Confidence-Based Global Attention Guided Network for Image Inpainting

Zhilin Huang, Chujun Qin, Lei Li, Ruixin Liu, and Yuesheng Zhu(✉)

School of Electronic and Computer Engineering, Peking University, Shenzhen, China
{zerinhwang03,chujun.qin,csleili,anne_xin,zhuys}@pku.edu.cn

Abstract. Most of recent generative image inpainting methods have shown promising performance by adopting attention mechanisms to fill hole regions with known-region features. However, these methods tend to neglect the impact of reliable hole-region information, which leads to discontinuities in structure and texture of final results. Besides, they always fail to predict plausible contents with realistic details in hole regions due to the ineffectiveness of vanilla decoder in capturing long-range information at each level. To handle these problems, we propose a confidence-based global attention guided network (CGAG-Net) consisting of coarse and fine steps, where each step is built upon the encoder-decoder architecture. CGAG-Net utilizes reliable global information to missing contents through an attention mechanism, and uses attention scores learned from high-level features to guide the reconstruction of low-level features. Specifically, we propose a confidence-based global attention layer (CGA) embedded in the encoder to fill hole regions with reliable global features weighted by learned attention scores, where reliability of features is measured by automatically generated confidence values. Meanwhile, the attention scores learned by CGA are repeatedly used to guide the feature prediction at each level of the attention guided decoder (AG Decoder) we proposed. Thus, AG Decoder can obtain semantically-coherent and texture-coherent features from global regions to predict missing contents. Extensive experiments on Paris StreetView and CelebA datasets validate the superiority of our proposed approach through quantitative and qualitative comparisons with existing methods.

Keywords: Image inpainting · Encoder-decoder · Attention mechanism

1 Introduction

Image inpainting is a task of restoring the missing or damaged parts of images in computer vision. In practice, many image inpainting approaches have been proposed in wide application ranges, such as photo editing, image-based rendering, etc. The main challenge of image inpainting is to generate semantically plausible and visually realistic results for missing regions [27].

J. Lokoč et al. (Eds.): MMM 2021, LNCS 12572, pp. 200–212, 2021.
https://doi.org/10.1007/978-3-030-67832-6_17

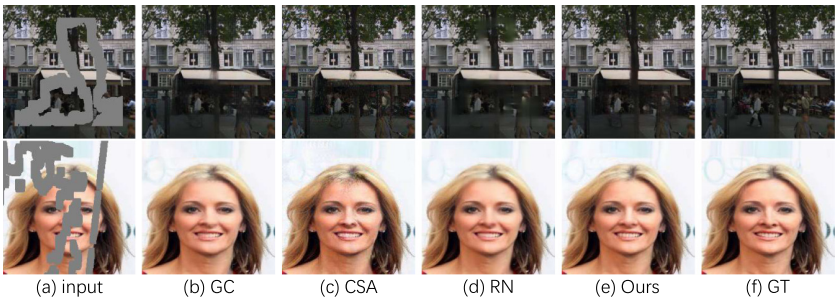

Fig. 1. Qualitative comparisons of inpainting results by Gated Conv (GC) [28], Coherent Semantic Attention (CSA) [13], Region Normalization (RN) [29], our model and ground truth (GT). [Best viewed with zoom-in.]

Traditionally, this task is settled with diffusion-based or patch-based approaches [1,4,17,30]. These methods only work well for stationary textural regions, and they fail to generate semantic information on non-stationary images. To make up for it, early learning-based methods [8,18,26] are proposed to formulate inpainting as a conditional image generation problem by using convolutional encoder-decoder network, where the encoder learns a latent feature representation of the image and the decoder reasons about the missing contents [18]. Unfortunately, these methods often create boundary artifacts and blurry results inconsistent with known regions. Recently, some approaches [21,27,28] adopt spatial attention mechanisms in encoder to effectively encode the latent representation by fully utilizing long-range contextual information. Firstly, they extract patches in known regions and hole regions of the high-level feature map, and then take known-region patches as references to calculate attention scores with hole-region patches. Finally, they fill the hole regions with known-region patches weighted by the attention scores.

However, most existing attention-based image inpainting methods [20,21,27] tend to completely ignore the impact of holes features which may be not well-inferred, or just model the correlation between adjacent hole-region features in a certain direction [13], which leads to discontinuous structures and textures in final results, as shown in the first row of Fig. 1. Moreover, due to limited size of the vanilla convolutional kernel and receptive field, they cannot effectively utilize distant information at each level of the vanilla decoder. Thus, they always fail to reason about realistic details in hole regions, as shown in the second row of the Fig. 1.

To handle these problems, we propose a confidence-based global attention guided network (CGAG-Net) which divides the inpainting task into coarse and fine steps as shown in Fig. 2(a). In the coarse step, a simple dilated convolutional network [27] generates preliminary results for the next step. And in the fine step, a confidence-based global attention layer (CGA) we proposed is applied to the high-level features of the encoder to reconstruct semantic continuous features in the hole regions by taking feature patches from both holes and known regions

as references. Considering the fact that indiscriminately model the correlation within the hole regions to reconstruct missing contents will introduce unreliable (i.e., poorly-inferred) information and results in blurriness, CGA automatically generates confidence values to measure the reliability of information for each channel at each spatial location of reference patches. The confidence values are able to highlight reliable information and suppress unreliable one.

In addition, we propose an attention guided decoder (AG Decoder) to fill hole regions from high-level to low-level by repeatedly applying a guided attention module (GA) we proposed to the decoder. Since the attention scores learned from high-level features reflect the correlation of spatial location between semantically-coherent features, they can be taken as the guidance of the attention mechanism to fill hole regions of low-level features with semantically-coherent and texture-coherent patches. By using the attention scores learned from high-level feature map to guide GA at shallow layers of the AG Decoder, our model can generate both semantically and visually plausible results. Furthermore, we propose a multi-scale gated block (MSGB) embedded in the encoder to capture valid information at various scales by adopting multiple gated convolutions [28] with different kernel sizes and connecting them in a hierarchical style. Extensive experiments on standard datasets Paris StreetView [3] and CelebA [15] demonstrate that the proposed approach can generate higher-quality inpainting results in irregular holes than existing methods.

The main contributions of this paper are summarized as follows:

- We propose a confidence-based global attention layer (CGA) to consider the impact of reliable global features on the reconstruction of missing contents, according to the automatically generated confidence values which can highlight reliable information of features and suppress unreliable one.
- An attention guided decoder (AG Decoder) is proposed to fill hole regions at each level with semantically-coherent and texture-coherent features under the guidance of attention scores from CGA.
- MSGB is designed to capture information at various scales by adopting multiple gated convolutions with different kernel sizes and connecting them in a hierarchical style.

2 Related Work

2.1 Learning-Based Image Inpainting

Learning-based methods for image inpainting [2,7,11,14,16,22,24] always use deep learning and adversarial training strategy [6] to predict the missing contents in hole regions. One of the early learning-based methods, Context Encoder [18] takes adversarial training into a encoder-decoder architecture to fill the holes in feature-level. On the basis of Context Encoder, Iizuka et al. [8] propose global and local discriminators to generate better results with regard to overall consistency as well as more detail. Yang et al. [26] propose a multi-scale neural patch synthesis approach to generate high-frequency details. Liu et al. [12] propose an automatic mask generation and update mechanism to focus on valid pixels in

the feature map for better results. Inspired by [12], Yu et al. [28] propose a gated convolution and SN-PatchGAN to better deal with irregular masks.

2.2 Attention-Based Image Inpainting

Recently, spatial attention mechanism is introduced in image inpainting task to model long-range dependencies within features [19,23]. Yan et al. [25] introduce a shift operation and a guidance loss to restore features in the decoder by utilizing the information in corresponding encoder layers. Yu et al. [27] propose a novel contextual attention layer to explicitly utilize the feature in known-regions as references to make better predictions. Liu et al. [13] propose a coherent semantic attention layer to model the correlation between adjacency features in hole regions for continuity results.

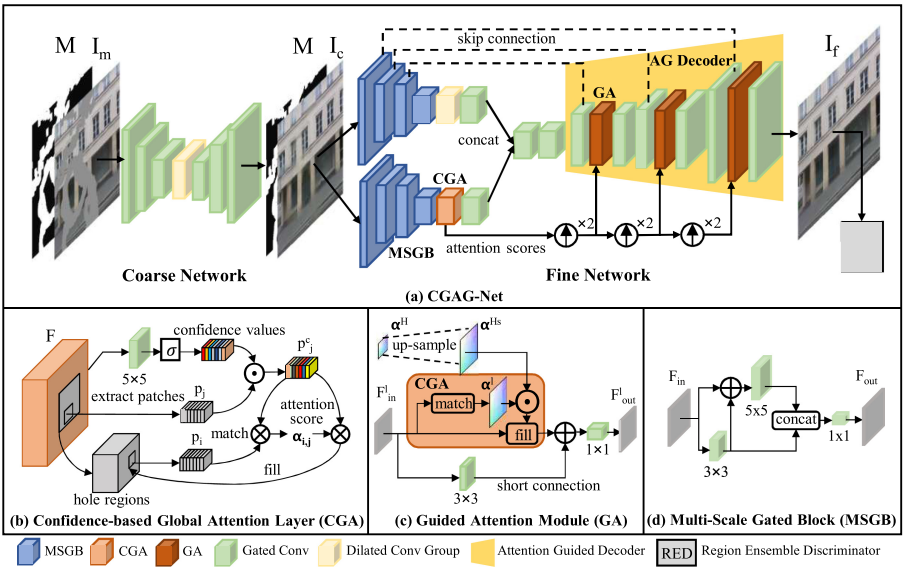

Fig. 2. The architecture of our proposed confidence-based global attention guided network (CGAG-Net).

3 Approach

3.1 Overview

Our model confidence-based global attention guided network (CGAG-Net) divides image inpainting task into coarse and fine steps, where each step is built upon the encoder-decoder architecture, as shown in Fig. 2(a). In the coarse step, we adopt the same structure of the coarse network in [27]. The coarse network

takes the concatenation of the masked image I_m and the binary mask M as input to generate the coarse prediction I_c. In the fine step, we take the concatenation of I_c and M as input of the fine network to obtain the finer result I_f. The fine network consists of two parallel encoders and an attention guided decoder (AG Decoder). The multi-scale gated block (MSGB) is embedded in each layer of two encoders to capture the information at different scales. The top encoder focuses on hallucinating contents with a dilated convolution group (i.e., stacked dilated convolutions). And the confidence-based global attention layer (CGA) is embedded in the deepest layer of the bottom encoder, which enables the encoder to reconstruct the semantically continuous contents with reliable patches from the global. Then, the output from two encoders are fused together and fed into the attention guided decoder (AG Decoder). The AG Decoder repeatedly uses a guided attention module (GA) to reconstruct hole-region features from high-level to low-level. Meanwhile, the attention scores learned in CGA is up-sampled to the corresponding resolution to guide the filling process of GA at each level of the AG Decoder. In addition, skip connections [13] are introduced to concatenate features from each layer of the top encoder and corresponding layers of AG Decoder. Finally, the region ensemble discriminator (RED) proposed in [19] is introduced to act as global and local discriminator simultaneously.

3.2 Confidence-Based Global Attention Layer

In order to model the correlation between hole-region features and avoid introducing unreliable information into hole regions, we propose a confidence-based global attention layer (CGA), as shown in Fig. 2(b).

CGA first takes the feature map F as input of a gated convolution [28] with an activation function, and then sigmoid function is applied on the output of the gated convolution to get confidence values in a confidence map C for each channel at each spatial location of F. The confidence values in C are between 0 and 1 (0 represents completely unreliable, and vice versa). The next, F^c is obtained by computing element-wise multiplication between C and F:

$$C_{x,y} = \sigma(\phi(W_c(F))) \tag{1}$$

$$F^c_{x,y} = C_{x,y} \odot F_{x,y} \tag{2}$$

where W_c denotes the convolutional filters, σ denotes the sigmoid function, $\odot$ denotes the element-wise multiplication and ϕ can be any activation functions. After extracting patches p and p^c from hole regions of F and the global (both hole regions and known regions) of F^c respectively, CGA takes patches p^c as references to calculate the cosine similarity with p:

$$s_{i,j} = < \frac{p_i}{||p_i||_2}, \frac{p^c_j}{||p^c_j||_2} > \tag{3}$$

where p_i and p_j^c are the i-th patch and the j-th patch of p and p^c respectively. Finally, the softmax is applied on the channel of similarities to obtain attention scores α and then CGA fills hole regions with patches p^c weighted by α:

$$\alpha_{i,j} = softmax(\frac{exp(s_{i,j})}{\sum_{i=1}^{N} exp(s_{i,j})}) \tag{4}$$

$$p_i = \sum_{i}^{N} \alpha_{i,j} \cdot p_j^c \tag{5}$$

Compared with existing attention mechanisms for image inpainting [21,25,27,28], our CGA additionally considers the impact of reliable information in whole hole regions to generate continuous results. The confidence values generated by CGA in an adaptive manner are able to highlight the reliable information for each channel at each spatial location of the feature map and suppress unreliable one. It is worth noting that confidence values are only applied on reference patches. In this way, our CGA can avoid the situation that a hole-region patch always have a large attention score with itself when CAG additionally takes hole-region patches as references. And the generalization ability of learned attention scores is enhanced simultaneously. The contextual attention layer proposed in [27] can be regarded as a special case of our CGA, where confidence values for each channel of features are 0 in hole regions and 1 in known regions.

3.3 Attention Guided Decoder

In order to generate semantically and texture plausible results, we propose an attention guided decoder (AG Decoder). Under the guidance of the attention scores learned from high-level features, a guided attention module (GA) is repeatedly applied on features at each level of AG Decoder to reconstruct the missing contents with semantically-coherent and texture-coherent information. The GA consists of CGA and a short connection which can ease the flow of information and stabilize the training process, as shown in Fig. 2(c).

In the l-th layer of AG Decoder, GA first obtains the attention score $\alpha_{i,j}^l$ between the patch pair, p_i^l and p_j^{cl}, by taking the same strategy as mentioned in Sect. 3.2, where p_i^l is the i-th hole-region patch and p_j^{cl} is the j-th reference patch. Furthermore, in order to maintain the semantic coherency between generated textures and surroundings, we use the attention map α^H learned from high-level features by CGA to guide the hole filling process of GA at each level of the AG Decoder, where α^H reflects the correlation of spatial location between semantically-coherent features. Thus, α^H is up-sampled to the corresponding resolution with scale factor s to obtain the up-sampled attention score map α^{Hs}. After that, softmax is applied on the result of the element-wise multiplication between α^{Hs} and α^l to get the guided attention score map α^{Gl}. In this way, only elements with high values in both α^{Hs} and α^l will have high values in α^{Gl}. That is to say, only if two patches in a patch pair have both high semantic and textural

coherency, can they obtain a high attention score in α^{Gl}. Finally, we reconstruct hole regions with p^{cl} weighted by α^{Gl}. The process can be formulated as follows:

$$\alpha^{Gl} = softmax(\alpha^{Hs} \odot \alpha^{l}) \tag{6}$$

$$p_i^l = \sum_{i=1}^{N} \alpha_{i,j}^{Gl} \cdot p_j^{cl} \tag{7}$$

3.4 Multi-scale Gated Block

Extracting features at different scales is essential for CNN models to capture important contextual information. Inspired by Res2Net [5] and Gated Conv [28], we propose multi-scale gated block (MSGB), as shown in Fig. 2(d), to extract valid features at various scales by adopting multiple gated convolutions with different kernel sizes and connecting them in a hierarchical style. The gated convolution proposed in [28] can distinguish valid pixels/features from invalid ones, thereby preventing predicted results from being affected by harmful information.

Let F_{in} and F_{out} be the input and the output feature map of MSGB, $GC_{i\times i}(\cdot)$ be the gated convolution [28] with kernel size i. MSGB first extracts features with a 3×3 gated convolution in the input feature map F_{in} to get the output $F_{3\times3}$. Instead of simply fusing features at different scales, MSGB uses element-wise sum operation between $F_{3\times3}$ and F_{in} before feeding F_{in} into a 5×5 gated convolution. After using a 1×1 gated convolution to reduce channels of the concatenation of $F_{3\times3}$ and $F_{5\times5}$, MSGB fuses information at different scales to obtain the output F_{out}. The process can be formulated as follows:

$$F_{3\times3} = \phi(GC_{3\times3}(F_{in})) \tag{8}$$

$$F_{5\times5} = \phi(GC_{5\times5}(F_{3\times3} + F_{in})) \tag{9}$$

$$F_{out} = \phi(GC_{1\times1}(concat([F_{3\times3}, F_{5\times5}]))) \tag{10}$$

where $\phi(\cdot)$ denotes the activation function. We select LeakyReLU as activation function in our experiments.

Compared with simply fusing multi-scale information in a parallel style, our MSGB can obtain larger receptive fields without using extra parameters. Specifically, when we take $F_{3\times3}$ as the input of a 5×5 gated convolution, the output will have a larger receptive field than the output obtained by taking F_{in} as the input of the 5×5 gated convolution due to the connection explosion effect [5].

3.5 Loss Function

To make constrains that the output of the coarse network and the fine network should approximate the ground-truth image, following [27], we use L1 distance as our reconstruction loss L_{rec}. Besides, We adopt region ensemble discriminator (RED) [19] as global and local discriminator to calculate the adversarial loss L_{adv} in each pixel individually, which drives our model to handle various holes with

arbitrary shapes and generate visually pleasing results. To address the gradient vanishing problem in generator, we employ the hinge version of the adversarial loss [13]. Moreover, we use the perceptual loss [9] L_{per} to generate plausible contents by measuring the similarity between high-level structure.

In summary, the overall loss function of the proposed CGAG-Net is as follows:

$$L_{total} = \lambda_{rec} \cdot L_{rec} + \lambda_{adv} \cdot L_{adv} + \lambda_{per} \cdot L_{per} \tag{11}$$

where λ_{rec}, λ_{adv}, λ_{per} are hyper-parameters for the reconstruction, adversarial and perceptual losses.

4 Experiments

4.1 Experiment Settings

We evaluate our model on two datasets: Paris StreetView [3] and CelebA [15]. For these two datasets, we use the original train, validation and test splits. And we obtain irregular masks which are classified based on different hole-to-image area ratios from Partial Conv [12]. The training and testing process are conducted on masks with 20%–40% hole-to-image area ratio. Besides, we follow [28] to adopt the same data augmentation such as flipping during training process. Our model is optimized by Adam algorithm [10] with learning rate of 1×10^{-4} and $\beta 1 = 0.5$. The hyper-parameters are set as $\lambda_{rec} = 1.0$, $\lambda_{per} = 1.0$, $\lambda_{adv} = 0.01$. And we train on an Nvidia Titan X Pascal GPU with a batch size of 1. All masks and images for training and testing are with the size of 256×256.

We compare our method with five methods: Partial Conv (PC) [12], Contextual Attention (CA) [27], Gated Conv (GC) [28], Coherent Semantic Attention (CSA) [13] and Region Normalization (RN) [29].

4.2 Qualitative Comparisons

Figure 3 present inpainting results of different methods on testing images from Paris StreetView and CelebA datasets. For all methods, no post-processing step is performed to ensure fairness. As shown in Fig. 3, PC, CA and RN are effective in generate semantically plausible results, but the results present distorted structures and lack realistic details. Compared with previous methods, GC and CSA can generate richer details, but the results still have discontinuous textures and boundary artifacts. This is mainly because they neglect the impact of hole-region features and the ineffectiveness of vanilla decoder in capturing distant information of low-level features. Compared with these methods, our model is able to generate semantically and visually plausible results with clear boundaries and continuous textures in hole regions.

Fig. 3. Example cases of qualitative comparison on the Paris StreetView and CelebA datasets. [Best viewed with zoom-in.]

Table 1. Quantitative comparison results over Paris StreetView [3] and CelebA [15] datasets with irregular masks between PC [12], CA [27], GC [28], CSA [13], RN [29] and Ours. $^{-}$ Lower is better. $^{+}$ Higher is better.

Datasets	Paris StreetView						CelebA					
Methods	PC	CA	GC	CSA	RN	Ours	PC	CA	GC	CSA	RN	Ours
MAE^{-}	3.884	3.386	3.283	3.245	3.491	**3.143**	3.307	3.144	2.887	3.060	2.849	**2.709**
SSIM^{+}	0.879	0.901	0.903	0.904	0.892	**0.906**	0.909	0.917	0.921	0.925	**0.927**	**0.927**
PSNR^{+}	27.981	28.810	29.082	29.112	28.691	**29.309**	28.131	28.470	29.057	29.427	29.448	**29.603**

4.3 Quantitative Comparisons

We use images from the testing set of Paris StreetView and CelebA datasets with irregular masks to make comparisons. We take MAE, PSNR, SSIM as evaluation metrics to quantify the performance of models. Table 1 lists the comparison results which present our method outperforms all other methods in these measurements on both Paris StreetView and CelebA datasets.

Table 2. Quantitative comparisons over Paris StreetView between CA [27] and CGA. $^{-}$ Lower is better. $^{+}$ Higher is better.

	MAE^{-}	SSIM^{+}	PSNR^{+}
With CA	3.283	0.903	29.082
With CGA (all 1)	3.345	0.903	29.077
With CGA	**3.271**	**0.904**	**29.154**

4.4 Ablation Study

Effect of CGA. In order to demonstrate the effect of our CGA, we adopt the architecture of Gated Conv [28] and replace the contextual attention layer (CA) [27] with CGA to make both qualitative and quantitative comparisons on the Paris StreetView testing set. Also, to validate the effect of confidence values, we set all confidence values in CGA as 1 to make a comparison. As present in Tab 2, by adopting CGA we proposed, the mode can achieve the best performance in all metrics. As shown in the areas marked with red bounding boxes in Fig. 4, CA fails to generate continuous structures and textures in hole regions. And directly modeling the correlation (all confidence values are set to 1) between hole-region features in CGA will cause blurriness. By adopting our CGA with automatically generated confidence values which can highlight reliable information of hole-region features and suppress unreliable one, the model is able to generate continuous structures and textures in hole regions.

Fig. 4. The effect of CGA. [Best viewed with zoom-in.]

Effect of AG Decoder. We make a comparison on the Paris StreetView testing set to demonstrate the effect of our AG Decoder. Figure 5 presents that the model can generate semantically plausible results but contain blurriness, when we replace the AG Decoder in our model with vanilla decoder (without using attention mechanisms). Without the guidance of attention scores learned from high-level features, AG Decoder fails to generate textures consistent with surroundings. When we adopt AG Decoder under the guidance of attention scores learned from high-level features, our model can generate semantically and visually plausible results.

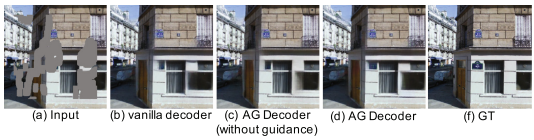

Fig. 5. The effect of AG Decoder. [Best viewed with zoom-in.]

Effect of MSGB. To verify the effect of MSGB, we replace the gated convolutions in the encoder of GC [28] with two types of MSGB which connects multiple gated convolutions in different styles (hierarchical and parallel) to make a comparison. For fair comparison, we keep the number of model parameters the same. As shown in Table 3, when adopting MSGB with gated convolutions connected in a hierarchical style, the model can obtain the best performance in all metrics.

Table 3. Quantitative comparisons between Gated Conv [28] and MSGB on CelebA dataset. $^{-}$ Lower is better. $^{+}$ Higher is better.

	MAE^{-}	SSIM^{+}	PSNR^{+}
With Gated Conv	2.887	0.921	29.057
With MSGB (parallel)	2.868	**0.924**	29.251
With MSGB (hierarchical)	**2.858**	**0.924**	**29.334**

5 Conclusion

In this paper, we propose a confidence-based global attention guided network (CGAG-Net) with two key components, a confidence-based global attention layer in the encoder and an attention guided decoder to synthesize missing contents in masked images. By measuring reliability of global features and predicting missing contents at each level of the attention guided decoder with semantically-coherent and texture-coherent features, our CGAG-Net can generate semantically and visually plausible results with continuous structures and textures. Extensive experiments on different datasets demonstrate that our methods can significantly outperforms other state-of-the-art approaches in image inpainting.

Acknowledgement. This work was supported in part by the Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing), and in part by the Key-Area Research and Development Program of Guangdong Province (2019B010137001).

References

1. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
2. Demir, U., Ünal, G.B.: Patch-based image inpainting with generative adversarial networks. CoRR abs/1803.07422 (2018)
3. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes Paris look like Paris? Commun. ACM **58**(12), 103–110 (2015)
4. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: ICCV, pp. 1033–1038. IEEE Computer Society (1999)
5. Gao, S., Cheng, M., Zhao, K., Zhang, X., Yang, M., Torr, P.H.S.: Res2Net: a new multi-scale backbone architecture. CoRR abs/1904.01169 (2019)
6. Goodfellow, I.J., et al.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)
7. Han, X., Wu, Z., Huang, W., Scott, M.R., Davis, L.: FiNet: compatible and diverse fashion image inpainting. In: ICCV, pp. 4480–4490. IEEE (2019)
8. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graph. **36**(4), 107:1–107:14 (2017)
9. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (Poster) (2015)
11. Liao, L., Hu, R., Xiao, J., Wang, Z.: Edge-aware context encoder for image inpainting. In: ICASSP, pp. 3156–3160. IEEE (2018)
12. Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 89–105. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_6
13. Liu, H., Jiang, B., Xiao, Y., Yang, C.: Coherent semantic attention for image inpainting. In: ICCV, pp. 4169–4178. IEEE (2019)
14. Liu, S., Guo, Z., Chen, J., Yu, T., Chen, Z.: Interleaved zooming network for image inpainting. In: ICME Workshops, pp. 673–678. IEEE (2019)
15. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV, pp. 3730–3738. IEEE Computer Society (2015)
16. Ma, Y., Liu, X., Bai, S., Wang, L., He, D., Liu, A.: Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation. In: IJCAI, pp. 3123–3129 (2019). ijcai.org
17. Newson, A., Almansa, A., Gousseau, Y., Pérez, P.: Non-local patch-based image inpainting. Image Process. Line **7**, 373–385 (2017)
18. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: CVPR, pp. 2536–2544. IEEE Computer Society (2016)
19. Shin, Y., Sagong, M., Yeo, Y., Kim, S., Ko, S.: PEPSI++: fast and lightweight network for image inpainting. CoRR abs/1905.09010 (2019)
20. Song, Y., et al.: Contextual-based image inpainting: infer, match, and translate. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 3–18. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_1

21. Wang, N., Li, J., Zhang, L., Du, B.: MUSICAL: multi-scale image contextual attention learning for inpainting. In: IJCAI, pp. 3748–3754 (2019). ijcai.org
22. Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J.: Image inpainting via generative multi-column convolutional neural networks. In: NeurIPS, pp. 329–338 (2018)
23. Xie, C., et al.: Image inpainting with learnable bidirectional attention maps. In: ICCV, pp. 8857–8866. IEEE (2019)
24. Xiong, W., et al.: Foreground-aware image inpainting. In: CVPR, pp. 5840–5848. Computer Vision Foundation/IEEE (2019)
25. Yan, Z., Li, X., Li, M., Zuo, W., Shan, S.: Shift-Net: image inpainting via deep feature rearrangement. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_1
26. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: CVPR, pp. 4076–4084. IEEE Computer Society (2017)
27. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: CVPR, pp. 5505–5514. IEEE Computer Society (2018)
28. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: ICCV, pp. 4470–4479. IEEE (2019)
29. Yu, T., et al.: Region normalization for image inpainting. In: AAAI, pp. 12733–12740. AAAI Press (2020)
30. Zhang, Q., Lin, J.: Exemplar-based image inpainting using color distribution analysis. J. Inf. Sci. Eng. **28**(4), 641–654 (2012)