Eye In-Painting with Exemplar Generative Adversarial Networks

Hongzhi Liu

August 24, 2018

Abstract

According to the statistics, around 300M pictures are captured and shared in social networks with a large percentage of them featuring people-centric content every day. Today I read a thesis written by Brian Dolhansky and Cristian Canton Ferrer who is from Facebook Inc.. This paper introduces a novel approach to in-painting where the identity of the object to remove or change is preserved and accounted for at inference time: Exemplar GANs (ExGANs). And they propose using exemplar information in the form of a reference image of the region to in-paint, or a perceptual code describing that object. Besides, the team show that ExGANs can produce photo-realistic personalized in-painting results that are both perceptually and semantically plausible by applying them to the task of closed-to-open eye in-painting in natural pictures.

1. Overview of Exemplar GANs

Previous approaches to opening closed eyes in photographs have generally used example photos, such as a burst of photographs of a subject in a similar pose and lighting conditions [1], and produced final results with a mixture of patch matching [2] and blending [6]. However, this technique does not take full advantage of semantic or structural information in the image, such as global illumination or the pose of the subject. Small variations in lighting or an incorrect gaze direction produce uncanny results, as seen in Fig. 1.

Recently, deep convolutional networks (DNNs) have produced high-quality results when in-painting missing regions of pictures showing natural scenery [4]. For the particular problem of facial transformations, they learn not only to preserve features such global lighting and skin tone, but can also encode some notion of semantic plausibility. Given a training set of sufficient size, the network will learn what a human face "shoul" look like [5], and will in-paint accordingly, while preserving the overall structure of the face image.

In this paper, Brian Dolhansky and Cristian Canton Ferrer focus on the particular problem of eye in-painting. While DNNs can produce semantically plausible, realistic-looking results, most deep techniques do not preserve the identity of the person in a photograph. For instance, a DNN could learn to open a pair of closed eyes, but there is no guarantee encoded in the model itself that the new eyes will correspond to the original person's specific ocular structure. Instead, DNNs will insert a pair of eyes that correspond to similar faces in the training set, leading to undesirable and biased results; if a person has some distinguishing feature, this will not be reflected in the generated part.

Figure 1. Comparison between the commercial state of the art eye opening algorithm in Adobe Photoshop Elements (c) and the proposed ExGAN technique (d). The exemplar and original images are shown in (a) and (b), respectively.

The motivation for the use of exemplar data is twofold. First, by utilizing extra information, ExGANs do not have to

hallucinate textures or structure from scratch, but will still retain the semantics of the original image. Second, output images are automatically personalized. For instance, to in-paint a pair of eyes, the generator can use another exemplar instance of those eyes to ensure the identity is retained.

2. Exemplar GANs for in-painting

Brian Dolhansky and Cristian Canton Ferrer propose two separate approaches to ExGAN in-painting. The first is reference-based in-painting, in which a reference image r_i is used in the generator as a guide, or in the discriminator as additional information when determining if the generated image is real or fake. The second approach is code-based in-painting, where a perceptual code c_i is created for the entity of interest. For eye in-painting, this code stores a compressed version of a person's eyes in a vector $c_i \in \mathbb{R}^N$, which can also be used in several different places within the generative and discriminator networks.

2.1. Reference image in-painting

Assume that for each image in the training set x_i, there exists a corresponding reference image r_i. Therefore the training set X is defined as a set of tuples $X = (x_1, r_1), \ldots, (x_n, r_n)$. For eye in-painting, r_i is an image of the same person in x_i, but potentially taken in a different pose. Patches are removed from x_i to produce z_i , and the learning objective is defined as Eq. 1:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x_i, r_i \sim p_{data}(x,r)}[\log D(x_i, r_i)] + \mathbb{E}_{r_i \sim p_r, G(\cdot) \sim p_z}[\log 1 - D(G(z_i, r_i))] + \| G(z_i, r_i) - x_i \|_1 \quad (1)$$

This objective is similar to the standard GAN formulation in [3], but both the generator and discriminator can take an example as input. For better generalization, a set of reference images R_i corresponding to a given x_i can also be utilized, which expands the training set to the set of tuples comprised of the Cartesian product between each image-to-be-in-painted and its reference image set, $X = x_1 \times R_1, \ldots, x_n \times R_n$.

2.2. Code in-painting

For code-based in-painting, and for datasets where the number of pixels in each image is $|I|$, assume that there exists a compressing function $C(r) : \mathbb{R}^{|I|} \rightarrow \mathbb{R}^N$, where $N \ll |I|$. Then, for each image to be in-painted z_i and its corresponding reference image r_i, a code $c_i = C(r_i)$ is generated using a r_i. Given the codified exemplar information,

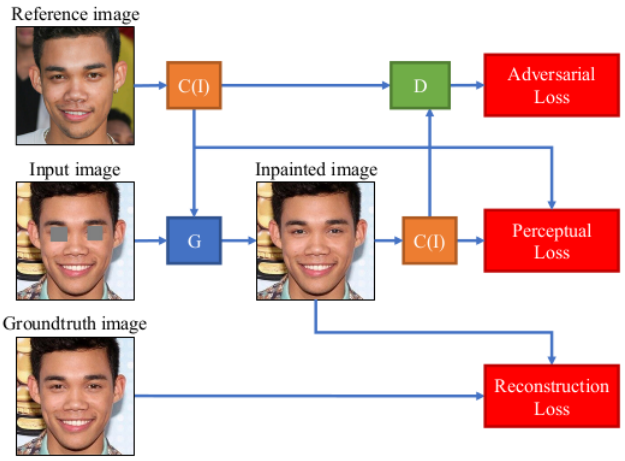

Figure 2. General architecture of an Exemplar GAN. The overall training flow can be summarized as (1) mark the eyes from the input image; (2) in-paint the image with the reference image or code as a guide; (3) compute the gradient of the generator's parameters via the content/reconstruction loss between the input image and the in-painted image; (4) compute the gradient of the discriminators parameters with the in-painted image, another real, ground truth image, and the reference image or code; (5) backpropagate the discriminator error through the generator. Optionally, (6) the generator's parameters can also be updated with a perceptual loss. For reference-based Exemplar GANs, the compressing functions C(I) are the identity function.

the team define the adversarial objective as Eq. 2:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x_i, c_i \sim p_{data}(x,c)}[\log D(x_i, c_i)] + \mathbb{E}_{c_i \sim p_c, G(\cdot) \sim p_z}[\log 1 - D(G(z_i, c_i))] + \| G(z_i, c_i) - x_i \|_1 + \| G(z_i, c_i) - c_i \|_2 \quad (2)$$

The compressing function can be a deterministic function, an auto-encoder, or a general deep network that projects an example onto some manifold. The final term in Eq. 2 is an optional loss that measures the distance of the generated

Table 1. Quantitative results for the 3 best GAN models. For all metrics except inception score, lower is better.

Model	L1	MS-SSIM	Inception	FID
Internal benchmark				
Non-exemplar	0.018	5.05E-2	3.96	11.27
Reference	0.014	3.97E-2	3.82	7.67
Code	0.015	4.15E-2	3.94	8.49
Celeb-ID				
Non-exemplar	7.36E-3	8.44E-3	3.72	15.30
Reference	7.15E-3	7.97E-3	3.82	15.66
Code	7.00E-3	7.80E-3	3.77	14.62

image $G(z_i, c_i)$ to the original reference image r_i.

2.3. Model architecture

The overall layout for both code- and reference-based ExGANs is depicted in Fig. 2. For most experiments, Brian Dolhansky and Cristian Canton Ferrer used a standard convolutional generator, with a bottleneck region containing dilated convolutions, similar to the generator proposed in [4], but with a smaller channel count in the interior layers of the network, as generating eyes is a more restricted domain than general in-painting. The input to the generator is an RGB image with the portions to in-paint removed, stacked with a one-channel binary mask indicating which regions to fill. The generator could take an additional four channels: the RGB values of the reference image, and another 1-channel mask indicating the eye locations. All eye locations are detected prior to training and stored with the dataset.

3. Experiment setup

In order to best judge the effects of both code- and reference-based ExGANs, Brian Dolhansky and Cristian Canton Ferrer avoided mixing codes and reference images in a single network. The FID score did correlate strongly with perceived quality. For the images in result, the FID score increased along with the blurriness in the image. The team therefore postulate that for eye in-painting in general, the best metric to compare models is the FID score, as it most accurately corresponds with sharpness and definition around the generated eye. A list of metrics for the three best GAN models is given in Tab. 1.

References

[1] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. *ACM TOG*, 2004. 1

[2] C. Barnes, E. Shechtman, A. Finkelstein, and B. G. Dan. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM TOG*, 2009. 1

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[4] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM TOG*, 2017. 1, 3

[5] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[6] P. Rez, M. Gangnet, and A. Blake. Poisson image editing. *ACM TOG*, 2003. 1