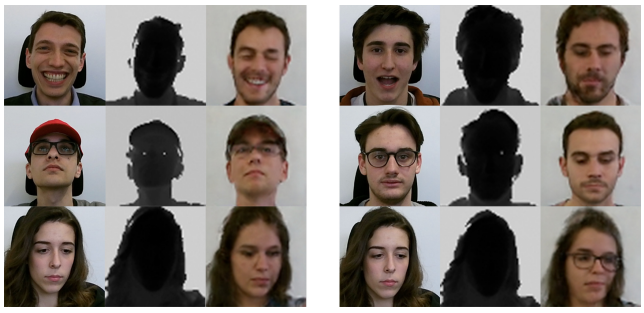

Fig. 4. Visual examples of generated images that preserve (left column) and do not preserve (right column) some attributes.

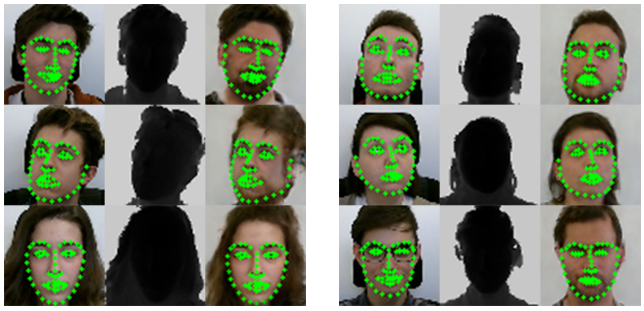

Fig. 5. Visual examples of landmark predictions on real and generated images.

TABLE II
PER ATTRIBUTE CONCORDANCE BETWEEN THE TRUE RGB FACE AND THE HALLUCINATED ONE USING VGG-FACE CNN.

Attribute	Accuracy	Precision	Recall	F1
Male	90.30	95.51	93.49	94.49
Young	93.01	97.69	95.09	96.37
Mouth Open	82.86	92.16	51.07	65.71
Smiling	96.25	99.54	66.48	79.72
Wearing Hat	98.40	99.38	58.05	73.29
Wavy Hair	98.46	95.28	48.44	64.24
No Beard	48.18	63.89	40.88	49.86
Straight Hair	79.12	07.78	57.76	13.71
Eyeglasses	80.12	24.91	08.14	12.27

the capabilities of our network to synthesize RGB face images directly from depth maps, a reconstruction comparison and two perceptual probes are performed. Firstly, we compared the performance of the proposed model with other *Image-to-Image* recent methods present in the literature, through metrics directly calculated over the reconstructed images. Secondly, we measured the capability of the proposed network of being able to preserve original facial attributes, like wearing hat and smiling, by exploiting a classification network trained with RGB face images. Thirdly, we measured whether or not reconstructed RGB face images are realistic enough that an off-the-shelf landmark localization system is able to localize accurate key-points. Eventually, in order to investigate how much the depth map information impacts the reconstruction task, we repeated the previous experiments testing our network trained with binary maps derived from the original depth maps.

A. Datasets

Experiments are conducted exploiting two publicly available datasets: *Pandora* [6] and *MotorMark* [13]. *Pandora* contains more than 250k frames, splitted into 110 annotated sequences of 22 different actors (10 males and 12 females), while *MotorMark* is composed of more than 30k frames of 35 different subjects, guaranteeing a great variety of face appearances. Subjects can wear garments and sunglasses and may perform driving activities actions like turning the steering wheel, adjust the rear mirror and so on. Both datasets have been acquired with a *Microsoft Kinect One*. In our

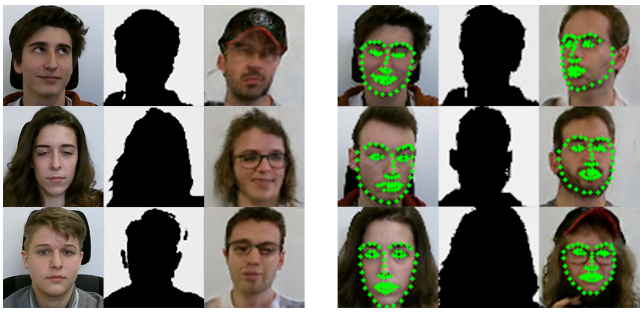

Fig. 6. Visual examples of issues using binary maps instead of depth maps: attributes are not preserved (left column) and landmark localization is not precise (right column).

TABLE III
QUANTITATIVE COMPARISON ABOUT THE AVERAGE ATTRIBUTES CONCORDANCE BETWEEN TRUE AND HALLUCINATED RGB FACES.

Method	Accuracy	Precision	Recall	F1
Autoencoder	75.21	61.84	40.55	51.38
pix2pix	84.57	74.42	56.01	60.78
Our	**85.19**	**75.13**	**57.71**	**61.07**
Our (Binary Maps)	60.51	49.48	29.12	42.76

experiments, *Pandora* has been used as the training set and *MotorMark* as the test set, performing a cross-dataset validation of the proposed method.

B. Reconstruction Comparison

Here, we check the capabilities of the proposed network to reconstruct RGB images from the correspondent depth ones. We exploited the metrics described in [22]: these metrics were originally used to evaluate depth images generated from RGB image sources. Results are reported in Table I. In particular, we compared our generative method with two other techniques: an Autoencoder trained with the same architecture as our Generator network, and *pix2pix* [9], a recent work that exploits the Conditional GAN framework. In the last line of Table I, is also reported the comparison with our network trained on binary maps, detailed at the end of this section. As shown, results confirm the superior accuracy of the presented method.

C. Attribute Classification

In the previous section, we checked the overall quality of the reconstructed RGB images. Here, we focus on the capability of our network to generate face images that specifically preserve the facial attributes of the original person. To this end, we exploited a pre-trained network, the *VGG-Face* CNN [23], trained on RGB images for face recognition purposes. In order to extrapolate only the attributes that can be carried by depth information, we fine-tuned the network with the *CelebA* Dataset [24].

By observing Table II, it is evident the good capability of the network to preserve gender, age, pose, and appearance attributes. Nevertheless, the depth sensor resolution fails at modeling hair categories such as curly or straight and glasses since such details are not always correctly captured in terms of depth. Glasses lenses, for example, are neglected by IR sensors and significantly captured only when the glasses structure is solid and visible. In all the other cases they tend to be confused by the network with the ocular cavities. Nonetheless, Table III exhibits the superiority of our proposal against state of the art generative networks also in attribute preservation. Moreover in Figure 4 are presented both successful and failure cases.

D. Landmark Localization

The intuition behind this experiment is that if the synthesized images are realistic and accurate enough, then a landmark localization method trained on real images will be able to localize key-points also on the generated images. To

TABLE IV
QUANTITATIVE COMPARATIVE RESULTS OF OUR PROPOSAL AGAINST THE AUTOENCODER AND PIX2PIX BASELINES IN TERMS OF FACE DETECTOR ACCURACY AND LANDMARK LOCALIZATION.

Method	**Accuracy**	$L2$ **Norm**
Autoencoder	54.03	2.219
pix2pix	85.21	2.201
Our	**86.86**	**2.089**
Our (Binary Maps)	62.37	2.980

this aim, we exploited the algorithm included in the *dLib* libraries [25], which gives landmark positions on RGB images. In Figure 5 qualitative examples that highlight the coherence of landmark predictions between original and generated images are presented. The last column of Table IV reports, for each method, the average $L2$ Norm between the position of landmarks predicted and the ground truth provided by the dataset. The results show that our method is able to produce outputs that can fool an algorithm trained on RGB face images.

E. Binary Maps

An ablation study is conducted to investigate the importance of depth information, by training our network providing as input binary maps instead of depth maps. Binary maps were gathered thresholding the depth maps. Figure 6 shows examples where the reconstructed face images are not coherent with the original images in terms of attributes preservation and landmark position. At the end of Tables I, III and IV are reported the results of the previous experiment where we used binary maps instead of depth maps. Results show that the depth information has a fundamental importance in the face generation task, to preserve coherent facial attributes and head pose orientation.

V. CONCLUSION

In this paper, a deterministic conditional GAN to reconstruct RGB face images from the correspondent depth one is presented. Experimental results confirm the ability of the proposed method to generate accurate faces, to preserve facial attributes and to maintain coherency in facial landmarking. Besides, we checked and shown the importance of depth data in all these tasks. Various future works can be planned, due to the flexibility and the accuracy of the presented method. For instance, it is possible to investigate how to generate or delete specific face attributes, or how the enhance the training capabilities of depth maps.

ACKNOWLEDGMENT

This work has been carried out within the projects "COSMOS Prin 2015" supported by the Italian MIUR, Ministry of Education, University and Research 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 CINECA for the availability of high-performance computing resources.

REFERENCES

[1] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," *Artificial Neural Networks and Machine Learning–ICANN 2011*, pp. 52–59, 2011.
[2] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections," *arXiv preprint arXiv:1606.08921*, 2016.
[3] S. A. Bigdeli and M. Zwicker, "Image restoration using autoencoding priors," *arXiv preprint arXiv:1703.09964*, 2017.
[4] C. Guillemot and O. Le Meur, "Image inpainting: Overview and recent advances," *IEEE signal processing magazine*, vol. 31, no. 1, 2014.
[5] M. S. Singh, V. Pondenkandath, B. Zhou, P. Lukowicz, and M. Liwicki, "Transforming sensor data to the image domain for deep learning-an application to footstep detection."
[6] 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.
[7] 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.
[8] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
[9] 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.
[10] M. Fabbri, S. Calderara, and R. Cucchiara, "Generative adversarial models for people attribute recognition in surveillance," in *14th IEEE International Conference on AVSS*, 2017.
[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017.
[12] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara, "Face-from-depth for head pose estimation on depth images," *arXiv preprint arXiv:1712.05277*, 2017.
[13] E. Frigieri, G. Borghi, R. Vezzani, and R. Cucchiara, "Fast and accurate facial landmark localization in depth images for in-car applications," in *International Conference on Image Analysis and Processing*, 2017.
[14] M. Venturelli, G. Borghi, R. Vezzani, and R. Cucchiara, "Deep head pose estimation from depth data for in-car automotive applications," *2nd International Workshop on Understanding Human Activities through 3D Sensors (UHA3DS'16)*, 2016.
[15] C.-H. Lee, Y.-C. Su, and L.-G. Chen, "An intelligent depth-based obstacle detection system for visually-impaired aid applications," in *International Workshop on Image Analysis for Multimedia Interactive Services*, 2012.
[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[17] M. Venturelli, G. Borghi, R. Vezzani, and R. Cucchiara, "From depth data to head pose estimation: a siamese approach," 2017.
[18] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Advances in neural information processing systems*, 2006.
[19] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2009.
[20] 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*. Curran Associates, Inc., 2014.
[21] 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.
[22] 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*.
[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.
[24] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.
[25] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.