Prior Guided GAN Based Semantic Inpainting

Avisek Lahiri[1]*, Arnav Kumar Jain[2*†], Sanskar Agrawal[1], Pabitra Mitra[1], Prabir Kumar Biswas[1]
[1]Indian Institute of Technology Kharagpur, [2]Microsoft

Abstract

Contemporary deep learning based semantic inpainting can be approached from two directions. First, and the more explored, approach is to train an offline deep regression network over the masked pixels with an additional refinement by adversarial training. This approach requires a single feed-forward pass for inpainting at inference. Another promising, yet unexplored approach is to first train a generative model to map a latent prior distribution to natural image manifold and during inference time search for the 'best-matching' prior to reconstruct the signal. The primary aversion towards the latter genre is due to its inference time iterative optimization and difficulty to scale to higher resolution. In this paper, going against the general trend, we focus on the second paradigm of inpainting and address both of its mentioned problems. Most importantly, we learn a data driven parametric network to directly predict a matching prior for a given masked image. This converts an iterative paradigm to a single feed forward inference pipeline with around 800× speedup. We also regularize our network with structural prior (computed from the masked image itself) which helps in better preservation of pose and size of the object to be inpainted. Moreover, to extend our model for sequence reconstruction, we propose a recurrent net based grouped latent prior learning. Finally, we leverage recent advancements in high resolution GAN training to scale our inpainting network to 256×256. Experiments (spanning across resolutions from 64×64 to 256×256) conducted on SVHN, Standford Cars, CelebA, CelebA-HQ and ImageNet image datasets, and FaceForensics video datasets reveal that we consistently improve upon contemporary benchmarks from both schools of approaches.

1. Introduction

Semantic inpainting refers to filling up of holes or masked regions with plausible pixel values coherent with the neighborhood context. Traditional techniques [4, 17] were mainly successful in inpainting background and scenes with repetitive textures by matching and copying background patches into holes. However, these methods fail on cases where patterns are unique or non-repetitive such as on faces and objects. Recent state-of-the-art generative models usually leverage advancements in deep generative models such as Variational Autoencoeder (VAE) [23] and Generative Adversarial Networks (GAN) [16]. There are mainly two schools of approach, a) *'single-pass-inference'* and b) *'iterative inference'* . The first approach has drawn majority of recent attraction [35, 21, 47, 29, 28] due to its fast inference speed and appreciable performance at high resolution. In general, this paradigm trains on a paired dataset of masked and unmasked images and is initially guided with traditional reconstruction loss over the masked regions. To get the finer details, the next step is to refine the reconstructions with an adversarial loss. The second approach is to first train a generative model such as a GAN on clean/unmasked images and then, based on the masked image at inference time, predict a suitable latent prior to complete the image. From a generative modeling research view point, this approach is appealing because the model is never explicitly guided by reconstruction loss over masked

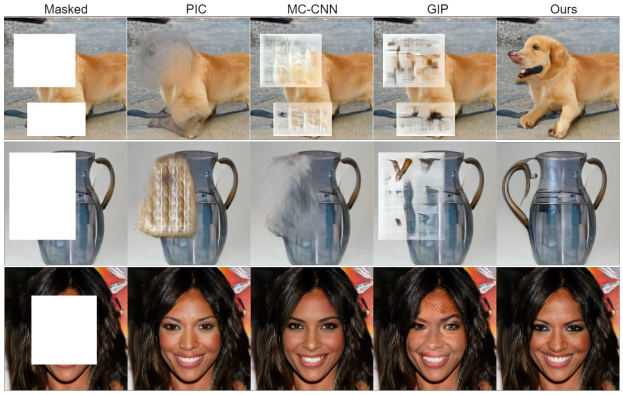


Figure 1: Examples of inpainting on ImageNet (first two rows) and CelebA-HQ (bottom row). On complex multi category dataset such as Imagenet, our network is significantly more capable of recovering semantic parts of the objects to be inpainted compared to state-of-the-art frameworks of MC-CNN [42], PIC [49] and GIP [47]. On simpler structures such as faces we perform comparable (sometimes better) to competing methods. All images are of 256×256. Zoom-in for better visualization.

*Denotes equal contribution.
†Work done when author was at IIT Kharagpur.
Correspondence to: avisek@ece.iitkgp.ac.in, arj@microsoft.com

pixels. However, absence of any reconstruction loss makes it harder to train these models because it has to 'hallucinate' an entire object with no information of masked/damaged pixels. Additionally, it also creates a run-time bottleneck due to iterative optimization during inference. Such an iterative framework prohibits real time applications.

In this paper, we want to encourage our readers to perceive inpainting as a search for best latent prior for a pre-trained generative model. This perspective is quite general and transcends across image and video domains. For this, we adopt the *iterative-inference* genre of approach and primarily aim to massively accelerate inference speed with simultaneous visual quality improvement. For this we follow a two-stage training strategy. In stage 1, we train a GAN network to map a noise distribution to the manifold of natural images. In stage 2, we fix the pre-trained GAN network and train another deep neural network to predict a suitable noise prior from a given masked image. Finally, during inference, we get a matching noise prior (in single feed-forward pass) for a given masked image and use the generator module of the pre-trained GAN to reconstruct an unmasked image.

Single image inpainting has multi-modal completion possibility. For example, a masked lip region can be inpainted to be neutral, smiling, angry etc. This is not an issue for current single image inpainting frameworks in which the primary objective is photo-realism. However, if we want to extend to videos such multi-modal possibilities leads to annoying jittering effects. In this paper, we present a conditional GAN setting [33] in which a structural prior is augmented with the noise prior while generating an image. We show that such structural priors not only help in improving sample quality but also force the generative model to better respect the pose and orientation of the object. To alleviate any human intervention during inference time, we also design a denoising auto-encoder [39] inspired network to automatically compute the structural priors from partially observed data points derived from masked images.

Contemporary single image inpainting models cannot be appreciably applied on videos. Though each frame might be photo-realistic, when viewed as a sequence, there are jitters due to temporal inconsistency of the models. We propose to subdue such inconsistencies with a recurrent neural net based grouped noise prior prediction. Such joint prediction of noise priors enables the network to respect the temporal dynamics of natural videos. Note, in this paper, by video inpainting we are referring to damaged regions of a frame. In traditional video coding literature this is referred to as 'error concealment' in videos [11, 2, 38]. HEVC [37], which is the current standard for video transmission is highly bandwidth efficient but is more susceptible to packet error compared to its successor, H.264 [32]. At the decoder (end user) side, HEVC cannot guarantee end-to-end reproduction. Since video streams are packaged and coded in rectangular patches, packet error manifests as rectangular holes on frames. Thus, video inpainting as an extension of image inpainting can serve as the 'concealment' module at the decoder side to reconstruct damaged blocks using spatio-temporal cues.

Our contributions are summarized as follows:

1. Converting *'iterative-inference'* pipeline of inpainting to a single feed-forward framework and, in-process, achieving up to **800**$\times$ simultaneous visual quality improvement

2. Augmenting structural priors (by automatically deriving from masked images) with noise priors to improve GAN samples which eventually results in better inpainting reconstructions. Such priors also regularize GAN training to respect pose and size of the object to be inpainted

3. Designing a recurrent neural net based grouped prior learning framework for video inpainting. This results in superior spatio-temporal characteristics compared to single-image baselines from both *'iterative-inference'* and *'single-pass-inference'* frameworks and also recent multi-frame approaches

4. Leveraging recent advancements in GAN training to scale up inpainting resolution to 256$\times$256 compared to visually plausible maximum resolution of 64$\times$64 available from current compared to *'iterative-inference'* baselines

2. Related works

Image Inpainting: Traditional image inpainting methods [3, 5, 13, 14] broadly worked with matching patches and diffusion of low level features from unmasked sections to the masked region. These method mainly worked on synthesis of stationary textures of background scenes where it is plausible to find a matching patch from unmasked regions. However, complex objects lack such redundancy of appearance features and thus recent methods leverage hierarchical feature learning capability of deep neural nets to learn higher order semantics of a scene.

Single-pass inference models $\rightarrow$ Initial frameworks of [24, 43] were mainly deep regression networks trained with the usual ℓ_2 reconstruction loss. With the advent of GANs, a common school of approach (*'hybrid'*) [35, 21, 47, 29] is to train a regression network with variants of ℓ_1 or ℓ_2 loss imposed on the masked regions. In Context Encoder(CE) [35] Pathak *et al.*, tried to learn scene representation along with inpainting. Iizuka *et al.* proposed 'Globally and Locally Consistent Image Completion' (GLCIC) in which a inpaint network is pitted against a local and global discriminator. Recently, Yu *et al.* [47] improved upon GLCIC,

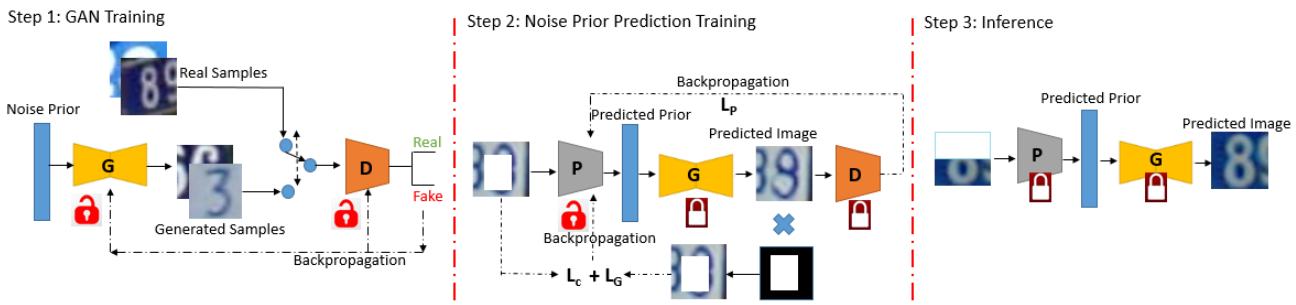


Figure 2: Our basic inpainting model. Step 1: Learn a GAN model. Step 2: Freeze GAN modules (shown as 'lock' symbols) and learn to infer noise prior based on masked input image. Step 3: During inference, given a masked image, predict a matching noise vector and use pre-trained GAN generator(G) to yield final output. The dashed arrows show flow of error gradients during training phase. Unlock symbols denote network modules which are being trained.

by incorporating contextual attention within inpainting network so that the net learns to leverage distant information from unmasked pixels. Wang *et al.* proposed Generative Multi-column CNN (GM-CNN) [42] for parallel synthesis of different image components. In [44], the authors introduce a shift-connection from encoder to decoder in an U-Net architecture. To handle random shaped holes, partial convolution [30] gated convolution [48] were proposed.

'Iterative-inference' baseline $\rightarrow$ Introduced by Yeh *et al.* [45], this approach obviates the need of pixel information inside masked region and instead relies on iterative inference time optimization by leveraging only unmasked pixels.

Video Inpainting: Though a major focus on deep learning based inpainting has been for single image, video inpainting still remains majorly unexplored. From a reconstruction point of view like ours, recently Wang *et al.* [40] presented a two stage video inpainting framework. In first stage, they train 3D CNN at half resolution for a course volumetric prediction followed by a 2D CNN branch for upsampling and refinement. The method still suffers from blurry reconstructions because it had only ℓ_1 reconstruction loss without adversarial loss setting. Some recent works [9, 10, 34, 27] focus on free form video inpainting mainly targeted for video editing.

3. Background

3.1. GAN Basics

Proposed by Goodfellow *et al.* [16], a GAN model consists of two parametrized deep neural nets, viz., generator, G_{θ_G}, and discriminator, D_{θ_D}. The task of the generator is to yield an image, $x \in \mathcal{R}^{H \times W \times 3}$ with a latent noise prior vector, $z \in \mathcal{R}^d$, as input. z is sampled from a known distribution, $p_z(z)$. A common choice [16] is, $z \sim \mathcal{U}[-1,1]^d$. The discriminator is pitted against the generator to distinguish real samples (sampled from p_{data}) from fake/generated samples. Specifically, discriminator and generator play the following two-player min-max game on $V(D_{\theta_D}, G_{\theta_G})$:

$$\min_{G_{\theta_G}} \max_{D_{\theta_D}} V(D_{\theta_D}, G_{\theta_G}) = \mathbb{E}_{x \sim p_{data}(x)}[\log D_{\theta_D}(x)] + \mathbb{E}_{z \sim p_z(z)}[1 - D_{\theta_D}(G_{\theta_G}(z))]. \quad (1)$$

With enough capacity, on convergence, G_{θ_G} fools D_{θ_D} at random [16].

3.2. Iterative inference baseline

According to the *'iterative-inference'* school of approach, given a masked/damaged image, I_d, corresponding to an original image, I, and a pre-trained GAN model, the idea is to iteratively find the 'closest-matching' z vector (starting randomly from $\mathcal{U}[-1,1]^d$) which results in a reconstructed image whose semantics are similar to corrupted image. z is optimized as,

$$\hat{z} = \underset{z}{\operatorname{argmin}} \; L(\overline{M} \odot G_{\theta_G}(z), \overline{M} \odot I) \quad (2)$$

, where M is the binary mask with ones on masked pixels else zeros, $\overline{M} = (1 - M)$, $\odot$ is the Hadamard operator and $L(\cdot)$ is any loss function. Interesting to note is that the loss function never makes use of pixels inside the masked region. Upon convergence, the inpainted image, $\hat{I}$, is given as, $\hat{I} = I_d + M \odot G_{\theta_G}(\hat{z})$, where $I_d = \overline{M} \odot I$.

4. Proposed Method

4.1. Data driven noise prior learning

Sluggish inference of Eq. 2 is a major bottleneck of the *'iterative-inference'* framework. Instead of iteratively optimizing for the noise prior, z, for each test image during inference, we propose to learn an offline parametric model, P_{θ_z}, for predicting z vector directly from masked image, I_d. The parameter set, θ_z, is optimized to jointly minimize the following losses:

Spatially Adaptive Contextual Loss: With this loss we want to penalize any mismatch between the unmasked pixels of I_d and the generated image, $G_{\theta_G}(P_{\theta_z}(I_d))$.

Usually ℓ_1 or ℓ_2 loss on the unmasked pixels can be used for this. However, to mitigate the requirement for any post-processing blending, we want to place more importance to visible pixels near hole boundaries for a better blend of of I_d and $G_{\theta_G}(P_{\theta_z}(I_d))$. Specifically let, S_M be a set of masked pixels; $S_M = \{(x,y)|M(x,y) = 1\}$. We define a spatially adaptive weighting mask, W, whose weight at location (i,j) is given by,

$$W(i,j) = \begin{cases} 0.99^{\ell};\ \ell = \min\limits_{(x,y)\notin S_M} |i-x| + |j-y|, \\ 0;\ \forall (i,j) \in S_M \end{cases} \tag{3}$$

We define, L_c, as:

$$L_c = W \odot |I - G_{\theta_G}(P_{\theta_z}(I_d)|_1. \tag{4}$$

Photo-realism Loss: This loss ensures that the inpainted output lies near the real data manifold and is measured by the log likelihood of belongingness to real class assigned by the pre-trained discriminator. We define, L_r, as:

$$L_r = \log(1 - D_{\theta_D}(G_{\theta_G}(P_{\theta_z}(I_d))). \tag{5}$$

Gradient Difference Loss: This loss is imposed between the masked gradient (horizontal and vertical) matrices of I_d and $G_{\theta_G}(P_{\theta_z}(I_d))$. This compels the network to predict noise priors which yield high frequency retaining samples and to further respect the structure of the original scene.

$$\begin{aligned} L_g = \overline{M} \odot |\nabla_x I_d - \nabla_x G_{\theta_G}(P_{\theta_z}(I_d))| \\ + \overline{M} \odot |\nabla_y I_d - \nabla_y G_{\theta_G}(P_{\theta_z}(I_d))|. \end{aligned} \tag{6}$$

In summary, parameter set, θ_z, is optimized to minimize the combined loss, L_z^{com},

$$L_z^{com} = L_c + \lambda_1 L_r + \lambda_2 L_g \tag{7}$$

, where λ_i's controls the relative importance of each loss factor. After convergence of training of P_{θ_z}, given a masked image, I_d, mask, M, we can get the inpainted output, $\hat{I}$, in one feed forward step. Inpainted image, $\hat{I}$, is given by,

$$\hat{I} = I_d + M \odot G_{\theta_G}(P_{\theta_z}(I_d)). \tag{8}$$

With Eq. 8, the iterative paradigm of [45] is converted to a single feed-forward framework leading to significant inference speedup. We provide a visualization of our proposed framework in Fig. 2.

4.2. Regularization with Structural Priors

We propose to further regularize our network by augmenting structural priors. Structural priors can be any representation which captures the pose and size of the object to be inpainted and thereby compelling the network to yield outputs by respecting such priors. Such additional priors can be seen as a conditional variable to the GAN framework. During GAN training, we condition both the generator and discriminator on such priors. Following this, the noise prior predictor network, P_{θ_z}, has to also optimize θ_z by respecting the structural prior as an additional constraint.

In this paper, without any loss of generalization, we have considered face inpainting with facial landmarks as structural priors.

Estimating structural priors on masked image. We initially tried to use the recent facial keypoint alignment benchmark of Adrian *et al.* [7] for landmark localization on masked images. However [7] gives erroneous detection on masked regions. The masking operation also degrades the localization efficacy on the unmasked pixels (see Fig. 4). This calls for a refinement stage following the estimation by [7].

We follow a refinement strategy inspired from denoising autoencoder [39] in which the idea is to recover the original signal from a noisy signal. On a masked image, j, we denote each of the N_K initially detected (by [7]) keypoints as, $k_i^j := [x_i^j, y_i^j]$, $i \in \{1,2,...,N_K\}$, x_i^j and y_i^j denote horizontal and vertical keypoint coordinates normalized between 0 and 1 based on the face bounding box region. For each k_i^j we have a switch vector, d_i^j such that, if a keypoint falls under masked region, then it is ignored and we set $d_i^j = x_i^j = y_i^j = 0$ before feeding to the prediction network, else $d_i^j = 1$. Let, $\mathbf{K} \in [0,1]^{N_K \times 2}$ encode the keypoints in a matrix. We then learn a parametrized function, $f_{\theta_f} : \mathcal{R}^{N_K \times 2} \rightarrow \mathcal{R}^{N_K \times 2}$ to predict the dropped keypoints (and refine the others) conditioned on detected keypoints. We realize, $f(\cdot)$ with a three hidden layer fully connected neural network. For training, $f_{\theta_f}(\cdot)$ we impose ℓ_2 loss between original and refined keypoints;

$$\theta_f^\star = \min_{\theta_f} \frac{1}{N * N_K} \sum_{j=1}^{N} \sum_{i=1}^{N_k} ||t_i^j - \widehat{k_i^j}||_2^2, \tag{9}$$

, where t_i^j is i^{th} target/ground truth keypoint on j^{th} image, $\widehat{k_i^j}$ is refined keypoint (starting from initially detected, k_i^j) and N is number of training images. The *'switch-on'* (d_i = 1) probability is set is 0.3 i.e., during training we created random masks to cover 70% of facial keypoints and were thus dropped during training $f_{\theta_f}(\cdot)$. The final vector of keypoints, $\mathbf{K_f}$, is given by $f_{\theta_f^\star}(\mathbf{K})$.

4.3. Grouped Noise Prior Learning for Sequences

A naive approach of applying the formulation of Eq. 7 on sequences is to inpaint individual frames independently. However, such an approach fails to learn the temporal dynamics of sequence and thereby yields jittering effects. In this regard, we propose to use a Long Short Term Memory (LSTM) network [20] to jointly predict z vectors for a

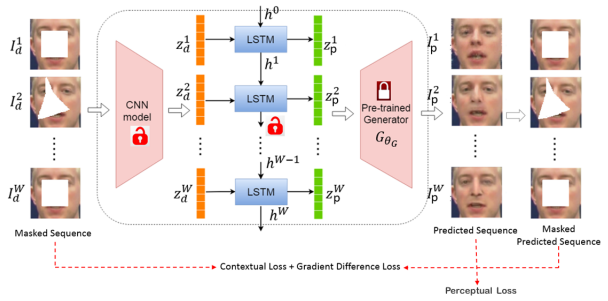


Figure 3: Grouped noise prior learning with a combined LSTM-CNN framework. Unlock sign means parameters to update.

group of W frames at a time. LSTM network has a hidden state, h^t, to summarize information observed upto that time step, t. The hidden state is updated after looking at the previous hidden state and the current masked image (with optional structural priors), leading to temporally coherent reconstructions.

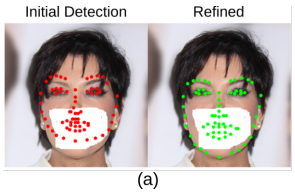
Fig. 3 shows our LSTM based framework for jointly recovering a group of frames. Let, $V = \{I_d^1, I_d^2, ..., I_d^W\}$ be a group of W corrupted successive frames. Initially, each frame I_d^t is passed through a shared CNN module (same architecture as of P_{θ_z}), to get an intermediate representation, z_d^t. z_d^t is the input to the LSTM model at time t and the obtained output is propagated through a feed forward network to get the latent prior z_p^t. The prior, z_p^t, is used for reconstructing I_p^t with the help of the pre-trained generator, G_{θ_G}. We use the loss function in Eq. 7, averaged over the grouped window of W frames to optimize the parameters of LSTM and the shared CNN module. Specifically, the grouped prior loss is defined by, L_z^{gr},

$$L_z^{gr} = \frac{1}{W}\sum_{i=1}^{W} L_z^{com}(I_d^i, I_p^i). \tag{10}$$

Please note, the parameters of pre-trained generator and discriminator are kept frozen.

5. Scaling up resolution

For photo-realistic inpainting at higher resolution we adopt the recent large scale GAN frameworks of Progressive-GAN (PG-GAN) [22] and BigGAN [6]. Since we just need pre-trained generator and discriminator, we use PG-GAN modules for CelebA-HQ and BigGAN modules for Imagenet. The authors of BigGAN have not released the models for CelebA-HQ and thus we adopted the modules from PG-GAN. However, PG-GAN is not scalable for ImageNet because it trains a different GAN for each class. On the contrary, BigGAN uses a single GAN across all 1000 Imagenet categories. We slightly change the nomenclature of our models to indicate high resolution models. For example, M_z is termed as $M_z^{\mathcal{H}}$, $M_{z+S} \rightarrow M_{z+S}^{\mathcal{H}}$ and so on.



(a)

Relative Localization Error (RLE) on LS3D-W test set @ 50% skin pixels masking

Methods	≤ 5%	≤ 10%
Bulat *et al.* (on unmasked)	52.2	85.8
Bulat *et al.* (on masked)	23.2	58.1
Our Refined (on masked)	51.4	84.3

(b)

Figure 4: (a:) Refinement of initial noisy landmark detection by Bulat *et al.* [7] on masked image. Note that our refinement stage rectifies even the landmarks on unmasked regions; (b) Comparing Relative Localization Error of facial keypoints detection on LS3D-W test set with 50% skin pixels masked.

Table 1: Comparison with the baseline *iterative-inference* inpainting baseline of Yeh *et al* [45] at different rates of hole-to-image ratio. Lower FID indicates more visually plausible reconstructions.

Method → Metric	SVHN		Cars		CelebA	
	10%	40%	10%	40%	10%	40%
Yeh *et al.* → PSNR	**21.3**	16.5	**15.1**	12.2	23.8	20.1
Ours (M_z) → PSNR	21.1	**17.0**	14.8	**12.5**	23.4	20.3
Ours (M_{z+S}) → PSNR	-	-	-	-	**24.1**	**21.7**
Yeh *et al. et al.* → FID	3.9	4.8	4.5	5.4	5.8	6.8
Ours (M_z) → FID	**3.6**	**4.3**	**4.1**	**5.0**	5.1	6.7
Ours(M_{z+S}) → FID	-	-	-	-	**4.9**	**6.0**

6. Experiments

6.1. Training details

All the loss functions are optimized with mini batch stochastic gradient descent using Adam optimizer. We implemented our models with Tensorflow 1.8.0, CUDA 9.0 and CUDNN 5.1 and executed on Intel(R) Xeon(R) E5-2650 v4 @ 2.2GHz with NVIDIA Tesla K40 GPU.

6.2. Structural Priors from masked image

We first demonstrate the efficacy of our model to predict the whole set of facial landmarks by observing only a subset of those detected on a given masked image. We used LS3D-W dataset [7] and adhered to the released partition of train/test set. To mitigate the issue of scale variation, we use Relative Localization Error (RLE), which is the ℓ_2 distance between predicted and original keypoint as a fraction of distance (inter-ocular distance) between two eye centers (inter-occular distance, IOD) [15]. In Fig. 4 we report the percentage of keypoints below a certain RLE with 50% of image masked by random shaped holes. It is encouraging to see that the prediction performance of our model on masked faces is comparable to that of [7] on unmasked faces specifically at stringent condition of RLE $\leq 5\%$. Advantage of our refinement state is also shown in Fig. 4.

6.3. Single Image Inpainting

One of the main motivations of the paper was to convert the iterative paradigm of [45] to a 'single-pass' framework, yet adhere to the underlying concept of 'best' z vector search for inpainting. In [45] and a recent follow up work in [46], the authors restrict to 64×64 resolution. This is at-

Figure 5: Benefit of proposed noise prior learning compared to iterative refinement. For each triad, first column is masked image, second column is the initial solution by *'iterative-inference'* baseline of Yeh *et al.* [45] and third column is our single-pass solution. Initial solutions of Yeh *et al.* lie far from natural data manifold and thus requires prohibitively long iterative refinements.

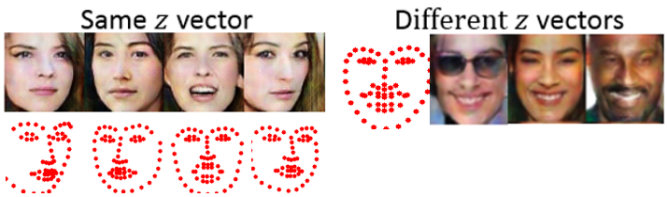


Figure 6: Proposed structural priors enables GAN to disentangle facial pose and appearance cues. **Left:** Faces sampled with same z vector but different structural priors. **Right:** Faces sampled with different z vectors for a given structural prior.

Figure 7: Visualizing inpainting on CelebA (top row), SVHN (bottom-left) and Standford Cars (bottom-right). For each triad, first column is masked image, second column is the final solution by *'iterative-inference'* baseline of Yeh *et al.* [45] and third column is our single-pass solution. Note, with large holes and out-of-plane rotated faces, our face model M_{z+S} is able to reconstruct pragmatic geometry and texture of the face.

tributed to a) ineptitude of 'DCGAN' framework to scale up to higher resolution and b) iterative search for z is not scalable with resolution. So, for a fair comparison we show the benefits of our components at 64×64. However, in Sec. 6.3.3 we will show that our framework also scales to higher resolution and outperforms several *'single-pass-inference'* methods.

Dataset Setup: For comparison with [45] we use the same datasets used by the authors; cropped SVHN[22], Standford Cars[25] and CelebA[31]. SVHN crops are resized to 64×64. On Standford Cars we use bounding box information to extract and resize cars to 64×64. Detected face on CelebA are center cropped to 64×64. On SVHN and Cars, we use the dataset provider's test/train split. On CelebA we test on 10000 samples. Holes of 32×32 at random locations are used for training and testing.

For comparing with recent *'single-pass-inference'* baselines of [47, 49, 42], we select CelebA-HQ [31] and ImageNet at 256×256 resolution. Images are resized to 256×256 for training and testing. During training random rectangular holes with smaller side ranging between 96-128 pixels are used. On CelebA-HQ holes are created at random locations. On ImageNet we use central holes because majority of images have the object of concern near the center of the image. For both CelebA-HQ and ImageNet we keep 10000 images (equally sampled across classes from validation set of ImageNet) for testing.

Evaluation Metrics: For quantitative comparison, we use PSNR (in dB). However, recent works [47, 30, 26] have suggested that reconstruction loss based metrics are not true reflections of photo-realism due to multi-modal image completion possibility. So, the current trend is to report the recently proposed Frechet Inception Distance (FID) metric which correlates well with photo-realism [19]. A lower value of FID is preferred.

6.3.1 Importance of predicting noise prior:

Faster Inference: The most important improvement that we achieve over [45] is a significant inference speedup. In Fig. 5, we compare the initial solution of [45] with our single feed forward solution. Without any mechanism to estimate noise prior from masked image, initial solutions of [45] lie far from real data manifold and thus require time consuming iterative updates. For convergence, total 1000 and 1500 iterations are required by [45] at 64×64 and 128×128 resolution respectively. Our approach just adds a noise predictor network and a negligible (optional) overhead for the structural priors. In Table 4 we compare the actual inference times on GPU. We achieve almost 780× and 820× speedup at 64×64 and 128×128 resolutions respectively.

Better generalization: Proposed framework of learning to predict noise priors from masked images generalizes better to novel images and masks than ad hoc iterative optimization of [45]. This is because, with evolution of training, our network learns to adapt parameters, P_{θ_z}, to map images with similar appearances to closely matched z vectors. Parameter updates for a given image thus implicitly generalizes to images with similar characteristics. On the contrary, every image is treated independently by [45] and chances are high to get stuck in some local minimum yielding inferior reconstructions. From Table 1, we see that our noise prior prediction model, M_z consistently outperforms consistently outperforms [45] in terms of FID. Some visual examples are provided in Fig. 7.

6.3.2 Importance of Structural Priors:

Control of Pose and Expression: During GAN training the structural priors enables the generator to disentangle appearance and pose. A given structural prior forces the generator to match the head pose and facial expression to that of the structural prior. On the other hand, appearance factor such as gender, skin texture are controlled by z vectors.

Table 2: Comparing FID metric of different inpainting methods on 256X256 resolution images of CelebA-HQ and ImageNet datasets. We report performances on masks of 96x96 and 128x128 at random locations. Lower FID metric is better.

Methods		CelebA-HQ		ImageNet	
	Holes→	96x96	128x128	96x96	128x128
GIP		4.9	11.2	8.7	23.3
PIC		4.0	9.1	7.6	38.2
MC-CNN		4.1	10.0	8.0	28.1
Ours ($M_z^{\mathcal{H}}$)		4.1	9.2	-	-
Ours ($M_{z+S}^{\mathcal{H}}$)		3.8	8.5	-	-
Ours ($M_z^{\mathcal{H}}$) (Actual Label)		-	-	4.5	14.5
Ours ($M_z^{\mathcal{H}}$) (Predicted Label)		-	-	6.5	17.8

Figure 8: Examples from ImageNet on which our fine-tuned ResNet-101 predicts correct class label on masked images. We overlay the spatial localization map for the top-1 class using Grad-CAM [36]; red = most important, blue = least important. Notice, even though substantial parts of the objects are missing, the network is still able to attend to unmasked important/complimentary cues for asserting correct class label.

In Figure 6, we show such disentanglement learned by our GAN model.

Improved GAN Samples and Reconstructions: During GAN training, conditioning on structural priors helped us in achieving more photo-realistic samples than [45]. If we assume natural images to belong to a joint distribution, $\mathcal{F}(\mathcal{T},\mathcal{P})$ of texture, $\mathcal{T}$, and pose, $\mathcal{P}$, then an unconditional GAN learns the following; $p_z \xrightarrow{G_{\theta_G}} \mathcal{F}(\mathcal{T},\mathcal{P})$. Under an additional pose constraint, it has to instead learn, $p_z \xrightarrow{G_{\theta_G}} \mathcal{F}(\mathcal{T}|\mathcal{P})$ which drastically reduces the mapping space for $G_{\theta_G}(\cdot)$ and easies training for generator. Additional benefit from structural priors for inpainting is evident from the lower FID scores reported in Tables 1 and 2. Lastly, from Table 4, we see that structural prior module adds negligible computational overhead.

6.3.3 Comparison to 'single-pass-inference' models

Next, we compare with some of the contemporary *'single-pass-inference'* baselines of PIC [49], GIP [47] and MC-CNN [42]. Here, we our high resolution models. We compare on CelebA-HQ and ImageNet at 256×256 resolution. In Table 2, we report FID metrics at different hole-to-image ratios. We perform comparably with other methods on CelebA-HQ but showcase a significant improvement on the more complex ImageNet dataset.

Please note that on ImageNet, the *'BigGAN'* generator uses class conditioned BatchNormalization [12]. So, class information is required during inference time. Initially, this might feel like an overhead. But considering the impressive inpainting performance, this auxiliary information seems worth it. Large number of object category is one of major issues in training a single end-to-end inpainting model on Imagenet. Unlike datasets such as Places2, which are mainly concerned with structures, Imagenet has much varied context. It thus helps to condition the network on auxiliary class information. However, to alleviate human intervention of providing class labels during inference, we also train a network to predict class labels from masked ImageNet. Specifically, we fine tune a ResNet-101 [18] pre-trained on ImageNet on masked images. We achieve a 75.3% top-1 accuracy (starting from 55%) on masked Imagenet validation set compared to 77% on unmasked version.

In Fig. 8 we show instances in which our fine-tuned network predicted correct class label from masked image. From Table 2 we see that the FID metric of our ImageNet model with predicted class labels is still better than competing models. In Fig. 1, we visualize some inpainting examples on ImageNet with correct class label predicted from masked image. Notice, how our model generates significantly better semantics consistent with the primary class of object. In Fig. 10 we show some example cases where the network could not predict correct and, subsequently, yielded unpragmatic reconstructions.

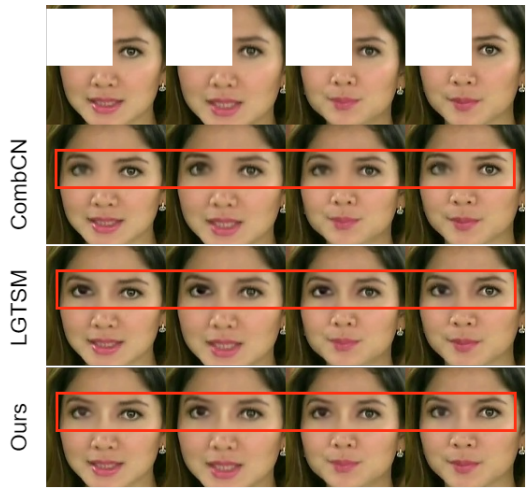


Figure 9: Example of facial video reconstruction on a sequence from FaceForensics dataset. Notice that our reconstructions preserve finer details compared to CombCN and perform comparable to LGTSM without gated convolution or temporal discriminator as used in LGTSM.

6.4. Sequence Inpainting

For sequence inpainting, we select the FaceForensics dataset [1] which has been one of the preferred facial video datasets for recent video reconstruction papers [40, 9, 10]. It contains 1004 face videos from YouTube and the YouTube-8m dataset [1]. Following the settings in [40, 9, 10] all faces

Table 3: Video FID metric by different inpainting methods on FaceForensics video dataset averaged over different mask-to-frame ratios between 10%-50%. Lower FID means better perceptual video quality.

Yeh *et al.* [45]	GLCIC [21]	GIP [47]	LGTSM [9]	3DGated [10]	CombCN [40]	Proposed			
						$M_z^{\mathcal{H}}$	$M_{z+S}^{\mathcal{H}}$	$M_{z+L}^{\mathcal{H}}$	$M_{z+S+L}^{\mathcal{H}}$
0.781	0.762	0.751	0.651	0.670	0.742	0.738	0.710	0.680	0.660

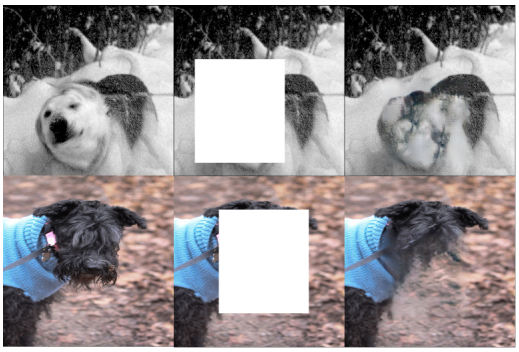


Figure 10: Some fail cases of our inpainting model. We predicted wrong class on each of these images and thus could not generate the true semantics of the main masked object. Left: Original image, Middle: Masked image, Right: Inpainted image,

are center cropped to 128×128 and trained with random rectangular masks in range $[0.35l, 0.5l]$, where $l = 128$. Total 150 videos were used for testing. We use fine-tuned (on FaceForensics frames) 'PG-GAN' modules for this experiment.

Comparing Methods: We compare against single-image frameworks of [45, 21, 47]. We also compare against recent video inpainting frameworks of CombCN [40], 3DGated [9] and LGTSM [10].

Quantitative Evaluation: PSNR calculated on individual frame does not reflect the temporal characteristics of a sequence. Following the settings in [9] we use the recently proposed video-FID metric [41] with I3D [8] pre-trained video recognition CNN. A lower video-FID is better and is an indicative of realistic spatio-temporal characteristics. In Table 3, we compare the average test set video-FID. It also provides an ablation study of different components of our model. It is encouraging to see that even our single image model, $M_z^{\mathcal{H}}$, performs better than the competing single image models. Since there is no temporal guidance, this can be attributed to better spatial reconstruction with *'BigGAN'* generator. With incorporation of structural priors ($M_{z+S}^{\mathcal{H}}$) and LSTM grouped prior ($M_{z+L}^{\mathcal{H}}$) the performance progressively improves. The combined model, $M_{z+S+L}^{\mathcal{H}}$ manifests the best performance. The video inpainting model of CombCN [40] is only trained with ℓ_1 losses without any adversarial refinement. Thus, even though the results are stable, the outputs are blurry and is finally penalized by high video-FID score. Our combined model has comparable performance to that of [9, 10] even though those models use time dimension specific convolutional concepts of temporal shifts and gated 3D kernels with temporal PatchGAN based discriminator. Those modules can be integrated to our model as well, but we leave that for a future work.

Table 4: Comparison of inpainting inference time (in ms). We dramatically improve the inference of our starting iterative baseline of Yeh *et al.* [45]. Our run time is also comparable with contemporary *'single-pass-inference'* methods of PIC [49], GIP [47] and MC-CNN [42]. Note, at 256×256 resolution, we are refering to the *'BigGAN'* generator network.

Res	Yeh *et al.*	PIC	GIP	MC-CNN	M_z (Ours)	M_{z+S} (Ours)
64X64	2175	-	-	-	2.7	2.8
128X128	10750	-	-		11.0	13.2
256X256	Not Converge	70	30	50	68	75

7. Discussion and Conclusion

In this paper we revisited an iterative inference framework for inpainitng with the objective of speeding up inference time. Towards this we showed the importance of data driven noise prior learning which gave about 800× speedup with simultaneous improvement of reconstruction compared to the baseline of [45]. We also extended our model for video inpainting and concepts of structural priors and LSTM driven grouped prior learning were introduced to significantly improve temporal dynamics. We also showed state-of-the-art performance against recent benchmarks in image inpainting and video reconstruction. Our paper instigates a new dimension to perceive inpainting as a search for 'best-matching' latent prior instead of the current trend of encoder-decoder driven 'image refinement' perspective.

We acknowledge that currently our model is not well suited for inpainting natural videos or outdoor scenes. This is not a drawback of our framework in general but a manifestation of inability of current GAN frameworks to synthesize natural scenes. However, with the release of PG-GAN and BigGAN the community is quite optimistic towards complex scene generation with GANs. As such, our framework is extremely modularized to accept any new GAN model and benefit from its generative capability. We leave this as a future work of exploration.

Acknowledgements

This project was funded by Google PhD Fellowship. Authors would like to thank Saurav Basu, for his insightful discussions.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 7

[2] Susanna Aign and Khaled Fazel. Temporal and spatial error concealment techniques for hierarchical mpeg-2 video codec. In *Proceedings IEEE International Conference on Communications ICC'95*, volume 3, pages 1778–1783. IEEE, 1995. 2

[3] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211, 2001. 2

[4] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (ToG)*, 28(3):24, 2009. 1

[5] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co., 2000. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 5

[7] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 4, 5

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 8

[9] Ya-Liang Chang, Zhe Yu Liu, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *ICCV*, 2019. 3, 7, 8

[10] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Learnable gated temporal shift module for deep video inpainting. In *BMVC*, 2019. 3, 7, 8

[11] Yan Chen, Yang Hu, Oscar C Au, Houqiang Li, and Chang Wen Chen. Video error concealment using spatio-temporal boundary matching and partial differential equation. *IEEE Transactions on Multimedia*, 10(1):2–15, 2007. 2

[12] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *ICLR*, 2017. 7

[13] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346. ACM, 2001. 2

[14] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *ICCV*, page 1033. IEEE, 1999. 2

[15] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2385–2392, 2014. 5

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 1, 3

[17] James Hays and Alexei A Efros. Scene completion using millions of photographs. In *ACM Transactions on Graphics (TOG)*, volume 26, page 4. ACM, 2007. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 6

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[21] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017. 1, 2, 8

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 5, 6

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[24] Rolf Köhler, Christian Schuler, Bernhard Schölkopf, and Stefan Harmeling. Mask-specific inpainting with deep neural networks. In *German Conference on Pattern Recognition*, pages 523–534. Springer, 2014. 2

[25] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 6

[26] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4, 2017. 6

[27] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[28] Chang Li, Youdong Ding, Bing Yu, Min Xu, and Qianqian Zhang. Inpainting of continuous frames of old movies based on deep neural network. In *2018 International Conference on Audio, Language and Image Processing (ICALIP)*, pages 132–137. IEEE, 2018. 1

[29] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017. 1, 2

[30] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, pages 85–100, 2018. 3, 6

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 6

[32] James Nightingale, Qi Wang, Christos Grecos, and Sergio Goma. The impact of network impairment on quality of experience (qoe) in h. 265/hevc video streaming. *IEEE Transactions on Consumer Electronics*, 60(2):242–250, 2014. 2

[33] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *ICML*, pages 2642–2651, 2017. 2

[34] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-peel networks for deep video completion. In *ICCV*, 2019. 3

[35] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 1, 2

[36] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*, pages 618–626, 2017. 7

[37] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012. 2

[38] Stephane Valente, Cecile Dufour, Francoise Groliere, and Daniel Snook. An efficient error concealment implementation for mpeg-4 video streams. *IEEE Transactions on Consumer Electronics*, 47(3):568–578, 2001. 2

[39] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103. ACM, 2008. 2, 4

[40] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In *AAAI*, volume 33, pages 5232–5239, 2019. 3, 7, 8

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 8

[42] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *NeurIPS*, pages 331–340, 2018. 1, 3, 6, 7, 8

[43] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *NeurIPS*, pages 341–349, 2012. 2

[44] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *ECCV*, pages 1–17, 2018. 3

[45] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *CVPR*, pages 5485–5493, 2017. 3, 4, 5, 6, 7, 8

[46] Raymond A Yeh, Teck Yian Lim, Chen Chen, Alexander G Schwing, Mark Hasegawa-Johnson, and MinhN Do. Image restoration with deep generative models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6772–6776. IEEE, 2018. 5

[47] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 1, 2, 6, 7, 8

[48] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, pages 4471–4480, 2019. 3

[49] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 1, 6, 7, 8