Analysis of SRGAM to Upscalling CCTV Image

1st Rahmatur Ramadhan
Faculty of Informatics
Telkom University
Bandung, Indonesia
rahmaturramadhan@students.telkomuniversity.ac.id

2nd Ema Rachmawati
Faculty of Informatics
Telkom University
Bandung, Indonesia
emarachmawati@telkomuniversity.ac.id

3rd Muhammad Haris
Faculty of Informatics
Telkom University
Bandung, Indonesia
harisrachmat@telkomuniversity.ac.id

Abstract
Detailed information from high-resolution images is needed when analyzing the content contained in the image. However, sometimes the image has a low resolution so that the image is difficult to interpret. Based on these problems, a research was carried out on improving the quality of image resolution especially in forensic field. CCTV video results that have low-quality images and degraded with many noises, bad illumination, distortions, and blurs. So, we can remove that using Super-Resolution (SR) method. This can help during to identification, image interpretation, and analysis process clearly. Using Super-Resolution methods, a high-resolution image is obtained from a set of low-resolution images. The research was conducted based on the Super-Resolution Generative Adversarial Network (SRGAN). SRGAN is a generative model method that can generate data (images) with good quality. From the experiments that have been carried out, the system that has been built is proven to be able to produce images with good quality with the highest PSNR value is 24,873 and the highest SSIM value is 0.831.

Keywords*: CCTV, Super-Resolution, Generative Adversarial Network, PSNR, SSIM*

I. INTRODUCTION

A. Background

Nowadays development of technology, information is very easy to obtain. All fields related to technology will be in line with the development of technology itself. One of the uses of information related to the representation of the use of images as authentic evidence or forensic evidence. Image is one component of multimedia has important role in the development of the latest information technology. Image has richer information so it can make it easier to delivering messages. However, not all images are in good quality, so they cannot display clear information. Low resolution images require improvement to produce images that have a higher resolution [1].

High resolution images are needed in various fields such as the medical field, target detection, and CCTV identification. For example, high-resolution medical images will greatly assist doctors in diagnosing patients in more detail. In addition, in the field of CCTV identification, high-resolution imagery also facilitates the image recognition process as well as the image data storage process in its stages. Image data storage process requires image reconstruction and restoration techniques to strengthen the resulting image so that it can be read. This is useful as evidence in presenting criminal law cases in places that have CCTV such as shops, banks, hotels and residental area.

A high-resolution image is a representation of an image with a high pixel density. Images have more detailed information and can help characterize an object from its equivalent image. The capabilities of a Computer Vision pattern recognition system can also be improved if high resolution imagery is available. One way to overcome the lack of high-resolution image availability is through the development of the Super-Resolution method. Super-Resolution algorithms generally solve this problem with a two-step approach - estimating motion between different images and projection from low resolution pixel values to high resolution grids [2]. The Super-Resolution reconstruction technique is fundamentally able to form one high-resolution image output based on a series of low-resolution images with different times for each low-resolution series of images [3].

B. Identification and Scope of the Problem

Based on the background that has been mentioned, the formulation of the problem in this research is how to implement the GAN method. to reconstruct CCTV images to have high resolution with PSNR and SSIM values exceeding the average PSNR and SSIM low resolution images. The system input is a low-resolution CCTV image, and the system output is a high-resolution CCTV image reconstructed by the generator network.

Some of the limitations of the problem in this study are:

1. The dataset is CCTV dataset which has been aligned and cropped.
2. The dataset used consists of 2000 non-CCTV images and 500 CCTV images.
3. The image is converted to a size of 640 X 640 pixels to be used as a label image and then done downscaling by a factor of 4 to a size of 160 x 160 pixels to be used as an input image
4. Only focuses on the implementation of the GAN method for the SISR case.

C. Purposes

The purpose of this study is to build a Single Image Super Resolution system using the GAN method to reconstruct low resolution CCTV images into high resolution and have PSNR and SSIM values that exceed the average PSNR and SSIM of low-resolution images.

D. Writing Organization

The order of writing this report is as follows: section 2 describes the studies related to this research. In section 3, the details of the system built to handle SISR cases using the OAK method will be explained. In section 4, we will discuss the test results and system evaluation. And the last part contains conclusions and suggestions for development for further research.

II. LITERATURE REVIEW

A. Single Image Super Resolution

SISR aims to reconstruct a low-resolution image into a high resolution image that has better detail. The methods used to deal with this problem can be categorized into three different approaches. The first approach is interpolation-based which is the easiest technique, namely applying interpolation to the sample data, but this method cannot restore high-frequency details from an image. The second approach is reconstruction based, by performing smoothing and downsampling on high-resolution images, and then utilizing these data to generate high-resolution images. In some cases, the resulting image has a lot of noise and artifacts. The third approach is example-based, which uses machine learning models to reconstruct the image. By far the example-based method is the best method to handle the SISR case. The previous exemplar-based method [4][5] studied the relationship between low resolution images and high resolution images. Several developments from this exemplar-based method include sparsity-based [6] and regression-based [7] methods.

After the development of deep learning methods, several studies began to implement CNK-based methods for the SISR case. The first deep CNN method for the SISR case, the Super Resolution Convolutional Neural Network (SRCNN) was introduced by Dong et al. in 2014 [8] and resulted in a significant performance improvement compared to the previous methods. a new CNN architecture developed either to improve the quality of the resulting image or to improve the training performance of the system.

B. Convolutional Neural Network

CNN has been a popular method since Krizhenky et al. [9] introduced "AlexNet" and won the ImageKet competition in 2012. Since then CNN has become a widely used method for computer vision cases such as image and ideo classification [10][11], object detection [4], segmentation object [5], face recognition [12] and estimation of human body pose [13]. Several factors that influence the development of this CNN method include: (i) the development of a Graphic Processing Unit (GPU) to speed up the training process, (ii) the emergence of a better regularization model, and (iii) the availability of datasets for a wider range of cases.

In the SISR system, recent research has implemented several variations of deep CNN architecture. In 2016, Dong et al. [10] introduced the Fast Super-Resolution Convolutional Neural Network (FSRCNN) implemented the hourglass-shaped CNN structure and succeeded in increasing the performance of the training process up to 40 times faster than the SRCNN method [4]. Another architecture that also managed to get better performance One of the best is the deeply-recursive CNN (DRCN) which was introduced by Kim et al.[11]. DRCN was successful in improving training performance without excessive parameters.

C. Generate Adversarial Network

Ian Goodfellow et al. [14] introduced the GAN method in 2014. The GAN method is a new training procedure for generative modeling that successfully generates realistic images from random noise sets. The GAN method consists of generators and discriminators that are trained simultaneously and compete during the training process. The discriminator is trained to distinguish between the original image and the artificial image, while the generator is trained to generate an image as close as possible to the original data so that it cannot be distinguished by the discriminator. Ledig et al. [15] implemented the GA method for the SISR case and introduced a perceptual loss function consisting of adversarial loss and content loss. Adversarial loss aims to make the generated image similar to the original image, and content loss aims to restore image details based on the similarity of features between the reconstructed image and the original image.

III. METHOD OF RESEARCH

A. System Overview

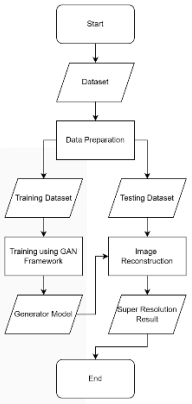

Figure 1 Framework System

The system development in this research consists of several stages. First, preprocessing the dataset is carried out by cropping and resizing the image to a size of 640 x 640 to be used as a label. Then the label image is resized to 160 x 160 to be used as an input image. Then the dataset is divided into training data and test data. Furthermore, SRGAN training was carried out using the GAN method. After the training process is complete, the generator model is taken for testing. Finally, the model generator is tested to reconstruct the test image and generate high resolution images which will then be evaluated.

B. System Architecture

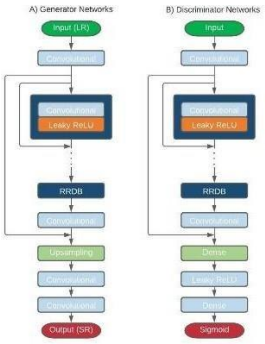

Figure 2 GAN Architecture

Generator Network architecture in general can be seen in Figure 2. The network schema is inspired by the lnceptionResNetV2 architecture introduced by Szegedy et al. [16]. The difference in architecture lies in the number and scheme of network blocks used. In addition, the Reduction block in the InceptionResNetV2 architecture is changed to an upscaling block because the generator network is designed to increase the resolution of the input image. The generator network input is an image with a resolution of 16 x 16 pixels. First, the stem block will extract low-level features from the input image, then it will be linked with the inception-residual block and upscaling block alternately. At the end of the network, 3 (three) convolution layers are used to rebuild the features into an image form. Each convolution layer uses the ReLU activation function, except for the last convolution layer using the hyper-tangent activation function.

Discriminator Network, the CNN architecture from Ledig et al is used. [15] with the change in the last laver before sigmoid, no dense layer was used but was replaced with global average pooling. The network consists of 7 convolution layers with a kernel size of 3 x 3. the number of kernels is increased by multiples of 2 from 64 to 512 as in the VGG architecture [17], followed by two convolution layers I x 1 and global average pooling. Each convolution layer uses the ReLU activation function, and at the end of the network the sigmoid function is used to obtain the classification probability.

C. Loss Function

To train the *generator* and *discriminator* networks simultaneously, adversarial loss is used which is the main *loss* function of the GAN method [14]. *Discrminator* D is trained to maximize the probability to determine the correct label for the original data and the data generated by *generator* G, while *generator* G is trained to minimize *log(1 - D(G(z)))* where *D(x)* is the probability that data x is the original image. : represents the low resolution image and *G(z)* represents the reconstructed image. Mathematically, the adversarial loss function is as follows:

$$\min_G \min_D \Rightarrow_{x-P_{data}(x)} [\log D(x)] + \Rightarrow_{z\sim P_z(z)} [\log(1 - D(G(z)))]$$

This formulation allows the generator to produce an image that is similar to the original image and is difficult to distinguish by the discriminator, and to improve the quality of the reconstructed image.

D. Testing Scenario

In this study, experiments were conducted to analyze the system performance. An experiment was conducted on the loss function to determine the effect of using different loss functions on cysteine. To determine the effectiveness of system training using the GAN method and the effect of using additional loss functions, four scenarios were carried out:

1. Generator was trained without discriminator and not using adversarial loss.
   a) Training datasets consist of 2000 non-CCTV images. Testing datasets consist of 500 CCTV images.
   b) Training datasets consist of 1500 non-CCTV images and 500 CCTV images. Testing datasets consist of 500 non-CCTV images.
2. Train the generator and discriminator networks simultaneously using adversarial loss.
   a) Training datasets consist of 2000 non-CCTV images. Testing datasets consist of 500 CCTV images.
   b) Training datasets consist of 1500 non-CCTV images and 500 CCTV images. Testing datasets consist of 500 non-CCTV images.

E. System Performance Measurement

3.5.1 PSNR (Peak Signal-to-Noise Ratio)

PSNR (*Peak Signal-to-Noise Ratio*) is used to measure the level of damage to the original image with noise as the human perception of the image. The higher the PSNR, the better the denoized image, especially with the same compression code. MSE (mean squared error) between the original image and the reconstructed image. Given m x n independent MSE is defined as.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} ||X(i) - X_{SR}(i)||_2$$

PSNR in decibel dB is defined as:

$$PSNR = 10.\log_{10}(\frac{L^2}{MSE})$$

3.5.2 SSIM (Structural Similarity Index Measure)

SSIM index quality assessment index is based on the calculation of three factors; luminance, contrast and structure. Can be defined as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_xy + C_2)}{(2\mu_x\mu_y + C_1)(2\sigma_xy + C_2)}$$

SSIM values range from 0 to 1 where 1 means a perfect match between the original image and the copy.

IV. RESULT OF RESEARCH

A. Testing Result

The optimizer used during the training with momenthun β_1 = 0.9, β_2 = 0.999, initial learning rate Ir = 0.0002 and 500 batches, with batch size 16. Training is using NVIDIA GTX 1050 TI.

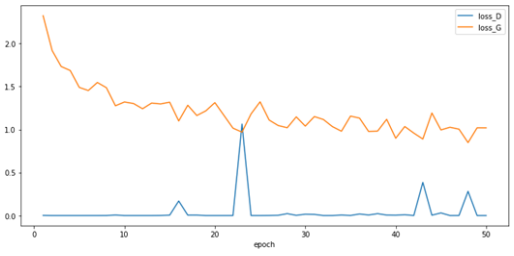

Figure 3 Training Results

As shown Figure 3 result of the training using the optimizer that already defined, in pixel-wise loss almost achieves convergence after 23 epochs. The additional reconstruction task with pixel-wise loss suggests a fast and stable training manner between the generator and the discriminator of GAN.

Scenario	1a	2a	1b	2b
PSNR	24.318	24.727	24.652	24.873
SSIM	0.769	0.811	0.782	0.831

Table 1 Performation Measurement Results

In Table 1 the highest PSNR and SSIM values are obtained from *scenario 1* only using generator and discriminator. However, the difference in PSNR and SSIM values in *scenario 2* is not very significant. And if the 640 x 640 directly without a reconstruction process and then the PSNR value is calculated against the label image, the average value of PSNR that obtained is 24,740 and the average value of SSIM that obtained is 0,798.

Figure 4 Image Input (Low Resolution)

Figure 5 SRGAN Result (*Scenario 1*)

Figure 6 SRGAN Result (*scenario 2*)

In Figure 5 and Figure 6 can be seen that in general, both scenarios have succeeded in reconstructing the image quite well and have more texture information. However, when viewed in more detail, the reconstructed image of the model from *scenario 1* has a smoother texture than the other scenarios. *Scenario 2* produces an image that has a more detailed texture than *scenario 1*, but sometimes the resulting image has unrealistic artifacts. Therefore, in *scenario* 2, the reconstructed image has the best texture compared to *scenario 1*.

B. Analysis of Testing Result

From the experimental results on the use of the loss function, it is shown that the reconstructed image of the model from *scenario 1* has a smooth texture and less sharp detail but has the highest PSNR and SSIM values. This is because in *scenario 1*, which should use loss function, the resulting image is required order·has a pixel value as close as possible to ground truth. In *scenario 2*, the resulting image is more detailed because the generator is trained to "dumb" the discriminator during the training process. In *scenario 2*, the generator is not only required to be able to "fool" the discriminator but also to have features similar to the original image, so that in some cases where the resulting image from *scenario 2* has artifacts, it can be overcome in scenario. And in loss function helps the generator to prevent noise and enrich the texture, so that the resulting image has better detail and texture. For *scenario a* and *scenario b*, the PSNR and SSIM result is not that significant, but the value of *scenario b* is more than *scenario a*, because scenario b is training the CCTV dataset. So, the result in *scenario b* better than *scenario a* since the experiment is using real CCTV images.

V. CONCLUSION

In this study, a Single Image Super Resolution system was created to reconstruct cctv images using the GAN method. The network generator that is built can reconstruct cctv images with good results using the inception-residual concept where visual information from the input image is extracted at different scales at the same time. The use of loss functions can help improve the quality of the resulting image. The experimental results prove that the system built is able to generate high-resolution images with realistic quality and texture, has good detail and low noise, the highest PSNR value is 24.873 and the highest SSIM value is 0.831.

For future work, is to extend this framework into complete CCTV forensics identification process specially to enhance the human face based on GAN algorithm.

REFERENSI

[1] B. C. Tom, N. P. Galatsanos, and A. K. Katsaggelos, “Reconstruction of a High Resolution Image from Multiple Low Resolution Images,” *Super-Resolution Imaging*, pp. 73–105, 2005, doi: 10.1007/0-306-47004-7_4.

[2] J. Zhu, C. Zhou, D. Fan, and J. Zhou, “A new method for superresolution image reconstruction based on surveying adjustment,” *J. Nanomater.*, vol. 2014, 2014, doi: 10.1155/2014/931616.

[3] Q. He and R. Schultz, “Super-Resolution Reconstruction by Image Fusion and Application to Surveillance Videos Captured by Small Unmanned Aircraft Systems,” *Sens. Fusion its Appl.*, 2010, doi:

10.5772/9978.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[5] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 2980–2988, 2017, doi: 10.1109/ICCV.2017.322.

[6] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, 2010, doi: 10.1109/TIP.2010.2050625.

[7] R. Timofte, V. De, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1920–1927, 2013, doi: 10.1109/ICCV.2013.241.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2016, doi: 10.1109/TPAMI.2015.2439281.

[9] B. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Cnn实际训练的," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2012.

[10] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9906 LNCS, pp. 391–407, 2016, doi: 10.1007/978-3-319-46475-6_25.

[11] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 1637–1645, 2016, doi: 10.1109/CVPR.2016.181.

[12] S. Sharma, K. Shanmugasundaram, and S. K. Ramasamy, "FAREC - CNN based efficient face recognition technique using Dlib," *Proc. 2016 Int. Conf. Adv. Commun. Control Comput. Technol. ICACCCT 2016*, no. 978, pp. 192–195, 2017, doi: 10.1109/ICACCCT.2016.7831628.

[13] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1653–1660, 2014, doi: 10.1109/CVPR.2014.214.

[14] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020, doi: 10.1145/3422622.

[15] C. Ledig *et al.*, "\href{https://ieeexplore.ieee.org/abstract/document/8099502}{Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network}," *Cvpr*, vol. 2, no. 3, p. 4, 2017, [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2017/papers/Ledig_Photo-Realistic_Single_Image_CVPR_2017_paper.pdf.

[16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," *31st AAAI Conf. Artif. Intell. AAAI 2017*, pp. 4278–4284, 2017.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.