# An End-to-end Facial Image Inpainting with Skip Connection

Liyuan Jiang
l.jiang1@student.tue.nl

Yiling Zhang
y.zhang2@student.tue.nl

## Abstract

*Although image inpainting has been around for many years, the development of deep learning provides more inspiration for image processing techniques. This research focuses on facial inpainting due to its small scale and application value. This paper briefly reviews the existing image inpainting approaches and their evaluation methods, then designs two experiments to research the effect of network structure on the CNN-based inpainting performance. Through the comparison under the same evaluation metric, we conclude the effects of different network structures and propose our final design of the CNN-based facial inpainting.*

## 1. Introduction

Image inpainting is the process of completing or recovering the missing region in an image or removing some object added to it. The input for the process should be an image with a rectangular or irregular mask, and the expected output should be the corresponding image in which the occluded part is restored, as shown in Fig 1.

Facial inpainting has a wide application in many fields, such as film and television special effects production, virtual and digital cultural heritage protections, and even criminal investigation, and such images on a small scale require less memory and time for training and test, so we would focus on the inpainting of facial images with some missing part in the research.

Compared with an irregular mask, facial inpainting for rectangular masks should be more difficult because all facial features are missing. **Therefore, we aim to achieve end-to-end CNN-based image inpainting for facial images with a rectangular mask and research on the effect of network structure on the inpainting results.**

## 2. Related work

Image inpainting research mainly includes tasks such as denoising, removing watermarks, removing text, removing scratches, repairing masks, target removal, and coloring

Figure 1. Facial inpainting

of old photos. Traditional image inpainting can be generally divided into patch-based and diffusion-based methods. Patch-based methods are based on techniques for patch-by-patch filling in the missing region by searching for well-matching replacement patches (i.e., candidate patches) in the undamaged part of the image and copying them to corresponding locations, such as the saliency map and gray entropy proposed by Zeng *et al*. [3]. Diffusion-based methods fill in the missing region (the "hole") by smoothly propagating image content from the boundary to the interior of the missing region, such as the Fractional-order derivative and Fourier transform based method proposed by Zeng *et al*. [1]. Although these methods do not rely on a large dataset for inpainting tasks, they perform poorly for context restoration in an image with a large missing part.

Recently, the strong potential of deep convolutional networks (CNNs) is being exhibited in all computer vision tasks, especially in image inpainting. The first GAN-based image completion network is Context Encoders (CE) proposed by Deepak *et al*. [4], where an encoder-decoder pipeline predicts missing parts of scene from their surroundings and an adversarial network performs self-supervision. Although it provides an important basis for the CNN-based image inpainting, it produces a semantically plausible but blurry result, and relies on the location and size of the missing part. Yan *et al*. [7] proposed a ShiftNet structure based on the UNet [6] and Context Encoder to fill in missing regions of any shape with sharp structures and fine-detailed textures. The additional shift connection in this structure offers a reference for removing the dependency on the size and location of masks in our research.

The evaluation metric is significant for image inpaint-

ing because the reconstructed results should be similar to the original images in the aspect of perception while the usual metrics may not be suitable. As shown in Fig 2, the better match in human perception cannot be recognized through the general pixel-wise metrics such as PSNR and SSIM (structural similarity). Zhang *et al*. [5] proposed a learned perceptual image patch similarity (LPIPS) metric, which evaluates the L2 distance between the features, resulting in a good simulation of human perception, so our research adopts LPIPS as the uniform metric.

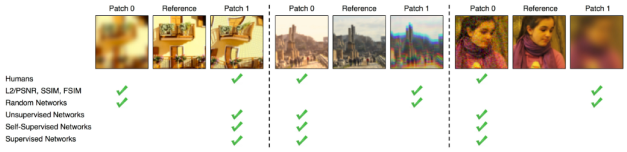



Figure 2. The comparison of different metrics [5]

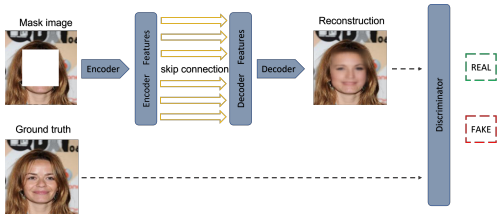



Figure 3. The final architecture

## 3. Method

Fig 3 illustrates our overall architecture for image inpainting, where the encoder-decoder pipeline performs image inpainting under the supervision of the adversarial discriminator and skip connection is adopted to supplement the existing information of masked images in order to achieve the end-to-end reconstruction.

### 3.1. Image inpainting network design

We design a network with the combination of an encoder-decoder pipeline and the skip connection as shown in Fig 4. The inpainting network includes an encoder which learns the features of an input image with a mask and produces a latent feature representation of that image by downsampling; and a decoder which rebuilds the missing image based on the features provided by the encoder using upsampling. On the basis of the encoder-decoder pipeline, there is a skip connection in each corresponding layer between encoder and decoder to pass the supplement features. The input images are RGB images ($128 \times 128$) with a $64 \times 64$ rectangular mask, and the corresponding output is the whole reconstructed image ($128 \times 128$).

**Skip connection is an important design for removing the dependence on the location and size of applied masks during the inpainting.** It copies the feature maps in the encoder and concatenates them with the feature maps in each decoder level so as to pass the detailed level information. In detail, we can get a feature map with the size of $w \times h \times c$ after the previous upsampling in a decoder layer, then the reconstructed feature map should be $w \times h \times 2c$ after the concatenating. By applying a convolution function on the reconstructed feature map, we can resize it into $w \times h \times c$ for the upsampling for the next layer.

### 3.2. Adversarial training

In addition to the image inpainting network, a discriminator is added to supervise the inpainting task. The output map of the discriminator should be 1 if the input image is a real image, and it should be 0 if the input image is a reconstructed image. The training for the discriminator aims to enable it to distinguish reconstructed images from real images, while the training for the image inpainting network aims to enable the network to generate images which can be recognized as real by the discriminator. So-called adversarial training [2] improves the global consistency of image

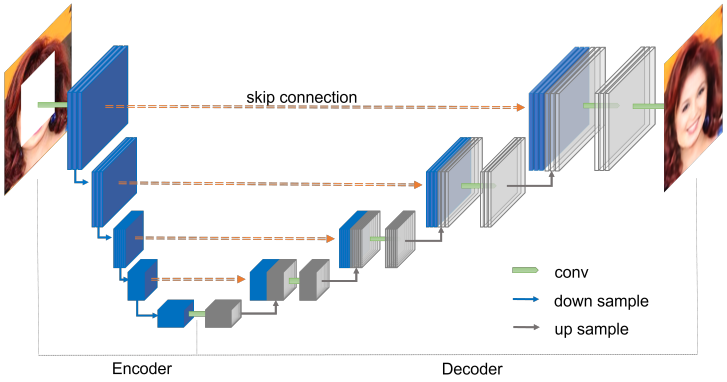



Figure 4. Image inpainting network

inpainting, including the below tasks.

- Training the image inpainting network
  We use a joint loss [4] to train the network, including a reconstruction loss which narrows the pixel-wise difference between the reconstructed results and the original images, and an adversarial loss which is used for evaluating the distance between a real output confirmed by the current discriminator and the inpainted result. The total loss for the training is

$$\mathcal{L}_1 = 0.001 \times \mathcal{L}_{adv} + 0.999 \times \mathcal{L}_{rec}$$

- Training the discriminator
  The adversarial loss makes an average of the real loss (i.e. the difference between the network output and the real output map) and the fake loss (i.e.the difference between the network output and the fake output map), which is used for training the discriminator, as shown in the below equation:

$$\mathcal{L}_2 = 0.5 \times \mathcal{L}_{real\_adv} + 0.5 \times \mathcal{L}_{fake\_adv}$$

# 4. Experiments

## 4.1. Dataset

We select the CelebFaces Attributes Dataset (CelebA) as the dataset for this task because it contains enough facial images for training (202,599) and the images are on a small scale with a size of 178×218. Given the images, we apply rectangular masks on them, which is supposed to be more challenging for image inpainting than irregular occlusion for facial images on a small scale as it provides less information about facial features.

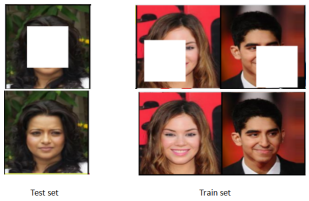



Figure 5. Dataset with masks

- Training set: images with a rectangular mask ($64 \times 64$) on a random region
- Test set: images with a rectangular mask ($64 \times 64$) on the center region

## 4.2. Metric

Learned Perceptual Image Patch Similarity (LPIPS) is selected for evaluation. Compared with other pixel-wise methods, LPIPS calculates the L2 distance between features of two images, which is closer to human perception. It could be good for facial image recovery, because we have a basic expectation for the blank region, such as the mouth or nose. A lower LPIPS means a higher similarity in perception between the original image and the restored image.

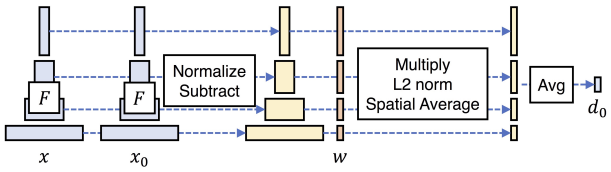



Figure 6. LPIPS map [5]

The algorithm is shown in Fig 6. To compute a distance between two images $x$ and $x_0$, deep embeddings are computed based on a given network, and the activations are normalised in the channel dimension. Each channel is scaled by vector $w$, and then we take the $\updownarrow_2$ distance and average across the spatial domain and across all layers.

LPIPS is the uniform metric for all experiments in this research. We use AlexNet as the network for distance computing because AlexNet is one of the models whose ability for learning feature distance has been confirmed by Zhang *et al.* [5].

## 4.3. Experimental settings

(i) Activation functions
The encoder uses ReLu as the activation function due to its easy computation, and the decoder and the discriminator use leaky ReLU to avoid the dying ReLU problem. Additionally, the last layer in the decoder uses Tahn because Tahn has a larger gradient and is 0-averaged.

(ii) Normalisation
In the image inpainting network, batch normalisation is adopted in each layer to stabilize the distribution during the propagation. In the discriminator network, instanceNorm is used because it is more suitable for a single image.

(iii) Loss functions
The reconstruction loss and the adversarial loss are L2 loss and L1 loss respectively.

(iv) Optimizer
Adam is used for both the image inpainting network and the discriminator since Adam generally performs well on most problems.

(v) Hyperparameters
The number of epoch is 40 for all experiments because the baseline network (i.e. Model 0 introduced in the

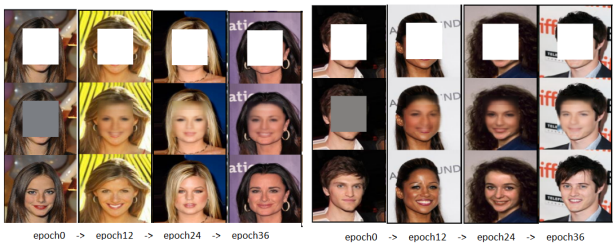
next Section) converges in the epoch 36, and an empirical value of learn rate is 0.0002.

## 5. Results and discussion

### 5.1. The effect of a channel-wise fully connection layer

| - | epoch0 | epoch12 | epoch24 | epoch36 |
|---|---|---|---|---|
| Model 0 | 9.96 | 3.21 | 2.99 | 2.91 |
| Model 1 | 9.98 | 3.32 | 2.98 | 2.83 |

Table 1. Experiment 1: LPIPS results




(a) Model 0 (b) Model 1

Figure 7. Experiment 1: The visualisation of inpainting results

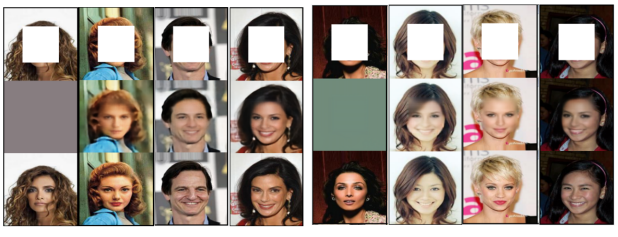
In the Context Encoder [4], a channel-wise fully connection layer is used to make all pixels in the previous map contribute to each location in the current layer. However, we were doubtful about the necessity of this FC layer. In order to decrease the model size and improve efficiency, we compared the performance with and without the layer. **Model 0** is a classical CE with a channel-wise FC layer, while we only remove the channel-wise fully connection layer to generate a new network, i.e.**model 1**, and the visualised results and the corresponding LPIPS are shown in Fig7 and Table1. We can see that there is no significant difference in both results, so it is not necessary to use a channel-wise fully connection layer. The possible reason is that an encoder-decoder pipeline does not need to make use of all high-level features for each pixel to rebuild a missing facial part. Based on the comparison, the channel-wise FC layer is removed from our final model to decrease the model size.

### 5.2. Generation of a whole image

If we change the reconstruction from filling in the missing part (i.e. Model 1) to generating a complete image directly (i.e. Model 2), the reconstruction would also blur the existing pixels in the input images as shown in Fig8(a), because the reconstruction is based on the decoded feature maps and will lose some detailed information somehow. Inspired by UNet [6], we researched the effect of skip-connection on this problem, as mentioned in Methods.

| - | epoch0 | epoch12 | epoch24 | epoch38 |
|---|---|---|---|---|
| Model 2 | 10.26 | 2.00 | 1.67 | 1.60 |
| Model 3 | 10.65 | 1.10 | 1.06 | 1.04 |

Table 2. Experiment 2: LPIPS results




(a) Model 2 (b) Model 3

Figure 8. Experiment 2: The visualisation of inpainting results

**Model 2** is an end-to-end generation without skip connection, while **model 3** contains the skip connection as shown in Fig4. Their results are shown in Fig8 and Table 2. With the comparison, we can see that the skip connection solves the blurring problem effectively. Even though in the earlier epoch, the existing part of the input is still clear in the output. Additionally, skip connection passes the detailed information in each level so that the LPIPS of results and the convergence speed are both improved. Therefore, skip connection is adopted in our final design.

## 6. Conclusion

We compared the results of different network structures and researched on the influence of a channel-wise FC layer and skip connection. Through the comparison, our final network removes the FC layer and uses skip connection, which decreases the model size and improves the performance in both visualisation and the LPIPS metric. At the current stage, the reconstruction of our model is relatively poor for front faces and Asian faces, which can be concentrated more on in future work.

## References

[1] S. S. Kumar G. Sridevi. Image inpainting based on fractional-order nonlinear diffusion for image reconstruction. *Circuits, Systems, and Signal Processing*, pages 1–16, 2019. 1

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[3] L. Leng C. Wang J. Zeng, X. Fu. Image inpainting algorithm based on saliency map and gray entropy. *Arabian Journal for Science and Engineering*, 44(4):3549–3558, 2019. 1

[4] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 1, 3, 4

[5] Alexei A. Efros Eli Shechtman Oliver Wang Richard Zhang, Phillip Isola. The unreasonable effectiveness of deep features as a perceptual metric. *Computer Vision and Pattern Recognition (cs.CV)*, 2018. 2, 3

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 4

[7] Mu Li Wangmeng Zuo Shiguang Shan Zhaoyi Yan, Xiaoming Li. Shift-net: Image inpainting via deep feature rearrangement. *Computer Vision and Pattern Recognition (cs.CV)*, 2018. 1