Semantic Image Completion through an Adversarial Strategy

Patricia Vitoria[0000−0002−2437−0191], Joan Sintes, and Coloma Ballester[0000−0001−6535−7367]

Department of Information and Communication Technologies, University Pompeu Fabra, Barcelona, Spain
{patricia.vitoria, coloma.ballester}@upf.edu, joansintesmarcos@gmail.com

Abstract. Image completion or image inpainting is the task of filling in missing regions of an image. When those areas are large and the missing information is unique such that the information and redundancy available in the image is not useful to guide the completion, the task becomes even more challenging. This paper proposes an automatic semantic inpainting method able to reconstruct corrupted information of an image by semantically interpreting the image itself. It is based on an adversarial strategy followed by an energy-based completion algorithm. First, the data latent space is learned by training a modified Wasserstein generative adversarial network. Second, the learned semantic information is combined with a novel optimization loss able to recover missing regions conditioned by the available information. Moreover, we present an application in the context of face inpainting, where our method is used to generate a new face by integrating desired facial attributes or expressions from a reference face. This is achieved by slightly modifying the objective energy. Quantitative and qualitative top-tier results show the power and realism of the presented method.

Keywords: Generative Models · Wasserstein GAN · Image Inpainting · Semantic Understanding.

1 Introduction

When looking at a censored photograph or at a portrait whose eyes, nose and/or mouth are occluded, our brain has no difficulty in hallucinating a plausible completion of the face. Moreover, if the visible parts of the face are familiar to us because, e.g., they remind us a friend or a celebrity, we mentally conceive a face having the whole set of attributes: the visible ones and the ones we infer. However, as an automatic computer vision task, it remains a challenging task.

The mentioned task falls into the so-called image inpainting where the goal is to recover missing information of an image in a realistic manner. Its applications are numerous and range from automatizing cinema post-production tasks enabling, e.g., the deletion of annoying objects, to new view synthesis generation for, e.g., broadcasting of sport events. Classical methods use the available

Fig. 1. Several inpainted images resulting from the proposed algorithm using different masks.

redundancy of the incomplete input image: smoothness priors in the case of geometry-oriented approaches and self-similarity principles in the non-local or exemplar-based ones. Nevertheless, they fail in recovering large regions when the available image redundancy is not useful, such as the cases of Figure 1. In this paper we capitalize on the understanding of more abstract and high level information that learning strategies may provide. We propose an extension of the semantic image inpainting model proposed in [34] to automatically complete any region of an image including those challenging cases. It consists of a deep learning based strategy which uses generative adversarial networks (GANs) to learn the image space and an appropriate loss for an inpainting optimization algorithm which outputs a semantically plausible completion where the missing content is conditioned by the available data. Our method can be applied to recover any incomplete image no matter the shape and the size of the holes of information.

In this work, following our previous paper [34], we will train an improved version of the Wasserstein GAN (WGAN) to implicitly learn a data latent space and subsequently to generate new samples from it. We incorporate in this paper the PatchGAN network structure [20] in the discriminator as well as we update the generator to handle images with higher resolution. The new discriminator is able to take decisions over patches of the image and it is combined with the original global discriminator that takes a decision for the whole image itself. With this purpose, we define a new energy function able to generate the missing content conditioned to a reference image. We deploy it on hallucinating faces

Fig. 2. Image inpainting results using three different approaches. (a) Input images, each with a big hole or mask. (b) Results obtained with the non-local method [14]. (c) Results with the local method [15]. (d) Our semantic inpainting method. Figure retrieved from our previous work [34].

conditioned by reference facial attributes or expressions. We quantitatively and qualitatively show that our proposal achieves top-tier results on two datasets: CelebA and Street View House Numbers. Additionally, we show qualitative results in our application for face hallucination. The code has been made publicly available.

The remainder of the paper is organized as follows. In Section 2, we review some preliminaries and state-of-the-art related work on that topic focusing first on generative adversarial networks and then on inpainting methods and face complation. Section 3 details both methods for image inpainting and face hallucination. In Section 4, we present quantitative and qualitative assessments of all parts of the proposed method. Section 5 concludes the paper.

2 Preliminaries and State-of-the-Art Related Work

2.1 Generative Adversarial Networks

GAN [16] learning strategy is based on a game theory scenario between two networks, the generator and the discriminator, competing against each other. The goal of the generator is to generate samples of data from an implicitly learned probability distribution that is aimed to be as closer as possible as the probability distribution the real data. On the other hand, the discriminator tries to distinguish real from fake data. To do so, the discriminator, denote here by D, is trained to maximize the probability of correctly distinguish between real examples and samples created by the generator, denoted by G, while G is trained to fool the discriminator and to minimize $\log(1-D(G(z)))$ by generating realistic examples. In other words, D and G play the following min-max game with value

function $V(G, D)$ defined as follows:

$$\min_G \max_D V(D, G) = \min_G \max_D \underset{x \sim P_{real}(x)}{E} [\log D(x)] + \underset{z \sim p_z(z)}{E} [\log(1 - D(G(z)))], \tag{1}$$

where P_{real} denotes the probability distribution the real data, and p_z represents the distribution of the latent variables z. The authors of [31] introduced convolutional layers to the GANs architecture, and proposed the so-called Deep Convolutional Generative Adversarial Network (DCGAN) able to learn more complex data.

GANs have been applied with success to many computer vision related tasks such as image colorization [8], text to image synthesis [32], super-resolution [21], image inpainting [39, 6, 11], and image generation [31, 25, 17, 28], to mention just a few. They have also been applied to other modalities such as speech, audio and text. However, three difficulties still persist as challenges. One of them is the quality of the generated images and the remaining two are related to the well-known instability problem in the training procedure. For instance, two problems can appear: vanishing gradients and mode collapse.

Aiming a stable training of GANs, several authors have promoted the use of the Wasserstein GAN (WGAN). WGAN minimizes an approximation of the Earth-Mover (EM) distance or Wasserstein-1 metric between two probability distributions. The authors of [2] analyzed the properties of this distance. They showed that one of the main benefits of the Wasserstein distance is that it is continuous. This property allows to robustly learn a probability distribution by smoothly modifying the parameters through gradient descend. Moreover, the Wasserstein or EM distance is known to be a powerful tool to compare probability distributions with non-overlapping supports, in contrast to other distances such as the Kullback-Leibler divergence and the Jensen-Shannon divergence (used in the DCGAN and other GAN approaches) which produce the vanishing gradients problem, as mentioned above. Using the Kantorovich-Rubinstein duality, the Wasserstein distance between two distributions, say a *real* distribution P_{real} and an estimated distribution P_g, can be computed as

$$W(P_{real}, P_g) = \sup E_{x \sim P_{real}} [f(x)] - E_{x \sim P_g} [f(x)] \tag{2}$$

where the supremum is taken over all the 1-Lipschitz functions f (notice that, if f is differentiable, it implies that $\|\nabla f\| \leq 1$). Let us notice that f in Equation (2) can be thought to take the role of the discriminator D in the GAN terminology. In [2], the WGAN is defined as the network whose parameters are learned through optimization of

$$\min_G \max_{D \in \mathcal{D}} E_{x \sim P_{real}} [D(x)] - E_{x \sim P_G} [D(x)] \tag{3}$$

where $\mathcal{D}$ denotes the set of 1-Lipschitz functions. Under an optimal discriminator (called a *critic* in [2]), minimizing the value function with respect to the generator parameters minimizes $W(P_{real}, P_g)$. To enforce the Lipschitz constraint, the authors proposed to use an appropriate weight clipping. The resulting WGAN

solves the vanishing problem, but several authors [17, 1] have noticed that weight clipping is not the best solution to enforce the Lipschitz constraint and it causes optimization difficulties. For instance, the WGAN discriminator ends up learning an extremely simple function and not the real distribution. Also, the clipping threshold must be properly adjusted. Since a differentiable function is 1-Lipschitz if it has gradient with norm at most 1 everywhere, [17] proposed an alternative to weight clipping by adding a gradient penalty term constraining the L^2 norm of the gradient while optimizing the original WGAN during training called WGAN-GP. The WGAN-GP minimization loss is defined as

$$\min_G \max_{D \in \mathcal{D}} E_{\tilde{x} \sim P_{real}} [D(\tilde{x})] - E_{x \sim P_G} [D(x)] - \lambda E_{\tilde{x} \sim P_{\tilde{x}}} \left[(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2 \right] \quad (4)$$

As in [17], $P_{\tilde{x}}$ is implicitly defined sampling uniformly along straight lines between pairs of point sampled from the data distribution P_real and the generator distribution P_G. Let us notice that the minus before the gradient penalty term in (4) corresponds to the fact that, in practice, when optimizing with respect to the discriminator parameters, one minimizes the negative of the loss instead of maximizing it.

In this work, we leverage the mentioned WGAN-GP improved with a new design of the generator and discriminator architectures.

2.2 Image Inpainting

In general, image inpainting methods found in the literature can be classified into two groups: model-based approaches and deep learning approaches. In the former, two main groups can be distinguished: local and non-local methods. In local methods, images are modeled as functions with some degree of smoothness (see [5, 26, 9, 7] to mention but a few of the related literature). These methods show good performance in propagating smooth level lines or gradients but fail in the presence of texture or large missing regions. Non-local methods exploit the self-similarity prior by directly sampling the desired texture to perform the synthesis (e.g., [10, 3, 13]. They provide impressive results while inpainting textures and repetitive structures even in the case of large regions to recover. However, as both type of methods use the redundancy present in known parts of the incomplete input image, through smoothness priors in the case of geometry-based and through self-similarity principles in the non-local or patch-based ones and thus fail in completing singular information.

In the last decade, most of the state-of-the-art methods are based on deep learning approaches [39, 11, 29, 38, 40]. The authors of [29] (see also [35, 33]) modified the original GAN architecture by inputting the image context instead of random noise to predict the missing patch. [39] proposes an algorithm which generates the missing content by conditioning on the available data given a trained generative model, while [38] adapts multi-scale techniques to generate high-frequency details on top of the reconstructed object to achieve high resolution results. The work [19] adds a local discriminator network that considers

only the filled region to emphasize the adversarial loss on top of the global discriminator. This additional network, called the local discriminator (L-GAN), facilitates exposing the local structural details. Also, the authors of [11] design a discriminator that aggregates the local and global information by combining a global GAN and a Patch-GAN.

2.3 Face Completion

Several works focus on the task of face completion. For example, the classical work of [18] completes a face by computing the least squares solution giving the appropriate optimal coefficients combining a a set of prototypes of shape and texture. [12] uses a spectral-graph-based filling-in algorithm to fill-in the occluded regions of a face. The occluded region is automatically detected and reconstructed in [23] by using GraphCut-based detection and confidence-oriented sampling, respectively. For a detailed account of the face completion and hallucination literature we refer to [37]. Deep learning strategies are used in [22].Their method is based on a GAN, two adversarial loss functions (local and global) and a semantic parsing loss.

3 Proposed Approach

We construct from our previous work on semantic image inpainting [34] which is built on two main blocks. First, given a dataset of (non-corrupted) images, it trains the proposed version of the WGAN [34] to implicitly learn a data latent space to subsequently generate new random samples from the dataset. Once the model is trained, given an incomplete image and the trained generative model, an iterative minimization procedure is performed to infer the missing content of the incomplete image by conditioning the resulting image on the known regions. This procedure consists on searching the closest encoding of the corrupted data in the latent manifold by minimizing the proposed loss which is made of a combination of contextual, through image values and image gradients, and prior losses.

In this paper we introduce two updates in the WGAN architecture. The first one allows us to train with images of higher resolution. The second improvement is aimed at the discriminator, where instead of having a single decision for each input image, the discriminator additionally takes decisions over patches of the image.

Additionally, we propose a method for conditional face completion. As in the previous method, given a dataset of images of faces, we train our network in order to learn the data distribution. Then, given an image of a face, say y_1 (that can be either complete or incomplete), and the previously trained generative model, we perform an energy-based optimization procedure to generate a new image which is similar to y_1 but has some meaningful visage portions (such as the eyes, mouth or nose) similar to a reference image y_2 by conditioning the generated image through the objective loss function.

3.1 Adversarial based Learning of the Data Distribution

The adversarial architecture presented in [34] was built on the top of the WGAN-GP [17]. Several improvements were proposed in [34] to increase the stability of the network:

- Deep networks can learn more complex, non-linear functions, but are more difficult to train. One of the improvements were the introduction of residual learning in both the generator and discriminator which eases the training of these networks, and enables them to be substantially deeper and stable. Instead of hoping each sequence of layers to directly fit a desired mapping, we explicitly let these layers fit a residual mapping. Therefore, the input x of the residual block is recast into $F(x) + x$ at the output. Figure 3 and 4 show the layers that make up a residual block in our model. Figure 5 and 6 display a visualization of the architecture of the generator (Figure 6) and of the discriminators (Figure 5).
- The omition of batch normalization in the discriminator. To not introduce correlation between samples, it uses layer normalization [4] as a drop-in replacement for batch normalization in the discriminator.
- Finally, the ReLU activation is used in the generator with the exception of the output layer which uses the Tanh function. Within the discriminator ReLU activation are also used. This is in contrast to the DCGAN, which makes use of the LeakyReLu.

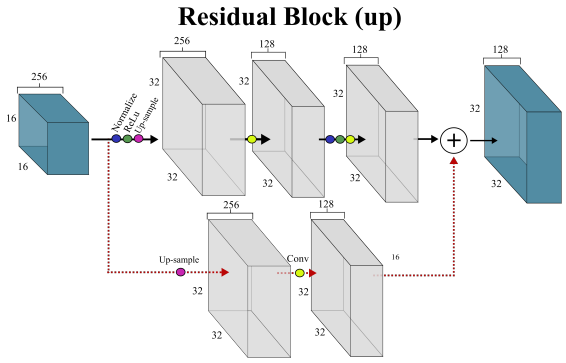

Fig. 3. An example of the residual block used in the generator. Figure retrieved from our previous work [34].

Additionally to the previous mentioned changes applied in the work [34], further modifications have been applied: an extra discriminator architecture based on PatchGAN and a modification on the model architecture to deal with higher resolution images.

PatchGAN Discriminator. Inspired by the Markovian architecture (PatchGAN [20]) we have added a new discriminator in our adversarial architecture.

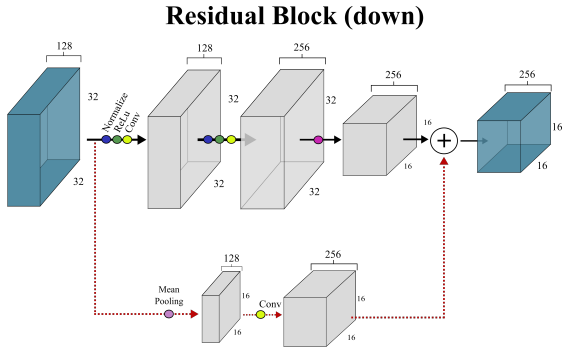

Fig. 4. An example of the residual block used in the discriminator. Figure retrieved from [34].

The PatchGAN discriminator keeps track of the high-frequency structures of the generated image by focusing on local patches. Thus, instead of penalizing at the full image scale, it tries to classify each patch as real or fake. Hence, rather than giving a single output for each input image, it generates a decission value for each patch.

3.2 Semantic Image Inpainting

Once the model is trained until the mapping from the data latent space to uncorrupted data has been properly estimated, semantic image completion can be performed. More precisely, after traning, the generator is able to take a random vector z drawn from p_z and generate an image mimicking samples from P_{real}. In order to perform inpainting or completion of an incomplete image, the aim is to recover the encoding $\hat{z}$ that is closest to the corrupted image while being constrained to the learned encoding manifold of z. Then, when $\hat{z}$ is found, the damaged areas can be restored by using the trained generator G on $\hat{z}$.

We formulate the process of finding $\hat{z}$ as an optimization problem. Let y be a damaged image and M a binary mask of the same spatial size as the input image y, where the white pixels (that is, the pixels i such that $M(i) = 1$) determine the uncorrupted areas of y. The closest encoding $\hat{z}$ can be defined as the optimum of the following optimization problem with the loss defined as [34]:

$$\hat{z} = arg\,\underset{z}{min}\{\mathcal{L}_c(z|y, M) + \eta\mathcal{L}_p(z)\} \tag{5}$$

where $\mathcal{L}_p$ stays for prior loss and $\mathcal{L}_c$ for contextual loss defined as

$$\mathcal{L}_c(z|y, M) = \alpha W\|M(G(z) - y)\| + \beta W\|M(\nabla G(z) - \nabla y)\| \tag{6}$$

where α, β, η are positive constants and ∇ denotes the gradient operator. In particular, the contextual loss $\mathcal{L}_c$ constrains the generated image to the color and gradients of the image y to be inpainted on the regions with available data

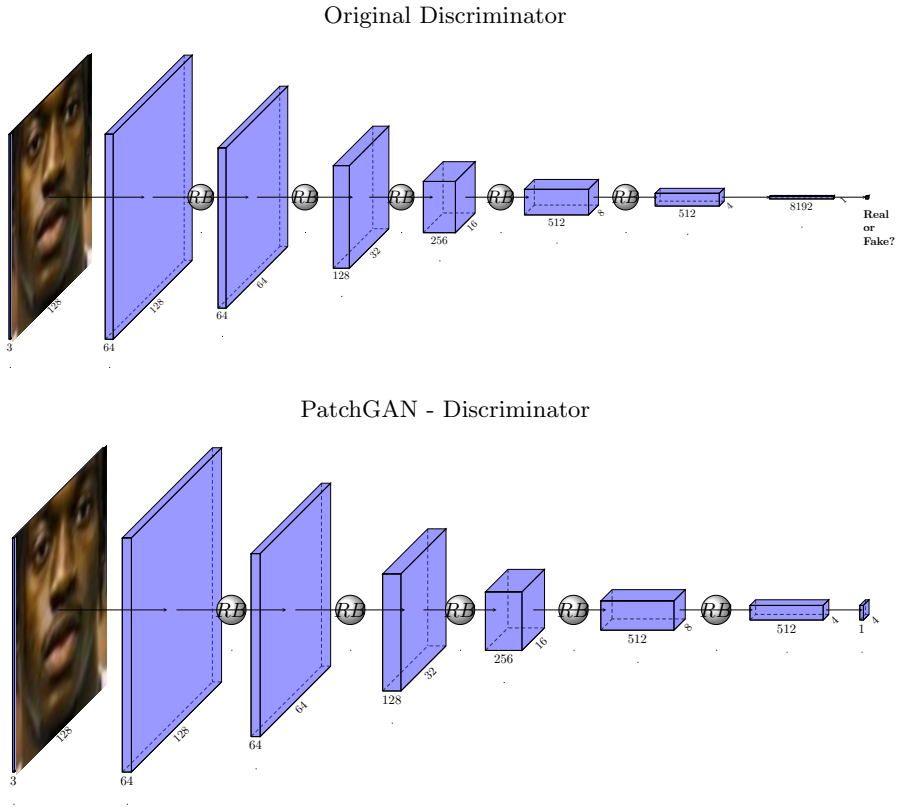

Fig. 5. Overview of the original discriminator architecture (above) and PatchGAN Discriminator (below). In our model we use both types of discriminators. RB stands for Residual Block.

given by $M \equiv 1$. Moreover, the contextual loss $\mathcal{L}_c$ is defined as the L^1 norm between the generated samples $G(z)$ and the uncorrupted parts of the input image y weighted in such a way that the optimization loss pays more attention to the pixels that are close to the corrupted area when searching for the optimum encoding $\hat{z}$. Notice, that the proposed contextual loss does not only constrain the color information but also the structure of the generated image given the structure of the input corrupted image. The benefits are specially noticeable for a sharp and detailed inpainting of large missing regions which typically contain some kind of structure (e.g. nose, mouth, eyes, texture, etc, in the case of faces). In practice, the image gradient computation is approximated by central finite differences. In the boundary of the inpainting hole, we use either forward or backward differences depending on whether the non-corrupted information is available.

The weight matrix W is defined for each uncorrupted pixel i as

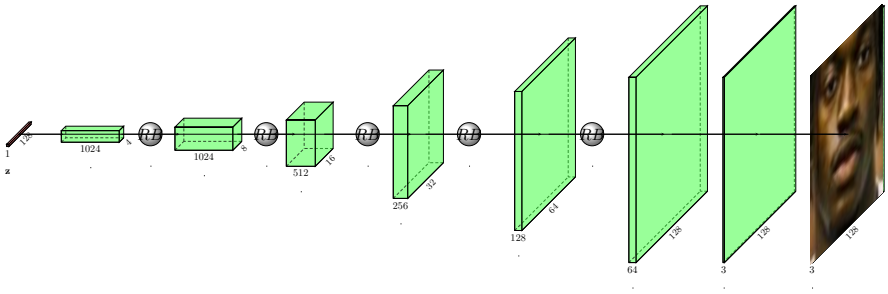

Fig. 6. Overview of the generator architecture.

$$W(i) = \begin{cases} \sum_{j \in N_i} \frac{(1 - M(j))}{|N_i|} & if M(i) \neq 0 \\ 0 & if M(i) = 0 \end{cases} \tag{7}$$

where N_i denotes a local neighborhood or window centered at i, and $|N_i|$ denotes its cardinality, i.e., the area (or number of pixels) of N_i. This weighting term has also been used by [39]. In order to compare our results with them, we have fixed the window size to the value used by them (7×7) for all the experiments.

Finally, the prior loss $\mathcal{L}_p$ is defined such as it favours realistic images, similar to the samples that are used to train our generative model, that is,

$$\mathcal{L}_p(z) = -D_{w_1}(G_\theta(z)) - D^{patch}_{w_2}(G_\theta(z)) \tag{8}$$

where D_{w_1} and $D^{patch}_{w_2}$ are the output of the discriminator D and D^{patch} with parameters w_1 and w_2 given the image $G_\theta(z)$ generated by the generator with parameters θ and input vector z. In other words, the prior loss is defined as the second WGAN loss term in (3) penalizing unrealistic images. Without $\mathcal{L}_p$ the mapping from y to z may converge to a perceptually implausible result. Therefore z is updated to fool the discriminator and make the corresponding generated image more realistic.

(a) (b) (c) (d)

Fig. 7. Images (b) and (d) show the results obtained after applying Poisson editing (equation (9) in the text) to the inpainting results shown in (a) and (c), respectively. Figure retrieved from [34].

The parameters α, β and η in Equation (6) allow to balance among the three losses. With the defined contextual, gradient and prior losses, the corrupted image can be mapped to the closest z in the latent representation space, denoted by $\hat{z}$. z is randomly initialized with Gaussian noise of zero mean and unit standard deviation and updated using back-propagation on the total loss given in the equation (6). Once $G(\hat{z})$ is generated, the inpainting result can be obtained by overlaying the uncorrupted pixels of the original damaged image to the generated image. Even so, the reconstructed pixels may not exactly preserve the same intensities of the surrounding pixels although the content and structure is correctly well aligned. To solve this problem, a Poisson editing step [30] is added at the end of the pipeline in order to reserve the gradients of $G(\hat{z})$ without mismatching intensities of the input image y. Thus, the final reconstructed image $\hat{x}$ is equal to:

$$\begin{aligned} \hat{x} &= arg\min_{x} \|\nabla x - \nabla G(\hat{z})\|_2^2 \\ &\text{such that } x(i) = y(i) \text{ if } M(i) = 1 \end{aligned} \tag{9}$$

In Figure 7 two examples can be seen where visible seams are appreciated in (a) and (c), but less in (b) and (d) after applying Poisson editing (9).

3.3 Conditional Face Completion

Let G be the previously defined generator mapping the noise vector z belonging to the latent space to an image $G(z)$ as obtained in Section 3.1. In order to generate a new face by integrating the desired facial attributes or expressions from a reference face, similar to the previous outline, we formulate the process of finding the closed encoding of the corrupted data in the latent manifold as an optimization problem. Let y_1 be an image of a face, y_2 a reference image, M a binary mask of the same spatial size as the image where the white pixels ($M(i) = 1$) determine the area to preserve of y_1 (last row, second column, of Figure 12 displays an example of M). We define the closest encoding $\hat{z}$ to y_1 conditioned by y_2 as the optimum of following optimization problem with the new proposed loss:

$$\hat{z} = \arg\min_{z} \mathcal{L}_{\mathrm{c}_1}(z|y_1, M) + \mathcal{L}_{\mathrm{c}_2}(z|y_2, I - M) + \beta\tilde{\mathcal{L}}_{\mathrm{p}}(z) \tag{10}$$

where the first two terms are contextual losses as defined in Section 3.2 that penalize on complementary regions (given by the masks M and $I-M$, where I is constant and equal to 1 on all the pixels). More specifically, the first contextual loss favours to maintain color and structure from the image to be inpainted, and the second contextual loss favours to maintain structure from the reference image. The third term is the prior loss as defined in 3.2. Let us write them again now in this conditional face completion context:

$$\mathcal{L}_{\mathrm{c}_1}(z|y_1, M) = \alpha_1 W_1 \|M(G(z) - y_1)\| + \alpha_2 W_1 \|M(\nabla G(z) - \nabla y_1)\|, \tag{11}$$

$$\mathcal{L}_{\mathrm{c}_2}(z|y_2, I - M) = \alpha_3 W_2 \|(I - M)(\nabla G(z) - \nabla y_2)\|, \tag{12}$$

$$\tilde{\mathcal{L}}_{\mathrm{p}}(z) = -D_w(G_\theta(z)) \tag{13}$$

where W_1 and W_2 denote the weights defined for each pixel i and its neighborhood N_i as

$$W_1(i) = M(i) \sum_{j \in N_i} \frac{(1 - M(j))}{|N_i|} \tag{14}$$

$$W_2(i) = (I - M)(i) \left(1 - \sum_{j \in N_i} \frac{M(j)}{|N_i|} \right) \tag{15}$$

4 Experimental Results

In this section we present qualitative and quantitative results of the proposed methods. We will show qualitative and quantitative results of our inpainting method proposed in [34]. The results will be compared with the ones obtained by [39] as both algorithms use first a GAN procedure to learn semantic information from a dataset and, second, combine it with an optimization loss for inpainting in order to infer the missing content. Additionally, further visual results in higher resolution images will be shown. To conclude, results on conditional face completion will be presented.

For all the experiments we use a fixed number of epochs equal to 10, batch size equal to 64, learning rate equal to 0.0001 and exponential decay rate for the first and second moment estimates in the Adam update technique, $\beta_1 = 0,0$ and $\beta_2 = 0,9$, respectively. Training the generative model required three days using an NVIDIA QUADRO P6000.

During the inpainting stage, the window size used to compute $W(i)$ in (7) is fixed to 7×7 pixels. In our algorithm, we use back-propagation to compute $\hat{z}$ from the latent space. We make use of an Adam optimizer and restrict z to fall into $[-1, 1]$ in each iteration, which we found it produces more stable results. In that stage we used the Adam hyperparameters learning rate, α, equal to 0.03 and the exponential decay rate for the first and second moment estimates, $\beta_1 = 0,9$ and $\beta_2 = 0,999$, respectively. After initializing with a random 128 dimensional vector z drawn from a normal distribution, we perform 1000 iterations.

The assessment is given on two different datasets in order to check the robustness of our method: the CelebFaces Attributes Datases [24] and the Street View House Numbers (SVHN) [27]. CelebA dataset contains a total of 202.599 celebrity images covering large pose variations and background clutter. We split them into two groups: 201,599 for training and 1,000 for testing. In contrast, SVHN contains only 73,257 training images and 26,032 testing images. SVHN images are not aligned and have different shapes, sizes and backgrounds. The images of both datasets have been cropped with the provided bounding boxes and resized to only 64x64 pixel size.

Remark that we have trained the proposed improved WGAN by using directly the images from the datasets without any mask application. Afterwards,

our semanting inpainting method is evaluated on both datasets using the inpainting masks. Notice that our algorithm can be used with any type of inpainting mask.

4.1 Qualitative Assessment

In [34] we have analyzed separately each step of our algorithm: The training of the adversarial model and the minimization procedure to infer the missing content. Since the inpainting result of the latter strongly depends on what the generative model is able to produce, a good estimation of the data latent space is crucial for our task. Notice that the CelebA dataset will be better estimated than SVHN dataset due to the fact that the number of images as well as the diversity of the dataset directly affects the prediction of the latent space and the estimated underlying probability density function (pdf). In contrast, as bigger the variability of the dataset, more spread is the pdf which difficult its estimation.

To evaluate the proposed inpainting method, a comparison with the semantic inpainting method by [39] was performed. While training our model, we use the proposed architecture (see Section 3.1) where the model takes a random vector, of dimension 128, drawn from a normal distribution. In contrast, [39] uses the DCGAN architecture where the generative model takes a random 100 dimensional vector following a uniform distribution between $[-1, 1]$. Some qualitative results are displayed in Figures 8 and 9. Focusing on the CelebA results (Figure 8), obviously the algorithm by [39] performs better than local and non-local methods (Figure 2) since it also makes use of adversarial models. However, although it is able to recover the semantic information of the image and infer the content of the missing areas, in some cases it keeps producing results with lack of structure and detail which can be caused either by the generative model or by the procedure to search the closest encoding in the latent space. It will be further analyzed with a quantitative ablation study. Since the proposed method takes into account not only the pixel values but also the structure of the image, this kind of problems are solved. In many cases, our results are as realistic as the real images. Notice that challenging examples, such as the first and sixth row from Figure 8, which image structures are not well defined, are not properly recovered with our method nor with [39].

Regarding the results on SVHN dataset (Figure 9), although they are not as realistic as the CelebA ones, the missing content is well recovered even when different numbers may semantically fit the context. As mentioned before, the lack of detail is probably caused by the training stage, due to the large variability of the dataset (and the size of the dataset). Despite of this, let us notice that our results outperform qualitatively the ones obtained by [39]. This may indicate that our algorithm is more robust when using smaller datasets than [39]. Some examples of failure cases found on both datasets are shown in Figure 11.

Additional Results in Higher Resolution Images

Figure 10 shows several resulting higher resolution images after applying the proposed algorithm in the corrupted regions of the image. Notice, that our algorithm is able to inpaint any region regardless of its shape. One can see that

the obtained results look realistic even in challenging parts of the image such as the eyes or nose. Also, it obtains good results when the observer does not see all the face, such in the middle example in the second row.

4.2 Quantitative Analysis and Evaluation Metrics

The goal of semantic inpainting is to fill-in the missing information with realistic content. However, with this purpose, there are many correct possibilities to semantically fill the missing information apart from the ground truth solution. Thus, in order to provide a thorough analysis and quantify the quality of our method in comparison with other methods, an ablation study was presented in [34]. We include it here for the sake of completeness. Different evaluation metrics were used: First, metrics based on a distance with respect to the ground truth and, second, a perceptual quality measure that is acknowledged to agree with similarity perception of the human visual system.

Table 1. Quantitative inpainting results for the central square mask, including an ablation study of our contributions in comparison with [39]. The best results for each dataset are marked in bold and the best results for each method are underlined. Table retrieved from [34].

	CelebA dataset			SVHN dataset		
Loss formulation	MSE	PSNR	SSIM	MSE	PSNR	SSIM
[39]	872.8672	18.7213	0.9071	1535.8693	16.2673	0.4925
[39] adding gradient loss with $\alpha = 0.1$, $\beta = 0.9$ and $\eta = 1.0$	832.9295	18.9247	0.9087	1566.8592	16.1805	0.4775
[39] adding gradient loss with $\alpha = 0.5$, $\beta = 0.5$ and $\eta = 1.0$	862.9393	18.7710	0.9117	1635.2378	15.9950	0.4931
[39] adding gradient loss with $\alpha = 0.1$, $\beta = 0.9$ and $\eta = 0.5$	794.3374	19.1308	0.9130	1472.6770	16.4438	0.5041
[39] adding gradient loss with $\alpha = 0.5$, $\beta = 0.5$ and $\eta = 0.5$	876.9104	18.7013	0.9063	1587.2998	16.1242	0.4818
Our proposed loss with $\alpha = 0.1$, $\beta = 0.9$ and $\eta = 1.0$	855.3476	18.8094	0.9158	631.0078	20.1305	**0.8169**
Our proposed loss with $\alpha = 0.5$, $\beta = 0.5$ and $\eta = 1.0$	**785.2562**	**19.1807**	**0.9196**	743.8718	19.4158	0.8030
Our proposed loss with $\alpha = 0.1$, $\beta = 0.9$ and $\eta = 0.5$	862.4890	18.7733	0.9135	**622.9391**	**20.1863**	0.8005
Our proposed loss with $\alpha = 0.5$, $\beta = 0.5$ and $\eta = 0.5$	833.9951	18.9192	0.9146	703.8026	19.6563	0.8000

Table 2. Quantitative inpainting results for the three squares mask including an ablation study of our contributions and a complete comparison with [39]. The best results for each dataset are marked in bold and the best results for each method are underlined. Table retrieved from [34].

	CelebA dataset			SVHN dataset		
Method	MSE	PSNR	SSIM	MSE	PSNR	SSIM
[39]	622.1092	20.1921	0.9087	1531.4601	16.2797	0.4791
[39] adding gradient loss with $\alpha = 0.1$, $\beta = 0.9$ and $\eta = 1.0$	584.3051	20.4644	0.9067	1413.7107	16.6272	0.4875
[39] adding gradient loss with $\alpha = 0.5$, $\beta = 0.5$ and $\eta = 1.0$	600.9579	20.3424	0.9080	1427.5251	16.5850	0.4889
[39] adding gradient loss with $\alpha = 0.1$, $\beta = 0.9$ and $\eta = 0.5$	580.8126	20.4904	0.9115	1446.3560	16.5281	0.5120
[39] adding gradient loss with $\alpha = 0.5$, $\beta = 0.5$ and $\eta = 0.5$	563.4620	20.6222	0.9103	1329.8546	16.8928	0.4974
Our proposed loss with $\alpha = 0.1$, $\beta = 0.9$ and $\eta = 1.0$	424.7942	21.8490	0.9281	168.9121	25.8542	0.8960
Our proposed loss with $\alpha = 0.5$, $\beta = 0.5$ and $\eta = 1.0$	380.4035	22.3284	0.9314	221.7906	24.6714	**0.9018**
Our proposed loss with $\alpha = 0.1$, $\beta = 0.9$ and $\eta = 0.5$	**321.3023**	**23.0617**	**0.9341**	**154.5582**	**26.2399**	0.8969
Our proposed loss with $\alpha = 0.5$, $\beta = 0.5$ and $\eta = 0.5$	411.8664	21.9832	0.9292	171.7974	25.7806	0.8939

In the first case, considering the real images from the database as the ground truth reference, the most used evaluation metrics are the Peak Signal-to-Noise Ratio (PSNR) and the Mean Square Error (MSE). Notice, that both MSE and PSNR, will choose as best results the ones with pixel values closer to the ground truth.

In the second case, in order to evaluate perceived quality, the Structural Similarity index (SSIM) [36] is used to measure the similarity between two images. It is considered to be correlated with the quality perception of the human visual system.

Given these metrics the obtained results are compared with the one proposed by [39] as it is the method more similar to ours. Tables 1 and 2 show the numerical performance of our method and [39]. To perform an ablation study of all our contributions and a complete comparison with [39], Tables 1 and 2 not only show the results obtained by their original algorithm and our proposed algorithm, but also the results obtained by adding our new gradient-based term $\mathcal{L}_g(z|y, M)$ to their original inpainting loss. We present the results varying the trade-off effect between the different loss terms (weights α, β, η).

By looking at the numerical results it can be seen that the proposed algorithm always performs better than the semantic inpainting method by [39]. For the case of the CelebA dataset, the average MSE obtained by [39] is equal to 872.8672 and 622.1092, respectively, compared to our results that are equal to 785.2562 and 321.3023, respectively. It is highly reflected in the results obtained using the SVHN dataset, where the original version of [39] obtains an MSE equal to 1535.8693 and 1531.4601, using the central and three squares mask respectively, and our method 622.9391 and 154.5582. On the one side, the proposed WGAN structure is able to create a more realistic latent space and, on the other side, the proposed loss takes into account essential information in order to recover the missing areas.

Regarding the accuracy results obtained with the SSIM measure, can be seen that the results obtained by the proposed method always have a better perceived quality than the ones obtained by [39]. In some cases, the values are close to the double, for example, in the case where the training dataset is small, namely, SVHN.

To conclude, the proposed method is more stable in smaller datasets such in the case of SVHN. Also, by decreasing the number of samples in the dataset does not mean to reduce the quality of the inpainted images in the proposed method. Contrary with what is happening in the case of [39]. Finally, in the cases where we add the proposed loss to the algorithm [39], in most of the cases the MSE, PSNR and SSIM improves. This fact clarifies the big importance of the gradient loss in order to perform semantic inpainting.

4.3 Conditional Face Completion

In this section, we evaluate our algorithm in the CelebA dataset [24] that consists on 202,599 face images, which are aligned and cropped to have pixel size equal to 64×64.

Some qualitative and quantitative results are shown in Figure 12. Our algorithm outputs a face hallucination of one of the images y_1 displayed in the first row, having as a reference the portion displayed in the second column of the image y_2 of the first column. More often than not, the results look natural and the combination of the target image together with its reference is plausible. As can be seen, our algorithm is robust in combining images with different skin tone, keeping the overall color of the target image. The last row shows results of our baseline semantic completion method showing that it can perceptually hallucinate a plausible completion without any reference image. In order to quantify the quality of our results, we have computed the Structural Similarity Index (SSIM) that is correlated with the quality perception of the human visual system. Notice that the SSIM is computed with respect to the target image y_1 although ours results are a combination of two images. Even so, the resulting SSIM is high in all the cases (above 0.85) which translates to a high perceived quality.

5 Conclusions

This paper proposes a semantic inpainting method based on an adversarial strategy. The method performs in two phases. First, the data latent manifold is learned by training a proposed improved version of the WGAN. Then, we propose a conditional objective loss. This loss is able to properly infer the missing content having into account the structure and pixel value of the data present on the image. Moreover, it takes also into account the perceptual realism of the reconstructed image. Additionally, a new loss is presented able to perform personalized face completion based on semantic image inpainting. By iteratively minimizing this new loss we are able to generate an image similar to a target image together with meaningful characteristics of a reference one. The presented experiments show the capabilities of the proposed method that is able to infer more realistic content for incomplete images than classical methods.

Acknowledgements

The authors acknowledge partial support by MICINN/FEDER UE project, reference PGC2018-098625-B-I00 and by H2020-MSCA-RISE-2017 project, reference 777826 NoMADS. We also thank NVIDIA for the Quadro P6000 GPU donation.

References

1. Adler, J., Lunz, S.: Banach wasserstein gan. In: Advances in Neural Information Processing Systems. pp. 6754–6763 (2018)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv:1701.07875 (2017)
3. Aujol, J.F., Ladjal, S., Masnou, S.: Exemplar-based inpainting from a variational point of view. SIAM Journal on Mathematical Analysis **42**(3), 1246–1285 (2010)

4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv:1607.06450 (2016)
5. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. pp. 417–424. SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA (2000). https://doi.org/10.1145/344779.344972, http://dx.doi.org/10.1145/344779.344972
6. Burlin, C., Le Calonnec, Y., Duperier, L.: Deep image inpainting (2017)
7. Cao, F., Gousseau, Y., Masnou, S., Pérez, P.: Geometrically guided exemplar-based inpainting. SIAM Journal on Imaging Sciences **4**(4), 1143–1179 (2011)
8. Cao, Y.e.a.: Unsupervised diverse colorization via generative adversarial networks. In: Machine Learning and Knowledge Discovery in Databases. Springer (2017)
9. Chan, T., Shen, J.H.: Mathematical models for local nontexture inpaintings. SIAM Journal of Applied Mathematics **62**(3), 1019–1043 (2001)
10. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based inpainting. IEEE Trans. on IP **13**(9), 1200–1212 (2004)
11. Demir, U., Unal, G.: Patch-based image inpainting with generative adversarial networks. arXiv preprint arXiv:1803.07422 (2018)
12. Deng *et al*, Y.: Graph laplace for occluded face completion and recognition. IEEE Transactions on IP (2011)
13. Fedorov, V., Arias, P., Facciolo, G., Ballester, C.: Affine invariant self-similarity for exemplar-based inpainting. In: Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. pp. 48–58 (2016)
14. Fedorov, V., Facciolo, G., Arias, P.: Variational Framework for Non-Local Inpainting. Image Processing On Line **5**, 362–386 (2015). https://doi.org/10.5201/ipol.2015.136
15. Getreuer, P.: Total Variation Inpainting using Split Bregman. Image Processing On Line **2**, 147–157 (2012). https://doi.org/10.5201/ipol.2012.g-tvi
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Adv in neural inf processing systems. pp. 2672–2680 (2014)
17. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Adv in Neural Inf Processing Systems. pp. 5769–5779 (2017)
18. Hwang, B.W., Lee, S.W.: Reconstruction of partially damaged face images based on a morphable face model. IEEE TPAMI
19. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graph. **36**(4), 107:1–107:14 (Jul 2017). https://doi.org/10.1145/3072959.3073659, http://doi.acm.org/10.1145/3072959.3073659
20. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
21. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR. vol. 2, p. 4 (2017)
22. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: CVPR. vol. 1, p. 3 (2017)
23. Lin, D., Tang, X.: Quality-driven face occlusion detection and recovery. In: IEEE CVPR

24. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
25. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: ICCV. pp. 2813–2821 (2017)
26. Masnou, S., Morel, J.M.: Level lines based disocclusion. In: Proc. of IEEE ICIP (1998)
27. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS workshop on deep learning and unsupervised feature learning
28. Nguyen, A., Yosinski, J., Bengio, Y., Dosovitskiy, A., Clune, J.: Plug & play generative networks: Conditional iterative generation of images in latent space. arXiv:1612.00005 (2016)
29. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (June 2016)
30. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH 2003
31. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434 (2015)
32. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: Proceedings of The 33rd Intern. Conf. Machine Learning. pp. 1060–1069 (2016)
33. Song *et al*, Y.: Contextual-based image inpainting: Infer, match, and translate. In: ECCV (2018)
34. Vitoria., P., Sintes., J., Ballester., C.: Semantic image inpainting through improved wasserstein generative adversarial networks. In: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,. pp. 249–260. INSTICC, SciTePress (2019). https://doi.org/10.5220/0007367902490260
35. Vo *et al*, H.V.: Structural inpainting. In: 2018 ACM Multimedia Conference (2018)
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. on IP **13**(4), 600–612 (April 2004). https://doi.org/10.1109/TIP.2003.819861
37. Wang *et al*, N.: A comprehensive survey to face hallucination. International journal of CV (2014)
38. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: CVPR. vol. 1, p. 3 (2017)
39. Yeh, R.A., Chen, C., Lim, T.Y., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: CVPR. vol. 2, p. 4 (2017)
40. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention

Fig. 8. Inpainting results on the CelebA dataset: Qualitative comparison with the method [39] (fourth and seventh columns, referenced as SIMDGM), using the two masks shown in the second and fifth columns, is also displayed.

Fig. 9. Inpainting results on the SVHN dataset: Qualitative comparison with the method [39] (fourth and seventh columns, referenced as SIMDGM), using the two masks shown in the second and fifth columns, is also displayed.

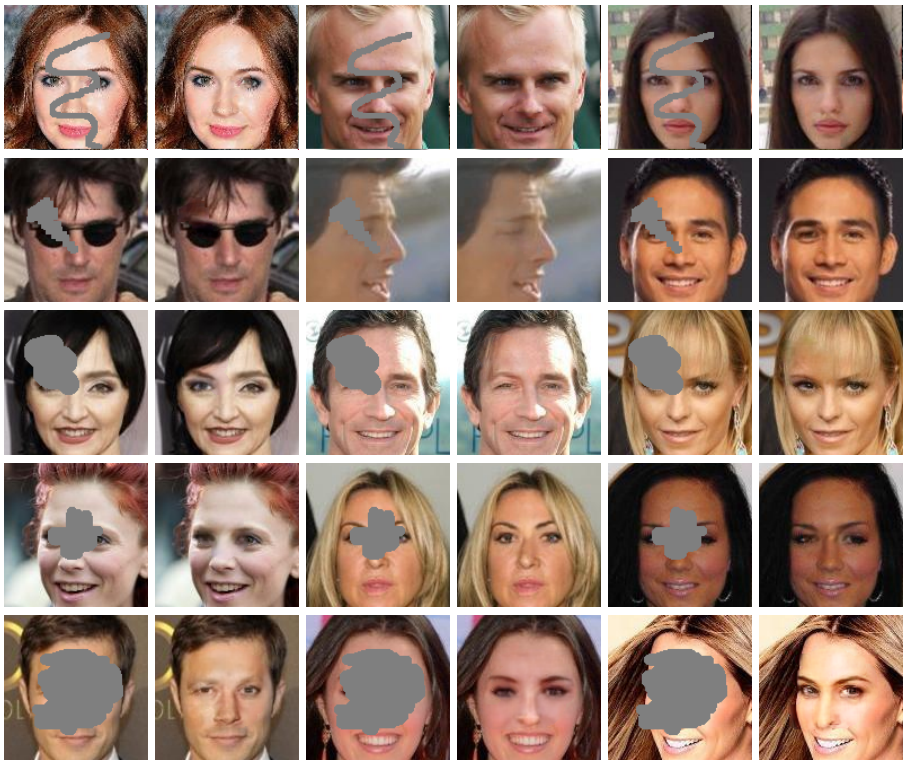

Fig. 10. Inpainting results on the CelebA dataset using the proposed architecture able to create images with higher resolution.

Fig. 11. Some failure cases in CelebA and SVHN dataset.

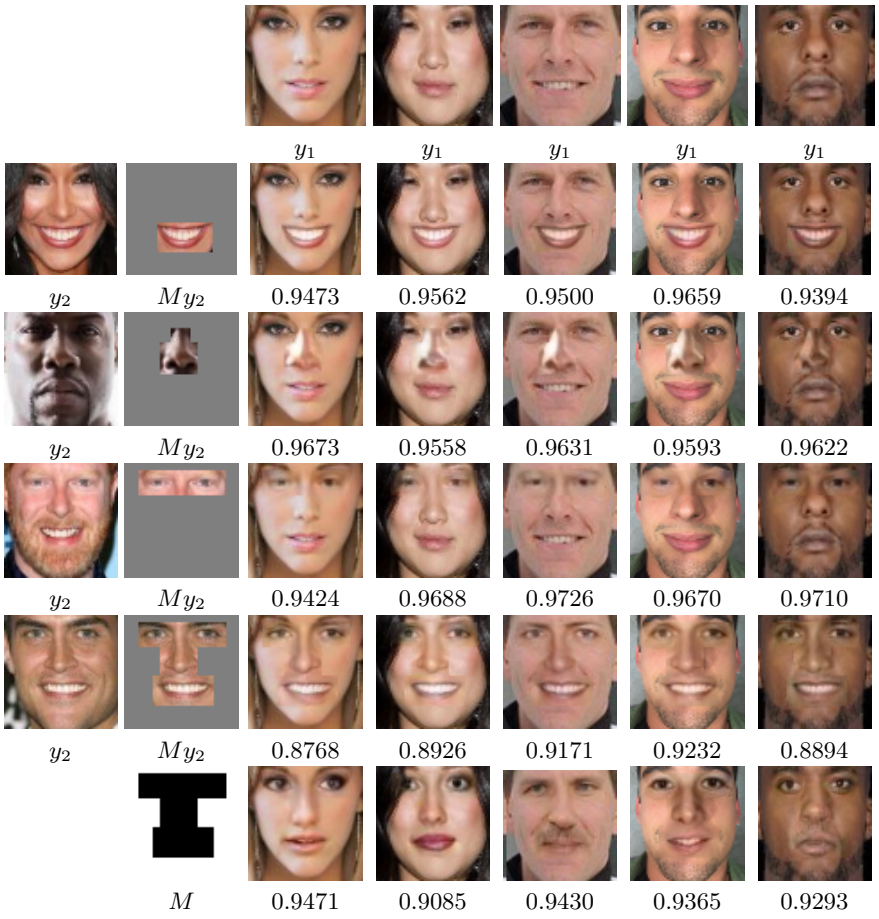

Fig. 12. Face halucination results obtained either using a reference image (from second to fourth rows) or no reference (last row). First row: image y_1 to change or complete, respectively. First column: reference image. Second column: inpainting mask together with the reference region (last row does not have any reference image). The corresponding SSIM is showed under each image.