Cloud Computing and Data Science
https://ojs.wiserpub.com/index.php/CCDS/

UNIVERSAL WISER
PUBLISHER

Research Article

Conditional-GAN Loss-Function Enhancement Utilizing Frobenius-Norm and Spatial Attention Mechanism in Pix2Pix Model

Negar Nekoui Naeini, Omid Sharifi-Tehrani*

Department of Electrical and Communication Engineering, Safahan Institute of Higher Education, Isfahan, Iran
E-mail: omidsht@gmail.com

Received: 14 April 2025; **Revised:** 16 May 2025; **Accepted:** 22 May 2025

Abstract: Nowadays, the use of artificial neural networks has gained a prominent position in various engineering applications. Generative Adversarial Networks (GANs) have attracted significant attention due to their unique capability in content generation. These networks have become foundational models for a wide range of applications across diverse fields such as art and design, medical sciences, engineering, education, and more. In this paper, we propose two distinct approaches to enhance the output image quality of the Image-to-Image Translation with Conditional Adversarial Networks (Pix2Pix) model. The first approach involves modifying parameter values and incorporating Frobenius loss, while the second integrates a spatial attention mechanism into the generator component of the GAN. Simulation results indicate improvements in Intersection Over Union (IOU), Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR) parameters compared to some existing techniques.

***Keywords*:** artificial neural network, generative adversarial network, frobenius norm, spatial attention mechanism

1. Introduction

The architecture of a Generative Adversarial Network (GAN) comprises two neural networks: a generator that attempts to produce data resembling real-world data, and a discriminator that aims to distinguish between real data and data generated by the generator. This process is modeled as a minimax game, where the generator attempts to deceive the discriminator, while the discriminator strives to distinguish between real and generated samples. This process is modeled as a minimax game, where the generator attempts to deceive the discriminator, while the discriminator strives to distinguish between real and generated samples [1].

GANs have gained prominence due to their unique ability to generate data. For instance, in image processing, GANs can produce realistic images, enhance image quality, convert black and white images to color, and even create new images, often outperforming previous neural networks [2]. Therefore, extensive research has been conducted in this area.

GANs have found applications in various areas of artificial intelligence, including image generation, text generation, image-to-image translation, and more. They are capable of producing highly realistic images; however, due to the use of random vectors as inputs, they lack precise control over the generated images, as each output is based on the stochastic nature of the input. To address this limitation, researchers introduced conditional Generative Adversarial Networks (cGANs) in [3], enabling data generation based on specific conditions. By incorporating conditional

DOI: https://doi.org/10.37256/ccds.6220256998

information into both the generator and discriminator networks, cGANs facilitate controlled and directed data generation processes. In [4], cGANs are trained as conditional generative models suitable for image-to-image translation tasks. The generator architecture is based on U-Net [5], while PatchGAN is employed for the discriminator. This approach enables the model to examine both local details and overall image structure by adjusting the patch size. In [6], the Pix2Pix model is introduced as a Generative Adversarial Network for image-to-image translation, emphasizing the preservation of fine-grained details with high accuracy. It is particularly suitable for applications such as map-to-photo translation, image colorization, and other tasks where paired datasets are available. To overcome the challenges faced by models like Pix2Pix in handling more complex image translation tasks, the CycleGAN model was proposed in [7]. CycleGAN is capable of generating high-quality translated images while preserving the essential characteristics of the input images. It achieves this by utilizing a mapping $G : X \rightarrow Y$ and an inverse mapping $F : Y \rightarrow X$, along with a cycle consistency loss function to learn the mappings between two domains without the need for paired image datasets. While models like CycleGAN have demonstrated success in tasks involving color and texture transformations, they exhibit limitations when addressing tasks that require geometric changes. This shortcoming is largely attributed to the architecture of the generator, which is primarily designed for appearance modifications rather than structural transformations. Addressing this challenge presents a promising avenue for future research. An alternative approach proposed in [8] to address challenges in GANs, such as inaccuracies in preserving details, color bleeding, and boundary blurring, involves integrating Gabor filters with an enhanced Pix2Pix model. This method leverages the multi-directional and multi-scale properties of Gabor filters to preprocess images, effectively retaining detailed features and mitigating the loss of crucial information. Additionally, by refining the Pix2Pix model's loss function and incorporating a penalty term, the training process is stabilized, resulting in the generation of high-quality color images. Integrating Gabor filters into image processing enhances image clarity and reduces boundary blurring. Furthermore, replacing the standard Vanilla GAN loss function with the Least Squares GAN (LSGAN) loss, combined with the addition of the Gradient Penalty from Wasserstein GANs (WGAN)-GP, improves model stability and mitigates convergence issues. A comparative analysis of the Pix2Pix and CycleGAN models for image-to-image translation, as presented in [9], indicates that CycleGAN is generally more suitable for practical applications due to its ability to operate without paired datasets. However, Pix2Pix remains advantageous for specific tasks requiring high fidelity between input and output, such as facial reconstruction and converting sketches to realistic images. In [10], the authors enhanced the Pix2Pix model's performance by incorporating a novel loss function based on a Multi-Layer Perceptron. This addition aims to produce images that more closely resemble reality and improve image classification accuracy. The results demonstrate that integrating this Multi-Layer Perceptron (MLP)-based loss function effectively reduces residual noise in reconstructed images and enhances image quality compared to the standard Pix2Pix model. Despite significant advancements and importance in various domains, GANs still face challenges and limitations, including issues related to image quality, training instability, computational time, algorithm complexity, detail preservation, model accuracy across diverse datasets, initialization, and others, as discussed in [11] and [12]. Authors address common challenges in underwater imaging [13], such as reduced sharpness, color distortion, and low illumination. They propose a two-stage deep learning framework that simultaneously utilizes the original Red, Green, and Blue (RGB) image and a corresponding depth map to correct color distortions and enhance overall image quality. The method is designed not only to improve the visual appearance of underwater images but also to preserve structural details. For performance evaluation, standard image quality metrics such as PSNR and SSIM are employed, demonstrating the effectiveness of the proposed approach. This study highlights the potential of deep learning techniques in addressing complex environmental distortions and enhancing image quality in underwater scenarios.

Machine learning models have been utilized for automated and efficient diagnosis across various domains, including healthcare. In [14], a model was proposed for disease detection based on input data and real-time analysis through a mobile application and cloud computing. This research highlights the effectiveness of machine learning in real-time decision-making and precise data processing, which can also be extended to image quality enhancement. In this context, the Pix2Pix model-recognized as one of the practical models for image reconstruction-has been employed and examined in this study to improve image quality.

With the further development of these networks, conditional models such as Pix2Pix have been introduced for image-to-image translation, producing more controlled outputs using specific input information. Pix2Pix has gained attention due to its ability to preserve the overall structure of images and its applicability in various tasks, including converting sketches to realistic images, image reconstruction, and enhancing visual quality. Despite the successes of the Pix2Pix model, challenges persist, including inaccuracies in detail preservation, output noise, color overlap, and boundary blurring. This study aims to improve the output quality of the Pix2Pix model through two proposed approaches:

1-Integrating the Frobenius norm into the generator's loss function of the Pix2Pix model aims to preserve the overall image structure by minimizing the sum of squared differences between corresponding pixels.

2-Incorporating a spatial attention mechanism into the generator of the Pix2Pix model enhances its focus on significant image details, leading to improved performance in image-to-image translation tasks.

The aforementioned approaches lead to reduced noise in outputs, increased model accuracy, and improved clarity of generated images. In this paper, after reviewing the theoretical foundations, the proposed methods are explained, followed by the presentation of experimental results and a comparison of the performance of the enhanced models.

2. Theoretical foundations and proposed method

The primary objective of this study is to convert graphic images into high-quality photorealistic images while preserving the geometric accuracy, color consistency, and fine architectural details of building facades. To achieve this, the Pix2Pix model, based on the conditional Generative Adversarial Network (cGAN) architecture first introduced by Isola et al. in [4], has been employed. In the following sections, the theoretical foundations related to the model are first presented, followed by a detailed explanation of the proposed method.

2.1 *Fundamentals of Pix2Pix model*

The Pix2Pix model is a conditional Generative Adversarial Network (cGAN) which, unlike traditional GANs, utilizes an input image instead of random noise to guide the output generation process. The architecture of this model is described in the following section.

Generator: The generator is typically a U-Net architecture based on an encoder-decoder structure, which employs skip connections to preserve fine-grained information and generate high-quality images [15].

Discriminator: The discriminator employs the PatchGAN architecture, which evaluates small patches of the image locally and computes their average to determine whether the final image is real or fake [8].

Loss Function: The model employs a combination of two loss functions: the $L1$ loss, which measures the pixel-wise difference between the real and generated images, encouraging the generated output to closely resemble the ground truth; and the conditional adversarial loss, which introduces a competition between the generator and the discriminator. This adversarial setup drives the generator to produce highly realistic images that can deceive the discriminator, while the discriminator simultaneously learns to distinguish between real and generated images. The overall objective function of the model is defined as Equation (1) [8], where G^* denotes the optimal generator, λ_1 is a regularization parameter that balances the influence of the $L1$ loss, and L_{cGAN} represents the conditional adversarial loss that models the competition between the generator and the discriminator.

$$G^* = \arg \min_{G} \max_{D} L_{cGAN}(G,\ D) + \lambda_1 L_{L1}(G) \tag{1}$$

Figure 1 illustrates the architecture of the Pix2Pix model for translating graphic images into photorealistic images. In this model, the generator processes a graphic input image and produces an output image resembling a real photograph. The discriminator then receives both real and generated images as input and determines whether the input image is real or fake. In this adversarial process, the generator aims to produce increasingly realistic images, while the discriminator strives to improve its ability to distinguish fake images from real ones. Ultimately, this interaction leads to the generation of higher-quality images, better preservation of geometric details, and reduced noise in the model's output [16].

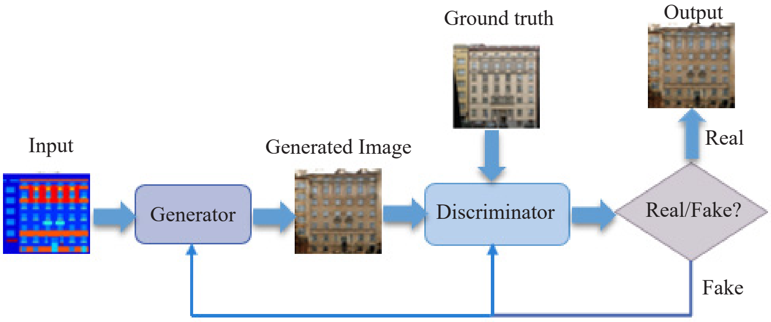

Figure 1. Pix2Pix model architecture

One of the main advantages of the Pix2Pix model is its stable training using paired images and its ability to reconstruct and translate images effectively. Given that the loss function has a significant impact on the model's performance, this study aims to enhance image reconstruction quality by modifying the loss function and integrating a spatial attention block into the generator, within the Pix2Pix architecture. The proposed loss function and the spatial attention block are described in the following sections.

2.2 *Frobenius norm*

The Frobenius norm is a commonly used metric in mathematics and image processing for measuring the difference between two matrices. It computes the square root of the sum of the squares of the elements of the matrix and is considered a generalization of the Euclidean norm for matrices. For a matrix A with dimensions $m \times n$ and elements a_{ij}, it is defined as follows:

$$F\left\|A\right\| = \sqrt[2]{\sum_{ij}\left|a_{ij}\right|^{2}} \tag{2}$$

2.3 *Attention mechanism*

The attention mechanism is an effective approach in Generative Adversarial Networks (GANs) that enhances the quality of generated images. This mechanism allows the network to learn weight maps for different parts of the input, enabling it to emphasize important areas while suppressing irrelevant ones [17].

2.3.1 *Spatial attention mechanism*

Spatial attention is a mechanism implemented using a soft attention layer that enables the network to simultaneously focus on all parts of the feature map. This is achieved by assigning different weights to various locations in the feature map, such that more important locations receive higher weights.

$$\mathrm{Ms}(F) = \sigma(f^{7\times 7}([\mathrm{AvgPool}(F);\ \mathrm{MaxPool}(F)])) = \sigma(f^{7\times 7}([F_{\mathrm{avg}}^{s};\ F_{\mathrm{max}}^{s}])) \tag{3}$$

According to Equation (3), this module first processes the input features F and applies the AvgPool and MaxPool operations independently. These two operations aggregate information along the channel dimension and generate two 2D maps, representing the average and maximum values of each channel at each spatial location, respectively. These two maps are then fused and processed through a 7 × 7 convolutional layer. Finally, a sigmoid function is applied to the output, normalizing the spatial attention map Ms(F) within the range of [0, 1]. Higher values in this map indicate more important regions. This final map is applied to the input features, enhancing relevant information while reducing unnecessary noise in the image. Additionally, the structure of this module is shown in Figure 2. In this model, the input features, previously processed by the channel attention module [18], are fed into the spatial attention block. Pooling operations are then applied to extract spatial information, and the output is processed through convolution and the sigmoid function. This mechanism models spatial relationships in the image, improving the network's performance in identifying important regions [19].

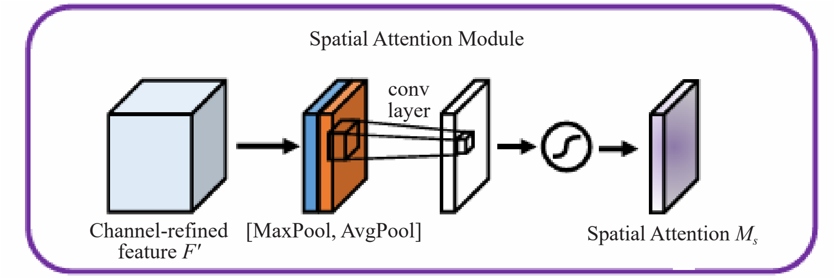

Figure 2. Spatial attention block [19]

2.4 *Proposed method*

Although the Pix2Pix model has made significant advancements in generating realistic images, as shown in Figure 3, the images produced by the model's original architecture still face challenges such as low resolution, lack of geometric detail preservation, and high noise levels. Since the generator's loss function plays a crucial role in enhancing image reconstruction quality, this study proposes the use of a composite loss function, incorporating the Frobenius norm and $L1$ loss, as defined in Equation (5), to address these challenges.

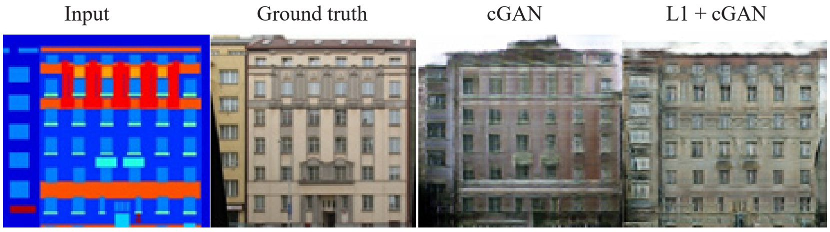

Figure 3. Output of the Pix2Pix

2.5 *Model architecture*

The implementation steps of the model aimed at improving the quality of the generated images include modifying the loss function by incorporating the Frobenius norm, tuning hyperparameters, and adding an attention mechanism to the generator section of the model.

2.5.1 *Loss function*

The inclusion of the frobenius norm in the loss function is due to its ability to measure the structural differences between the generated image and the real image. This norm, as a matrix-based metric, is defined by Equation (4) [20]:

$$L_{\text{frobenius}}(G) = \|y - G(x)\| \tag{4}$$

Where $L_{\text{frobenius}}(G)$ is the Frobenius loss function, y represents the real image, $G(x)$ is the generated image, and $\|\cdot\|$

denotes the Frobenius norm, which measures the difference between the real and generated images.

Unlike the $L1$ loss function, which focuses on pixel-wise differences, the frobenius norm is capable of providing more general and precise information about the structural differences between images. It evaluates the overall variations between two images and encourages the model to generate more accurate geometric structures with less distortion. As a result, the combination of these two loss functions can lead to improved image reconstruction quality and reduced noise. The final generator loss function, incorporating both of these metrics, is defined as follows:

$$G^{*} = \arg \min_{G} \max_{D} \ L_{\text{cGAN}}(G, D) + \lambda_1 L_{L1}(\text{G}) + \lambda_2 L_{\text{frobenius}}(G) \tag{5}$$

Where G^{*} is the optimal generator, $L_{\text{cGAN}}(G, D)$ is the conditional adversarial loss function, $L1(G)$ is the $L1$ loss function for reducing pixel-wise differences, $L_{\text{frobenius}}(G)$ is the frobenius loss function for improving the preservation of structural features, and λ_1 and λ_2 are the regularization coefficients for these functions.

By adjusting the value of λ_1, the influence of the $L1$ loss on reducing pixel-wise differences between the generated and real images can be controlled. Similarly, modifying the value of λ_2 can affect the preservation of the overall image structure and the reduction of geometric distortions.

In the proposed method, the effective use of this loss function has led to a reduction in geometric distortions and an improvement in the accuracy of detail reconstruction. As a result, the model is capable of generating images with higher resolution, reduced noise, and more precise geometric structures.

2.5.2 *Generator architecture of the proposed model*

The spatial attention mechanism helps the model focus more effectively on architectural details (such as windows and doors) during the process of transformation and reconstruction of building facades. This enables the model to process image features more efficiently, thereby enhancing the accuracy and quality of the reconstructed images.

In this study, to enhance the performance of the Pix2Pix model in image translation and transformation tasks, a spatial attention mechanism has been integrated into the generator component of the model. The generator consists of two main parts: an encoder with 8 downsampling layers and a decoder with 7 upsampling layers. As illustrated in Figure 4, a spatial attention module is added after each upsampling layer. This module leverages the spatial information present in the image to increase the model's focus on more important regions, thereby improving the quality and accuracy of the generated image. This modification enables the model to better preserve fine details in the output, resulting in a final image that more closely resembles the real appearance of the building façade.

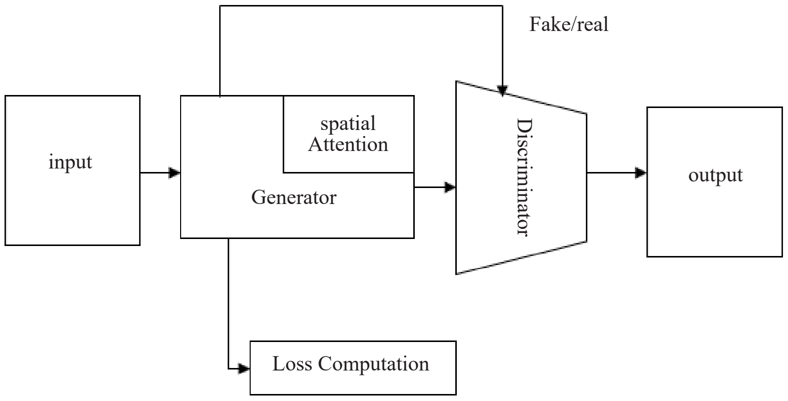

Figure 4. Enhanced Pix2Pix architecture

3. Simulation results

In this study, to improve the quality of the output images generated by the Pix2Pix model (as shown in Figure 3), a spatial attention mechanism and modifications to the loss functions were implemented by appropriately tuning the coefficients λ_1 and λ_2. To evaluate the performance of the model, the PSNR, SSIM, and IoU metrics were employed.

3.1 *Implementation details*

To implement and evaluate the Pix2Pix model, the free cloud-based platform Google Colab, which is based on Jupyter Notebook, was utilized. One of the key advantages of this environment is its free access to GPU resources, the ability to run Python code with pre-installed libraries, and the automatic saving of results to Google Drive. The implementation and training process was carried out using the Facades dataset, which includes 400 images for training, 106 images for testing, and 72 images for validation, all depicting building facades [21]. The dataset consists of RGB images in JPG format, each with a resolution of 256 × 256 pixels.

3.1.1 *Evaluation metrics*

In this study, various evaluation metrics were employed to assess the accuracy of the generated images:

PSNR: A numerical metric used to evaluate the quality of image reconstruction by measuring the pixel-wise difference between the generated image and the real image. A higher value of this metric indicates better quality and greater similarity between the generated image and the real image [13, 22].

SSIM: A metric used to evaluate the similarity between two images in terms of structural features, contrast, and brightness. A higher value of this metric indicates greater similarity between the generated image and the real image [13, 22].

IOU: Measures the overlap between the area predicted by the model and the real area. The value of this metric ranges from 0 to 1, with higher values of IoU indicating greater prediction accuracy [22].

3.1.2 *Parameter tuning*

In the proposed method, the initial settings of the model are based on the original Pix2Pix architecture[4]. Additionally, in this study, the coefficients λ_1 and λ_2 are adjusted according to Equation (5) to compare images at four different stages:

First Proposed Stage: The adjusted values of the coefficients are set as $\lambda_1 = 0$ and $\lambda_2 = 0.1$.

Second Proposed Stage: The values of the coefficients are changed to $\lambda_1 = 100$ and $\lambda_2 = 0.1$.

Third Proposed Stage: In this stage, spatial attention is separately added to the generator of the original Pix2Pix model, along with the $L1$ + cGAN loss function.

Fourth Proposed Stage: In this stage, spatial attention is separately added to the generator of the model from the second proposed stage. These constraints may lead to a reduction in the similarity between the generated images and the real data.

3.2 *Evaluation of the proposed model's output*

In this section, the impact of the four stages mentioned will be evaluated in improving challenges such as high noise, inadequate preservation of geometric details, reduced visual quality, and other limitations in the original Pix2Pix model structure.

The first stage of the proposed model: In this stage, the coefficients in Equation (5) are set as $\lambda_1 = 0$ and $\lambda_2 = 0.1$. Under these conditions, the $L1$ loss is removed, and the Frobenius norm has the most significant impact on the training process.

Qualitative comparison of the images in Figure 5 shows that the cGAN model generates an image with low resolution and noticeable noise, with some architectural details not properly reconstructed. By adding the $L1$ loss function, the image quality improves, edges become clearer, and noise is reduced; however, some details like windows and wall textures still show distortion. The addition of the Frobenius norm with the settings $\lambda_1 = 0$ and $\lambda_2 = 0.1$ results in improved resolution, reduced noise, and better reconstruction of architectural structures and details, such as windows. These results demonstrate that combining the Frobenius norm with the model's loss function has a positive effect on enhancing image quality and reducing distortion.

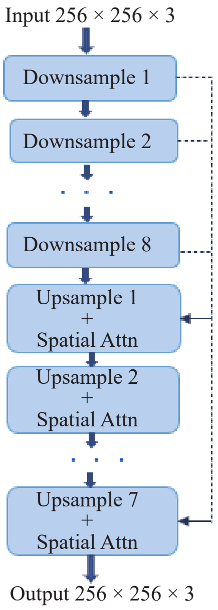
Input 256 × 256 × 3
Downsample 1
Downsample 2
Downsample 8
Upsample 1 + Spatial Attn
Upsample 2 + Spatial Attn
Upsample 7 + Spatial Attn
Output 256 × 256 × 3

Figure 5. Generator architecture of the proposed method

Table 1. Quantitative evaluation results of the output of the first stage of the proposed model with settings $\lambda_2 = 0.1$ and $\lambda_1 = 0$

Model	IOU	SSIM	PSNR
cGAN	0.57	0.87	48.09
($L1$ + cGAN) Pix2Pix	0.46	0.88	48.31
First stage of the proposed model	0.54	0.89	48.63

By comparing the values in Table 1, it is observed that the first proposed model shows significant improvement in some metrics compared to the cGAN and ($L1$ + cGAN) Pix2Pix models. Specifically, the PSNR and SSIM values in the first proposed model have increased, indicating overall image quality improvement and noise reduction. While the IOU value shows slight improvement over the ($L1$ + cGAN) Pix2Pix model, it has decreased compared to the cGAN model. This suggests that while the first proposed model performs better in preserving geometric details and reducing noise, its accuracy in reconstructing key image areas is lower than that of the cGAN model. Overall, the proposed model has succeeded in improving visual quality and reducing noise, but challenges remain in accurately reconstructing key areas of the image.

The second stage of the proposed model: In this stage, to improve reconstruction accuracy and better preserve image details, the coefficients are set as $\lambda_1 = 100$ and $\lambda_2 = 0.1$ in Equation (5), so that the $L1$ loss has the greatest impact on the training process. The $L1$ loss helps improve reconstruction accuracy by minimizing the pixel-wise difference between the generated image and the real image. Combining it with the Frobenius norm not only reduces overall discrepancies but also preserves finer details.

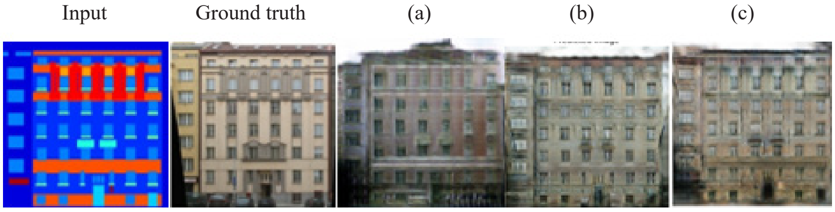

Figure 6. (a) Output of the baseline Pix2Pix model, (b) Output of the Pix2Pix model, (c) First stage of proposed model with the settings $\lambda_2 = 0.1$ and $\lambda_1 = 0$

The qualitative comparison of the images in Figure 6 shows that the image generated by the cGAN model is accompanied by high noise and low resolution, with many architectural details such as windows and wall textures not being properly reconstructed. Adding the $L1$ loss improves the image resolution and reduces noise, but some details still suffer from distortion. Finally, setting the coefficients as $\lambda_1 = 100$ and $\lambda_2 = 0.1$ results in the generation of an image with more accurate reconstruction of architectural details, further noise reduction, and better resolution. These results indicate that adding the Frobenius norm to the Pix2Pix model's loss function enables more precise reconstruction of building structures and improves the visual quality of the output.

The quantitative comparison of the results in Table 2 shows that the second proposed model performs better than the cGAN and Pix2Pix models. The PSNR value in the second proposed model has increased compared to the cGAN and ($L1$ + cGAN) Pix2Pix models, indicating improved image quality and reduced noise. Additionally, the SSIM in this model shows significant improvement over the cGAN and ($L1$ + cGAN) Pix2Pix models, suggesting better preservation of the image structure. The increase in IoU compared to the ($L1$ + cGAN) Pix2Pix model indicates that the overall structure of the reconstructed image is closer to the real image. These results suggest that the proposed model exhibits significant improvements in image quality and structural similarity. These findings indicate that the combination of the $L1$ and Frobenius norm loss functions has played a key role in enhancing image reconstruction quality and reducing noise.

Table 2. Quantitative evaluation results of the output of the second stage of the proposed model with settings $\lambda_2 = 0.1$ and $\lambda_1 = 100$

Model	IOU	SSIM	PSNR
cGAN	0.57	0.87	48.09
($L1$ + cGAN) Pix2Pix	0.46	0.88	48.31
Second stage of the proposed model	0.56	0.92	49.67

The third stage of the proposed model: In this stage, to improve reconstruction accuracy and focus more on local details, the spatial attention mechanism has been added to the generator of the Pix2Pix model with the cGAN + $L1$ loss function. The goal of this addition is to enhance the model's ability to preserve structural information and reduce noise in the image reconstruction.

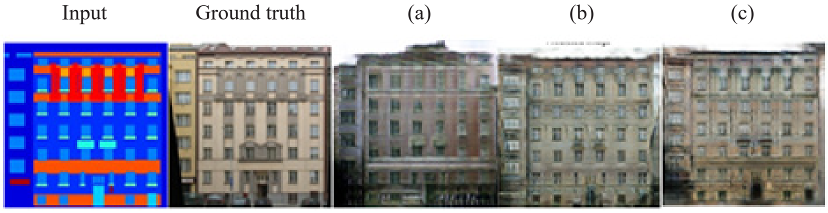

Figure 7. (a) Output of the baseline Pix2Pix model, (b) Output of the Pix2Pix model, (c) Second stage of proposed model with the settings $\lambda_2 = 0.1$ and $\lambda_1 = 100$

As shown in Figure 7, the third stage of the proposed model provides better output compared to the previous models, with clearer edges, more realistic building textures, and reduced image noise. As a result, the integration of the spatial attention mechanism into the generator of the original Pix2Pix model with the cGAN + $L1$ loss function can help preserve important image structures and increase the output's similarity to the real image.

Table 3. Quantitative evaluation of the third stage output of the proposed model

Model	IOU	SSIM	PSNR
cGAN	0.57	0.87	48.09
($L1$ + cGAN) Pix2Pix	0.46	0.88	48.31
Third stage of the proposed model	0.44	0.98	55.9

The quantitative results in Table 3 show that by adding the spatial attention mechanism to the generator, the model's performance has improved; both PSNR and SSIM values have increased, indicating an enhancement in the structural similarity of the reconstructed image to the real image. However, the IoU value has decreased compared to ($L1$ + cGAN) Pix2Pix and cGAN. This reduction may indicate that the proposed method has focused more on improving the overall image quality, leading to a decrease in the accuracy of the overlap between the original and reconstructed image regions. Overall, the third stage of the proposed method has shown better performance in improving image clarity and structural similarity, though it has resulted in a slight reduction in the IoU metric.

The fourth stage of the proposed model: In this stage, the spatial attention mechanism is added to the generator of the second stage proposed model. This model utilizes the cost function cGAN + $\lambda_1 L1$ + λ_2 Lfrobenius with λ_1 = 100 and λ_2 = 0.1. The aim of this change is to enhance the impact of the spatial attention mechanism on key image regions and improve the preservation of structural features such as edges and textures.

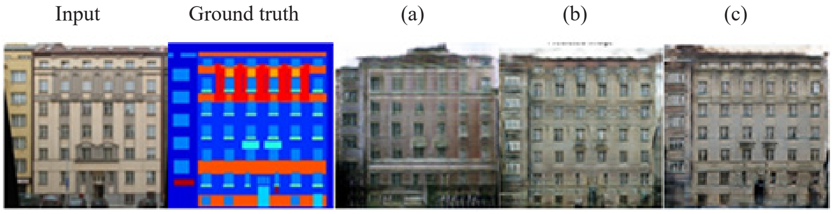

Figure 8. (a) Output of the baseline Pix2Pix model, (b) Output of the Pix2Pix model, (c) Third stage of proposed model

As shown in Figure 8, the output image from the fourth proposed stage demonstrates noticeable improvements in preserving the overall structures and architectural details compared to the cGAN and Pix2Pix models. The changes introduced in this stage result in clearer regions of key features, such as windows and building edges. While previous models experienced distortions and loss of certain details, the proposed model provides a more organized structure and is closer to the real image. This indicates that the addition of the spatial attention mechanism enhances the model's focus on important parts of the image, thus improving the overall quality of the generated output.

The quantitative results in Table 4 indicate that the proposed model in the fourth stage demonstrates a significant improvement in PSNR and SSIM metrics compared to previous models. The PSNR value has increased compared to the PSNR values in cGAN and Pix2Pix, which signifies a reduction in noise and an enhancement in image reconstruction quality. Additionally, the SSIM value, which measures structural similarity, has improved from 0.87 and 0.88 in previous models to 0.98, indicating better preservation of structural details and image texture. Although the IoU value in the fourth stage has reached 0.50, showing a decrease compared to the IoU in cGAN, it still outperforms the Pix2Pix model, suggesting that the proposed model, while emphasizing structural feature preservation, still has a strong capability in reconstructing key parts of the image. Overall, these results confirm that the addition of the spatial attention mechanism and the new loss function has significantly enhanced the quality of the generated images (Figure 9).

Table 4. Quantitative evaluation of the fourth stage output of the proposed model

Model	IOU	SSIM	PSNR
cGAN	0.57	0.87	48.09
($L1$ + cGAN) Pix2Pix	0.46	0.88	48.31
Fourth stage of the proposed model	0.50	0.98	56.11

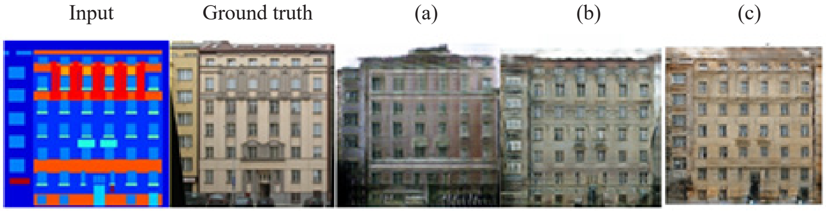

Figure 9. (a) Output of the baseline Pix2Pix model, (b) Output of the Pix2Pix model, (c) Fourth stage of proposed model

4. Conclusion

In this study, the Pix2Pix model, based on cGAN, was used for converting graphic images into real photos of building facades. By improving the Pix2Pix model through the addition of a new loss function and a spatial attention mechanism, the quality of the generated images was enhanced. The results showed that these modifications led to a reduction in noise, increased image clarity, and better preservation of details such as edges and windows. Moreover, evaluations using metrics such as PSNR and SSIM demonstrated that the proposed model generates more realistic images compared to previous models. Therefore, the use of an optimized loss function combination and the attention mechanism can be considered an effective approach to improving the performance of GAN-based models for converting graphic images into real photos in various applications.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. *Advances in Neural Information Processing Systems*. 2014; 27: 1-9.

[2] Dash A, Ye J, Wang G. A review of generative adversarial networks (GANs) and its applications in a wide variety of disciplines: From medical to remote sensing. *IEEE Access*. 2023; 12: 18330-18357.

[3] Mirza M, Osindero S. Conditional generative adversarial nets. *arXiv:1411.1784*. 2014. Available from: https://doi.org/10.48550/arXiv.1411.1784.

[4] Isola P, Zhu JY, Zhou T, Efros AA, Shechtman E, Wang O, et al. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

p.1125-1134.

[5] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015*. Cham: Springer; 2015. p.234-241.

[6] Henry J, Natalie T, Madsen D. Pix2pix GAN for image-to-image translation. *ResearchGate*. 2021; 1-5. Available from: https://doi.org/10.13140/RG.2.2.32286.66887.

[7] Zhu JY, Park T, Isola P, Efros AA, Zhang R, Shechtman E, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the Ieee International Conference on Computer Vision*. 2017. p.2223-2232.

[8] Chen H, Guan M, Li H. ArCycleGAN: Improved CycleGAN for style transferring of fruit images. *IEEE Access*. 2021; 9: 46776-46787.

[9] Chen L, Chen P, Lin Z. Artificial intelligence in education: A review. *IEEE Access*. 2020; 8: 75264-75278.

[10] Christovam LE, Shimabukuro MH, Galo MLBT, Silva J, Pereira R, Almeida C, et al. Pix2pix conditional generative adversarial network with MLP loss function for cloud removal in a cropland time series. *Remote Sensing*. 2021; 14(1): 144.

[11] Wang L, Chen W, Yang W, Zhang Y, Liu H, Li X, et al. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*. 2020; 8: 63514-63537.

[12] De Souza VLT, Marques BAD, Batagelo HC, Oliveira L, Santos M, Lima P, et al. A review on generative adversarial networks for image generation. *Computers & Graphics*. 2023; 114: 13-25.

[13] Kumar N, Manzar J, Shivani, Gupta S, Sharma R, Singh A, et al. Underwater image enhancement using deep learning. *Multimedia Tools and Applications*. 2023; 82(30): 46789-46809.

[14] Kumar N, Narayan Das N, Gupta D. Efficient automated disease diagnosis using machine learning models. *Journal of Healthcare Engineering*. 2021; 2021(1): 9983652.

[15] Jiang Y, Zhang Y, Luo C, Wang Z, Li H, Chen Q, et al. A generalized image quality improvement strategy of cone-beam CT using multiple spectral CT labels in Pix2pix GAN. *Physics in Medicine & Biology.* 2022; 67(11): 115003.

[16] Langr J, Bok V. *GANs in Action: Deep Learning with Generative Adversarial Networks*. USA: Manning; 2019.

[17] Tang H, Liu H, Xu D, Yan S, Li Y, Wang T, et al. AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Transactions on Neural Networks and Learning Systems*. 2021; 34(4): 1972-1987.

[18] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. p.7132-7141.

[19] Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Munich, Germany: Springer; 2018. p.3-19.

[20] Ghasemi A, Sharifi-Tehrani O. Single-tone continuous-wave interference detection and estimation using matrix features. In: *2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)*. Iran: IEEE; 2024. p.1-5.

[21] Efros AA. *Index of /pix2pix/datasets*. Available from: https://efrosgans.eecs.berkeley.edu/pix2pix/datasets/ [Accessed 18 December 2024].

[22] Pang Y, Lin J, Qin T, Liu Z, Li X, Wang L, et al. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*. 2021; 24: 3859-3881.

Appendix A

This appendix presents a collection of test images in Figure 10 to demonstrate the performance of the proposed model. These images have been selected to visually evaluate the reconstruction quality and to compare the model's results under various conditions. Additionally, Figure 11 illustrates the model outputs at different training steps, including steps 1, 10, 20, and 30.

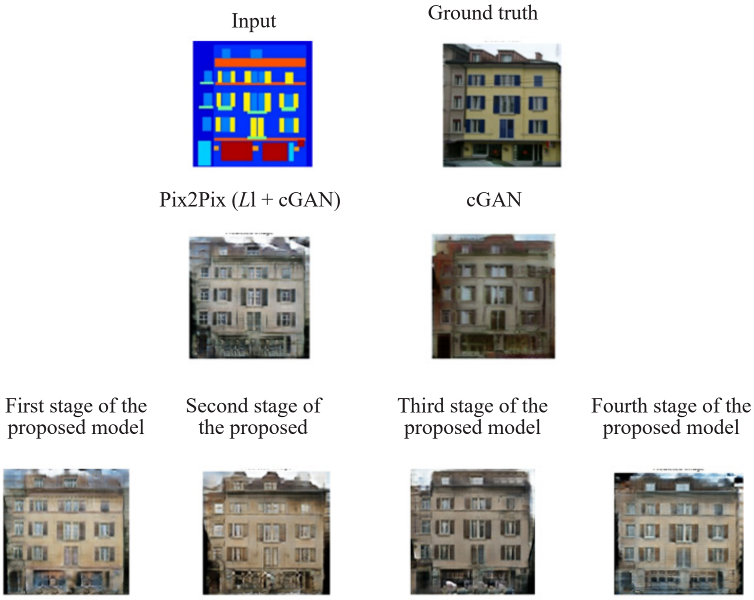
Input
Ground truth
Pix2Pix ($L1$ + cGAN)
cGAN
First stage of the proposed model
Second stage of the proposed
Third stage of the proposed model
Fourth stage of the proposed model

Figure 10. Qualitative comparison of test outputs from the four proposed methods

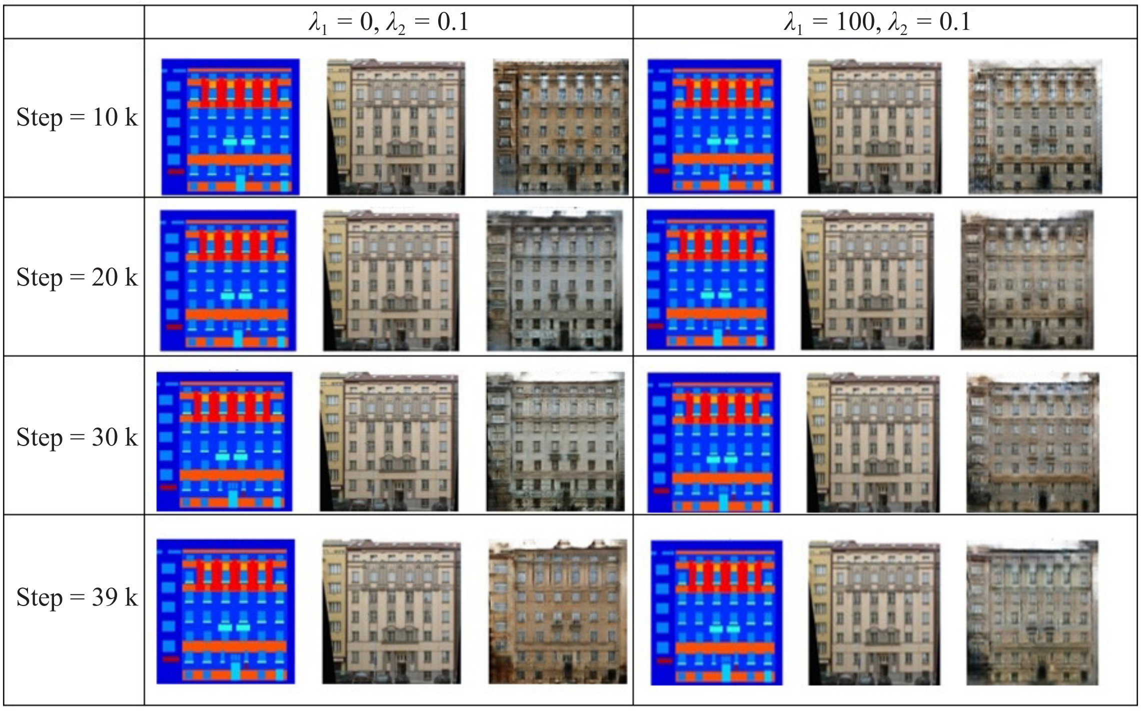
$\lambda_1 = 0, \lambda_2 = 0.1$
$\lambda_1 = 100, \lambda_2 = 0.1$
Step = 10 k
Step = 20 k
Step = 30 k
Step = 39 k

Figure 11. Model outputs at different training steps (10, 20, 30, and 39)

Appendix B

This appendix presents the results of evaluating the impact of the regularization coefficients λ_1 and λ_2 in the proposed loss function, summarized in the three tables (Table A-C). For each combination of λ_1 (columns) and λ_2 (rows), the average values of three evaluation metrics-PSNR, SSIM, and IoU-over the entire test dataset are reported. This evaluation was conducted to select the most effective coefficient values and provides an empirical basis for identifying the optimal combination.

Table A. IoU values for different combinations of λ_1 and λ_2

0		λ_2				
		0/1	10	40	80	
λ_1	0	0.57	0.54	0.58	0.55	-
	20	-	-	-	0.55	-
	100	0.46	0.56	0.46	-	-
	120	-	-	-	-	0.55

Table B. SSIM values for different combinations of λ_1 and λ_2

0		λ_2				
		0/1	10	40	80	
λ_1	0	0.87	0.89	0.98	0.87	-
	20	-	-	-	0.90	-
	100	0.88	0.92	0.87	-	-
	120	-	-	-	-	0.93

Table C. PSNR values for different combinations of λ_1 and λ_2

0		λ_2				
		0/1	10	40	80	
λ_1	0	48.09	48.63	55.46	48.2	-
	20	-	-	-	49.3	-
	100	48.31	49.67	48.2	-	-
	120	-	-	-	-	