Towards Automatic Image Editing: Learning to See another You

Amir Ghodrati[3,1]
http://homes.esat.kuleuven.be/~aghodrat/
Xu Jia[3,1]
http://homes.esat.kuleuven.be/~xjia/
Marco Pedersoli[2]
marco.pedersoli@inria.fr
Tinne Tuytelaars[1]
http://homes.esat.kuleuven.be/~tuytelaa/

[1] ESAT-PSI
KU Leuven, iMinds
Leuven, Belgium

[2] THOTH
INRIA Grenoble
Grenoble, France

[3] equal contribution to this work and listed in alphabetical order

Problem Definition. We propose a method that aims at automatically editing an image by altering its attributes. More specifically, given an image of a certain class (*e.g.* a human face), the method should generate a new image as similar as possible to the given one, but with an altered visual attribute (*e.g.* the same face with a new pose or a different illumination).

Contributions. The main contributions of this paper are: i) definition of a new problem, where the goal is to generate images as similar as possible to a source image yet with one attribute changed; ii) a solution that follows an encoder-decoder pipeline, where the desired attribute modification is first encoded then integrated at feature map level; iii) the insight that the result can be refined by adding another convolutional encoder-decoder model; and iv) good qualitative and quantitative results on different tasks on MultiPIE dataset.

How. We propose a model following the encoder-decoder fashion. It takes a face image as input and encodes it into several feature maps; takes a desired attribute vector as input and encodes it into several feature maps; then combines and deeply fuse these two flows of information; finally generates a new image with a convolutional decoder module. The image output of this network produces already a reasonable result, but it still has some missing details and some artifacts. Therefore, we adopt a coarse-to-fine scheme, dividing the problem in two stages. In the second stage, we add another convolutional encoder-decoder network to refine the previously generated image which takes as input the source image and the generated image of the first stage. In summary, the first stage is in charge of rendering a global representation of the desired object, while the second stage focuses on local refinements to remove some artifacts.

Evaluation. We evaluate our method on three different tasks on the MultiPIE [1] dataset. The main task is to rotate a face. We extensively evaluate our method for this task, showing both qualitative and quantitative results which is measured in terms of per-pixel mean squared error (MSE) between generation and ground-truth image. Our method shows better performance when compared with the method of [2]. The other two tasks are generating faces with different illumination and filling in the missing part of a face image on synthetic data generated from MultiPIE.

Figure 1: Qualitative results of our image generation from test data of MultiPIE. From left to right are the input image, the ground-truth target image, the output of the first stage and the output of the second stage.

Figure 2: Qualitative results for the task of image inpainting. From left to right are the input image, images generated with our method and the original images without the occluding pattern.

[1] Ralph Gross, Iain Matthews, Jeffrey F. Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image Vision Comput.*, 28(5):807–813, 2010.

[2] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Du-Sik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *CVPR*, 2015.