eISSN: 2976-1190

Journal of Engineering and Sciences

Vol. 2 Issue 1 November 30, 2023

Journal homepage: **Journal of Engineering and Sciences (nepjol.info)**

Publisher: Research Management Cell, Pashchimanchal Campus, Lamachour, Pokhara

Sketch to Image Translation using Generative Adversarial Network

Ramchandra Giri[1*], Badri Raj Lamichhane[1], Biplove Pokhrel[1]

[1] *Department of Electronics and Computer Engineering IOE, Pashchimanchal Campus, Tribhuvan University, Nepal*

(Manuscript Received:30/08/23; Revised:/26/10/2023; Accepted: 02/11/2023)

Abstract

Using a Generative Adversarial Network (GAN) has proven its ability to successfully implement realistic images in image translation fields. It has its successful implementation in the sketch-to-image translation, too. Generative adversarial networks are widely used for the purpose of image translation. Most discriminators in generative adversarial networks use encoder or decoder blocks for image segmentation and classification tasks. U-net-based architecture is mostly used in the generator but rarely in the discriminator. If used in the discriminator, it is used for image resolution increment and segmentation tasks. In this research, a U-net-based discriminator is used for image translation tasks. U-net-based discriminator uses local and global differences between the real and fake images, which helps maintain global and local data representation. Resnet-9, used in the generator, uses skip connections, shortcuts, and concatenations, enabling information to flow from earlier to later layers. This helps preserve the original image features and solves the vanishing gradient problems in normal generators. The use of a strong discriminator and effective generator helps in the improvement system's performance. The available dataset was unpaired at the same time. Datasets from various sources were combined and formed a sketch-image pair. The input is a 512x256 human sketch and a corresponding real image pair. The image pair is split into sketch and image with dimensions 256x256. The system's output is the human face image of the corresponding sketch.

Keywords: Concatenation; Generative Adversarial Network; Resnet-9 Generator; Skip Connections; U-Net Discriminator

1. Introduction

1.1 Background

Generating an image from the sketch is an important task in computer vision. The human sketch-to-image translation is very useful in criminal investigation and image identification tasks. We propose a GAN method to generate real human face images from the human sketch. A network is trained to generate

impressive real human face results from many sketch image pairs. Generative adversarial network(GAN) is a method that generates a new set of data based on the training dataset. It consists of two blocks Generator and discriminator which compete with each other for capturing, copying and analyzing the variations in a dataset. The generator describes how data is generated using the probabilistic model. It explains the visualization of the data in the dataset. The term adversarial means the model is trained in an adversarial manner. GAN uses neural networks for the purpose of training. Noise is added in GAN, increasing the chances of misclassifying the images by the discriminator. A small rise in noise leads to implementing something neural networks can start visualizing new patterns like sample train data. GAN generates new fake results similar to the original data. Discriminators identify the abnormalities in the images generated by the generator and classify them as fake or real. This competition between the generator and the discriminator continues until perfection is achieved, in which the generator wins by fooling the discriminator for the classification of the image.

For sketch-to-image translation, the sketch provides less profile information, and the realistic image provides high-level semantic information such as color and details. The proposed methodology network provides the attribute information for reconstructing a color photographic face or image from the sketch image used to generate high-level semantic information. A combination of GAN and skip connection layers is used in the convolution layers of the generator and the discriminator. The output of the convolution layer is back-propagated and simultaneously concatenated with the next layer. A discriminator is a supervised approach trained on real data that provides feedback to the generator. It is a classifier that classifies the image as real or fake based on the data present on the dataset.

[*] Corresponding author. Tel.: +977- 9856032003
E-mail address: ramchandragiri48@email.com

The main aim of the GAN is to study the training examples and the probability distribution, which could generate more samples from the learned data. In most of the research, U-Net architecture is used in the generator part of the GAN and least used in the discriminator part. Using U-Net architecture in the discriminator helps maintain global and local realism. U-Net-based discriminator performs per image classification and per pixel class decision, providing coherent decision to the generator. Patch GAN in the discriminator[1,2] uses an encoder or decoder for image classification. This only maintains either local or global realism. As a discriminator, U-Net uses both an encoder and decoder that helps maintain local and global realism. Using local and global realism helps to increase the performance of the discriminator in image classification tasks. The stronger the discriminator is, the more effective the generator becomes. Resnet-9 used in the generator uses concatenation and skip connection that helps to solve the gradient vanishing problem[5].

1.2 Related Work

Image translation using generative adversarial network Image-to-image translation with conditional Adversarial Networks by Philip Isola, Jun Yan Zhu, Tinghui Zhou and Alexei A.Efros 2018[1] is the pioneer. They used different datasets like cityscapes label to photos, architectural labels to photos, black and white to color, edges to shoes, day to night, and thermal to color photos for image translation. Conditional pix2pix GAN is used as an architecture. U-net architecture with skip connection is used in the generative part and patch GAN with receptive field is used in the discriminator for image translation. Most of the research is compared with the outcome of this research.
Comparing GANs for Translating Satellite Images to Maps by Arnav Parekhji, Mansi Pandya, and Pratik Kanani 2021[2] used cycle GAN and pix2pix GAN for the purpose of satellite image to map translation. The result obtained from both the models is compared with the help of the FID score and training time. On the basis of comparison, they found that the FID score for Pix2pix is less than that of the cycle GAN. Also, the training time of cycle GAN was more than that of the pix2pix GAN for 80 epochs. Finally, they concluded that pix2pix is a better method for image-to-image translation as compared to the cycle GAN based on the FID calculation and training time Image to image translation with conditional-GAN by Jason Hu,Weini YU and Zhouvchangwan Yu 2018 used Generative adversarial network for aerial to map conversion. They used Resnet architecture for the generator and patch GAN with a receptive field for the discriminator architecture. The result was compared with the result obtained from the pre-trained C-GAN model from Isola [1].
A U-Net-Based Discriminator for Generative Adversarial Networks by Edgar Schonfeld and Bernt Schiele 2020[3] used a U-net-based discriminator for image translation. The result obtained from the U-net-based discriminator was compared with the baseline paper, which used big Gan architecture. The comparison between these two concluded that the average improvement is found on the U-net discriminator that big GAN. The comparison was based on the FID points across FFHQ, CelebA, and COCO animals' datasets. This type of discriminator provides global and local feedback to the generator. U-net has demonstrated state-of-the-art performance in various tasks. Using a U-net discriminator can recover more details than the normal encoder discriminator. The stronger the discriminator is, the more powerful the data representation, making the generator's task of fooling the discriminator more challenging and improving the quality of samples of generated images.
Sketch2face: Generative Adversarial Networks for Transferring Face Sketches into Photo Realistic Images by Julia Gong and Matthew Mistele,2021[4] used conditional GANs iterative Refinement with four variants.
Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks by Jun-Yan Zhu,Taesung Park, Philip Isola and Alexei A.Efros 2020 used [6] cycle GAN for image translation for unpaired images. Unpaired labels to photos and maps to image datasets are used for image translation. A convolutional layer with residual blocks is used in the discriminator, and patch GAN is used for image conversion.
Generating Photographic faces from the Sketch Guided by attribute using GAN by Jian Zhao,Xie Xie,Lin Wang, Meng CAO and Miao Zhang 2019[7] used a generative adversarial network for image translation. They used convolutional layers with skip connection for the generator and Stride 2 convolutional with batch normalization followed by leaky Relu in the discriminator for image translation. The result is compared with peak signal-to-noise ratio(PSNR) and structural similarity(SSIM).

1.3 Contribution:

The main contribution of this research is using U-Net architecture in the discriminator for image translation tasks. U-Net architecture is mostly used in the generator part and less used in the discriminator part. If used in the discriminator, it is used for image segmentation and resolution increment tasks. Most discriminators use the encoder or decoder part for image translation tasks. However, the U-Net-based discriminator used an encoder and decoder for image translation, which helps to maintain local and global realism. Resnet-9, used in the generator, uses skip connections, shortcuts, and concatenations, enabling information flow from earlier to later layers. This helps to preserve the original image features and solves the vanishing gradient problems. The use of a strong discriminator and effective generator helps in the improvement system's performance. The available datasets used in the research were first unpaired. They must be paired so the system learns from the sketch and corresponding real image. Different datasets are combined and resized to fit the system. The system performance is improved based on SSIM and PSNR metric evaluations.

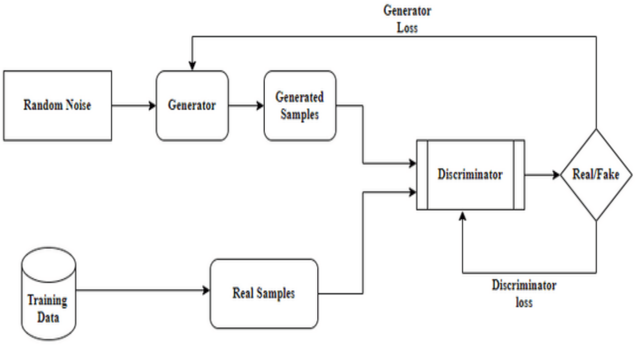
2. Materials and Method

2.1 Methodology

In this research, we use the generative adversarial network. It uses an architecture that comprises a generator and discriminator. The generator outputs a new plausible synthetic image, and the discriminator classifies the data as real or fake. The discriminator is updated automatically, but the generator is updated by the discriminator. Two models are trained simultaneously in an adversarial manner. The generator seeks to fool the discriminator better and the discriminator seeks to identify the counterfeit images better. The output image generation is based on an input, i.e., a source image. The source and the target image are provided to the discriminator, which determines whether the target is a plausible transformation of the source image. In normal GAN, images are generated by a generator on the basis of a random vector and the discriminator classifies it as real or fake with the help of a real sample. The loss function updates the generator and discriminator weights.

Figure 1:Normal GAN

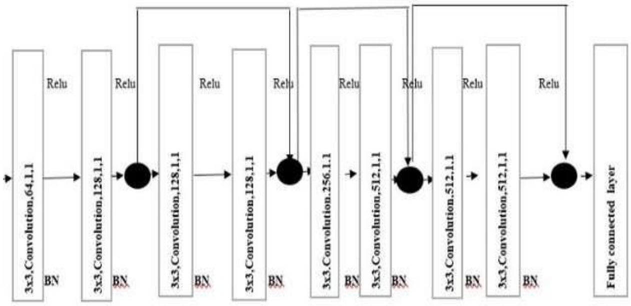
2.1.1 Resnet-9 Generator



Residual network is the network that helps to solve the vanishing gradient problem. Residual blocks, along with the skip connections, are used in this network. Residual blocks are made by connecting the activations of a layer to the next layer using skip connections. The sketch image of size 256x256 is fed in the residual network and convolution operation along with the batch normalization is done to generate the image. Skip connections are also used for the purpose of generation. It uses 9 convolutional layers, batch normalization, relu activation function and skip connections to solve the vanishing gradient problem. Batch normalization used in resnet is a technique that increases the stability of a network. It helps normalize the values in a certain range such that the neural network works faster and provides efficient results. Convolution operation in resnet is a simple matrix multiplication having weights called kernel. The kernel moves over the data, multiplying each element of the input. The product is summed up to generate single-pixel output. This results in a new 2D matrix creation. The kernel size determines the input features combined to produce the output. Convolution is customized by changing a number of filters, size, shape, padding, stride and activation functions applied to the output.

Figure 2: Resnet-9 Generator

2.1.2 U-Net Discriminator



The U-net discriminator performs classification based on pixel-pixel and segments the image as real and fake with the help of generated and real images. The encoder performs per image classification and the decoder performs per pixel class decision that provides coherent feedback to generator. Thus, the discriminator is enabled to learn the global and local difference that exists between the real and fake images. The loss on the discriminator is calculated:

$$\mathrm{L}_{\mathrm{D}}^{\mathrm{U}} = \mathrm{L}_{\mathrm{D}}^{\mathrm{U}}\mathrm{enc} + \mathrm{L}_{\mathrm{D}}^{\mathrm{U}}\mathrm{dec} \qquad (1)$$

The loss in the encoder of the discriminator is calculated from the scalar output as

$$\mathrm{L}_{\mathrm{D}}^{\mathrm{U}}\mathrm{enc} = -\mathrm{E}_{\mathrm{X}}[\log \mathrm{L}_{\mathrm{D}}^{\mathrm{U}}\mathrm{enc}(\mathrm{x})] - \mathrm{E}_{\mathrm{Z}}[\log[1 - \mathrm{L}_{\mathrm{D}}^{\mathrm{U}}\mathrm{enc}(\mathrm{G}(\mathrm{Z}))]] \qquad (2)$$

And the loss in the decoder of the discriminator is calculated as mean decision over all the pixels as

$$\mathrm{L}_{\mathrm{D}}^{\mathrm{U}}\mathrm{dec} = -\mathrm{E}_{\mathrm{x}}[\sum \log(\mathrm{D}_{\mathrm{dec}}^{\mathrm{U}}(\mathrm{x})_{\mathrm{i,j}} - \mathrm{E}_{\mathrm{z}}[\sum \log[1 - \mathrm{D}_{\mathrm{dec}}^{\mathrm{U}}[\mathrm{G}(\mathrm{z})]_{\mathrm{i,j}} \qquad (3)$$

$[D_{dec}^{U}(x)]_{i,j}\ and[\ D_{dec}^{U}(G(z))]_{i,j}\ \ refers\ to\ the$

Discriminator output at pixel i,j

2.1.3 Discriminator with skip connections and system architecture:

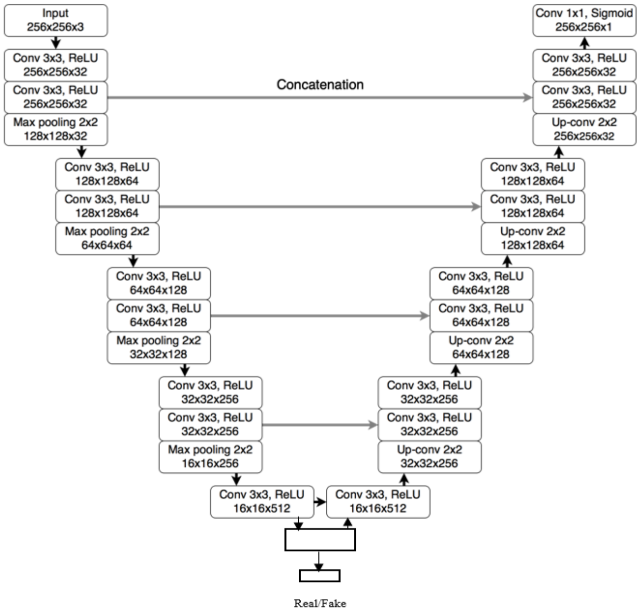


Figure 3: U-Net discriminator Architecture

Mapping of a high-resolution input grid to a high-resolution output grid is the defining feature of the image-to-image translation. We design discriminator architecture in which the input image is roughly aligned with the structure in the output. The input and output images differ in their surface appearance. Skip connection is used in the architecture to share low-level information between the input and the output. Skip connection is followed by the general shape of the u-net architecture. Skip connections are added between each layer i and layer n-i. Where n is the total number of layers, each skip connection concatenates all channels at layer i with those at layer n-i. The overall system architecture includes a resnet-9 generator and U-Net discriminator. The generator generates the images based on the inputs fed to the U-net-based discriminator and the real input. The discriminator classifies the image as real or fake based on the real input of real samples. Generator and Discriminator are trained with loss functions. Binary Cross entropy and L1 loss are taken for the generator, whereas encoder and decoder loss are taken for the discriminator to update their weights.

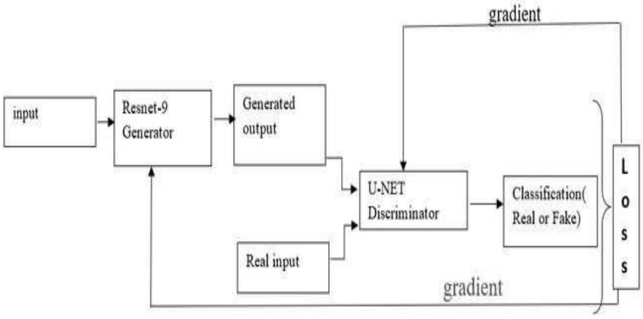


Figure 4: System Architecture

2.2 Datasets

For this research, person face sketch dataset, pretty face dataset and CUHK dataset is combined into a single dataset. The datasets were combined to make a larger dataset of different kinds. The dataset was unpaired, and the size of each image was 512 x 512. The dataset was first paired individually such that image and sketch forms a pair. And the dataset was resized into 512 x 256. The number of datasets used for the training purpose was 4000 and 250 was used for the testing purpose. A random flip was done for data augmentation. The dataset is converted to pixel values ranging from -1 to 1.

Figure 4: System Architecture

3. Results and Discussion

The constructed model is able to translate the sketch of an image into a real image. The output is obtained by using the following hyper parameters.
Input Image Size:512x256(Paired Image)
Real Input Size: 256x256
Batch Size: 1
Learning Rate: 0.0004 for both Discriminator and Generator
Optimizer: Adam’s Optimizer
Training Steps: 50 thousand
Discriminator loss: Encoder loss +decoder Loss
Generator Loss: Binary Cross Entropy Loss + L1
The image translation model was built and successfully tested with different parameters. The of the model output of the system is then visualized by plotting losses on different hyper parameters. The optimum result was obtained at 5000 thousand training steps. Due to the limited resources and time, we used only 4000 datasets.

3.1 Output

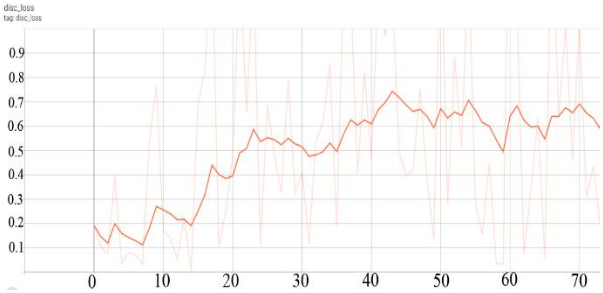


Figure 5: Discriminator loss (x-axis: steps in thousands, y-axis: loss value)

The principal objective of GAN is to train the generator network to produce realistic human face images. The discriminator is trained for classifying the image as real of fake produced by the generator. At first, the generator produces samples that are different from the real samples. The discriminator easily classifies these samples as real or fake. So, initially, the discriminator loss is lower than the generator loss. As training steps increase, generator networks learn to produce more realistic samples to improve the discriminator network for classification. As a result, discriminator loss increases and generator loss decreases. The initial low discriminator loss and high generator loss are the part of GAN training which is expected to happen. Training must be continued until the losses reach a steady state. The total GAN loss, discriminator loss and generator loss, along with outputs at different training steps, are shown below.

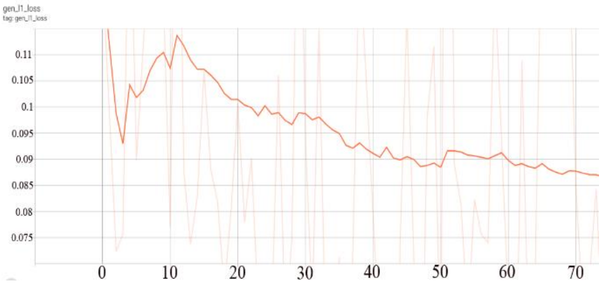


Figure 6: Generator loss (x-axis: steps in thousands, y-axis: loss value

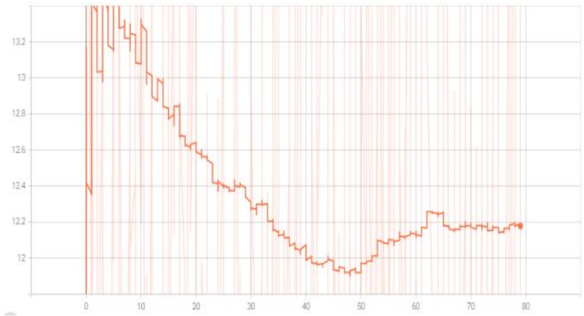


Figure 7: Total GAN loss (x-axis: steps in thousands, y-axis: loss value

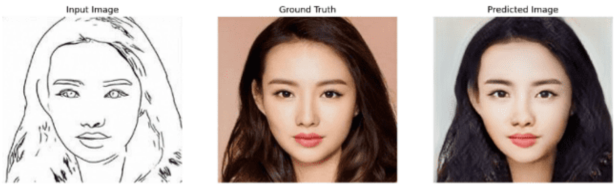


Figure 8: Output at 50K training steps

3.1 Evaluation

Peak Signal-to-Noise Ratio (PSNR) is a commonly used evaluation metric for image quality in GANs (Generative Adversarial Networks), including the pix2pix model. PSNR measures the difference between the generated image and the ground truth image in terms of the peak signal-to-noise ratio, which is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation.
PSNR is calculated as follows:

PSNR = 20 * log10 (MAX_I / sqrt (MSE)) (4)
Where MAX_I represents the difference between the maximum and minimum allowed pixel values, and MSE is the minimum square error.
The structural similarity index (SSIM) is used to find the similarity between the real and predicted image. It helps to find the pixel inter-dependencies which carries important information about structural information of the images. It measures perceptual differences between real and predicted images. Luminance, contrast, and structural information are considered as the important parameters for quantitative measurement. A score from -1 to 1 is generated by SSIM, in which 1 means images are indistinguishable and -1 means they are entirely different.

Table 1: Metric Comparison

Models	PSNR	SSIM
U-net generator and Patch GAN discriminator[1]	-	0.487
Conditional GANs with Iterative Refinement[4]	-	0.608
Generating Photographic Faces from the Sketch Guided by Attribute Using GAN Model [6]	16.306	0.579
Resnet-9 Generator and U-Net Discriminator	**28.478**	**0.647**

4. Conclusions and Future Works

4.1 Conclusions

In most of the research, the generator architecture is changed for the purpose of the performance comparison between the models, but the least change was made in the discriminator architecture. U-net based architecture is mostly used in the generator and rarely used in the discriminator. If used in the discriminator it is mainly for image generation and resolution increment tasks. In this research, U-net discriminator is used for image translation tasks. It uses local and global differences between the real and fake images. The use of a strong discriminator and effective generator architecture helps in the improvement of the performance of the system.

4.2 Future Works

This research contains only human sketches for realistic image translation. Future work may include translating other sketches into realistic images, including human sketches. Also, diversified data sets with larger numbers may be used for training purposes, increasing the system's performance.

Acknowledgment

We are highly indebted to the Departments of Electronics and Computer Engineering, Pashchimanchal Campus, Lamachour, Pokhara, for providing us the opportunity for research so that we could explore our capabilities in implementing real-world applications.

References

[1] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern* recognition, (2018) 1125-1134.

[2] Parekhji, A., Pandya, M., & Kanani, P. Comparing GANs for translating satellite images to maps. In 2021 5th *International Conference on Computing Methodologies and Communication (ICCMC), IEEE*, (2021) 1267-1274.

[3] Schonfeld, E., Schiele, B., & Khoreva, A. A u-net-based discriminator for generative adversarial networks. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* (2020) 8207-8216.

[4] Julia Gong, Matthew Mistele,2021. Sketch2face: Generative Adversarial Networks for Transferring Face Sketches into Photo Realistic Images.

[5] Hu, J., Weini, Y.U., and Zhouvchangwan Y. Image to image translation with conditional-GAN, CS230:Deep learning, Spring 2018, Standford University,CA (2018).

[6] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, (2017) 2223-2232.

[7] Zhao, J., Xie, X., Wang, L., Cao, M., & Zhang, M. Generating photographic faces from the sketch guided by attribute using GAN. *IEEE Access*, 7 (2019) 23844-23851.

[8] Li, Z., Deng, C., Yang, E., & Tao, D. Staged sketch-to-image synthesis via semi-supervised generative adversarial networks. *IEEE Transactions on Multimedia*, 23 (2020) 2694-2705.

[9] Arefeen Sultan, K. M., Imrul Jubair, M., Nahidul Islam, M. D., & Hossain Khan, S. (2021). toon2real: Translating Cartoon Images to Realistic Images. arXiv e-prints, arXiv-2102.

[10] Ghosh, A., Zhang, R., Dokania, P. K., Wang, O., Efros, A. A., Torr, P. H., & Shechtman, E. Interactive sketch & fill: Multiclass sketch-to-image translation. *In Proceedings of the IEEE/CVF International Conference on Computer Vision,* (2019) 1171-1180.

[11] Jiang, S., Yan, Y., Lin, Y., Yang, X., & Huang, K. Sketch to Building: Architecture Image Translation Based on GAN. *In Journal of Physics: Conference Series*, 2278(1) (2022) 012036.

[12] Liu, Y., Zhao, Q., & Jiang, C. Conditional image generation using feature-matching GAN. *In 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE,* (2017) 1-5.

[13] Lu, Y., Wu, S., Tai, Y. W., & Tang, C. K. Image generation from sketch constraint using contextual GAN. *In Proceedings of the European conference on computer vision (ECCV),* (2018) 205-220.

[14] Jo, Y., Yang, S., & Kim, S. J. Investigating loss functions for extreme super-resolution. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, (2020) 424-425.