LPI-Net: Lightweight Inpainting Network with Pyramidal Hierarchy

Siyuan Li[1][0000−0002−2354−4233], Lu Lu[1], Kepeng Xu[1], Wenxin Yu[1*], Ning Jiang[1], and Zhuo Yang[2]

[1] Southwest University of Science and Technology
[2] Guangdong University of Technology
[*] yuwenxin@swust.edu.cn

Abstract. With the development of deep learning, there are a lot of inspiring and outstanding attempts in image inpainting. However, the designed models of most existing approaches take up considerable computing resources, which result in sluggish inference speed and low compatibility to small-scale devices. To deal with this issue, we design and propose a lightweight pyramid inpainting Network called LPI-Net, which applies lightweight modules into the inpainting network with pyramidal hierarchy. Besides, the operations in the top-down pathway of the proposed pyramid network are also lightened and redesign for the implementation of lightweight design. According to the qualitative and quantitative comparison of this paper, the proposed LPI-Net outperforms known advanced inpainting approaches with much fewer parameters. In the evaluation inpainting performance on 10-20% damage regions, LPI-Net achieves an improvement of at least 3.52 dB of PSNR than other advanced approaches on CelebA dataset.

Keywords: Image Inpainting · Lightweight Network · Deep Convolution.

1 Introduction

Digital image inpainting technology, which aims at completing the missing contents of damaged images, is a basic and critical research task in the field of computer vision. In recent years, with the rapid progress of deep convolution neural networks (DCNN), digital image inpainting technology has attracted extensive interest and achieve great progress. Up to now, most of these advanced approaches adopt the lengthy and complicated image inpainting generators. Therefore, these methods are greatly dependent on considerable computational resources and have a high inference latency. But in many popular vision applications, such as mobile applications, website tools, and batch image processing, the image inpainting tasks are expected to be carried out on small-scale platforms with limited computation.

Among the recent literature about efficient architecture, Andrew G. *et al.* [4] proposes a small, low latency general network constructed by depthwise separable convolutions modules, which is called as MobileNet. Afterward, Mark

Sandler *et al.* [14] further extend MobileNet to the second version (MobileNet V2) by applying the linear bottlenecks block and inverted residuals. Due to the principal tasks of these general architectures are image classification and object detection tasks, these networks lack a forceful decoder, and therefore they cannot directly be applied to the image inpainting task.

This paper proposes an efficient pyramid network that is specifically designed for image inpainting and resource-constrained environments, which is named as Lightweight Pyramid Inpainting Network (LPI-Net). The proposed architecture combines the modified pyramid network with the novel lightweight ResBlock to adapt the image inpainting task. Due to the exquisite design of the pyramid network for the inpainting task, LPI network achieves excellent inpainting performance with a small number of parameters.

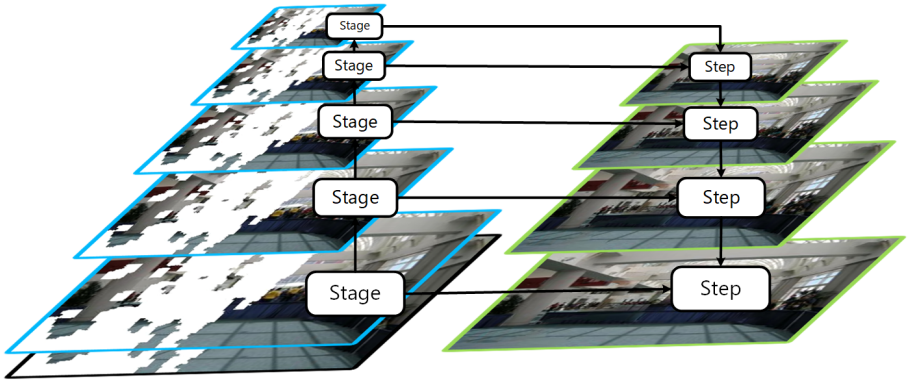

Fig. 1. The general design of the Lightweight Pyramid Inpainting Network, which consists of 5 stages in the bottom-up pathway (left half of the figure) and 4 steps in the top-down pathway (right half).

In addition, the proposed architecture is evaluated on Place2 [19] and CelebA [9] dataset. Compared with other inpainting methods, the proposed architecture obtains equivalent state-of-art performance in qualitative and quantitative aspects. In conclusion, the main contributions of this paper are as follows:

- We propose a tailored pyramid generator network specifically for image inpainting tasks. The structure and first layer of generator are specifically designed for inpainting.
- To reduce the parameters and computation of the designed model, the depthwise separable convolutions module and linear bottleneck module are embedded in the suitable position of the LPI-Net, the united layers are named as lightweight ResBlock.
- Experiments on two public datasets show that proposed LPI network achieves competitive inpainting results and only consumed little computing resources.

2 Related Work

In the last few years, a variety of image inpainting literature introduces the deep convolution neural network, which can extract semantic information for structural inpainting. Besides, because the Generative Adversarial Network (GAN) [1] has the ability to generate novel content similar to the counterpart existing in the training dataset, most of the advanced inpainting methods introduce adversarial learning.

Context Encoder (CE) [12] is one of the earlier work that introduces the GAN to the learning-based inpainting, the encoder of CE embeds the corrupted image into the high-level compact features with low-resolution and multiple channels, then the decoder utilizes these compact features to reconstruct the high-resolution features.

This approach has demonstrated that the excellent ability of the GAN-based convolution neural network in the understanding image context. Because Context Encoder can craft realistic objects in the missing region, its inpainting performance is far superior to the non-learning inpainting method [2]. However, due to the excessive compression of features and sequential forwarding process, some contents generated by CE are visually obscured.

Following the proposal of CE [2], Iizuka *et al.* [5] proposes a large-scale GAN-based inpainting architecture with two sibling discriminators, which aim to improve the local authenticity and overall rationality of inpainted image. However, because style inconsistencies exist in the generated image, the inpainting performance is greatly dependent on the post-processing in GLCIC [5]. In addition, due to the complexity of the architecture, the training process of GLCIC is both time-consuming and unstable.

Recent literature [17,18,10] exploit a two-stage architecture to complete images in two steps. Although these methods achieve state-of-the-art inpainting performance, the intricate workflow and considerable computation limit the wider application of image inpainting.

Feature Pyramid Network (FPN) [7] is originally proposed to detect objects in different scales, the decoder of FPN maintains a multi-scale feature representation, where all levels are semantically stronger than Unet-like [13] networks, including high-resolution level features in the decoder. The strong semantics with the high-resolution feature is conducive to the reconstruction of the overall scene and detailed information. Therefore, the architecture similar to FPN can be applied to the task of image inpainting. However, due to the use of multiple residual blocks [3] in each stage of FPN, FPN has a large number of parameters and takes up considerable memory.

MobileNet [4] builds the network blocks with depthwise separable convolutions, which is designed to reduce the redundancies of filter weights. The depthwise separable convolutions reform the standard convolution to two layers - a depthwise convolutions layer and a pointwise convolution. The depthwise convolution layer only adopts a single 2D filter per each input channel (input depth), instead of using multiple 3D filters in standard convolution. And the pointwise convolution is equivalent to a simple 1×1 standard convolution, which is used

to employ a linear transformation of the output of the depthwise layer. More specifically, the standard convolutional layer takes as input a feature map F with the shape of $x \times y \times M$. It produces an $x \times y \times N$ feature map G. Rather in depthwise separable convolutions, the depthwise convolution inputs an $x \times y \times M$ feature and temporarily produce an $x \times y \times M$ intermediate feature by M 2D filters. And then, the pointwise convolution process the intermediate feature and finally generate an $x \times y \times N$ output. Because depthwise convolution only filters input channels with M 2D filters, depthwise separable convolution is extremely efficient relative to standard convolution.

Afterward, MobileNetV2 [14] extends the depthwise separable convolution to the linear bottlenecks and residual block, which reduce both the operating space and computing parameters. As we observe, this enhanced residual bottleneck is feasible to replace the residual block in the pyramid network. It can greatly reduce the complexity of computation and parameters of the pyramid model.

3 Proposed Method

3.1 Architecture of LPI-Net

The raw Feature Pyramid Networks (FPN) [7] consists of the bottom-up pathway, top-down pathway, and lateral connections, each pathway involves features at several scales with a scaling step of 2. At each feature block with the same scale, each pathway process the feature maps with more than three residual blocks. In order to facilitate the application of object detection, all stages in the top-down pathway maintain the dimension of features at 256. All these original designs for object detection make FPN a gigantic network, and it can not be directly applied to lightweight image inpainting.

In the paper, we proposed a lightweight inpainting network named Lightweight Pyramid Inpainting Network, which has a general architecture similar to the FPN but reduces some intricate and redundant designs. Moreover, the first layer and inside operations of LPI network are modified to better adapt to image inpainting, which can efficiently extract the high-resolution features.

As shown in Figure 1, the general components of LPI network include the bottom-up pathway, top-down pathway, and lateral connections. Similar to the FPN [7], the bottom-up pathway consists of 5 stages, where each stage contains one standard convolution layer and N residual bottleneck layer with depthwise separable convolutions [14]. Inside each stage, the operations manipulate and produce feature maps of the same size. Thus we thumbnail each stage into one rectangle in the diagram (Figure 1) for better visualization. Moreover, the operation group in the top-down pathway of LPI network is simplified and referred to as 'steps'.

3.2 The Internal Structure of Each Stage in Bottom-up Pathway

Figure 2 shows the concrete internal process of each stage in the LPI network, which includes a standard convolution layer and N novel lightweight residual

bottlenecks (lightweight ResBlocks). The N is the number of superimposed lightweight blocks, and it is a hyperparameter that determines the scale of the LPI network.

The first convolution layer in each stage is designed to control the variation of resolution of all features in that stage, and it can expand the channel dimension of feature to a specific amount. In the lightweight ResBlock, the $tanh(x)$ activation function is adopted to deliver feature values into the range between -1 and 1, and the same paddings in each layer of lightweight ResBlock are used to maintain the same space dimension.

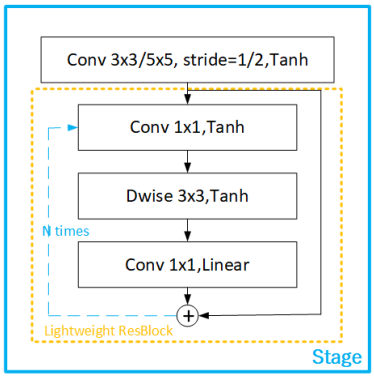

Fig. 2. The forward process of each stage in the bottom-up pathway of the pyramid inpainting network, the 'Dwise' in the diagram refers to depthwise convolution [4].

To obtain the detailed information from the original image, we set the stride of the first convolution layer in the first stage of the bottom-up pathway as 1, so that the first stage adequately extracts and operates the full-resolution features from the corrupted image. This novel layer is specially designed for image inpainting tasks. Because the bottom-up pathway in the pyramid network aims to encode the image to compact features, the stride in other stages of the bottom-up pathway is set to 2, which gradually reduces the space dimension of features in the forwarding process.

3.3 Top-down Pathway and Lateral Connections

The top-down pathway of LPI network reconstructs the high-resolution features by lateral connections, bilinear upsampling layer, and standard convolution layers.

The detailed operations of each step in the top-down pathway are shown in Figure 3. The channel reduction layer reduces the dimension of features before the subsequent operations, the smooth layer maintain the channel of feature unchanged, which aims to process and fuse the upsampled features and the transferred features from the bottom-up pathway. Thus there are only two standard

convolution layers in each step of the top-down pathway. These designs retrench a great deal of computation and space resources from the pyramid network.

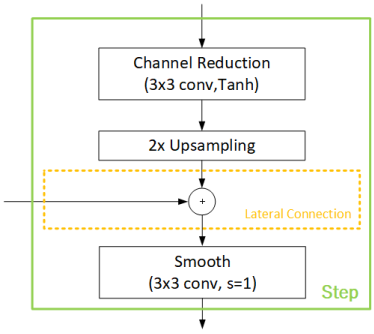

Fig. 3. The forwarding process of lateral connections and detailed per step in the top-down pathway, the upsampling layer refers to bilinear interpolation.

3.4 Loss Function and Adversarial Learning

Similar to those advanced inpainting work [5,12,18,10], we also exploit the adversarial learning [1] to improve the inpainting performance. In this paper, the function of the LPI network (generator) and discriminator are respectively denoted as $G(x)$ and $D(x)$, and the design of $D(x)$ is similar to one of the recent advanced work [10].

The G and D play the two-player minimax game in adversarial learning, the adversarial loss is described in Equation 1. The distribution p_{pred} is the generated data ($\boldsymbol{z} = G(x)$) from the LPI network, and the p_{gt} is the ground truth image distribution that needs to be learned. The maximization of $\mathcal{L}_D$ makes the discriminator D better assign the correct label to both ground truth images and generated images as accurately as possible.

$$\max_D \mathcal{L}_D = \mathbb{E}_{\boldsymbol{x}\sim p_{gt}(\boldsymbol{x})}\left[\log(D\left(\boldsymbol{x}\right))\right] + \mathbb{E}_{\boldsymbol{z}\sim p_{pred}(\boldsymbol{z})}\left[\log\left(1 - D\left(\boldsymbol{z}\right)\right)\right] \tag{1}$$

Then the total loss function of the generator in LPI network can be formalized as Equation 2, it only includes the l_{ℓ_1} distance between the inpainted image and truth image and the adversarial loss feedback from discriminator .

$$\min_G \mathcal{L}_G = \mathcal{L}_{\ell_1} + \mathcal{L}_D \tag{2}$$

If respectively denote ground truth image and the inferred image as $\mathbf{I}_{gt}$ and $\mathbf{I}_{pred}$, the l_{ℓ_1} loss can be specifically interpreted as Equation 3.

$$\mathcal{L}_{\ell_1} = \frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{I}_{gt}(i) - \mathbf{I}_{pred}(i)\right\|_1 \tag{3}$$

where the n refers to the total number of pixel in the image sample, and $\mathbf{I}(i)$ represents ith pixel value in image. For example, if the input image is a 256×256 RGB image, the denominator $n = 256 \times 256 \times 3 = 196,608$.

4 Experiments

4.1 Implementation Details

The proposed LPI network and adversarial training process are based on Pytorch framework [11]. And the performance evaluation of the proposed LPI network is based on two public image datasets - Places2 [19] and CelebA [9], and the resolution of samples in both datasets is 256×256. All the quantitative results in this section are based on the comparison between the real image sample $\mathbf{I}_{gt}$ and the processed composite sample $\mathbf{I}_{comp} = \mathbf{I}_{pred} \odot (1 - M) + \mathbf{I}_{gt} \odot M$. And in the evaluation stage, we measure the performance of $10,000$ randomly selected samples from the test data set.

The irregular mask maps used in the experiment are provided by work [8], which labeled the corrupted area with values 1. For the convenience of the direct use of masks in the training process, our experiment inverts the masks and resizes them into the resolution of 256×256. In this experiment, the generative loss (Equation 2) and adversarial loss (Equation 1) of LPI network are both updated and minimized by Adam [6] optimizers with the learning rate of 1×10^{-4} and 1×10^{-3} respectively, and the parameter of Adma optimizers $\beta 1$ and $\beta 2$ are set as 0.0 and 0.9. It is worth noting that we simultaneously update the learnable parameters of the generator and discriminator at each step, and thus the optimization degree of the network (generator/discriminator) is only determined by different learning rates.

4.2 Quantitative Comparison and Analysis

To evaluate the inpainting performance on different degrees of image damage, we calculate the quantitative performance of the models under the four different damage ratios in terms of peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [15], and average absolute error (ℓ_1 distance) [16]. The computation of SSIM is based on the window size of 11. In the following diagrams, the LPI-Larger is a large version of LPI network because it set the N (involved in Section 3.2) to 4, the standard LPI network set the N to 1.

Places2 [19] is a large-scale public dataset that collected numerous natural scenes, which contains more than 1,000,000 image samples for training and more than 300,000 images for evaluation. Compared to Places2, the scale of CelebA [9] is smaller, and it consists of 202,599 number of face images. In our research, we adopt the align and cropped face image provided by CelebA to evaluate the inpainting performance of face images.

The performance comparison on test set of Places2 [19] is shown in Table 1, it can be observed that the proposed LPI network outperforms these state-of-art image inpainting approach in the listed four methods. And the Table 2

records equivalent inpainting comparison on the CelebA [9] of different advanced approaches, these data demonstrate that the excellent inpainting ability of LPI network on face completion task and medium-scale dataset.

Table 1. The comparison of 3 kinds of quantitative evaluation over the Places2 [19], these existing recorded data are taken from the literatures [8,10]. These data are calculated over 10,000 samples from Places2 test dataset, and the correspond mask maps are provided by PConv [8].

Method	10%-20% Damage			20%-30% Damage			30%-40% Damage			40%-50% Damage		
	PSNR	SSIM	ℓ_1(%)	PSNR	SSIM	ℓ_1(%)	PSNR	SSIM	ℓ_1(%)	PSNR	SSIM	ℓ_1(%)
GLCIC [5]	23.49	0.862	2.66	20.45	0.771	4.70	18.50	0.686	6.78	17.17	0.603	8.85
CA [17]	24.36	0.893	2.05	21.19	0.815	3.52	19.13	0.739	5.07	17.75	0.662	6.62
Pconv [8]	28.02	0.869	1.14	24.9	0.777	1.98	22.45	0.685	3.02	20.86	0.589	4.11
EdgeCnt [10]	27.95	0.920	1.31	24.92	0.861	2.26	22.84	0.799	3.25	21.16	0.731	4.39
LPI-Large	**30.55**	**0.951**	**0.71**	**26.44**	**0.900**	**1.51**	**23.78**	**0.842**	**2.47**	**21.54**	**0.775**	**3.79**
LPI	30.32	0.948	0.75	26.31	0.894	1.56	23.66	0.834	2.56	21.44	0.766	3.91

Table 2. The comparison of 3 kinds of quantitative evaluation over the CelebA [19], these existing recorded data are taken from the literature [10].

Method	10%-20% Damage			20%-30% Damage			30%-40% Damage			40%-50% Damage		
	PSNR	SSIM	ℓ_1(%)	PSNR	SSIM	ℓ_1(%)	PSNR	SSIM	ℓ_1(%)	PSNR	SSIM	ℓ_1(%)
GLCIC [5]	24.09	0.865	2.53	20.71	0.773	4.67	18.50	0.689	6.95	17.09	0.609	9.18
CA [17]	25.32	0.888	2.48	22.09	0.819	3.98	19.94	0.750	5.64	18.41	0.678	7.35
EdgeCnt [10]	33.51	0.961	0.76	30.02	0.928	1.38	27.39	0.890	2.13	25.28	0.846	3.03
LPI-Large	**37.03**	**0.978**	**0.33**	**32.00**	**0.952**	**0.76**	**28.75**	**0.920**	**1.31**	**25.74**	**0.874**	**2.19**
LPI	36.44	0.976	0.36	31.65	0.949	0.79	28.37	0.914	1.37	25.33	0.867	2.32

Table 3 indicates the comparison of the parameters over these aforementioned approaches. It is noticeable that all these statistics only consider the generators in the whole network framework. Because in the real application of mobile devices, we can only deploy the pre-trained generators to accomplish the image inpainting task. Another point to note is that the CA [17] and EdgeConnect [10] are two-stage inpainting approaches, and thus we need to calculate the parameter number of two generators.

It can be observed that LPI-Large is a medium-scale network because it is only larger than the network of CA [17], the LPI is the most lightweight network and the paramters numbers is only 0.2879 times of CA [17] and GLCIC [5]. Moreover, the inpainting performance of the LPI approximate LPI-Large and outperforms the other state-of-art inpainting approaches (observe from Table 1,2).

Table 3. The numbers of network parameters with involved methods, the 'M' in 'Parameters' column equal to 2^{20}. All the statistics only consider the generators. (*): The statistics of Pconv [8] is based on the unofficial implementation.

Method	Parameters	Occupied Capacity of Disk	Network Type
GLCIC [5]	5.8M	46.3MB	Two-Discriminator GAN
CA [17]	2.9M(×2)	13.8MB	Two-stage GAN
Pconv [8]	31.34M*	393 MB*	Only Generator
EdgeCnt [10]	10.26M(×2)	41.1MB	Two-stage GAN
LPI-Large	5.47M	22.0MB	Generative Adversarial Netwrok
LPI	**1.67M**	6.74MB	Generative Adversarial Netwrok

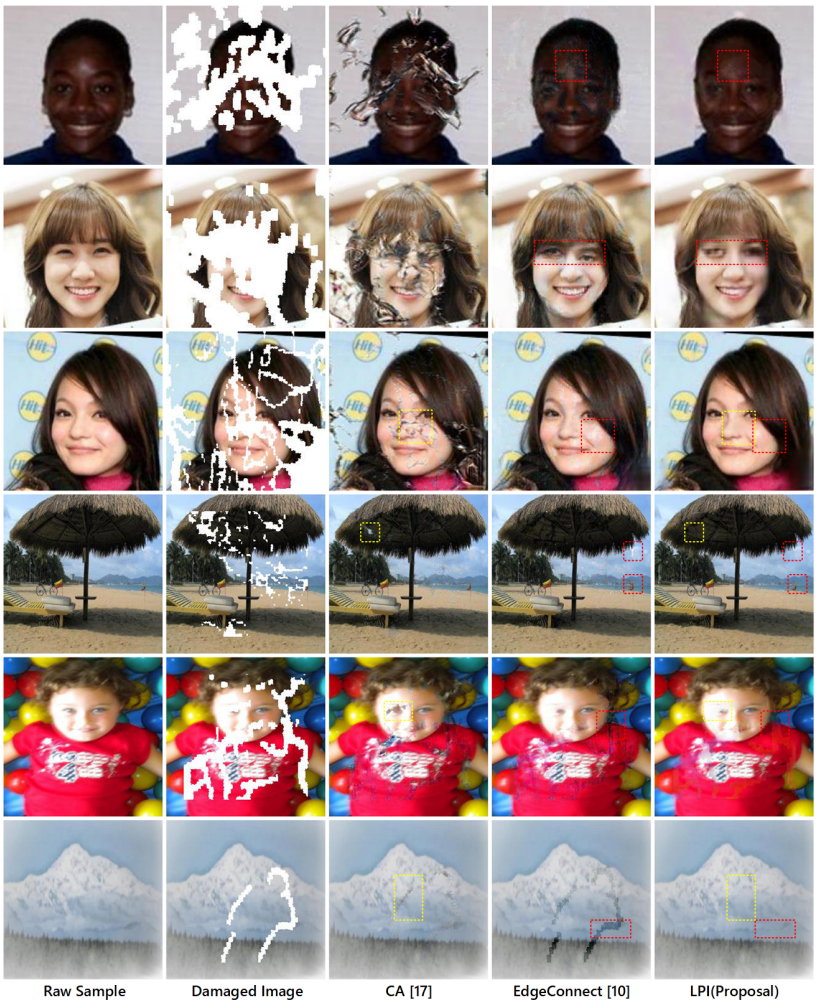

Fig. 4. The qualitative comparison on test dataset of Places2 [19] and CelebA [9] over state-of-art inpainting methods.

4.3 Qualitative Comparison and Observation

As shown in Figure 4, it illustrates the visual inpainting results of different methods on the test set of Places2 [19] and CelebA [9]. The images in the first three rows are sampled from CelebA [9], and the samples in the last three rows are from Places2 [19]. The colored dotted rectangles display the detailed difference between our proposal and other approaches. The result demonstrates that the LPI network can produce more reasonable and legible results while occupying lower memory space.

Specifically, the red rectangle in the second row of Figure 4 indicates that the inpainting result of the LPI network is more realistic and reasonable, the eyes in inpainting results of LPI are at least parallel. And the yellow rectangle and red rectangle in the third row show that the LPI network produces more clear face texture.

Acknowledgements This research is supported by Sichuan Provincial Science and Technology Department (No. 2019YFS0146, No. 2019YFS0155), National Natural Science Foundation of China (No. 61907009), Science and Technology Planning Project of Guangdong Province (No. 2019B010150002).

5 Conclusions

To promote the application of image inpainting on mobile devices and embedded devices, this paper designs and proposes a lightweight pyramid inpainting network named LPI-Net. Benefited from the lightweight ResBlock and lightweight design of each step in the top-down pathway, the LPI network takes up at least 71.2% (CA [17] and GLCIC [5]) fewer parameters than these well-known state-of-art inpainting method.

In addition, the experiments and comparisons show that the proposed LPI network achieves the most advanced inpainting results in terms of qualitative images and quantitative data, and the improvement of the inpainting of smaller damage regions is most obvious. However, the improvement of large-damaged-region inpainting is limited, and thus we can pay more attention to large-damaged-region inpainting in further study.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
2. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Transactions on Graphics (SIGGRAPH 2007) **26**(3) (2007)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

4. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
5. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics (ToG) **36**(4), 107 (2017)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
8. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 85–100 (2018)
9. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
10. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019)
11. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
12. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
14. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
15. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
16. Willmott, C., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research **30**, 79–82 (2005). https://doi.org/10.3354/cr030079, https://doi.org/10.3354%2Fcr030079
17. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5505–5514 (2018)
18. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4471–4480 (2019)
19. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)