ChinaRxiv 中X

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202204.00071

Asymmetric Periodic Inference Cyclic Progressive Face Restoration Algorithm Research Postprint

Authors: Li Yaqian, Zhang Xuyao, Li Qilong

Date: 2022-04-07T15:01:57+00:00

Abstract

To address the issue that inpainting networks in generative adversarial networks fail to simultaneously preserve global and local consistency of images while incurring substantial computational load, we incorporate the concept of progressive inpainting into an asymmetric U-Net architecture. First, we propose an asymmetric cyclic feature reasoning module that enhances the correlation between inpainted content and surrounding known pixels, thereby improving the global consistency of restored images. Second, we introduce a novel U-Net-structured generator network that prevents unknown pixels in the encoder from entering the decoder, thus averting feature corruption within the decoder. Finally, the incorporation of perceptual loss and style loss further enhances the inpainting performance under subjective evaluation. Experimental results on face image datasets demonstrate that the proposed algorithm achieves significant improvements in both subjective visual quality and objective metrics.

Full Text

Preamble

Vol. 39 No. 8
Application Research of Computers
ChinaXiv Partner Journal

Asymmetric Periodic Inference Cyclic Progressive Face Completion Algorithm Research

Li Yaqian, Zhang Xuyao†, Li Qilong
(School of Electrical Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China)

Abstract: To address the problem that inpainting networks in generative adversarial networks cannot simultaneously maintain global and local consistency while suffering from high computational load, this paper introduces the concept of progressive inpainting based on an asymmetric U-Net architecture. First, we propose an asymmetric periodic feature inference module that enhances the correlation between inpainted content and surrounding known pixels, thereby improving global consistency in restored images. Second, we present a novel U-Net structured generator network that prevents unknown pixels from the encoder from entering the decoder and corrupting features. Finally, we incorporate perceptual loss and style loss to enhance inpainting quality under subjective evaluation. Experiments on face image datasets demonstrate significant improvements in both subjective visual effects and objective metrics.

Keywords: generative adversarial network; gradual inpainting; asymmetric periodic feature inference; image inpainting

0 Introduction

Images serve as a crucial information carrier in daily life, and when they become damaged, aged, or partially lost, they can easily lead to misinterpretation. Since real-world images are captured under unconstrained conditions, they frequently suffer from occlusion, stains, or damage, which negatively impacts image recognition, detection, and segmentation tasks. Consequently, image inpainting technology has become urgently needed, particularly as image recognition and segmentation tasks have entered everyday usage scenarios.

Traditional artisanal photo restoration typically begins with repairing image lines or object contours, followed by color and texture restoration. Li et al. [1] first applied GANs to face completion, drawing inspiration from the artisanal approach of "contour first, texture later." By incorporating both global and local loss functions in the discriminator, their method generated images that were not only semantically valid but also visually harmonious. This pioneered a class of hierarchical inpainting approaches that decompose the restoration task into stages.

Nazeri et al. [2] proposed a two-step edge-color inpainting algorithm that first uses a PatchGAN [3] for edge restoration, followed by another PatchGAN for color restoration. This tandem GAN architecture enables hierarchical processing of the inpainting network, yielding finer restoration results. Xiong et al. [4] argued for explicit separation between structure inference and content restoration. Their model guides inpainting through precise boundary prediction, where a boundary restoration module infers reasonable structures within the region to be inpainted, and an image restoration module generates content based on these predicted boundaries, while also integrating coarse-to-fine network restoration [5] into each GAN.

A similar approach employs a hierarchical VQ-VAE based multiple-solution inpainting method [6], which separates structure and texture by learning a conditional autoregressive network for the distribution of structural features. For texture generation, it proposes a structural attention module to capture long-range correlations in structural features, improving structural consistency and texture realism [7]. Li et al. designed an edge restoration module that simplifies the edge restoration process, enabling edge-color restoration within a single GAN. Through a Visual Structure Reconstruction (VSR) module [8], edge restoration is performed in the initial layers of the generator while color restoration occurs in the main generator pathway. Embedding edge restoration as a module within the network effectively avoids the high computational load and convergence difficulties associated with multiple cascaded GANs in hierarchical restoration.

Subsequently, Li et al. [9] argued that edge-color hierarchy is not optimal for image inpainting and designed the Recurrent Feature Reasoning (RFR) module, which adopts a region-graded restoration approach that progressively inpaints from the outside inward. The RFR module controls the restoration region at each step, using the content from the previous restoration step as the basis for inference in the next step. Knowledge Consistent Attention (KCA) [9] mechanisms are designed across feature maps at each layer to enhance inter-layer correlation.

While most existing models achieve better inpainting results, they also increase computational burden and fail to resolve the convergence issues of multiple cascaded GANs. A simple yet effective solution is to introduce hierarchical restoration concepts without increasing training difficulty by designing a plug-and-play feature reasoning module for inpainting tasks. The RFR module provides a new modular network design approach for progressive inpainting. Inspired by the RFR design philosophy, this paper reconstructs the RFR module and combines it with an asymmetric U-Net framework to design an asymmetric periodic feature inference module, thereby further enhancing the module's feature reasoning capability.

The main contributions of this paper are as follows:

(1) To address the problem that image inpainting networks cannot simultaneously maintain global and local consistency while requiring high computational load, we compare existing methods and find that the RFR module demonstrates good performance in reducing unknown pixels in the encoder portion and improving the decoder. We reconstruct and improve the RFR module into an asymmetric periodic feature module and apply it within a U-Net structure, proposing the Asymmetric Periodic Feature Inference (APFI) module.

(2) We propose a progressive inpainting [10] face completion network. We introduce PatchGAN to train the inpainting network, use gated convolutional layers as the convolution function in the encoder of the U-Net

structured generator, modify the U-Net bottom layer to use dual-channel dilated convolution, and combine perceptual loss and style loss to achieve excellent face completion results.

(3) Experiments on the CelebA face dataset use random masks for training and testing, and comparisons with state-of-the-art inpainting models including Edge-Connection [2], PConv [11], GatedConv [12], GFP-GAN [13], and LaMa [14] demonstrate the effectiveness of our algorithm in improving both subjective quality and objective metrics.

1.1 Cyclic Inpainting Network Framework

The periodic feature reasoning module draws inspiration from progressive inpainting concepts and the restoration region localization mechanism in partial convolutions, locating each step' s restoration region through mask channel updates. The periodic reasoning module consists of four components: area identification [15], feature reasoning [16], feature merging, and knowledge consistent attention. The module structure is shown in Figure 1.

Unlike current popular image inpainting methods, the RFR model does not use GANs and is overall a CNN structure. In our designed asymmetric periodic feature inference module, we introduce GANs' excellent unsupervised learning capability by designing a PatchGAN structure and employing a Markovian discriminator trained based on Wasserstein distance [17] to train the U-Net generator embedded with the RFR module.

Area identification uses the mask [18] update mechanism from partial convolutions to locate the size and position of the restoration region in each cycle. Since conventional convolution is unsuitable for image hole filling—because spatially shared convolution filters treat all input pixels or features as equally valid, leading to visual artifacts such as color differences, blurriness, and noticeable edge reflections around holes—this paper adopts gated convolution [12] as the convolution kernel for the feature reasoning module to enhance feature reasoning capability for inpainted content, constructing an asymmetric structure. The framework of our proposed asymmetric periodic feature inference module is shown in Figure 2.

By combining the above asymmetric feature reasoning module with the skip connection [19] mechanism of the U-Net network structure, we enhance the decoder' s image generation capability, thereby improving image visual performance. We adopt a VGG-Net pretrained on face datasets to construct perceptual loss and style loss, which enhances the generator' s performance in visual perception and style. The designed asymmetric periodic inference cyclic progressive face completion network framework is shown in Figure 3. The cyclic progressive inpainting network first embeds the RFR module after the input layer, ensuring that images entering the encoder have no unknown regions. The feature map without unknown regions is then fed into the U-Net network. This network operates similarly to an RNN: the feature map output from the first pass through the RFR module is re-input into the RFR module a second time. The second pass further fills in or predicts more reasonable values based on the feature map from the first pass. After several such cycles, the process proceeds to the feature fusion stage. This cyclic mechanism enables repeated utilization of features in the RFR module, allowing the model to achieve lightweight design while predicting more reasonable feature parameters and further enhancing module feature reasoning capability.

To improve the network’ s ability to extract semantics from high-level feature maps, the U-Net bottom layer is designed as a parallel dilated convolution structure. This algorithm uses gated convolutional layers as the convolution function in the encoder of the U-Net structured generator and modifies the U-Net bottom layer to dual-channel dilated convolution to prevent the image spatial structure from being corrupted at the generator’ s bottom layer.

1.2 Markovian Discriminator

Conventional GANs employ a discriminator to determine whether input comes from the real data distribution or generated data distribution. Unlike standard GAN discriminators, the Markovian discriminator is designed as a fully convolutional structure that outputs an $N \times N$ matrix X, where each element $X_{i,j}$ represents the authenticity of an image patch corresponding to a receptive field of size $M \times M$ in the original image. The final authenticity score is obtained by averaging all values in the matrix. Network training employs a metric based on Wasserstein distance to ensure convergence.

The Markovian discriminator offers several advantages over conventional discriminators: (1) It can focus on inpainting results at the patch level, significantly improving image detail performance. (2) By focusing on different regions of the input, it can learn to consider the contribution of different regions to identifying whether the image is real, enabling targeted attention to regions with high contribution. For example, when inpainting faces, the discriminator focuses on facial regions while reducing attention to background areas.

To improve PatchGAN convergence speed and enhance restoration diversity, we incorporate spectral normalization based on Miyato et al.’ s research. WGAN convergence requires the network to satisfy the Lipschitz constraint [20]. To meet this condition, the original WGAN employed gradient clipping to artificially constrain all convolutional layer weights to a range, which caused uneven gradient distribution. Subsequently, WGAN-GP [21] alleviated this issue using gradient penalty regularization. Later, Miyato et al. proposed spectral normalization regularization, which enables the network to satisfy the 1-Lipschitz condition without interfering with gradients. Our Markovian discriminator employs spectral normalization regularization during training to improve discriminator performance.

1.3 Perceptual Loss and Style Loss

For face inpainting tasks, this paper introduces perceptual loss and style loss [22]. Our algorithm uses a face recognition network pretrained on face datasets as supervision to compute perceptual differences and style consistency between inpainted and original images.

The restoration network output and original image have dimensions of 256$×256, $whileVGG16networkinputis224\times224images.Therefore, duringtraining, boththerestorationnetwo$ when computing losses. The VGG16 network structure is shown in Figure 4.

Perceptual loss is computed at the relu3_3 layer of VGG16, as shown in Equation (1):

$$\mathcal{L}_{\text{perceptual}} = \sum_{j} \frac{1}{C_j H_j W_j} \|\phi_j(I_{\text{gt}}) - \phi_j(I_{\text{out}})\|_2^2$$

where ϕ represents the VGG network, ϕ_j denotes using the feature map from the j-th layer of VGG network as output, and C_j, H_j, W_j represent the number of channels, height, and width of the feature map output from the j-th layer of VGG network.

Using VGG16 as a supervision network identifies pixel-level differences between two input images at the feature map level. Compared to direct pixel-level difference computation on images, feature maps represent high-level semantic features of images, so pixel-level computation on feature maps can represent perceptual differences between two images.

Style loss computation is shown in Equations (2) and (3):

$$\mathcal{L}_{\text{style}} = \sum_{j} \|\phi_j(I_{\text{gt}}) - \phi_j(I_{\text{out}})\|_2^2$$

$$G_{\phi_j}(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h,w} \phi_j(x)_{c,h,w} \cdot \phi_j(x)_{c',h,w}$$

where G_{ϕ_j} represents the Gram matrix computation, and other symbols maintain the same meaning as in perceptual loss.

1.4 Overall Loss

The total training loss consists of four components: PatchGAN loss, perceptual loss, style loss, and L1 reconstruction loss. Define the network input as $I_{\text{in}} = I_{\text{gt}} \odot M$, where M is the mask image. In M, regions with pixel value 0 represent unknown areas, while regions with pixel value 1 represent known areas. The input damaged image is obtained by pixel-wise multiplication of the original image and mask image. After inpainting through the cyclic progressive face completion network, the output is $I_{\text{out}} = G(I_{\text{in}})$. The total network loss function is shown in Equation (4):

$$\mathcal{L}_{\text{total}} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{per}}\mathcal{L}_{\text{per}} + \lambda_{\text{style}}\mathcal{L}_{\text{style}} + \lambda_{l1}\mathcal{L}_{l1}$$

where $\mathcal{L}_{\text{adv}}$ represents the generative adversarial network loss [23], $\mathcal{L}_{\text{per}}$ represents perceptual loss, $\mathcal{L}_{\text{style}}$ represents style loss, and $\mathcal{L}_{l1}$ represents L1 reconstruction loss.

Our network optimization employs the Adam optimizer to optimize the total loss. First, forward computation is performed to obtain the output inpainted image, then network loss is calculated according to Equation (4), and Adam is used to optimize the loss. The weight update is shown below:

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{n}_t} + \epsilon}\hat{m}_t$$

where first-order momentum $\hat{m}_t$ and second-order momentum $\hat{n}_t$ are given by Equations (6) and (7):

$$\hat{m}_t = \frac{m_t}{1 - \gamma_1^t}, \quad m_t = \gamma_1 m_{t-1} + (1 - \gamma_1)\frac{\partial \mathcal{L}}{\partial w_t}$$

$$\hat{n}_t = \frac{n_t}{1 - \gamma_2^t}, \quad n_t = \gamma_2 n_{t-1} + (1 - \gamma_2)\left(\frac{\partial \mathcal{L}}{\partial w_t}\right)^2$$

where γ_1, γ_2, and ϵ are optimizer parameters, and $\mathcal{L}$ is the total network loss. The backpropagation process first computes unbiased estimates of first-order and second-order momentum from the gradient of network loss with respect to weights, then substitutes them into Equation (5) to update weights.

2.1 Deep Learning Environment Configuration

Our algorithm's experimental code is implemented using the PyTorch framework. PyTorch provides functionality for computing derivatives of each node based on computational graphs, enabling dynamic construction of computational flow graphs and facilitating gradient backpropagation. It also integrates numerous commonly used deep learning functions, simplifying network construction and experimentation.

Our experimental platform operates on a 64-bit Linux kernel system, Ubuntu 18.04 LTS, with PyTorch version 1.2.0, OpenCV 4.4.0.46, numpy 1.16.5, scipy 1.1.0, and scikit-image 0.13.1. The GPU is an Nvidia 1080Ti with 8GB memory.

2.2 Dataset Introduction and Hyperparameter Configuration

CelebA [24] is a large-scale face dataset collected and created by the Chinese University of Hong Kong, containing 202,599 face images. Since the unprocessed dataset contains images of varying sizes and face positions, making learning difficult, we first preprocess using the author's provided face bbox regression box annotation files to obtain 202,599 aligned images, then resize them to 256$×$256. We randomly select 1,000 images as the test set, with the remaining images used as the training set.

Our experiments use the Adam optimizer with a learning rate of 4×10^{-4} during training and 1×10^{-5} during fine-tuning, batch size of 6, and 450,000 training iterations. Hyperparameters are set as $\lambda_{\text{adv}} = \lambda_{\text{per}} = 0.05$, $\lambda_{\text{style}} = 120$, and $\lambda_{l1} = 1$.

2.3 Experimental Results and Evaluation Metrics

The selected comparison models all demonstrate excellent performance on face data, including Edge-Connection [2], PConv [11], GatedConv [12], GFP-GAN [13], and LaMa [14]. We employ the irregular mask dataset proposed in PConv for training and testing masks. This dataset provides random occlusions with different occlusion ratios and shapes, simulating the randomness of real-world occlusions. Using this dataset for test masks ensures greater objectivity for horizontal comparison. Different mask ratios are illustrated in Figure 5, and images to be restored are shown in Figure 6.

The subjective performance of each algorithm on the face dataset is shown in Figure 7. Observation reveals that PConv, GatedConv, Edge-Connection, GFP-GAN, LaMa, and our algorithm all achieve good effects in restoring global semantics, but our algorithm demonstrates more prominent performance in restoring detailed structures, which the first five algorithms cannot accurately restore completely.

We conduct horizontal comparisons of objective metrics across different mask ratios. The irregular mask dataset categorizes different mask ratios, with each ratio containing 2,000 random mask images. To test our algorithm' s restoration performance under different mask ratios, we select three different mask ratio datasets: (0.1,0.2], (0.3,0.4], and (0.5,0.6] for testing. The horizontal comparison results of objective metrics between our algorithm and other algorithms under different mask ratios are shown in Tables 2-4.

Table 2 shows the metrics for different models when mask ratio is in (0.1,0.2]. **Table 3** shows the metrics when mask ratio is in (0.3,0.4]. **Table 4** shows the metrics when mask ratio is in (0.5,0.6].

Tables 2-4 demonstrate the objective effectiveness of our algorithm under different mask occlusion ratios. We employ three different metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and L1 error. Higher PSNR and SSIM values indicate better performance, while lower L1 error is preferable. L1 error represents the L1 norm between restored and original images. Our algorithm shows superiority across these three objective metrics on different datasets.

Experiments demonstrate that our algorithm achieves certain improvements over the original algorithm. We find that the iteration count of the non-periodic feature reasoning affects restoration results. Through experiments with different iteration counts, we determine that network restoration performance is optimal at 6 iterations, as shown in Table 5.

To explore the effectiveness of our proposed asymmetric periodic feature inference module and the network framework trained with PatchGAN, we design comparison experiments with RFR. The comparison results are shown in Table 1.

The above experiments prove the effectiveness of our innovations. Figure 8 shows our algorithm’ s restoration performance under different occlusion ratios, and Figure 9 shows our algorithm’ s performance on the same image with different occlusion ratios. Subjectively, our algorithm achieves good restoration results across large-range arbitrary occlusions.

---

3 Conclusion

To address the problem that inpainting networks cannot simultaneously maintain global and local consistency while requiring high computational load, this paper proposes an asymmetric periodic feature inference module that enhances the correlation between inpainted content and surrounding known pixels. First, we introduce progressive inpainting concepts by combining a progressive inpainting module (RFR) with an asymmetric network architecture and reconstructing the progressive inpainting module. Second, we employ a U-Net framework with an embedded asymmetric RFR module, introduce PatchGAN to train the inpainting network, use gated convolutional layers as the convolution function in the encoder of the U-Net structured generator, and modify the U-Net bottom

layer to dual-channel dilated convolution to prevent image spatial structure corruption at the generator' s bottom layer. Finally, combining perceptual loss and style loss yields excellent face completion results. Experimental results demonstrate that our algorithm shows significant improvements in objective metrics including PSNR, SSIM, and L1 error, as well as in subjective visual quality compared to existing algorithms.

References

[1] Li Yijun, Liu Sifei, Yang Jimei, et al. Generative face completion [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 2017: 3911-3919.

[2] Nazeri K, Ng E, Joseph T, et al. Edgeconnect: Generative image inpainting with adversarial edge learning [J]. arXiv preprint arXiv: 1901.00212, 2019.

[3] Du Xuemei. Research on image completion algorithm based on GAN [D]. ChengDu: University of Electronic Science and Technology of China, 2021.

[4] Xiong Wei, Yu Jiahui, Lin Zhe, et al. Foreground-aware image inpainting [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5840-5848.

[5] Moskalenko A, Erofeev M, Vatolin D. Deep Two-Stage High-Resolution Image Inpainting [J]. CEUR Workshop Proceedings, 2020, 2744.

[6] Peng Jialun, Liu Dong, Xu Songcen, et al. Generating Diverse Structure for Image Inpainting With Hierarchical VQ-VAE [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10775-10784.

[7] Liao Liang, Xiao Jing, Wang Zheng, et al. Image Inpainting Guided by Coherence Priors of Semantics and Textures [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 6539-6548.

[8] Li Jingyuan, He Fengxiang, Zhang Lefei, et al. Progressive reconstruction of visual structure for image inpainting [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[9] Li Jingyuan, Wang Ning, Zhang Lefei, et al. Recurrent Feature Reasoning for Image Inpainting [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7760-7768.

[10] Wang Tengfei, Ouyang Hao, Chen Qifeng. Image Inpainting with External-internal Learning and Monochromic Bottleneck [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 5120-5129.

[11] Liu Guilin, Reda F A, Shih K J, et al. Image inpainting for irregular holes using partial convolutions [C]// Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2018: 85-100.

[12] Yu Jiahui, Lin Zhe, Yang Jimei, et al. Free-form image inpainting with gated convolution [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4471-4480.

[13] Wang Xintao, Li Yu, Zhang Honglun, et al. Towards real-world blind face restoration with generative facial prior [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9168-9178.

[14] Suvorov R, Logacheva E, Mashikhin A, et al. Resolution-robust Large Mask Inpainting with Fourier Convolutions [C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.

[15] Gong Yunpeng. A general multi-modal data learning method for Person Re-identification [J]. arXiv preprint arXiv: 2101.08533, 2021.

[16] Ai Mingxi, Xie Yongfang, Tang Zhaohui, et al. Deep learning feature-based setpoint generation and optimal control for flotation processes [J]. Information Sciences, 2021, 578: 644-658.

[17] Jam J, Kendrick C, Drouard V, et al. Symmetric Skip Connection Wasserstein GAN for High-resolution Facial Image Inpainting [C]// 16th International Conference on Computer Vision Theory and Applications.

[18] Zeng Wenwen, Yang Yang, Zong Xiaopin. An improved Mask R-CNN based method for segmentation of on-shelf book spine image instances [J]. Computer Applications Research, 2021, 38(11): 3456-3459+3505.

[19] Li Dahai, Wang Yufeng, Wang Zhendong. A novel U-Net model for remote sensing image cloud segmentation problem [J]. Computer Application Research, 2021, 38(11): 3506-3509+3516.

[20] He Zhiyu, He Jianping, Chen Cailian, et al. CPCA: A chebyshev proxy and consensus based algorithm for general distributed optimization [C]// 2020 American Control Conference (ACC). IEEE, 2020: 94-99.

[21] Wei Xiang, Gong Boqing, Liu Zixia, et al. Improving the improved training of wasserstein gans: A consistency term and its dual effect [J]. arXiv preprint arXiv: 1803.01541, 2018.

[22] Cai Shaoyu, Zhu Kening, Ban Y, et al. Visual-Tactile Cross-Modal Data Generation using Residue-Fusion GAN with Feature-Matching and Perceptual Losses [J]. IEEE Robotics and Automation Letters, 2021, 6(4): 7525-7532.

[23] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139-144.

[24] Liu Ziwei, Luo Ping, Wang Xiaogong, et al. Deep learning face attributes in the wild [C]// Proceedings of the IEEE international conference on computer

vision. IEEE, 2015: 3730-3738.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.