International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-10 Issue-4, February 2021

A New Image Completion Method Inserting an Image Generated by Sketch Image

Gilhee Choi, Kyounghak Lee, Hyung-Hwa Ko

Abstract*: *Recently, many studies on the image completion methods make us erase obstacles and fill the hole realistically but putting a new object in its place cannot be solved with the existing Image Completion. To solve this problem, this paper proposes Image Completion which filled a new object that is created through sketch image. The proposed network use pix2pix image translation model for generating object image from sketch image. The image completion network used gated convolution to reduce the weight of meaningless pixels in the convolution process. And WGAN-GP loss is used to reduce the mode dropping. In addition, by adding a contextual attention layer in the middle of the network, image completion is performed by referring to the feature value at a distant pixel. To train the models, Places2 dataset was used as background training data for image completion and Standard Dog dataset was used as training data for pix2pix. As a result of the experiment, an image of dog is generated well by sketch image and use this image as an input of the image completion network, it can generate the realistic image as a result.

Keywords*: *Image Completion, sketch, pix2pix, WGAN.

I. INTRODUCTION

Due to the recent development of devices and social media, many people have taken pictures frequently in their daily life. They want to remove unwanted people or unsightly structures from the photos we took. To modify these photos, users can directly edit photos using photo edit-tool such as Photoshop. In this case, users want to delete some portion of the original image and put the new image object naturally into the original image. To eliminate the sense of incongruity between the existing image and the new image object, there needs some technique to generate a desired photo by smoothing the edges of the object along with color control.

However, instead of Photoshop in which individual ability is important, the public's interest in a computer program that automatically corrects images has grown. Image Completion is a technique that naturally fills empty areas of an image. The Image Completion or Inpainting techniques came under the spotlight recently. In the existing Image Completion, the method of filling the empty space by a simple patch or dispersion method from information around the empty region was mainly used, but in recent years, deep learning has developed and Image Completion methods using the generation model have been developed. Image Completion using the generation model can alleviate the disadvantages that are difficult to solve when the empty area mentioned as a problem of the existing method is large [1,2,3]. This Image Completion technique is appropriate when a user wants to naturally fill a picture by learning from surrounding information after erasing an unwanted part from a picture. However, after clearing the part you do not want to place, when you want to put a new object, a conventional Image Completion cannot achieve the desired results. In this paper, we proposed a new Image Completion model and combined the pix2pix method to solve the limitations of the existing Image Completion based on deep learning [4,5,6]. The pix2pix method is that can be used for various image-to-image translation. Since the pix2pix model has a good performance despite the relatively simple structure of the generation model, we use this model to create a new image object by the sketch image [4]. In this paper, a new image completion model is proposed by applying WGAN- GP, contextual attention, and gated convolution [5,6]. This paper is organized as follows. In chapter 2, sketch-to- Image translation and the Image Completion method are introduced. Chapter 3 offers a proposed image completion network, and chapter 4 shows experiments and results. After discussion in chapter 5, we conclude in chapter 6.

Manuscript received on January 06, 2020.
Revised Manuscript received on January 11, 2021.
Manuscript published on February 28, 2021.
* Correspondence Author
***Hyung-Hwa Ko,** Department of Electronics and Communications Eng., Kwangwoon Univ., Seoul, Korea. Email: hhkoh@kw.ac.kr
GilHee Choi, Department of Electronics and Communications Eng., Kwangwoon Univ., Seoul, Korea. Email: rlfgmltkfkd@kw.ac.kr
KyoungHak Lee, IACF, Kwangwoon Univ., Seoul, Korea. Email: goldbug@kw.ac.kr

II. RELATED WORKS

A. Image-to-image translation

When a user wants to put a new image in an empty area, information on the new object must be notified and an image object must be created according to the information. The Sketch-to-Image method creates an RGB image using a sketch image as input. Since the user must draw a sketch and put it as an input, the effect of the completeness of the input sketch is large, but the result is designed to be like the result desired by the user. The method used in this study, pix2pix, is a model developed by Berkeley AI research lab in 2018 and based on the conditional GAN structure. This model can solve various image-to-image translation problems by not only learning the mapping from the input image to the output image, but also setting different loss functions to learn this mapping [4].

Training is performed by providing appropriate and sufficient dataset for the image-to-image translation from sketch image. Also, the generated image by pix-to-pix method becomes the condition information of the conditional GAN structure, and an output is generated from the input image.

B. GAN-based Image Completion Method

Image Completion or Image Inpainting refers to a technique that completely transforms an incomplete image by naturally filling the area where there is no value in the image. Image Completion is literally used to correct damaged images, but it can be used to naturally fill the empty spots after erasing obstacles to be erased from within the image. As for the image filling method, a diffusion-based method that fills the spot by spreading the value around the vacancy and a patch-based method that fills the spot by attaching a patch with a high probability of being similar from other areas except the vacancy have been widely used.[1,2] However, with the development of deep learning, the method of image completion using the generative model showed great results.[3,9] The former method has the disadvantage of not being able to understand the structure of the vacancy as well as creating new information because it fills the vacancy by taking values from the surrounding area. However, using a generative model can solve this drawback. However, the method using the generative model has not yet produced perfectly natural results, so Image Completion is still a challenge. GAN is a structure in which the generator outputs data of a distribution similar to that of the training data through competitive learning between the generator network and the discriminator network [7]. Since various types of data can be generated using this GAN model, in most recent image generation models, good image quality can be obtained through GAN-based models. Accordingly, Image Completion is also developed based on the image generation model in that it eventually generates an image of an empty area, and most of the recent Image Completion models are based on GAN. In the case of the generator network, it is composed of an encoder-decoder structure, and it learns the background characteristics of the input image in the process of encoding the image received as an input and learns to naturally fill the empty area through decoding. Among the existing GAN-based Image Completion models, the most basic model is the Context Encoder model proposed by Berkeley University in 2016.[9] With the development of deep learning, the method of configuring the Image Completion network with Convolutional Neural Networks (CNN) improved the result compared to the conventional patch methods in the case where the empty area was large. But the blur phenomenon became more severe. Context Encoder, the first method of image completion, learned the network using the adversarial loss of GAN. By using this adversarial loss together with the existing L2 loss in Context Encoder, the blur phenomenon of the generated image is noticeably reduced.[9] The second image completion method, Globally and Locally Consistent Image Completion, is a model proposed by Iizuka. It is a model that uses GAN for Image Completion like the Context Encoder, and the block diagram is shown in Fig. 1.[3] The difference from Context Encoder is that it uses dilated convolution in the middle. Unlike general convolution, dilated convolution increases the size of the kernel by intentionally creating an empty area between the kernels performing convolution. Through this, it is better to capture the overall characteristics of the image as the receptive area increases. In addition, instead of using only one discriminator of GAN, it uses global discriminator and local discriminator. The local discriminator learns losses to minimize the difference between the image of the empty area and the ground truth image. On the other hand, the global discriminator helps to improve the quality of the created image by learning to make it natural with the surrounding image [3]. The third image completion method, Image Inpainting with Contextual Attention, is published in 2018. The basic structure is based on the model proposed by Iizuka [3] and using WGAN-GP and contextual attention additionally. [6,10]. A general discriminator is difficult to learn with a problem like the mode collapse. To solve this problem, WGAN-GP is adopted. Additionally, in the case of Image Completion, the most appropriate feature value for filling an empty area may be around it, but it may be far away, so it is necessary to check the feature value of a pixel that is spatially distant. Therefore, contextual attention is used. Contextual attention divides the image into a blank area and a background area. And the correlation between them is calculated to obtain the attention score of each pixel in the blank area. Instead of inefficient layer stacking, using this attention score can result natural image. The fourth image completion method is Image Inpainting with Gated Convolution, which is published in 2019 and adding gated convolution and SN-Patch- GAN based on the structure of the Image Inpainting with Contextual Attention method [5]. When a general convolution is used in the Image Completion network, the same filter processing is performed meaningless for the empty area which has no value. Considering the generated result, the best result cannot be produced. To solve this problem, when convolution is performed, the weight of the region without a value should be given less and the weight of the region with a value should be given to each region. Gated convolution is a method that gives different weights for each region according to the network layer.[12] SN-PatchGAN discriminates all coordinates and structures of the created image by discriminating the image from the fully convolutional feature map itself and learns through the result [5]. By using SN-PatchGAN in this way, the whole network learns faster and more stable.

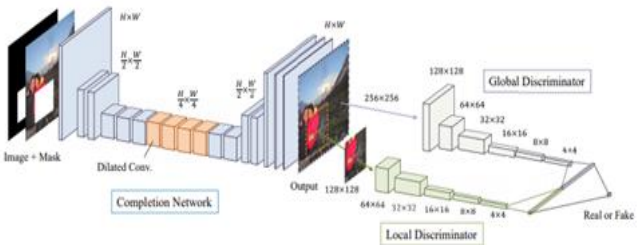

Fig. 1. Globally and Locally Consistent Completion Network

III. PROPOSED IMAGE COMPLETION METHOD

A. Overall network configuration

In this paper, we proposed a structure that creates a new object in the blank area of the input image and naturally fills the surrounding area.

The entire network is largely composed of the pix2pix network and the Image Completion network as shown in Fig. 2. Among these, the pix2pix network generates an RGB image when a sketch image is received as information on an object that a user wants to generate. In the Image Completion network, a new 256*256 RGB image object created by the previous pix2pix network is entered. Then, binary mask data that designates the empty space where this object will be placed is put in. In the Image Completion network, the erased original image and the above two inputs are used to create a naturally filled image.

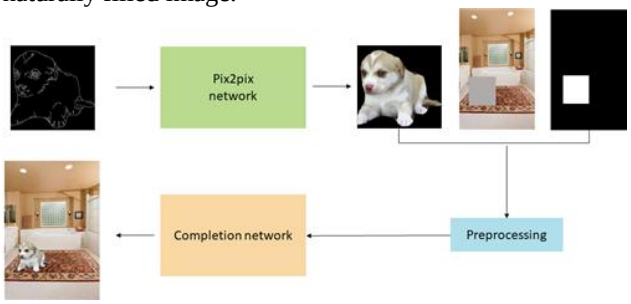

Fig. 2. Proposed Image Completion Network

B. Image Completion network configuration

Image Completion Network consists of Coarse Network and Refinement Network as shown in Fig. 3. As described above, the images received by the Image Completion network after passing through the pix2pix network are the generated RGB image object, an image with a blank area, and a binary mask that indicates where the blank area is. And, put the newly created image object with the background image where the empty area exists.

The first coarse network consists of a simple encoder-decoder method, and the loss function for learning this uses only the L1 loss between the result generated through the coarse network and the ground truth, thus resulting in severe blur. However, in this paper, the reason for proceeding with the refinement network after passing through the coarse network is that by generating rough results from the coarse network and putting the results into the input of the refinement network, the refinement network can focus on learning in more detailed portion. Eventually, it produces good results.

The refinement network is trained through the global and local WGAN-GP loss as well as the L1 loss between the generated result and the ground truth. As a result, blur is reduced, and the appearance is not broken when the texture of the surrounding area and that of the created area are compared. Here, the L1 loss is the sum of the L1 loss between the ground truth and the local patch between the generated result and the L1 loss for the entire image of 256*256 size. If the empty area in the image is a square, the result can be improved because the internal loss is taken care of once more than using only the L1 loss for the entire image. For this reason, the WGAN-GP loss also contributed to improve the result by calculating the summation of loss for the global patch and the local patch. In addition, the refinement network processes the encoder and uses the contextual attention technique. Using contextual attention, the attention score between each pixel in the background and the blank area is calculated, and accordingly, the pixels in the blank area refer to the characteristics of the appropriate area in the background to make the image quality much natural. Both the coarse network and the refinement network use gated convolution instead of general convolution. This avoids loading information from meaningless pixels during convolution and improves the result of Image Completion.

IV. EXPERIMENT AND RESULTS

A. Experimental environments

The experiment was performed on a desktop PC equipped with two GeForce RTX 2080 Ti boards. The operating system is Ubuntu 16.5, and tensorflow 1.13 and python3.5 are used. As a training data, the Stanford Dogs dataset [13] is used to learn the Sketch-to-Image of the pix2pix model, and the Places2 dataset [14] is used as the background image.

The Stanford Dogs dataset is a dataset that collects 120 kinds of dog images around the world. It consists of a total of 20,580 images and provides class labels and bounding boxes for each image. In this paper, 70% of the Stanford Dogs dataset is used for training, and the remaining 30% is used for test. Also, sketch image was created by processing with Canny edge detector from Stanford Dogs dataset. To remove the background except for the dog, a sketch image was created after instance segmentation by the mask R-CNN learned with the MS COCO data set, saving separately as shown in Fig. 4. Places2 dataset [14] consists of over 400 categories and more than 10 million images. Among them, a dataset consisting of high-quality images was used, but only 1/10 of the total dataset was used, with 144,276 images as training images and 36,072 as test images. In addition, a blank area was created at a random location and used for network learning. According to these settings, the time taken to train each model was about 100 hours for pix2pix, and about 160 hours for Image Completion.

B. Results and Analysis

The experiment first confirmed the result of creating a new image object from the sketch image in the pix2pix network. After that, we checked the result of the Image Completion network that received the image
created by the pix2pix network. Fig. 5(a) and 5(b) are the resulting images of the pix2pix network for the sketch image created by the canny edge detector and the sketch image drawn by hand, respectively. The result of the sketch image drawn by hand is somewhat degraded. Discussion on this was dealt with in Chapter 5.

Background image and a sketch image that contains the control information for image generation are used.

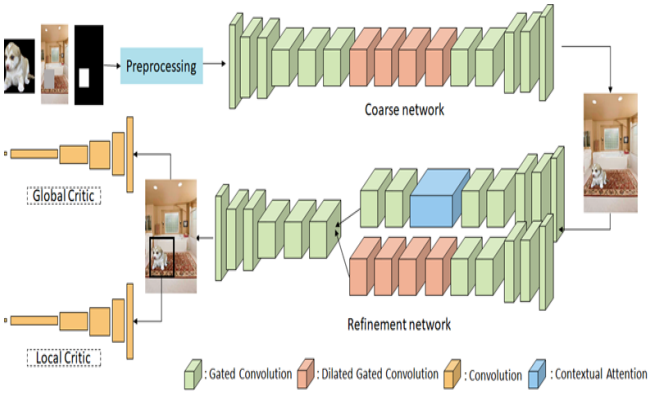

Fig. 3. Proposed Completion Network

Fig. 4. Segmented dog image

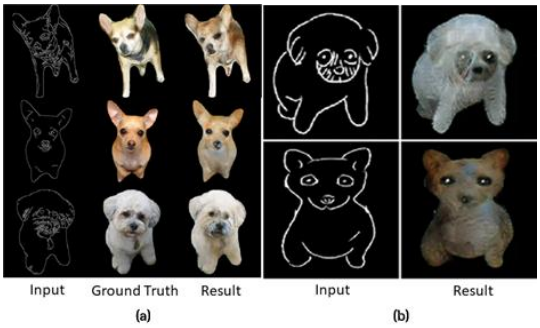

Fig. 5. (a) Generated image by Canny edge image, (b) Generated image by hand-sketch

In the background image, the area to be erased is designated with the mouse, and the binary mask and the image with the area specified in the default background are created.
Fig. 6(a) shows the original image patched with generated image object, Fig. 6(b) shows the mask, Fig. 6(c) shows the completion image, and Fig. 6(d) shows the enlarged image.

Fig. 7 is an example of using the method proposed in this paper for the purpose. When the original image as shown in (b) exists, the cat from the original image is deleted, and a new object, dog, is created from the sketch (a), and naturally filled like (c). If you enlarge the modified part, you can see that the object has changed naturally as shown in (d), so you can see the usefulness of this paper.

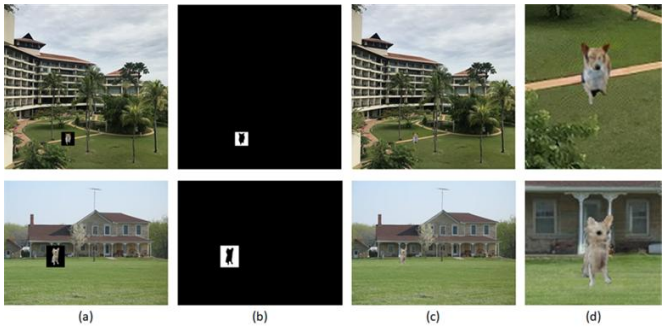

Fig. 6. (a) original image with patched generation image, (b) mask, (c) completion image, (d) enlarged image

V. RESULT AND DISCUSSION

Sketch-to-Image translation model using pix2pix has a relatively simple structure, but it is a good network. However, this was done when adequate training data were given enough. In this paper, to deal with the Image Completion network, the pix2pix network simply receives an edge image through a Canny edge detector and uses it for training, resulting in satisfactory quality results. However, using a sketch image drawn by hand as an input, it was found that the details were degraded. This problem can be solved by training the pix2pix model by providing training data for the hand-drawn sketch image. Compared with the Image Inpainting with Contextual Attention algorithm [6], this paper uses gated convolution instead of general convolution. When calculating the L1 loss for training, there is a difference that not only the L1 loss of the global patch but also the L1 loss of the local patch is used together. This difference results in Fig. 8. To compare our proposed method with Image Inpainting with Contextual Attention model, the code provided in GitHub site in the Attention model was simulated at the same learning environment and learning data. Fig. 8(a) and (b) are the results of the proposed model and the contextual attention model, respectively, and Fig. 8(c) and (d) are enlarged around the objects created in Fig. 8(a) and Fig. 8(b), respectively.

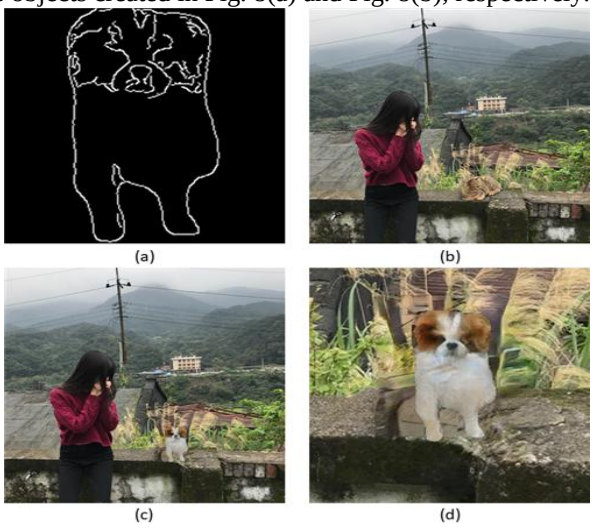

Fig. 7. Application of the proposed image completion method

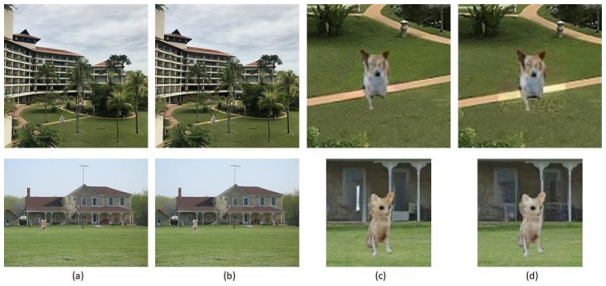

Fig. 8. Comparison with contextual attention method. (a) proposed algorithms, (b) contextual attention model, (c) enlarged image of (a), (d) enlarged image of (b).

As a result of comparing the enlarged results of the two models, it was confirmed that the proposed model produced better results.

VI. CONCLUSION

In this paper, after creating a new image object from sketch information, we proposed a network that naturally fills the surroundings after inserting the object into the background image. The pix2pix model, used as a model to create a new image object from sketch information, is a relatively easy and high-performance model that can handle various image-to-image translation problems. However, it is difficult to get enough image data to match the hand-drawn sketch image, so we applied the Canny edge image of Stanford Dogs data set and obtained good result. But on the hand-drawn sketch image, the result was somewhat deteriorated. However, it seems that it will not be difficult if the learning data for the sketch image is provided sufficiently. In addition, it was confirmed that new image objects and background images can be made as natural as possible by combining WGAN-GP, contextual attention, and gated convolution based on the newly proposed methods for image completion. However, due to the lack of learning data, there is a slight discontinuity between the two areas, so it seems necessary to mitigate this result.

REFERENCES

1. Gepshtein, Shai, and Yosi Keller, "Image Completion by Diffusion Maps and Spectral Relaxation," *IEEE Transactions on Image Processing* 22.8, 2013, pp. 2983-2994.
2. Fang, Chih-Wei, and Jenn-Jier James Lien, "Rapid Image Completion System Using Multiresolution Patch-based Directional and Nondirectional Approaches," *IEEE Transactions on Image Processing* 18.12, 2009, pp. 2769-2779.
3. Iizuka, Satoshi, Edgar Simo-Serra, and Hiroshi Ishikawa, "Globally and Locally Consistent Image Completion," *ACM Transactions on Graphics (ToG)* 36.4, 2017, pp. 1-14.
4. Isola, Phillip, et al., "Image-to-image Translation with Conditional Adversarial Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
5. Yu, Jiahui, et al., "Free-form Image Inpainting with Gated Convolution," *Proceedings of the IEEE International Conference on Computer Vision*, 2019.
6. Yu, Jiahui, et al., "Generative Image Inpainting with Contextual Attention," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
7. Goodfellow, Ian, et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, 2014.
8. Mirza, Mehdi, and Simon Osindero, "Conditional Generative Adversarial Nets," *arXiv preprint arXiv:1411.1784*, 2014.
9. Pathak, Deepak, et al., "Context encoders: Feature Learning by Inpainting," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
10. Gulrajani, Ishaan, et al., "Improved Training of Wasserstein GANs," *Advances in Neural Information Processing Systems*, 2017.
11. Arjovsky, Martin, Soumith Chintala, and Léon Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.
12. Liu, Guilin, et al. "Image Inpainting for Irregular Holes Using Partial Convolutions," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
13. Aditya Khosla et al., Stanford Dog Datasets, Available: http://vision.stanford.edu/aditya86/ImageNetDogs/
14. Bolei Zhou et al., Places Dataset, Available: http://places2.csail.mit.edu/index.html

AUTHORS PROFILE

GileHee Choi received his bachelor's degree in Electronics and Communication Eng. from Kwangwoon University in 2018 and received master's degree from Kwangwoon University in 2020. Her research interests are in Image Processing, Object Detection, Image Completion and Deep Learning algorithms.

Kyounghak Lee is Associate Professor at Kwangwoon University. He received Bachelor's Degree, Master's Degree and Ph.D in Kwangwoon University. He was senior researcher at KEIT from 1994 to 2011 and assistant professor at Namseoul University from 2011 to 2016. His research interests are in vision system and AI platform.

Hyung-Hwa Ko is Professor at Kwangwoon University. He received Bachelor's Degree, Master's Degree and Ph.D from Seoul National University. He joined Kwangwoon Univ. as a faculty member from 1985, and he served as dean of college of Electronics and Information Eng. from 2008 to 2010. His research interests are in Image Compression, JBIG2, HEVC, Machine Learning and Deep Learning applications.