A Survey on Improved GAN based Image Inpainting for Different Aims

Nanxiang Wang *

School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

* Corresponding author email: 2020211863@stu.cqupt.edu.cn

Abstract. Computer Vision (CV) is an important field of Artificial Intelligence (AI) that focuses on processing and analyzing the visual data input like images or videos. And image inpainting technique is a task of reconstructing image with missing parts by using the information of the existing part, which becomes a crucial part in CV. It has various image processing applications such as digital cultural heritage preservation, old photograph restoration and object removal. The researchers found that using generative adversarial network can extract the feature from image without huge amount of computation which are two main challenges of image inpainting, therefore GAN-based inpainting method has recently been a hub for research. In this paper, we review most of the latest improved image inpainting methods and divide them into three categories according to their different aims: for semantic image inpainting, for generating higher diversity and for irregular or free-form image input. Firstly, we review the methods from structure and algorithm. Then we analyze the advantages and disadvantages of each method. Finally, we conclude the challenges or limitations raised from our findings and some potential development trend.

Keywords: Image Inpainting; GAN; Deep Learning; Computer Vision.

1. Introduction

Image inpainting refers to the process of selecting appropriate pixels to fill the defect area or occluded area by extracting and screening the features of the known area. At present, the development of inpainting algorithms have become a research hotspot of CV field, which can be widely applicated in medical imaging, military research, image segmentation, and the film and television industry. Although traditional image inpainting methods perform well with images with simple texture, small missing area and low resolution, they can't extract higher-level semantic features and often lack semantic consistency. After GAN was introduced by Goodfellow in 2014[1], researchers found it could be a good solution to the above problems, therefore it is now a widely used research technique in image inpainting. To produce better results, the GAN model has been improved in various perspectives such as network architecture, loss function and convolution. Although GAN model has made great breakthroughs over the recent years, there are still some existing limitations such as the inability to handle high resolution images or the images with large missing region. Our paper makes a detailed summary of GAN-based image inpainting methods in recent years, aiming to provide a macro perspective for the current research status and future work.

In Section 2, a few improved GAN-based techniques for inpainting are introduced and summarized in three categories. We focus mainly on semantic image inpainting, which refers to inpainting an image with a large missing region according to information around the missing region. Pathak et al. [2] first propose context encoder which uses encoder-decoder architecture for generator. Based on context encoder, Iizuka et al. [3] divide discriminator into global and local one to ensure the consistency. Then we introduce contextual attention-based methods [4,5] and WGAN-based methods [6,7,8]. Unlike the other two ideas, WGAN-based methods use a new loss function without changing the architecture or convolution and also achieve top-tier results. As for the rest of the Section 2, we summary the methods for two other directions of optimization: for higher diversity and for irregular or free-form image input.

Section 3 summarizes the datasets used for the methods mentioned in our paper; Section 4 concludes the picture inpainting process, and we also find some limitations and predict the future research directions.

2. Improved GAN-based Image Inpainting

With the development of image inpainting, traditional GAN models no longer meet the needs from many perspectives such as training speed, quality of results and cost. Researchers have therefore proposed many improved methods to get better performance.

Based on their goals, these enhanced techniques can be divided into three groups: for semantic image inpainting, for generating higher diversity and for irregular or free-form image input. The summary of the methods that we have reviewed are shown in Table 1.

2.1 Context Encoder-based

In recent years, semantic image inpainting has drawn great attention. Accurate semantic inpainting techniques have many benefits. Pathak et al. [2] propose context-encoder, the over view of this architecture is shown in Fig .1. The model includes an encoder that extracts the feature representation from the input picture and a decoder that uses the feature map from the encoder to create the missing region content. The encoder is derived from the Alex-Net architecture, but the paper connects the encoder features to the decoder via a channel-wise fully-connected layer to learn more feature information while avoiding an exponential growth of parameters. Five up-convolutional layers, each with a ReLU activation function and learned filters, make up the decoder.

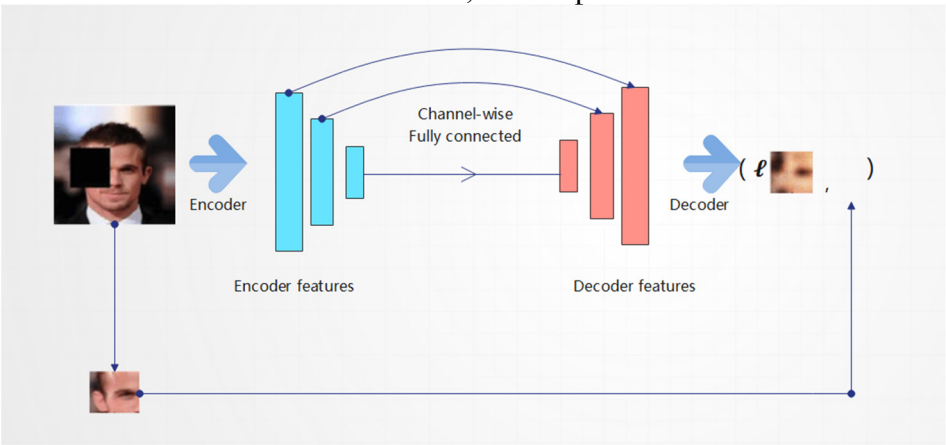

Fig 1. The framework of context encoder

The paper also divides loss function into two parts: $\ell 2$ loss and adversarial loss. The high-frequency detailed information for inpainted images can be recovered from the reconstruction loss, which is a masked $\ell 2$ distance. And the adversarial loss can enhance the coherence of inpainted images.

By revisiting Context Encoder, Iizuka et al. [3] find that there were some existing problems. It doesn't fully describe how to handle the image with random masks nor how to inpaint high resolution images. What's more, the inpainted image isn't consistent with the surrounding region. To solve these problems, they develop a new approach based on Context Encoder. The discriminator is split into two pieces in the paper: global discriminator network and the local discriminator (shown in Fig .2). Both discriminators are trained to recognize if a picture is real or fake, while the generator is trained to trick both discriminator networks.

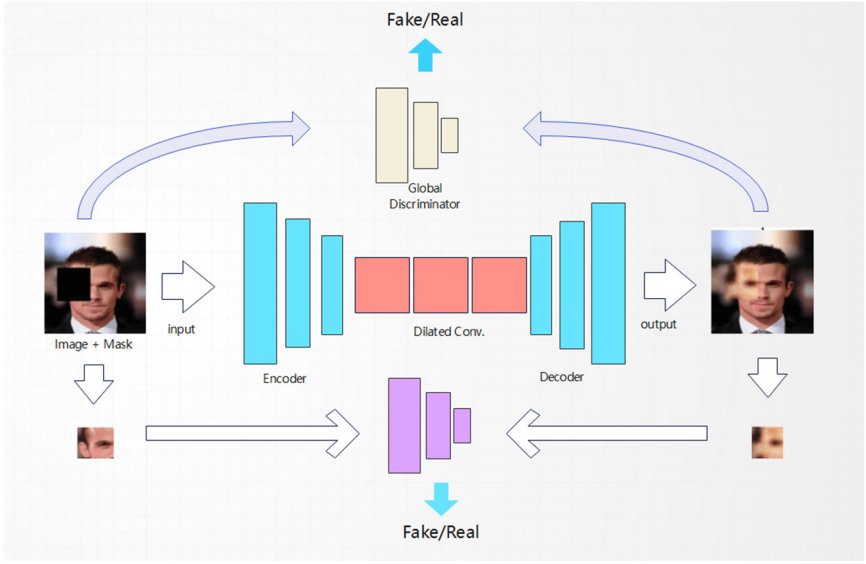

Fig 2. The framework of the architecture used in [3]

The local discriminator network uses a quarter-size region of the inpainted area as input to verify local consistency while the global discriminator network uses the full picture as input to ensure global consistency.

2.1.1 Contextual Attention-Based

To address an issue that previous deep learning algorithms frequently produced deformed structures or fuzzy textures that were discordant with the surroundings, the paper [4] proposes a new generative inpainting Network named coarse-to-fine network architecture (as shown in Fig .3). A picture with white pixels and a binary mask identifying the gaps are used as input by the coarse network to make an initial coarse prediction and trained with reconstruction loss. Then, using the coarse predictions as inputs, the refinement network forecasts refined results. The dilated convolution is used in the both stages. Similar to Iizuka et al., the paper set global critic and local critic. What's more, to make the result local and global consistent, they use the WGAN-GP loss to train the refinement network on a global and local level.

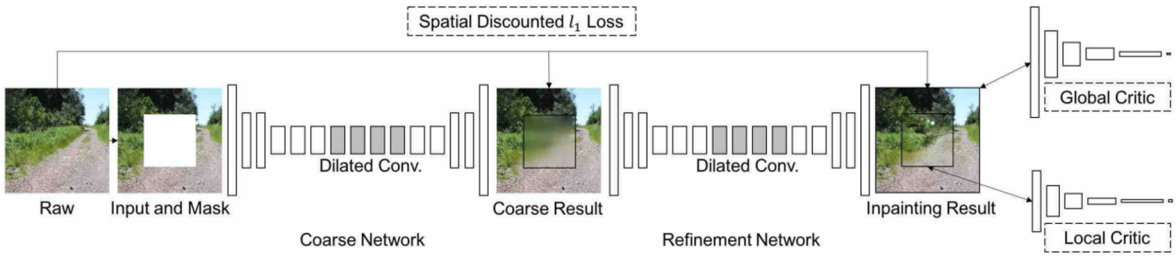

Fig 3. The coarse-to-fine network architecture used in contextual attention [4].

But there's a limitation: Convolutional neural networks that use local convolutional kernels layer by layer find it challenging to extract information from far-off spatial locations. To solve that, the paper introduces CA layer and consider attention, which are the main innovations of the paper. The CA contents initially employ convolution to determine the foreground patches' matching score with background patches. Then they use attention propagation to encourage the coherency of attention.

To combine the contextual attention layer with the generator network, they introduced two parallel encoders in the refinement network. To extract the background region, feature, the top encoder concentrates on filling the contextual attention layer. The lower encoder concentrates on the imagined

missing area's content using the dilated conversion. To generate the inpainted image, the output feature maps from both encoders are combined and sent into a decoder.

The approaches based on contextual attention can't locate the proper features and lead to artifacts in the results. Besides, they always incur a considerable computational burden. Aiming at these problems, Zeng et al. [5] propose an auxiliary contextual reconstruction task. The paper attaches an extra branch for contextual reconstruction as a new training loss termed CR loss to the network. The generator has a similar coarse-to-fine architecture to the one described in gated convolution, but the paper replaces the contextual attention with the CR loss. They train coarse network with $\ell 1$ loss while using the $\ell 1$, adversarial, and CR losses to train the refinement network. The generator features are trained to be as similar as possible to the recognized picture features with the minimum inpainting loss by reducing CR loss. The experiment demonstrates that CR loss produces superior outcomes to the official CA layer implementation while avoiding computational overhead.

2.1.2 WGAN-based

While the earliest GANs are capable of producing decent results, it is frequently discovered that they falter when attempting to do more challenging calculations. A mode collapse or a failure of convergence or are two of the most typical causes. When a model fails to yield optimum or high-quality outcomes, it is said to have failed to converge. When a mode collapses, the model is unable to create distinct pictures that replicate the same pattern or quality. Hence, to solve some of these issues, a novel approach called WGAN is presented by Arjovsky et al. [6] as an alternative to conventional GAN training. No longer using a logarithm, the paper instead employs a novel loss function derived from the Wasserstein distance. The discriminator assists in determining the Wasserstein measure between the distribution of actual and generated data rather than acting as a direct critic. No sigmoid function is employed to restrict the values to the genuine or false values of 0 or 1. The WGAN discriminator networks instead return a value that falls inside a range. The weights on the critic function will be clamped to a narrow-defined range following each gradient update. With the new model, learning stability can be improved, problems like mode collapse can be resolved, and informative learning curves.

However, Wasserstein GAN is not perfect. Wasserstein GAN can occasionally produce only subpar results or can't converge, and Gulrajani et al. [7] addressed this issue and discovered that the application of weight clipping in WGAN to place the critic under a Lipschitz restriction is frequently to blame for these issues. Therefore, as a replacement to clipping weights, they suggest: punish the critic's gradient norm in relation to its contributions. The new model, known as WGAN-GP, outperforms the traditional WGAN in terms of training efficiency and sample quality.

Inspired by WGAN and WGAN-GP, Vitoria et al. [8] provide a novel technique for semantic image inpainting in order to restore significant lost portions of an image. The paper improves the architecture of WGAN-GP such as the removal of the fully-connected layers over the top of convolutional features, the introduction of the residual learning framework in generator as well as discriminator and the switch from the frequently used batch normalization by a layer normalization. These improvements ease the training and make the networks stable. Additionally, they introduce brand-new optimization loss that incorporates contextual information that utilizes the image values and gradients as well as semantic information that intrinsic in the image. Their proposal achieved top-tier results.

2.2 For Generating Higher Diversity

PEN-Net [9] makes great breakthrough that it can generate visually-realistic results, but the limitation is that the results lack diversity. Due to the irregular content production in the hole region, image inpainting may in fact yield multiple outputs. Aiming at this diverse image inpainting is less discussed by PEN-Net and other existing inpainting methods, Liu et al. [10] propose PD-GAN based on vanilla GAN to generate diverse image result from several inputs for random noise. The paper propose that the diversity is improved while approaching the hole center and is reduced towards the

hole boundary. Based on this theory, they provide the Hard SPDNorm and Soft SPDNorm, which make up SPDNorm ResBlock. While the soft SPDNorm makes the training stable and dynamically learns the condition of the prior information but lacks diversity, the hard SPDNorm enhances the possibility of producing diverse results but cannot guarantee the quality of the results. They can work well together with the help of the SPDNorm ResBlock. To increase network diversity, they use perceptual diversity loss to train the PD-GAN.

Experiments show that PD-GAN is effective for visually realistic image inpainting.

2.3 For Irregular or Free-form Image Input

The majority of the current deep learning-based picture inpainting techniques rely heavily on pricey post-processing and only inpaint the image using a regular rectangular area mask. In addition, these techniques frequently produce artifacts like blurriness and color disparity. To solve these problems, partial convolutional layer [11] is introduced. Thanks to the mask update technique, it can tell whether the input contains valid pixels compared to earlier convolutional layers. As to the network, they design a U-Net architecture. However, they employ closest neighbor up-sampling at the decoding step and substitute partial convolutional layers for all of the convolutional layers. The loss function contents perceptual, per-pixel and style losses. The new approach can get SOTA inpainting results.

By reviewing the partial convolution, Yu et al. [12] find several limitations. First, classifying all locations as either invalid or legitimate will ignore some crucial information. Second, if we apply this method to user-guided picture inpainting, where users give sparse sketches inside the mask, the pixel positions cannot be regarded as valid or invalid. Third, the partial convolution can't tell if the current locations are within or outside of the hole therefore the pixels in hole can't be synthesized.

To overcome the above limitations Yu et al. propose a new image inpainting system. Compared to the partial convolution, the following improvements have been made: (1) They replaced partial convolution with gated convolution. To provide superior inpainting results, it may be trained to highlight the masked areas and sketch information using several channels. Additionally, it is capable of autonomously learning soft masks from data. (2) They used coarse and refinement networks similar to the one in CA (contextual attention) instead of U-Net used in partial convolution because for non-narrow mask, the U-Net design has no discernible influence. (3) As to the loss function, the paper used SN-PatchGAN loss and $\ell 1$ reconstruction loss.

The new system produces outcomes that are more adaptable and of greater quality than the older ones. Additionally, this study focuses on the image inpainting of pictures using a free-form mask and user-provided instructions, which will have further applications in the future.

Inspired by partial convolution and gated convolution, Jo et al. [13] propose a novel image editing system aiming at inpainting facial images while users provide color, sketches and free-form masks as inputs. They use a U-net encoder-decoder architecture for the generator. The generator has 16 convolutional layers, and all of them—aside from the input and output layers—are subjected to local signal normalization (LRN). And the generator is trained with perceptual loss, style loss, per-pixel losses, generic GAN loss and total variance loss. Referring to gated convolution, they train the network with the SN-PatchGAN discriminator structure, using a style loss.

In compared to existing systems, the network design and loss algorithms greatly increase inpainting outcomes while requiring minimal effort.

Table 1. Summary of the viewed methods

Paper title	Description	Loss function	Dataset	Merits	Demerits
Context-encoder[2]	Encoder–decoder network architecture, using context encoder as generator.	$\ell 2$ reconstruction, Adversarial loss.	Paris Street View, ImageNet, PASCAL VOC2007.	● Provides high Semantic consistency. ● Performs well for center mask regions	Can't inpaint irrgular missing regions.

Globally and Locally Consistent[3]	An encoder–decoder architecture with dilated convolutional layer is used as the generator, while the discriminator is divided into global and local one.	Weighted MSE, Adversarial loss.	Place2, CelebA.	• Can inpaint realistic image. • Can inpaint images of faces.	Can't inpaint irrgular missing regions.
Contextual attention[4]	Coarse-to-fine network architecture based on dilated convolution combined with attention mechanism and contextual attention layer.	Reconstruction loss, Global and local Wasserstein GAN loss.	Places2, CelebA, CelebA-HQ, DTD and ImageNet.	• Better extract the feature. • Higher consistency.	Can't inpaint irrgular missing regions.
CR-Fill[5]	Uses a coarse-to-fine architecture similar to DeepFillv2 as generator while replacing CA layer as CR loss.	ℓ1 loss, adversarial loss and the CR loss	Places2.	• Avoids artifacts in the results. • Avoids computational overhead.	Needs to improve the speed of handling irregular holes.
WGAN[6]	Removes the logarithm of the loss function and the sigmoid activation function of the discriminator. The weights will be clamped to a small fixed range after every gradient update.	Wasserstein GAN loss.	LSUN bedrooms.	• Improves the stability. • Avoids mode collapse. • Offer informative learning curves.	Can still only produce subpar samples or collapse.
WGAN-GP[7]	Improves gradient based on WGAN.	Wasserstein GAN loss.	CIFAR-10 and LSUN bedrooms.	Improves the speed and quality of generated samples.	Not applicable to the filling of larger regions.
WGAN_GP based Semantic image inpainting[8]	Based on WGAN-GP, removes the fully connected layers on top of convolutional features and replaces the widely used batch normalization by a layer normalization.	Contextual loss, prior loss and gradient loss.	CelebA and Street View House Numbers(SVHN).	Performs well for large missing areas	• Is limited to images of low resolution. • Can't extract more complex features
PEN-Net [9]	Uses PEN-Net and ATN for the generator while PatchGAN is used as the discriminator.	Adversarial loss and pyramid ℓ1 loss.	Facade, DTD, CelebA-HQ, Places2.	• Can provide more visually realistic and semantically plausible results. • Effective.	Not applicable to the filling of larger regions.
PD-GAN[10]	Hard SPDNorm combined with Soft SPDNorm.	Diversity loss, reconstruction loss, feature matching loss and hinge adversarial loss.	Paris StreetView , CelebA-HQ , and Places2.	Effective to generate diverse and visually realistic contents	Needs higher consistency.
Partial convolutions[11]	Uses a U-NET architecture and replaces normal convolutions with partial convolutions.	Perpixel , perceptual , style and the TV losses.	CelebA, CelebA-HQ, Places2 and ImageNet.	• Achieves state-of-the-art inpainting results. • Performs well for irregular missing regions.	• Fails for some sparsely structured images or the largest of holes. • Ignores some important information. • Unable to say if present positions are inside or outside of the hole.
Gated convolution[12]	Uses coarse and refinement network and replaces the partial convolution with gated convolution.	SN-PatchGAN loss and ℓ1 reconstruction loss.	Places2 and CelebA-HQ.	• Produces outcomes that are more adaptable and of greater quality. • Can inpaint the images under guidence provided by users. • The pixels in hole can be synthesized.	Doesn't support high resolution image.

SC-FEGAN[13]	Uses U-net architecture as generator, SN-PatchGAN structure for the discriminator. Style loss	The lossed of total variance,Per-pixel, style, perceptual, and generic GAN.	CelebA-HQ.	• Works with free-form masks, sketches and color provided by users. • Requires minimal effort.	Needs to improve the consistency.

3. Dataset

Datasets play a key role in training and evaluating methods, which contents training datasets and test datasets. We have analyzed the datasets used in the reviewed papers from different perspectives: category, number of images, purpose and resolution. The summary is shown in Table 2. and some samples of the datasets are shown as Fig. 4- Fig. 8.

Table 2. Dataset used in the reviewed papers

Dataset	Category	Number of images	Purpose	Resolution
CelebA[14]	Celebrity face	202,599	Facial inpainting	178 × 218
CelebA-HQ[15]	Celebrity face	30,000	Facial inpainting	1024 × 1024
ImageNet[16]	Multiclass	1,281,167	Classification and localization	Variable
Paris Street View[17]	Street iew	14,900	Urban	936 × 537
Places2[18]	Multiclass	>10 Million	Urban	Variable
DTD[19]	Texture images	5,640	Recognition	300 × 300-640 × 640
PASCAL VOC2007[20]	Multiclass	9963	Classification	Variable
CMP facade[21]	Building facade	606	Inpainting	Variable
SVHN[22]	House Number	> 600,000	Recognition	32 × 32
LSUN bedrooms[23]	Bedroom scenes	303,125	Classification	Variable
CIFAR-10[24]	Multiclass	80 million	Classification	32 × 32



Fig 4. CelebA [14]

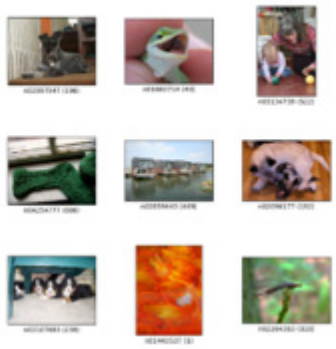

Fig 5. Imagenet [16]

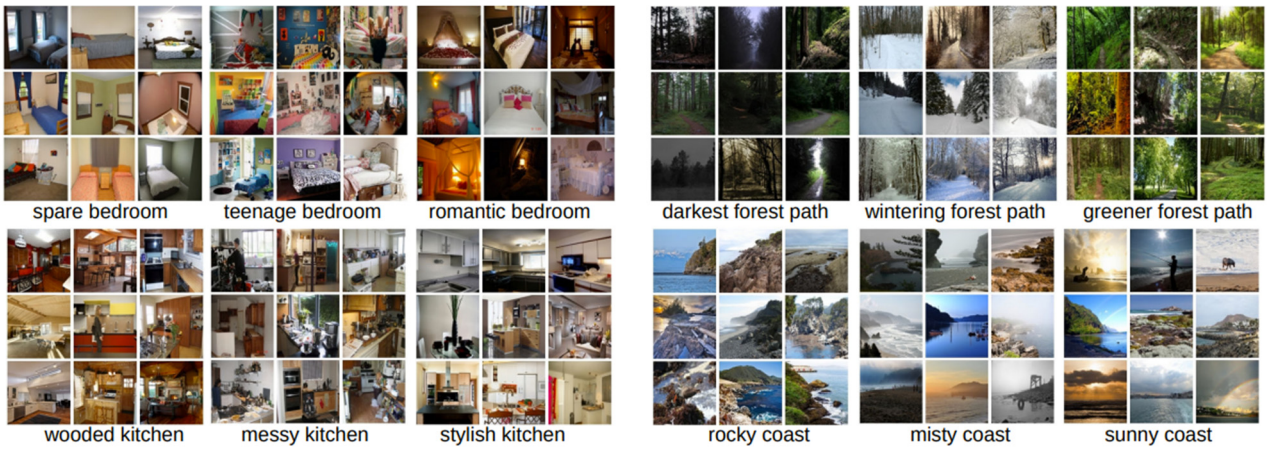

Fig 6. Places 2[18]

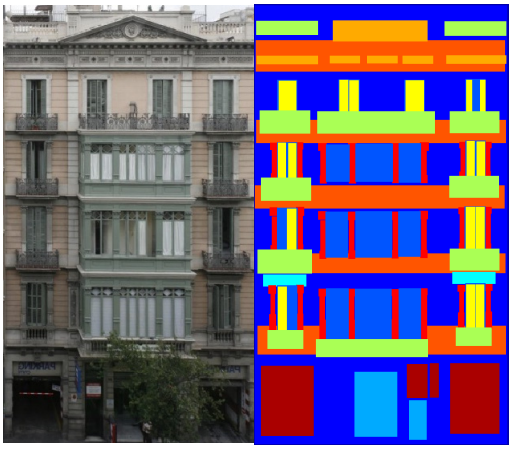

Fig 7. CMP Facades [21]



Fig 8. SVHN [22]

We discovered that the pictures in the image inpainting datasets are separated into truth images and mask images. Pictures from public databases like ImageNet, Places2, CelebA, and Paris StreetView are considered truth images. Mask images are images with masked region. There are three commenly used methods to mask the images: Center area, random blocks and random area. Center area is to zero the square patch in the center. Random blocks are to set up multiple overlapping square masks, taking up to 1/4 of the image. Random area is masking an area of random shape. Most of the methods use center area because it's the easiest way.

In terms of purpose, CelebA and CelebA-HQ are mainly used for facial inpainting, ImageNet, Paris Street View and Places2 datasets are applicable for realistic view image inpainting, DTD, SVHN are used for recognition and PASCAL VOC2007, LSUN bedrooms, CIFAR-10 are used for classification. For different training and evaluation purposes, each of these methods is based on specific dataset for training and testing.

4. Conclusion

At present, the development of GAN based image inpainting algorithms has become a key focus. To better understand the current research status and provide potential development direction, we have reviewed and analyzed some existing improved methods from some main aspects: prior guidance, network architecture (generator, discriminator, convolution and loss function), dataset, performance, application, limitation and development Potential. Three goals are the primary emphasis of the most recent GAN-based image inpainting advancements: for semantic image inpainting, for generating higher diversity and for irregular or free-form image input. By reviewing the paper, we found that many improved models have reached significant performance. But there are still some limitations such as how to inpaint some sparsely structured images or the largest of holes and the higher resolution images. The method of inpainting video stream input is also not mentioned in most papers. What's more, inspired by Gated convolution and SC-FEGAN, we think it's a development trend to consider user needs while inpainting. So, we proposed text-to-image inpainting, which refers to generating the more personalized results under the textual guidance provided by users. We hope that this review can provide some valuable information for how to overcome the limitation and carry out the future work.

References

[1] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[2] Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: Feature learning by inpainting[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2536-2544.

[3] Iizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics (ToG), 2017, 36(4): 1-14.

[4] Yu J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5505-5514.

[5] Zeng Y, Lin Z, Lu H, et al. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 14164 -14173.

[6] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//International conference on machine learning. PMLR, 2017: 214-223.

[7] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[J]. Advances in neural information processing systems, 2017, 30.

[8] Vitoria P, Sintes J, Ballester C. Semantic image inpainting through improved wasserstein generative adversarial networks[J]. arXiv preprint arXiv:1812.01071, 2018.

[9] Zeng Y, Fu J, Chao H, et al. Learning pyramid-context encoder network for high-quality image inpainting [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1486-1494.

[10] Liu H, Wan Z, Huang W, et al. Pd-gan: Probabilistic diverse gan for image inpainting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9371-9381.

[11] Liu G, Reda F A, Shih K J, et al. Image inpainting for irregular holes using partial convolutions [C]// Proceedings of the European conference on computer vision (ECCV). 2018: 85-100.

[12] Yu J, Lin Z, Yang J, et al. Free-form image inpainting with gated convolution[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 4471-4480.

[13] Jo Y, Park J. Sc-fegan: Face editing generative adversarial network with user's sketch and color [C]// Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1745-1753.

[14] Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3730-3738.

[15] Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation[J]. arXiv preprint arXiv:1710.10196, 2017.

[16] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211-252.

[17] Doersch C, Singh S, Gupta A, et al. What makes paris look like paris? [J]. ACM Transactions on Graphics, 2012, 31(4).

[18] Zhou B, Lapedriza A, Khosla A, et al. Places: A 10 million image database for scene recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(6): 1452-1464.

[19] Cimpoi M, Maji S, Kokkinos I, et al. Describing textures in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 3606-3613.

[20] Everingham M, Eslami S M, Van Gool L, et al. The pascal visual object classes challenge: A retrospective [J]. International journal of computer vision, 2015, 111(1): 98-136.

[21] Tyleček R, Šára R. Spatial pattern templates for recognition of objects with regular structure[C]//German conference on pattern recognition. Springer, Berlin, Heidelberg, 2013: 364-374.

[22] Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning[J]. 2011.

[23] Yu F, Seff A, Zhang Y, et al. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop[J]. arXiv preprint arXiv:1506.03365, 2015.

[24] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J]. 2009.