

# Symmetric Skip Connection Wasserstein GAN for High-resolution Facial Image Inpainting

Jireh Jam<sup>1</sup><sup>a</sup>, Connah Kendrick<sup>1</sup><sup>b</sup>, Vincent Drouard<sup>2</sup><sup>c</sup>, Kevin Walker<sup>2</sup><sup>d</sup>, Gee-Sern Hsu<sup>3</sup><sup>e</sup> and  
Moi Hoon Yap<sup>1</sup><sup>f</sup>

<sup>1</sup>Manchester Metropolitan University, Manchester, U.K

<sup>2</sup>Image Metrics Ltd, Manchester, U.K

<sup>3</sup>National Taiwan University of Science & Technology, Taipei, Taiwan

**Keywords:** Inpainting, Generative Neural Networks, Hallucinations, Realism.

**Abstract:** The state-of-the-art facial image inpainting methods achieved promising results but face realism preservation remains a challenge. This is due to limitations such as; failures in preserving edges and blurry artefacts. To overcome these limitations, we propose a Symmetric Skip Connection Wasserstein Generative Adversarial Network (S-WGAN) for high-resolution facial image inpainting. The architecture is an encoder-decoder with convolutional blocks, linked by skip connections. The encoder is a feature extractor that captures data abstractions of an input image to learn an end-to-end mapping from an input (binary masked image) to the ground-truth. The decoder uses learned abstractions to reconstruct the image. With skip connections, S-WGAN transfers image details to the decoder. Additionally, we propose a Wasserstein-Perceptual loss function to preserve colour and maintain realism on a reconstructed image. We evaluate our method and the state-of-the-art methods on CelebA-HQ dataset. Our results show S-WGAN produces sharper and more realistic images when visually compared with other methods. The quantitative measures show our proposed S-WGAN achieves the best Structure Similarity Index Measure (SSIM) of 0.94.

## 1 INTRODUCTION

Historically, inpainting is an ancient technique that was performed by professional artists to restore damaged paintings in museums. These defects (scratches, cracks, dust and spots) were inpainted by hand to restore and maintain the image quality. The evolution of computers in the last century and its frequent daily use has encouraged inpainting to take a digital format (Efros and Leung, 1999; Bertalmio et al., 2000; Criminisi et al., 2004; Pathak et al., 2016; Liu et al., 2018a; Yang et al., 2017; Yan et al., 2018) as an image restoration technique. Image inpainting aims to fill in missing pixels caused by a defect based on pixel similarity information (Bertalmio et al., 2000).

The state-of-the-art approaches are two categories: conventional and deep learning-based methods. Conventional methods (Efros and Leung, 1999;

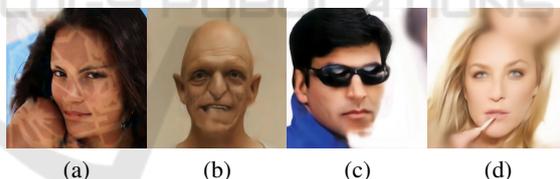


Figure 1: Images showing some issues by state of the art: (a) Poor performance on holes with arbitrary sizes; (b) Lack of edge-preserving technique; (c) Blurry artefacts; and (d) Poor performance on high-resolution images and image completion with mask at the border region.

Criminisi et al., 2004; Barnes et al., 2009; Sun et al., 2005) use image statistics of best-fitting pixels to fill in missing regions (defects). However, these approaches often fail to produce images with plausible visual semantics. With the evolution in research, deep learning-based methods (Pathak et al., 2016; Liu et al., 2018a; Iizuka et al., 2017; Yu et al., 2018; Yan et al., 2018; Yang et al., 2017; Park et al., 2020) encode the semantic context of an image into feature space and fill in missing pixels on images by hallucinations (Yang et al., 2017) through the use of generative neural network. Although deep learning approaches achieve excellent performance in facial in-

<sup>a</sup> <https://orcid.org/0000-0003-2309-4655>

<sup>b</sup> <https://orcid.org/0000-0002-3623-6598>

<sup>c</sup> <https://orcid.org/0000-0001-7055-9609>

<sup>d</sup> <https://orcid.org/0000-0002-3009-3311>

<sup>e</sup> <https://orcid.org/0000-0003-2631-0448>

<sup>f</sup> <https://orcid.org/0000-0001-7681-4287>

painting, there are some limitations of state of the art as illustrated in Figure 1. These are cases, where Figure 1(a) shows poor performance on holes with arbitrary sizes; Figure 1(b) illustrates the lack of edge-preserving using the existing technique; Figure 1(c) depicts the blurry artefacts; and Figure 1(d) demonstrates the poor performance on high-resolution images and image completion with mask at border region.

To correctly predict missing parts of a face and preserve its realism, we propose S-WGAN with the following contributions:

- We propose a new framework with Wasserstein Generative Adversarial Network (WGAN) that uses symmetric skip connection to preserve image details.
- We define a new combined loss function based on RGB and feature space.
- We demonstrate that our loss, combined with our S-WGAN, can achieve better results than the state-of-the-art algorithms.

## 2 PREVIOUS WORK

Pathak et al. (Pathak et al., 2016) proposed to use GANs (Goodfellow et al., 2014) with a context-encoder similar to (Vincent et al., 2010; Le et al., 2011) and AlexNet (Krizhevsky et al., 2012) for image inpainting despite poor hallucinations. Results show more artefacts and blur with randomised hole-to-image mask regions. Iizuka et al. (Iizuka et al., 2017) used a local and global discriminator to assess coherency and consistency of predicted pixels, and replaced the fully-connected layer of the generator with dilated convolutions (Yu and Koltun, 2015). Iizuka et al. (Iizuka et al., 2017) method failed to capture long-ranged textured information, however they used Poisson Blending by Perez et al. (Pérez et al., 2003) to process the output image. Yang et al. (Yang et al., 2017) proposed a multi-scale neural patch synthesis based on style transfer (Johnson et al., 2016; Ulyanov et al., 2016; Li and Wand, 2016), but failed to guarantee content and texture of high-resolution images with difficulty on irregular mask inpainting task. Yeh et al. (Yeh et al., 2017) introduced a spatial attention mechanism in deep convolutional GAN (Radford et al., 2015) combined with context loss but this algorithm suffers misalignment on closest encoding in latent space. It performs poorly in handling of high-resolution and complex scene images. Li et al. (Li et al., 2017) used face parsing network combined with a generator (encoder-decoder) and two discriminators

optimised by a semantic parsing loss to ensure local-global consistency and pixel fidelity. However, despite excellent performance, neighbouring pixels fail to establish spatial connections leading to colour inconsistencies. Li et al. (Li et al., 2018) introduced reflection symmetry into face completion and used two networks, to establish a correspondence between missing pixels on two half-faces optimised using a symmetry loss defined on VGGFace (Parkhi et al., 2015). However, this network fails to preserve structural information and is computationally costly.

Liu et al. (Liu et al., 2018a) used partial convolution to replace typical convolutions (Ulyanov et al., 2018) with an automatic mask-updating step. This technique masks and renormalise convolutions to target only valid pixels. However, it performs poorly on sparsely structured images and binary masks with huge holes and no quantitative evaluation report on facial images. Yan et al. (Yan et al., 2018) used deep feature rearrangement by adding a particular shift-connection layer to the U-Net architecture (Ronneberger et al., 2015), but lacks efficiency with no guarantee in computational speed. Yu et al. (Yu et al., 2018) proposed a dual-stage network convolutional network combined with a contextual attention layer that learns the location of feature information from background patches to generate missing content. However, this network lacks pixel-wise consistency on high-resolution images. Liu et al. (Liu and Jung, 2019) proposed a multi-scale feature extraction powered by a multi-level generative network optimised by content and texture losses based on Mean Square Error ( $\ell_2$ ) and Structure Similarity Index (MS-SSIM), to capture features at various levels. This model struggles with larger masks and fails to preserve structure in unaligned facial images. Li et al. (Li et al., 2019) proposed a nested GAN for facial inpainting, that uses a residual connection structure to transport information and interpolate feature map in deeper layer and shallow layer. Wang et al. (Wang et al., 2019) introduced a Laplacian approach based on residual learning (He et al., 2016) to propagate high-frequency details and predict missing information. Despite the significant contributions by the methods above to the field of inpainting, the absence of preserved realism on facial images from a compact latent feature is still challenging due to larger and irregular masks.

## 3 PROPOSED FRAMEWORK

Our proposed model uses skip connections with dilated convolution across the network, to perform image inpainting. We discuss the architecture and loss

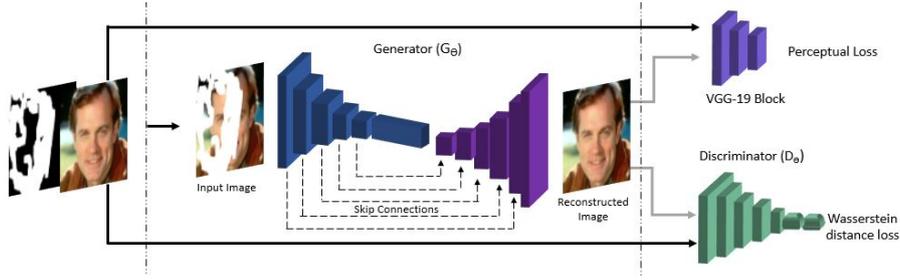


Figure 2: S-WGAN framework. The dilated convolution and deconvolution with the element-wise sum of feature maps (skip connection) combined with a Wasserstein network. The skip connections in the diagram ensure local pixel-level accuracy of the feature details to be retained.

function of S-WGAN in the following sections

### 3.1 Architecture

Figure 2 shows the overall framework of our proposed S-WGAN. The network is designed to have a generator ( $G_\theta$ ) and a discriminator ( $D_\theta$ ). We define  $G_\theta$  as an encoder-decoder framework with dilated convolutions and symmetric skip connections. Figure 3 shows the process of dilated convolution. Dilated convolutions (Yu and Koltun, 2015), combined with skip connections, are critical to the design of our model as:

- It broadens the receptive fields to capture more contextual information without parameter accretion and computational complexity, which are preserved and transferred by skip connections to corresponding deconvolution layers.
- It detects fine details and maintains high-resolution feature maps, and achieves end-to-end feature learning with a better local minimum (high restoration performance).
- It has shown considerable improvement of accuracy in segmentation task (Yu and Koltun, 2015; Chen et al., 2017a; Chen et al., 2017b).

**Generator ( $G_\theta$ ).** The effectiveness of feature encoding is improved by having an encoder of ten-convolutional layers, with a kernel size of 5 and dilation rate of 2, designed to match the size of the output image. This technique enables our model to learn larger spatial filters and help reduce volume (Rosebrock, 2019). Each block of convolution in exception of the final layer has Leaky ReLU activation and max-pooling operation of pool size  $2 \times 2$ . We apply a dropout regularisation with a probability value of 0.25 in the 4th and final layer of the encoder. The dropout layer randomly disconnects nodes and adjust the weights to propagate information to the decoder without overfitting.

**Decoder.** The decoder are five blocks of deconvolutional layers, with learnable upsampling layers that

recover image details using the same kernel size and dilation rate of the generator. The corresponding feature maps in the decoder are asymmetrically linked by element-wise skip connections to reach an optimum size. The final layer in the decoder is Tanh activation.

**Dilated Convolutions.** We express the dilated convolution based on the network input in Equation 1:

$$I'_{(m,n)} = \sum_{i=1}^M \sum_{j=1}^N (\mathbf{M}_I)(m+d_r \times i, n+d_r \times j) * \omega_{(i,j)} \quad (1)$$

where  $I'_{(m,n)}$  is the output feature map of the dilated convolution from the input  $\mathbf{M}_I = (I \odot (1 - M)) + M$  and the filter is given by  $\omega_{(i,j)}$ . The dilation rate parameter ( $d_r$ ) reverts to normal when  $d_r = 1$ .

It is advantageous to use dilated convolution compared to using typical convolutional layers combined with pooling. The reason for this is that a small kernel size of  $k \times k$  can enlarge into  $k + (k - 1)(d_r - 1)$  based on the dilated stride  $d_r$ , thus allowing a flexible receptive field of fine-detail contextual information while maintaining high-quality resolution.

The inpainting solver  $G_\theta$  may result in predictions  $G_\theta(\hat{\mathbf{z}})$  of the missing region, that may be reasonable or ill-posed. We include as part of our network  $D_\theta$ , adopted from (Arjovsky et al., 2017) to provide improved stability and enhanced discrimination for photo-realistic images. With ongoing adversarial training, the discriminator is unable to distinguish real data from fake ones. Equation 2 shows the reconstruction of the image during training from  $G_\theta$ :

$$G_\theta(I_R) = I \odot M + (1 - M) \odot G_\theta(\hat{\mathbf{z}}) \quad (2)$$

where  $I_R$  is the reconstructed image,  $I$  is the ground-truth,  $(\hat{\mathbf{z}})$  is the predictions,  $\odot$  is the element-wise multiplication and  $M$  is the binary mask, represented in 0 and 1. In our case 0 is the context of the entire image and 1 is the missing regions.

Equation 3 adopted from (Arjovsky et al., 2017) refers to the Wasserstein discriminator.

$$\max_D V_{WGAN} = E_{x \sim p_r} [D_\theta(I)] - E_{z \sim p_z} [D(G_\theta(I_R))] \quad (3)$$

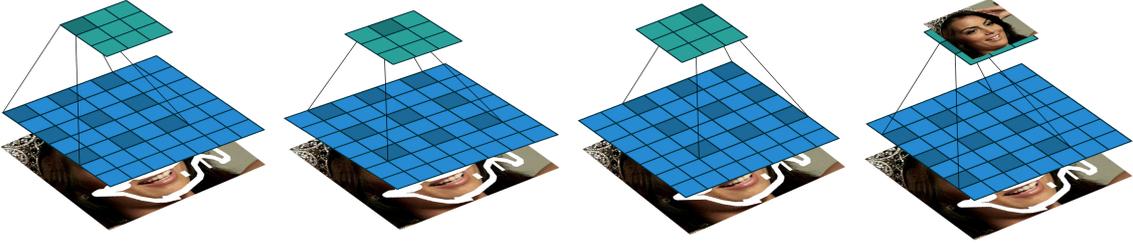


Figure 3: Illustration of dilated convolution process. Convoluting a  $3 \times 3$  kernel over a  $7 \times 7$  input with a dilation factor of 2 (i.e.,  $i = 7$ ,  $k = 3$ ,  $d_r = 2$ ,  $s = 1$  and  $p = 0$ ) (Dumoulin and Visin, 2016). The accretion of receptive field is in linearity with the parameters (Yu and Koltun, 2015). A  $5 \times 5$  kernel will have the same receptive field view as over a  $7 \times 7$  input at dilation rate=2 whilst only using 9 parameters over a  $512 \times 512$  input.

where  $D$  is the discriminator and  $P_r$  is real data distribution.  $G$  is the generator of our network and  $P_z$  is the distribution.

### 3.2 Loss Function

**Perceptual Loss.** Instead of using the typical  $\ell_2$ -loss function used in (Pathak et al., 2016), we use a new combination of loss functions, luminance ( $L_l$ ) and feature loss. Pixel-wise reconstruction and feature space loss are not new to inpainting (Yeh et al., 2017; Yu et al., 2018; Johnson et al., 2016). We define a luminance guided  $L_l$  that uses  $\ell_1$ -loss as a base to compute the loss using a range of constant pixel values in the RGB space. This preserves colour and luminance and does not over penalise large errors (Zhao et al., 2016). We use the  $L_l$  to adjust our perceptual loss, thus minimising any error  $>1$ . Also, the  $L_l$  allows better evaluation of the predictions to match the ground-truth. More specifically, we express the luminance loss ( $L_l$ ) based on  $\ell_1$  as:

$$L_l = \|K \odot (x_i - \hat{z}_i)\|_1 \quad (4)$$

where  $i$  is the pixel index with  $x_i$  and  $\hat{z}_i$  as pixel values of the ground-truth and the predictions, constraint by a constant  $K$ . Our feature loss  $L_f$  is a feature based  $\ell_2$ -loss, rather than being computed directly on the image we computed the loss in a feature space. To achieve this, we adopt a pre-trained VGG-16 model trained on ImageNet (Krizhevsky et al., 2012), and use it as a feature extractor in our loss function. More specifically we use the output of block3-convolution3 of this model to generate image feature. We use the  $\ell_2$  as base to compute our loss function, which is the same as the perceptual loss proposed by Johnson et al. (Johnson et al., 2016). The advantage of using feature space is that a particular filter determines the extraction of feature maps, from low-level to high-level sophisticated features. To reconstruct quality images, we compute our loss function with feature maps determined by block3-conv3, resized to the same size as masks and generated images. The reason is that using

another output for example block4-conv4 or block5-conv5 will result in poor quality, as the network starts to expand the view at these layers due to more filters used. Our feature loss is expressed as follows:

$$L_f = \frac{1}{N} \sum_{i \in \Phi} (\phi_i[M_{I_i}] - \phi_i[I_{R_i}])^2 \quad (5)$$

where  $M_I$  is the input image,  $I_R$  is the reconstructed image and  $N$  is dimensions obtained from  $\Phi$  feature maps with high-level representational abstractions extracted from the third block convolution layer. By combining  $L_l$  and  $L_f$  we obtained:

$$L_p = L_l + L_f \quad (6)$$

By using  $L_p$  the model learns to produce finer details in the predicted features and output without any blurry artefacts. We add the Wasserstein loss ( $L_w$ ) improves convergence in GANs and its the mean difference between two images. Finally the entire model trains end-to-end with back-propagation and uses the global Wasserstein-perceptual loss function ( $L_{wp}$ ) defined in Equation 7, to optimise  $G_\theta$  and  $D_\theta$  to learn reasonable predictions. Our goal is to reconstruct an image  $I_R$  from  $M_I$  by training the generator  $G_\theta$  to learn and preserve image details.

$$L_{wp} = L_w + L_p \quad (7)$$

## 4 EXPERIMENT

This section describes the dataset, binary masks and the implementation.

### 4.1 Dataset and Irregular Binary Mask

Our experiment focuses on high-resolution face images and irregular binary masks. The benchmark dataset for high-resolution face images is CelebA-HQ dataset (Karras et al., 2017), which was curated from the CelebA dataset (Liu et al., 2018b) and contained



Figure 4: Sample images from CelebA-HQ Dataset (Karras et al., 2017).



Figure 5: Process of input generation: a) CelebA-HQ image; b) Binary mask image (Iskakov, 2018); and c) Corresponding masked image (input image).

30,000 images. Figure 4 shows a few samples from the CelebA-HQ dataset.

To create irregular holes on images, we use the Quickdraw irregular mask dataset (Iskakov, 2018), available for public use and is divided into 50,000 train and 10,000 test masks. The images are of size  $512 \times 512$  pixels.

## 4.2 Implementation

We used the Keras library with TensorFlow backend to implement and design our network. With our choice of the dataset, we followed the experiment settings of state of the art (Liu et al., 2018a) and split our data into 27,000 images for training and 3,000 images for testing.

We perform normalised floating-point representation on the image to set the intensity values of the pixels in the range  $-1, 1$  and apply the mask on the image to obtain our input, as shown in Figure 5. We initialize pre-trained weights from VGG-16 to compute our loss function. We use a learning rate of  $10^{-4}$  in  $G_\theta$  and  $10^{-12}$  in  $D_\theta$  and optimise the training process using the Adam optimiser (Kingma and Ba, 2014). We use a Quadro P6000 GPU machine to train these models. According to our hardware conditions, we use a batch-size of 5 in each epoch for input images with shape  $512 \times 512 \times 3$ . It takes 0.193 seconds to predict missing pixels of any size created by binary mask on an image and ten days to train 100 epochs.

## 5 RESULTS

We assess the performance of the inpainting methods qualitatively and quantitatively in this section.

### 5.1 Qualitative Comparisons

Consider the importance of visual and semantic coherence; we conducted a qualitative comparison of our test dataset. First, we implemented a WGAN approach with  $L_f$  and  $L_w$ . We observed an induced pattern and pitiable colour on the images, as shown in Figure 6(d). We introduced dilated convolution, skip connections combined with end-to-end training using  $L_{wp}$  to handle the induced pattern and match the luminance of the original images.

We compare our model with three popular methods:

- **CE:** Context-Encoder method by Pathak et al. (Pathak et al., 2016).
- **PConv:** Image Inpainting for irregular holes using partial convolutions by Liu et al. (Liu et al., 2018a).
- **WGAN:** Wasserstein GAN method with perceptual loss.

We test our S-WGAN against state of the art on CelebA-HQ  $512 \times 512$  test dataset and show the results in Figure 6. Based on visual inspection, Figure 6(b) illustrates blurry generated by the Pathak et al.'s CE method (Pathak et al., 2016). On the other hand, PConv (Liu et al., 2018a) generates clear images but with residues of the binary mask left on the images as shown in Figure 6(c). WGAN induced pattern and low-contrast images, shown in Figure 6(d). Overall, our proposed S-WGAN, as shown in Figure 6(e), produced the best visual results when compared to the ground-truth in Figure 6(f).

### 5.2 Quantitative Comparisons

We select some popular image quality metrics including  $\ell_1$ ,  $\ell_2$ , Peak Signal to Noise Ratio (PSNR), SSIM to evaluate the performance quantitatively. Table 1 shows the results from our experiment compared to state of the art (Pathak et al., 2016; Liu et al., 2018a) for image inpainting with our S-WGAN in bold.

For  $\ell_2$  and  $\ell_1$ , the lower the value, the better the image quality.  $\ell_2$  measures the average squared intensity difference of pixels while  $\ell_1$  measures the magnitude of error between the ground-truth image and the reconstructed image. Conversely, for PSNR and SSIM, the higher the value, the closer the image quality to the ground-truth. Based on observation from Table 1, S-WGAN achieves lower  $\ell_1$ ,  $\ell_2$ ,

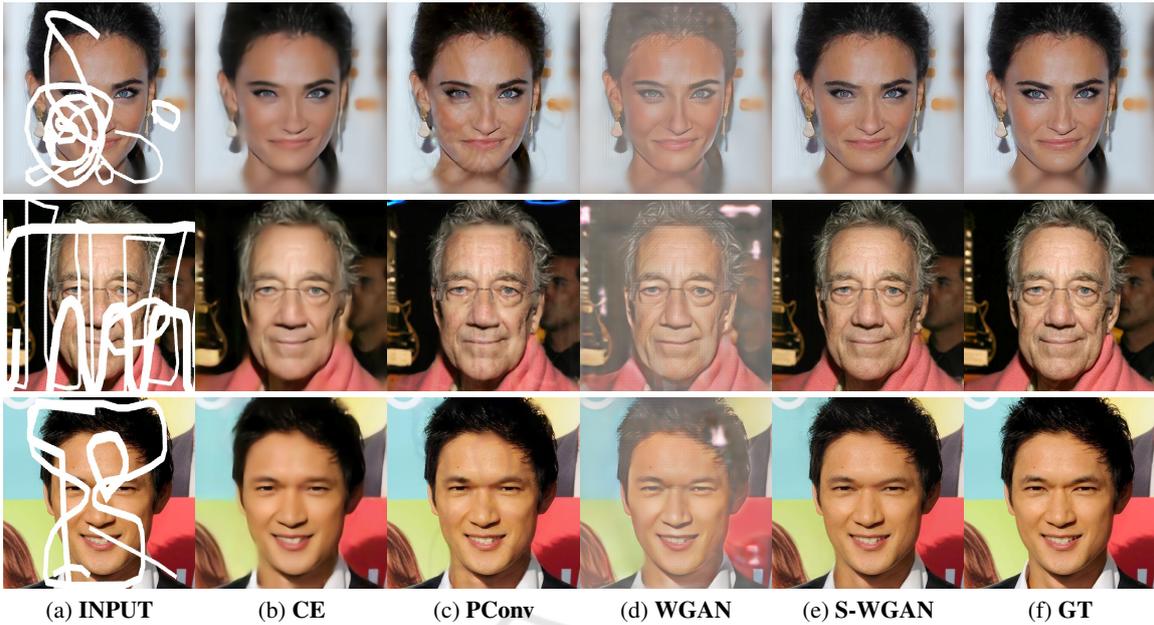


Figure 6: Qualitative comparison of our proposed **S-WGAN** with the state-of-the-art methods on CelebA-HQ: (a) **Input masked-image**; (b) **CE** (Pathak et al., 2016); (c) **PConv** (Liu et al., 2018a); (d) **WGAN**; (e) **S-WGAN** (proposed method); and (f) Ground-truth image.

Table 1: Quantitative comparison of various performance assessment metrics on 3,000 test images from the CelebA-HQ dataset. † Lower is better. ‡ Higher is better.

Method	$\ell_2$ †	$\ell_1$ †	PSNR ‡	SSIM ‡
<b>WGAN</b>	3562.13	87.03	13.50	0.56
<b>CE</b>	133.481	129.30	27.71	0.76
<b>PConv</b>	124.62	105.94	28.82	0.90
<b>S-WGAN</b>	<b>81.03</b>	<b>66.09</b>	<b>29.87</b>	<b>0.94</b>

higher PSNR and higher SSIM values in comparison with **CE** (Pathak et al., 2016) and **PConv** (Liu et al., 2018a), which suggests that S-WGAN provide more accurate predictions than the state-of-the-art inpainting algorithm.

## 6 ABLATION STUDY

To justify the S-WGAN framework and validate the effectiveness of  $L_p$ , we conduct experiments and show intermediate results using different alterations of the S-WGAN on CelebA-HQ dataset. Firstly, we conduct investigations on the WGAN and WGAN with skip connection (WGAN-S) using the  $L_f$ , and observed a slight improvement in texture and structure of the reconstructed masked regions of the images. Figure 7 (b) and (c) show changes influenced by skip connections. We observed that visually and quantitatively, the WGAN-S performs better than WGAN model but not satisfactory as shown in

the first part of Table 2. Secondly, we improve the WGAN-S model by including dilated convolutions to each block, and additional convolution layers to obtain our WGANSD model. We train the WGANSD with the  $L_f$  and train the S-WGAN model with our new combined loss function. We noticed that training with the  $L_f$  improved our results slightly, but not satisfactorily. To verify the differences of these models, we conduct a qualitative and quantitative evaluation. Visually, within the yellow rectangle on Figure 7 comparing columns (d) and (e), the S-WGAN result in column (e) improved with significantly enhanced local detail when compared with column (d) and the original on column (f). Also, in quantitative evaluation shown in Table 2, we observe that S-WGAN trained end-to-end with  $L_{wp}$  predicts reasonable outputs with finer details. We also show more qualitative results in Figure 8 to demonstrate the S-WGAN produces images with preserved realism.

To validate our S-WGANs' representational ability generalised to other masks e.g. Nvidia mask (Liu et al., 2018a), we use the various architectures of our model to conduct experiments during the ablation studies. We apply the Nvidia mask as the masking method and show our results in Figure 9.

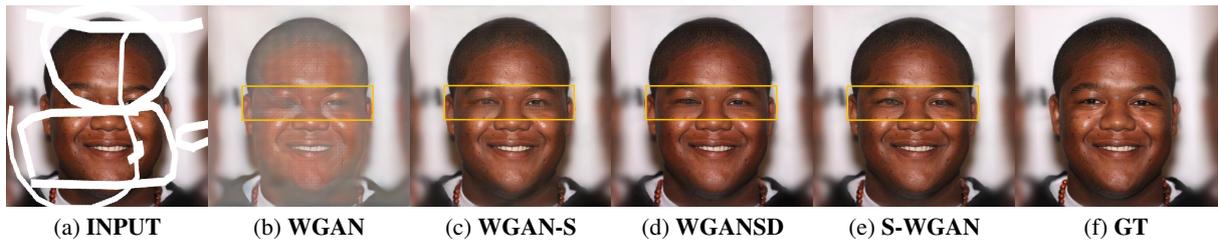


Figure 7: Qualitative comparison of results using different architectures (Johnson et al., 2016) on CelebA-HQ (Karras et al., 2017). (a) Input masked image (b) Inpainted image by WGAN (c) Improved WGAN with skip connections (WGAN-S) (d) Improved WGAN with skip connection and dilated convolution (WGANSD) (e) Complete network with  $L_p$  (f) Ground-Truth image. The yellow box indicates the region where other models failed to inpaint successfully completely. This region in (e) shows the effectiveness of  $L_p$  on the inpainted image.

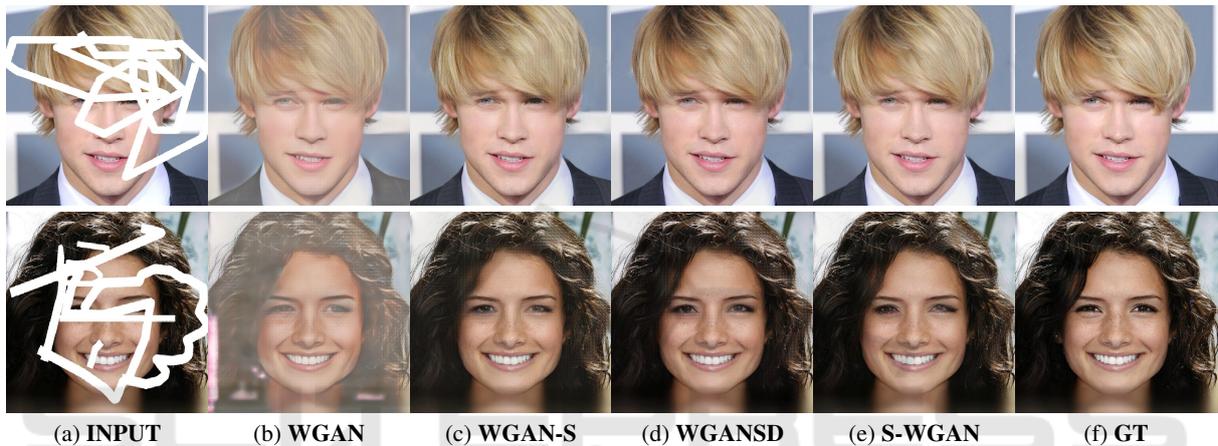


Figure 8: Qualitative comparison of results using different architectures with the perceptual loss (Johnson et al., 2016) on CelebA-HQ (Karras et al., 2017). (a) Input masked image; (b) inpainted image by WGAN; (c) Improved WGAN with skip connection; (d) improved WGAN with skip connection and dilated convolution (e) Complete network with  $L_p$ ; (f) The ground-Truth image.

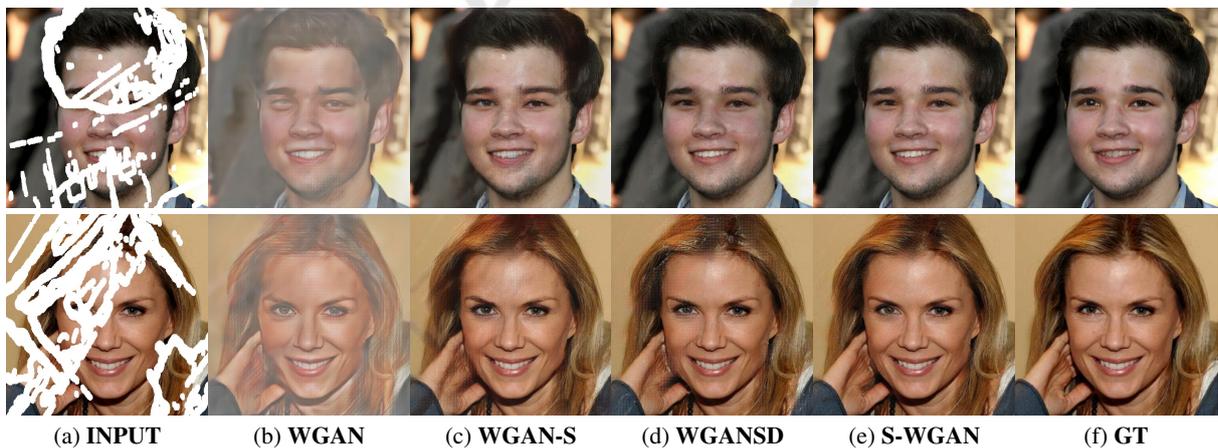


Figure 9: Qualitative evaluation of different architectures with perceptual loss (Johnson et al., 2016) on CelebA-HQ (Karras et al., 2017) and Nvidia Mask. (a) Input masked image; (b) Inpainted image by WGAN; (c) Improved WGAN with skip connection; (d) Improved WGAN with skip connection and dilated convolution; (e) Complete network with  $L_p$ ; (f) The ground-Truth image.

Table 2: Quantitative difference of results based on different architectures (WGAN), WGAN-S, WGANSD with  $L_f$ , and S-WGAN trained with  $L_p$ . † Lower is better. ‡ Higher is better.

Method	$\ell_2$ †	$\ell_1$ †	PSNR ‡	SSIM ‡
WGAN	3562.13	87.03	13.50	0.56
WGAN-S	151.4	69.59	27.01	0.87
WGANSD	145.82	65.15	29.26	0.92
S-WGAN	<b>81.03</b>	<b>66.09</b>	<b>29.87</b>	<b>0.94</b>

## 7 DISCUSSION

Our proposed S-WGAN with dilated convolution and skip connections trained end-to-end with Wasserstein-perceptual loss function outperforms the state-of-the-art. Our model can learn the end-to-end mapping of input images from a large-scale dataset to predict missing pixels of the binary mask regions on the image. Our S-WGAN automatically learns and identifies missing pixels from the input and encodes them as feature representations, to be reconstructed in the decoder. Skip connections help to transfer image details forwardly and find local minimum by backward propagation.

Our experiments show the benefit of skip connection combined with Wasserstein-perceptual loss for image inpainting. We have visually compared our proposed method with state of the art (Pathak et al., 2016; Liu et al., 2018a) in Figure 6. To verify the effectiveness of our network, we carried out experiments with regular convolutions and used the  $L_f$ . We noticed that the generated images had checkboard artefacts with pitiable visual similarity compared to the original image, as shown in Figure 6(d). We introduced skip connections with dilated convolution and our new loss function, and obtained improved results that are semantically reasonable with preserved realism in all aspects.

Compared to existing methods, the generator of our S-WGAN learns specific structures in natural images by minimising  $L_p$  with an enhanced hallucinating ability powered by symmetric skip connections. Based on Figure 6, our S-WGAN can handle irregularly shaped binary mask without any blurry artefacts and has shown edge-preserving and mask completion at border regions on the output images. Additionally, using the Wasserstein discriminator enables the overall network to perform better. This boost the experimental performance of our network to achieve state-of-the-art results in inpainting task on high-resolution images.

One limitation is a consistent practice of other inpainting methods in the preprocessing step. Most

preprocessing ignores the fact that the image has to be converted into normalised floating points representations and an inverse-normalisation on the output image, which contributes to the colour discrepancies on the output image, that leads to expensive post-processing. We have been able to solve this using S-WGAN with a new combination of the loss function that preserves colour and image detail.

## 8 CONCLUSION AND FUTURE WORK

In this paper, we propose S-WGAN. Our network can generate images, which are semantically and visually plausible with preserved realism of facial features. We achieved this with a network structure that can widen the receptive field in each block to capture more information and forward to the corresponding deconvolutional blocks. Additionally, we introduced a new combined loss function based on luminance and feature space combined with Wasserstein loss. Our network was able to generate high-resolution images from input covered with arbitrary binary mask shape and achieve a better performance compared to the state-of-the-art methods. The proposed network has shown the effectiveness of skip connections with dilated convolutions as a capture and refining mechanism of contextual information combined with WGAN. For future work, we aim to extend our model to inpaint coarse and fine wrinkles extracted from wrinkle detectors (Yap et al., 2018) with preserved realism.

## ACKNOWLEDGEMENTS

The authors would like to thank The Royal Society (Grant number: IF160006 and INF/PHD/180007). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P6000 and SCAN UK for providing DGX A100 servers used for this research.

## REFERENCES

- Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.
- Barnes, C., Shechtman, E., Finkelstein, A., and Goldman, D. B. (2009). Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (ToG)*, 28(3):24.

- Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. (2000). Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co.
- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017a). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848.
- Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017b). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.
- Criminisi, A., Pérez, P., and Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212.
- Dumoulin, V. and Visin, F. (2016). A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*.
- Efros, A. A. and Leung, T. K. (1999). Texture synthesis by non-parametric sampling. In *iccv*, page 1033. IEEE.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Iizuka, S., Simo-Serra, E., and Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107.
- Iskakov, K. (2018). Semi-parametric image inpainting. *arXiv preprint arXiv:1807.02855*.
- Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer.
- Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.
- Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., and Ng, A. Y. (2011). On optimization methods for deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 265–272. Omnipress.
- Li, C. and Wand, M. (2016). Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2479–2486.
- Li, X., Liu, M., Zhu, J., Zuo, W., Wang, M., Hu, G., and Zhang, L. (2018). Learning symmetry consistent deep cnns for face completion. *arXiv preprint arXiv:1812.07741*.
- Li, Y., Liu, S., Yang, J., and Yang, M.-H. (2017). Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3911–3919.
- Li, Z., Zhu, H., Cao, L., Jiao, L., Zhong, Y., and Ma, A. (2019). Face inpainting via nested generative adversarial networks. *IEEE Access*, 7:155462–155471.
- Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., and Catanzaro, B. (2018a). Image inpainting for irregular holes using partial convolutions. *arXiv preprint arXiv:1804.07723*.
- Liu, J. and Jung, C. (2019). Facial image inpainting using multi-level generative network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1168–1173. IEEE.
- Liu, Z., Luo, P., Wang, X., and Tang, X. (2018b). Large-scale celebfaces attributes (celeba) dataset. Retrieved August, 15:2018.
- Park, T., Zhu, J.-Y., Wang, O., Lu, J., Shechtman, E., Efros, A., and Zhang, R. (2020). Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33.
- Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition.
- Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544.
- Pérez, P., Gangnet, M., and Blake, A. (2003). Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318.
- Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.
- Rosebrock, A. (2019). *Deep Learning for Computer Vision with Python*. PyImageSearch.com, 2.1.0 edition.
- Sun, J., Yuan, L., Jia, J., and Shum, H.-Y. (2005). Image completion with structure propagation. In *ACM Transactions on Graphics (ToG)*, volume 24, pages 861–868. ACM.
- Ulyanov, D., Lebedev, V., Vedaldi, A., and Lempitsky, V. S. (2016). Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, pages 1349–1357.
- Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2018). Deep image prior. In *Proceedings of the IEEE Conference*

- on *Computer Vision and Pattern Recognition*, pages 9446–9454.
- Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408.
- Wang, Q., Fan, H., Sun, G., Cong, Y., and Tang, Y. (2019). Laplacian pyramid adversarial network for face completion. *Pattern Recognition*, 88:493–505.
- Yan, Z., Li, X., Li, M., Zuo, W., and Shan, S. (2018). Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–17.
- Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., and Li, H. (2017). High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3.
- Yap, M. H., Alarifi, J., Ng, C.-C., Batool, N., and Walker, K. (2018). Automated facial wrinkles annotator. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0.
- Yeh, R. A., Chen, C., Lim, T.-Y., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. (2017). Semantic image inpainting with deep generative models. In *CVPR*, volume 2, page 4.
- Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.
- Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention. *arXiv preprint*.
- Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57.