MUSICAL: Multi-Scale Image Contextual Attention Learning for Inpainting

Conference Paper · August 2019

DOI: 10.24963/ijcai.2019/520

4 authors, including:

Wang Ning
Wuhan University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Jingyuan Li
Wuhan University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Bo Du
Wuhan University
165 PUBLICATIONS **2,498** CITATIONS
SEE PROFILE

MUSICAL: Multi-Scale Image Contextual Attention Learning for Inpainting

Ning Wang, **Jingyuan Li**, **Lefei Zhang*** and **Bo Du**
School of Computer Science, Wuhan University
{wang_ning, jingyuanli,zhanglefei,remoteking}@whu.edu.cn

Abstract

We study the task of image inpainting, where an image with missing region is recovered with plausible context. Recent approaches based on deep neural networks have exhibited potential for producing elegant detail and are able to take advantage of background information, which gives texture information about missing region in the image. These methods often perform pixel/patch level replacement on the deep feature maps of missing region and therefore enable the generated content to have similar texture as background region. However, this kind of replacement is a local strategy and often performs poorly when the background information is misleading. To this end, in this study, we propose to use a multi-scale image contextual attention learning (MUSICAL) strategy that helps to flexibly handle richer background information while avoid to misuse of it. However, such strategy may not promising in generating context of reasonable style. To address this issue, both of the style loss and the perceptual loss are introduced into the proposed method to achieve the style consistency of the generated image. Furthermore, we have also noticed that replacing some of the down sampling layers in the baseline network with the stride 1 dilated convolution layers is beneficial for producing sharper and fine-detailed results. Experiments on the Paris Street View, Places, and CelebA datasets indicate the superior performance of our approach compares to the state-of-the-arts.

1 Introduction

Image inpainting is a research hotspot in computer vision and machine learning communities, it refers to restoring or reconstructing images which have missing regions [Guillemot and Le Meur, 2014]. In practice, many inpainting approaches have been proposed in wide application ranges, e.g., the removal of unwanted objects, eye inpainting, identities obfuscation, and shape inpainting [Criminisi *et al.*, 2004; Sun *et al.*, 2018].

*Corresponding author

In general, the current image inpainting methods are designed based on the assumption that the missing area should contains similar patterns of the background region. Prior to the deep learning era, nearly all the methods employ the certain statistics of the remaining image to recover the corrupted region [Levin *et al.*, 2003; Criminisi *et al.*, 2004]. In particular, as the one of the once state-of-the-art methods, the PatchMatch [Barnes *et al.*, 2009] matches and copies the background patches into holes starting from low-resolution to high-resolution or propagating from hole boundaries. While this approach generally produces smooth results, especially in background inpainting tasks, it is limited by the available image statistics and not able to capture high-level semantics or global structure of the image. Furthermore, as the traditional diffusion-based and patch-based methods assume missing patches can be found somewhere in the background regions, they cannot produce novel image contents for complex inpainting regions where involve intricate structures like faces [Yu *et al.*, 2018].

In the recent years, the deep learning based methods have been reported to overcome the limitations above by the support of the large volume of training images [Gao and Grauman, 2017]. In particular, the deep convolutional neural networks (CNN) and generative adversarial networks (GAN) have been introduced to solve the image inpainting task [Iizuka *et al.*, 2017; Yeh *et al.*, 2017]. These deep learning based approaches can be simply divided into two categories including one-stage methods [Pathak *et al.*, 2016; Iizuka *et al.*, 2017] and two-stage methods [Yang *et al.*, 2017; Yu *et al.*, 2018]. Due to the difficulty of using one latent code to represent high-dimensional distribution of the complex real scene, two-stage methods are proposed to do content generation and texture refinement separately, but they are quite time consuming [Xiao *et al.*, 2019; Liu *et al.*, 2017]. What's more, two-stage methods still generate some boundary artifacts, distorted structures and blurred textures inconsistent with surrounding areas like one-stage methods, since the neural patch is a mixture of content and style, copying them from known region into target region in later stage will introduce change to the originally generated content. More detailed literature review of the existing image inpainting methods can refer to the next section.

In this paper, we propose a novel one-stage image inpainting model, i.e., the multi-scale image contextual attention

learning (MUSICAL). Technically, our model adopts the U-Net [Ronneberger *et al.*, 2015] architecture as the baseline to propagate the global and local style coherency and detailed texture information to the missing region. In the MUSICAL algorithm, we develop a special multi-scale attention module by which the feature maps of each scale attention output are merged by the structure of the Squeeze-and-excitation net [Hu *et al.*, 2018], in this way, we can better capture the background information in multiple scales and produce content with elegant details. Then the output images are sent into two networks, including (1) the VGG16 for calculating style loss and perceptual loss, and (2) the discriminator whose structure is the DenseNet [Huang *et al.*, 2017; Liu *et al.*, 2019]. This will help to generate details in consistent with the global style. Furthermore, we replace the bottom layer of the U-Net with the stride 1 dilated convolution without downsampling to make our results sharper. Experiments on three standard datasets (i.e., Paris Street View, Places and CelebA) demonstrate that the proposed approach generates higher quality results compare to the existing competitors. The main contributions of this paper are summarized as follows:

- We develop a novel multi-scale attention module into the proposed MUSICAL architecture. By merging the feature maps produced by attention module of different matching patch sizes, we can capture information in multiple scales and flexibly take advantage of background information to balance the needs of different styles of images.
- We introduce the style loss, perceptual loss, and adversarial loss to construct the proposed loss function, in which the style loss and perceptual loss are conducive to generating the consistent style, and the adversarial loss can help the network to generate sharper results with better details.
- Experiments on the Paris Street View, Places and CelebA datasets demonstrate the superiority of our approach compares to the existing state-of-the-art approaches.

The rest of the paper is organized as follows: section 2 reviews some related works, then section 3 introduces our proposed MUSICAL algorithm in detail. After that, the experimental results and analysis on three public available datasets are reported and discussed in section 4, followed by the conclusions in section 5.

2 Related Works

2.1 Image Inpainting

A variety of different approaches have been proposed for the image inpainting tasks, and these works can be summarized into the following two groups.

Traditional Diffusion and Patch Based Approaches

These methods usually introduce variational algorithms based on the patch similarities to fill target regions with local image information propagating from background regions [Levin *et al.*, 2003; Ding *et al.*, 2019]. Among which, the diffusion based approaches can only fill small or narrow holes, while the patch based methods may be performed on more complicated image inpainting scenes and can fill large holes in natural images. Specifically, a fast nearest neighbor field algorithm called PatchMatch [Barnes *et al.*, 2009] has shown significant practical values for image editing applications including inpainting. It is worth to note that these traditional methods may work well for stationary textures but are not effective to fill in holes on complicated structures, since they mainly depend on low-level features. Furthermore, they are unable to generate novel objects which not exist in the source image.

Learning-based Approach

In the recent years, the deep learning based approaches have appeared as a remarkable exemplar for image inpainting. Context Encoder [Pathak *et al.*, 2016] attempts to inpaint the center region (64×64) of 128×128 images, which is the first parametric inpainting algorithm that is able to give reasonable results for semantic hole-filling (i.e. large missing regions). [Iizuka *et al.*, 2017] introduces global and local context discriminators as adversarial losses to improve the network to be more consistent. Furthermore, [Yang *et al.*, 2017] and [Snelgrove, 2017], regard the image inpainting task as an optimization problem. For example, [Yang *et al.*, 2017] proposes a multi-scale neural patch synthesis (MNPS) approach that matches and adapts patches with the most similar mid-layer feature correlations of a deep classification network. More recently, the Shift-Net [Yan *et al.*, 2018] introduces a special shift-connection layer for the U-Net architecture [Ronneberger *et al.*, 2015] to fill missing regions of any shape with sharp structures and fine-detailed textures. Compared with MNPS, the Shift-Net can uncover better results and takes much less time in the training procedure.

2.2 Attention Modeling

In the case of limited computing resources, the attention mechanism is a resource allocation scheme that solves the problem of information overload, thus it may allocate the computing resources to more important tasks. Attention can be incorporated as an operator following one or more layers representing higher-level abstractions for adaptation between modalities. Researches on the spatial attention in deep convolutional neural networks has emerged a lot in the recent years.

[Jaderberg *et al.*, 2015] introduces a parametric spatial attention module called spatial transformer network (STN) for neural networks. The STN model can predict parameters of global affine transformation to warp features with a localization module, but not suitable for modeling patch-wise attention due to its global transformation. By introducing an appearance flow, [Zhou *et al.*, 2016] predicts offset vectors that specify which pixels in the input image should be moved to reconstruct the target region for novel view synthesis. This method is effective for matching related views of the same objects but is not effective in predicting a flow field from the background region to the target region. More recently, [Yu *et al.*, 2018] proposes a contextual attention layer to explicitly attend on related feature patches at distant spatial locations, which also has spatial propagation layer to encourage

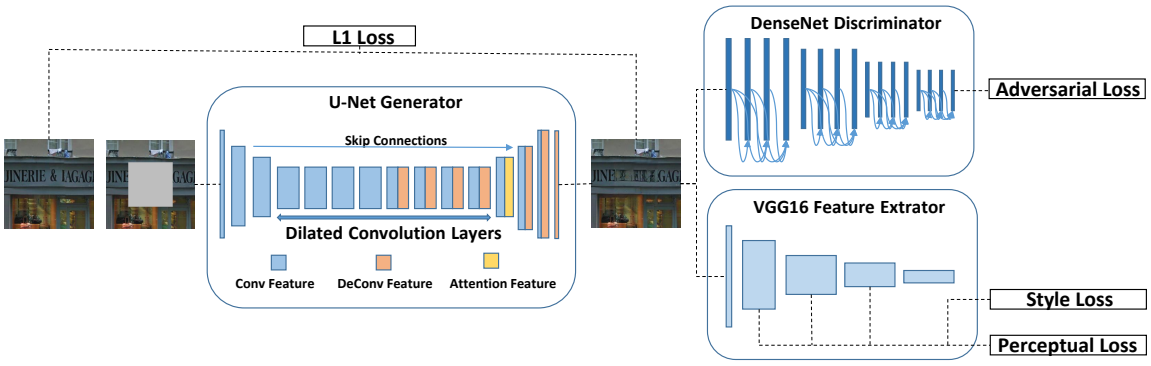

Figure 1: Overall model architecture

spatial coherency of attention. However, the original contextual attention uses a fixed patch size when calculating patch similarity and doing patch swap. Such strategy may perform poorly when the background is misleading or lacks similar content and is therefore unable to filexibly handle different backgrounds. Unlike [Yu *et al.*, 2018], we use a multi-scale attention module to perform feature map swaping in different scales. By doing so, we could have better confidence that some of the attention layers have the expected feature map and therefore eliminate the problem of misusing of information. In addition, our model can solve various scenes and balances the style and detail level, but Shift-Net may perform not well in some datasets.

3 Proposed Approach

The image inpainting task aims to generate the plausible content given masked input I_m. The generated image should not only exhibit global and local style coherency but also detailed texture that is consistent with foreground. It's easy to notice that images from different scenes have various requirements on style and detail level. For example, the Paris Street View dataset contains large amount of latent structure information like the size and location of windows and doors, under which condition the model should take more care of global style and should not try to generate too many details. While for another datasets, the style consistency may be less important while generating the detailed texture may help to improve visual effect more.

For the one-stage inpainting methods, the U-Net [Akeret *et al.*, 2017] has been widely used as baseline network as skip connections can preserve low level informations and enable the rest of network to focus on recovering masked area. In this study, we also use the U-Net as our baseline but with some modifications. By merging the output feature maps from attention modules of different matching patch sizes, our model is able to generate content with fine details while the generated images are also consistent in style. One of the most significant benefit of our model is that it can flexibly take the advantage of background information without modifying the model or changing any hyperparameter. In the following subsections, a novel multi-scale attention module which helps to better make use of the foreground information and our proposed loss function which balanced generated details and global style coherency will be introduced, respectibely.

3.1 Overview of the MUSICAL Algorithm

Our model follows a one-stage and end-to-end architecture, which indicates that it works much faster than the two-stage methods. In detail, as mentioned above, we use the U-Net as our baseline network. Furthermore, we further extend the architecture of U-Net. A series of downsampling layers are contained on U-Net architecture and the size of feature map is reduced to 1×1 in the inner most layers, which only contains a trival amount of information. To preserve more detail information, we replace the inner downsampling layers with stride 1 dilated convolution [Yu and Koltun, 2015] layers, which have larger receptive field without losing too much information. To avoid attention module using misleading or even incorrect information from background feature map, the input and masked area should be large enough. So we place our multi-scale attention layers before the third last deconvolution layer, where the size of feature map is 64×64.

3.2 Multi-scale Attention Module

Image Inpainting task requires GANs to generate style consistent images given foreground. A promising way to ensure style consistency is to make use of background information(unmasked area) such as the Patch-match [Barnes *et al.*, 2009]. Previous works have shown that doing patch/pixel match on deep feature maps helped to improve quality of generated images. However, an appropriate size for patch match is hard to determine as the requirements on detail and style level various from image to image. In general, larger patch size helps ensure style consistency while smaller patch size is more flexible on using background feature map. Patch matching on a single fixed scale seriously limits the capability to fit the model into different scene. To this end, we propose a novel multi-scale attention module that helps to make use of background content flexibly based on the overall image style.

Given an input feature map ϕ_{in}, we firstly replace foreground feature map using attention machnism. Instead of using fixed patch size and propagate size, we choose to use two different patch size and therefore two feature maps ϕ_{att11} and ϕ_{att33} are generated. For each attention map, we use similar strategy to calculate scores as [Yu *et al.*, 2018], the calculation of attention score could be implemented as convolution

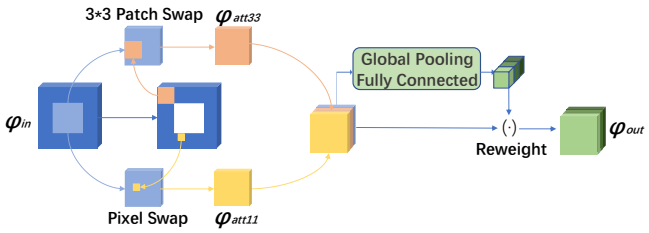

Figure 2: Multi-scale Attention Module: In this attention module, the input feature map is sent to two different attention modules, on the top there is a 3×3 attention module and the bottom is a pixel-wise one, the stride and propagation sizes are the same for two modules. The output feature maps are then reweighted by a Squeeze-and-Excitation module. In the end, the channel number is reduced by the Pixel-Wise Convolution.

calculation.

$$s_{x,y,x',y'} = \langle \frac{f_{x,y}}{||f_{x,y}||}, \frac{b_{x',y'}}{||b_{x',y'}||} \rangle \tag{1}$$

To calculate the weight of each patch, we use softmax on the channel of score map, and get softmax score s^*. Since a shift in foreground patch is likely corresponding to an equal shift in background patch for attention, we adopt a left-right propagation followed by a top-down propagation with kernel size of k. Then we propagate the score to better merge patchs.

$$s'_{x,y,x',y'} = \sum_{i,j \in \{-k,\ldots,k\}} s^*_{x+i,y+j,x'+i,y'+j} \tag{2}$$

Finally, deconvolution operation is used to recover the attention feature map with. By doing so, our model could capture information in multiple scales and fit into various background better. We then concatenate the generated feature maps and original feature maps, denoted by $\langle \phi_{in}, \phi_{att11}, \phi_{att33} \rangle$. To decide which level of detail is the most important one on current image, the feature maps are then fed into a squeeze-and-excitation [Hu *et al.*, 2018] module to reweight different channels. The squeeze-and-excitation function is denoted as $f_{SE}()$ in our paper. The SE module firstly computes the average pooling value of the feature map, and then put it into a fully connected neural network to calculate the weight of each channel of the original feature map, and add weight to it. The output of SE module could be expressed by $f_{SE}(\langle \phi_{in}, \phi_{att11}, \phi_{att33} \rangle)$. Note that Squeeze-and-excitation module above could not be replaced by convolutional computation as the convolution kernels are a set of fixed parameters and lack the ability to add weight to each channel variously based on background information. In the end of the module, we use pixel-wise convolution operation to finally merge all feature maps and reduce the channel numbers to the original channel number. As the output channel number is the same as input, it's easy for our proposed module to be added to any other inpainting model. The final output of the module could be denoted as:

$$\phi_{out} = f_{Conv}(f_{SE}(\langle \phi_{in}, \phi_{att11}, \phi_{att33} \rangle)) \tag{3}$$

3.3 Loss Function

Similar to the design of our model, the style consistency and detail level are also taken into the consideration of our loss function. In general, our loss function consists of two parts, i.e., the perceptual and style losses for generating plausible style and adversarial loss for generating sharper and detailed content, respectively. For perceptual loss ($L_{perceptual}$), we put the generated image into a $VGG16$ feature extractor and compare the feature maps from $pool1$, $pool2$ and $pool3$ with the ones corresponding to the ground truth image. In our model, we use the perceptual loss to measure the similarity between the high level structures. In Equation 4, H, W, C refer to the height, weight and channels number for a feature map, respectively. And N is the number of feature maps generated by the VGG16 feature extractor.

$$L_{perceptual} = \sum_{i=1}^{N} \frac{1}{HWC} |\phi^{gt}_{pool_i} - \phi^{pred}_{pool_i}|_1 \tag{4}$$

Perceptual loss helps to capture high level structure but it still lacks the ability to preserve style consistency. To address this issue, we further employ the style loss (L_{style}) as a part of our loss function. With the help of the style loss, our model could learn color and overall style information from background.

$$L_{style} = \sum_{i=1}^{N} \frac{1}{C*C} \Big| \frac{1}{HWC} (\phi^{style_{gt}}_{pool_i} - \phi^{style_{pred}}_{pool_i}) \Big|_1 \tag{5}$$

$$\phi^{style}_{pool_i} = \phi_{pool_i} \phi^T_{pool_i} \tag{6}$$

With the loss functions discussed above, our model is ready to generate plausible content where details are less important than structure. However, the generated area tends to be blurry when our model tries to learn more details and we notice that having a discriminator is still necessary for generating fine details. In our model, we choose to use a pretrained DenseNet121 [Huang *et al.*, 2017] as our discriminator for its relatively smaller size and high accuracy in recognizing objects. With the help of the adversarial loss (L_{adv}), the level of sharpness and detail becomes controllable by using different weights of adversarial loss. The total variation loss (L_{tv}) [Rudin *et al.*, 1992] is originally introduced to address the checkboard artifact brought from style loss. In our model, we have also employed it to enhance the smoothness of the generated content. In summary, the overall loss function of the proposed MUSICAL algorithm is as follows:

$$\begin{aligned} L_{total} = \lambda_{sty} L_{style} + \lambda_{perc} L_{perceptual} \\ + \lambda_{adv} L_{adv} + \lambda_{tv} L_{tv} + \lambda_{l1} L_{l1} \end{aligned} \tag{7}$$

4 Experiments

4.1 Experimental Settings

Dataset

In this section, we conduct experiments to investigate the effectiveness of our MUSICAL algorithm on three public image datasets including the Paris Street View [Doersch *et al.*,

(a) (b) (c) (d) (e)

Figure 3: Qualitative comparisons on Paris Street View dataset. From left to right are: (a) Input, (b) CE, (c) GLC, (d) SN, (e) Ours.

(a) (b) (c) (d) (e)

Figure 4: Qualitative comparisons on Places dataset. From left to right are: (a) Input, (b) CE, (c) GLC, (d) SN, (e) Ours.

2012], Places [Zhou *et al.*, 2018], and CelebA [Liu *et al.*, 2015]. The Paris Street View contains 14,900 training images and 100 test images. The Places dataset is the canyon scene selected from Places365-Standard dataset, and this category has 5,000 training images, 900 test images and 100 validation images. In our experiment, we use the training set for training and the validation set for testing. And the CelebA dataset contains 162,770 training images, 19,867 validation images and 19,962 test images. We use both of the training set and validation set for training, and use the test set for testing. We use the same dataset setting for all experiments, including our experiment and comparison experiments.

Training Details

For both Paris Street View and Places, we resize each training image to let its minimal length/width be 350, and randomly crop a subimage of size 256×256 as input to our model. As for CelebA, we resize each training image to let its minimal length/width be 256, and crop a subimage of size 256×256 at the center as input to our model. The size of mask is 128×128 for each image, and the mask is located in the center of the image. We train the model with a batch size of 5 for each epoch. For all the datasets, the tradeoff parameters are set as $\lambda_{sty} = 250$, $\lambda_{perc} = 0.07$, $\lambda_{tv} = 0.001$ and $\lambda_{l1} = 100$. While the λ_{adv} is different in these datasets. Specifically, for the CelebA dataset, we have set $\lambda_{adv} = 0.0$, while we set $\lambda_{adv} = 0.3$ for the other two datasets. All the experiments are conducted with the Python on Ubuntu 17.10 system, with i7-6800K 3.40GHz CPU and 12G NVIDIA Titan Xp GPU.

4.2 Experimental Results

We compare the proposed MUSICAL algorithm with the following three state-of-the-art methods:

- CE: Context Encoder [Pathak *et al.*, 2016]
- GLC: Globally and Locally Consistent Image Completion [Iizuka *et al.*, 2017]
- SN: Shift-Net [Yan *et al.*, 2018]

(a) (b) (c) (d) (e)

Figure 5: Qualitative comparisons on CelebA dataset. From left to right are: (a) Input, (b) CE, (c) GLC, (d) SN, (e) Ours.

Qualitative Comparisons

Figure 3 shows the comparisons of our method with the three state-of-the-art approaches on Paris Street View. The images are all shown at the same resolution (256×256) except CE (128×128). Context encoder is effective in semantic inpainting, but the results seem blurry and detail-missing due to the effect of bottleneck. Compared with CE, the proposed method can handle much larger images and the synthesized contents are much sharper. GLC is effective in understanding the context of entire image, but the results tend to be less realistic or recognizable. Shift-Net enable the generated content to have similar texture as background region. However, it will performs poorly when the background information is misleading. Our inpainting results have significantly fewer artifacts than these methods and in particular better than the state-of-the-art methods most of the time for large holes. In comparison to these methods, our proposed MUSICAL algorithm is able to generate more visual-pleasing and more elegant results.

In addition, we have also evaluated our method on the

Method	SSIM	PSNR	Mean l_1 Loss
CE	0.7879	22.87	2.943%
GLC	0.7925	23.01	2.771%
SN	0.8292	23.85	2.482%
Our Method	0.8428	24.42	2.264%

Table 1: Numerical comparison on Paris Street View dataset.

Method	SSIM	PSNR	Mean l_1 Loss
CE	0.7794	22.30	3.157%
GLC	0.7886	20.74	3.551%
SN	0.7944	21.14	3.360%
Our Method	0.8025	21.84	3.063%

Table 2: Numerical comparison on Places dataset.

Places dataset (see Figure 4) and CelebA dataset (see Figure 5). It is observed that our MUSICAL algorithm performs favorably in generating fine-detailed, semantically plausible, and realistic images.

Quantitative Comparisons

We have also compared our model quantitatively with the comparison methods on three datasets. Three quality measurements that we adopted are the structural similarity index (SSIM), peak signal-to-noise ratio (PSNR) and mean l_1 loss, respectively [Liao *et al.*, 2018]. Note that the results of the CE are based on the inputs and outputs of 128×128 images since the codes only accept 128×128 images as inputs.

Table 1,Table 2 and Table 3 show numerical comparison results among our approach, CE, SN and GLC on Paris Street View, Places, and CelebA datasets, respectively. As shown in Table 1, our method produces decent results with best SSIM, PSNR and mean l_1 loss on Paris Street View dataset. On the Places dataset, our approach has better SSIM and mean l_1 loss, but the PSNR is lower than CE. As for the CelebA faces dataset, we yield best SSIM, PSNR and mean l_1 loss among these methods. As reported above, our MUSICAL algorithm achieves the best numerical performance on the Paris Street View, Places, and CelebA datasets.

Internal Analysis of MUSICAL Algorithm

As highlighted before, the main contributions of our MUSICAL algorithm are the multi-scale attention module and the combination of different losses. To clearly present the effectiveness of these operations, the following experiment settings are applied on Paris Street View dataset.

Experiment 1: maintaining the overall model architecture, but eliminating the adversarial loss in loss function.

Experiment 2: keeping other settings but replacing our multi-scale attention module with a single-scale attention module, i.e. using single fixed patch size and propagate size.

Method	SSIM	PSNR	Mean l_1 Loss
CE	0.8631	24.82	2.208%
GLC	0.8776	24.02	2.360%
SN	0.8796	25.13	1.958%
Our Method	0.9008	26.64	1.629%

Table 3: Numerical comparison on CelebA dataset.

(a) (b) (c) (d) (e)

Figure 6: Qualitative comparisons of Internal analysis. From left to right are: (a) Input, (b) Ground Truth, (c) Experiment 1, (d) Experiment 2, (e) Ours.

Experiment 1 (see Figure 6(c)) shows that L_{adv} is important for generating sharper images. Compared to our algorithm (see Figure 6(e)), the results without L_{adv} exhibit more artifacts and distortions. For our MUSICAL algorithm, L_{adv} is introduced for inpainting with clear content as L_{style} and $L_{perceptual}$ mainly help us to produce structure of images. Thus, the L_{adv} helps to generate sharper results.

By comparing Experiment 2 (see Figure 6(d)) with our algorithm (see Figure 6(e)), we can notice that, we get general structure with single-scale attention module but quality inferior to the multi-scale result. Given an input feature map ϕ_{in}, our method uses two different patch sizes and generates two feature maps ϕ_{att11} and ϕ_{att33}, while Experiment 2 only uses single patch size and generates one feature map ϕ_{att33}. Therefore, multi-scale attention module can get more comprehensive information from ϕ_{in} than single-scale. The difference between Figure 6(d) and Figure 6(e) indicates the fact that multi-scale attention module acts as a refinement and enhancement role in recovering clear and fine details.

5 Conclusion

In this paper, we propose the MUSICAL algorithm for image inpainting. The novel points of our model lie in that we develop a multi-scale attention module and introduce several losses (including style loss, perceptual loss and adversarial loss) to ensure the style consistency and fine-detailed content. Various image inpainting experiments show that the proposed MUSICAL algorithm generates sharp and fine-detailed images, and achieves the state-of-the-art performance across different datasets. In the future research, the proposed MUSICAL algorithm can also be generalized to the similar image restoration tasks including the image denoising, conditional image generation, and image editing.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61771349 and 61822113, and in part by the Key R & D Program of China under Grant 2018YFA0605501 and the Natural Science Foundation of Hubei Province under Grants 2018CFA050, 2018CFB432.

References

[Akeret *et al.*, 2017] Joel Akeret, Chihway Chang, Aurelien Lucchi, and Alexandre Refregier. Radio frequency interference mitigation using deep convolutional neural networks. *Astronomy and Comput.*, 18:35–39, 2017.

[Barnes *et al.*, 2009] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM TOG*, 28(3):24, 2009.

[Criminisi *et al.*, 2004] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE TIP*, 13(9):1200–1212, 2004.

[Ding *et al.*, 2019] Ding Ding, Sundaresh Ram, and Jeffrey J Rodríguez. Image inpainting using nonlocal texture matching and nonlinear filtering. *IEEE TIP*, 28(4):1705–1719, 2019.

[Doersch *et al.*, 2012] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? *ACM TOG*, 31(4):101, 2012.

[Gao and Grauman, 2017] Ruohan Gao and Kristen Grauman. On-demand learning for deep image restoration. In *Proc. CVPR*, pages 1086–1095, 2017.

[Guillemot and Le Meur, 2014] Christine Guillemot and Olivier Le Meur. Image inpainting: Overview and recent advances. *IEEE SPM*, 31(1):127–144, 2014.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. CVPR*, pages 7132–7141, 2018.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proc. CVPR*, pages 4700–4708, 2017.

[Iizuka *et al.*, 2017] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM TOG*, 36(4):107, 2017.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Proc. NIPS*, pages 2017–2025, 2015.

[Levin *et al.*, 2003] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In *Proc. ICCV*, pages 305–312, 2003.

[Liao *et al.*, 2018] Liang Liao, Ruimin Hu, Jing Xiao, and Zhongyuan Wang. Edge-aware context encoder for image inpainting. In *Proc. ICASSP*, pages 3156–3160, 2018.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. ICCV*, pages 3730–3738, 2015.

[Liu *et al.*, 2017] Weiwei Liu, Ivor W. Tsang, and Klaus-Robert Müller. An easy-to-hard learning paradigm for multiple classes and multiple labels. *JMLR*, 18:94:1–94:38, 2017.

[Liu *et al.*, 2019] Weiwei Liu, Donna Xu, Ivor W. Tsang, and Wenjie Zhang. Metric learning for multi-output tasks. *IEEE TPAMI*, 41(2):408–422, 2019.

[Pathak *et al.*, 2016] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proc. CVPR*, pages 2536–2544, 2016.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, pages 234–241, 2015.

[Rudin *et al.*, 1992] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1-4):259–268, 1992.

[Snelgrove, 2017] Xavier Snelgrove. High-resolution multi-scale neural texture synthesis. In *Proc. SIGGRAPH Asia Technical Briefs*, page 13, 2017.

[Sun *et al.*, 2018] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proc. CVPR*, pages 5050–5059, 2018.

[Xiao *et al.*, 2019] Jing Xiao, Liang Liao, Qiegen Liu, and Ruimin Hu. Cisi-net: Explicit latent content inference and imitated style rendering for image inpainting. In *Proc. AAAI*, pages 1–7, 2019.

[Yan *et al.*, 2018] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proc. ECCV*, pages 3–19, 2018.

[Yang *et al.*, 2017] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proc. CVPR*, pages 6721–6729, 2017.

[Yeh *et al.*, 2017] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proc. CVPR*, pages 5485–5493, 2017.

[Yu and Koltun, 2015] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[Yu *et al.*, 2018] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proc. CVPR*, pages 5505–5514, 2018.

[Zhou *et al.*, 2016] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A. Efros. View synthesis by appearance flow. In *Proc. ECCV*, pages 286–301, 2016.

[Zhou *et al.*, 2018] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 40(6):1452–1464, 2018.