Survey on Image Inpainting

Tianyuan Zhang
Department of Computer Science
Peking University
1600012888@pku.edu.cn

Abstract

Image inpainting, a task of synthesizing alternative contents in missing regions, can be used in many applications. Although many approaches have been proposed, it remains a challenging problem. In this paper, we give a brief review over some of the most effective deep-learning based method.

1 Introduction

Image inpainting is a task of synthesizing visually realistic and semantically correct contents in missing regions. There are mainly two broad approaches: classical patch-based ones[1] and convolutional networks based deep generative models. The former approach can synthesize plausible stationary textures but typically does not work well in cases where semantics are needed. The later approach can utilize semantics learned from large scale datasets and seems more promising. In this paper, we mainly focus on the later deep learning based approach.

1.1 Datasets

Large scale datasets with various scenes such as ImageNet dataset[2] and Places2 datasets[3], are needed in order to help the networks learn sufficient semantics. Datasets with strong priors shuch as CelebA-HQ[4, 5] can also be used.

1.2 Evaluation

PSNR, SSIM[6] and MSSIM[7] can be used in evaluation, but these perceptual motivated simple distance metric often fail to account for many nuances of human perception. Though some effective deep learning based metric has been proposed and have been widely used in many image synthesizing tasks, the most reliable known method for evaluating the realism of synthesized images remains perceptual experiments with human observers, such experiments have been used in [8].

2 Utilizing different losses

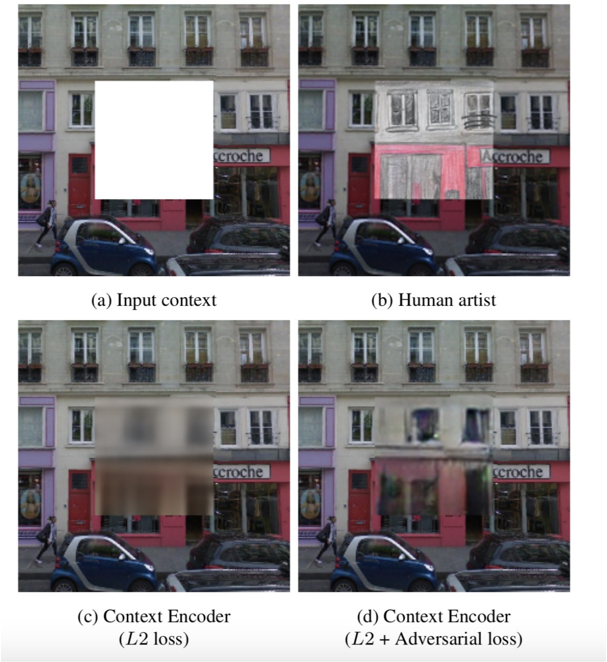
For inpainting problems, no single metric can assess the results of proposed method, so no simple loss can be used to train a feed-forward generative networks. Many loss functions or metrics have been used in this area.

2.1 Adversarial Loss

GAN [9] was first proposed to learn a generative model of a data distribution, where a critic was jointly trained and provide loss gradients to the generative models.

CVDL course project.

Context encoder[10] was the first to introduce adversarial loss into image inapinting task, using similar GAN loss proposed in [8] and L2 reconstruction loss. Figure 1 shows the effectiveness of adversarial loss.



(a) Input context

(b) Human artist

(c) Context Encoder ($L2$ loss)

(d) Context Encoder ($L2$ + Adversarial loss)

Figure 1: Demonstration of the effectiveness of adversarial loss used in Context Encoder[10]. This figure was from [10].

Though context encoder can capture high-level recognition of scenes, but it's difficult to synthesize visually plausible textures. Iizuka et al.[11] introduced Global & Local GAN to help improve the image quality, the global critic looks at the entire image to assess if it is coherent as a whole and the local discriminator looks only at the completed region to ensure the local consistency of the generated patches. Equipped with this powerful critic, the generator achieved state-of-the-art results in both regular and irregular hole inpainting tasks. Results can be seen in figure 2.

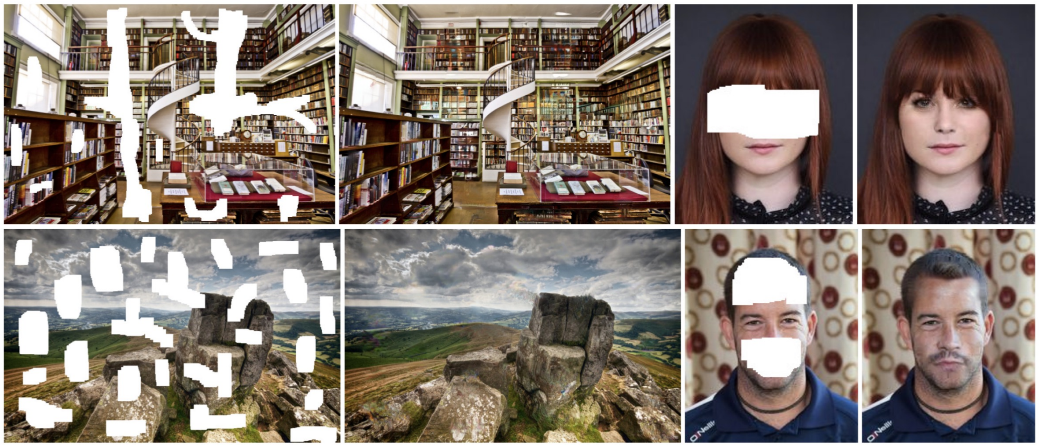


Figure 2: Results obtained by [11]. The masked region is shown in white.This figure was from [11].

Although Global & Local GAN has shown prominent results for large hole inpaitning problems, but there are some thing not appropriate with the Local GAN, it only emphasis on the completed region, which may push the generator to produce independent textures that are incompatible with the whole image.

Demir[12] attempted to ease this problem by paying attention to every patches. They proposed PGGAN(Patch & Global GAN). The Global-GAN does the same thing as before, and the Patch-GAN try to assess every synthesized patches. The proposed method are shown in figure 3.

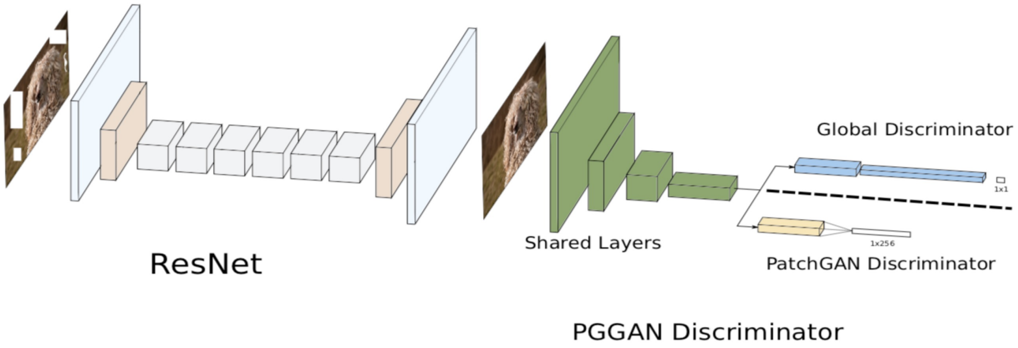


Figure 3: Archtecture of Generative network and PGGAN discriminator.

2.2 Perceptual loss

While quickly measuring the perceptual similarity between two images is nearly impossible, researchers have designed many metrics such as *PSNR* and *SSIM*, and these simple functions often fail to account for many nuances of human perception. Recently researchers in deep learning community found that features extracted by deep CNNs are extremely useful for measuring perceptual similarity and can be used for image synthesis.

Gatys et al.[13] refered to the feature responses in higher layers of VGG[14] as the content representation and the gram matrix of theses features as style representation, and succeeded to capture both the semantic and textures information of an image. Johnson et al.[15] proposed the use of perceptual loss function for training feed-forward networks for image transformation tasks, which can give similar results as in [13] but is three orders of magnitude faster. More impressively, Chen[16] showed that photographic images can be synthesized from semantic layouts by a single feed-forward network trained with only perceptual loss.

Given a GAN, Yeh et al.[17] designed a context distance by combining perceptual loss and adversarial loss, and searched for the closest encoding of the corrupted image in the latent manifold, then passed the encoding through the generative model to infer the missing content. Different from context encoder[10] and other method, where corrupted images and patch information was directly feed into the generator, the proposed method learns the representation of training data and predict meaningful content for missing regions. But due to the limitation of training a powerful generative model that can well capture the data distribution, this method does not work well for images with complex scenes or high resolution images.

Liu et al.[18] conducted tons of hyper-parameter tuning and finally successfully trained one feed-forward convolution networks using perceptual loss, TV loss and L1 reconstruction loss. Combined with partial convolution , the they achieved state-of-the-art results in image inpainting with irregular holes. Results are shown in figure 4.

3 Something Wrong with vanilla convolution

Vanilla convolution was used as a feature extractor in a sliding window fashion that treat all pixels equal, but treating pixels in holes and outsides holes or synthesized pixels equally was inherently wrong and may lead to visual artifacts such as color discrepancy and blurriness. Generative models using vanilla convolutions for image inpainting often need post processing to reduce such artifacts.

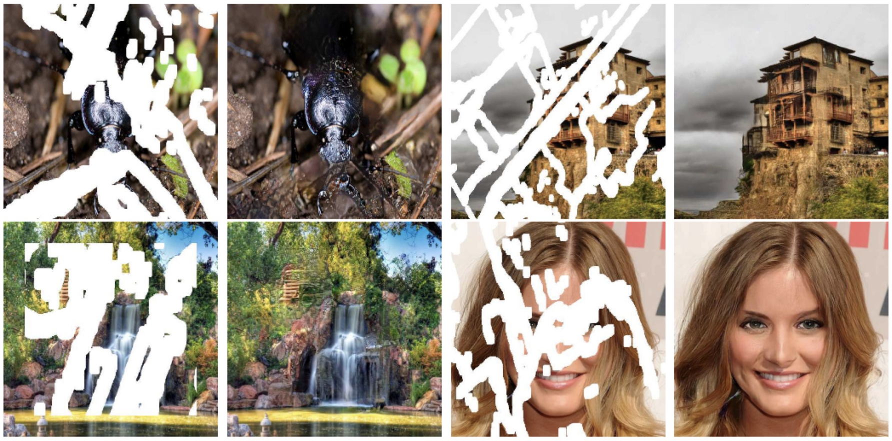


Figure 4: Results obtained by [18]. The masked region is shown in white.This figure was from [18].

Liu et al.[18] recently proposed partial convolutions where convolution is masked and normalized to be conditioned only on valid pixels.

Partial convolution is essentially a hard-gating single-channel un-learnable layer multiplied to input feature maps[19], and Yu et al.[19] generalized partial convolution by proposing gated convolution that learns a dynamic feature selection mechanism for each channel and each spatial location. Combining a powerful PatchGAN loss, the method achieved state-of-the-art results in free-from image inpainting and user-guided inpainting. The proposed gated convolution is illustrated in figure 5, and qualitative comparisons can been seen in figure 6.

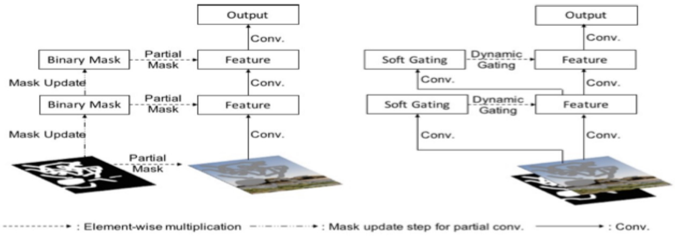


Figure 5: Illustration of partial convolution(left)and gated convolution(right).This figure was from [19].

Spatial attention modules can also be used to rescue vanilla convolution. Inspired by early works[1] which match and copy background patches into the hole region, Yu et al.[20] proposed a contextual attention layer which match and copy background patches in a learnable way. Illustration of the contextual attention layer can been seen in figure 7. This attention module was also exploited in [19].

3.1 Footnotes

4 Two stage strategy

Some classic inpainting techniques propagate in a gradually refine manner, first regress to the pixels near the boundary and then the pixels near the "new" boundary. Many deep learning based method also consists of multiple stages. Coarse-to-fine strategy are exploited in [20,22]

Inspired by the work done by Chen[13] that a visually realistic images can be synthesized from semantic layouts by a feed-forward network, Song et al.[21] factorized the image inpainting process into segmentation prediciton and segmentation guidance inpainting by first predicting the segmentation labels in the missing region then generate segmentation guidied inpainting results. This two step pipeline can be trained end-to-end with perceptual loss and adversarial loss.

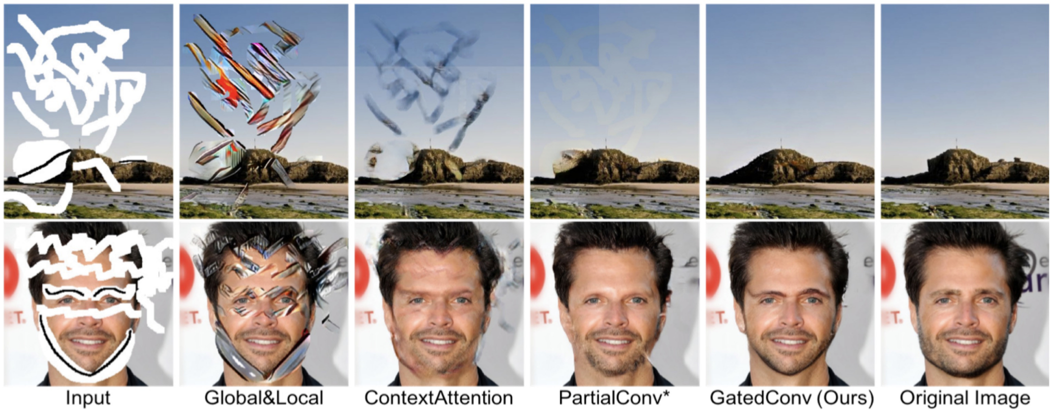


Figure 6: Qualitative Comparisons on the Places2 and CelebA-HQ validation sets.This figure was from [19].

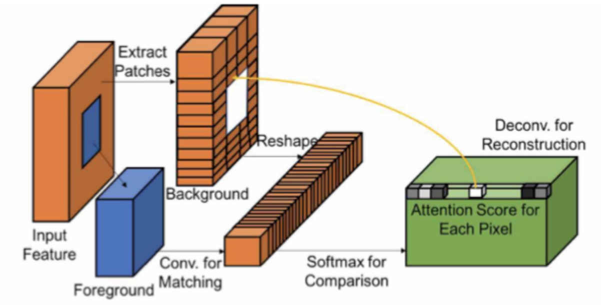


Figure 7: Illustration of the contextual attention layer. They used convolution and softmax to compute attention score of foreground pixels and background patches, then reconstruct foreground by performing deconvolution with background patches with attention score. This figure was from [20].

5 Acknowledgments

Thanks Xiangyu Zhang for insightful discussions.

References

[1] Barnes C, Shechtman E, Finkelstein A, et al. PatchMatch: A randomized correspondence algorithm for structural image editing[J]. ACM Transactions on Graphics (ToG), 2009, 28(3): 24.

[2] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115(3) (2015) 211–252

[3] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).

[4] Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV). (December 2015) [5] Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV). (December 2015)

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. TIP, 2004. 1, 2, 8, 12, 14

[7] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale struc- tural similarity for image quality assessment. In Signals, Sys- tems and Computers. IEEE, 2004. 1

[8] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a Laplacian pyramid of ad- versarial networks. In NIPS, 2015

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Gen- erative adversarial nets. In NIPS, 2014.

[10] Pathak, Deepak, et al. "Context encoders: Feature learning by inpainting." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[11] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. ACM Transactions on Graphics (TOG) 36, 4 (2017), 107.

[12] Demir, U., Unal, G. (2018). Patch-Based Image Inpainting with Generative Adversarial Networks. arXiv preprint arXiv:1803.07422.

[13] Gatys L A, Ecker A S, Bethge M.A neural algorithm of artistic style[J].arXiv preprintarXiv:1508.06576, 2015.

[14] Simonyan K, Zisserman A.Very deep convolutional networks for large-scale image recogni-tion[J].arXiv preprint arXiv:1409.1556, 2014.

[15] Johnson J., Alahi A., Fei-Fei L. (2016) Perceptual Losses for Real-Time Style Transferand Super-Resolution. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer VisionECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9906. Springer, Cham

[16] Chen Q, Koltun V. Photographic image synthesis with cascaded refinement net-works[C]//IEEE International Conference on Computer Vision (ICCV). 2017, 1(2): 3

[17] Yeh, Raymond A., et al. "Semantic image inpainting with deep generative models." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017

[18] Liu G, Reda F A, Shih K J, et al. Image inpainting for irregular holes using partial convolutions[J]. arXiv preprint arXiv:1804.07723, 2018.

[19] Yu, Jiahui, et al. "Free-Form Image Inpainting with Gated Convolution." arXiv preprint arXiv:1806.03589 (2018).

[20] Yu, Jiahui, et al. "Generative image inpainting with contextual attention." arXiv preprint (2018).

[21] Song, Yuhang, et al. "SPG-Net: Segmentation Prediction and Guidance Network for Image Inpainting." arXiv preprint arXiv:1805.03356 (2018).

[22] Yang, Chao, et al. "High-resolution image inpainting using multi-scale neural patch synthesis." The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 1. No. 2. 2017.