Learning to Generate Facial Depth Maps

Stefano Pini, Filippo Grazioli, Guido Borghi, Roberto Vezzani, Rita Cucchiara
University of Modena and Reggio Emilia, Italy
name.surname@unimore.it

Abstract

In this paper, **an adversarial architecture** for **facial depth map estimation** from monocular intensity images is presented.

By following an *image-to-image* approach, we combine the advantages **of supervised learning** and **adversarial training**, proposing a **conditional Generative Adversarial Network** that effectively learns to translate intensity face images into the corresponding depth maps.

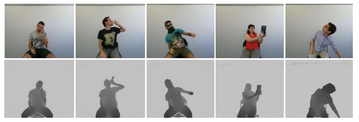
Furthermore, we show that the model is capable of **predicting distinctive facial details** by testing the generated depth maps through a deep model trained on authentic depth maps for **the face verification task.**

Datasets

Pandora dataset

This dataset has been presented in [1] and it has the following features:

- **250k frames** from **110 sequences**
- **22 subjects** (10 males and 12 females).
- Both **depth** and **RGB** frames and **skeleton** annotations frame by frame.
- Acquired by **Microsoft Kinect One v2**
- Original tasks: **head and shoulder pose estimation**. Subjects can vary their head appearance wearing **garments** and **objects** like smartphones, tablets and plastic bottles that can generate **head and body occlusions**.



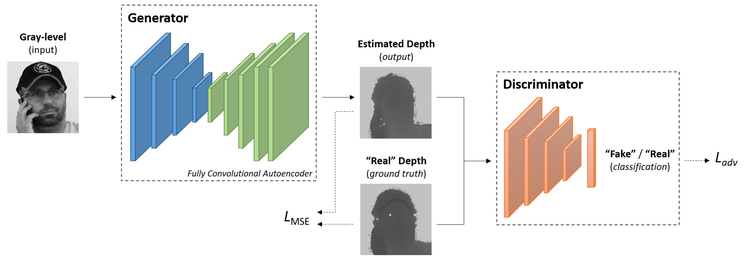
Biwi Kinect Head Pose

This dataset [2] contains:

- **15k frames**
- **20 subjects** (14 males and 6 females)
- Both **depth** and **RGB** frames
- Acquired by **Microsoft Kinect v1**
- Original task: head pose estimation only

Proposed Method

We propose a **Conditional Generative Adversarial Network** that effectively learns to translate intensity face images into the corresponding depth maps.



Dynamic Face Crop → a new window of size (w_h, h_H) for each frame

$f_{x,y}$: focal lengths D: estimated distance $R_{x,y}$: face size

$$w_H = \frac{f_x \cdot R_x}{D} \qquad h_H = \frac{f_y \cdot R_y}{D}$$

Input: RGB images (100x100)
Ouput: depth maps (100x100 8-bit images)
Loss Functions:

$$L_{MSE}(\mathbf{s}^g, \mathbf{s}^d) = \frac{1}{N}\sum_{i=1}^{N} \|G(s_i^g) - s_i^d\|_2^2$$

$$L_{adv}(\mathbf{y}, \mathbf{t}) = -\frac{1}{N}\sum_{i=1}^{N} [t_i \log y_i + (1 - t_i)\log(1 - y_i)]$$

Activation Functions:

- *ReLU + tanh*

Network Architecture:

- Generator: *Fully Convolutional Network*
- Discriminator: *conv + fc* network

Face Verification Accuracy

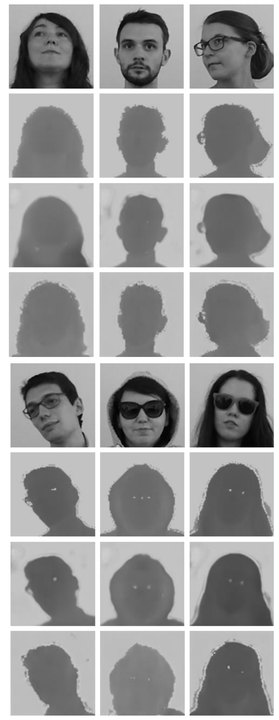
We create **subsets of the Pandora dataset** to test the *Face Verification* accuracy.
$\{S_1, S_2, S_3\}$: actions performed with constrained movements
$\{S_4, S_5\}$: complex movements, occlusions and challenging camouflage
$\{A_1\} = \{s_{\varrho\theta\sigma} \mid \forall\gamma \in \{\varrho, \sigma, \theta\}: -10° \leq \gamma \leq 10°\},\ \varrho = roll,\ \sigma = pitch, \theta = yaw$
$\{A_2\} = \{s_{\varrho\theta\sigma} \mid \exists\gamma \in \{\varrho, \sigma, \theta\}: \gamma < -10° \vee \gamma > 10°\},\ \varrho = roll,\ \sigma = pitch, \theta = yaw$

	$\{S_i\}_{i=1,2,3}$		$\{S_i\}_{i=4,5}$		$\{S_i\}_{i=1,2,3,4,5}$	
	original	*generated*	*original*	*generated*	*original*	*generated*
A_1	0.8184	0.8614	0.7685	0.7155	0.7917	0.7950
A_2	0.7928	0.7499	0.7216	0.6586	0.7576	0.7007
$\{A_1, A_2\}$	0.8034	0.7851	0.7271	0.6696	0.7664	0.7247

Pixel-wise Metrics

Metrics comparison between our *cGan*, *Autoencoder* (AE), *pix2pix* [4] and Cui *et al.* [5]

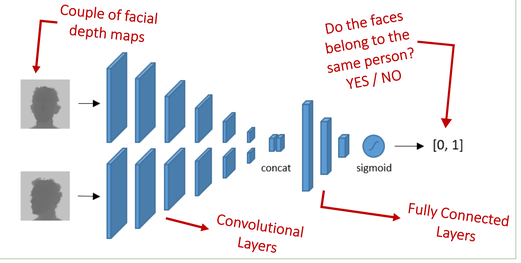
		Pandora				Biwi			
Metrics		*cGAN*	*AE*	pix2pix	Cui et al.	*cGAN*	*AE*	pix2pix	Cui et al.
L_1 Norm	↓	**11.792**	16.185	18.172	19.046	10.503	**10.444**	47.191	16.507
L_2 Norm	↓	**1,678.2**	2,224.8	3,109.0	2,093.3	2,368.5	2,342.5	6,661.3	**2,319.8**
Absolute Diff	↓	**0.1019**	0.1441	0.1512	0.1465	**0.1838**	0.1936	0.9062	0.2836
Squared Diff	↓	**2.9974**	5.3891	8.6444	3.9084	**8.7122**	9.0332	100.89	9.3032
RMSE_{lin}	↓	**18.677**	25.213	33.526	22.599	24.865	24.699	72.084	**24.521**
RMSE_{log}	↓	**0.1744**	0.2752	1.0864	0.2105	**0.2932**	0.2970	1.2240	0.3390
$\text{RMSE}_{\text{scale-inv}}$	↓	0.1345	0.2018	1.0774	**0.1301**	0.2687	**0.2642**	1.1759	0.2867
$\delta < 1.25$	↑	**0.8529**	0.6854	0.7802	0.7556	**0.7393**	0.7230	0.4149	0.6395
$\delta < 1.25^2$	↑	**0.9642**	0.8728	0.8978	0.9554	**0.9224**	0.9064	0.5298	0.7943
$\delta < 1.25^3$	↑	0.9915	0.9651	0.9638	**0.9935**	**0.9609**	0.9557	0.6360	0.9311
$\delta < 1.25^{\frac{1}{4}}$	↑	**0.5619**	0.4763	0.5053	0.0740	0.5293	**0.5532**	0.3128	0.3580
$\delta < 1.25^{\frac{1}{3}}$	↑	**0.6113**	0.5047	0.5615	0.1036	0.5727	**0.6034**	0.3274	0.4549
$\delta < 1.25^{\frac{1}{2}}$	↑	**0.6954**	0.5547	0.6496	0.5509	0.6348	**0.6606**	0.3513	0.5189
Face Verification	↑	**0.7247**	0.6570	0.5315	0.6442	**0.6251**	0.6043	0.5422	0.5966



Sample outputs of the proposed method
[RGB, Depth, AE and cGAN]

Face Verification test

We introduce a **Face Verification *Siamese*** architecture [3], trained on the original face depth images, to check if the generated images maintain the facial **distinctive features** of the original subjects, **not only when visually inspected by humans, but also when processed by deep convolutional network**.



Limitations

- The detail accuracy of the proposed model is quite good, compared with the tested competitors;
- δ *-metrics* commonly used ($\delta < 1.25, \delta < 1.25^2$, $\delta < 1.25^3$) [6, 7], are effective to check the overall quality of depth maps generated from landscapes or wide-angle scenes, but **the threshold value is too high to take fine details into account**;
- We introduce **a new set of $\boldsymbol{\delta}$*-metrics*** ($\delta < 1.25^{\frac{1}{2}}$, $\delta < 1.25^{\frac{1}{3}}$, $\delta < 1.25^{\frac{1}{4}}$) with harder thresholds

Detail Accuracy is still an open problem with Conditional GANs!

Finally, we note that our approach is able to produce overall accurate views of the generated facial depth maps, **preserving the shape of the face and the garments.**

Check the Code and Models

Download Pandora Dataset

3DV 2018 International Conference on 3DVision
Verona, Italy
September 5 - 8, 2018

References

[1] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara "Poseidon: Face-from-depth for driver pose estimation" (CVPR 2017)
[2] G. Fanellli, M. Dantone, J. Gall, A. Fossati, L. Van Gool "Random Forests for Real Time 3D Face Analysis" (IJCV 2013)
[3] G. Borghi, S. Pini, F. Grazioli, R. Vezzani, R. Cucchiara "Face Verification from Depth using Privileged Information" (BMVC 2018)
[4] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. "Image-to-image translation with conditional adversarial networks." (CVPR 2017)
[5] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen. "Improving 2d face recognition via discriminative face depth estimation" (ICB 2018)
[6] A. Atapour-Abarghouei and T. Breckon. "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer" (CVPR 2018)
[7] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci. "Structured attention guided convolutional neural fields for monocular depth estimation" (CVPR 2018)

Acknowledgements

This work has been carried out within the project "*FAR2015 - Monitoring the car driver's attention with multisensory systems, computer vision and machine learning*" funded by the University of Modena and Reggio Emilia. We also acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support