Image Inpainting with Informed Mask Generation

Anonymous submission

Paper ID

Abstract

Deep learning approaches have resulted in considerable improvements in image inpainting during the last few years. However, having accurate masks is difficult in practice in a variety of settings. For example, naturally occurring Image deformations are random and do not have a defined mask. Passing an explicit mask to solve the image inpainting problem is therefore tricky. This work proposes a novel method to solve image inpainting task without the need for any explicit mask and also modifying image denoising task as image inpainting task using gray scale masking. As our network architecture, we proposed a robust pipeline. Initially, a "Autoencoder" mask prediction network gets an incomplete color image as input and generates masks. The resulting mask will then be used to predict the complete edge map using "Gated Convolutions". This predicted edge map as well as an incomplete color image are sent to the refinement network for image inpainting. As a result, semantically convincing and visually appealing information is generated. We test our model end-to-end using publically available datasets CelebA and Paris StreetView, and show that it outperforms the existing methods.

1. Introduction

The process of filling missing pixels in an image is known as image inpainting [13]. It can be used for picture editing, image-based rendering, and computational photography. The main challenge of image inpainting is to produce visually realistic and semantically believable pixels for missing regions that are consistent with those that already exist.

Conventional methods to image inpainting are roughly characterized as sequential-based, CNN-based, and GAN-based [2]. The sequential-based approach is divided into two groups: patch-based approaches and diffusion-based methods. Patch-based solutions rely on techniques for patch-by-patch filling in the missing portion by searching for and copying well-matching replacement patches (i.e., candidate patches) in the undamaged part of the image. By contrast, diffusion-based techniques fill in the empty region (the "hole") by smoothly conveying visual content from the boundary to the interior of the missing region. These methods, however, are effective for simpler photos; nonetheless, when the image is complicated, such as containing a lot of texture and objects, or when covering a vast portion in the image, these methods fail.

Recent works have been inspired by rapid advances in deep convolutional neural networks (CNN) and generative adversarial networks (GAN) [3] [13] to formulate inpainting as a conditional image generation problem in which high-level recognition and low-level pixel synthesis are formulated into a convolutional encoder-decoder network that is jointly trained with adversarial networks to encourage coherency between generated and existing pixels. These research show that it is possible to generate realistic new material in highly ordered images such as faces, objects, and sceneries. Several methods for image inpainting that use convolutional neural networks (CNNs) or CNN-based encoder-decoder networks have been developed. Unfortunately, CNN-based techniques usually result in boundary distortions, deformed structures, and blurry textures that are inconsistent with the surrounding areas. This is partly due to the convolutional neural networks' inability to explicitly borrow or replicate data from remote (faraway) spatial locations.

In general, some of the image inpainting methods takes a corrupted image as input, as well as a mask that marks missing pixels, and restores it using the semantics and textures of uncorrupted parts. However, the mask requirement is tough to meet. Creating a mask by hand for such images is a time-consuming task. The work in this study is inspired by our observation that many existing image inpainting approaches need explicit masking. However, our idea is focused with recreating images utilizing only the image as input and solving the reconstruction problem by executing inpainting and image enhancements simultaneously. We also believed that our proposal method would work since it is sequential, beginning with prediction of simple structures like edges and then utilizing a reconstruction network(RN) to fill in texture and color. Furthermore, when combined

with informative mass, this should improve reconstruction outcomes since information in the masked regions will help in obtaining a weighted reconstruction.

Experiments on a variety of datasets, including faces and natural images, show that the proposed method produces higher-quality inpainting results than previous methods. Our contributions are summarized below: 1) An end-to-end pipeline that combines mask prediction network and edge prediction network and refinement network to fill missing regions and results the output as visually realistic image. 2) We also proposed a novel way for solving the Image Denoising problem by generating continuous valued masks.

2. Related Work

Several methods for image inpainting have been proposed. Patch-based [13] techniques have historically been used to find and paste the most similar image patch by gradually extending pixels adjacent to hole edges based on low-level information. These techniques work well in stationary textural regions but frequently fail in non-stationary images. Simakov et al. also propose a bidirectional similarity synthesis approach [10] to improve the capture and summarization of non-stationary visual data.

Recently, image inpainting systems based on deep learning are proposed to directly predict pixel values inside masks. A significant advantage of these models is the ability to learn adaptive image features for different semantics. Thus they can synthesize more visually plausible contents especially for images like faces, objects and natural scenes. Context encoder [8], which uses an encoder-decoder architecture and made considerable progress in deep convolutional neural networks, was one of the first deep learning systems built for picture inpainting (CNNs). Pathak et al. proposed training an encoder-decoder CNN while simultaneously decreasing pixel-wise reconstruction loss and adversarial loss. Aside from the encoder-decoder structure, a U-net-like structure was also used [11].

Yang et al. [12] employ a pre-trained VGG network to improve the context-encoder output by decreasing the feature difference of image background. Additionally, Yang et al. propose a method to tackle inpainting of large sections on large images. The problem with previous methods for these tasks is that they give blurry outcomes, with visible edges between context and reconstruction. To achieve high resolution inpainting, they use multi-scale approaches to generate high-frequency details on top of the reconstructed object. Furthermore, Yang et al. [12] and Yu et al. [13] introduced coarse-to-fine CNNs for image inpainting. To improve inpainting results from the coarse CNN, a combination of CNN and Markov Random Field [12] was used as a post-process to provide more realistic and detailed texture.

We employ an autoencoder for the Mask prediction network (MPN), followed by network architectures for the edge prediction network (EPN) and reconstruction network (RN) influenced by Kamyar et al. and Jiahui et al. Autoencoders and Variational Autoencoders (VAEs) have become prominent approaches for unsupervised learning of complex distributions. When compared to GANs, VAEs produce images that are too smooth, which is undesirable for inpainting applications. The free-form inpainting method presented in [14] is possibly the most close to the edge connect. It guides the inpainting process with hand-drawn sketches. However, our method predicts the masks automatically and learns to hallucinate edges in missing regions.

3. Method

3.1. Mask Prediction Network

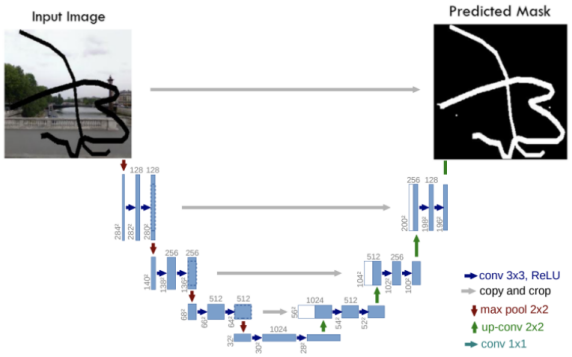

Figure 1. Mask Prediction Network

MPN is aimed at learned regions of image which would possibly require inpainting or reconstruction in. It is based on U-Net [9] which is a residual connection based encoder-decoder architecture.

3.2. Edge Prediction Network

The second module in our pipeline is the Edge Prediction Network. As the name suggests, this network is used to predict edges for the masked region in the image. These masks would later serve as an exemplar information to the subsequent inpainting network module.

Initially, we use Canny Edge Detector to generate edges for the unmasked region using the input (masked) image converted to grayscale. The generated edge, along with the grayscale input (masked) image and the mask predicted using the Mask Prediction Network is passed as an input to the Edge Connection Network. The Network then outputs the predicted edges for the masked areas, concatenated with the edges from the Canny Edge Detector. During the training process the predicted edges, along with the edges from the original (unmasked) image are passed through the discriminator.

The network architecture has been inspired from the Edge Connect paper [7]. In the generator architecture, we

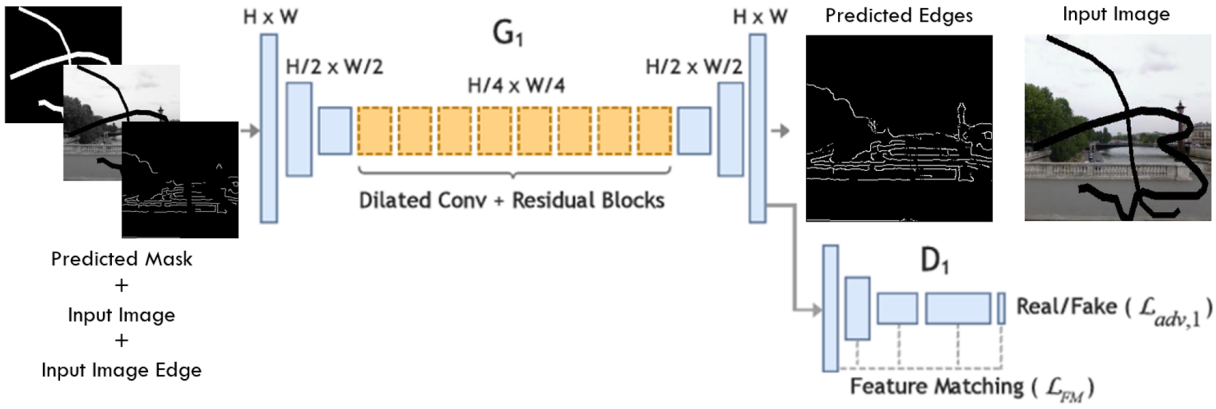

Figure 2. Edge Prediction Network

use a simple encoder-decoder network. The encoders downsample the input twice, which is the followed by 8 residual blocks, with dilated convolutions and ultimately the decoder upsamps the feature map twice. Dilated convolutions play a major role in the residual blocks, as it increases the receptive field. Also, to make the complete training procedure more stable, we add spectral normalisation [6] to all the traditional convolutions. Spectral Normalisation is a normalisation technique which provides a major advantage when training the discriminator, as it has a convenient property that Lipschitz constant is the only hyper-parameter that needs to be trained.

As for the discriminator part, we use a 70×70 PatchGAN architecture [4]. Let C(m)(n) denotes a 4x4 convolution layer with m filter and a stride of n. The discriminator has the following architecture: C-(64)(2), C-(128)(2), C-(256)(2), C-(512)(1), C-(1)(1). The end output is the score predicting if the 70x70 overlapping edge patches are real or fake. Instance normalization is used across all layers of the discriminator. .

3.3. Image Inpainting Model

The third and the final module for the complete pipeline is the image inpainting network. This network takes as input

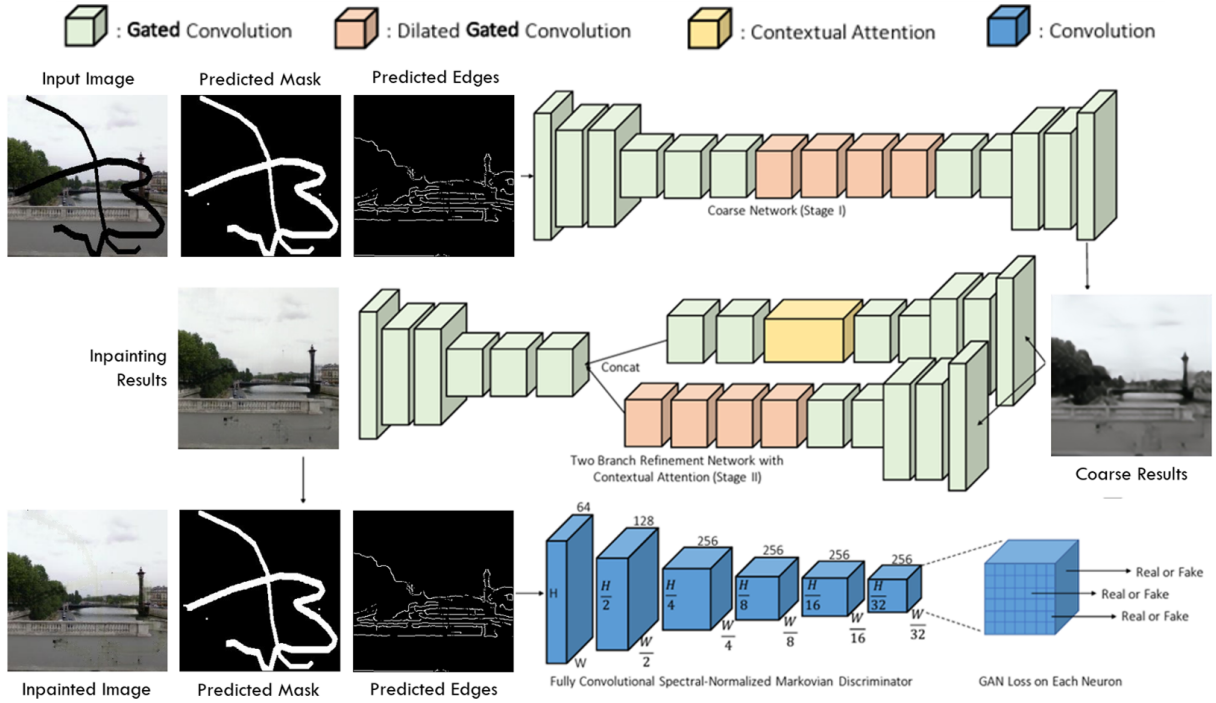

Figure 3. Image Inpainting Model

the output from the previous two models, which is the MPN and EPN. Hence, the three inputs to this model is the Predicted Mask, Predicted Edge and the input Image. The network is a two step generation process, where first a coarse inpainted image is generated, this coarse image is further refined to generate the final inpainted results. During the training phase, the refined inpainted results, along with the predicted edges and the predicted masks is passed through the discriminator network.

As mentioned above, for the inpainting generator we use a two-stage network with Gated Convolutions giving first coarse then fine results [14]. It is noteworthy, that the skip connections here have no significant impact in this generator architecture. This is mainly because for inner masked area, the information from these skip connections would be nearly zero. Hence, they do not pass-on detailed texture or color information to the decoder for those region.

This is why we use a simple auto-decoder architecture, without any skip connections for both the coarse and fine generators. Instead of regular convolutions, we use gated convolutions here. For better and improved refinement we use an additional contextual attention network [13] branch here, which which branched out from the refinement encoder and concatenates right before the refinement decoder.

For the discriminator part, we use a fully convolution network again with spectral normalization. Here, the output of the discriminator is a downsampled 1D vector. Each neuron in the output has a receptive field of the complete image, therefore we do not need any global discriminator here.

4. Experiments Results and Analysis

	PSV		CelebA	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
MPN				
EPN	22.2051	0.9569		
IPN	19.15	0.84		

Table 1. This table shows results for separately trained Mask Prediction, Edge Prediction and Inpainting network.

5. Conclusions

We have proposed a robust inpainting and reconstruction model and framework. This network is trained to inpaint a destroyed image. This image could be destroyed in certain regions either completely or partially. To achieve that our data preparation uses masks that contain continuous values such that it can be used as information indicating the amount of the data not destroyed. This is done alongside relaxing the need of an explicit mask which allows scalable task of image inpainting to be performed. Our inpainting network combines the hypothesis of blind but guided inpainting and thus our future work will revolve around the using better guide heuristics as well as making the result

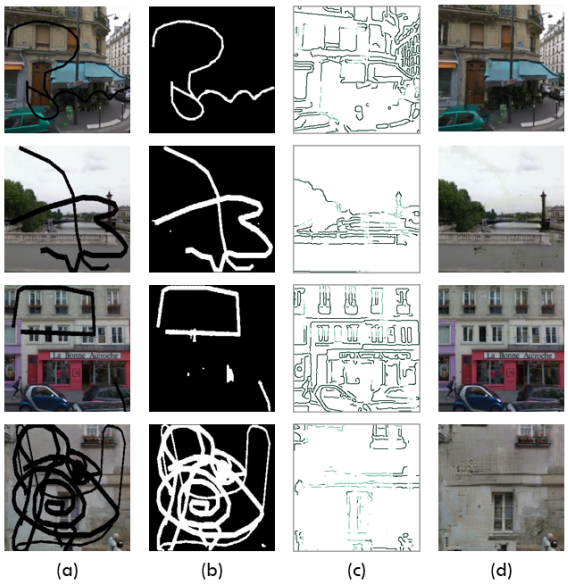

Figure 4. (a) Masked Image from Paris Street View [1] (b) Mask predicted from MPN (c) Edges: black is Canny Edges, while green shows the edges predicted by EPN (d) Final inpainted results

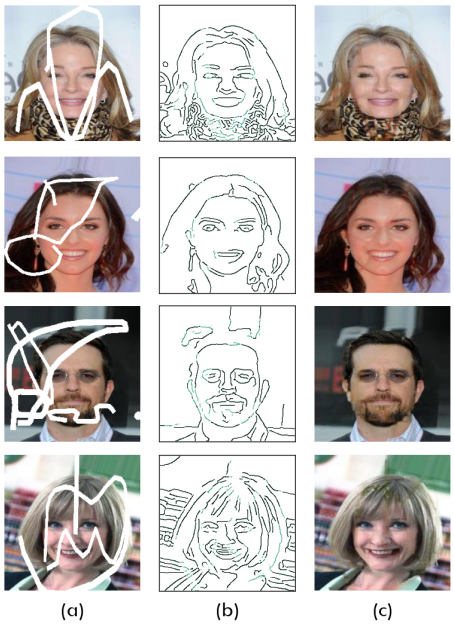

Figure 5. (a) Masked Image from Celeb Faces Attributes (CelebA) Dataset [5] (b) Edges: black is Canny Edges, while green shows the edges predicted by EPN (c) Final inpainted results

	PSV	
	PSNR↑	SSIM↑
EPN	22.2051	0.9569
MPN +	23.1022	0.9750

Table 2. Results from Edge Network being trained alone and jointly with Mask Prediction Network.

look more coherent while keeping the network robust.

References

[1] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012. 4

[2] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. Image inpainting: A review. *Neural Processing Letters*, 51(2):2007–2028, 2020. 1

[3] 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. 1

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 4

[6] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 3

[7] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. 2

[8] 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. 2

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. volume 9351, pages 234–241, 10 2015. 2

[10] Denis Simakov, Yaron Caspi, Eli Shechtman, and Michal Irani. Summarizing visual data using bidirectional similarity. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[11] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European conference on computer vision (ECCV)*, pages 1–17, 2018. 2

[12] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6721–6729, 2017. 2

[13] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 1, 2, 4

[14] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019. 2, 4