A Resnet and U-Net Hybrid Discriminator Based on GANs for Facial Images Inpainting

Yue Zhu
Data Science and Big Data Technology, Shanghai University of Engineering Science, Shanghai, China

Keywords: Reconstructing Image, Inpainting, GANs, Hybrid Discriminator.

Abstract: Image inpainting technique is an application of computer vision technology, whose main task is to complete reconstructing image with missing or damaged areas by using the information of the existing part, which becomes a crucial part in computer vision (CV). This research presents a method for optimizing discriminators in Generative Adversarial Networks (GANs) inpainting application. By introducing a hybrid discriminator comprised by Residual Network-50 (Resnet50) and U-Net network, it has better perception of reconstructed facial image textures and regional contours, thereby reducing the common blurring problem of reconstructed images. Through random blurring images pre-training for the proposed discriminator, only the fine-tuned process is required in practical applications, resulting in faster training time. Through comparative experiments, this research can achieve smaller training losses, while the subjective image effect after reconstruction is also significantly improved, specifically in areas such as the nose and eyes, where the edge resource is more pronounced. The discriminator optimization design proposed in this research, combined with pre- training methods, can be applied to GANs image inpainting application, significantly improving the reconstruction effect of complex texture areas in damaged images, especially facial areas.

1 INTRODUCTION

Image inpainting is an image processing technique which involves the selection of suitable pixels to fill the defect area by extracting and screening the features of the known area. At present, the development of inpainting algorithms have become a research hotspot of Computer Vision (CV) field, which can be widely applied in multimedia, medical, entertainment and film industry (Shah et al 2022). The basic idea of image inpainting techniques is inferring the damaged content from the content of undamaged areas in the same image. The traditional techniques use many kinds of graphic filters to reconstruct areas which can generally perform well for repairing small and regular damaged areas, but it is not effective for complex graphic textures (such as faces), large or irregular damaged areas. They can't extract higher-level semantic features and often lack enough context and semantic information.

After Deep learning techniques was introduced in 2014, researchers found it could be a good solution to the above problems, therefore it is now a widely used research technique in image inpainting (Shah et al 2022). In 2016, Pathak et al. introduced the concept of Context Encoders (CE) based on Goodfellow's research on Generative Adversarial Networks (GANs) (Pathak et al 2016 & Goodfellow et al 2014). CE can extract semantic features and the global structure. It has a good effect on reconstruction. After that, many improved versions of CE have emerged. For example, Lizuka et al. simultaneously used global and local discriminators to improve the coherence between reconstructed pixels and surrounding pixels (Iizuka et al 2017). Convolution Neural Networks (CNN), as the core algorithm of deep learning, bring a bright perspective to image processing related application such as millions of image classifications. The convolutional algorithm has better generalization for different tasks due to their better interpretability in image processing (Donahue et al 2014). If context related information is added in image generation, the effect of image generation will exceed that of autoencoder (Doersch et al 2015). Yeh et al. introduced two new loss functions with no need for masking during training. However, this method still cannot reconstruct irregular damaged areas (Yeh et al 2017).

The main purpose of this research is to introduce pre-trained context discriminator to improve the inpainting performance of facial images based on

Zhu, Y.
A ResNet and U-Net Hybrid Discriminator Based on GANs for Facial Images Inpainting.
DOI: 10.5220/0012819400003885
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 192-196
ISBN: 978-989-758-705-4
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

GANs image reconstruction algorithms. Specifically, first, a CNN structured convolutional network is used to mine image features, generate image feature vectors, and then use deconvolution to generate predicted image pixel information. Secondly, in adversarial generation networks, discriminator is crucial for the pixel-wised authenticity and comprehensibility of generated images, as it can infer the overall semantic information of the image. The discriminator adopts a hybrid network structure of Residual Network-50 (Resnet50) and U-net, and pre-trains through random blurring images. By introducing adversarial loss functions, it mitigates the common blurring problems in facial images for GANs; Thirdly, the predictive performance of different models is analyzed and compared. In this research, a ResNet and U-net hybrid discriminator is introduced to highlight the generation of context and segment structural information for facial regions such as nose, eyes etc., enhancing facial local texture coherence in the reconstructed image. The proposed model has significant improvements in terms of accuracy. The experiments show that this research can effectively decrease the blur phenomenon in reconstructed facial images.

2 METHODOLOGY

2.1 Dataset Description and Preprocessing

The dataset used in this study is CelebFaces Attribute (CelebA-HQ) images, containing 30,000 high-quality celebrity photos resized to 256 x 256 pixels (Dataset). This dataset has been often used in tasks like face recognition, facial attribute analysis and face generation. Data transformations and augmentations are applied in the experiments, including converting images into Tensor format, random cropping, normalizing etc. These operations can improve the generalization ability of the model and reduce the overfitting in training process.

2.2 Proposed Approach

Using GANs to generate images, if only pixel wise loss (L1 or L2) is used to train parameters, it is easy to cause image blur in the generated image. This may be due to this prediction method simply comparing pixel wise errors without considering the contextual information of the image, so the predictor tends to minimize the pixel-wise L2 loss, but also generates blurry images. Generally, the discriminator in GANs trained together with the generator, which just recognizes original and artifact image parts, hasn't pre-knowledge for blurred images and facial segmentation features, so loss function has no related parts. ResNet enhances face context perception ability, while U-Net can enhance its semantic consistency based on facial feature segmentation. By using resnet50 and U-Net as hybrid discriminator, corresponding reconstruction error parts are also considered for loss function. The blurriness of reconstructed can be reduced. The process is shown in the Figure 1.

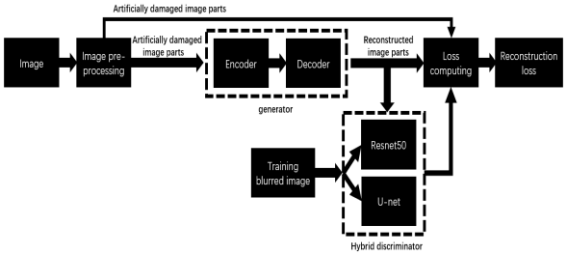

Figure 1: The pipeline of the model (Picture credit: Original).

2.2.1 ResNet

Unlike traditional neural networks, the outputs of each layer in ResNet are residuals of any previous layer output. ResNet is a neural network structure that solves the problem of gradient explosion and gradient vanishing in the training of deep neural networks by applying residual connectivity, thus allowing the network to be deeper and easier to train. The internal feature is the introduction of residual block which contains two or three convolutional layers, where the first convolutional layer is used to extract the features, the second convolutional layer is used to fuse the features, and finally the inputs and outputs are added by using the residual connection to get the output of the residual block. This residual join helps to keep more information in the training process, thus increasing the accuracy of the network. Each stage contains multiple residual blocks, and each residual block contains multiple convolutional layers and residual connections inside. The input of the whole network passes through a convolutional layer and pooling layer and enters the first stage, and then passes through multiple stages sequentially, and finally the output is obtained through a fully connected layer and a global average pooling layer. ResNet's framework process consists of the input passing through a convolutional layer and a pooling layer to get the input for the first stage; The first stage contains multiple residual blocks, each containing

two or three convolutional layers and residual connections; Stages two through four are similar to stage one and contain multiple residual blocks; Finally, the output is obtained through a global average pooling layer and a fully connected layer. The internal idea of ResNet is to solve the problem of gradient vanishing and gradient explosion during the training of deep neural networks through residual connectivity, thus allowing the network to be deeper and easier to train. In each residual block, the residual connection adds the inputs and outputs to get the output of the residual block, and this residual connection allows the network to retain more information in the training process, thus increasing the accuracy of the network.

Residual refers to that the inputs can skip computations of intermediate layers and go directly to the output. The approach of skipping layers is also called shortcut connection. This design helps neural networks to avoid problems such as vanishing gradient or exploding gradient which tends to appear with the increasing number of layers. ResNet enables training robustness and reliability while improving speed and accuracy. Considering computation complexity, ResNet50 is used. ResNet50, as its name suggests, is a 50-layer deep Residual Network which consists of six layers: input layer, convolution layer, residual block, pooling layer, fully-connected layer and output layer. The convolution layer executes convolution calculations on the input image to extract features. Then, the convolutional results are fed into the residual blocks. This allows more effective extraction of high-dimensional image features. The pooling layer down samples the feature maps. Finally, in the fully connected layer, the feature maps are connected and classified (He et al 2016).

2.2.2 U-Net

U-Net is a type of CNN that is specifically designed for image segmentation tasks. The structural characteristics of U-Net is its ability to connect the feature maps of traditional encoder-decoder structures with corresponding low-level feature maps. This connection allows U-Net to retain more information during the feature extraction process, which makes it highly effective at handling incomplete or blurry boundaries in images. The U-Net network includes a symmetric encoder and decoder structure, with the segmentation accuracy improved by connecting the encoder and decoder feature maps. The encoder is made up of multiple convolutional and max-pool layers, which progressively reduce the size and channel of feature maps while extracting high-level features. On the other hand, the decoder comprises multiple convolutional and up-sampling layers, which gradually restore the size and channel of feature maps. Overall, U-Net is a effective tool for image segmentation tasks, especially when dealing with complex or challenging images. Its ability to retain more information during feature extraction and connect encoder and decoder feature maps makes it a highly efficient and effective solution for a wide range of image segmentation applications. Additionally, by connecting low-level feature of the encoder directly with the corresponding feature maps of the decoder, U-Net can retain even more information (Shen et al 2018 & Ronneberger et al 2015).

2.2.3 Loss Function

Generator joint loss is a loss function that is obtained by weighting and summing the pixel-wise reconstruction loss and adversarial loss. Adversarial loss is comprised of adversarial context loss and semantic loss. The generator related loss functions are defined as show:

$$L_{pixel} = \|Img - G(Mask\_Img)\|_2^2 \quad (1)$$

$$L_{adv} = \|Ones - D(G(Mask\_Img))\|_2^2 \quad (2)$$

$$L_{gen\_joint} = \alpha * L_{pixel} + \beta * L_{adv} \quad (3)$$

where G and D respectively represent generator and discriminator network processing. L_{pixel} and L_{adv} represent pixel-wise loss and adversarial loss. They are both L2 loss functions. Img and Mask_Img respectively represents original image and masked or damaged image, Ones represents all 1 vector. The linear combination L_{gen_joint} of the 2 losses makes joint generator loss to train only generator parameters. α and β are weights that are obtained during training to obtain the optimal value.

Discriminator joint loss is a loss function that is obtained by weighting and summing the both the real and fake image discrimination loss. The discriminator related loss functions are defined as show:

$$L_{real} = \|Ones - D(Img)\|_2^2 \quad (4)$$

$$L_{fake} = \|Zeros - D(G(Mask\_Img))\|_2^2 \quad (5)$$

$$L_{disc\_joint} = \gamma * L_{real} + \varepsilon * L_{fake} \quad (6)$$

where L_{real} represents loss of judging real image as fake, and L_{fake} represents loss of judging fake image as real. They are both discriminator loss and take L2 loss functions. Zeros represents all 0 vector. The linear combination L_{disc_joint} of the 2 losses makes joint discriminator loss to train only discriminator parameters. γ and ε are weights that are obtained during training to obtain the optimal value.

2.3 Implementation Details

This experiment uses Python version 3.9 and imports Python libraries such as NumPy, pandas, torch and scikit-learn. The entire process is trained on CPU and Adam optimizer is used for optimization. The Adam optimizer can adaptively adjust the learning rate and handle sparse gradients, thus improving the speed and effectiveness of the model. The hybrid discriminator pre-training uses 10000 images from the CelebA-HQ training set. In order to make the discriminator more general in recognition, the images are resized to 128x128 using blurring processing, and then select a square area with upper left corner coordinates (80,80) and random integer 64~80 width/height. A two-dimensional blurring Gaussian filter with a mean of 0 and a variance of random 5~15 numbers is applied into the area. After processing, the Resnet50 and U-Net networks are pre-trained. The target is binary classification to accurately recognize original and blurred images. The pre-trained discriminator model is trained for 5 epochs and other model is trained for 15 epochs with a batch size of 16. The learning rate of the Adam optimizer is 0.0008. The number of CPU threads is 4.

3 RESULTS AND DISCUSSION

Here are some experiments results. Table 1 shows the pre-training data for hybrid discriminator. Table II shows the loss comparison between CNN discriminator model and this research. Furtherly comparing reconstructed images quality of CelebA-HQ dataset by using the discriminator structure in (Pathak et al 2016) and the pre-trained hybrid discriminator structure proposed in this paper, fig. 2 shows research in this paper has better quality.

Table 1: Pre-Training Data for Hybrid Discriminator.

Epochs	Test loss for pre-train discriminator
1	0.013
2	0.008
3	0.007

Table 2: Loss Comparison in Last 3 Epochs Between (Pathak et al 2016) and Research.

Epochs	Loss data in (Pathak et al 2016)			research		
	gen_adv_loss	gen_pixel_loss	disc_loss	gen_adv_loss	gen_pixel_loss	disc_loss
11	0.999	0.196	0.0013	0.322(-68%)	0.149(-24%)	0.241
12	0.999	0.194	0.0013	0.321(-68%)	0.147(-24%)	0.246
13	0.999	0.191	0.0012	0.328(-68%)	0.136(-28%)	0.238

According to Table 1, the hybrid discriminator is highly effective in distinguishing between the original and blurred images, with a minimal loss value of 0.008. Additionally, Table II shows that the research in this paper resulted in a 68% decrease in gen_adv_loss, a 24% decrease in gen_pixel_loss, and a significant reduction in error. As shown in Figure 2, the introduction of the discriminator through context and segmentation mechanisms not only reduces the loss but also significantly improves the subjective recognition effect of some reconstructed images, particularly in areas such as the nose and eyes, where the edge contour is more pronounced. These results indicate that the hybrid discriminator is a powerful tool for image

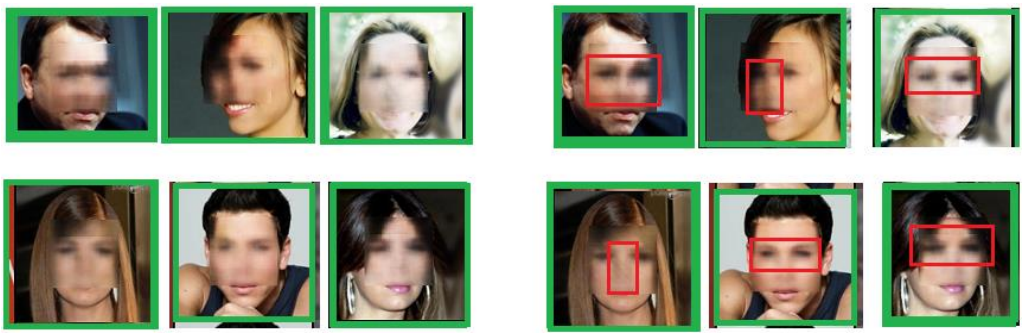

Figure 2: Reconstruction images comparison between Pathak et al's study (2016) and research, the left 6 images is from Pathak et al's study (2016), the right 6 images from this paper research (Picture credit: Original).

reconstruction tasks. For the reconstruction of 64x64 regions on 128x128 resolution images, the entire GANs training typically requires 20 epochs and takes approximately 45 minutes on a single Nvidia RTX3090 GPU. However, by parallelizing multiple GPUs, the training time can be significantly reduced.

4 CONCLUSION

Intuitively, for better repairing defective areas of an image, it is essential to understand the content of the surrounding images, and use the correlation of the image area content to infer the defective content. Previous research from Pathak et al's study (2016) found that the repaired images areas generally are blurry, especially contours around the eyes and nose. During the training process, I also found that even after multiple rounds of training, even if the image contour is blurry, the discriminator's error remains basically unchanged. That is, the discriminator is no longer sensitive to this kind of blurring, so I decided to use relevant pre-blurred image as training set and also combine context with segmentation analysis to enhance the discriminator's sensitivity to blurry contour.

This research proposes a hybrid neural network discriminator network structure of resnet50 and U-net for image inpainting, which has good perception ability for facial textures and segmentation structures. When combined with GANs, it can significantly reduce the blurring in reconstructed images. Specifically, firstly, a 2D Gaussian filter is used to randomly blur the images to generate a pre-trained blurred image set. Secondly, through the pre-training by using the blurred image set, the hybrid discriminator can discriminate the original image from the blurred image. Thirdly, build GANs networks for image generation training, and compare the reconstructed loss values and the quality of reconstructed images between this study and other papers. The experimental results indicate a significant decrease in the loss, while the subjective recognition effect of some reconstructed images is also significantly improved, particularly in areas such as the nose and eyes, where the edge source is more pronounced.

In the future, further work about optimizing the discriminator network structure, pre-training method and loss models design will be continued to recognize different size texture and thus achieving better general image reconstruction results.

REFERENCES

R. Shah, A. Gautam and S. K. Singh, "Overview of Image Inpainting Techniques: A Survey," 2022 IEEE Region 10 Symposium (TENSYMP), 2022, pp. 1-6.

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A.A. Efros. "Context encoders: Feature learning by inpainting," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets," Proceedings of Advances in neural information processing systems (NIPS), vol. 27, 2014, pp. 2672–2680.

S. Iizuka, E. Simo-Serra, and H. Ishikawa. "Globally and locally consistent image completion," ACM Transactions on Graphics, vol. 36, 2017, pp. 1–14.

J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. "Decaf: A deep convolutional activation feature for generic visual recognition," Proceedings of the 31st International Conference on Machine Learning (ICML), 2014, pp. 647-655.

C. Doersch, A. Gupta, and A.A. Efros. "Unsupervised visual representation learning by context prediction," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1422-1430.

R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. "Semantic image inpainting with deep generative models," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5485–5493.

Dataset https://www.kaggle.com/datasets/badasstechie/celebahq-resized-256x256

K. He, X. Zhang, S. Ren and J. Sun. "Deep Residual Learning for Image Recognition," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

Z. Y. Shen, W.S. Lai, T.F. Xu, J. Kautz, Y. Ming-Hsuan, "Deep Semantic Face Deblurring," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8260-8269.

O. Ronneberger, P. Fischer, T. Brox, "U-net: convolutional networks for biomedical image segmentation," Proceedings of Springer International conference on medical image computing and computer-assisted intervention, 2015, pp 234–241.