Computer Vision and Image Understanding 204 (2021) 103155

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Multi-scale attention network for image inpainting

Jia Qin, Huihui Bai *, Yao Zhao

Institute of Information Science & Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China

ARTICLE INFO

Keywords:
Image inpainting
Multi-scale neural network
Attention mechanism
Spatial attention
Channel attention

ABSTRACT

Recently, deep learning based inpainting methods have shown promising performance, in which some multi-scale networks are introduced to characterize image content in both details and structures. However, few of these networks explore local spatial components under different receptive fields and internal connection between multi-scale feature maps. In this paper, we propose a novel multi-scale attention network (MSA-Net) to fill the irregular missing regions, in which a multi-scale attention group (MSAG) with several multi-scale attention units (MSAUs) is introduced for fully analysing the features from shallow details to high-level semantics. In each MSAU, an attention based spatial pyramid structure is designed to capture the deep features from different receptive fields. In this network, attention mechanisms are explored for boosting the representation power of MSAU, where spatial attention is combined with each scale to highlight the most probably attentive spatial components and the channel attention is used as a globally semantic detector to build the connection between the multiple scales. Furthermore, for better inpainting results, a max pooling based mask update method is utilized to predict the missing parts from the border regions to the inside. Finally, experiments on Places2 dataset and CelebA dataset demonstrate that the proposed method can achieve better results than the previous inpainting methods.

1. Introduction

Image inpainting is the task to fill the missing pixels in a corrupted image, which can be used in numerous applications, such as image editing (Tang et al., 2014; Patrick et al., 2003; Portenier et al., 2018), object removal (Criminisi et al., 2003; Li et al., 2017; Xiao et al., 2012), noise removal (Zhang et al., 2017; Rakhshanfar and Amer, 2018), and the restoration of old photos (Chang et al., 2005). As an ill-posed inverse problem (Guillemot and Meur, 2013), researchers focus on predicting the missing areas realistically and accurately by analysing the known parts of the corrupted image.

Early inpainting methods are divided into diffusion-based inpainting and exemplar-based inpainting. Diffusion-based inpainting means to generate the local structure via parametric models or partial differential equations (Shen and Chan, 2002; Chan and Shen, 2001). Although the diffusion-based methods can generate the connected edges, it is difficult for these methods to restore the large missing region or the missing region with complex textures. On the other hand, exemplar-based inpainting methods try to fill the missing regions by exploiting image statistical and self-similarity priors (Criminisi et al., 2003; Efros and Leung, 1999). However, these methods are effective only when the priors and the missing parts have the similar textures.

Considering that the convolutional neural network (CNN) and the generative adversarial network (GAN) (Goodfellow et al., 2014) can obtain better visual quality, these deep learning technologies are adopted in image inpainting (Pathak et al., 2016; Yeh et al., 2017; Yu et al., 2018; Nazeri et al., 2019; Wang et al., 2018). Pathak et al. (2016) firstly introduce GAN into image inpainting. They present an unsupervised visual feature learning algorithm driven by context-based pixel prediction, which can capture not only appearance but also the semantics of visual structures. And then, Yeh et al. (2017) present the inpainting algorithm based on global GAN to introduce image semantics. Subsequently, Yu et al. (2018) and Wang et al. (2018) complete the corrupted image by both global and local GAN, in which the small region around the missing areas are adopted in discriminator to improve the performance of training. Additionally, Nazeri et al. (2019) propose the PatchGAN based inpainting network to focus on the patch details. Although the above methods have obtained promising performance, most of these networks tend to use a standard structure, in which the convolutional layers are stacked and only one size of kernel is selected in each layer. Furthermore, few of these methods adopt the multi-scale structure, especially to explore the locally spatial information and internal semantic characteristics of the multi-scale features.

To address these problems, we propose a multi-scale attention network (MSA-Net) for image inpainting, in which a multi-scale attention group (MSAG) is presented to improve the performance of inpainting network. Here, several multi-scale attention units (MSAUs) are included

* Corresponding author.
E-mail address: hhbai@bjtu.edu.cn (H. Bai).

https://doi.org/10.1016/j.cviu.2020.103155
Received 31 December 2019; Received in revised form 27 November 2020; Accepted 30 November 2020
Available online 1 December 2020

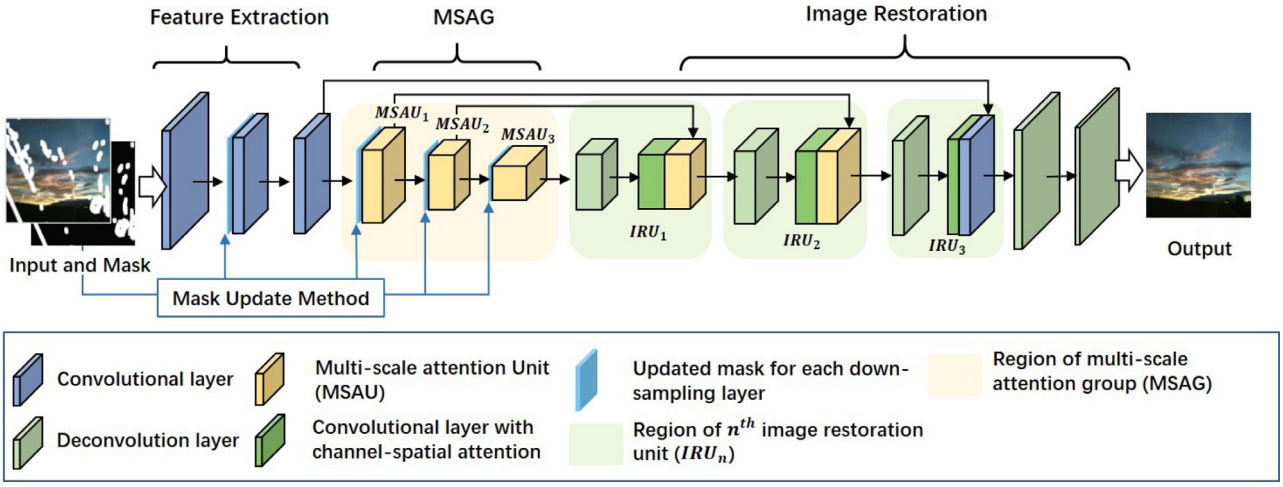

Fig. 1. The framework of MSA-Net.

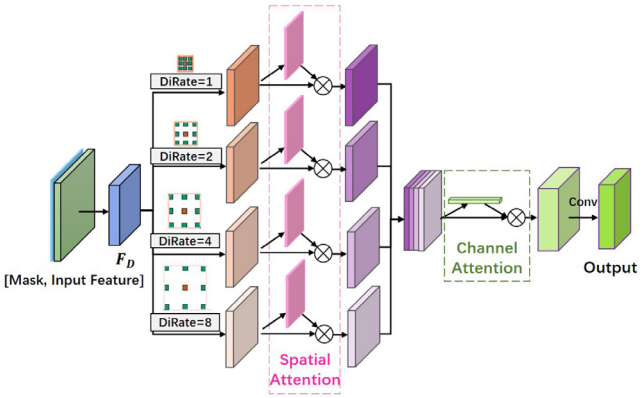

Fig. 2. The structure of multi-scale attention unit (MSAU).

in MSAG to catch the deep context from low-level details to high-level semantics gradually. In each MSAU, an attention based spatial pyramid structure is presented to analyse the image context from different receptive fields. In the structure, the obtained multi-scale features are strengthened by the attention mechanisms. Here, a fusing spatial attention is designed to combine average information, high activation and deep context of local neurons, which can distinguish the important spatial components from the feature in a scale. Furthermore, an augmented channel attention is presented to describe the semantics of features in all scales, which can emphasis informative maps and suppress useless deep context. Finally, in order to generate the missing parts from the border regions to the inside, a max pooling based mask update method is explored to define the location of the missing region for each downsampling layer of MSA-Net.

In summary, the contributions of our work can be described as follows:

- We propose a multi-scale attention network (MSA-Net) for image inpainting to restore the irregular missing regions, in which both the internal connection between multi-scale feature maps and the spatial characteristics of each scales are explored.
- In the MSA-Net, an MSAG with several MSAUs is proposed, in which an attention based spatial pyramid structure is designed in an MSAU to capture multi-scale features from appropriate receptive fields. And for boosting the representation power of multi-scale context, attention mechanisms are adopted in MSAU to construct more representative features by fusing both local spatial components in each scale and global channel connections in all multiple scales.
- For the downsampling layers of MSA-Net, a novel mask update method (MUM) is utilized to fill the missing parts from the border regions to the inside, which can mark the spatially valid features in current layer according to the irregular missing region.

This paper is organized as follows. In Section 2, some related works about image inpainting methods, multi-scale structure and attention models are introduced. In Section 3, the details of proposed inpainting network will be illustrated. And in Section 4, the experimental results will be displayed and analysed in details. Finally, the conclusion and future work is summarized in Section 5.

2. Related work

2.1. Image inpainting

Previous image inpainting researches generally fill the missing regions by the diffusion-based inpainting (Shen and Chan, 2002; Chan and Shen, 2001; Mainberger et al., 2011; Boscain et al., 2018; Zhang et al., 2014; Amrani et al., 2017) and exemplar-based inpainting (Criminisi et al., 2003; Efros and Leung, 1999; Jin and Bai, 2019; Kumar et al., 2016; Ružić and Pižurica, 2015). Here, Shen and Chan (2002) propose a total variation (TV) based general mathematical model for local non-texture inpainting. Chan and Shen (2001) propose a new inpainting model based on curvature-driven diffusions (CDD) to improve TV inpainting by realizing the connectivity principle. Though the above diffusion-based inpainting methods can ensure local intensity smoothness, they are not suitable to fill large missing regions for producing blurring artefacts. For better details of textures, the exemplar-based algorithms try to synthesize textures by directly copying similar patches from the input images (Akl et al., 2018). Efros and Leung (1999) propose a non-parametric method for texture synthesis, which can preserve local structure and produce good results for a wide variety of synthetic and real-world textures. Criminisi et al. (2003) propose an algorithm for removing large objects, in which the confidence in the synthesized pixel values is propagated in a manner similar to the propagation of information in inpainting. Since the exemplar-based algorithms fill the holes with suitable image patches, they are effective only when the priors and the missing parts have the similar textures.

In recent years, many deep learning methods have made dramatic achievements in image inpainting (Pathak et al., 2016; Yeh et al., 2017; Yu et al., 2018; Liu et al., 2018; Nazeri et al., 2019; Wang et al., 2018; Zheng et al., 2019; Guo et al., 2019; Hong et al., 2019). Yang et al. (2017) propose a multi-scale neural patch synthesis approach,

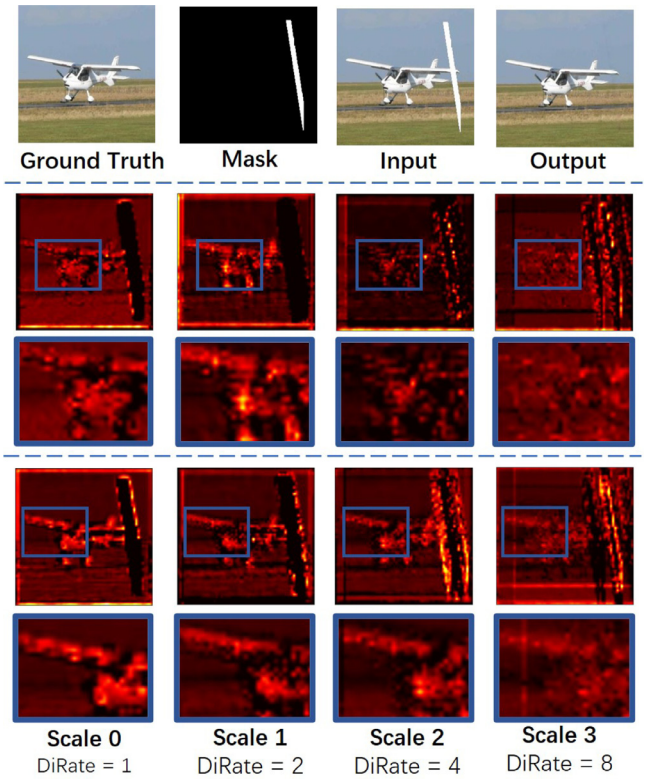

Fig. 3. Feature visualizations of multiple scales.

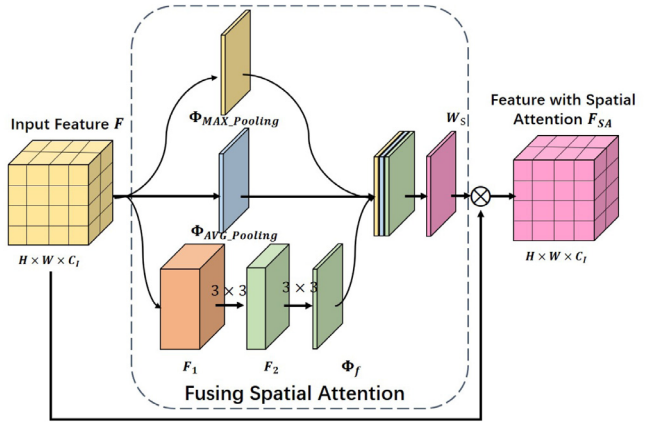

Fig. 4. The architecture of fusing spatial attention.

which can produce high-frequency details by matching and adapting patches with the most similar mid-layer feature correlations of a deep classification network. Yeh et al. (2017) propose a method for semantic image inpainting, which predicts information in large missing regions and achieves pixel-level photorealism. Liu et al. (2018) employ partial convolutions to avoid the colour discrepancy and blurriness and design a mechanism to automatically generate an updated mask for the next layer. Nazeri et al. (2019) propose a two-stage adversarial model to reproducing filled regions with fine details, which includes an edge generator and an image completion network. Guo et al. (2019) propose full-resolution residual network (FRRN) to fill irregular holes, which is effective for progressive image inpainting. And Hong et al. (2019) propose a concise deep fusion network (DFNet), which can achieve more accurate structure information accompanying by the adjustable loss constraints on each layer. However, few of these methods explore the locally spatial components and internal structure of deep features.

2.2. Multi-scale structure

Inspired by a neuroscience model of the primate visual cortex, Christian et al. (2015) propose a deep convolutional neural network codenamed Inception, which can improve the performance by increasing the depth and width of the network while keeping the computational budget constant. Subsequently, Christian et al. (2016) explore ways to scale up networks in ways that aim at utilizing the added computation as efficiently as possible by suitably factorized convolutions and aggressive regularization.

In addition, to generate more realistic and complex results, image inpainting models (Yu et al., 2018; Wang et al., 2018) can benefit from incorporating the same types of features, which can be captured from different receptive fields, configurations or stages. Yu et al. (2018) propose a coarse-to-fine inpainting network, which can not only synthesize novel image structures but also explicitly utilize surrounding image features as references during network training to make better predictions. And then, Wang et al. (2018) point out the limitation of the coarse-to-fine architecture, which is that errors in the coarse-level already influence refinement. For overcoming it, they present a generative multi-column network for image inpainting, which can produce visual compelling results even without previously common post-processing.

In this paper, we adopt the network with only one stage to avoid the limitation of coarse-to-fine architecture. Besides, motivating by the astounding performance of Inception structure, we propose an MSAG to improve the performance of multi-scale structure by analysing the internal characteristics of the feature, especially investigation of spatial component and semantic descriptor.

2.3. Attention model

As an important role in human perception, attention model is widely utilized to improve the performance of networks, in which channel attention and spatial attention work in global semantics and local context respectively. For spatial attentions (Woo et al., 2018; Chen et al., 2017; Xu et al., 2015a; Zhu et al., 2016), they follow the idea that humans selectively focus on salient parts rather than process a whole scene at once (Larochelle and Hinton, 2010), which can improve the representational power of a layer by enhancing the performance of spatial encodings throughout its deep feature. Xu et al. (2015a) propose the first attention based model to describe the content of images, in which the learned alignments correspond very well to human intuition. To further improve the spatial attention, the channel attention is adopted as a semantic detector to preserve the globally semantic consistency. Hu et al. (2017) focus on the channel relationship and propose a novel architectural unit, which adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. Chen et al. (2017) introduce a convolutional neural network (SCA-CNN) to incorporate spatial and channel attentions, which can dynamically emphasize the attentive in multi-layer feature maps. And Woo et al. (2018) propose convolutional block attention module (CBAM), a lightweight and general attention module for feed-forward convolutional neural networks, in which the feature is refined along channel axis and spatial axis.

For sharp and accurate results, the recent inpainting networks (Yan et al., 2018; Yang et al., 2017; Yu et al., 2018; Liu et al., 2019) present the attention mechanisms with the thought of patch-match to build the connection between holes and known context. For correlating feature patches at distant spatial locations, Yu et al. (2018) design contextual attention to match the generated patches with known contextual patches, in which channel-wise softmax is used to weight relevant patches. And Liu et al. (2019) propose a coherent semantic attention (CSA) layer to model the semantic relevance between the holes features, in which the CSA layer is embedded for refinement of inpainting network.

Though above methods have achieved significant progress in local pixel continuity, few of them analyse the deep feature of each scale in both global semantics and local textures. Moreover, they also fail to model the connection between scales with different receptive fields.

3. The proposed method

In this section, the proposed MSA-Net will be described in details. Firstly, an overview of MSA-Net will be shown, in which the generation of missing contents is introduced briefly. Then, MSAG is presented to extract the multi-scale features, in which MSAUs with attention based spatial pyramid structure will be described. Furthermore, the attention mechanisms will be elaborated in the proposed network, in which augmented channel attention, fusing spatial attention and progressive channel-spatial attention are used for stronger representation power. Besides, an effective mask-update method is displayed to generate specific masks for the downsampling layers in MSA-Net. Finally, the loss function is discussed for better training.

3.1. Overview of MSA-Net

In this paper, MSA-Net is proposed to predict the missing regions by analysing the available parts of the corrupted image from different receptive fields. As shown in Fig. 1, the network is divided into 3 parts: feature extraction, multi-scale attention group (MSAG) and image restoration (IR). In the network, several convolutional layers are firstly used for shallow feature extraction. And then in MSAG, several MSAUs are designed to analyse the deep features from different receptive scales, in which an attention based spatial pyramid structure is combined with attention mechanisms to strengthen the representations of multi-scale context. Here, spatial attention encodes the spatial component locally, while channel attention describes a feature map globally from a semantic viewpoint to emphasis effective channels. In the downsampling layers of feature extraction and MSAG in Fig. 1, a novel mask update method (MUM) is introduced to mark the valid location of irregular missing region. Finally, in IR part, several image restoration units (IRUs) are designed with channel-spatial attention to focus on the valuable features in both channel axis and spatial axis.

3.2. Attention based spatial pyramid structure

As shown in Fig. 1, MSAG contains several MSAUs to aggregate multi-scale features from the low-level details to high-level semantics gradually. In Fig. 2, the attention based spatial pyramid structure in each MSAU is designed to analyse the features encoding from the viewpoint of the receptive fields.

From the figure, for extracting the shallow feature F_D, a downsampling convolution is firstly introduced in MSAU as follows:

$$F_D = \tau\left(f_{K\times K}\left(F_{Input}\right)\right) \tag{1}$$

Here, F_{Input} is the input feature maps, which consists of feature from previous layer and corresponding mask update layer. $f_{K\times K}(\cdot)$ is the convolution with a $K \times K$ kernel. $\tau(\cdot)$ denotes ReLU (Xu et al., 2015b) activation function.

For analysing the locally spatial components from different receptive fields, the dilated convolutions with different dilation rates (DiRate) are considered as multiple scales. Here, four parallel dilated convolutions with the dilation rates of 1, 2, 4 and 8 are selected for the computation of multi-scale context. In each scale, the extracted feature F_r can be represented as:

$$F_r = \tau\left(d^r_{K\times K}\left(F_D\right)\right) \tag{2}$$

Here, $d^r_{K\times K}(\cdot)$ is the operator of dilated convolution, in which r is the dilation rate and $K \times K$ is the size of filter.

And then, the feature with spatial attention can be calculated as follows:

$$F^s_r = F_r \otimes f_s\left(F_r\right) \tag{3}$$

Here, $f_s(\cdot)$ is the function to compute map of spatial attention. $\otimes$ means element-wise multiplication of spatial attention and each channel of F. The weighted result by spatial attention is F^s_r. After achieving the features from all scales, the multi-scale features are concatenated as $F^s = \left[F^s_1, F^s_2, F^s_4, F^s_8\right]$. And the channel attention is computed as semantic descriptors to select important channels of F^s as:

$$F^{s,c} = F^s \odot f_c\left(F^s\right) \tag{4}$$

$f_c(\cdot)$ is the function of channel attention to calculate the weight of each channel. $\odot$ is the channel multiplication for each channel of feature and the corresponding channel weight. Finally, the output is obtained according to the weighted multi-scale features which are further aggregated by a convolutional operator in Fig. 2. The details of the attention mechanisms for MSA-Net will be introduced in the next subsection.

In Fig. 3, we visualize the feature of each scale in $MSAU_1$ of Fig. 1 to illustrate the efficiency of proposed multi-scale structure. From Fig. 3, although the features from larger receptive fields (e.g., DiRate = 8) tend to focus on more spatial information, they are not sensitive on the textures and edges. On the other hand, features with smaller receptive fields (e.g., DiRate = 1) can detect image details. Specifically, in Fig. 3, the features of aircraft fuselage are highlighted with blue rectangles in the second row and the fourth row. Besides, the third row and the last row are the enlarged details of these blue rectangles. From the figure, it can be seen that the feature maps with DiRate = 8 can smooth the texture of fuselage for considering more spatial information from larger receptive fields, while the feature maps with DiRate = 0 can display more detailed information, such as sand and grassland next to the fuselage. Therefore, for better inpainting result, the proposed MSAU is designed to embed local textures into more spatial information.

3.3. Attention mechanism for MSA-Net

In Fig. 1, spatial attention, channel attention and channel-spatial attention are introduced to improve the representation of features in MSAG and image restoration. In MSAG, a fusing spatial attention is combined with each scale to explore the spatial components from different receptive fields, while channel attention is introduced to emphasis the informative channels by modelling the internal relevance of multi-scale features. And in image restoration, channel-spatial attention is presented to analyse the deep feature in both global semantics and local details.

3.3.1. Fusing spatial attention

In Fig. 4, for the computation of spatial attention, the deep map from a gradual feature extraction model is fused with two pooling features. Here, average-pooling operator and max-pooling operator are used to consider the spatial locations of all neurons and to highlight the spatial locations with high activations, respectively (Woo et al., 2018).

It is noted that the gradual feature extraction model is utilized to detect the complex textures or important details by deep learning operators. In Fig. 4, F_1 and F_2 are extracted firstly as the summary of the spatial information with less feature maps. And then, the deep feature Φ_f is used to represent input feature map with only a channel.

For integrating the deep feature into spatial locations, a 2D map W_s containing the spatial information of input feature map F is calculated as:

$$W_s = s_1^{3\times 3}\left(\left[\Phi_f, \Phi_{Max\_Pooling}, \Phi_{AVG\_Pooling}\right]\right) \tag{5}$$

Here, $\Phi_{Max\_Pooling}$ and $\Phi_{AVG\_Pooling}$ are the results of max pooling operator and average pooling operator of input feature F. Φ_f is the

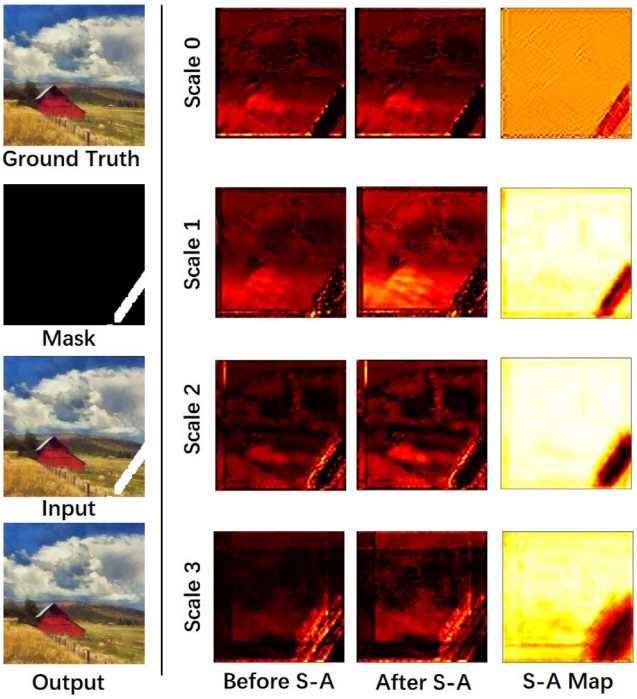

Fig. 5. Feature visualizations of fusing spatial attention.

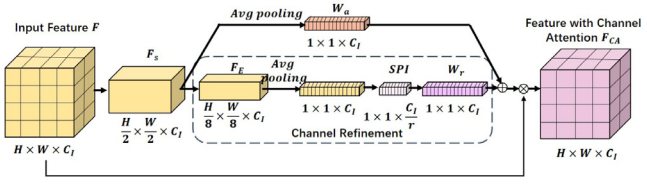

Fig. 6. The architecture of augmented channel attention.

result of gradual feature extraction model. $s_n^{m\times m}(\cdot)$ is a convolutional operator with a $m \times m$ kernel and n channels. W_s is a map of spatial attention, which is the combination of $\Phi_{AVG\_Pooling}$, $\Phi_{Max\_Pooling}$ and Φ_f.

Finally, the process of spatial attention can be summarized as:

$$F_{SA} = F \otimes W_s \tag{6}$$

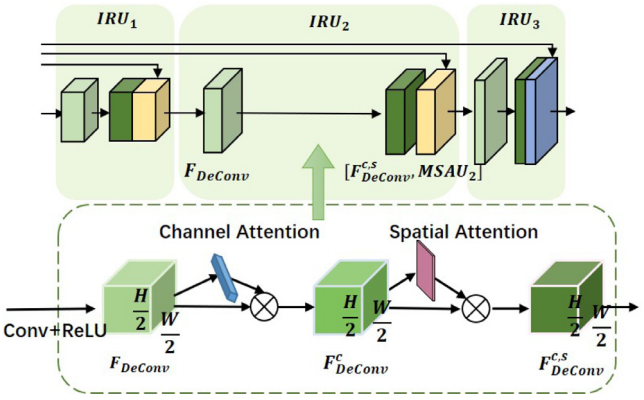

Fig. 8. The structure of image restoration.

where the input deep feature F is weighted by the spatial attention map W_s. $\otimes$ is the element-wise multiplication of spatial attention and each channel of F. And F_{SA} is the obtained feature processed by the spatial attention.

In Fig. 5, we visualize the features from different scales of $MSAU_1$, in which the feature before spatial attention, the feature after spatial attention and spatial attention map are displayed to describe the correlation between them. From this figure, after spatial attention, the available regions with more details or complex textures can be highlighted by obtaining higher weights of spatial attention map, while the smooth areas will be suppressed. Therefore, the important spatial components from known areas can be distinguished.

3.3.2. Augmented channel attention

In this paper, a channel attention is utilized to model the interdependencies between the channels explicitly, which can improve representation power and preserve important semantic features. In order to obtain the channel attention accurately, a shallow feature is extracted by the convolutional operator, which is represented by F_s in Fig. 6. Subsequently, the channel weights are mainly assigned by two aspects: globally average activation and channel refinement.

In global average activation, an average pooling is utilized to catch the global weights W_a as roughly channel-wise statistics. And in channel refinement, we squeeze the height and width of input feature firstly to aggregate semantic information as F_E, which can be calculated as follows:

$$F_E = \tau\left(f_n^{3\times3}\left(F_s\right)\right) \tag{7}$$

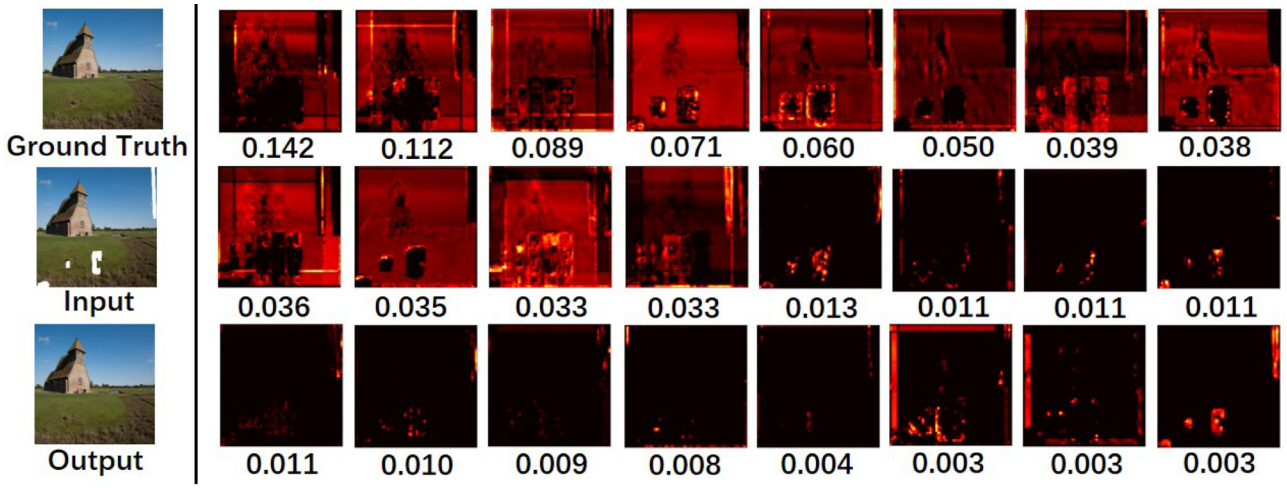

Fig. 7. The weights of augmented channel attention.

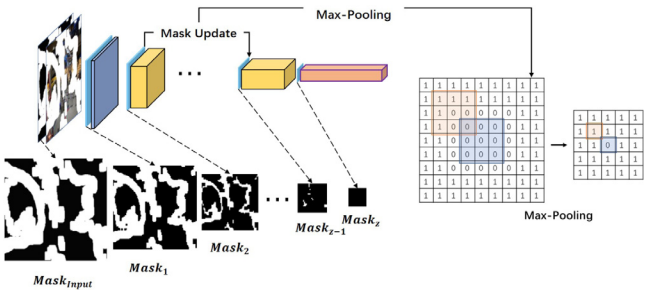

Fig. 9. The structure of mask update method.

where F_s is the squeezed feature with n channels. $f_n^{3\times3}(\cdot)$ is the convolutional layer with a 3×3 kernel and n channels. After the analysis of input feature F_s, an average-pooling operator is used for further semantic statistics and generation of channel descriptors. Here, $\tau(\cdot)$ means the ReLU activation function. In addition, a 1×1 convolution is adopted for obtaining the shared perception information (SPI) with size of $\frac{C_I}{r}$, which can merge the channel descriptors. In this process, r denotes the reduction ratio, which is set to 4 in this paper. After that, a layer is used in the shared channel information to obtain a feature vector $W_r = \left\{ w_1, \ldots, w_i, \ldots, w_{C_I} \right\}$, which makes the number of channel in restored vector consistent with input feature.

For improving the performance of channel attention, W_r is used to further refine roughly channel-wise statistics W_a as follows:

$$W_c = W_a + W_r \tag{8}$$

Finally, W_c is the output channel weights. In short, the adoption of channel attention is shown as the following:

$$F_{CA} = W_c \odot F \tag{9}$$

where $\odot$ is the channel multiplication for each channel of feature and its corresponding channel weight.

In Fig. 7, we select some representative maps and their channel weights to illustrate the efficiency of augmented channel attention. From this figure, the informative maps with complex context will be aligned higher weights, while the useless maps will obtain lower weights.

3.3.3. Progressive channel-spatial attention

For the deep feature, spatial attention works locally in each channel, and channel attention is globally for all feature maps from a spatial viewpoint. As shown in Fig. 8, for realistic results, a sequentially progressive channel-spatial attention is introduced in image restoration to focus on the vital features in both spatial axis and channel axis. Each IRU_i of image restoration can be denoted as follows:

$$U_i^{c,s} = s\left(c\left(h_i\left(U_{i-1}\right)\right)\right), (0 < i \leq 3) \tag{10}$$

where $s(\cdot)$ and $c(\cdot)$ mean the operators of spatial attention and channel attention respectively. $h_i(\cdot)$ is the ith deconvolutional layer for image restoration. U_i is the extracted feature of ith restoration layer h_i.

3.4. Mask update method

Since the missing region is irregular, it is difficult for the filter to define the location of missing region. Here, for marking the valid location of each downsampling layer, a space based mask update method (MUM) is realized by max-pooling operator, which involves the feature extraction and MSAG in Fig. 1. Inspired by Uhrig et al. (2017), a 0–1 matrix is applied to represent the damaged image effectively, in which the missing regions are equal to 0, and the available regions of corrupted image are equal to 1. As shown in Fig. 9, in MSAG, the masks are introduced to match with the multi-scale feature maps, which are

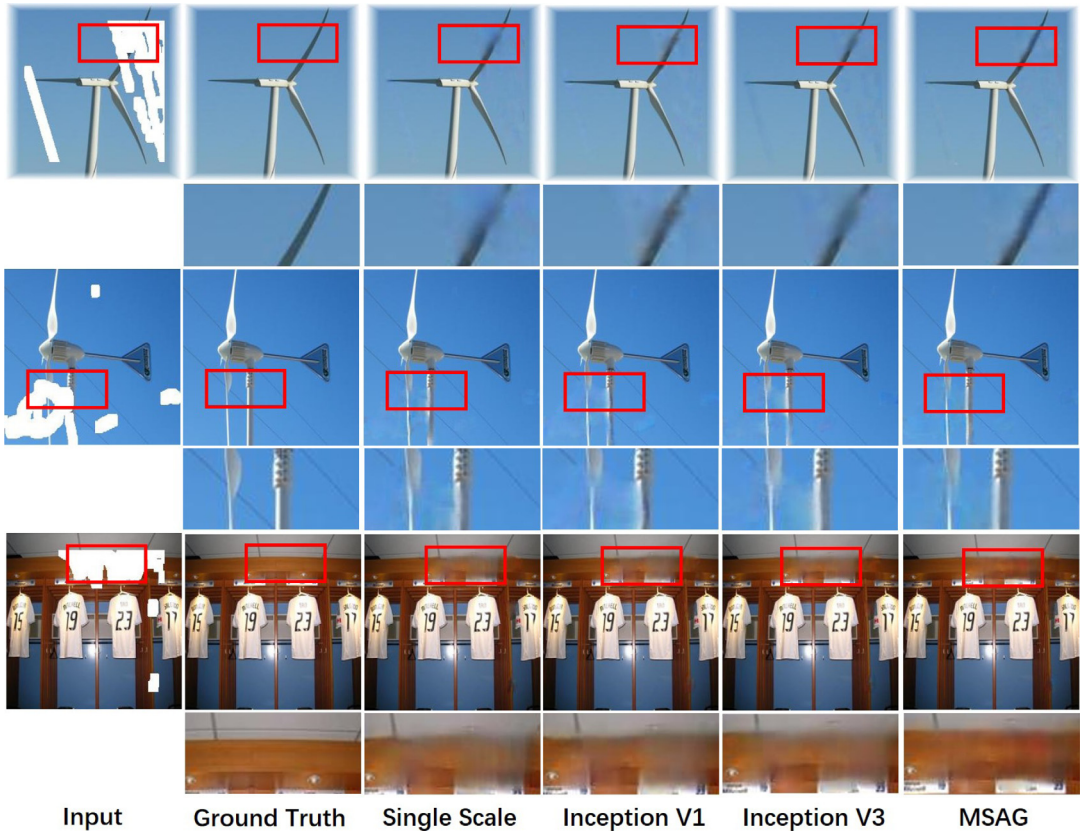

Fig. 10. The comparisons of the model without MSAG (Single Scale) and the models with Inception V1, Inception V3 and MSAUs in Places2 dataset.

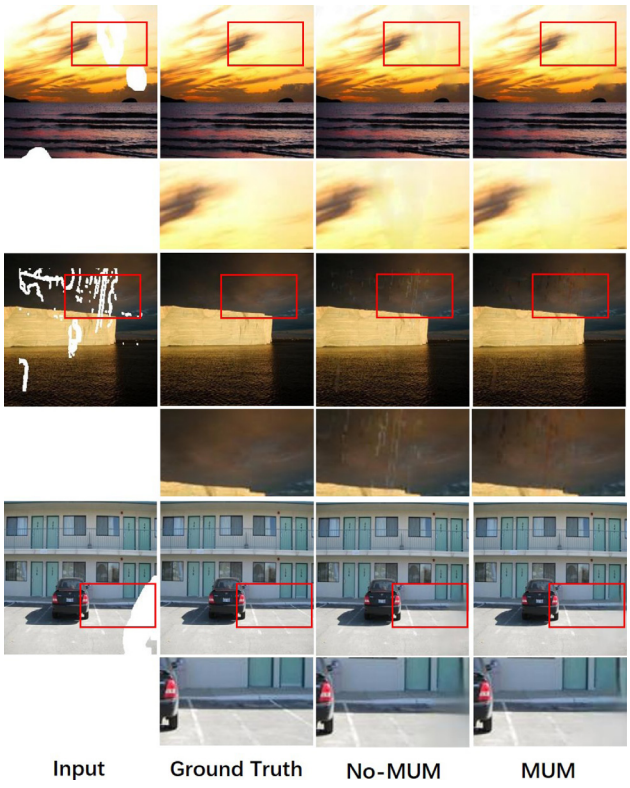

Fig. 11. The comparisons of the model without mask update method (No-MUM) and the MSA-Net in Places2 dataset.

updated as unfilled regions and valid regions. Therefore, the multi-scale masks can carry the valid information to subsequent layers to guide the restoration of damaged image from the boundary of missing areas to centre gradually.

3.5. Loss function

For better training of MSA-Net, L_2 loss, perceptual loss and style loss are used to represent the differences in pixel level and feature level. These loss functions are described in the following.

Here, L_2 loss is used to represent the pixel-level difference between the filled image and the original image, which is defined as follows:

$$L_2 = \frac{1}{W * H * C}\|I^* - I\|_2 \tag{11}$$

where I^* denotes the filled image and I means the original image. C, H and W are the channel size, height and width of the image I. And $\|\cdot\|_2$ is a 2 norm. L_2 loss is effective in the image inpainting to capture the overall structure of the missing region (Pathak et al., 2016). However, it is difficult for L_2 loss to recover sharp edges, which may lead to overly-smooth results (Lim et al., 2017).

For a better restoration, perceptual loss and style loss are further added in the proposed MSA-Net to consider the correlations between the original images and the damaged images in the feature level.

For the perceptual loss, it can be then calculated as follows:

$$L_{perceptual\_loss} = \sum_{i=1}^{P} \frac{1}{W_i \times H_i \times C_i}\|F_i^{Out} - F_i^{GT}\|_2 \tag{12}$$

where F_i^{Out} and F_i^{GT} are the ith feature maps with size of $H_i \times W_i \times C_i$. For our work, F_i corresponds to feature maps from aggregated multi-scale layer and deconvolutional layers in image restoration, which are shown in Fig. 1.

Furthermore, for the style loss, Gram Matrix is introduced firstly by Gatys et al. (2015) as:

$$G_i^{Out} = \gamma\left(F_i^{Out}\right)^T \gamma\left(F_i^{Out}\right), (0 < i \le 3) \tag{13}$$

$$G_i^{GT} = \gamma\left(F_i^{GT}\right)^T \gamma\left(F_i^{GT}\right), (0 < i \le 3) \tag{14}$$

Here, F_i^{Out} is the feature obtained by the input of damaged image and F_i^{GT} is the feature map with the input of original image. γ is the vectorization process of F_i^{Out} and F_i^{GT} with sizes of $\left(H_i \times W_i\right) \times C_i$, and the sizes of G_i^{Out} and G_i^{GT} are $C_i \times C_i$. Then, the style loss is defined as follows:

$$L_{style\_loss} = \sum_{i=1}^{P} \frac{1}{C_i \times C_i}\|K_i\left(G_i^{Out} - G_i^{GT}\right)\|_1, (0 < i \le 3) \tag{15}$$

where G_i^{Out} and G_i^{GT} are the Gram Matrices of the ith selected layers F_i^{Out} and F_i^{GT} respectively. K_i is the normalization factor $\frac{1}{H_i \times W_i \times C_i}$ of G_i^{Out} and G_i^{GT}. And $\|\cdot\|_1$ is the 1 norm.

Finally, the total loss of this inpainting method can be denoted as follows:

$$L = \alpha \times L_2 + \beta \times L_{perceptual\_loss} + \gamma \times L_{style\_loss} \tag{16}$$

Here, α, β and γ are used to weight these three types of loss functions. In this paper, for limiting L_2, $L_{perceptual\_loss}$ and L_{style_loss} in appropriate data ranges, α, β and γ are 1, 0.001 and 250 respectively.

4. Experimental results

In order to demonstrate the validity of the proposed inpainting network, it is conducted on Places2 dataset (Zhou et al., 2018) and CelebA dataset (Liu et al., 2015). We use the original train set and test set for Places2. As for the CelebA-HQ, 28 K and 2 K images are used as training set and test set respectively. Moreover, we test the proposed MSA-Net in an irregular mask dataset released by Liu et al. (2018). All the masks and images are with the size of 256 × 256.

In the training phase, we set the learning rate to 0.0003. And after training, the model is refined by 0.0001. We use peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) (Wang et al., 2004) and the visual quality of the filled image to evaluate the performance of MSA-Net. PSNR is the comparison in pixel level, while SSIM is the holistic similarity between the original image and the restored image. We train on a single GPU of Titan XP (12GB) with a batch size of 16. The model of Places2 dataset is trained for 7 days, whereas CelebA-HQ for 3 days. And the average test time of an image is about 0.026 s in both Places2 and CelebA-HQ.

4.1. Ablation study

In this subsection, for discussing the importance of MSAG, we firstly compare the model with several scales and the model with only one scale. And then, we discuss the efficiency of the max pooling based adaptive mask update method. Afterwards, we further illustrate the performance of attention mechanisms, which contain fusing spatial attention, augmented channel attention and progressive channel-spatial attention. Finally, we also compare our MSA-Net with previous methods.

4.1.1. Investigation of MSAG

For improving the performance of MSA-Net, we design the structure of MSAG to achieve the multi-scale context. For making a comparison with MSAG, we also test the models that replaces the MSAG structure with the single-scale dilated convolutions and two representative multi-scale structures from Inception V1 (Christian et al., 2015) and Inception V3 (Christian et al., 2016). The multi-scale structures of $MSAU_1$ and these in Inception V1 or Inception V3 are shown in Table 1, in which the parameters of related structures are unified to similar parameters. Here, an inception unit in V1 is divided into four scales, which have

Table 1
The multi-scale structures of MSAU, Inception V1 and Inception V3.

			Inception V1		Inception V3		MSAU	
			Operator	Output	Operator	Output	Operator	Output
Multiple scales	Scale 1	Layer 1	Conv 1 × 1	192	Conv 3 × 3	192	DiConv 3 × 3 (rate = 1)	48
	Scale 2	Layer 1	Conv 1 × 1	96	Conv 1 × 1	32	DiConv 3 × 3 (rate = 2)	48
		Layer 2	Conv 3 × 3	192	Conv 3 × 3	48	–	–
		Layer 3	–	–	Conv 3 × 3	48	–	–
	Scale 3	Layer 1	Conv 3 × 3	24	Max pooling 3 × 3	128	DiConv 3 × 3 (rate = 4)	48
		Layer 2	Conv 1 × 1	64	Conv 1 × 1	64	–	–
	Scale 4	Layer 1	Max pooling 3 × 3	128	–	–	DiConv 3 × 3 (rate = 8)	48
		Layer 2	–	–	Conv 1 × 1	64	–	–
Multi-scale maps			512		304		192	
Parameter of multi-scale structure			240K		268K		221K	
Reduce			Conv 1 × 1		Conv 1 × 1		–	
Output maps			192		192		192	
Total parameter			338K		326K		221K	

Table 2
The comparisons of MSA-Net with the single-scale structure, Inception V1 and Inception V3 in Places2 dataset.

	Single scale	Inception V1	Inception V3	MSAG
PSNR	26.513	26.535	26.553	26.580
SSIM	0.8768	0.8761	0.8771	0.8775

the same number of scales with MSAU. And the structure of Inception V3 is a way to expand the filter banks with less parameters. At the end of Inception V1 and V3, in order to ensure the output map number of inception units are the same as MSAU, we added a convolution layer to aggregate the multi-scale feature and realize the output map number of inception units.

From Table 2, it can be seen that the usage of MSAG can obtain more accurate results. In addition, the visual quality of the MSA-Net and the network without MSAG is also shown in Fig. 10, in which we can see that the results in MSAG are better restoration in texture than the model with other multi-scale structures.

4.1.2. Investigation of mask update method

A mask update method (MUM) is used in down-sampling layers of MSA-Net to guide the feature extraction from the boundary of missing areas to centre gradually. For displaying the effectiveness of this mask update method, the results of the proposed MSA-Net are compared with the network without mask update in Table 3, which shows that MSA-Net has better inpainting performance than the model without MUM. Furthermore, some results of the comparison are shown in Fig. 11, from which we can find that the visual quality of the proposed method is more realistic and accurate than the model without MUM.

4.1.3. Analysis of attention mechanism

We compare the MSA-Net with the models that apply different attention mechanisms in Table 4 to clearly show their performance,

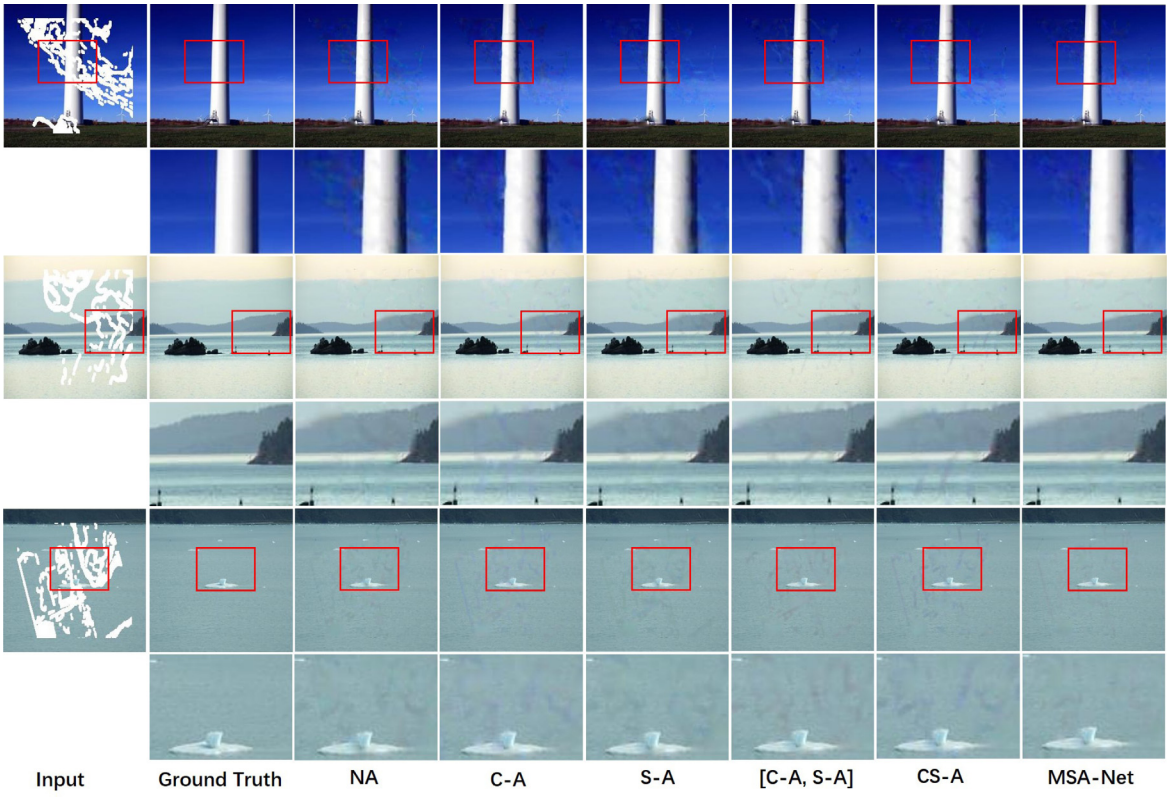

Fig. 12. The effectiveness of channel attention in MSAG (C-A), spatial attention in MSAG(S-A) and channel-spatial attention in IR (CS-A).

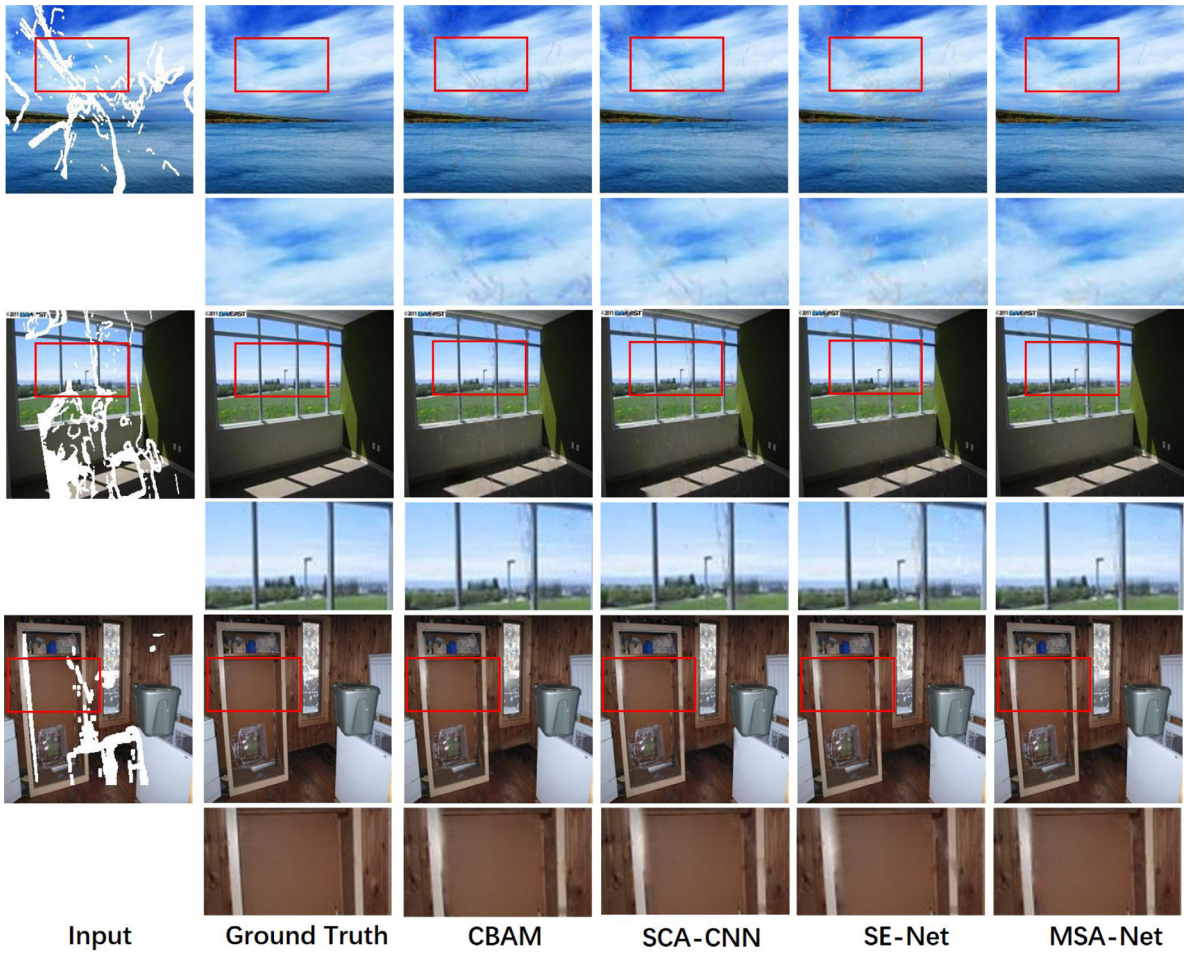

Fig. 13. The comparisons of the proposed attentions in Places2 dataset with them in CMBA, SCA-CNN and SE-Net.

Table 3
The comparisons of MSA-Net in Places2 dataset with the network without mask update.

	No-MUM	MUM
PSNR	26.580	26.615
SSIM	0.8775	0.8788

Table 4
The effectiveness of channel attention (C-A), spatial attention (S-A) and channel-spatial attention (CS-A) in Places2 dataset.

C-A	S-A	CS-A	PSNR	SSIM
✗	✗	✗	26.615	0.8788
✓	✗	✗	26.633	0.8802
✗	✓	✗	26.690	0.8803
✓	✓	✗	26.692	0.8809
✗	✗	✓	26.724	0.8816
✓	✓	✓	26.802	0.8820

which contains channel attention in MSAG (C-A), spatial attention in MSAG (S-A) and channel-spatial attention in IR (CS-A). It is shown that with the addition of C-A, S-A and CS-A, the results are gradually getting better in PSNR and SSIM.

In addition, for illustrating the efficiency of our proposed attention mechanisms, we compare them with CBAM (Woo et al., 2018), SCA-CNN (Chen et al., 2017) and SE-Net (Hu et al., 2017). Here, CBAM and SCA-CNN design both spatial attention and channel attention, while SE-Net only proposes a channel attention. Therefore, as shown in Table 5, the results of CBAM and SCA-CNN are the models that both spatial attention and channel attention are replaced. And for the model of SE-Net, only channel attention is replaced in MSA-Net.

Table 5
The comparisons of the proposed attention mechanism with other attention methods in Places2 dataset.

	CBAM	SCA-CNN	SE-Net	MSA-Net
PSNR	26.710	26.728	26.758	26.802
SSIM	0.8811	0.8805	0.8817	0.8820

Furthermore, the visual quality is displayed in Figs. 12 and 13, in which it can be observed that the proposed attention based network can obtain more accurate results in colour consistency and the image contents.

4.2. Comparisons with other inpainting methods

In order to evaluate our proposed MSA-Net, it is compared with CA (Yu et al., 2018), PConv (Liu et al., 2018), EdgeConnect (Nazeri et al., 2019) and GMCNN (Wang et al., 2018), in which PSNR and SSIM are used as image quality metrics. The trained models are compared on the Places2 dataset and CelebA dataset. The results are displayed in Table 6, in which all the methods are tested in an irregular mask dataset (Liu et al., 2018). The irregular mask dataset is further categorized by the size of missing regions, which can generate six categories with different missing region ratios: (0.01, 0.1], (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5] and (0.5, 0.6] mask for all sizes. Furthermore, we also test the model in regular masks with fixed size, in which the

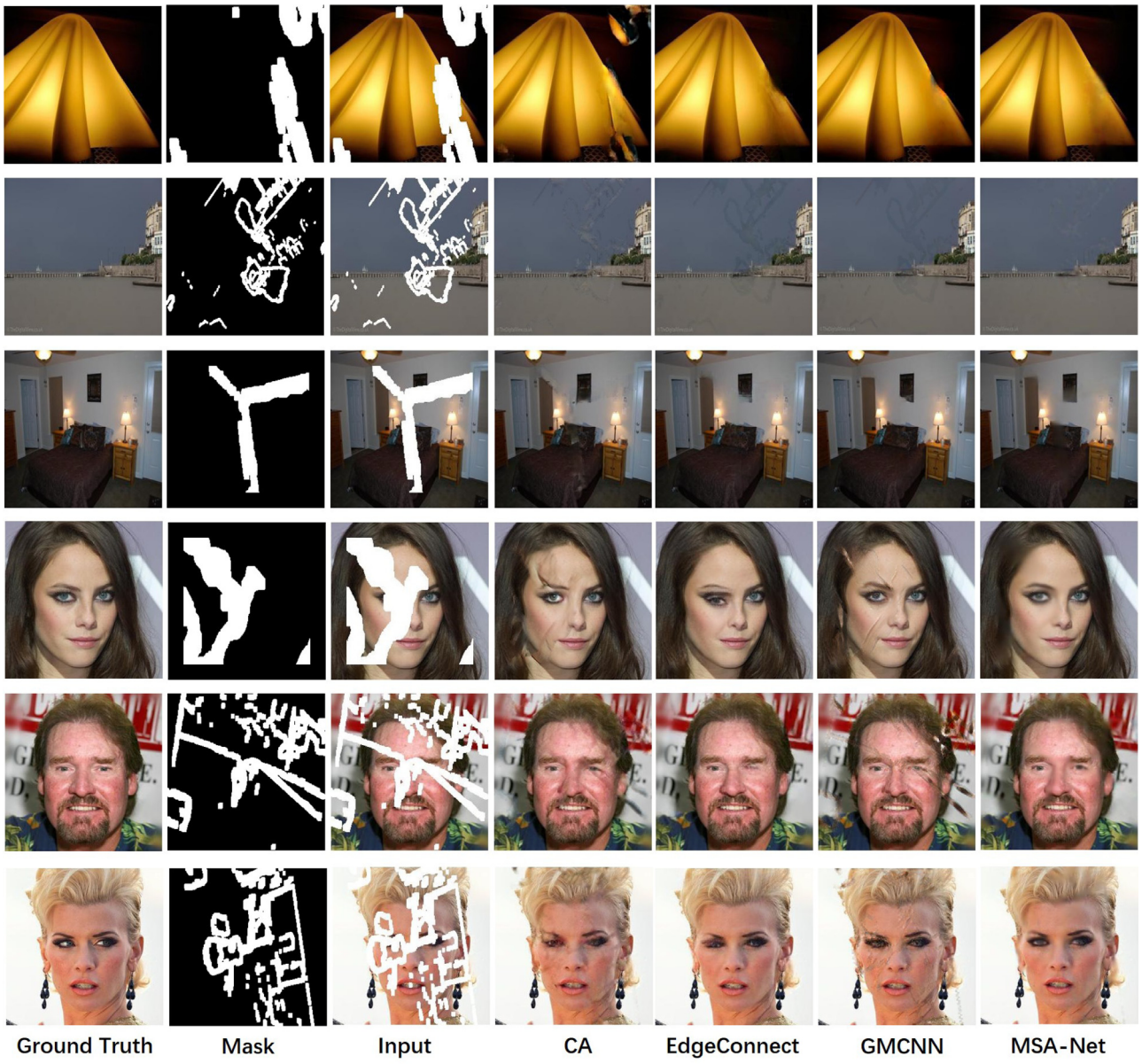

Fig. 14. The comparisons of the final results with CA (Yu et al., 2018), EdgeConnect (Nazeri et al., 2019) and GMCNN (Wang et al., 2018) in Place2 dataset and CelebA dataset.

Table 6
The quantitative evaluation of MSA-Net in Places2 dataset and CelebA dataset.

		(0.01, 0.1]		(0.1, 0.2]		(0.2, 0.3]		(0.3, 0.4]		(0.4, 0.5]		(0.5, 0.6]		Fixed	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Places2	CA	30.71	0.964	24.56	0.898	21.34	0.813	19.27	0.725	17.84	0.636	16.42	0.523	20.66	0.774
	PConv	34.05	0.946	28.02	0.869	24.90	0.777	22.45	0.685	20.86	0.589	18.63	0.476	–	–
	EdgeConnect	–	–	27.95	0.920	24.92	0.861	22.84	0.799	21.16	0.731	–	–	21.75	0.823
	GMCNN	34.84	0.986	28.81	0.957	25.42	0.912	22.96	0.854	20.86	0.778	17.20	0.593	18.88	0.737
	MSA-Net	**35.80**	**0.988**	**30.03**	**0.965**	**26.88**	**0.929**	**24.69**	**0.884**	**22.91**	**0.826**	**20.51**	**0.701**	**22.89**	**0.810**
CelebA	CA	33.37	0.981	27.71	0.946	24.66	0.902	22.29	0.844	20.37	0.775	18.11	0.667	23.12	0.861
	EdgeConnect	**39.60**	0.985	**33.51**	0.961	30.02	0.928	27.39	0.890	25.28	0.846	22.11	0.771	25.49	0.891
	GMCNN	32.66	0.978	26.63	0.938	23.50	0.890	21.23	0.832	19.66	0.773	17.75	0.690	25.00	0.905
	MSA-Net	38.55	**0.994**	33.12	**0.982**	**30.20**	**0.967**	**27.90**	**0.945**	**25.99**	**0.917**	**23.22**	**0.852**	**26.29**	**0.916**

missing areas account for 25% of all image pixels. And the fixed masks are centred at a random location within the test image.

In Table 6, the results of Places2 dataset for Liu et al. (2018) and Nazeri et al. (2019), and CelebA for Nazeri et al. (2019) are taken from their paper. And other results are generated by their pre-trained weights respectively, if they are available online. From the tables, it can be seen that our method can produce more accurate results than others. Moreover, it can be observed clearly from Fig. 14 that the proposed MSA-Net also outperforms other inpainting algorithms in subjective vision.

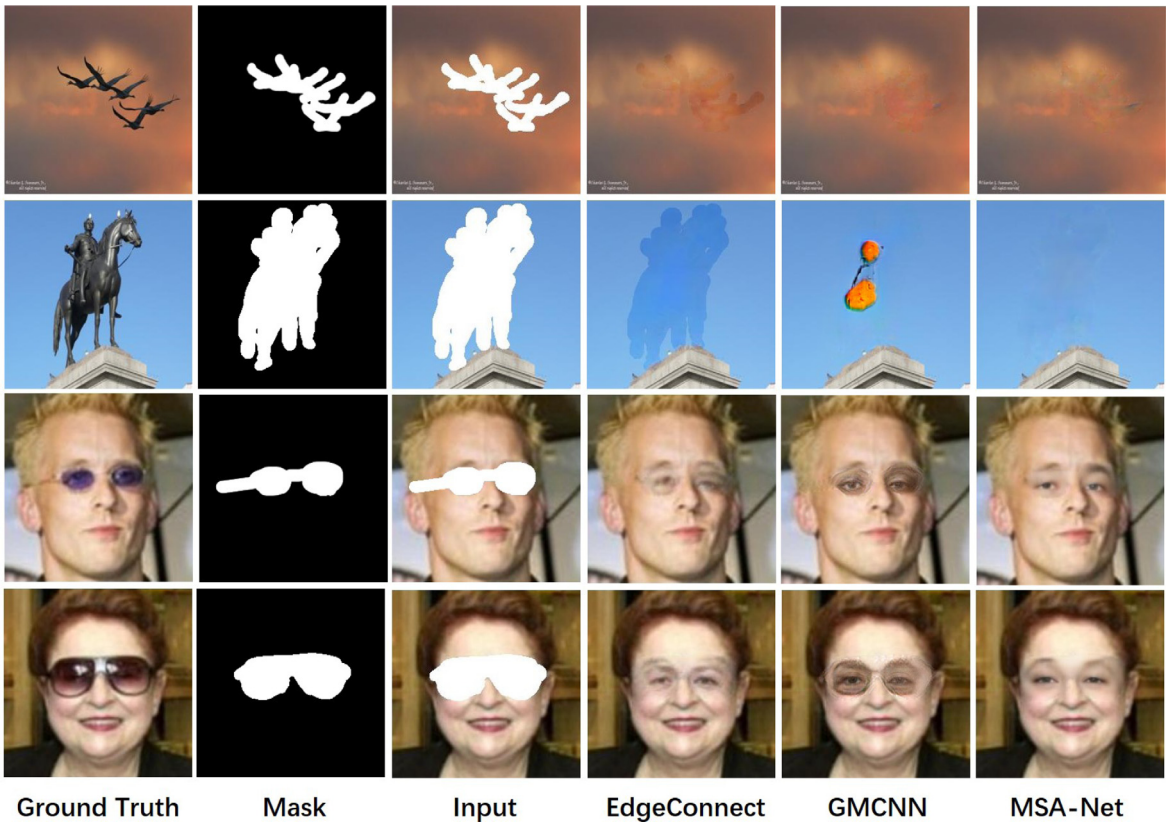

Fig. 15. The results of removing occlusions on ImageNet dataset and LFW dataset.

4.3. Object removal

Object removal is one of main applications in image inpainting, which means to remove the unnecessary objects or unwanted occlusions, such as the passing-by person or some scratches. Here, in order to show the efficiency of our network, we also display the results of occlusion removal in Fig. 15, in which the inpainting models in natural images and face images are tested in ImageNet (Deng et al., 2009) and labelled faces in the wild (LFW) (Huang et al., 2012) respectively. From this figure, it can be seen that the proposed network can remove the unnecessary objects in images naturally.

5. Conclusion and future work

In this paper, we propose a novel multi-scale attention network (MSA-Net) for image inpainting to fill the irregular missing regions. For extracting the multi-scale context gradually, we design a multi-scale attention group (MSAG), which consists of several multi-scale attention units (MSAUs). MSAU is the structure to capture features from various receptive fields, in which dilated convolutions with different dilation rates can be regarded as the various scales. Furthermore, three attention mechanisms are introduced to analyse the locally spatial components of each scale and internal semantic characteristics of multi-scale features, which consist of the fusing spatial attention, augmented channel attention and progressive channel-spatial attention. Moreover, in order to get a realistic and accurate results, the max pooling based mask update method is introduced to predict the missing parts from the border regions to the inside. Finally, the experimental results have demonstrated the superior performance of our proposed MSA-Net on restoration of damaged image.

However, the proposed algorithm may exist blurriness in the generated contents when the missing areas are large, which is still a challenge in the image inpainting to restore the large missing regions accurately and realistically. Aiming at this problem, we will further extend the work to explore the connection between the missing regions and the available information of the corrupted image, in which the texture and multi-scale structure will be combined to improve the performance of inpainting network.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by Fundamental Research Funds for the Central Universities (2018YJS027 and 2019JBZ102) and National Natural Science Foundation of China (No. 61972023).

References

Akl, A., Yaacoub, C., Donias, M., Da Costab, J.P., Germain, C., 2018. A survey of exemplar-based texture synthesis methods. Comput. Vis. Image Underst. 172, 12–24.

Amrani, N., Serra-Sagrista, J., Peter, P., 2017. Diffusion-based inpainting for coding remote-sensing data. IEEE Geosci. Remote Sens. Lett. 1–5.

Boscain, U.V., Chertovskih, R., Gauthier, J.P., 2018. Highly corrupted image inpainting through hypoelliptic diffusion. J. Math. Imaging Vision 60, 1231–1245.

Chan, T.F., Shen, J., 2001. Nontexture inpainting by curvature-driven diffusions. J. Vis. Commun. Image Represent. 12, 436–449.

Chang, R., Sie, Y., Chou, S., 2005. Photo defect detection for image inpainting. In: IEEE International Symposium on Multimedia, ISM.

Chen, L., Zhang, H., Xiao, J., 2017. SCA-CNN: Spatial and Channel-wise attention in convolutional networks for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR.

Christian, S., Vincent, V., Sergey, I., Jonathon, S., Zbigniew, W., 2016. Rethinking the inception architecture for computer vision. In: IEEE conference on computer vision and pattern recognition, CVPR, pp. 2818–2826.

Christian, S., Wei, L., Yangqing, J., Pierre, S., 2015. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR.

Criminisi, A., Perez, P., Toyama, K., 2003. Object removal by examplar-based inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 721–728.

Deng, J., Dong, W., Socher, R., 2009. ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.

Efros, A., Leung, T., 1999. Texture synthesis by non-parametric sampling. In: IEEE International Conference on Computer Vision, ICCV, pp. 1033–1038.

Gatys, L.A., Ecker, A.S., Bethge, M., 2015. A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., 2014. Generative adversarial nets. In: Neural Information Processing Systems. NIPS, pp. 2672–2680.

Guillemot, C., Meur, O.L., 2013. Image inpainting: Overview and recent advances. IEEE Signal Process. Mag. 31, 127–144.

Guo, Z., Chen, Z., Yu, T., Chen, J., Liu, S., 2019. Progressive image inpainting with full-resolution residual network. In: Proceedings of the 27th ACM International Conference on Multimedia. ACM, pp. 2496–2504.

Hong, X., Xiong, P., Ji, R., Fan, H., 2019. Deep fusion network for image completion. In: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, pp. 2033–2042.

Hu, J., Shen, L., Albanie, S., 2017. Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell.

Huang, G., Mattar, M., Honglak, L., 2012. Learning to align from scratch. In: Proceedings of the Neural Information Processing Systems, NIPS.

Jin, D., Bai, X., 2019. Patch-sparsity-based image inpainting through a facet deduced directional derivative. IEEE Trans. Circuits Syst. Video Technol. 29, 1310–1324.

Kumar, V., Mukherjee, J., Mandal, S.K.D., 2016. Image inpainting through metric labeling via guided patch mixing. IEEE Trans. Image Process. 25, 5212–5226.

Larochelle, H., Hinton, G.E., 2010. Learning to combine foveal glimpses with a thirdorder boltzmann machine. In: Neural Information Processing Systems. NIPS.

Li, Y., Baciu, G., Han, Y., Li, C., 2017. Indoor localization with occlusion removal. In: International Conference on Cognitive Informatics & Cognitive Computing, ICCI*CC, pp. 191–198.

Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M., 2017. Enhanced deep residual networks for single image super-resolution, In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1132–1140.

Liu, H., Jiang, B., Xiao, Y., 2019. Coherent semantic attention for image inpainting. In: IEEE International Conference on Computer Vision, ICCV.

Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep learning face attributes in the wild. In: IEEE International Conference on Computer Vision, ICCV.

Liu, G., Reda, F.A., Shih, K.J., Wang, T., Tao, A., Catanzaro, B., 2018. Image Inpainting for irregular holes using partial convolutions. In: European Conference on Computer Vision, ECCV, pp. 85–100.

Mainberger, M., Hoffmann, S., Weickert, J., 2011. Photo defect detection for image inpaintingoptimising spatial and tonal data for homogeneous diffusion inpainting. In: International Conference on Scale Space and Variational Methods in Computer Vision.

Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M., 2019. EdgeConnect: Generative image inpainting with adversarial edge learning. In: IEEE International Conference on Computer Vision, ICCV.

Pathak, D., Krahenbuhl, P., Donahue, J., 2016. Context Encoders: Feature learning by inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2536–2544.

Patrick, P., Gangnet, M., Blake, A., 2003. Poisson Image editing. ACM Trans. Graph. 22, 313–318.

Portenier, T., Hu, Q., Szabó, A., 2018. FaceShop: Deep sketch-based face image editing. ACM Trans. Graph..

Rakhshanfar, M., Amer, M.A., 2018. Low-frequency image noise removal using white noise filter. In: IEEE International Conference on Image Processing, ICIP, pp. 3948–3952.

Ružić, T., Pižurica, A., 2015. Context-aware patch-based image inpainting using Markov random field modeling. IEEE Trans. Image Process. 24, 444–456.

Shen, J., Chan, T.F., 2002. Mathematical models for local nontexture inpaintings. SIAM J. Appl. Math. 62, 1019–1043.

Tang, C., Hu, X., Chen, L., 2014. 2014. sample-based image completion using structure synthesis. in: ieee international conference on image processing, icip.

Uhrig, J., Schneider, N., Schneider, L., 2017. Sparsity invariant CNNs. In: 2017 International Conference on 3D Vision, 3DV.

Wang, Z., Bovik, A.C., Sheikh, H.R., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13.

Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J., 2018. Image inpainting via generative multi-column convolutional neural networks. Adv. Neural Inf. Process. Syst. 331–340.

Woo, S., Park, J., Lee, J.Y., 2018. CBAM: Convolutional block attention module. In: European Conference on Computer Vision, ECCV, pp. 3–19.

Xiao, X., Daneshpanah, M., Javidi, B., 2012. Occlusion removal using depth mapping in three-dimensional integral imaging. J. Disp. Technol. 8 (8), 483–490.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., 2015a. Show, attend and tell: Neural image caption generation with visual attention, In: International Conference on Machine Learning, ICML.

Xu, B., Wang, N., Chen, T., 2015b. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853.

Yan, Z., Li, X., Li, M., Zuo, W., Shan, S., 2018. Shift-net: Image inpainting via deep feature rearrangement. In: European Conference on Computer Vision, ECCV.

Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H., 2017. High-Resolution image inpainting using multi-scale neural patch synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 6721–6729.

Yeh, R.A., Chen, C., Lim, T.Y., 2017. Semantic image inpainting with deep generative models. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 5485–5493.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018. Generative image inpainting with contextual attention. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 5505–5514.

Zhang, H., Dong, Y., Fan, Q., 2017. Wavelet frame based Poisson noise removal and image deblurring. Signal Process. 363–372.

Zhang, J., Zhao, D., Xiong, R., 2014. Image restoration using joint statistical modeling in a space-transform domain. IEEE Trans. Circuits Syst. Video Technol. 24, 915–928.

Zheng, C., Cham, T.J., Cai, J., 2019. Pluralistic image completion. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR.

Zhou, B., Lapedriza, A., Khosla, A., 2018. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40, 1452–1464.

Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L., 2016. Visual7w: Grounded question answering in images. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR.