Image Editing using Block-wise Refined Training with Annealed Adversarial Counterpart

Anonymous ECCV submission

Paper ID ***

Abstract. We introduce a novel architecture and training scheme, called block-wise refined training, to solve for a variety of image editing tasks, including inpainting, harmonization, composition and super resolution. This general network, which gradually adds residual blocks during training, achieves state-of-the-art in all of these tasks. We also found that leveraging the powerful perceptual similarity loss and decaying the weight of the adversarial loss is critical to the success of training. We conducted a variety of experiments, showing the effectiveness of our method in a broad range of common image editing problems.

Keywords: Image inpainting, harmonization, composition and super resolution.

1 Introduction

people often wish to remove undesired content from their photos in a releastic way. often this involves automatied image inpainting, the task of filling in th e lost part of an image Inpainting is one of the most common operations in image editing, which is used in different scenarios such as fixing random holes, object removal, or image composition. The first scenario, which is mostly seen as repairing a broken photography. This is challenging as it is unstrained, meaning it involves object completion and background fill in. Object removal usually removes unwanted object and fills in with background. Image composition is more specific, which is more like taking part of another image (could be an object, and background) to combine with the original image. This involves harmonization, and inpainting to fill in the unmatched regions. Regardless of the specific oepration involved, the goal is to make the final image look applausible and real in terms of color, textures and contents. Therefore, it is essential to generate realistic textures. If we were able to render photorealistic images to remove unwanted objects or add new objects, it is very easy for us to generate fake images.

Traditional methods tackling image inpainting, harmonization, etc usually by using very separate approaches such as patch propagation (ref), statistics transfering (ref), etc. These methods often invoves texture synthesis techniques which search for similar patches and synthesis the content from the surrounding regions. These methods recover missing regions by copying existing patterns or structures from surround regions. Wilc specify desire search regions to automatically detect better match patches, Barnes proposed the patch match model

which searches nearest neightbor patch to reconstruct missing regions. For image composition, ? often involve cutting and pasting a semantically similar patch from a large database of images, and then blend them togeter. Although these methods are good at propagating high-requency details from the guidance image to the hole, there are often inconsistent regions humans can easily detect. For other alternative methods try to hallucinate missing image regions from surrounding context. Yet, traditional texture synthesis approaches cannot capture global structure by extending texture from surrounding regions. This means that sometimes the inpainted regions will ignore borrow patches from unwanted or inconsistent space. For hallucination it usually is simply statstic methods.

More recently, the powerful capability of deep neural networks has exhibited excellent performance in texture synthesis and image restoration. deep neural networks have been proved to be successful in a variety of computer vision tasks, most notably classification, segmentation, and image synethesis. Also has progressed in image inpaintnng. [] trained an encoder-decoder CNN to predict missing image regions from surrounding pixels. In [], Yang et al. proposed a multi-scale neural patch synthesis method which uses feature extracted from middle layers of convolutional neural network to optimize the process of patch match. Although this generates high-resolution images, the optimization makes it very slow for large images. [] instead uses feed-forward neural network to generate and could produce satisfying results for smaller images and smaller holes. However, it falils to generate higher resolution images. Among the dnn based image inpainting approaches, most stateof the art methodss improve the performance of contexture encoder through capturing more context or texture syntheis. Although these appraochaes can achive good performance by difference network topology and training procedure, they have major limitations: 1. lack of high-frequency details. 2. perceptual discontinuity. For image composition, there is work [][]. However, image inpainting and composition are viewed as separate tasks and no efforts have been made to joitnly solve them.

Given a wide range of applications, we discuss a new approach that produces high-quality inpainting results for the aforementioned discussed tasks. Image editing is commonly desired tools, which can consist of a variety of tasks including object removal, object addition, denoising, super resolution, etc. These steps typically involve steps like image inpainting, image harmonization, segmentation, super-resolution, etc. In particular, image inpainting is the task of fill in the lost part of an image, and the harmonization is to change the color of the object such that it is consistent with the background. This is especially important steps for object removal, image composition, image repairing etc.

Traditional methods tackling image inpainting, harmonization, etc usually by using very separate approaches such as patch propagation, statistics transfering, etc. These method generally encodes prior knowledge of human such that the holes should share similar patches from the background, and the mean and standard deviation encodes the image color information. Recently, deep neural network has been very successful to learn from big data and succeed in a variety of computer vision tasks such as classification, image synthesis, etc. It

has also been successfully applied to several other tasks such as image inpainting and harmonization. Deep learning is good at classification problem such as segmentation. The advantage of deep neural network is that it is able to learn the knowledge from the trained data without handcrafting the features or rules, therefore it is easier to apply to different tasks. It is also better in terms of being creative, able to generate more diverse and dynamic results. However, unlike traditional methods, in addition to use huge amout of date for training, deep neural networks is usually hard to generate results of high resolutions and high quality.

To overcome the limitations above, we discuss a new approach that produces high-quality inpainting results for the aforementioned discussed tasks, refer to as Block-wise Trained Adversarial Model for Image Inpainting. Specifically, we decompose the generator into the generator head followed by multiple residual blocks. We first train the generator head for inpainting, and then gradually add the residual blocks during training one at a time. In this way, we are able to train a very deep network for inpaitning and stabalize the training process. In addition to the block-wise training, we also add a novel loss function which is called perceptual similarity loss, inspired by [?] that measures the distance of two images using perceptual similarity, we use it as an additional loss while training. The third critical point is that it is essential to gradually reduce the weight of the adversarial loss during block-wise training. We observe that this greatly reduces the noise and artifacts of the final result. Moreover, we address the three inpainting scenarios in the unified framework, showing that it is easy to adapt to has multi-output including inpainting, harmonization and composition results, and we can train these tasks jointly.

To evaluate the proposed model, we conduct extensive experiments on different datasets in different settings of inpainting. The dataset we use is COCO, ADE20K and Face. We show that in the general inpainting setting, our model's result is better than the state-of-the-art results, both quantitatively and qualitatively. In the guided inpainting setting, our composition and harmonization result is also superior with state-of-the-art, showing the generalization ability of our model. Finally, we show the power of our approach in real user cases, where people want to remove or add an object into an image. We also perform user-study to show that our approach is better.

In summary, in this paper we present:

1. A high performance network model that is end-to-end, and can be used for different inpainting and composition settings.
2. A novel training approach that progressively increases the depth of the network and stablize the training.
3. Practical use of our approach, which is simple to use yet present state-of-the-art results for inpainting, haromization and guided inpainting.

References