Image Inpainting Through Coherent Semantic Attention

Jialiang Yan

Abstract—**The latest deep learning-based approaches have shown promising results for the challenging task of inpainting missing regions of an image. However, the existing meth ods often generate contents with blurry textures and dis torted structures due to the discontinuity of the local pix els. From a semantic-level perspective, the local pixel dis continuity is mainly because these methods ignore the se mantic relevance and feature continuity of hole regions. To handle this problem, we investigate the human behavior in repairing pictures and propose a fifined deep generative model-based approach with a novel coherent semantic at tention (CSA) layer, which can not only preserve contex tual structure but also make more effective predictions of missing parts by modeling the semantic relevance between the holes features. The task is divided into rough, refifine ment as two steps and model each step with a neural net work under the U-Net architecture, where the CSA layer is embedded into the encoder of refifinement step. To sta bilize the network training process and promote the CSA layer to learn more effective parameters, we propose a con sistency loss to enforce the both the CSA layer and the corresponding layer of the CSA in decoder to be close to the VGG feature layer of a ground truth image simulta neously. The experiments on CelebA, Places2, and Paris StreetView datasets have validated the effectiveness of our proposed methods in image inpainting tasks and can ob tain images with a higher quality as compared with the existing state-of-the-art approaches.**

Index Terms—**convolutional neural network,deep learning, generative adversarial network,image inpainting.**

I. Introduction

Image inpainting is the task to synthesize the missing or damaged parts of a plausible hypothesis, and can be uti lized in many applications such as removing unwanted ob jects, completing occluded regions, restoring damaged or corrupted parts. The core challenge of image inpainting is to maintain global semantic structure and generate realistic texture details for the missing regions.

Traditional works [2, 3, 11, 12, 34] mostly develop tex ture synthesis techniques to address the problem of hole fifill ing. In [2], Barnes et al. propose the Patch-Match algorithm which iteratively searches for the best fifitting patches from hole boundaries to synthesize the contents of the missing parts. Wilczkowiak et al. [34] take further steps and detect desirable search regions to fifind better match patches. How ever, these methods fall short of understanding high-level semantics and struggle at reconstructing patterns that are locally unique. In contrast, early deep convolution neural networks based approaches [17, 24, 30, 39] learn data dis tribution to capture the semantic information of the image, and can achieve plausible inpainting results. However, these methods fail to effectively utilize contextual information to generate the contents of holes, often leading to the results containing noise patterns.

Some recent studies effectively utilize the contextual in formation and obtain better inpainting results. These meth ods can be divided into two types. The fifirst type [32,36,42] utilizes spatial attention which takes surrounding image fea tures as references to restore missing regions. These meth ods can ensure the semantic consistency of generated con tent with contextual information. However, they just focus on rectangular shaped holes, and the results always tend to show pixel discontinuous and have semantic chasm (See in Fig 1(b, c)). The second type [26, 41] is to make the pre diction of the missing pixels condition on the valid pixels in the original image. These methods can handle irregular holes properly, but the generated contents still meet prob lems of semantic fault and boundary artifacts (See in Fig 1(g, h)). The reason that the above mentioned methods do not work well is because they ignore the semantic relevance and feature continuity of generated contents, which is cru cial for the local pixel continuity.

II. Procedure for Paper Submission

In order to achieve better image restoration effect, we investigate the human behavior in inpainting pictures and fifind that such process involves two steps as conception and painting to guarantee both global structure consistency and local pixel continuity of a picture. To put it more concrete, a man fifirst observes the overall structure of the image and conceives the contents of missing parts during conception process, so that the global structure consistency of the im age can be maintained. Then the idea of the contents will be stuffed into the actual image during painting process. In the painting process, one always continues to draw new lines and coloring from the end nodes of the lines drawn previ ously, which actually ensures the local pixel continuity of the fifinal result.

Inspired by this process, we propose a coherent semantic attention layer (CSA), which fifills in the unknown regions of the image feature maps with the similar process. Initially, each unknown feature patch in the unknown region is ini tialized with the most similar feature patch in the known regions. Thereafter, they are iteratively optimized by con sidering the spatial consistency with adjacent patches. Con sequently, the global semantic consistency is guaranteed by

Manuscript received October 23, 2021

Jialiang Yan, School of Computer Science and Technology, Tiangong University,Tianjin,China

the fifirst step, and the local feature coherency is maintained by the optimizing step.

Similar to [42], we divide the image inpainting into two steps. The fifirst step can be constructed by training a rough network to rough out the missing contents. A refifinement network with the CSA layer in encoder guides the second step to refifine the rough predictions. In order to make net work training process more stable and motivate the CSA layer to learn more effective features, we propose a con sistency loss to measure not only the distance between the VGG feature layer and the CSA layer but also the distance between the VGG feature layer and the the correspond ing layer of the CSA in decoder. Meanwhile, in addition to a patch discriminator [18], we improve the details by introducing a feature patch which is simpler in formula tion, faster and more stable for training than conventional one [29]. Except for the consistency loss, reconstruction loss, and relativistic average LS adversarial loss [28] are incorporated as constraints to instruct our model to learn meaningful parameters. We conduct experiments on standard datasets CelebA [27], Places2 [44], and Paris StreetView [8]. Both the qualitative and quantitative tests demonstrate that our method can generate higher-quality inpainting results than existing ones. (See in Fig 1(d, i)). Our contributions are summarized as follows: We propose a novel coherent semantic attention layer to construct the correlation between the deep features of hole regions. No matter whether the unknown region is irregular or centering, our algorithm can achieve state-of-the-art inpainting results.

To enhance the performance of the CSA layer and training stability, we introduce the consistency loss to guide the CSA layer and the corresponding decoder layer to learn the VGG features of ground truth. Mean while, a feature patch discriminator is designed and jointed to achieve better predictions.

Our approach achieves higher-quality results in com parison with [26,36,41,42] and generates more coher ent textures. Even the inpainting task is completed in two stages, our full network can be trained in an end to end manner.

III. APPROACH

Our model consists of two steps: rough inpainting and refifinement inpainting. This architecture helps to stabi lize training and enlarge the receptive fifields as mentioned in [42]. The overall framework of our inpainting system is shown in Fig 2. Let Igt be the ground truth images, Iin be the input to the rough network, the M and M denote the missing area and the known area in feature maps respec tively. We fifirst get the rough prediction Ip during the rough inpainting process. Then, the refifinement network with CSA layer takes the Ip and Iin as input pairs to output fifinal result Ir. Finally, the patch and feature patch discriminators work together to obtain higher resolution of Ir.

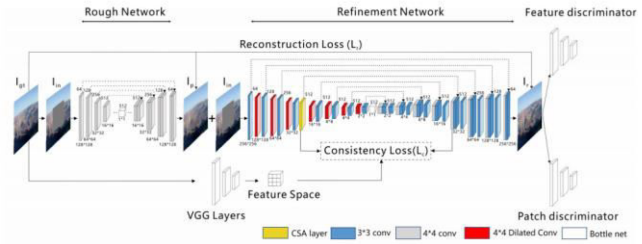

Figure 1: our network

IV. EXPERIMENTS

Our proposed model is evaluated on the datasets CelebA .Results are compared against the current state-of-the-art methods both qualitatively and quantitatively. Shown in Figure 2 and Figure 3. The image inpainting indicators are compared in Table 1

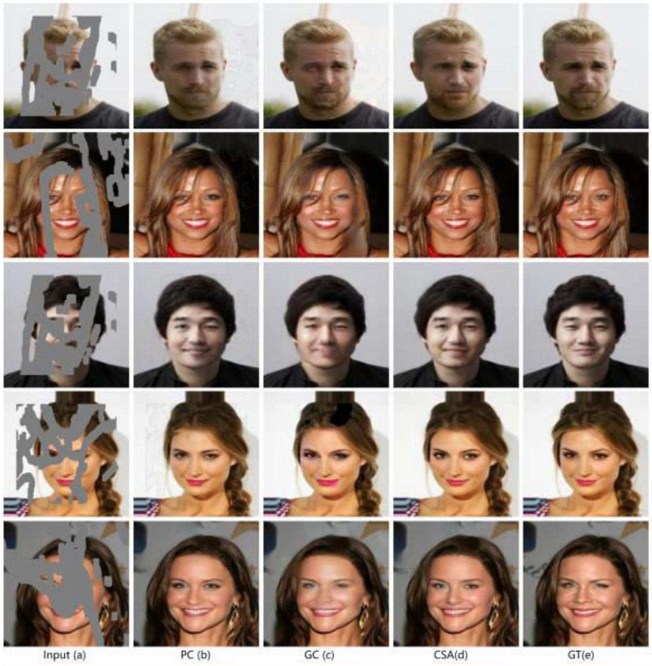

Figure 2 Our results show in celebA

Method	PSNR	SSIM
PC	21.34	0.814
GC	21.42	0.819
Our	**21.75**	**0.823**

V. CONCLUSION

In this paper, we proposed a fifined deep generative model based approach which designed a novel Coherent Semantic Attention layer to learn the relationship between features of missing region in image inpainting task. The consistency loss is introduced to enhance the CSA layer learning ability for ground truth feature distribution and training stability. Moreover, a feature patch discriminator is joined into our model to achieve better predictions. Experiments have verified the effectiveness of our proposed methods. In future, we plan to extend the method to other tasks, such as style transfer and single image super-resolution.

TRANSFER AND SINGLE IMAGE SUPER-RESOLUTION.REFERENCES

[1] Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu,D. Warde-Farley, S. Ozair, A. Courville, and Y . Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014. 1, 3
[2] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? ACM Transactions on graphics (TOG), 31(4):44 – 1,2012. 2
[3] Z. Liu, P . Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), 2015. 2, 5
[4] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba.Places: A 10 million image database for scene recognition.IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. 2, 5
[5] J. Canny. A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence,(6):679 – 698, 1986. 3
[6] S. Nowozin, C. H. Lampert, et al. Structured learning and prediction in computer vision. F oundations and Trends R ? inComputer Graphics and Vision, 6(3 – 4):185 – 365, 2011. 3
[7] S. Xie and Z. Tu. Holistically-nested edge detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1395 – 1403, 2015. 3, 7,12
[8] P . Sangkloy, J. Lu, C. Fang, F. Y u, and J.Hays.Scribbler:Controlling deep image synthesis with sketch and color. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 2, 2017. 3
[9] P . Isola, J.-Y . Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 3, 11
[10] M. Mirza and S. Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014. 3
[11] J.-Y . Zhu, T. Park, P . Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In The IEEE International Conference on Computer Vision (ICCV), 2017. 3, 11
[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu,D. Warde-Farley, S. Ozair, A. Courville, and Y . Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672 – 2680, 2014. 1, 3
[13] J. Johnson, A. Alahi, and L. Fei-Fei.Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision (ECCV), pages 694 – 711.Springer, 2016. 3, 4, 11
[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),pages 770 – 778, 2016. 3
[15] P . Isola, J.-Y . Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 3, 11
[16] D. Ulyanov, A. V edaldi, and V . Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 3