WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 5, Number 1, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v5n1.02
Journal homepage: https://wepub.org/index.php/IJCSIT/index

General Multi-modal Image Fusion Transformer Network

Ao Dong *, Zhi Wang

College of Computer Science and Technology, Qingdao University, Qingdao, 266071, China
*Corresponding Author: dongao613@163.com

ABSTRACT

In the field of image fusion, images obtained from multiple different sensors are fused into one image that contains more complementary information and fewer redundant features to produce a single image with enhanced information. In order to be accurate and concise for different Fusion tasks, this paper proposes a General end-to-end Multi-modal image fusion Transformer Network (GMTN). A two-branch feature extraction module is designed, which integrates the self-attention mechanism of the improved convolutional neural network CNN and Transformer to extract the short-range features and remote dependencies of the image respectively, and take into account various information of the fused image in a more comprehensive way. A large number of experimental results show that the proposed method achieves the same or even better performance than the existing image fusion on a variety of multi-modal medical image data sets, and also achieves good results on infrared and visible light, multi-focus and multi-exposure image fusion tasks.

KEYWORDS

Image fusion; Transformer; End-to-end network

1. INTRODUCTION

As an increasingly mainstream research field in the field of computer vision and image processing, image fusion technology has been widely used in the real world. The purpose of image fusion is to improve image quality visually, extract significant features from two or more source images with complementary information and fuse them into a comprehensive image to make up for the possible defects of a single source image and provide more accurate and reliable information [1, 2]. A variety of image fusion algorithms have been proposed for different image fusion tasks, and these algorithms can be roughly divided into two categories, based on traditional methods and methods based on deep learning [3, 4]. Traditional fusion methods extract image features in spatial domain or transform domain, and realize image fusion according to specific fusion rules. However, manual feature extraction is heavily relied on, the algorithm's generalization ability is poor, and the fusion of complex source images may have low contrast effect, thus failing to achieve the best fusion effect, and other problems are particularly obvious in these traditional methods [5]. With the rapid development of deep learning neural networks, technologies such as convolutional neural networks (CNN), networks (GAN) and Transformer model have also been widely used in the field of image fusion. However, an image fusion network model that can extract short-range features and remote dependencies at the same time is applicable to a variety of image fusion tasks, and is trained in an end-to-end self-supervised way, so that the model parameters can jointly optimize the fusion results, such a model is still missing. Therefore, this paper proposes a General end-to-end Multi-modal image fusion Transformer Network (GMTN) method to solve the above problems. We have made a number of meaningful optimizations to the CNN and Transformer branches of the network. In the CNN branch, we add a specially modulated CNN convolution block to facilitate the feature extraction of different

types of images. In Transformer branch, we use trainable weights and the specially designed embedding layer to enable different types of images to get better remote feature extraction in Transformer. GMTN is a trainable end-to-end fusion network that takes source image pairs from different fusion tasks as inputs, uses a unified model framework, and trains different parameters to get the best fusion results for each task.

2. PROPOSED METHOD

2.1. Network Architecture

The multi-purpose image fusion method proposed in this paper is a fusion network for multiple image fusion tasks, and its network architecture is shown in Figure 1.

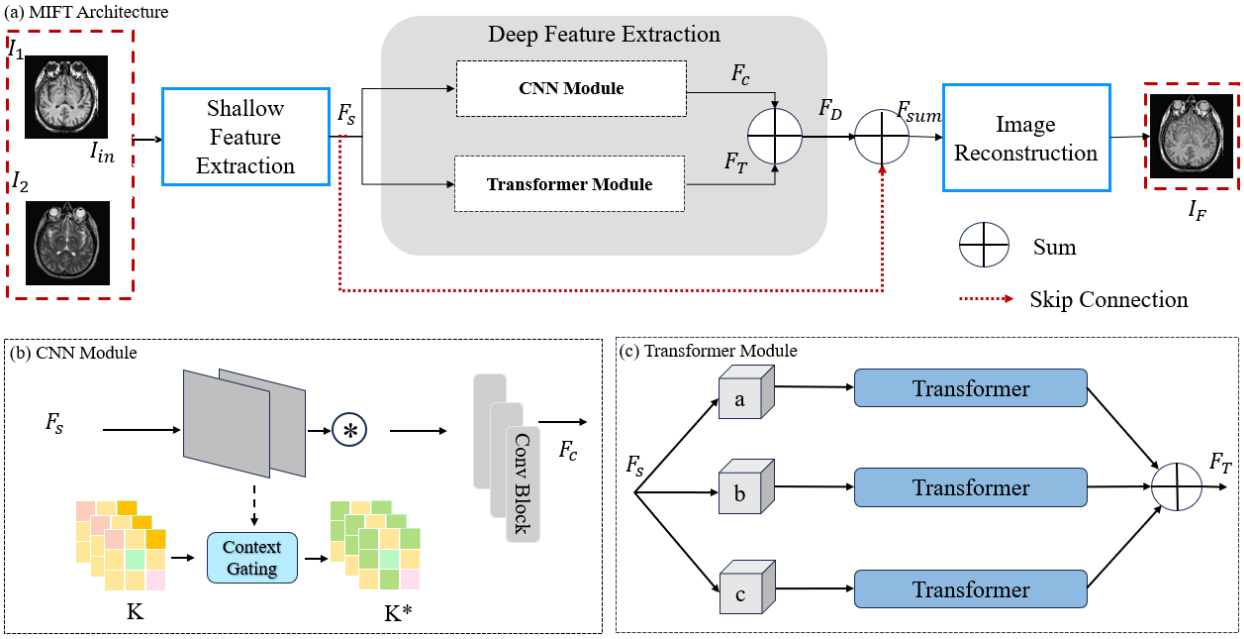

Figure 1. GMTN architecture

The multi-purpose image fusion network GMTN is composed of three parts: shallow feature input module, deep feature extraction module and image reconstruction module. The shallow feature input module includes image input and corresponding conversion operation. And a subsequent convolutional layer with a kernel size of 3 × 3. The last part of the model is the ReLU activation function and the pooling operation. For a given source input, we extract the input features of the input image from the convolution block and residual block of the module. The depth feature extraction module is then used to extract multi-scale depth features from the two images. The depth feature extraction module consists of a spatial CNN branch and a self-attention Transformer branch. The spatial CNN branch consists of a special convolution with 3 layers of Conv-Blocks to capture local features. Moreover, a pixel difference convolution block is added to extract edge features better. The Transformer branch consists of an axial attention-based Transformer block that captures remote dependencies and global context information. Finally, we spliced the multi-scale features extracted from the first two modules into the image reconstruction module to generate the final fusion result. In particular, for different fusion tasks, the image generation module will make different targeted processing.

2.2. Preliminary Feature Input Module

Taking the medical image fusion task as an example, the image to be fused is first input into the preliminary feature input module. Considering that medical images have both single-channel

grayscale images and multi-channel color images. Therefore, aiming at the fusion problem of gray image and color image, the widely used RGB-to-YUV color conversion method is used to solve the channel mismatch problem [6]. Specifically, the brightness information in the three-channel color image is extracted and converted into the Y component of the YUV color space. Then the hue and color saturation information are extracted and placed into U component and V component. Then, the extracted Y component is spliced on the channel dimension and input into the proposed MIFT multifunctional image fusion framework.

2.3. Deep Feature Extraction Module

Global context information plays a key role in interpreting visual scenes. Based on this consideration and motivation, the spatial CNN branch and the attention branch are introduced respectively. The two-branch structure has great advantages in extracting the global information of images. Different from previous extraction methods based on feature mapping, the improved CNN branch can directly modulate convolution kernel to represent features adaptively under the guidance of global context information. Some adjustments to Transformer also provide better performance.

2.3.1. CNN Module.

As shown in Figure 1 (b), we use Adaptive Convolution (AC) and Conv Block to extract short-range features from images. AC is a specially modulated adaptive convolution, which adaptively adjusts the convolution mode for different image types, thereby modulating the convolution kernel and extracting image information more conducive to subsequent fusion. Conv Block is a specially designed convolution layer, which consists of two CNN convolution layers with kernel 7 and step 2 and activation function ReLU. Through three extraction operations, the multi-scale extraction of depth features is gradually strengthened.

2.3.2. Transformer Module.

The self-attention mechanism relates the different labels of a single sequence to compute the representation of the same sequence. Considering the assumption that x and y are input and output features respectively, where H and W represent the length and width of the image, and the attention value y_o of a certain arbitrary pixel point x_o at any position p in the whole picture is calculated as follows:

$$qo = WQxo,\ ko = WKxo,\ vo = WVxo \tag{1}$$

$$yo = \sum_{p \in N} softmax(q_o^T k_p) v_p \tag{2}$$

Where q, k, and v are three linear transformations of the input x_o respectively. However, the computation of this primitive attention mechanism is very costly in terms of time and computing power, which has a great negative impact on the performance of our model. Therefore, this paper no longer applies the attention mechanism globally to the whole picture, but draws on the design method of the axis attention mechanism [7]. The attention mechanism is applied separately in height and width to greatly reduce computational complexity.

Specifically, in the calculation of axial attention, the self-attention calculation is performed first on the height axis of the feature graph, and then on the width axis to reduce the computational complexity. Therefore, for a given input x_o, the formula for calculating high self-attention is as follows:

$$y_{ij} = \sum_{h=1}^{H} softmax(q_{ij}^T k_{ih} + q_{ij}^T r_{ih}^q + k_{ij}^T r_{ih}^k)(v_{ih} + r_{ih}) \tag{3}$$

In the Transformer branch, we use the axial attention mechanism to model remote dependencies to learn global context characteristics. The specific design is shown in Figure 1 (c). The Transformer

branch is divided into 3 layers to help extract remote dependency information for multi-scale images. First, trainable parameters a, b, and c are used separately for each branch to obtain different parameters for different tasks, thus ensuring more accurate fusion quality. In each layer of Transformer, we no longer take the traditional Transformer architecture to calculate the attention mechanism, but take the width and high attention mechanism to calculate separately. This greatly reduces the computational complexity of the model and provides support for the expansion of the model.

3. EXPERIMENT

3.1. Dataset

As a general model for image fusion methods, this model was trained and tested on four types of data sets (medical images, infrared and visible images, multi-exposure images, and multi-focus images). However, note that due to the good generalization of our method, training on one data set and testing on other data sets can also achieve better fusion results in terms of visual effects. In this work, we selected 516 pairs of medical images from the Harvard Mainstream Medical Library website [8] for medical image fusion experiments, of which 36 pairs were used for validation tasks. For the infrared and visible image fusion task, we trained our model on 60,000 pairs of infrared and visible images in the KAIST dataset. For multi-focus and multi-exposure image fusion tasks, we selected 30,000 image pairs of various scenes in the large dataset MS-COCO [9] as the training set to unify the training network. We set the learning rate, epoch, and batch size to 0.001, 120, and 16, respectively.

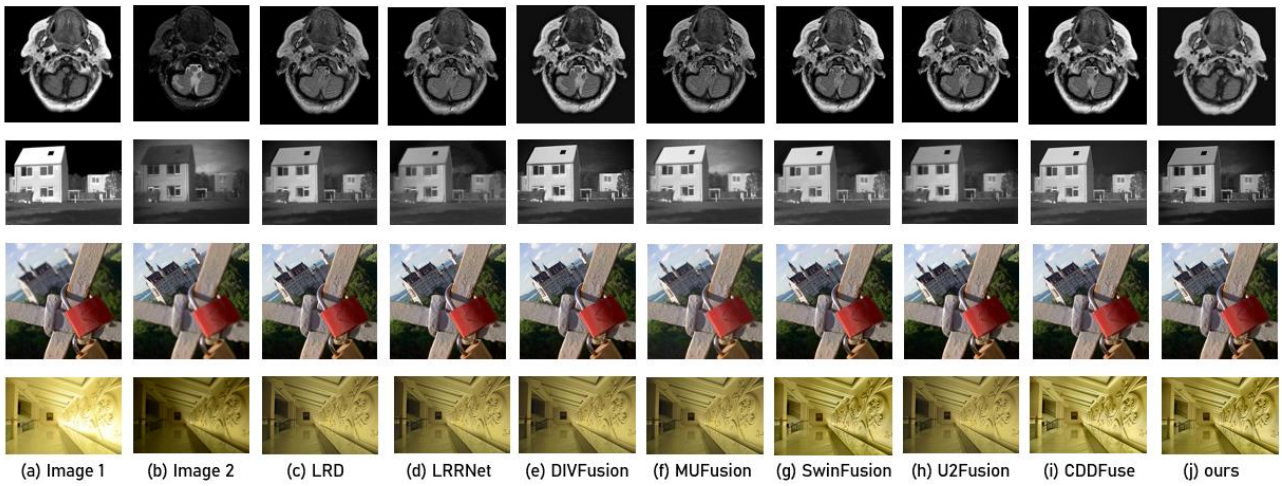

Figure 2. Fusion result

Our model is compared with a variety of traditional and state-of-the-art methods for image fusion. These include two latest single-task fusion methods DivFusion [10], LRRNet [11] and five general-purpose model fusion methods LRD [12], SwinFusion [13], MUFusion [14], U2Fusion [15], CDDFuse [16]. Our method was tested with these methods on the same dataset and obtained the best fusion performance in some metrics.

3.2. Qualitative and Quantitative Analysis

Our model focuses on the medical image fusion task. In the experiment specially trained for the medical image fusion task, our model obtains good results in multiple indicators, which has great practical significance. From Figure 2, we first performed a large number of experiments focusing on medical image fusion tasks and evaluated image fusion algorithms on medical image datasets. For two groups of different types of fusion experiments, texture structure and visual effect are relatively obvious in our method. It also achieves good results in other fusion tasks. We can infer that our

approach achieves optimal performance in MI, STD, SF and other metrics by preserving local and remote information. These results show that the image generated by GMTN method has higher fusion quality and clearer image resolution, and achieves better results.

4. SUMMARY

This paper proposes a General end-to-end Multi-modal image fusion Transformer Network (GMTN) for optimizing attention mechanism based on Transformer. This model uses unsupervised end-to-end training to solve the fusion task of multiple image types. The model proposed in this paper has the following two advantages: (1) Based on the current mainstream model network, the model is fully adapted to a variety of image fusion tasks, rather than limited to a single task. GMTN is a tradeoff network, which can effectively enhance complementary information to optimize the fusion process and obtain excellent fusion results for different fusion images. (2) Our network performs a variety of meaningful optimizations on the prevailing two-branch structure. Different types of images can be adapted to better remote feature extraction in Transformer. The details of other aspects of the model further help to improve the performance of the model.

However, although a large number of experiments have proved the effectiveness of this model method, there are still several aspects that need to be further improved. First of all, the latest Transformer has a variety of improvements. The performance of this model can be further improved by applying various optimized Transformer models. Secondly, how to further improve the generalization of the model is still an important topic. The image model of fusion registration is designed to fit more kinds of fusion tasks. Third, the two-branch feature extraction can be further optimized. It may be a simple and effective scheme to further improve the performance of the model to extract image features through deeper neural networks. In the future, we will further design a two-branch optimized fusion registration image model, and further solve the general problem of multi-exposure and multi-aggregation fusion task performance.

REFERENCES

[1] Zhang, Hao, et al. "Image fusion meets deep learning: A survey and perspective." Information Fusion 76 (2021): 323-336.

[2] Ardeshir, Goshtasby A., and S. Nikolov. "Image fusion: Advances in the state of the art." Information Fusion 8.2 (2007): 114-118.

[3] Li, Hui, **anbiao Qi, and Wuyuan **e. "Fast infrared and visible image fusion with structural decomposition." Knowledge-Based Systems 204 (2020): 106182.

[4] Zhang, Hao, et al. "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 07. 2020.

[5] Wu, Minghui, et al. "Infrared and visible image fusion via joint convolutional sparse representation." JOSA A 37.7 (2020): 1105-1115.

[6] Tang, Wei, et al. "A phase congruency-based green fluorescent protein and phase contrast image fusion method in nonsubsampled shearlet transform domain." Microscopy research and technique 83.10 (2020): 1225-1234.

[7] Ho, Jonathan, et al. "Axial attention in multidimensional transformers." arxiv preprint arxiv:1912.12180 (2019).

[8] Harvard medical website. http://www.med.harvard.edu/AANLIB/home.html

[9] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.

[10] Tang, Linfeng, et al. "DIVFusion: Darkness-free infrared and visible image fusion." Information Fusion 91 (2023): 477-493.

[11] Li, Hui, et al. "Lrrnet: A novel representation learning guided fusion network for infrared and visible images." IEEE transactions on pattern analysis and machine intelligence 45.9 (2023): 11040-11052.

[12] X. Li, X. Guo, P. Han, X. Wang, H. Li and T. Luo, "Laplacian Redecomposition for Multimodal Medical Image Fusion," in IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 9, pp. 6880-6890, Sept. 2020, doi: 10.1109/TIM.2020.2975405.

[13] Ma, Jiayi, et al. "SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer." IEEE/CAA Journal of Automatica Sinica 9.7 (2022): 1200-1217.

[14] Cheng, Chunyang, Tianyang Xu, and **ao-Jun Wu. "MUFusion: A general unsupervised image fusion network based on memory unit." Information Fusion 92 (2023): 80-92.

[15] Xu, Han, et al. "U2Fusion: A unified unsupervised image fusion network." IEEE Transactions on Pattern Analysis and Machine Intelligence 44.1 (2020): 502-518.

[16] Zhao, Zixiang, et al. "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.