Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 16, No. 1, 2025

Dual Generative Adversarial Network for Infrared and Visible Image Fusion

Zhi Wang *, Ao Dong [a]

College of Computer Science and Technology, Qingdao University, Qingdao, China
* **Corresponding author**: Zhi Wang (Email: wangzhi2210@163.com), [a]dongao613@163.com

Abstract: The objective of infrared and visible image fusion is to integrate the prominent targets from the infrared image with the background information from the visible image into a single image. Many deep learning-based approaches have been employed in the field of image fusion. However, most methods have not been able to sufficiently extract the distinct features of images from different modalities, resulting in fusion outcomes that lean towards one modality while losing information from the other. To address this, we have developed a novel method based on generative adversarial network for infrared and visible image fusion. We have designed two sets of generative adversarial networks. The first set is utilized for preliminary feature extraction, generating intermediate results and discriminating features with the infrared image. The second set is employed for deep feature extraction, generating the fused image and discriminating features with the visible image. Through the adversarial training of the two sets of generators and discriminators, we ensure the comprehensive extraction of diverse features from images of various modalities. Extensive qualitative and quantitative experimental results indicate that our approach can retain more information from the source images. Compared to seven other prominent methods, our approach achieves superior quality.

Keywords: Image fusion; Infrared image; Visible image; Generative adversarial network.

1. Introduction

The aim of image fusion is to extract information from different modalities and generate a single image that carries more comprehensive information and has better readability. Infrared and visible image fusion is an important research direction in this field. Infrared images and visible images exhibit significant differences in physical properties and information representation. Generally, infrared images carry rich thermal radiation information, focus more on prominent bright areas in the images, However, infrared images often overlook environmental information surrounding the targets. In contrast, visible images are better at expressing environmental information and details such as textures [1]. Therefore, fusing them can provide more comprehensive and accurate visual information, which is crucial for many applications.

In recent years, with the advancement of neural networks, deep learning-based methods have started to be applied in image fusion. Compared to traditional methods, deep learning-based methods have more powerful feature extraction capabilities. As image fusion tasks usually lack ground truth, Generative Adversarial Network (GAN) proposed by J. Goodfellow et al. [2] are suitable for unsupervised learning and perform better in image fusion tasks. Ma et al. introduced FusionGAN [3], the first application of GAN to image fusion. However, FusionGAN tends to lose a significant amount of environmental information from visible images. To address this limitation, Ma et al. subsequently proposed DDcGAN [4], utilizing two discriminators to analyze features from the two source images.

Despite the notable fusion results achieved by the aforementioned methods, there are still shortcomings in certain aspects. Most current methods fail to adequately consider the different features of infrared images and visible images, leading to fusion images visually biased towards one modality.

To address these issues, this paper proposes a new model based on a GAN network framework. Our model consists of two generators and two discriminators. Through adversarial training between the generators and discriminators, the generators ultimately produce fusion images with multiple image features. The main contributions of this paper include:

(1) We propose a new GAN-based image fusion method where two sets of networks are responsible for different feature extraction and fusion to minimize the loss of feature information from a particular image.

(2) The proposed discriminators apply a multi-scale discrimination algorithm to focus more on the local detailed features of images.

(3) The proposed method not only makes changes to the network architecture but also applies multiple discriminative losses in the loss function to adapt to our model structure and guide the fusion process.

The remaining structure of this paper is organized as follows. Section 2 presents our research methodology, including the model structure and loss functions. Section 3 introduces the experimental results and analysis of the proposed method. Finally, in Section 4, we provide our conclusions.

2. Proposed Method

This section provides a detailed introduction to the method we propose. Firstly, we will discuss the overall structure of the model, followed by a discussion on the network architectures of the generators and discriminators.

2.1. Overall Framework

Our fusion framework consists of two generators and two discriminators. After the adversarial training between the generators and discriminators, the generators can produce fused images carrying rich information. The overall model structure is illustrated in Fig. 1. Initially, the source images are concatenated and fed into the first generator. After multi-

scale feature extraction and fusion, an intermediate result is generated. This intermediate result, along with the infrared image from the source images, is input to the first discriminator. In this discriminator, a comparison of features between the intermediate result and the infrared image is performed to ensure that the intermediate result retains the features of the infrared image as much as possible. Subsequently, the intermediate result and the two source images are concatenated and fed into the second generator to generate the final fused result. The fused image and the visible image are then input to the second discriminator to ensure that the fused image possesses the distinct features of images from different modalities. During the testing phase, by inputting the source images into the trained generators, the fused image can be obtained directly. Our proposed method fully leverages the characteristics of GAN. Through the adversarial training between the discriminators and generators, the capabilities of each module are continuously enhanced, ensuring that the fused image maintains consistency with the source images in terms of overall style and structure, as well as local texture and detail features as much as possible.

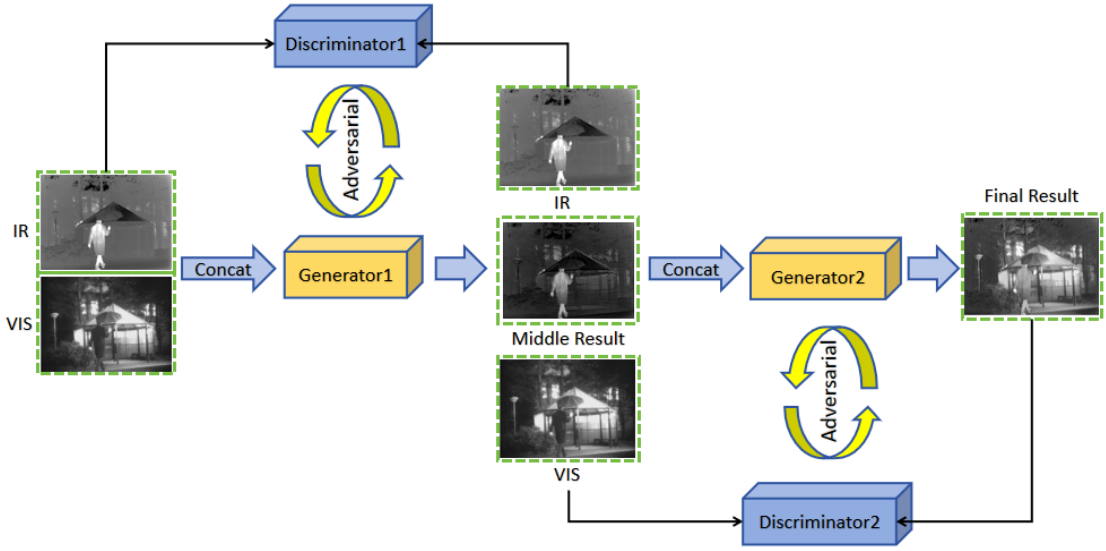

Fig. 1 Overall structure of the proposed model.

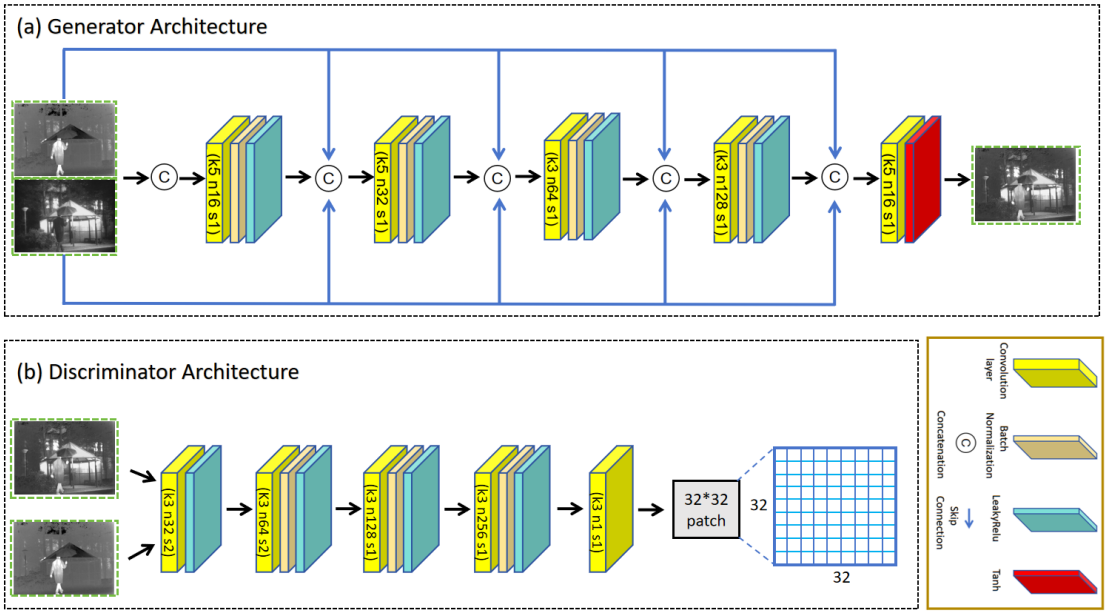

Fig. 2 Architecture of Generator and Discriminator

2.2. Architecture of Generator

The generator structure is illustrated in Fig. 2(a). The generator network is based on a convolutional neural network (CNN), where each layer consists of a convolutional layer, a BatchNorm layer, and a LeakyReLU layer. The first two layers use 5x5 convolutional kernels to expand the receptive field of the shallow network, while the subsequent layers use 3x3 convolutional kernels to reduce the network parameters. The input to the network includes the infrared image and the visible image, with image-related information concatenated along the channel dimension to provide rich initial data for further processing. Following the input layer, multiple convolutional layers operate successively. Each convolutional layer controls the extraction and transformation of image features based on annotated parameters such as "k5 n16 s1" (representing a 5x5 kernel size, 16 output channels, and a stride of 1). As the network progresses through the convolutional layers, the number of output channels gradually increases from 16 to 32, 64, and up to 128, allowing the model to learn increasingly complex and abstract image features. After each convolutional layer, batch normalization layers are

included to normalize the data, enhancing the stability of model training and accelerating convergence. LeakyReLU activation functions are incorporated in each layer to introduce non-linearity, enabling the model to handle more complex patterns. Importantly, skip connections from the original images are added at each convolutional layer. This is done to reduce information loss in the feature maps during the convolution process, ensuring that each layer's feature map retains the original information of the source images as much as possible. These skip connections serve to transmit feature information from earlier layers to subsequent layers, preventing gradient vanishing issues and allowing the model to comprehensively utilize features from different levels, further optimizing the feature fusion effect. After a series of processing steps, the data is finally mapped to an appropriate range through a Tanh activation function layer to generate the fused image.

2.3. Discriminator Architecture

The discriminator we designed is based on the Markovian discriminator. In contrast to conventional discriminators that output a scalar to represent the overall similarity between the generated image and the source image, the Markovian discriminator divides the generated image into several blocks and outputs a two-dimensional matrix. Each value in the matrix represents the similarity between each small block and the corresponding position in the source image, emphasizing local features more effectively. As shown in Fig. 2(b), our discriminator segments the image into a 32x32 matrix. Within each matrix, the discriminator evaluates the features of the source image and the fused image, further enhancing the quality of the fused image by focusing on local features.

2.4. Loss Function

The loss function of the generator consists of two parts, namely the adversarial loss L_{adv} and the content loss $L_{content}$. The generator loss is as follows:

$$L_G = L_{adv} + \lambda_1 L_{content} \tag{1}$$

Where λ_1 is the weight parameter that controls the balance between the two loss functions. We set $\lambda_1 = 3$. L_{adv} represents the adversarial game loss between the generator and the two discriminators, and is defined as follows:

$$L_{adv} = \frac{1}{2N}\sum_{n=1}^{N}(D(G(I_f^n)) - c)^2 \tag{2}$$

Where D denotes the discriminator, G denotes the generator, $\mathrm{n} \in \mathbb{N}_\mathrm{N}$, N denotes the number of fused images, and I_f^n denotes the n_{th} fused image. c denotes the label that the generator hopes the discriminator will assign to false data to be judged as true data, and we set $c = 0$. $L_{content}$ represents the content loss of the generator during the training process. Due to the differences in features between infrared images and visible images, we have also made corresponding treatments in the loss function. Specifically, $L_{content}$ is defined as follows:

$$L_{content} = \xi \parallel I_f - I_r \parallel_F^2 + (1-\xi) \parallel I_f - I_v \parallel_F^2 + \gamma \parallel \nabla I_f - \nabla I_r \parallel_F^2 \tag{3}$$

Where I_r and I_v represent the infrared image and the visible image respectively, $\|\cdot\|_F^2$ represents the Frobenius norm, ∇ represents the gradient operator, ξ and γ are positive parameters that control the trade-off between the terms. In this article, $\xi = 0.6,\ \gamma = 15$.

To improve the stability of the training process, we used the loss function of the Least Squares Generative Adversarial Network (LSGAN). The LSGAN loss function is defined as follows:

$$\min_D V_{LSGAN}(D) = \frac{1}{2}\mathbb{E}_{x\sim P_{data}(x)}[(D(x)-b)^2] + \frac{1}{2}\mathbb{E}_{z\sim P_z(z)}[(D(G(z))-a)^2] \tag{4}$$

$$\min_G V_{LSGAN}(G) = \frac{1}{2}\mathbb{E}_{z\sim P_z(z)}[(D(G(z))-c)^2] \tag{5}$$

In this paper, the discriminator loss function L_D is defined as follows:

$$L_D = \frac{1}{N}\sum_{n=1}^{N}[D_1(I_r^n)-b)^2 + D_2(I_v^n)-b)^2 + (D_{(1,2)}(I_f^n)-a)^2] \tag{6}$$

Where D_1 represents the first discriminator, D_2 represents the second discriminator. a and b represent the labels of false data and real data respectively, and $a = 0,\ b = 1$.

3. Experimental results and analysis

In this section, we present experimental details, including dataset, introduction to image quality metrics, and comparison experiments with seven current mainstream methods to demonstrate the effectiveness of our work.

3.1. Dataset and the Implementation Details

Dataset: The TNO dataset contains infrared images from various military, security, and other scenarios, along with corresponding registered visible light images. It can be used for researching and evaluating image fusion algorithms.

Implementation Details: Comparison Experiment and Fusion Metrics: In the comparison experiment, we selected 7 state-of-the-art fusion methods to compare with our method, including Laplacian pyramid (LP) [5], Wavelet [6], DenseFuse [7], FusionGAN [3], PMGI [8], SDNet [9], and U2Fusion [10]. We conducted subjective and objective evaluations of these 7 methods along with our approach. We utilized 9 evaluation metrics to comprehensively assess the quality of the fused images. Using these 9-quality metrics, we quantitatively compared our fusion method with the other 7 state-of-the-art fusion methods. The metrics include Entropy (EN) [11], Standard Deviation (SD) [12], Spatial Frequency (SF) [13], Peak Signal-to-Noise Ratio (PSNR) [14], Mutual Information (MI) [15], Structural Similarity Index Measure (SSIM) [16], Visual Information Fidelity (VIF) [17], Chen-Blum metric (Qcb) [18] and Mean Squared Error (MSE) [19].

3.2. Results on TNO Dataset

Qualitative analysis: Infrared images exhibit distinct infrared target features, primarily characterized by bright human outlines, while visible light images contain more complex background details and texture features. In these fused images, LP, Wavelet, DenseFuse, and U2Fusion failed to capture the contours and brightness features of the infrared targets effectively, and FusionGAN did not adequately retain the background information from the visible light images. This comparative result indicates that our method effectively preserves the infrared target features from the infrared images and the texture features from the visible light images. As shown in the two sets of images in Fig. 3, for the first set of images, the contours and colors of the trees marked in red and the buildings marked in yellow are clearer in our fused images compared to other methods, making them more perceptually appealing. In the second set of images, our fusion result effectively retains prominent features of the target individuals and clearly displays the texture information of the ground lawn marked in yellow. Additionally, for the tree contours

marked in blue, other methods produce white contour artifacts at the edges, while our method avoids such artifacts.

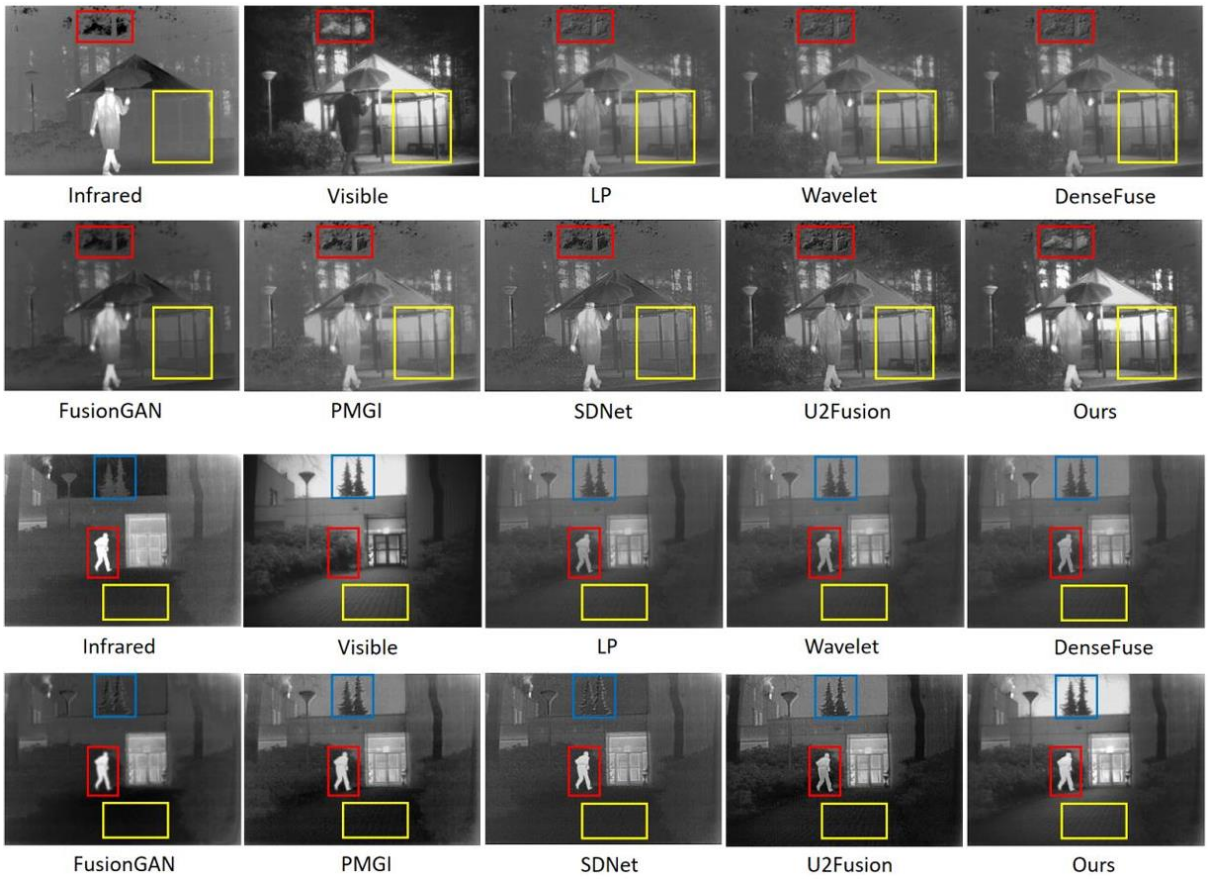

Fig. 3 Details of the comparison between our method and the 7 methods:

Table 1. The objective evaluation metric values for each algorithm on 41 pairs of images in the TNO dataset, with bold indicating the best results.

	EN	SD	SF	PSNR	MI	SSIM	VIF	Qcb	MSE
LP	6.1671	24.9134	6.3001	15.9979	1.2123	0.7607	0.2721	3.6106	10.1466
Wavelet	6.3176	24.9138	6.2862	15.9962	1.2163	0.7609	0.1980	3.6111	10.1504
DenseFuse	6.3558	24.8918	6.2532	15.9972	1.2136	0.7615	0.2726	3.6104	10.1487
FusionGAN	6.5710	30.8745	6.2451	11.2798	0.7078	0.6120	0.1589	3.6484	10.3533
PMGI	**7.0267**	38.2873	8.5572	12.4465	0.8399	0.6549	0.2146	3.7688	10.1889
SDNet	6.7069	33.8910	11.3965	11.6643	0.6290	0.6549	0.2143	3.7485	10.0414
U2Fusion	6.9630	36.6015	**11.5067**	13.7455	0.9559	0.6605	0.2476	3.7853	10.0954
Ours	6.9666	**41.2392**	8.9105	**16.6619**	**1.5006**	**0.7877**	**0.3125**	**3.7871**	**10.0229**

Quantitative analysis: We further conducted quantitative comparisons on 41 pairs of images from the TNO dataset to validate the effectiveness of our method. Table 1 displays the average values of various metrics for each method across the 41 pairs of images. From the results, it is evident that our method achieved the best results in terms of SD, PSNR, MI, SSIM, VIF, QCB, and MSE metrics. The high PSNR value indicates that the fused images generated by our method are less affected by noise and contain rich information. The highest SD value suggests that our fused images contain abundant detailed information. The SSIM evaluation results also indicate that our fused images maintain a high level of structural consistency with the source images. The VIF and QCB results suggest that our results align more closely with human visual perception. The optimal MI and MSE results indicate that our fusion results contain the most detailed information from the source images. In terms of the EN metric, our method ranks closely behind PMGI, indicating that the images carry a significant amount of information. Overall, our method achieved the best results in quantitative evaluations.

4. Conclusion

This article proposes a novel end-to-end fusion network based on GAN, consisting of two sets of generators and discriminators. Through adversarial training of each module, the capabilities of the components are continuously improved to encourage the generators to produce high-quality fusion images. The first generator is responsible for initial feature extraction, generating intermediate results with basic features from both input images. The first discriminator analyzes the intermediate results with the infrared image to enhance the infrared features in the intermediate results. The second generator performs further feature extraction and generates the final fusion image. The second discriminator analyzes the fusion image with the visible light image to incorporate different features from multiple source images into the final fusion image. Skip connections from the two source images are incorporated into the feature extraction process to reduce

feature information loss during training and preserve more original information as much as possible. In the discriminator, a multi-scale analysis approach is employed to further enhance the quality of the fusion image. Additionally, during the training process, the loss functions are modified to adapt to the model structure and further minimize the differences between the fusion image and the source images. Qualitative and quantitative experiments on public datasets demonstrate that our fusion results retain more environmental information and texture details. The experimental results validate the effectiveness of the proposed method.

References

[1] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," Information Fusion, vol. 76, pp. 323–336, 2021.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[3] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "Fusiongan: A generative adversarial network for infrared and visible image fusion," Information fusion, vol. 48, pp. 11–26, 2019.

[4] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," IEEE Transactions on Image Processing, vol. 29, pp. 4980–4995, 2020.

[5] W. Wang and F. Chang, "A multi-focus image fusion method based on laplacian pyramid.," J. Comput., vol. 6, no. 12, pp. 2559–2566, 2011.

[6] R. Chao, K. Zhang, and Y.-j. Li, "An image fusion algorithm using wavelet transform," ACTA ELECTONICA SINICA, vol. 32, no. 5, pp. 750, 2004.

[7] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2614– 2623, 2018.

[8] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in Proceedings of the AAAI conference on artificial intelligence, vol. 34, pp. 12797–12804, 2020.

[9] H. Zhang and J. Ma, "Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion," International Journal of Computer Vision, vol. 129, pp. 2761–2785, 2021.

[10] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, pp. 502–518, 2020.

[11] D. Y. Tsai, Y. Lee, and E. Matsuyama, "Information entropy measure for evaluation of image quality," Journal of digital imaging, vol. 21, pp. 338–347, 2008.

[12] G. Hur, S. W. Hong, S. Y. Kim, Y. H. Kim, Y. J. Hwang, W. R. Lee, and S. J. Cha, "Uniform image quality achieved by tube current modulation using sd of attenuation in coronary ct angiography," American Journal of Roentgenology, vol. 189, no. 1, pp. 188–196, 2007.

[13] R. N. Youngworth and B. D. Stone, "Simple estimates for the effects of midspatial-frequency surface errors on image quality," Applied optics, vol. 39, no. 13, pp. 2198–2209, 2000.

[14] A. Tanchenko, "Visual-psnr measure of image quality," Journal of Visual Communication and Image Representation, vol. 25, no. 5, pp. 874–878, 2014.

[15] E. Matsuyama, D.-Y. Tsai, and Y. Lee, "Mutual information-based evaluation of image quality with its preliminary application to assessment of medical imaging systems," Journal of Electronic Imaging, vol. 18, no. 3, pp. 033011–033011, 2009.

[16] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," IEEE transactions on Image Processing, vol. 20, no. 8, pp. 2378–2386, 2011.

[17] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," IEEE Transactions on Image Processing, vol. 15, no. 2, pp. 430–444, 2006.

[18] Y. Chen and R. S. Blum, "A new automated quality assessment algorithm for image fusion," Image and vision computing, vol. 27, no. 10, pp. 1421–1432, 2009.

[19] U. Sara, M. Akter, and M. S. Uddin, "Image quality assessment through fsim, ssim, mse and psnra comparative study," Journal of Computer and Communications, vol. 7, no. 3, pp. 8–18, 2019.