Improving Consistency and Correctness of Sequence Inpainting using Semantically Guided Generative Adversarial Network

Avisek Lahiri*[1] Arnav Kumar Jain* [2] Prabir Kumar Biswas[3] Pabitra Mitra[4]
Indian Institute of Technology Kharagpur
{[1]avisek, [3]pkb}@ece.iitkgp.ernet.in, {[2]arnavkj95, [4]pabitra}@iitkgp.ac.in

Abstract

Contemporary benchmark methods for image inpainting are based on deep generative models and specifically leverage adversarial loss for yielding realistic reconstructions. However, these models cannot be directly applied on image/video sequences because of an intrinsic drawback- the reconstructions might be independently realistic, but, when visualized as a sequence, often lacks fidelity to the original uncorrupted sequence. The fundamental reason is that these methods try to find the best matching latent space representation near to natural image manifold without any explicit distance based loss. In this paper, we present a semantically conditioned Generative Adversarial Network (GAN) for sequence inpainting. The conditional information constrains the GAN to map a latent representation to a point in image manifold respecting the underlying pose and semantics of the scene. To the best of our knowledge, this is the first work which simultaneously addresses consistency and correctness of generative model based inpainting. We show that our generative model learns to disentangle pose and appearance information; this independence is exploited by our model to generate highly consistent reconstructions. The conditional information also aids the generator network in GAN to produce sharper images compared to the original GAN formulation. This helps in achieving more appealing inpainting performance. Though generic, our algorithm was targeted for inpainting on faces. When applied on CelebA and Youtube Faces datasets, the proposed method results in a significant improvement over the current benchmark, both in terms of quantitative evaluation (Peak Signal to Noise Ratio) and human visual scoring over diversified combinations of resolutions and deformations.

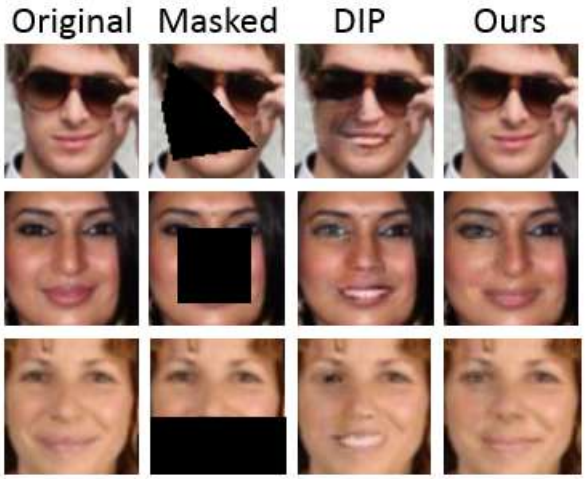

Figure 1. Exemplary success of our model in simultaneously preserving facial semantics(appearance and expressions) and improving inpaiting quality. Benchmark generative models such as DIP [49] are agnostic to holistic facial semantics and thus generate independently realistic, yet structurally inconsistent solutions.

1. Introduction

Semantic inpainting refers to reconstructions of damaged portions of an image using available neighborhood information. In this paper, we are interested in investigating the role of automated dense semantic conditioning to generative adversarial networks (GAN) [12] for the specific task of semantic inpainting. We have focused on the special case of semantic inpainting of faces because faces are tough to inpaint due to presence of finer semantic details. Also, due to contemporary proliferation of multimedia services, video calling is deemed to become a frequent mode of communication and in such video streams, human faces occupy major part of a frame. Thus computer vision guided facial sequence inpainting is the call of the hour. Specifically, we wish to study and improve upon two aspects, viz., a) consistency and b) correctness. Consistency is applica-

*Denotes equal contribution.

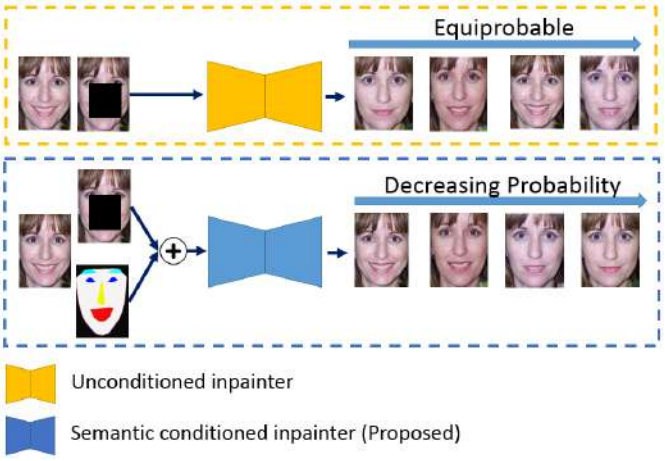

Figure 2. Illustration of multi model image completion possibility of GAN based inpainting methods. Given a corrupted image, an unconditioned inpainting algorithm such as [49] samples from a uniform distribution of viable inpainted images. However, if conditioned by facial semantics, the sampling distribution is biased towards samples which preserve original facial semantics.

ble in case of sequence inpainting, in which we measure the coherence among a group of reconstructed frames. If not accounted for, generative models render abrupt structural changes and unpleasing flickering effects over stationary portions of frames. This is an intrinsic nature of generative model because the forward process of mapping a corrupted section to a valid image manifold is multimodal. An illustration is shown in Figure 2 (Refer to Figure1 for actual comparison of outputs), wherein a generative model has multiple independent and equiprobable options to semantically fill in the corrupted portion of the image. However, if the model is applied on a stream of video frames, then such independent reconstructions renders the sequence unrealistic, because, for example, a smiling face has very low probability of transitioning into a neutral face in next frame. Our intuition to tackle this problem is to constrain the possible models of generation by an auxiliary conditional information. Such conditioning can be in the form of shape priors as used by Fišer *et al.* [11] for synthesis of stylized facial animations or consistency in optical flow field [13] for video style transfer. IIzuka *et al.* only concentrated on consistency of reconstruction at a local and global scale within a single frame [14], but did not address the issue of multimodal image completion in sequence inpainting. We illustrate, both numerically and visually, the inconsistencies in GAN based inpainting methods and offer a simple yet computationally frugal solution to enforce consistency.

Regarding correctness: Correctness refers to a similarity metric quantifying the fidelity of reconstructed output to original version. As we are building upon the recent "DCGAN"[36] based inpainting method by Yeh *et al.*[49](we abbreviate this as 'DIP' in rest of the paper), the quality of reconstruction depends on the success of training the generator to approximate the underlying data distribution. Recent work by [38] shows that conditioning the GAN framework on positional constraints fosters in better sample generation. Our idea of improving upon [49] is to condition the GAN framework with automatically extracted facial semantics and thereby enabling (can be treated as constraining) the generator to generate specific facial components adhering to this conditioning input. In §5.2, we show that this simple yet elegant solution significantly improves quality of generated samples and also plays a pivotal role in achieving consistent reconstruction.

Specifically, our key contributions in the paper are:

1. To the best of our knowledge, this is the first time the dual concept of "correctness" and "consistency" is being explicitly studied in the context of GAN based inpainting.

2. We present a novel semantically guided GAN architecture for generating more appealing images at 64×64 and 128×128 resolutions compared to [49, 12](§4.2).

3. We show that the facial semantic conditional information enables our generative model to disentangle between appearance and pose cues (§5.1).

4. A new framework is presented for assessing consistency of reconstruction by generative models. We show both that our model is able to reconstruct more consistent images compared to DIP (§5.4).

5. We present extensive quantitative, qualitative and subjective evaluations to establish the superiority of samples generated by our proposed GAN model (§5.2) and subsequent inpainted reconstructions (§5.3).

2. Related Works

With the advent of Variational Auto-Encoder (VAE)[21] and GAN [12], there has been a recent surge in interest towards automated image/video generation and subsequent unsupervised feature learning [9, 34, 23]. GANs are known to generate sharper images compared to VAE because VAE is based on the principle of ℓ_2 loss based between generated images and posterior distribution and thereby producing blurry outputs. In [36], the authors recommend an empirically tested, stable architecture framework for GAN training and it became popular as the "DCGAN". While the original GAN formulation allowed the generator network to unrestrainedly sample from generated distribution, recently, researchers have utilized additional conditional information to constrain the output space of GANs for controlled sample generation. Class conditional GAN [31] was the natural extension, wherein the generator was forced to generate

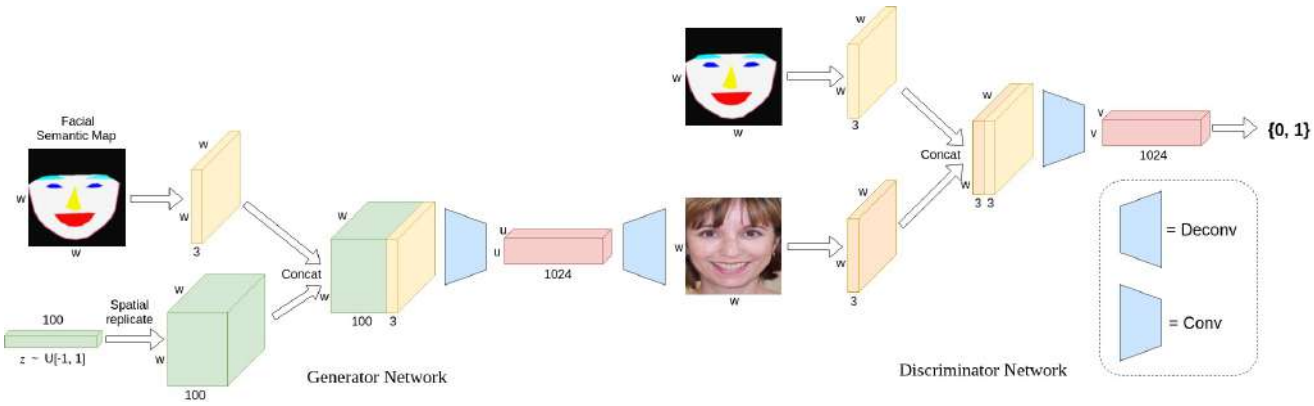

Figure 3. Proposed architecture of training semantically guided GAN. The generator network takes in a noise vector, z, an facial semantic map and generates a facial image. The discriminator is also conditioned on both real/generated images with corresponding facial maps and distinguishes between (image, map) pair belonging to real or generated distribution. The z vector is spatially replicated to a spatial resolution of W×W. Convolutional net of generator section reduces spatial resolution to $U = \frac{W}{16}$. The convolutional of Discriminator reduces spatial resolution of combined image and semantic map to, $U = 4 \times 4$. Detailed explanation is in §4.2

samples of a given class. Denton *et al.* [8] extended this idea in a class conditional Laplacian pyramid GAN setting. Such hierarchical conditioning information aided in better sample quality. Apart from discrete class labels, continuous attributes such as 'smile', 'age', etc., have been used in [18] to interactively modify a given image. Such continuous conditioning have also been leveraged by [4, 51] for making semantically consistent photo editing on faces. Conditioning on natural text was leveraged by Reed *et al.* [37] to directly map an informal description of a flower and bird to pixel space. Later, Zhang *et al.* [50] used a stacked GAN architecture to generate sharper images at higher resolution by conditioning on both text and first level of GAN generated image.

Another contemporary practice is to condition a GAN on an auxiliary image, specially for the task of image-to-image translation [16, 7], style transfer [52, 6, 17], video/sequence generation [29, 43], image denoising [47], real time texture synthesis [26], image super resolution [25], semantic inpainting [27], unsupervised visual domain alignment [3, 40] to list a few.

Conditional information is not only restricted to GANs. Recent works on VAE have also explored such auxiliary conditions for predicting future state from a single static image [44], attribute based face editing [48] and in learning to represent structured output [41]. Conditional inputs have also been used as discriminative regularizers [24] for improving VAE sample quality.

Recently, Reed at al [39] showed that providing sparse localization information to the generator network aids the generator in producing better samples. Our idea is mainly motivated from this observation. However, our approach is computationally more scalable because we use an automated facial fiducial points detection framework based on the real time face alignment with ensemble of regression trees [19]. The authors in [39] instead had to manually mark the parts of the objects before training the conditional GAN. Also, we provide a dense semantic guide to the generator instead of sparse body joint or bounding box locations.

3. Preliminaries

3.1. Generative Adversarial Networks

Generative adversarial network engages two parametrized models, viz., discriminator and generator in a two-player min-max game. Realized as a feed forward neural net, the generator network takes a latent noise vector z drawn from a prior noise distribution $p_z(z)$. Following [49, 12], $z \sim \mathcal{U}[-1, 1]$ (uniform distribution) and generator maps it onto an image, y; $G : z \rightarrow y$. The other network, discriminator, has the task to discriminate samples coming from the true data distribution p_D and the generated distribution, p_G. Specifically, generator and discriminator play the following game on $V(D, G)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[1 - D(G(z))] \quad (1)$$

This min-max game has global optimum when $p_{data} = p_G$ and this happens when both discriminator and generator have enough capacity [12]. Empirically, it has been observed that for generator, it is prudent to maximize $\log(D(G(z)))$ instead of minimizing $\log[1 - D(G(z))]$.

3.2. Conditional generative adversarial networks

In conditional GANs, an extra input, c is also fed to the generator in addition to the z vector and thus $G : (c \times z) \rightarrow$

y. Under this conditioning, the modified objective for GAN training becomes,

$$\min_G \max_D V(D,G) = \mathbb{E}_{x,c\sim p_{data}(x,c)}[\log D(x,c)] + \mathbb{E}_{z\sim p_z(z),c\sim p_{data}(c)}[1 - D(G(z,c))] \quad (2)$$

The GAN framework is flexible in accepting different genres of conditioning inputs such as class labels [33], natural language description [37], localization information[39] and even an entire image [50, 16] or a sequence of images [29]. In our case, we condition the GAN framework with a facial semantic map capturing the pose(head orientation, size) and coarse facial expressions.

4. Method

4.1. Facial Semantic Map Extraction

The first requirement to train our semantic guided GAN framework is to extract facial semantics. In that regard, we make use of the real time face alignment framework of Kazemi *et al.* [19]. However, detection of facial key points alone does not explicitly give semantic information of face. To mitigate this issue, semantically similar facial components are grouped together and given the same RGB color encoding. As shown in Figure3, this semantic map acts an conditional information during GAN training and inference phases.

4.2. Training semantic conditioned GAN

The basic architecture of our proposed conditioned GAN training is shown in Figure 3. We draw the noise prior $z \sim \mathcal{U}[-1,1]$ and tile to it all spatial locations to match the resolution of the conditional map. Next, the tiled z vector and facial semantic maps are concatenated and fed to a conv-deconv [1] network. The convolutional network has 5 layers of convolution of stride 2, kernel size 5 and number of filters doubles at every stage. Next, the transposed convolutional section consists of 4 layers of fractionally strided convolution. Each layer upscales the previous layer's output by 2 and halves the number of filters. The discriminator is also conditioned on the semantic map by concatenating the generated/real images with the corresponding maps. This forces the generator not only to generate realistic samples but also to adhere to the face pose and expressions constraints imposed by the semantic map. Discriminator consists of series of stride 2, kernel size 5 convolutions till the spatial resolution is reduced to 4×4, followed by a linear layer which outputs the probability of joint combination of (face, map) belonging to real/fake distribution. Following the recommendations in [36], we apply Batch Normalization [15] after all layers of the discriminator followed by ReLu non-linearity. Exception is the last deconvolutional layer which is followed by tanh non linearity without Batch Normalization. In case of discriminator, except the first and last layer, Batch Normalization is applied after all the convolutional layers. We use LeakyReLu[28] non linearity activation after each convolutional layer. The final layer is followed by sigmoid non linearity.

[1] Deconv layer should ideally be termed as transposed convolution layer

4.3. Semantic inpainting with appearance and pose constraint GAN

It was shown in [36] that a linear interpolation in the z space results in smooth transition in semantic space. This indicates that semantically similar looking images can be created from 'close' z vectors (here, we define "close" as per Equation 3). We build upon the work of [49], wherein the idea is to find the approximate z vector related to the semantically "closest" natural image compared to the corrupted image. However, in our proposed case, the semantic closeness between corrupted and uncorrupted image is constrained by both appearance and pose criteria; such joint constraint helps in visually correct and structurally aligned inpainitng. Specifically, given a damaged image, I^d, the corruption mask, D, and the semantic map conditioning, c, we aim to find the best fit z vector by iterative optimization of the following loss function,

$$\hat{z} = \underset{z}{\operatorname{argmin}} \{L_{con}(z|I^d, c, D) + \eta L_{per}(z|c)\} \quad (3)$$

where η strikes a trade off between $L_{con}(z|I,c,D)$ and $L_{per}(z|c)$. $L_{con}(\cdot)$ is the contextual loss which penalizes for changing the appearance of the uncorrupted pixels.

$$L_{con}(z|I,c,D) = ||D \odot G(z,c)||_1 , \quad (4)$$

where $\odot$ is the Hadamard product operator. $D(x,y) = 1$ for uncorrupted pixels and 0 otherwise. $L_{per}(\cdot)$ is the perceptual loss coming from the pre-trained discriminator of §4.2 and penalizes if the joint combination of generated image and the semantic map lies away from the natural image manifold.

$$L_{per}(z|c) = \log[1 - (G(z,c))] \quad (5)$$

For a given corrupted image, we start with a random $z \sim \mathcal{U}[-1,1]$ and iteratively update z with stochastic gradient descent to minimize the loss in Equation 3. This enables us to approximately find the $\hat{z}$ vector which approximately maps the corrupted image to its closest semantic neighbor. After calculating $\hat{z}$, the inpainted image, I^{inp}, is formed by overlaying the corrupted image, I^d, with the reconstructed image, $G(\hat{z}|c)$.

$$I^{inp} = D \odot I^d + (1 - M) \odot G(\hat{z}, c) \quad (6)$$

Authors in [49] reported that such overlaying leads to subtle local appearance mismatch and can be mitigated by post processing with Poisson blending [35].

4.4. Implementation Details

Both experiments of GAN training and inpainitng were performed on 64×64 and 128×128 resolutions. We have followed mini-batch stochastic gradient descent optimization with mini-batch size of 64 using Adam [20] optimizer. During GAN training, learning rate was kept constant at 2×10^{-4} and Adam momentum parameters, β_1 and $\beta_2 = 0.5$. During semantic inpainitng, learning rate was set to 5×10^{-2} and iterative back propagation was carried on for 1500 iterations, after which the loss in Equation 3 saturates. z vector was restricted to be within $[-1, 1]$ and η was set to 0.1. Momentum parameters of Adam, were set to β_1=0.9 and β_2=0.99. Same parameters were used for both 64×64 and 128×128 resolutions and all deformation types. For the framework of DIP [49], we have used the parameter settings as reported by the authors.

4.5. Assumptions

Our model assumes the presence of facial semantic map on a corrupted image. As of today, this is not an over restrictive assumption because current state-of-the-art facial landmark localizers [42] are able to perform appreciably under significant occlusions. Also, recent work of Luc *et al.* [32] has shown significant promise of predicting future semantic maps of a scene. Thus, our assumption of having facial is significantly pragmatic. Moreover, we envision that the concepts of this paper will be extended for video inpainting. In videos, temporal redundancy makes it possible to reuse facial maps from a preceding voxel. It is a frequent practice [10, 22] in traditional video coding literature to reuse spatio-temporal information from nearby voxels for appearance and motion vector based error concealment. However, the main question we ask in this paper is, "if *somehow* we provide semantic maps to GAN, will that improve overall inpainting quality and consistency?". Predicting facial semantic map under occlusion is an independent research and deviates readers from central theme of this paper.

5. Experiments

Before delving into experimental analysis, we feel, a justification is required for selecting DIP [49] as our comparing baseline and restricting ourselves to the original GAN formulation of Goodfellow *et al.* [12] for our GAN training. First, DIP, as of today, is the benchmark for GAN based image inpainitng. Superiority of GAN paradigm of inpainting over the previous state-of-the-art method of context encoders [34] has already been shown in [49]. The core objective of this paper is to aware the readers of the intrinsic sequence inpainting inconsistency of DIP and to examine whether semantic conditioning aids in mitigating this drawback. Second, we could have exploited other variants of GAN formulation such as Wasserstein GAN [1], boundary equilibrium GAN [2] and unrolled GAN [30] as these variants have shown to produce better samples compared to original GAN formulation. However, in this paper we are interested in showing that conditional semantic map is effective to improve sample qualities compared to original unconstrained GAN formulation. Amalgamation of better GAN loss function and semantic conditioning might not be a fair comparison to the framework of DIP.

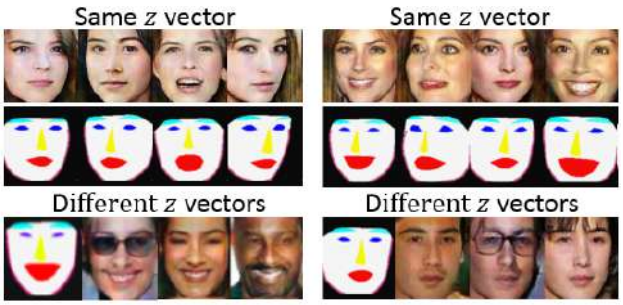

Figure 4. Illustration of our proposed model learning to disentangle facial pose and appearance cues. **Top setting:** Faces generated with same z vector but different semantic maps. **Bottom setting:** Faces generated with different z vectors for a given semantic map. See supplementary document for more examples.

Figure 5. Visual comparison of random samples generated by our semantically conditioned GAN model and DIP[49]. Samples generated by our proposed network are superior in terms of visual quality (§5.2). See supplementary document for more examples.

5.1. Independence of appearance and pose

Main hypothesis of our semantic conditioned GAN is that the generator should learn to disentangle appearance and pose cues for generating images. Intuitively, the semantic map should force the generator to create face with

Table 1. Mean correctness of competing inpainting models measured in terms of PSNR (in dB) at 64×64 and 128×128 resolutions on CelebA test dataset. Ground truth images were corrupted with 4 types of masks, viz., Central, Checkboard, left and Freehand and fed to inpainting networks.

Resolution@64X64				
	Central	Checkboard	Left	Freehand
DIP[49]	24.96	18.51	17.13	24.53
Proposed	26.32	19.97	18.42	25.86
Resolution@128X128				
	Central	Checkboard	Left	Freehand
DIP	23.91	18.36	17.23	23.72
Proposed	24.84	19.12	18.21	24.68

matching head pose and facial expression while two nearby z vectors should result in similar facial textures. In Figure 4(Top setting) we show groups of images which have been generated using same z vectors but different semantic maps. Appearance factors such as gender, skin textures, hair color/styles are preserved; yet the facial expressions and pose closely adhere to the semantic map. In Figure 4(bottom setting), images along a row are generated with different z vectors for a given facial map. Changes in appearances can be appreciated but the facial expression/orientation remains constant. Such independence of appearance and shape is key in success of our inpainting method. Given a semantic map, the z vector mainly focuses on perfecting the appearance.

5.2. Generated image quality and visual turing test

Success of GAN based inpainting framework depends on the capability of the generator in approximating the real image manifold. So, a generator yielding more realistic samples is expected to perform better inpainting. Towards this end, we visually compare the quality of random samples from our proposed semantic conditioned GAN and [49] at resolutions of 64×64 and 128×128. As shown in Figure 5, samples from our proposed model are usually sharper and structurally more coherent. To quantitatively compare the visual appearance of the two models, we perform a visual turing test as followed in [40]. A human annotator is randomly shown total 200 images(100 real and 100 generated) in groups of 20 and asked to label each sample as real or fake. Decisions from 10 annotators are taken. On average, at 64×64 resolution, the classification accuracy is 5.8% higher for DIP($p = 10^{-3}$) and 4.2% higher($p = 10^{-2}$) at 128×128 resolution. Thus, human annotators found it more difficult to distinguish samples from our dataset compared to DIP. This finding advocates the use of semantic conditioning for improving GAN samples without any significant overhead of architecture and loss function modification.

5.3. Image inpainting

As we interested in face inpainting, we have used the CelebA dataset which contains 202,599 face images. Following [49], we separated 2000 images for testing. GAN training (both [49] and ours) was done on the remaining images.

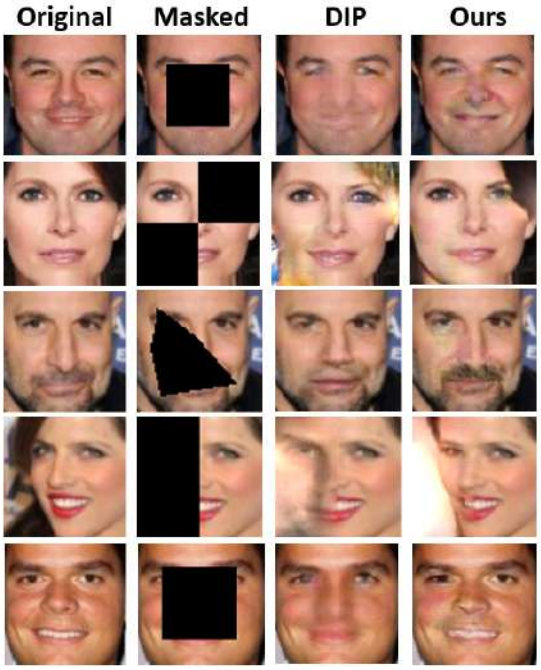

Figure 6. Inpainting comparison with DIP [49] at 128×128 resolution. See supplementary document for more examples.

5.3.1 Correctness of inpainting: Quantitative and visual evaluation

For evaluating correctness, we measure the PSNR between an uncorrupted image and its inpainted version. While reporting PSNR, we have not used Poisson blending post processing because such post processing obscures the true performance of a generative deep model. We have performed extensive experiments on different types of corruption masks as shown in Figure 6 and 7 to compare the generalization capability of each model. The mean PSNR for each setting is reported in Table 1. At 64×64 resolution for Central, Checkboard, Left and Freehand masks, our method outperforms DIP by margins of 1.36dB, 1.46dB, 1.29dB and 1.33 respectively. At 128×128 resolution, corresponding margins are 0.93dB, 0.76dB, 0.98dB and 0.92dB. Statistical significance of the observations reveals p-value $\leq 10^{-5}$ on all cases; this shows that our model significantly outperforms DIP. We show some exemplary inpaintings in Figure6 and Figure7. It can be appreciated that finer facial structural details are preserved in our model. Our recon-

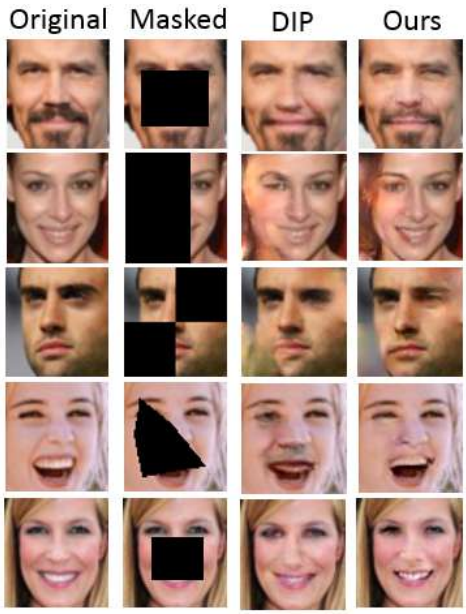

Figure 7. Inpainting comparison with DIP [49] at 64×64 resolution. See supplementary document for more examples.

structions are also sharper due to the intrinsic superiority of the underlying semantic conditioned GAN model.

However, we acknowledge the fact that PSNR (or even Structural Similarity Metric [45](SSIM)) might not be the best metric to compare generative models because these models are not trained explicitly to minimize ℓ_2 loss. Such observations were pointed out in recent works on image super resolution [25, 17]. To complement out findings in Table 1, we perform a human visual testing experiment. Each subject is shown the original image, the corrupted version and the two inpainted images without revealing the identity of the underlying algorithm. The subject has to vote for the inpainted image with better visual quality. Each subject was shown a random selection of 100 images. In a study with 10 participants, our algorithm selected 69.7% of times which is significantly better ($p \leq 10^{-5}$) than chance.

5.4. Consistency in inpainting

The strategy behind enumerating consistency of inpainting is to corrupt a given image using different(or same) deformations and pass the corrupted images independently to the inpainting model. Ideally, each of the inpainted images should be coherent with each other. It is to be noted that such study of consistency is not recommended on real videos because there are unknown transformations between two successive frames. We can use an approximate motion compensation [5] to align two frames, but then the evaluation system will have an intrinsic motion compensation noise which is not separable from the generative modeling noise. Thus, we perform this study on pseudo sequences generated from the 2000 test images of CelebA. Examples of pseudo sequences are shown in Figure 8.

Table 2. Mean consistency (Refer to Equation7) on CelebA test set measured in terms of PSNR(in dB). A sequence was randomly perturbed by either one of Random Central, Random Freehand or constant 50% Left masks. Higher consistency is better.

Resolution@64X64			
	Random Central	Random Freehand	Left
DC-PAINT	21.42	22.21	16.08
Proposed	26.14	26.15	16.72
Resolution@128X128			
	Random Central	Random Freehand	Left
DC-PAINT	21.97	22.40	16.10
Proposed	23.81	25.60	17.15

To formalize, given an uncorrupted image, I^u, we create a sequence comprising of N different (or same) corrupted images, $I^u_{c_i}$ given by, $I^u_{c_i} = D_i(I_u)\ i \in \{1, 2, .., N\}$; $D_i(\cdot)$ is corruption operator on I^u. Following Equation 3, for each $I^u_{c_i}$ we converge at a $\hat{z}_i$ and get the inpainted image, $G(\hat{z}_i)$, from the generative model. For calculating consistency, we enumerate PSNR between all possible pairwise inpainted frames in the sequence. Consistency, η^{Iu}, for the pseudo sequence seeding from I^u, is calculated as,

$$\eta^{I_u} = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N} \sum_{j=1;j\neq i}^{N} PSNR(G(\hat{z}_i),\ G(\hat{z}_j)). \qquad (7)$$

In Table 2 we report the average values of consistency calculated over the 2000 sequences at 64×64 and 128×128 resolution with different deformation masks as shown in Figure 8. Random center and Freehand masks depict the condition where each frame in sequence is corrupted by a different deformations (center mask varies from 50%-70%, Freehand mask corrupts random 25% in 3 different freehand shapes). Constant left mask corrupts the left 50% of a given frame for all frames in the pseudo sequence.

From Table 2, we see that our method is more consistent in reconstruction under all the different deformations. At 64×64 resolution, for random central crop, mean consistency for our method is 4.62 dB higher than that of DIP . The corresponding margins are 3.94 dB and 0.64 dB for random Freehand and left masks respectively. At resolution of 128×128, the average margins of success for our method are 1.84 dB, 3.2 dB and 1.05 dB respectively. Statistical significance of the observations reveals p-value $\leq 10^{-5}$ for each of the deformation settings and on both 64×64 and 128×128 resolution. Thus our proposed reconstructions are significantly more consistent than DIP. To appreciate the numerical findings, we also show example cases in Figure 8. It can be seen that DIP often fails to maintain consistency of the subtle yet important semantics such as facial

Figure 8. Comparison of consistent sequence inpainting. **Top row:** Pseudo sequence by corrupting a given image of CelebA dataset with random deformations. **Middle Row:** Inpainted output sequence by DIP [49]. **Bottom Row:** Inpainted output sequence by proposed model. Our model is more consistent in maintaining facial expressions and textures. See supplemental GIFs for better appreciation of consistency.

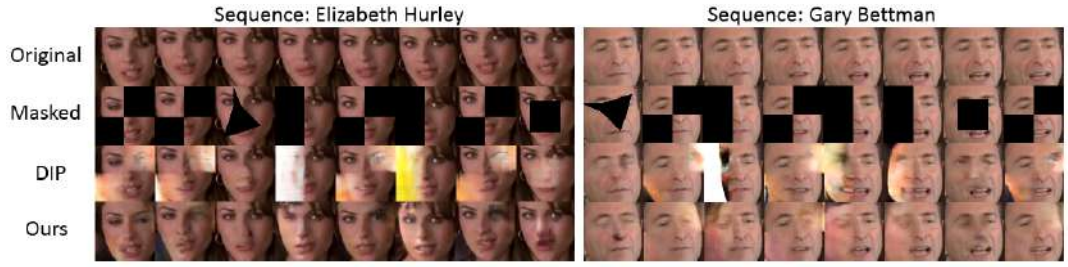

Figure 9. Visual comparison of inpainting on Youtube Faces video sequences.

Table 3. Comparison of PSNR (in dB) on Youtube Faces videos

Sequence	DIP	Ours
Gary Bettman	22.48	26.15
Elizabeth Hurley	13.05	23.67

expression (smile/neutral face), extent of eye opening, skin texture. Independently, each of the reconstructed frames by DIP might be acceptable, but when perceived as a sequence , the performance lacks realism due to abrupt change of facial semantics. Our model, however, faithfully retains not only the pose and expressions but also the skin texture of a subject. Such observation strongly bolsters our hypothesis that semantic guide is crucial in incorporating consistency in generative models.

6. Video inpainting on Youtube Faces

As a proof of concept of viability of our model for video inpainting, we conducted preliminary experiments on the challenging Youtube Faces [46] dataset with the pretrained GAN models on CelebA. It is to be noted that in CelebA, the resolution of faces are bigger than in Youtube Faces dataset. The latter mainly captures celebrity videos in the wild. Thus there is an intrinsic domain difference between the distribution on which we trained GAN models and the distribution in which we are trying to do inpainting. As a result, visual quality of inpainting results are not at par with results on CelebA sequences. But, for this preliminary study, we were interested in examining the raw performances of the GAN models without any domain adaptation or retraining on Youtube videos. We randomly chose video sequences of 2 celebrities, viz. Elizabeth Hurley and Gary Bettman. We extracted the facial region from each frame, resized it to 64×64 and corrupted randomly. The PSNR performance is reported in Table 3 and a short snippet of qualitative visualizations are shown in Figure 9. Even on real life video sequences, our model outperforms DIP, both visually and quantitatively. To our knowledge, this is the first time, a GAN based semantic inpainting framework has been applied on real life videos and our model shows a promising pathway in this regard.

7. Discussion and Conclusion

In this work we presented a simple yet effective framework for improving consistency and correctness in sequence inpainting by conditioning original GAN formulation with semantic mapping. Such conditioning also significantly improved the visual quality of generated samples both at 64×64 and 128×128. We showed that our model learns to disentangle appearance and pose information during sample generation and this helps in preserving both pose and appearance during inpainting. Also, we showed initial success of our model on Youtube Faces videos.

An important lesson here is that generative models are not naively suitable for video/sequence applications due to the multimodal nature of inference pipelines.

The results in this paper thus advocates the use of semantic inpainting for improving GAN performance, specially for video applications. In future we wish to study the combination of advanced variants of GAN [1, 2] with semantic conditioning. Another immediate extension would be to incorporate frameworks for predicting semantic mapping on corrupted portions of frames.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017. 5, 9

[2] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 5, 9

[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *CVPR*, pages 3722–3731, 2017. 3

[4] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016. 3

[5] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. *CVPR*, 2016. 7

[6] A. J. Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*, 2016. 3

[7] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. *arXiv preprint arXiv:1707.09405*, 2017. 3

[8] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, pages 1486–1494, 2015. 3

[9] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 2

[10] M. Ebdelli, O. Le Meur, and C. Guillemot. Video inpainting with short-term windows: application to object removal and error concealment. *IEEE Transactions on Image Processing*, 24(10):3034–3047, 2015. 5

[11] J. Fišer, O. Jamriška, D. Simons, E. Shechtman, J. Lu, P. Asente, M. Lukáč, and D. Sỳkora. Example-based synthesis of stylized facial animations. *ACM Transactions on Graphics (TOG)*, 36(4):155, 2017. 2

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 1, 2, 3, 5

[13] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang, X. Zhu, Z. Li, and W. Liu. Real-time neural style transfer for videos. In *CVPR*, pages 783–791, 2017. 2

[14] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017. 2

[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 4

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, pages 1125–1134, 2016. 3, 4

[17] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 3, 7

[18] T. Kaneko, K. Hiramatsu, and K. Kashino. Generative attribute controller with conditional filtered generative adversarial networks. In *CVPR*, pages 6089–6098, 2017. 3

[19] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pages 1867–1874, 2014. 3, 4

[20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[22] W.-Y. Kung, C.-S. Kim, and C.-C. Kuo. Spatial and temporal error concealment techniques for video transmission over noisy channels. *IEEE transactions on circuits and systems for video technology*, 16(7):789–803, 2006. 5

[23] A. Lahiri, K. Ayush, P. K. Biswas, and P. Mitra. Generative adversarial learning for reducing manual annotation in semantic segmentation on large scale miscroscopy images: Automated vessel segmentation in retinal fundus image as test case. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 42–48, 2017. 2

[24] A. Lamb, V. Dumoulin, and A. Courville. Discriminative regularization for generative models. *arXiv preprint arXiv:1602.03220*, 2016. 3

[25] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *CVPR*, pages 4681–4690, 2016. 3, 7

[26] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, pages 702–716. Springer, 2016. 3

[27] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. *NIPS Workshop on Adversarial Learning*, 2016. 3

[28] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, volume 30, 2013. 4

[29] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *ICLR*, 2016. 3, 4

[30] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016. 5

[31] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[32] N. Neverova, P. Luc, C. Couprie, J. Verbeek, and Y. LeCun. Predicting deeper into the future of semantic segmentation. *arXiv preprint arXiv:1703.07684*, 2017. 5

[33] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In *ICML*, pages 2642–2651, 2017. 4

[34] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 2, 5

[35] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. In *ACM Transactions on graphics (TOG)*, volume 22, pages 313–318. ACM, 2003. 4

[36] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016. 2, 4

[37] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In *ICML*, 2016. 3, 4

[38] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In *NIPS*, pages 217–225, 2016. 2

[39] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In *NIPS*, pages 217–225, 2016. 3, 4

[40] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. *CVPR*, pages 2107–2116, 2017. 3, 6

[41] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, pages 3483–3491, 2015. 3

[42] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, pages 3476–3483, 2013. 5

[43] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *NIPS*, pages 613–621, 2016. 3

[44] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*, pages 835–851. Springer, 2016. 3

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[46] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. *CVPR*, pages 529–534, 2011. 8

[47] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Isgum. Generative adversarial networks for noise reduction in low-dose ct. *IEEE Transactions on Medical Imaging*, 2017. 3

[48] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, pages 776–791. Springer, 2016. 3

[49] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In *CVPR*, pages 5485–5493, 2017. 1, 2, 3, 4, 5, 6, 7, 8, 11, 14, 15, 16, 17, 18, 19, 20

[50] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *CVPR*, pages 5077–5086, 2016. 3, 4

[51] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, pages 597–613. Springer, 2016. 3

[52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 3

Additional Results

8. Visualization: Independence of appearance and pose

As mentioned in Section 5.1 of main paper, independence of appearance and pose refers to the fact that our generator learns to disentangle appearance and pose cues for generating images. In Figure 10, we show example cases in which different faces are generated for a given semantic map. In Figure 11, we show example cases in which similar looking faces are generated from the same z vector but conditioned on different facial maps.

9. Visualization : Inpainting performance

In Figures 12 and 13 we visually compare some cases of semantic impainting on CelebA dataset at 64×64 resolution by our model and DIP [49]. In Figures 14 and 15 we show comparisons of inpainting at 128×128 resolution.

10. Visualization : Consistency of inpainting

In Figures 16 and 17 we visually compare consistency of inpainting of a pseudo sequences at 64×64 resolution. Figures 18 and 19 show the visualizations for 128×128 resolution. Given an uncorrupted image, a pseudo sequence is created by deforming the original image with different(or same) masks. Refer to Section 5.4 in main paper for more details.

11. Visualization: Quality of generated samples from GAN

In Figures 20 and 21 we show some samples generated by our semantically conditioned GAN at 64×64 and 128×128 resolution respectively. In Figures 22 and 23 the corresponding samples from DCGAN architecture used by DIP are shown. Qualitatively, sample qualities from our GAN model is better. Refer to Section 5.2 of main paper for a detailed analysis.

*Denotes equal contribution.

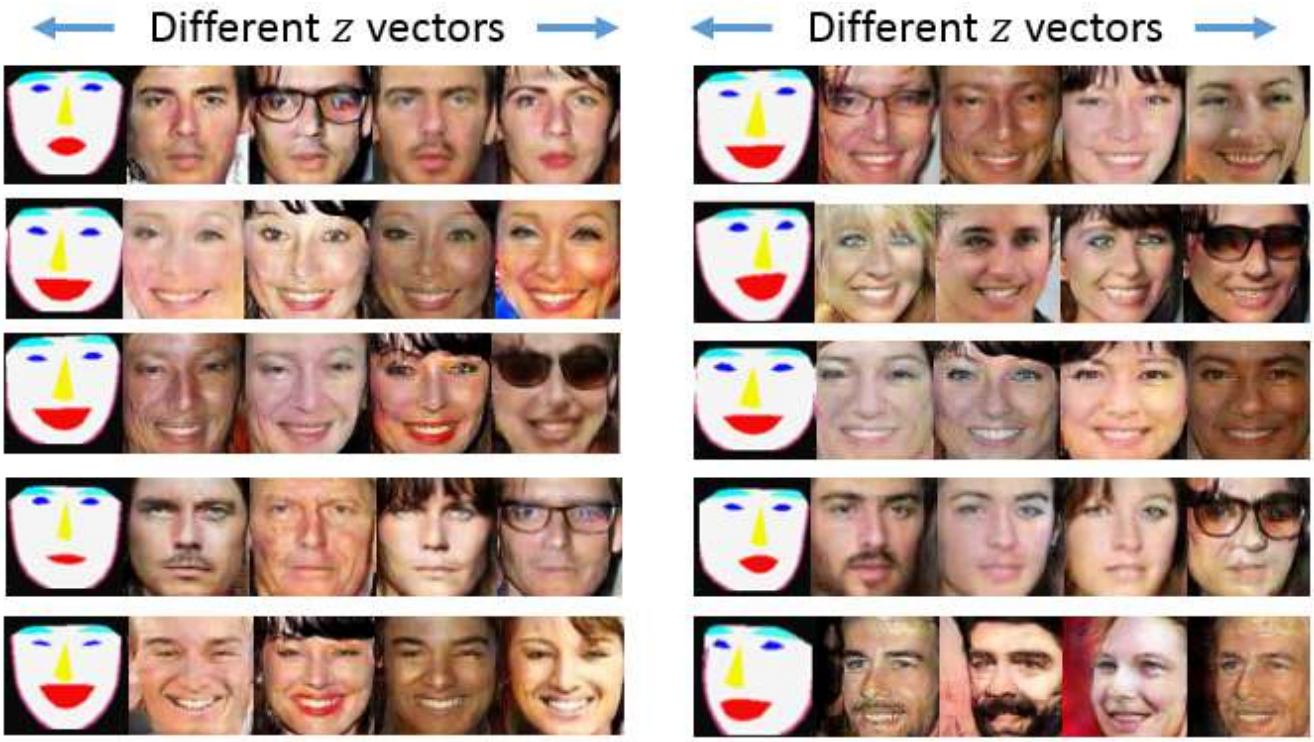

Figure 10. Faces generated for different z vectors but a given semantic facial map. Each set of images is generated using different z vectors but conditioned on a given facial map. Note, how the appearance of the faces changes across the columns for a given set, but the facial expression and orientation is modulated by the conditioning map.

Figure 11. Faces generated for different facial semantic maps but a given z vector. Each set of faces is created with same z vector but different facial semantic map. Note, how facial expressions and orientations are modulated by conditioning maps but facial appearance cues such as texture, gender, hair color etc., are preserved.

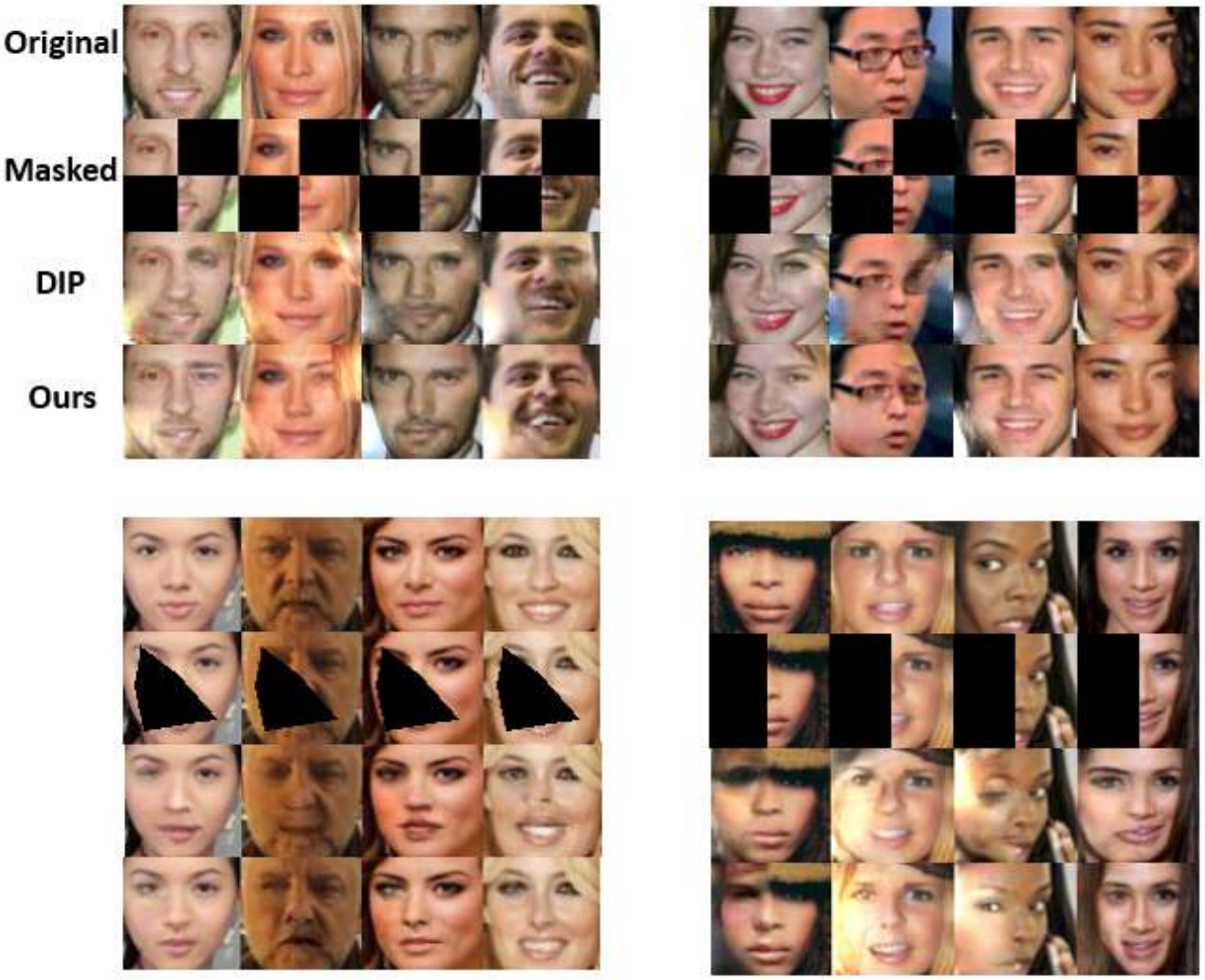

Figure 12. Comparison of inpainting at 64×64 resolution with DIP [49]. **1^{st} row:** Original image; **2^{nd}** row: Damaged image; **3^{rd} row:** Inpainting by DIP; **4^{th} row:** Inpainting by our method.

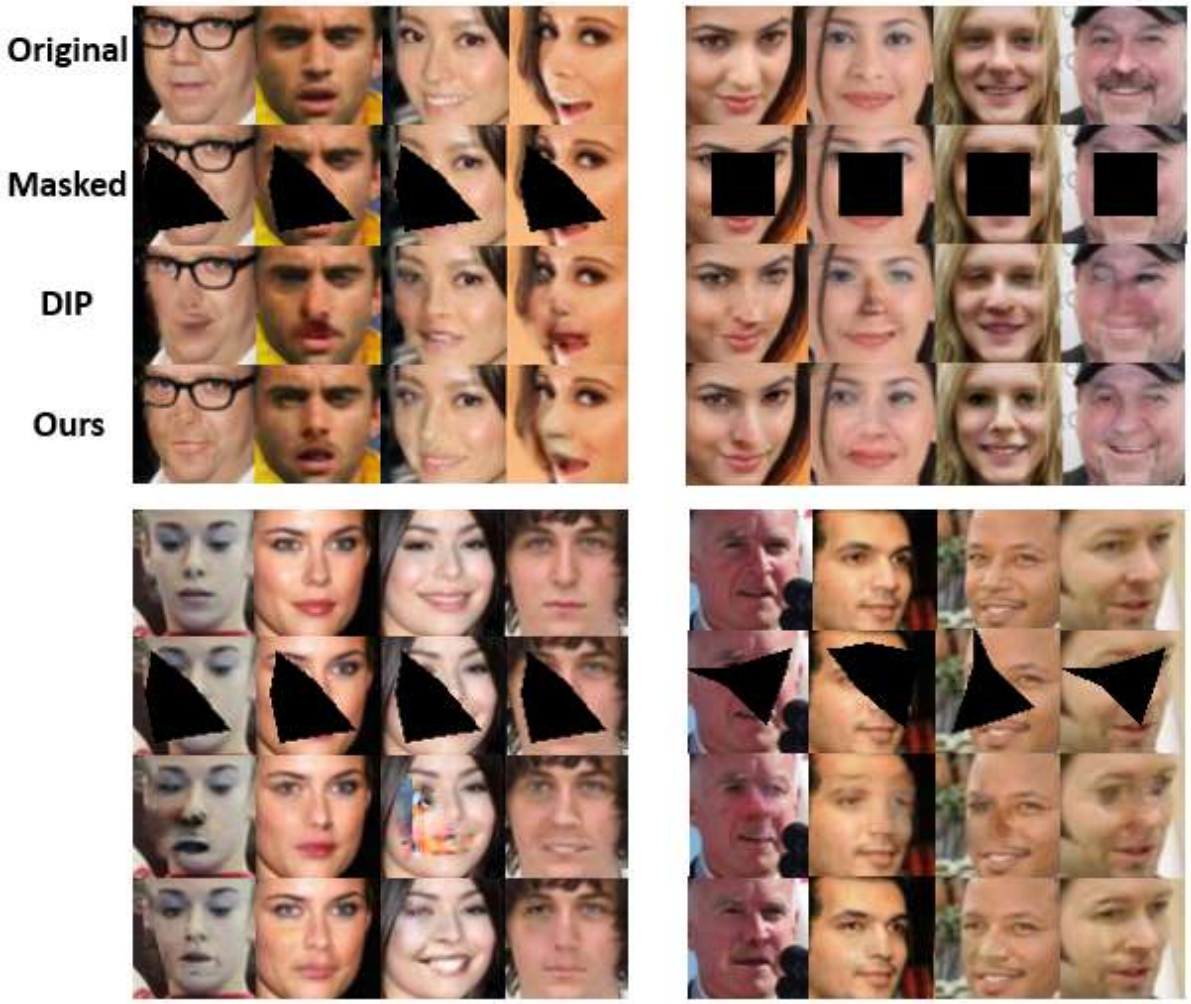

Figure 13. Comparison of inpainting at 64×64 resolution with DIP [49]. **1**st **row:** Original image; **2**nd row: Damaged image; **3**rd **row:** Inpainting by DIP; **4**th **row:** Inpainting by our method.

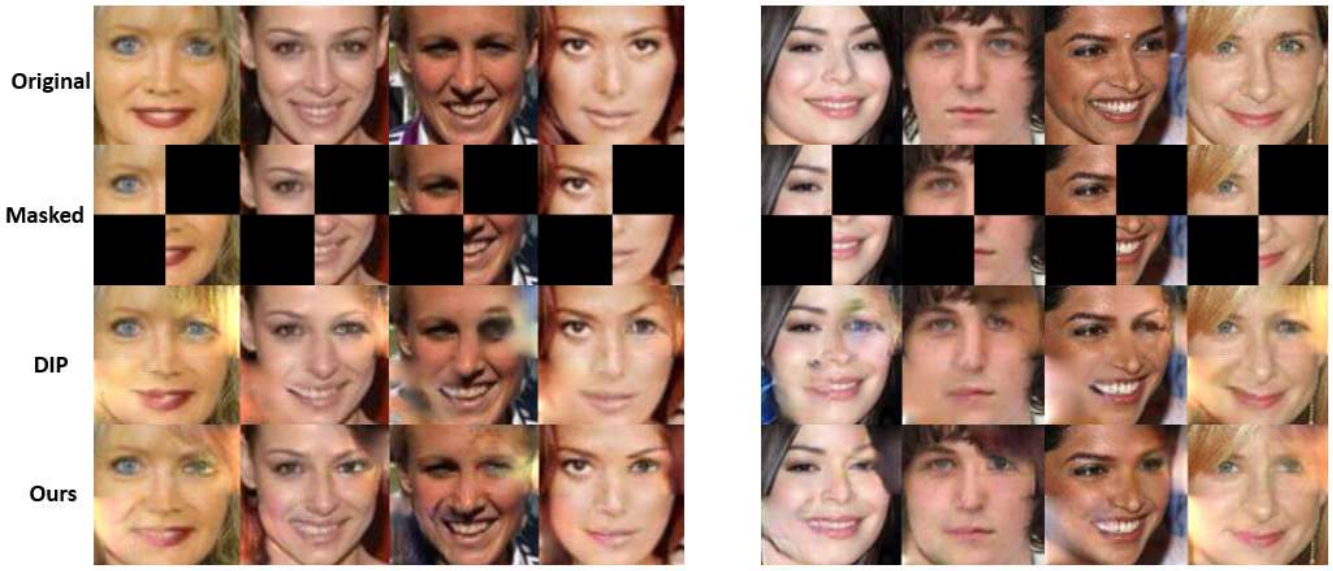

Figure 14. Comparison of inpainting at 128×128 resolution with DIP [49]. **1**st **row:** Original image; **2**nd row: Damaged image; **3**rd **row:** Inpainting by DIP; **4**th **row:** Inpainting by our method.

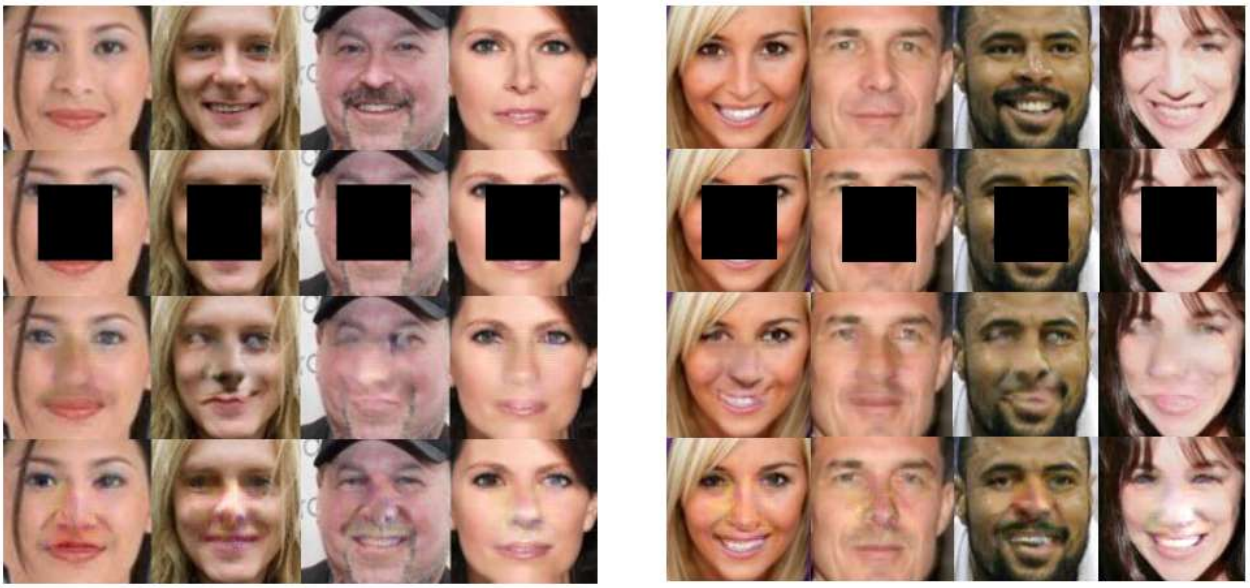

Figure 15. Comparison of inpainting at 128×128 resolution with DIP [49]. **1**st **row:** Original image; **2**nd row: Damaged image; **3**rd **row:** Inpainting by DIP; **4**th **row:** Inpainting by our method.

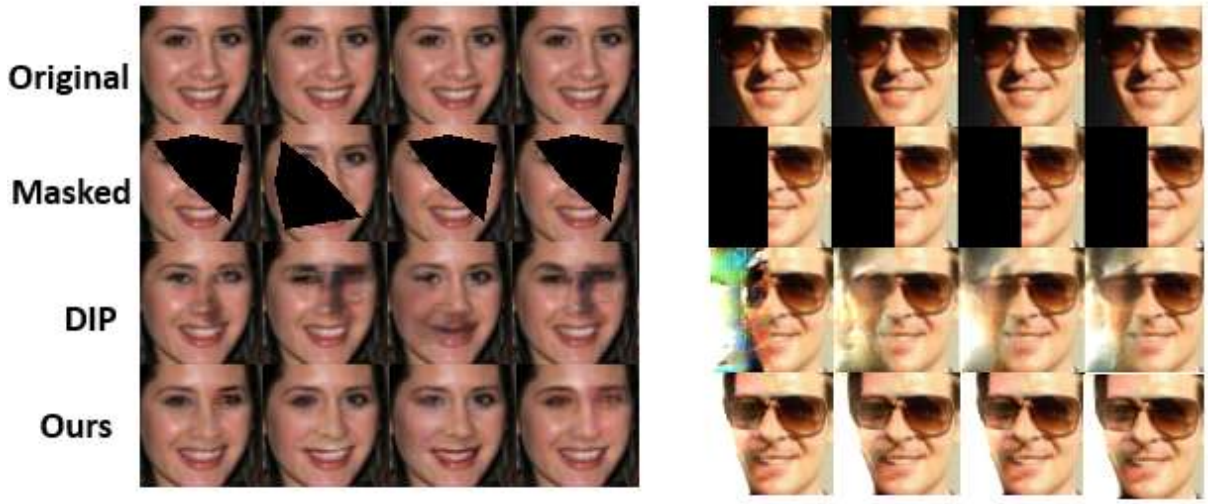

Figure 16. Comparison of consistency of inpainting with DIP [49] at 64×64 resolution. **1**st **row:** Original image; **2**nd row: Damaged image; **3**rd **row:** Inpainting by DIP; **4**th **row:** Inpainting by our method. It can be appreciated visually that our method is more consistent.

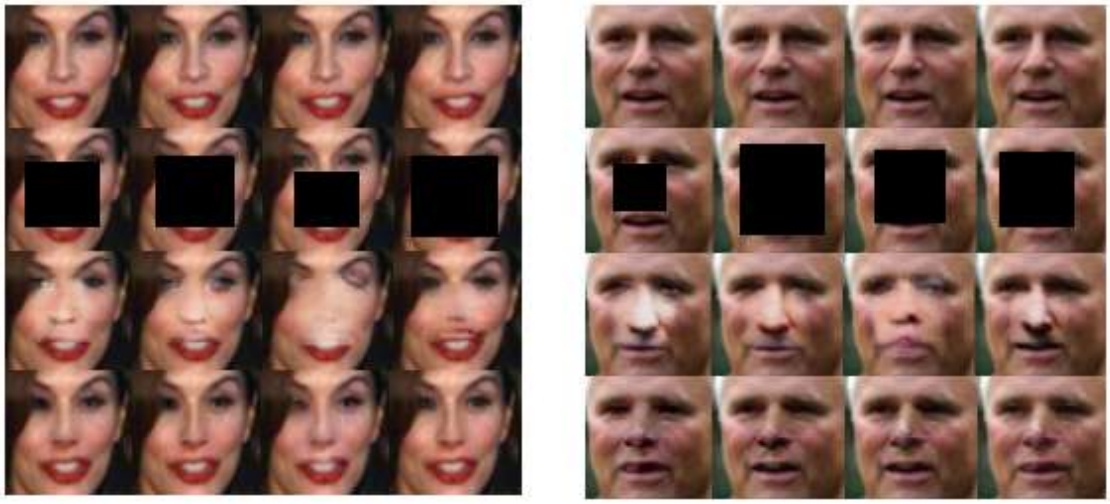

Figure 17. Comparison of consistency of inpainting with DIP [49] at 64×64 resolution. **1**st **row:** Original image; **2**nd row: Damaged image; **3**rd **row:** Inpainting by DIP; **4**th **row:** Inpainting by our method. It can be appreciated visually that our method is more consistent.

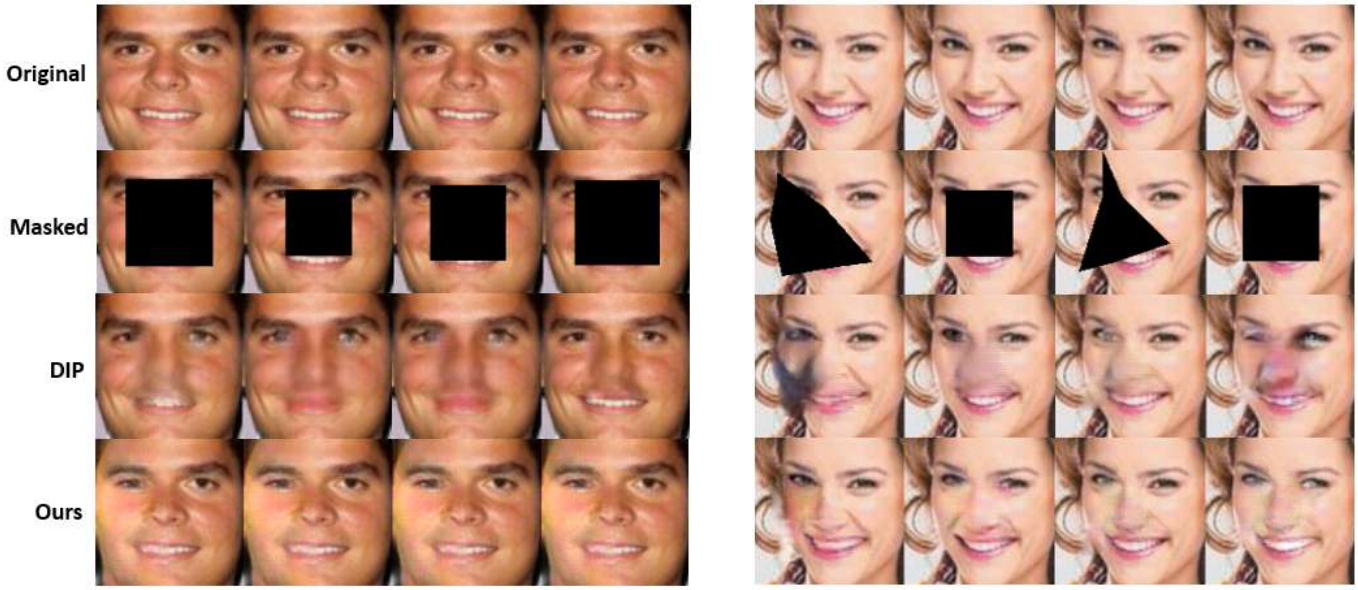

Figure 18. Comparison of consistency of inpainting with DIP [49] at 128×128 resolution. **1**st **row:** Original image; **2**nd row: Damaged image; **3**rd **row:** Inpainting by DIP; **4**th **row:** Inpainting by our method. It can be appreciated visually that our method is more consistent.

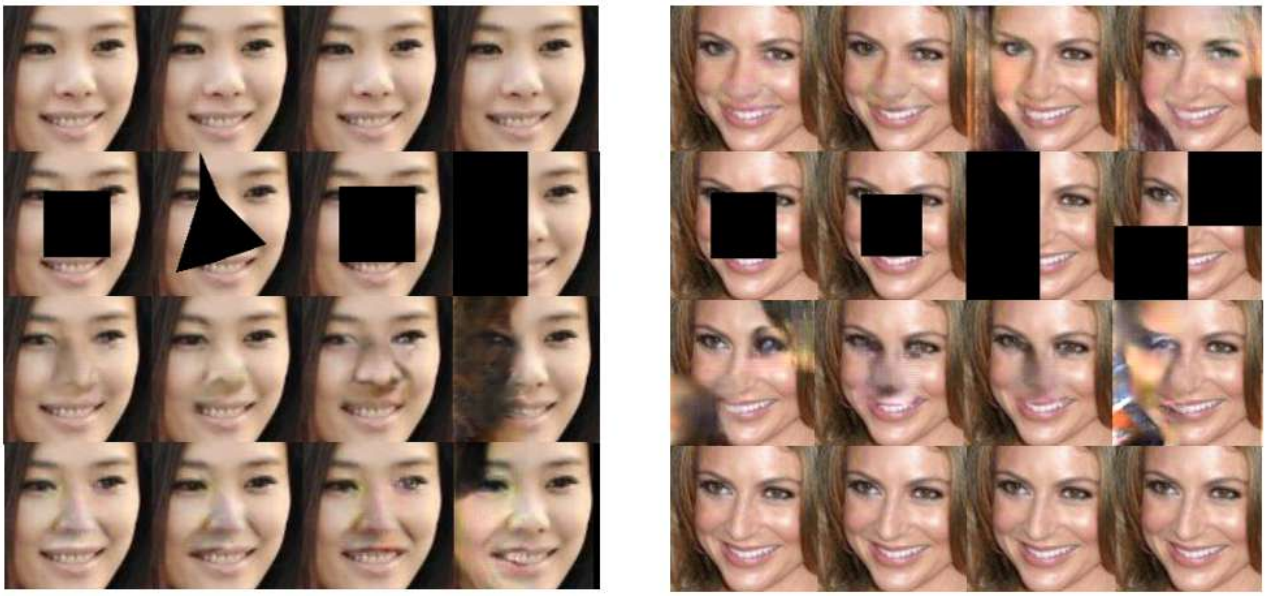

Figure 19. Comparison of consistency of inpainting with DIP [49] at 128×128 resolution. **1**st **row:** Original image; **2**nd row: Damaged image; **3**rd **row:** Inpainting by DIP; **4**th **row:** Inpainting by our method. It can be appreciated visually that our method is more consistent.

Figure 20. Samples from our proposed semantic conditioned GAN model at 64×64 resolution. Visual quality of samples is better than those produced by DIP [49].

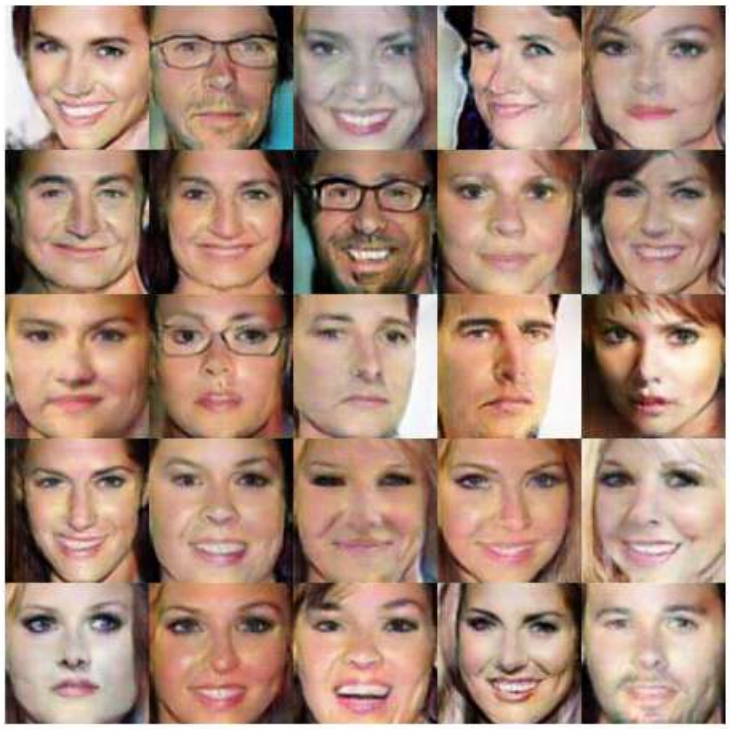

Figure 21. Samples from our proposed semantic conditioned GAN model at 128×128 resolution. Visual quality of samples is better than those produced by DIP [49].

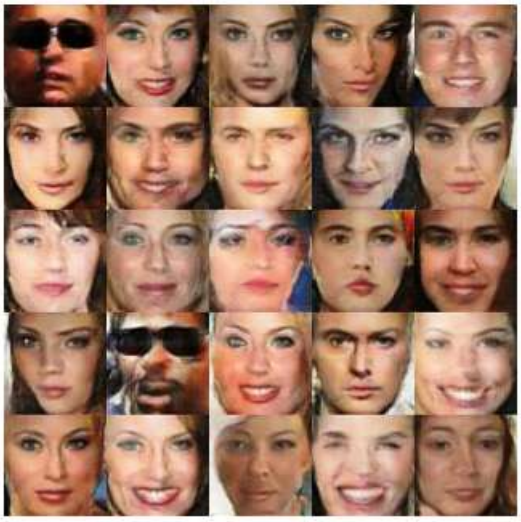

Figure 22. Samples from DCGAN model followed by [49] at 64×64 resolution.

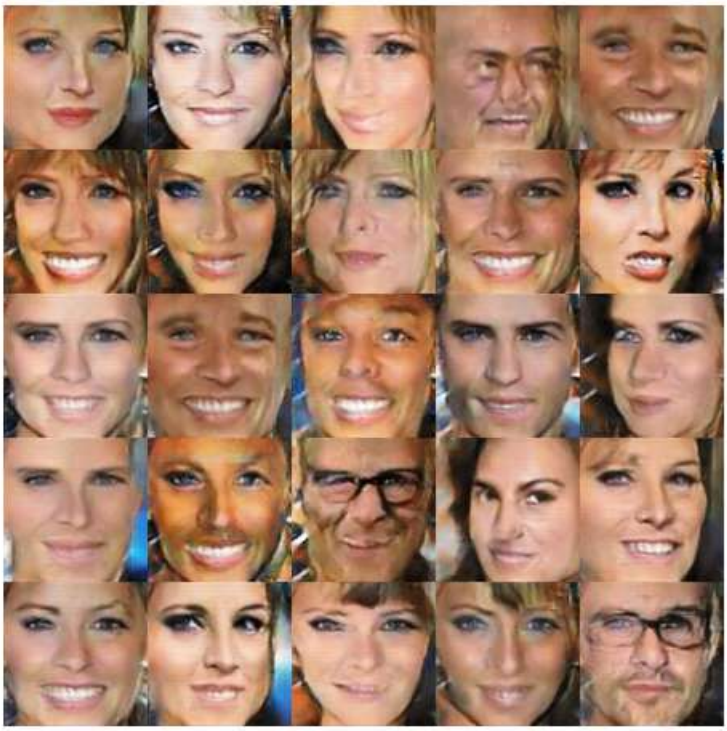

Figure 23. Samples from DCGAN model followed by [49] at 128×128 resolution.