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

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:

- 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
- A solution that follows an encoder-decoder pipeline
- The insight that the result can be refined by adding another convolutional encoder-decoder model
- Good qualitative and quantitative results on different tasks

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 second stage, e propose to add another convolutional encoder-decoder network to perform image refinement. The second stage takes as input the source image and the generated image of the first stage. These two inputs are first concatenated channel-wise. Then we apply several convolutional, ReLU and max-pooling layers in the encoding process followed by unpooling, convolutional and ReLU layers in the decoding process. In summary, The first stage is in charge of rendering a global representation of the desired object, while the second focuses on local refinements to remove some artifacts.

Dataset. We evaluate our method on the MultiPIE [1] dataset. MultiPIE is a large face dataset with a variety of attributes (e.g. pose and illumination) annotated and varied systematically for each individual in the database. We crop faces to 60×60 and align them for our method.

Tasks. We evaluate our method for three different tasks. The main task is to rotate the face. We extensively evaluate our method for this task, showing both qualitative and quantitative results. The other two tasks are generating faces with different illumination and filling in the missing part for a face image on synthetic data generated from MultiPIE.

Rotating Face. In this task we have 7 different poses ($-45°$ to $+45°$). The input to our method is an image and one target pose vector out of 7 possible poses. In Figure 1 we show some qualitative results. Also, we quantitatively validate the effectiveness of our method in terms of per-pixel mean squared error (MSE) between generation and ground-truth image in Table 1. We compare our method with the method of [2] as well on a subset of test faces.

First stage	Second stage	CPI [2]
578.5	570.5	884.4

Table 1: Mean Squared Error on a subset of 510 images with neutral illumination, for a fair comparison with [2].

Figure 1: Some qualitative results of our image generation from test data of MultiPIE. In each row, first column is input image, last column is ground-truth target image, 2nd column is output of first stage and 3rd column is generated image of second stage network.

Changing illumination. We train another model to generate a face with a specific illumination out of 20 different illumination conditions. Some examples of the generated faces are shown in Figure 2. Quantitatively, the per-pixel MSE of test set is 193.3 and 146.6 for the first and second stage respectively which indicate this task is easier than rotating faces.

Figure 2: Qualitative results for the task of changing illumination. The most right image is input face. For each identity, the first row shows 20 faces generated under different illuminations and second row is the corresponding ground-truth face.

Image Inpainting. For this task, we randomly generate 10 black blocks of different shapes as occlusion patterns which is overlaid on the face image at a random location. We train the proposed model to learn to inpaint the occluded face image. As shown in Figure 3 our method can generate reasonably good images also for this task considering the high variability of the input image. The model fills in the occluded region using the knowledge it learned during training such as the continuity of local region and the symmetry within the face.

Figure 3: Qualitative results for the task of image inpainting. The first column shows the input image, the second column shows images generated with our method and the third column shows the complete image 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.