Journal of Image Processing Theory and Applications (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2025.080105
ISSN 2560-6239 Vol. 8 Num. 1

Based on Improved U2-Net for Visible to Infrared Image Translation

Yinxiang Fan*

Yunnan Normal University, Kunming, China
2224100001@ynnu.edu.cn
**Corresponding author*

Keywords: Image Translation, U2-Net, Criss-Cross Attention Module, Frequency Domain Constraint, Infrared Image Generation

Abstract: With the rapid development of artificial intelligence technology, significant progress has been made in the field of image generation and translation. However, traditional Generative Adversarial Networks[1] (GAN) face issues such as training instability, mode collapse, and artifacts in image translation tasks. This paper proposes an improved method based on U2-Net[2] for visible to infrared image translation. By introducing the Criss-Cross Attention (CCA) [3]module into the deep layers of U2-Net, the model's ability to capture global information is enhanced, and a frequency domain loss function is used to optimize the image quality. Experimental results show that the improved model outperforms the original U2-Net in multiple evaluation metrics.

1. Introduction

In modern society, image information plays a major role in how people acquire information, especially in fields such as transportation and healthcare, where cameras are ubiquitous. However, under harsh weather conditions (e.g., heavy fog, low light), traditional visible light cameras struggle to capture clear image information. In contrast, infrared cameras, based on infrared light with wavelengths ranging from 3000nm to 15000nm, can clearly capture objects in low-light conditions by detecting their thermal emissions.

With the advancement of artificial intelligence technology, generative models have been widely applied in image generation and translation tasks. How to obtain infrared images using generative adversarial networks (GANs) without adding specialized infrared capture equipment has achieved significant results in image translation tasks. Excellent models like Pix2Pix[4] and CycleGAN [5]are based on cGAN[6], which is an optimization of GAN, and are widely applied in image generation,.but GAN-based models face instability during training and are prone to mode collapse[7]. Additionally, these methods only focus on pixel-level differences and neglect frequency domain constraints.

2. Research Objectives

The research objectives of this paper are to improve the U2-Net model to enhance the quality of visible to infrared image translation. The specific goals are:

1) Introduce the Criss-Cross Attention (CCA) module into the deep layers of U2-Net to enhance the model's ability to capture global information.

2) Improve the loss function by introducing frequency domain constraints to optimize the quality of the generated images.

3) Validate the effectiveness of the improved model through experiments and compare it with existing methods.

3. Related Work

3.1. Image Generation and Translation

Image generation and translation is an important research direction in the field of computer vision[8]. Generative Adversarial Networks (GANs) have been widely used in image translation tasks as the foundation of generative models. Conditional Generative Adversarial Networks (CGAN)-based models like Pix2Pix and CycleGAN can achieve image translation across domains. CycleGAN can even perform image translation across different styles without aligned image pairs, and its model's capability increased rapidly during its development. Notable advancements in image generation quality include models such as StyleGAN[9],, which has been thoroughly analyzed and improved to achieve superior image quality, and Glow[10], a generative flow model that utilizes invertible 1x1 convolutions to enhance image generation .

3.2. On of U2-Net in Image Segmentation

U2-Net is an improved model based on U-Net, widely used in image segmentation tasks. Compared with traditional U-Net[11], U2-Net has significant improvements in accuracy and detail capturing ability. The multi-level structure of U2-Net allows it to better capture global and local information of the image, which is especially useful in medical image segmentation and other fields. Moreover, U2-Net has also been used in generative models like CycleGAN as part of the image generator. In the broader field of image segmentation, several important methods have been proposed, including Fully Convolutional Networks (FCNs)[12], which significantly advanced the field of semantic segmentation, and SegNet[13], a deep convolutional encoder-decoder architecture that improved segmentation performance .

4. Research Method

4.1. Model Structure Improvement

First, we replace the original U-Net with U2-Net because U2-Net performs excellently in image segmentation tasks. U2-Net primarily relies on convolution operations, which effectively capture shallow information but have limitations in handling deeper image information. To enhance the model's ability to capture global information, we introduce the Criss-Cross Attention (CCA) module into the deep layers of U2-Net. The CCA module can focus on the relationships between pixels, enhancing the model's understanding of global features.

4.2. Loss Function Improvement

Traditional image generation tasks usually use L1 loss, which calculates the pixel difference between the generated image and the target image. However, L1 loss only considers pixel-level differences and ignores frequency domain information. To further optimize the quality of generated images, we introduce frequency domain constraints in the loss function by calculating the difference between the generated image and the target image in the frequency domain, enhancing the model's ability to capture global frequency information.

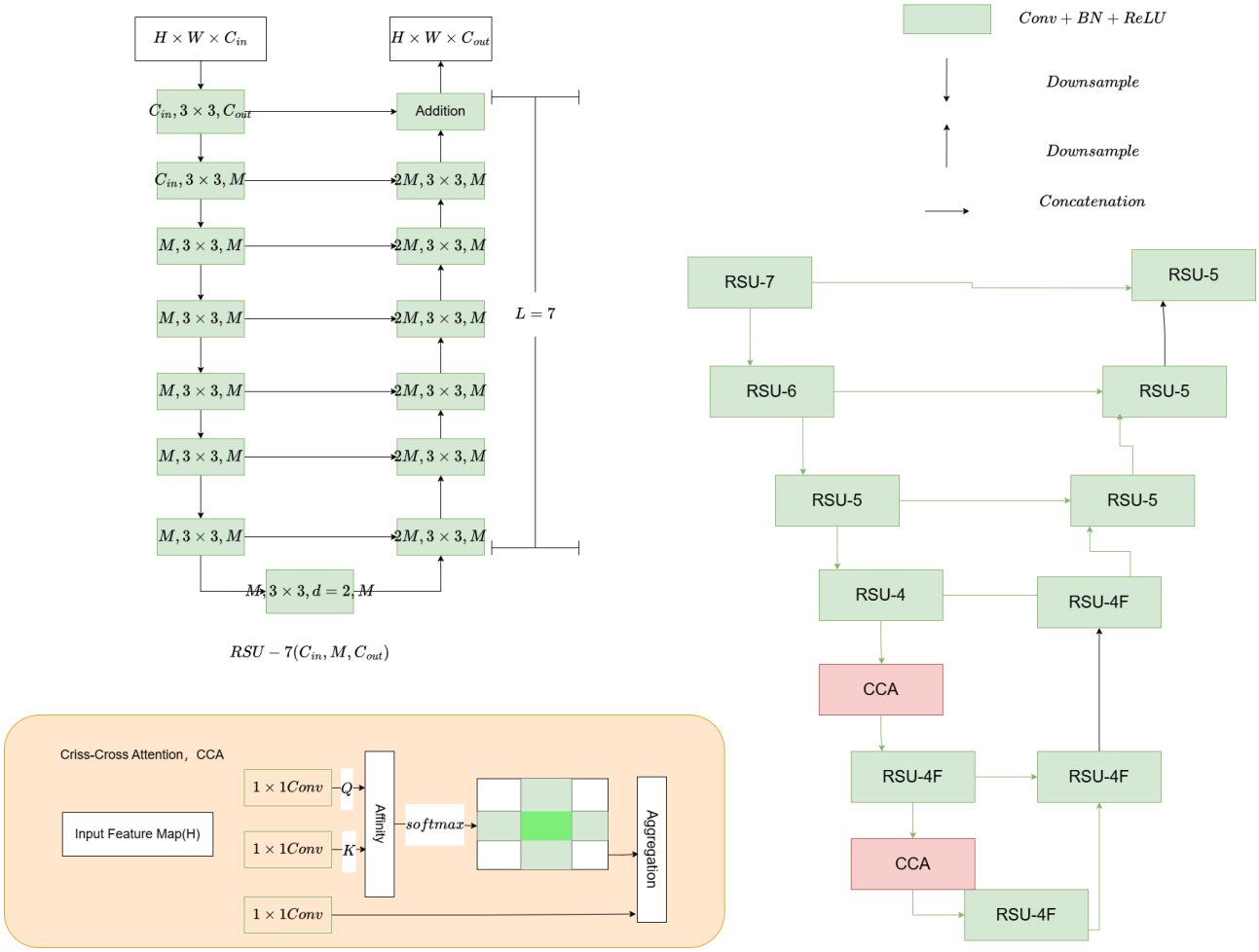

Figure 1: Architecture of U²-Net with Criss-Cross Attention (CCA)

As figure 1, the green parts represent the original components of U2Net. The mode lU2NET_CCAconsists of the two parts shown on the right side of the figure. The upper part,RES-L, adjusts L to obtainRSU-6,RSU-5, and RSU-4. These modules, together withCCA, form theU2NET_CCAmodel. The top-most part is RSU-7, which includes RSU-6, RSU-5, and RSU-4, varying only by the number of layers. RSU-4F differs from the others as it uses dilated convolutions to maintain the input and output sizes, serving only to constrain the features. Additionally, the CCA module is incorporated, which does not alter the input-output dimensions.

The entire experimental process starts with the input visible light image I_o. The image is first transformed into the frequency domain via FFT, resulting inf_o. The visible light imageI_o is passed through the model, generating the image I_o_, which corresponds to the frequency domain representation F_o. The original dataset image I_o corresponds to the real infrared image I_r, and its frequency spectrum is denoted asF_r. The improved loss function consists of the total output, which

is the L1 loss combined with the L1 loss after the frequency domain transformation.

5. Experimental Results

5.1. Dataset and Evaluation Metrics

We used the publicly available LLVIP[14] dataset for the experiments, and additional data under low-light conditions were added to improve the model's generalization ability. Evaluation metrics include:

- MAE (Mean Absolute Error): Measures the pixel-wise differences between the generated and target images.
- MSE (Mean Squared Error): Measures the pixel-wise differences between the generated and target images.
- PSNR (Peak Signal-to-Noise Ratio): Measures the quality of the generated image.
- SSIM (Structural Similarity Index): Measures the structural similarity between the generated and target images.
- FFT-L1: Frequency domain L1 loss, measures the difference between the generated and target images in the frequency domain.

5.2. Results Analysis

The experimental results show that the improved model outperforms the original U2-Net across multiple evaluation metrics. The specific results are shown in the table 1 below:

Table 1: Ablation Study on CCA and FFT Los

Method	MAE	MSE	PSNR	PSNR	FFT-L1	FFT-L1
U²-Net	9.832	5. 28	35.12	7.87	9. 25	34.87
+ CCA	9.730	5. 25	36.21	8.89	9.23	35.65
+ FFT Loss	8.028	4. 23	37.34	8.91	8.21	36.12

From the table, we can see that after introducing the CCA module and frequency domain constraints, the model's performance improves significantly, particularly in PSNR and SSIM.

6. Conclusion

In this paper, we propose an improved method for visible to infrared image translation, building upon the U2-Net architecture. The key innovation of our approach is the incorporation of the Criss-Cross Attention (CCA) module into the deeper layers of U2-Net, which significantly enhances the model's ability to capture long-range dependencies and global context information, thus improving the quality of the generated infrared images. In addition to this, we introduced frequency domain constraints into the loss function, leveraging the spatial frequency information to guide the model toward more accurate and realistic image translation. This combination of spatial and frequency domain optimization improves the model's performance, especially in terms of preserving fine details and mitigating artifacts during translation.

Experimental results demonstrate that the improved model outperforms the original U2-Net on multiple evaluation metrics, including Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), which are critical indicators of image quality. The incorporation of the CCA module and frequency domain constraints leads to a noticeable enhancement in both visual quality and quantitative performance, making the model more effective in various practical applications that require high-quality image translation.

6.1. Future Research Directions

While the current improvements have shown promising results, there are several exciting avenues for future research in this field:

Integrating with Image-to-Image Translation Models: One of the key future directions is to apply the improved U2-Net model as a generator within advanced image-to-image translation frameworks such as Pix2Pix or CycleGAN. By doing so, we aim to generate even more accurate and detail-rich infrared images by leveraging the model's enhanced capability in both spatial and frequency domains. This could lead to better results in tasks such as style transfer, domain adaptation, and image synthesis across different modalities.

Application in Denoising Networks and Diffusion Models: Another promising direction is applying the improved U2-Net model in denoising networks, particularly for use in diffusion models. These models are increasingly used in fields such as medical imaging, where noise reduction plays a crucial role in enhancing image clarity and diagnostic accuracy. The enhanced feature extraction capabilities of the improved U2-Net model can help preserve fine details while effectively removing noise from the images, leading to more accurate reconstructions.

Optimization of Image Translation Effects: Further optimization of image translation effects could be achieved by comparing the performance of the improved U2-Net model against various other image distributions, such as images with different contrast levels, illumination conditions, or spatial resolutions. By identifying the most effective distribution for specific tasks, future models can be trained to handle a wider variety of input conditions and thus provide better performance across more diverse real-world scenarios. Additionally, optimizing the model's ability to deal with challenging image conditions, such as low-light or high-noise environments, could significantly enhance its practical utility.

Real-Time Image Translation: As a future enhancement, the real-time application of visible-to-infrared image translation could be explored. While the current model achieves great performance, further optimization in terms of speed and computational efficiency is needed for real-time applications, such as surveillance, autonomous vehicles, or remote sensing. This could involve exploring model compression techniques, faster inference strategies, or using specialized hardware to accelerate the process.

Exploration of Multi-Modal Translation: Another potential direction is to explore multi-modal image translation, where the model is trained to translate not only from visible light to infrared but also across other modalities, such as thermal images, medical scans, or multi-spectral imagery. This would make the model highly versatile and applicable to a broader range of image translation tasks, making it a powerful tool in fields such as security, healthcare, and environmental monitoring.

In conclusion, while the improvements presented in this paper represent a significant step forward in the field of visible-to-infrared image translation, further research and exploration are required to enhance its applicability and robustness in various real-world scenarios.

References

[1] Labaca-Castro R .Generative Adversarial Nets[M].Springer Vieweg, Wiesbaden,2023.

[2] Qin X, Zhang Z, Huang C, et al.U2-Net: Going deeper with nested U-structure for salient object detection[J]. Pattern Recognition, 2020, 106:107404.

[3] Huang Z, Wang X, Wei Y ,et al.CCNet: Criss-Cross Attention for Semantic Segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, PP(99):1-1.

[4] Henry J, Natalie T, Madsen D. Pix2pix gan for image-to-image translation[J]. Research Gate Publication, 2021: 1-5.

[5] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.

[6] Mirza M, Osindero S .Conditional Generative Adversarial Nets[J].Computer Science, 2014:2672-2680.
[7] Wang Kunfeng, Gou Chao, Duan Yanjie, etc Research progress and prospects of Generative Adversarial Networks (GANs) [J]. Journal of Automation, 2017, 43 (3): 321-332.
[8] Chen Foji, Zhu Feng, Wu Qingxiao, etc A Review of Generative Adversarial Networks and Their Applications in Image Generation [J]. Journal of Computer Science, 2021, 44 (2): 347-369.
[9] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8110-8119.
[10] Kingma D P, Dhariwal P. Glow: Generative flow with invertible 1x1 convolutions[J]. Advances in neural information processing systems, 2018, 31.
[11] Ronneberger O, Fischer P, Brox T .U-Net: Convolutional Networks for Biomedical Image Segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention.Springer International Publishing, 2015, 234-241.
[12] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.
[13] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.
[14] Jia X, Zhu C, Li M, et al. LLVIP: A visible-infrared paired dataset for low-light vision[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 3496-3504.