Image Outpainting: Hallucinating Beyond the Image

QINGGUO XIAO\textsuperscript{1}, GUANGYAO LI\textsuperscript{1}, AND QIAOCHUAN CHEN\textsuperscript{2}

\textsuperscript{1}College of Electronics and Information Engineering, Tongji University, Shanghai 200092, China
\textsuperscript{2}School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China

Corresponding authors: Guangyao Li (lgy423@126.com) and Qingguo Xiao (qingguoxiao@126.com)

This work was supported by the National Nature Science Foundation of China (NSFC) under Grant 61771346.

\section*{ABSTRACT}
Image inpainting is a technique that aims to fill in the missing regions with visually plausible content. However, an opposite idea, which is painting outside images, receives little work. In this study, we investigate the issue of image outpainting. Considering that the model needs better prediction ability as there is less neighboring information in image outpainting, the study proposes a novel image outpainting architecture that can obtain both deep model performance and detailed information. To fully take advantage of residual learning, dense residual (DR) learning is proposed and the image generative network is built on DR. To avoid losing subtle information caused by downsampling in encoder-decoder, shortcuts are added for transferring previous knowledge. Different from vanilla U-Net, we propose a skip method of the semi-complete form. Experimental results show that the proposed method achieves excellent performance.

\section*{INDEX TERMS}
Image outpainting, residual learning, adversarial learning, generation model.

\section*{I. INTRODUCTION}

Image inpainting allows removing distracting objects in photos and repairing flawed regions or defects of digital images in a reasonable manner \cite{1}, \cite{2}. It has been widely investigated in the applications of digital effect (e.g., object removal), image restoration (e.g., scratch or text removal in a photograph), image coding and transmission (e.g., recovery of the missing blocks). For today’s consumer-level multi-megapixel cameras, image inpainting is still a challenging problem due to the inherent ambiguity and complexity of natural images.

Image inpainting has been studied for a long time. Early exemplar-based methods sample the pixels from a known region of the image and copy them to a damaged region. These methods have a limitation. It is that the synthesized textures only come from the input image. This will be an obvious problem when convincing completion requires textures that are not found in the input image.

Deep learning tools have shown exciting promise for kinds of tasks. This leads to another orientation for image inpainting. By analogy with auto-encoders, Pathak et al. \cite{3} presented an unsupervised visual feature learning algorithm driven by context-based pixel prediction. Yeh et al. \cite{4} investigated semantic inpainting and achieved pixel-level photorealism for face completion using Generative Adversarial Network (GAN). Iizuka et al. \cite{5} presented a novel generative adversarial network that has two discriminators for both locally and globally consistent image completion. Yu et al. \cite{6} introduced a new generative model to not only synthesize novel image structures but also explicitly utilize surrounding image features as references to make predictions. As some methods are specialized to inpaint images of regular holes, Liu et al. \cite{7} proposed partial convolutions to better deal with images with irregular holes. Image inpainting by deep learning is widely researched at present.

However, the relation to its cousin, which is image outpainting, receives little work. Different from those previous researches, the goal of painting outside images is to generate contents on the peripheral regions of images. We call this issue image outpainting. Given an image, we want to synthesize the contents on its outer boundaries. It can be seen as an image extrapolation problem. The idea is basically opposite to image inpainting or hole-filling problem where the missing regions are generated based on its surrounding context. In the inpainting task, there is a large amount of surrounding context for representing features to finish the completion. However, painting outside requires extrapolation to unknown areas with less neighboring information. As shown...
in Figure 1, the regions indicated with a red rectangle are relatively hard to generate preferable extrapolations. The only available knowledge is from the adjacent generated contents. Except for that, they have no other reference information. Compared with inpainting, image outpainting is much more challenging.

To tackle this challenging issue, we believe that the ability of model prediction should be powerful as it is stated above that image outpainting is more troublesome. Previous work demonstrates that deeper neural networks bring higher model performance. Residual learning has been demonstrated as a powerful tool for designing deep models.

In this work, we propose dense residual learning to take advantage of residual learning to an extreme. It is an evolutionary form of residual learning. We build the extrapolation network with dense residual learning. Encoder-decoder is a baseline for image-to-image translation models. We build on this architecture to design our generative network. Different from conventional encoder-decoder, the designed generative network is built with dense residual neural layers. The depth of the model is deeper. Subtle information is lost in encoder-decoder as there are consecutive downsamples. Ronneberger et al. [8] proposed the architecture of U-Net. They used skip connections for transferring previous knowledge. Isola et al. [9] demonstrated that adding skip connections to an encoder-decoder to create a “U-Net” results in higher quality results. However, standard U-Net is not appropriate for our image-to-image synthesis pipeline. This is because our image-to-image generative network is deep form, skip connecting all layers of deep neural networks will cause enormous parameters, feature redundancy, and a huge increase in computation.

To address this problem, we propose the framework of Semi U-Net. In Semi U-Net, skip shortcuts are only added between the early layers where feature maps are down-sampled in the encoder and the corresponding deconvolutional layers in the decoder. The feature blocks passed by the skip connections carry much detail information. Copying and combining the paired layers’ outputs make the feature maps have low- and high-frequency information. The middle bottleneck is without skip connections. It is built with dense residual learning. Our semi-skip connecting strategy is simple but useful. The model performance is improved without bringing large amounts of parameters and high computation cost.

The contribution of this study is threefold: (1) A novel image outpainting network that obtains deep model performance and high-quality texture is proposed. (2) Dense residual learning is proposed for more high-level feature representation and better model performance. (3) We show that our method can well tackle the issue of image outpainting.

II. RELATED WORK

A. IMAGE INPAINTING

1) IMAGE INPAINTING BY PATCH-BASED METHODS

Efros and Leung [10] first proposed patch-based method. They used a non-parametric approach for texture synthesis, which grows a new image outward from an initial seed, one pixel at a time. Criminisi et al. [11] proposed to fill the unknown areas by searching the most closely matching patches in the rest of the image. In the following years, some methods improved the completion results by modifying the patch priority and search strategy [12]–[14]. Some methods regard this problem as a patch-based optimization problem with an energy function. Thus, the image completion problem can be solved using existing optimization methods such as the deterministic Expectation-Maximization (EM)-like scheme and Markov Random Field (MRF) models. Barnes et al. [15] proposed a novel randomized algorithm to quickly find the approximate nearest neighbor matches by propagating the neighborhood information using neighbor patches. Darabi et al. [1] used patch transformations, e.g., rotation, scaling, and affine transformations to extend the patch search space. However, these methods lack a high-level understanding of images, they are hard to generate semantic and novel results.

2) DEEP GENERATIVE MODELS FOR IMAGE INPAINTING

Deep generative models for image inpainting fill missing regions at the feature-level. They usually encode an image into a latent feature and then decode the feature back into an image. By using Generative Adversarial Network (GAN), Yeh et al. [4] investigated semantic inpainting and achieved pixel-level photorealism for face completion. Iizuka et al. [5] presented a novel generative adversarial network that has two discriminators for both locally and globally consistent image completion. Yu et al. [6] introduced an attention model
to utilize surrounding image features as references to make predictions. Liu et al. [7] proposed partial convolutions to better deal with images with irregular holes. Song et al. [16] divided the task into inference and translation, and used simple heuristics to guide the propagation of local textures from the boundary to the hole. There have been many image inpainting deep generative models.

B. IMAGE OUTPAINTING

There is limited work about image outpainting. Wang et al. [17] used a data-driven approach and combined it with a graph representation of the source image. They treated images as graphs to find candidates for image extrapolation from a large-scale image library. This approach is a way of image patch matching. Indeed, for such an approach, extrapolation is easier than inpainting, since there are less to match at the boundary. The results depend on sizable databases and image retrieval technology. Recent learning-based research is confined to low resolution, the image input size is not bigger than 128 × 128 [18], [19]. The input in [19] is also heavily predefined. Sabini et al. [20] attempted to tackle this challenging problem by using GAN tools. They adversarially train convolutional neural networks to hallucinate past image boundaries, but the results are far from satisfactory.

III. METHODS

A. DENSE RESIDUAL BLOCK

The idea of Resnet [21] and Highway Networks is to create short paths from early layers to later layers. They reduce the generation of model degradation [22] by constructing an isomorphic mapping bypass. The identity mapping function in Resnet is constructed by:

\[ x_l = H_l(x_{l-1}) + x_{l-1} \]  

(1)

By learning the residual of \( H_l(x_{l-1}) \) to be near zero, the identity mapping can be got. It is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. To further alleviate the problem of gradient vanishing and model degradation, we propose dense residual learning. In dense residual learning, the \( l \) th layer receives the feature-maps of all preceding layers, in a feed-forward fashion \( x_0, \ldots, x_{l-1} \), as input:

\[ x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) + x_{l-1} \]  

(2)

It is different from that in Densenet [22]. In Densenet, \([x_0, x_1, \ldots, x_{l-1}]\) refers to the concatenation of the feature-maps. Ours follows residual forms yet. There is no feature concatenation. This kind of bypassing maximizes the superiority of residual learning. Figure 2 demonstrates the comparison of prior network structures and our dense residual block.

B. NETWORK STRUCTURE

The generator is used to extrapolate contents on the peripheral regions of the image. Figure 3 shows its overall architecture. The generator \( G \) follows the symmetric encoder-decoder baseline and it is specifically a dense residual network. The encoder stage is to learn image features. It is a process of down-sampling and characterizing images. The decoder stage is to decode previously learned features to desired images. It is a process of restoring and upscaling characterized maps. As stated above, much subtle information is inevitably lost during the down-sampling process. Conventional encoder-decoder architecture is not appropriate for the task of image inpainting and outpainting, which needs to keep the original detailed information of the image.

We make full use of the symmetry and add skip connections to hold high-frequency details and obtain more underlying information to generate more precise images. The skip connections are used to transfer previously learned knowledge and maintain fidelity with the input. Different from vanilla U-Net, our skip connections are optionally added between the convolutional layer \( i \) which precedes downsampling and the appropriate deconvolutional layer \( N - i - 1 \). The middle bottleneck is constructed by dense residual blocks. This generator can not only retain initial detailed information well but also possess the ability of high-level feature representation.
Generative Adversarial Networks (GAN) make the completed image natural as a whole and encourage the generated synthetic samples to be closer to the realistic image manifold. Figure 4 shows the architecture of the discriminator. The discriminator $D$ is used to estimate the probability that a sample came from the real data and distinguish whether an image is generated by a network or is real.

Inspired by [23] and [9], we conditioned the input of $D$ using $I_p$, which is an input image to be painted, to force $G(I_p)$ to be related with $I_p$. Our discriminator is a conditional one. The paired inputs are sent to $D$ to learn a mapping $D: (I_p, G(I_p)) \rightarrow 0; (I_p, I) \rightarrow 1$, where $I$ is the ground truth.

C. OBJECTIVE LOSS FUNCTION

During training, with regards to its context, the reconstruction loss is responsible for capturing the overall structure of the missing region. Given an input image to be painted, denoted as $I_p$, a binary mask $M$ corresponding to the input with a value of 1 and 0 for the dropped image regions, the network generation $I_p$ or $G(I_p)$, and the ground truth image $I$, we define the reconstruction loss:

$$L_{rec} = \| M \odot (I - G(I_p))\|_1 + M \odot (I - G(I_p))$$

(3)

where $\odot$ is the element-wise product operation, the first and second are respectively the inside and the non-outside pixels loss, $L_{outside}$ and $L_{outside}$.

The adversarial loss serves as a high-level loss penalty factor and trains the generator $G$ to capture the real-data distribution. By training $G$ and $D$ alternately, a generated image that is very similar to the real data could be obtained. It would be highly desirable if we can make the generated results look real. Mathematically, the objective loss based on GAN with conditions can be expressed as follows:

$$\min_{G} \max_{D} L_{GAN}(G, D)$$

(4)

$$E_{I_p}[\log(D(I_p, I))] + E_{I_p}[\log(1 - D(I_p, G(I_p)))]$$

Besides, we follow prior work on feature inversion [24], [25]. We make use of total variation regularizer $L_{TV}$ to encourage spatial smoothness in the output. Thus, the final objective function consists of three parts: reconstruction loss, adversarial loss, and total variation loss:

$$L = L_{rec} + \lambda_{cGAN}L_{cGAN} + \lambda_{TV}L_{TV}$$

(5)

IV. EXPERIMENTS AND ANALYSIS

We carry out experiments on Paris StreetView [26] and Places2 [27]. Paris StreetView consists of 14,900 images for training and 100 images for testing. Places2 is a large-scale dataset. There are 1.8 million train images. The experiments are conducted with multiple NVIDIA GeForce GTX-1080 Ti GPUs. The adopted deep learning platform is Pytorch. Adam [28] is utilized as the learning method to optimize the objective loss function. The optimization algorithm is based on adaptive estimates of lower-order moments. It is computationally efficient, has few memory requirements, and is well suited for problems that are large in terms of data or parameters. Formally, the Adam update rule can be expressed as follows:

$$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot m_t / \sqrt{v_t} + \varepsilon$$

(6)

where $m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot \nabla f(\theta_{t-1})$, $v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot (\nabla f(\theta_{t-1}))^2$, and $\alpha_t = \alpha \cdot \sqrt{1 - \beta_2^t}/(1 - \beta_1^t)$. $f(\theta)$ is our objective function with parameter $\theta$, $t$ denotes the timestep, and $\nabla f(\theta_{t-1})$ is the gradient at timestep $t$. The algorithm updates the exponential moving averages of the gradient $m_t$ and the squared gradient $v_t$, where the hyperparameters $\beta_1$ and $\beta_2$ are the exponential decay rates for the moment estimates and $\alpha$ is the step size. For optimization, we adopt the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. The learning rate is set 1e-4. The weighting parameters of $L_{rec}$, $L_{cGAN}$ and $L_{TV}$ are 1.0, 0.01 and 0.01 respectively.

A. RESULTS AND ANALYSIS

We use existing inpainting methods for carrying out the comparisons and compare with the following algorithms: Photoshop Content-Aware Fill (PatchMatch) [15], Image Melding (IM) [1], Context Encoder (CE) [3], U-Net [8], Pconv [7]. The first two are the nearest neighbor (NN) filling algorithms.
The latter three are deep learning approaches and we retrained their models for the task of image outpainting. Figure 5 shows the image outpainting results by our approach. It is observed that our approach can outpainting a wide diversity of scenes such as mountain ranges, rivers, and buildings. The results look natural even when the regions to be extrapolated are located in multiple directions.

Figure 6 shows the comparison results. It can be seen that the typical patch-based methods, PatchMatch and IM, can generate clear textures but with distorted structures inconsistent with surrounding areas. In the first row, it is obvious that distorted structures appear in the target region. Although their synthesized repeated textures look well, the generated contents are not semantic. The completed scenes are abnormal. The way of searching the approximate nearest neighbor patches in the known region to match the unknown region with the patches have limitations. It can not generate novel objects and semantic contents. Learning-based methods including CE, U-Net, and PConv tend to generate blurry textures in the final results. The results of CE is the most blurry. The results of U-Net and Pconv are better than CE, the generated textures are still blurry. Benefit from dense residual learning and skip connections, the proposed deep generative model is able to generate semantically-reasonable and visually realistic results with clear textures and consistent structures with semantic context. For better evaluation and comparison, more samples are randomly selected for evaluation and the results with PSNR and SSIM are computed.

Table 1 shows the average numerical comparison. We observe that the proposed method provides a better score than the others.

**B. EFFECT OF SKIP CONNECTIONS**

We also study the effect of skip connections. Two models with different architectures have been experimented. Figure 7 compares the proposed semi skip connection architecture against the encoder-decoder without skip connections.
TABLE 1. PSNR and SSIM comparison.

| Methods          | PatchMatch [15] | IM [1] | CE [3] | U-Net [8] | Pconv [7] | Ours  |
|------------------|-----------------|--------|--------|-----------|-----------|-------|
| PSNR             | 19.3152         | 20.1926| 19.9494| 23.4281   | 22.6222   | 24.1352|
| SSIM             | 0.7936          | 0.8122 | 0.6513 | 0.8484    | 0.8523    | 0.8671|

![FIGURE 8. Effect of bottleneck depth. (a) Input, (b) results with two dense residual blocks, (c) results with three dense residual blocks, (d) ground truth. The results of (c) are more preferable.](image)

![FIGURE 9. Image inpainting result generated by our approach.](image)

Figure 7 (b) are the results by conventional encoder-decoder and Figure 7 (c) are the results by semi skip connection architecture. For the convenience of comparison, we randomly select a region and zoom in to clearly show the outputs. For the results of conventional encoder-decoder, the abstractions of the image are learned, but many details are lost. The outputs of the pre-existing region are blurry. Whereas, the results of semi skip connection architecture hold exactitude. Compared with the architecture without skip connections, the proposed semi skip connection architecture extrapolates higher quality texture and reserves the initial information of the undamaged regions well.

Experiments show that skip connections are important for these tasks which have tougher requirements for high-quality texture and exactitude of the input such as image reconstruction, inpainting, and outpainting, as they provide knowledge including both low- and high-frequency information.

C. ANALYZE OF THE BOTTLENECK

The bottleneck of the proposed architecture is constructed using dense residual blocks, and they are without skip shortcuts. We find that deeper models bring better performance under certain conditions. Here we set two and three blocks for evaluation. One block is constructed with five dense residual convolution modules. A deeper bottleneck with more blocks can be trial with enough GPU rams. In Figure 8, we show the effect of bottleneck depth.
Figure 8 (b) are the results with two blocks, Figure 8 (c) are the results with three blocks, and Figure 8 (d) are the ground-truth images. We box out the extrapolation with a green rectangle in the output image. Figure 8 shows that although both models can not extrapolate the exact same content that was in the original image, the results are acceptable and the contents are realistic-looking. On the side of bottleneck depth, results indicate that a deeper bottleneck helps produce more preferable and consistent outputs. In Figure 8 (b), it can be seen that the extrapolated contexts are not complete and consistent. The window bars are desultory and crooked. Whereas the results in Figure 8 (c) are more semantic consistent. Moreover, the extrapolated textures in (c) are clearer than that in (b). We ascribe significant performance improvement to deeper and larger networks. The bottleneck depth can be built flexibly as the proposed semi skip connection architecture does not connect middle shortcuts. It can avoid the defects in conventional U-Net architecture.

D. EXTENSIVE EXPERIMENTS ON IMAGE INPAINTING

We applied our technique to image inpainting. Figure 9 gives the experimental results. It is seen that the missing regions are filled with visually reasonable content.

One of the applications of image inpainting is to be able to remove unwanted objects in the image. Figure 10 shows the simulation effect of the proposed algorithm on object removal. The object to be removed is marked with faint red color. As shown in the first example in Figure 10, we want to remove the text below, in the second example, the bird in the grass is to be removed. We also tested the effect of removing multiple targets simultaneously. As shown in the third example, we want to remove the person in the scene and the ship in the background all at the once. From the results, we observe that the images after object removal maintain consistency and continuity with the surrounding unbroken area. The results look realistic and natural.

E. DISCUSSION

There are two common failed cases for image outpainting. As stated above, there is less neighboring information. This leads to that feature far away from the border can not be well represented and the corresponding extrapolation is not as well as the near one. As shown in Figure 11, for a given source image (a), (b) is the outpainting results. We mark out a portion of the results using two red rectangles. It shows that the result of the right portion which is closer to the border is preferable. The feature of the left portion is not represented as well as the right. This is due to the nature of image outpainting. Another case is that the textures can not be so rich as that in the original image when extrapolating large regions. For a given source image as (c), (d) is the outpainting results, it can be seen that persons who are at the bottom right in the ground truth (Figure 11 (e)) are not reconstructed. The results are simple. Note that it is still acceptable as the goal is not to fully reconstruct the ground-truth but to obtain visually consistent extrapolations. We believe that better models and more samples can boost improvement.

V. CONCLUSION

In this study, we propose a novel image-to-image synthesis architecture to deal with a challenging issue - image outpainting. The designed model using this architecture generates high-quality results. The proposed dense residual learning achieves outstanding performance and realistic image extrapolations. In the future, we plan to investigate the application of image outpainting. Processing homography transformed images needs outpainting and slightly extending images or recursive outpainting to expand images can fulfill panorama construction.

REFERENCES

[1] S. Durabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, “Image melding: Combining inconsistent images using patch-based synthesis,” *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–82, 2012.

[2] K. He and J. Sun, “Image completion approaches using the statistics of similar patches,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2423–2435, Dec. 2014.

[3] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[4] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, “Semantic image inpainting with deep generative models,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5485–5493.
[5] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Globally and locally consistent image completion,” *ACM Trans. Graph.*, vol. 36, no. 4, p. 107, 2017.

[6] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, “Generative image inpainting with contextual attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[7] G. Liu, F. A. Reda, K. J. Shih, T. Wang, A. Tao, and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 89–105.

[8] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[10] A. A. Efros and T. K. Leung, “Texture synthesis by non-parametric sampling,” in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1033–1038.

[11] A. Criminisi, P. Perez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[12] Z. Xu and J. Sun, “Image inpainting by path propagation using patch sparsity,” *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1153–1165, May 2010.

[13] R. Martinez-Noriega, A. Roumy, and G. Blanchard, “Exemplar-based image inpainting: Fast priority and coherent nearest neighbor search,” in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Sep. 2012, pp. 1–6.

[14] V. Kumar, J. Mukhopadhyay, and S. K. D. Mandal, “Modified exemplar-based image inpainting via primal-dual optimization,” in *Proc. Int. Conf. Pattern Recognit. Mach. Intell.*, 2015, pp. 116–125.

[15] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, “PatchMatch: A randomized correspondence algorithm for structural image editing,” *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

[16] Y. Song, C. Yang, Z. Lin, X. Liu, Q. Huang, H. Li, and C. C. J. Kuo, “Contextual-based image inpainting: Infer, match, and translate,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–18.

[17] M. Wang, Y.-K. Lai, Y. Liang, R. R. Martin, and S.-M. Hu, “BiggerPicture: Data-driven image extrapolation using graph matching,” *ACM Trans. Graph.*, vol. 33, no. 6, pp. 1–13, Nov. 2014.

[18] B. Van Hoorick, “Image outpainting and harmonization using generative adversarial networks,” 2019, arXiv:1912.10960. [Online]. Available: https://arxiv.org/abs/1912.10960

[19] L. Zhang, J. Wang, and J. Shi, “Multimodal image outpainting with regularized normalized diversification,” 2019, arXiv:1910.11481. [Online]. Available: https://arxiv.org/abs/1910.11481

[20] M. Sabini and G. Rusak, “Painting outside the box: Image outpainting with GANS,” 2018, arXiv:1808.08483. [Online]. Available: https://arxiv.org/abs/1808.08483

[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[23] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” 2014, arXiv:1411.1784. [Online]. Available: http://arxiv.org/abs/1411.1784

[24] A. Mahendran and A. Vedaldi, “Understanding deep image representations by inverting them,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5188–5196.

[25] E. d’Angelo, A. Alahi, and P. Vanderheynst, “Beyond bits: Reconstructing images from local binary descriptors,” in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 935–938.

[26] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, “What makes paris look like paris?” *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–9, Aug. 2012.

[27] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[28] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

QINGGUO XIAO received the B.S. degree in communication engineering from Zhejiang Normal University, in 2012, and the M.S. degree in electronics and communication engineering from Ningbo University, in 2015. He is currently pursuing the Ph.D. degree with the College of Electronics and Information Engineering, Tongji University. His current research interests include deep learning, image generation, and image inpainting.

GUANGYAO LI received the B.S., M.S., and Ph.D. degrees from the Nanjing University of Aeronautics and Astronautics, China, in 1986, 1989, and 1997, respectively. He is currently a Professor and the Doctoral Supervisor with the College of Electrical and Information Engineering, Tongji University. His research interests include image processing and virtual reality.

QIAOCHUAN CHEN received the B.S. degree in information security from the Hefei University of Technology, in 2013, the M.S. degree in computer technology from Yunnan University, in 2015, and the Ph.D. degree from Tongji University, in 2020. He is currently a Research Associate with Shanghai University. His current research interests include deep learning, image generation, and image inpainting.