Spatially adaptive multi-scale contextual attention for image inpainting

Xueting Wang1,2 · Yiyan Chen1 · Toshihiko Yamasaki1

Received: 14 December 2020 / Revised: 29 March 2021 / Accepted: 25 January 2022 / Published online: 11 April 2022
© The Author(s) 2022

Abstract
Image inpainting is the task to fill missing regions of an image. Recently, researchers have achieved a great performance by using convolutional neural networks (CNNs) with the conventional patch-matching method. Existing methods compute the attention scores, which are based on the similarity of patches between the known and missing regions. Considering that patches at different spatial positions can convey different levels of detail, we propose a spatially adaptive multi-scale attention score that uses the patches of different scales to compute scores for each pixel at different positions. Through experiments on the Paris Street View and Places datasets, our proposal shows slight improvement compared with some related methods on the quantitative evaluation metrics commonly used in the existing methods. Moreover, we found that these quantitative metrics are not appropriate enough considering the subjective impressions of the generated images. Therefore, we conducted subjective evaluation through user study for comparison, which shows that our proposal has superiority of performance generating much more detailed and subjectively plausible images.

Keywords Image inpainting · Spatially adaptive · Contextual attention · Multi-scale attention

1 Introduction
Image inpainting is a task to complete the image by generating plausible contents to replace the missing regions as the examples given in Fig. 1. This technique can be applied to many different fields, such as photo editing and image restoration.

Springer
Fig. 1 Examples of our image inpainting. For each image sample, the ground truth is shown on the left, the one with missing region in the middle, and the rightmost is the inpainting result by our proposal

Conventional methods assume that the missing regions can be filled by content from the known regions. Some approaches focused on using the surrounding known regions to replace the missing regions [3, 17]. Others [9, 10, 27] proposed filling the missing regions with similar patches from other images, sometimes from other databases. Conventional methods can achieve good results for some specific images having simple/repetitive structures. However, they cannot generate semantic and natural contents for complex images without a higher-level understanding of image contents.

Recently, deep learning techniques have been applied to image inpainting. CNNs have been used to extract high-level semantic features of images to achieve great performance on many image processing tasks. Generative adversarial networks (GAN) [7] were introduced to generate natural images and achieved great performance. Most of methods have used the encoder and decoder as the generator where the encoder abstracts the image to a high-level feature, from which the decoder generates the image.

These methods can be divided into two categories: one-stage methods [25, 31, 37, 38] using one generator, and two-stage methods [24, 39, 40] using two successive generators. One-stage methods, such as Yan et al. [37] have tried combining the conventional methods with deep learning by comparing patches from the known regions and missing region in feature map and shifted the patches from the known region to the missing region. For two-stage methods, the output of the first generator becomes input into the second generator. They assume that the output of the first stage can be a suitable prior for the second stage. Thus, they divide the process of generation into coarse generation and textual refinement separately to restore the complex scene. However, these methods are very time consuming and tend to generate blurred images and artifacts as one-stage methods do. Therefore, we focus on the challenge of achieving semantic and natural image inpainting with the balance of coarse and textual generation using efficient one-stage architecture in this paper.

To better generate the coarse style and detailed textual information for various scenes, attention mechanism has also been applied to image inpainting. Yu et al. [39] proposed to use contextual attention scores to attend to feature patches in different spatial areas. Because the patch size is fixed, the performance could be limited when the known region is misleading or lacks similar content. To flexibly handle different known regions, Wang et al. [31] proposed a multi-scale attention module to use different scales of patches to compute the attention scores. The obtained feature maps in different scales provide more confidence to avoid misusing of known regions information. However, they simply combined the multi-scale attention scores without considering the spatial differences. Thus, it
may fail to generate visually good images with enough details considering that patches in different spatial positions can convey different levels of information.

In this paper, we propose a spatially adaptive multi-scale contextual attention based one-stage image inpainting method. As shown in Fig. 2, the basic framework of our model consists of a generator based on the U-Net \[28\] architecture, a VGG16 network, and a discriminator with the structure of DenseNet \[13\] to balance the coarse style and textual details similar to \[31\]. Especially, in the generator, we propose a spatially adaptive multi-scale contextual attention module, where the patch-based attention scores are computed using different scales of patches and obtained multi-scale attention scores are combined with spatially aware reweighting scores considering that patches in different spatial positions can convey different levels of information. Two convolutional layers are used to predict the spatially adaptive reweighting scores. Compared to \[31\], which reweighted multi-scale attentions without considering the spatial differences, our proposal automatically considers the scale to use for each position. Therefore, our proposal generates much more detailed and plausible images.

Moreover, for the performance evaluation, we found that the quantitative metrics commonly used in the existing methods are not appropriate enough considering the visual impression of generated results is important for the inpainting target, as detailed discussed in Section 5.2. For example, during the experiments, Shift-Net (SN) \[37\] achieved the best performance on quantitative metrics such as structural similarity (SSIM) and peak signal-to-noise ratio (PSNR), while the generated images were blurred and clearly inferior. Therefore, we conducted the user study to acquire qualitative evaluation on the generated images of different methods. The user study results using Places and Paris Street View datasets show that our proposed method is significantly subjectively better than the previous approaches although we can observe only slight performance improvements in objective measures such as SSIM and PSNR. Although technical contributions compared to \[31\] may seem small, we experimentally demonstrate that our spatially adaptive multi-scale contextual attention significantly contributes to higher-quality image inpainting. We also show the inpainting results for CelebA-HQ dataset.

The technical contributions of this paper can be summarized as follows.

Fig. 2 The overview of our proposed framework. It consists of a generator based on the U-Net \[28\] architecture, a VGG16 network, and a discriminator with the structure of DenseNet \[13\] to balance the coarse style and textual details. In the generator, the proposed spatially adaptive multi-scale attention module is added to generate the spatial-adaptive multi-scale attention scores based on multiple patch sizes considering their spatial differences.
We propose spatially adaptive multi-scale contextual attention module into the inpainting framework. By combining multi-scale attentions variably for each pixel in different spatial positions, different levels of information of multiple scale patches in different spatial positions and more details can be better represented.

Experiments on Places and Paris Street View datasets show that our proposal has the superiority of performance on generating images that are much more detailed and plausible compared to the state-of-the-art methods, which is confirmed by user study.

2 Related work

2.1 Image inpainting

Image inpainting has been studied for several decades. Existing methods can be divided into non-semantic and semantic methods.

Non-semantic methods are based on the assumption that the information of the missing region can be found in the known region. Therefore, they search the known region either from the surrounding of the same image [3, 17] or from other similar images [9, 10, 27]. However, these kinds of methods cannot generate semantical and natural contents and lead to poor performance. Therefore, the proposal aims at semantic and natural image inpainting by a learning based approach.

Deep learning based models (e.g. CNNs) have been applied to many different tasks of computer vision [11, 19, 26, 29, 33] to extract high-level semantic contextual information of images. Gao et al. [6] proposed a specially-designed cloud-edge distributed framework for salient object detection, which can effectively capture the global contextual features while preserving the local detailed features. For image inpainting, Pathak et al. [25] introduced an encoder-decoder pipeline. The encoder took images having missing regions as input and generated a contextual feature representations. The decoder then generated the content of the missing regions using the contextual feature representations as inputs. Yan et al. [37] combined conventional methods with deep learning techniques. Their Shift-Net comprised a U-Net architecture enhanced with a defined shift-connection layer. This shift-connection layer was used to compare the latent feature of missing regions with one of the known region and then shifted the most similar one to the missing regions. Some methods focused on distinguishing the valid pixels and invalid pixels. Liu et al. [20] proposed a partial convolution such that given a binary mask, their convolutional results only depended on the non-hole regions at each layer. To implement it, they designed a mechanism to automatically update the mask, which could then easily be integrated into any deep learning framework. Yu et al. [40] claimed that it was not proper to directly categorize each pixel as valid or invalid. They instead classified pixels as valid, masked, neurons, or synthesized. Instead of using a rule-based mask update, they allowed the neural network to learn the mask automatically. This dynamic feature gating mechanism utilized the information adaptively for the different types of pixels.

Some other methods focused on using more input information to enhance the inpainting quality. Xie et al. [35] also considered to distinguish the valid pixels and holes. They proposed a learnable attention map module for mask-updating. Nazeri et al. [24] proposed to use the edge generator first. This edge generator completed the missing edge far more easily than using complex RGB information. The input image with missing regions was then input
into the second generator having the completed edge. Xiong et al. [36] proposed a similar idea to predict the foreground contour first and used the contour information to improve the inpainting performance. Zeng et al. [41] proposed a pyramid-context encoder and transferred the learned attention from deep to shallow in a pyramid way.

2.2 Contextual attention

To better balance the coarse style and detailed textual information for various scenes, the proposed method focus on applying attention mechanism to image inpainting.

Recently, attention mechanism has attracted a lot of attention from researchers. The attention mechanism has achieved great performance in many fields, especially in natural language processing [4, 30]. Apart from that, the attention mechanism has also been applied to computer vision tasks [12, 15, 34]. Jaderberg et al. [15] proposed a parametric spatial attention module called spatial transformer network (STN) for classification tasks. Hu et al. [12] proposed a squeeze-and-excitation module, which was added after the convolutional layer to compute the channel-wise attention score. Other methods computed the spatial attention so that the neural network could consider the relationship between two pixels that were far away from one another. Attention mechanism has also been applied in video object segmentation [21] and the co-attention simaese network shows the effectiveness especially for unsupervised learning situations [22, 23].

This attention mechanism was also applied to image inpainting. Yu et al. [39] introduced a contextual attention layer to utilize the attention mechanism. Their contextual attention layer computed the similarity of known and missing patches and then obtained the attention scores between pixels. This contextual attention layer was implemented via convolution. Therefore, it was differentiable and fully-convolutional. However, the existing work used the fixed size of patch to compare the similarity of pixels. When the known patch contains some errors, this kind of method could perform poorly. Therefore, Wang et al. [31] proposed a multi-scale image contextual attention method. Their method computed attention scores using multi-scale patches. In this way, their method flexibly handled richer background information instead of being misled by it. In addition to square masked and small proportion damaged images, Wang et al. [32] proposed a multistage attention module by adding two pixelwise attention layers in the decoding stage in a coarse to-fine manner specifically for images with large proportions of irregular defects.

However, these method simply combined the multi-scale attention scores without considering the spatial differences while patches in different spatial positions can convey different levels of information. Thus, our proposal provide a spatially adaptive multi-scale contextual attention module to generate visually good images with enough details considering the spatial differences.

3 Proposed method

Image inpainting focuses on generating the plausible images. A plausible generated region is expected to be consistent with the other non-missing (known) regions. However, each known region usually carries different levels of importance in terms of plausibility. Additionally, some regions may convey a large amount of details whereas others may lack many details. Wang et al. [31] proposed multi-scale contextual attention to flexibly use the information of the known regions from multi-scale patches. However, they combined multi-scale
attention ignoring the spatial differences. Thus, in this paper, we propose a spatially adaptive multi-scale contextual attention based image inpainting method.

3.1 Framework of the one-stage architecture

In this paper, we propose a spatially adaptive multi-scale contextual attention based one-stage image inpainting framework as shown in Fig. 2. The baseline of the framework consists of a generator based on the U-Net [28] architecture, a VGG network, and a discriminator with the structure of DenseNet [13] to balance the coarse style and textual details similar to [31]. We use the one-stage U-Net architecture [28] because it is the most common baseline network for image inpainting tasks with skip connections, which promises high quality of generated images by letting the network focus on the masked area. The bottom downsampling layers of the U-Net are replaced with stride 1 dilated convolution layers which have larger receptive field without losing too much information the same as [31]. Then the output images of the generator are sent into the two networks. The VGG16 network calculates style loss and perceptual loss for coarse style information consistent. The discriminator (DenseNet) calculates adversarial loss to generate details.

Most importantly, in the generator, one attention layer and the proposed spatially adaptive layer (spatially adaptive multi-scale attention module) are added before the third last deconvolution layer to generate the spatial-adaptive multi-scale attention scores. Our model concatenates the output feature maps from the upsampling and attention layer. In this way, attention scores based on multiple patch sizes can be used automatically, considering their spatial differences. In the following subsections, we will introduce the spatially adaptive multi-scale attention module in detail.

3.2 Spatially adaptive multi-scale attention module

3.2.1 Multi-scale attention

Image inpainting is a task to complete the missing region of an image by generating consistent and natural content. Conventional methods search the known region to fill the missing regions [2, 3, 17]. Yu et al. [39] combined the conventional patch match and deep learning image generation method GANs. A contextual attention score is computed to measure the similarity between different patches. This contextual attention layer is added into the middle of upsampling layers. They consider the missing regions and the known regions as the foreground and the background. A set of $3 \times 3$ patches in the background are extracted and reshaped as convolutional filters to measure the similarity between the foreground patches, $\{f_{x,y}\}$, and the background patches, $\{b_{x',y'}\}$, using the normalized inner product (i.e., cosine similarity).

$$sx,y,x',y' = \frac{f_{x,y}}{||f_{x,y}||} \cdot \frac{b_{x',y'}}{||b_{x',y'}||} > .$$

(1)

The similarities for all positions are then weighed to get an attention score for each pixel $s^*_{x,y,x',y'} = softmax_{y'}(\lambda sx,y,x',y')$, where $\lambda$ is a constant value. However, they selected fixed-scale patches to compute the attention score. Wang et al. [31] proposed using multi-scale patches to compute the attention score and flexibly use the background content. They used two different patch sizes ($1 \times 1$ and $3 \times 3$) and generated two feature maps ($\pi_{att1}$ and $\pi_{att3}$). For each attention map, they use similar strategy to calculate attention scores as Yu et al. [39] introduced above.
However, pixels at different positions shared the same reweighting score, because [31] combined multi-scale attention scores without distinguishing their spatial differences. Therefore, in this paper, we propose the spatially adaptive multi-scale contextual attention, which is detailed introduced in the following subsection.

### 3.2.2 Spatially adaptive multi-scale attention

The proposed spatially adaptive multi-scale attention also computes the attention scores with multi-scale of patches. Then rather than uniformly treating pixels at different spatial positions, the attention scores are combined differently for each pixel. Patches with different scales ($1 \times 1$ and $3 \times 3$) are selected to compute the attention scores (generated feature maps $\pi_{\text{att}1}$ and $\pi_{\text{att}3}$) using similar strategy to calculate attention scores as Yu et al. [39] introduced in the above subsection for an input feature map $\pi_{\text{in}}$. As Fig. 3 shows, our proposal combines multi-scale attentions differently according to their spatial locations. Two convolutional layers are used to predict reweighting scores ($\phi_{\text{wt}}$). This reweighting scores are 2D map with channels. The scores at the different positions are variable. In this way, multi-scale contextual attention scores are weighted differently for each pixel.

$$\phi_{\text{wt}} = f([\pi_{\text{att}1}, \pi_{\text{att}3}]),$$

where $f(\cdot)$ represents two convolutional layers, and the dimensions of $\phi_{\text{wt}}$ is $[2 \times c, h, w]$, $c$ is the channel of attention map, $h$ and $w$ are the size of the attention map.

### 3.3 Loss function

We use the following loss function for generating plausible style and detailed content similar to Wang et al. [31].

**Reconstruction loss** We use L1 loss as the reconstruction loss, which has been used by many generative models [14, 43]. Compared with other generative tasks, image inpainting

![Fig. 3 Our spatially adaptive multi-scale contextual attention layer: two layers of convolutional neural network are used to compute the spatial reweighting $\phi_{\text{wt}}$ for each pixel in the different positions. The inputs are two attention maps ($\pi_{\text{att}1}$ and $\pi_{\text{att}3}$) computed by multi-scale of patches ($1 \times 1$ and $3 \times 3$). The output feature maps are reweighted spatially differently.](image-url)
focuses on filling the missing regions of a given image. Therefore, the reconstruction loss can be computed only by the generated region.

$$L_{rec} = |(G(x_{in}) - x_{gt}) \odot M|_1,$$

where $G$ is the generator, $x_{in}$, $x_{gt}$ and $M$ represent the input image, the ground truth image, and the mask of missing regions, respectively. The operator $\odot$ means element-wise production so that it can focus on filling the missing region.

**Perceptual and style loss** Perceptual and style loss is computed from the high level feature. A pre-trained VGG16 network is used to extract the feature map from multiple layers for the generated images. Perceptual loss ($L_{perceptual}$) is obtained by comparing the extracted feature maps with ones of the ground truth image.

$$L_{perceptual} = \sum_{i=1}^{N} \frac{1}{HWC} \left| \phi_{\text{layer}_i}^{gt} - \phi_{\text{layer}_i}^{\text{pred}} \right|_1,$$

where $C$, $H$, and $W$ represent the channel, the height, and the width of the feature map, respectively. $N$ is the number of feature maps ($\phi$) generated by the VGG16 feature extractor. While the perceptual loss focuses on the high level structure, the style loss ($L_{style}$) can help to focus on the overview style information.

$$L_{style} = \sum_{i=1}^{N} \frac{1}{C \times C} \left| \frac{1}{HWC} \left( \phi'_{\text{layer}_i}^{gt} - \phi'_{\text{layer}_i}^{\text{pred}} \right) \right|_1,$$

where $\phi'_{\text{layer}_i} = \phi_{\text{layer}_i} \phi_{\text{layer}_i}^T$.

**Total variation loss** Total variation (TV) loss can help the generated image to become smoother [1, 38].

$$L_{tv} = \sum_{m,n} \left( (y_{m,n+1} - y_{m,n})^2 + (y_{m+1,n} - y_{m,n})^2 \right),$$

where $m$ and $n$ represent the coordinate position of a pixel, and $y = G(x_{in})$ is the generated image.

**Adversarial loss** Adversarial loss is proposed in generative adversarial networks (GAN) [8]. The generator is trained to fool the discriminator and the discriminator is trained to distinguish the generated and the real samples.

$$L_{adv} = \min_G \max_D \mathbb{E}_{x_{gt}} \log(D(x_{gt})) + \mathbb{E}_{x_{in}} \log(1 - D(G(x_{in}))),$$

where $G$ and $D$ are the generator and the discriminator, respectively. $\mathbb{E}$ is the expected value. The adversarial loss can then improve the quality of the generated image.

Our total loss consists of all the losses described above:

$$L_{total} = \lambda_1 L_{rec} + \lambda_2 L_{perceptual} + \lambda_3 L_{style} + \lambda_4 L_{tv} + \lambda_5 L_{adv}.$$

Springer
4 Experiments

4.1 Dataset

The experiments were conducted on three datasets: Paris Street View [5], Places [42], and CelebA-HQ [16]. Paris Street View dataset includes 14,900 train images and 100 test images. For the Places dataset, we selected images of five categories (canyon, rainforest, waterfall, valley, and rock_arch) from Places365-Standard dataset. Places dataset includes 25,000 train images and 500 test images. As for CelebA-HQ, we select 2,000 images for testing and 28,000 images for training.

4.2 Implementation details

During the training process, we resized each image to make its minimal length be 350 for Paris Street View and Places datasets. We then cropped each image randomly into 256 × 256 pixels. For CelebA-HQ, we resized each image to make its minimal length be 256 and cropped the center of it to 256 × 256. 1 and 3 were selected as patch sizes, which were used to compute the attention scores. Two convolutional layers were used to compute the reweighting map $\phi_{init}$. Their kernel sizes were 5 and 3. For GAN, the discriminator is DenseNet [13] as in MUSICAL [31]. The learning rate was 0.0002 and the batch size was 5.

Fig. 4  Inpainting results on Paris Street View dataset. From (b) to (e): CE, SN, MUSICAL and our proposal
4.3 Evaluation metrics

We conducted both quantitative evaluation and qualitative evaluation to verify the effectiveness of the proposed method. For quantitative evaluation, we used three most commonly used metrics: the structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and mean $l_1$ loss [18]. For qualitative evaluation, we conducted subjective evaluation via a user study and compared the generated result images. The detail is introduced in Section 5.3.

4.4 Comparative methods

We compared our proposal with three state-of-the-art methods: Context Encoder (CE) [25], Shift-Net (SN) [37], and Multi-Scale Attention (MUSICAL) [31]. These methods are all one stage. CE was the first one to apply deep CNN to image inpainting tasks using GANs. It was designed for images with resolutions of 128 × 128. Thus, the resolution of images shown in Figs. 4, 5, and 6 were all 256 × 256 except for those of CE. SN combined the traditional patch-match method and deep CNNs. A shift layer was added into the convolutional layers, which shifts similar patches to the missing regions. MUSICAL computes the attention scores between multi-scale patches and combines them without considering their spatial difference. Actually, there are numerous methods on image inpainting. In our paper, in order to clarify what point is compared, we compared with the most related and latest

![Fig. 5 Inpainting results on Places dataset. From (b) to (e): CE, SN, MUSICAL, and our proposal](image-url)
methods such as MUSICAL, which is one-stage, utilizing the surrounding information and with attention-based mechanism.

5 Results

We verify the effectiveness of the proposal comparing to the state-of-the-art methods in the following three aspects: analysis on the generated examples of the proposed method, subjective evaluation via a user study, and objective evaluation on several metrics.

5.1 Comparison on generated results

Images in Fig. 4 are generated by (b) CE, (c) SN, (d) MUSICAL, and (e) our proposal on the Paris Street View dataset. CE generated semantic content but it was not highly consistent with the surrounding image, especially in the border region. SN generated a blurred structure of a building and failed to generate details. MUSICAL generated a clearer structure of a building but with artifacts and failed to generate plausible images when it included a lot of details. Our proposal generated plausible images with far more details.

Figure 5 shows the results on the Places dataset. CE generated style-consistent image, but it was far from visual plausibility. SN generated blurred images. In some cases, such as
the one in the second row from the top, it was visually pleasing and acceptable. MUSICAL generated clear results but still failed to deal with complex content. Our proposal generated visually pleasing images with highly consistent content in most cases.

On the CelebA-HQ dataset, four methods achieved acceptable results as shown in Fig. 6. SN, MUSICAL, and our proposal had comparable results. Considering that the known regions included less information about the missing regions, methods stemming from patch match did not utilize the strengths of matching patches. Therefore, it led to comparable results.

5.2 Comparison on quantitative metrics

We have also compared the quantitative results on the three metrics: the structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and mean $l_1$ loss in Tables 1, 2 and 3. Compared with CE and MUSICAL, our proposal showed improved performance on the Paris Street View and Places datasets. On the CelebA-HQ dataset, we achieved comparable results with MUSICAL. However, on all three datasets, SN achieved the best results on SSIM, PSNR, and mean $l_1$ loss. This was caused by the fact that SN generates blurred results, (i.e., only low-frequency components are reconstructed), compared to MUSICAL and our proposal. Highly detailed images having high frequency components tended to have lower objective scores if they did not match the ground truth image, even though the resultant images were acceptable. Therefore, these objective measures are not sufficient to discuss the quality of image inpainting.

5.3 User study

As discussed in the previous section, some images which were more blurred achieved higher performance with quantitative metrics, which does not coincide with our intuition. Therefore, we also conducted a subjective evaluation by asking crowd workers. As shown in Fig. 7, we showed images generated by CE, SN, MUSICAL, and our proposal together and requested workers to rank them in an order from the best to the worst on plausibility.

| Method           | SSIM  | PSNR  | $l_1$ loss |
|------------------|-------|-------|------------|
| CE [25]          | 0.7948| 23.50 | 2.733%     |
| SN [37]          | 0.8726| 25.77 | 1.847%     |
| MUSICAL [31]     | 0.8441| 24.80 | 2.188%     |
| Our Proposal     | 0.8471| 24.92 | 2.144%     |

| Method           | SSIM  | PSNR  | $l_1$ loss |
|------------------|-------|-------|------------|
| CE [25]          | 0.8130| 21.69 | 3.057%     |
| SN [37]          | 0.8099| 21.59 | 3.094%     |
| MUSICAL [31]     | 0.8130| 21.69 | 3.057%     |
| Our Proposal     | 0.8099| 21.59 | 3.094%     |
Table 3: Numerical comparison on CelebA-HQ dataset

| Method           | SSIM  | PSNR  | $l_1$ loss |
|------------------|-------|-------|------------|
| CE [25]          | 0.8673| 25.17 | 2.039%     |
| SN [37]          | 0.9110| 27.38 | 1.407%     |
| MUSICAL [31]     | 0.8818| 25.73 | 1.777%     |
| Our Proposal     | 0.8817| 25.85 | 1.762%     |

Fig. 7: Example of images for user study. We show images generated by CE, SN, MUSICAL, and our proposal in a random order. Workers are required to give a rank like ABCD

Table 4: Scores for plausibility on Paris Street View dataset from the first time user study

| Dataset           | CE   | SN   | MUSICAL | Proposal |
|-------------------|------|------|---------|----------|
| Paris Street View | 1.38 | 2.26 | 3.15    | 3.21     |
| Places            | 1.25 | 2.23 | 3.18    | 3.22     |

Table 5: Results of the second time user study between the MUSICAL and our proposal

| Dataset           | MUSICAL better | Same | Proposal better |
|-------------------|----------------|------|-----------------|
| Paris Street View | 36 (36%)       | 9 (9%) | 55 (55%)       |
| Places            | 196 (39%)      | 42 (8%) | 262 (52%)      |
methods voted from the best to the worst respectively obtained the scores of 4, 3, 2, and 1. Each image received votes from 18 persons. Considering that methods stemming from patch match cannot show obvious improvements on the CelebA-HQ dataset, we did not include it in the experiment. In total, 256 people participated in the evaluation. From Table 4, we can see that, although SN achieved better quantitative scores, it was obviously worse than MUSICAL and our proposal. It is also shown that our proposal achieved the best result.

We then only showed images generated by MUSICAL and our proposal, requesting the workers to choose the subjectively better reconstructed images. Doing this experiment ensured that people could focus only on the differences between images from MUSICAL and our proposal. Each image was evaluated by 20 workers, and the total number of workers was 162. Table 5 shows a comparison between MUSICAL and our proposal. From Table 5, our proposed method was equivalent to or better than MUSICAL in most cases (64% and 60%, respectively). This result showed that, although the objective scores of MUSICAL and our proposed method were close to each other, the qualitative superiority of our proposal was notable.

6 Conclusion

In this paper, we proposed a spatially adaptive multi-scale contextual attention for image inpainting. Our proposal considered spatial differences and combine the multi-scale attention scores with variable reweighting based on the spatial positions of each pixel, which can represent different levels of information to generate more details. Our experiments showed the improvement of our proposal on the Paris Street View and Places datasets. In particular, results of the user study show that our proposal can generate visually better and more plausible images with more details. The results for CelebA-HQ dataset were also reasonable. In the future work, the object of inpainting can be more diverse, such as not only natural generation but also generation with high aesthetic scores, and can be applied to other similar tasks such as image editing.

Acknowledgements This work was partially financially supported by the Grants-in-Aid for Scientific Research Numbers JP19H05590 and JP19K20289.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Ballester C, Bertalmio M, Caselles V, Sapiro G, Verdera J (2001) Filling-in by joint interpolation of vector fields and gray levels. IEEE Trans Image Process 10(8):1200–1211
2. Barnes C, Shechtman E, Finkelstein A, Goldman DB (2009) Patchmatch: a randomized correspondence algorithm for structural image editing. In: ACM Transactions on graphics (ToG), vol 28, p 24. ACM
3. Bertalmio M, Sapiro G, Caselles V, Ballester C (2000) Image inpainting. In: Proceedings of the 27th annual conference on computer graphics and interactive techniques, pp 417–424

Springer
4. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv:1406.1078
5. Doersch C, Singh S, Gupta A, Sivic J, Efros A (2012) What makes paris look like paris?
6. Gao Z, Zhang H, Dong S, Sun S, Wang X, Yang G, Wu W, Li S, de Albuquerque VHC (2020) Salient object detection in the distributed cloud-edge intelligent network. IEEE Netw 34(2):216–224
7. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems, vol 27. Curran Associates, Inc., pp 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
8. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. arXiv:1406.2661
9. Hays J, Efros AA (2007) Scene completion using millions of photographs. ACM Trans Graph (TOG) 26(3):4–es
10. He K, Sun J (2012) Statistics of patch offsets for image completion. In: European conference on computer vision, pp 16–29. Springer
11. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
12. Hu J, Sheng L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
13. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
14. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
15. Jaderberg M, Simonyan K, Zisserman A et al (2015) Spatial transformer networks. In: Advances in neural information processing systems, pp 2017–2025
16. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196
17. Levin A, Zomet A, Weiss Y (2003) Learning how to inpaint from global image statistics. In: null, p 305. IEEE
18. Liao L, Hu R, Xiao J, Wang Z (2018) Edge-aware context encoder for image inpainting. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 3156–3160. IEEE
19. Liao X, Li K, Zhu X, Liu KR (2020) Robust detection of image operator chain with two-stream convolutional neural network. IEEE J Sel Top Sig Process 14(5):955–968
20. Liu G, Reda FA, Shah KJ, Wang TC, Tao A, Catanzaro B (2018) Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European conference on computer vision (ECCV), pp 85–100
21. Lu X, Wang W, Danelljan M, Zhou T, Shen J, Van Gool L (2020) Video object segmentation with episodic graph memory networks, arXiv:2007.07020
22. Lu X, Wang W, Ma C, Shen J, Shao L, Porikli F (2019) See more, know more: unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3623–3632
23. Lu X, Wang W, Shen J, Crandall D, Luo J (2020) Zero-shot video object segmentation with co-attention siamese networks. IEEE Transactions on Pattern Analysis and Machine Intelligence
24. Nazari K, Ng E, Joseph T, Qureshi F, Ebrahimi M (2019) Edgeconnect: structure guided image inpainting using edge prediction. In: The IEEE International conference on computer vision (ICCV) workshops
25. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA (2016) Context encoders: feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2536–2544
26. Peng L, Liao X, Chen M (2020) Resampling parameter estimation via dual-filtering based convolutional neural network. Multimedia Systems, 1–8
27. Pritch Y, Kav-Venaki E, Peleg S (2009) Shift-map image editing. In: 2009 IEEE 12th international conference on computer vision, pp 151–158. IEEE
28. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, pp 234–241. Springer
29. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556
30. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need
31. Wang N, Li J, Zhang L, Du B (2019) Musical: multi-scale image contextual attention learning for inpainting. In: Proceedings of the 28th international joint conference on artificial intelligence, pp 3748–3754. AAAI Press
32. Wang N, Ma S, Li J, Zhang Y, Zhang L (2020) Multistage attention network for image inpainting. Pattern Recognition, p 107448
33. Wang W, Lai Q, Fu H, Shen J, Ling H, Yang R (2021) Salient object detection in the deep learning era: an in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence
34. Woo S, Park J, Lee JY, So Kweon I (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
35. Xie C, Liu S, Li C, Cheng MM, Zuo W, Liu X, Wen S, Ding E (2019) Image inpainting with learnable bidirectional attention maps. In: The IEEE International conference on computer vision (ICCV)
36. Xiong W, Yu J, Lin Z, Yang J, Lu X, Barnes C, Luo J (2019) Foreground-aware image inpainting. In: The IEEE Conference on computer vision and pattern recognition (CVPR)
37. Yan Z, Li X, Li M, Zuo W, Shan S (2018) Shift-net: image inpainting via deep feature rearrangement. In: Proceedings of the European conference on computer vision (ECCV), pp 1–17
38. Yang C, Lu X, Lin Z, Shechtman E, Wang O, Li H (2017) High-resolution image inpainting using multi-scale neural patch synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6721–6729
39. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2018) Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5505–5514
40. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2019) Free-form image inpainting with gated convolution. In: Proceedings of the IEEE international conference on computer vision, pp 4471–4480
41. Zeng Y, Fu J, Chao H, Guo B (2019) Learning pyramid-context encoder network for high-quality image inpainting. In: The IEEE conference on computer vision and pattern recognition (CVPR)
42. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2017) Places: a 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1452–1464
43. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.