Visible Light Polarization Image Fusion based on Dense Connection Generative Adversarial Network

Tao Wang1, Jianghua Cheng1*, Tong Liu1, Kangcheng Zhao1, Bang Cheng1, Zilong Liu2

1. College of Electronic Science, National University of Defense Technology, Changsha, Hunan, 410005, China
2. 32065 Troops, Shenyang, Liaoning, 110162, China

*Corresponding author’s e-mail: jianghua_cheng@nudt.edu.cn

Abstract. Polarized images have rich high-frequency information. The texture, contours and edges of objects in the image are very obvious, while the intensity images contain the main energy of the image and the background field is rich in information. Therefore, the polarization image is combined with Light intensity image fusion is of great significance. This article uses deep learning methods, based on the training method of generating adversarial networks, using densely connected generator networks, using SSIM loss and gradient loss functions, and experiments have shown that the ideal fusion effect can be achieved.

1. Introduction

The polarization characteristics of different materials are different, which makes it possible to distinguish objects with the same color but different materials by polarization image. However, compared with visible image, polarization image lacks low-frequency information, which is more representative of the essence of the image object. Image fusion is the fusion of two modal images that carry different information into an image that contains important information from two source images. During the process, the crucial information of the original image must not be lost, and redundant information must not be generated. In this paper, this method combines linear polarization image and light intensity image, which can make the edge texture of the image clear and improve the contrast of the image. Utilizing the characteristics of the degree of polarization, the image based on image fusion can also detect the target camouflaged in nature, extract the target information from the background more easily, and improve the recognition rate of object detection.

There are several traditional fusion methods for different images. Methods based on transform domain include wavelet transform [1], curvelet transform (CVT) [2] and multi-resolution singular value decomposition (MSVD) [3]. The spatial domain fusion method based on gradient information has a fusion method based on guided filtering [4]. In addition, there are methods based on sparse representation (SR) [5] and gradient transfer fusion (GTF) [6]. Traditional fusion methods usually include three key steps: image conversion, activity level measurement and fusion rule design. Most methods adopt the same transformation for different source images, so they are not necessarily suitable for fusing light intensity and polarization images. In addition, the activity level measurement and fusion rules in most methods are designed artificially and difficult to adapt to different scenarios. As the design of fusion rules becomes more complicated, the calculation load increases accordingly.
Currently, deep learning methods shine in the field of image fusion. These methods have been widely used in infrared and visible light fusion, multi-focus fusion, multi-exposure, multi-modal and remote sensing image fusion. Compared with traditional methods, deep learning methods do not require artificial design of fusion standards and fusion rules. It only needs to design the corresponding network structure and loss function, then the network can get better fusion results by learning and has high robustness. The PFnet designed by Zhang [7] of Central South University is used for polarization image fusion.

This paper proposes a dense connection generator model by the method of generating images based on GAN network. In addition to the GAN loss, the SSIM loss and gradient loss function are added in the generation model to constrain the generation of the image. The gradient information is assigned to SSIM as a weight to guide the generated image to move closer to the original image with more gradient information. This method refers to densenet dense connection module as a generator to improve the utilization rate of grid feature images in the network. Based on the training method of GAN network, high-quality images are generated through adversarial learning.

2. Fusion method

2.1. GAN network structure

Figure 1 shows the GAN network model. This network can enhance the generating ability of the generator by establishing a continuous game between the generator and the discriminator. Unlike the original GAN network, the generator changes to conditional input. During the training phase, the visible light source image and the polarization image are input to the generator network. The generator’s fusion result and original image of visible light are input to the discriminator simultaneously in the next phase. The major function of the discriminator is to judge the visible light source image as a true image and the generated fusion image as a false image. In this way, the counterfeiting ability of the generator can continuously improve. Meanwhile, the discriminator will continue to update its parameters along with the training to continuously improve its ability to distinguish the true image and false image. When the discriminator cannot distinguish the generated image and the source image, the training process is deemed to be over.

During the testing phase, the discriminator is not required. The fusion result is obtained by inputting the polarization and visible light source image into the generator. Figure 2 shows the test model:
2.2. Generator network structure

Figure 3 shows the generator model. This paper uses a densenet-like network structure. It has denser feature image connections than the resnet residual module and can retain the detailed features of the source image efficiently. Down-sampling processing is not used to reduce pixel loss. In the structure, feature sampling is performed on the two source images respectively in the front three layers. Particularly, the feature image of the next layer in each channel of the two sampling channels is obtained by cascading the upper layer's feature images of two channels and connecting the original images of the respective channels to ensure potential connectivity. This design can retain both the original image information and feature information. The lack of a down-sampling operation makes the size of the feature image of each layer as same as the source image. The input of the last four layers of denseblock is the cascade of all the previous feature images and two original images are added. Two original images are connected in each subsequent layer's feature image, and then successively passes through the 4 * 4 convolution kernel, BN layer and LeakyReLU activation layer. The LeReLU activation function has the advantages of accelerating convergence and solving the gradient disappearance. Compared with the relu activation function, it can effectively reduce the risk of neuronal invalidity. The final output layer's activation function is Tanh, which normalizes all output elements from -1 to 1.

2.3. Discriminator network structure
Figure 4 shows the discriminator part. The figure illustrates the difference between the discriminator network and the generator network. The reason for the difference is that the discriminator is essentially a classifier. It uses the convolution of the input image to obtain a feature image to determine whether the input is true or false. This paper draws on the idea of the VGG network to do two-classification. A BN layer will be added to accelerate the convergence after each convolution operation. It's similar to the generator. The fully connected layer and the pooling layer are not used. Since the fusion task does not require the characteristics of the full image at a certain point, and the pooling layer loses some pixel information. The convolutional layer is only used to change the dimensionality in the entire network, which makes the network more robust.

2.4. The design of the loss function

The core of GAN is that the performance of the generator and discriminator improve continuously in the adversarial training. The generator generates real data as much as possible and tries to make the discriminator judges the generated result as true, then a smaller discrimination loss is obtained. On the contrary, the discriminator needs to distinguish the generated data from the real data as much as possible, that is, the result of the generator is judged as false, and the real data of the source image is judged as true. Since unsupervised learning does not have a corresponding label result for training, the S0 image is used as the true control image label of the discriminator. Ideally, the generated data is judged as 0 and the S0 image is judged as 1. The adversarial loss is as follows:

\[
\min_G \max_D E_{x \sim P_{data}} \left[ \log D(x) \right] + E_{x \sim P_z} \left[ \log (1 - D(G(z))) \right] 
\]

Where \( x \) is the source image from the input, \( D(x) \) is the output of the discriminator, \( z \) is the input of the generator, and \( G(z) \) is the output of the generator. The generator tries to make the \( P_z \) distribution obeyed by the generated image result consistent with the \( P_{data} \) obeyed by the real image. Since the real data, that is, the discrimination result of the source image when solving the gradient does not affect the parameters of the generator, the adversarial loss of the generator is:

\[
V_G = \frac{1}{2N} \sum_{n=1}^{N} (D(I^n_s) - 1)^2 
\]

Where \( N \) is the number of batches, and \( D \) is the result of the discriminator.

In addition to the counter loss, the loss function of the generator also has content loss. The content loss function designed in this paper includes similarity loss and gradient loss. Similarity loss uses structural similarity SSIM [8], the formula is as follows:

\[
Loss_{ssim} = w_{Dolp} \left( 1 - SSIM_{Dolp} \right) + w_{S0} \left( 1 - SSIM_{S0} \right) 
\]

Where \( w \) is the weight. Here, the gradient operator is used to calculate the gradient information of the two source images, and the information is allocated to the two SSIMs in proportion. The formula is as follows:

\[
w_{Dolp} = \frac{\nabla I_{Dolp}}{\nabla I_{Dolp} + \nabla I_{S0}} 
\]

\[
w_{S0} = \frac{\nabla I_{S0}}{\nabla I_{Dolp} + \nabla I_{S0}} 
\]

Where \( \nabla I_{Dolp} \) is the gradient information image of the input Dolp original image, \( \nabla I_{S0} \) is the gradient information image of the input light intensity image.

The gradient loss calculates the L1 norm of the gradient image of the generated image and the two source images, the formula as follows:

\[
Loss_{grad} = \frac{1}{NW} \left\| \nabla I_{fused} - \nabla I_{Dolp} \right\|_1 + \frac{1}{NW} \left\| \nabla I_{fused} - \nabla I_{S0} \right\|_1 
\]
Where $N$ and $W$ are the length and width of the image respectively.

3. Training

3.1. Data set production

The training data used in the paper is a self-made data set. FLIR's polarization camera is chosen, which uses SONY IMX250 Polarized sensor. The calculation result of the original linear polarization image has a lower gray value, darker image brightness and lower contrast. Inspired by the method of the paper [9], this paper uses gray-scale stretching to increase the contrast of the image, making the contrast of gray-scale information more obvious. The effect of gray-scale stretching is as follows:

![Comparison of gray-scale stretched images](image)

Figure 5. Comparison of gray-scale stretched images; (a) Original Dolp image, (b) Dolp image after gray-scale stretch

Match the visible light and linear polarization diagrams one by one. The size of each image is 1024*1224. The network input used in this paper is 128*128. Each source image is randomly cropped to 128*128 size to expand the data set.

3.2. Training process:

The experiment adopts the Mini-Batch training strategy, and each batch has a training volume of 32. When the generator generates a batch, the discriminator is trained twice to make the discriminator perform better than the generator in the initial stage of training.

The training steps are as follows: 1. The generator fuses the image to generate a fusion image and inputs the fusion image and corresponding visible light to the discriminator. The discriminator judges and calculates the average value of the loss function $L_D$ to update the network parameters through the Adam optimizer; 2. After step 1 runs twice for the same set of data, the average value of $L_G$ is calculated once, and the Adam optimizer is used to update the network parameters once; 3. Start the next round of image training until the end of the training.
4. Comparison and analysis of experimental results

4.1. Visual comparison

Figure 6. Visual comparison chart

Figure 6 shows the four pairs of fusion results selected randomly. The front two rows are the fusion source images, and the following are the fusion results of each method. The visual comparison methods used in this paper are PFnet, Gradient Transfer Fusion (GTF), Low-pass Pyramid Ratio Fusion (RP) [10]. The style of the 4 fusion results of PFnet is more biased towards the DOLP image, and the contrast between light and dark is more obvious. It can be seen that the light and dark imbalance of the first fusion result is more obvious. The image generated by GTF is much smoother, but there will be some pixels in the image that are blurry, and the original information has been lost. The RP method has obvious distortion. Especially, there are obvious pixel black blocks in the third result.

4.2. Comparison of quantitative indicators

Figure 7 shows the numerical index comparison of fusion images generated by various methods. In addition to PFnet, GTF, and RP, there is also multi-resolution singular value decomposition (MSVD). 10 fusion results are randomly selected. The indexes of the image used are SSIM, CC, PSNR, EN. SSIM reflects the similarity between the generated image and the source image. This paper takes the average value of the SSIM of the fusion result and the two source images. CC reflects the relevance of the two images, which is similar to the effect of SSIM. When the value is 1, it means that the two images are completely the same. PSNR (Peak Signal-to-noise Ratio) reflects the ratio of image information to noise. The larger the value is, the smaller the noise ratio is. EN is information entropy. The larger the EN is, the higher the information content of the image has.
Figure 7. Comparison of quantitative indicators; (a) SSIM value, (b) CC value, (c) PSNR value, (d) EN value

Table 1 shows this paper's method is at the forefront of CC and PSNR indexes, while the values of SSIM and EN are slightly behind. It can prove that this method has certain advantages in the quality of generated images when compared to the method based on deep learning and three traditional methods. Compared with the PFnet, each has its own advantages in some indexes.

|      | GTF  | MSVD | RP   | PFnet | Ours |
|------|------|------|------|-------|------|
| SSIM | 0.388| 0.445| 0.418| 0.520 | 0.470|
| CC   | 0.270| 0.575| 0.509| 0.524 | 0.563|
| PSNR | 12.928| 14.112| 13.006| 13.497| 13.973|
| EN   | 6.943| 7.190| 7.551| 7.545 | 7.176|

PFnet uses a dual-way desblock to extract features from the DOLP image and the S0 image respectively. The feature extracted from two source images is more abundant. PFnet uses not only the SSIM loss function but also the L1 norm of the average of the two source images and the fusion result image, both of which can make the extracted feature image have a high reduction rate to the source image. In this dataset, the average image of the two source images may lose information and increase noise, the L1 loss function of gradient information is used instead. The signal-to-noise ratio has a significant improvement.

5. Conclusion
This paper is based on the training method of generative adversarial network. The dual-channel-input generator consists of 3 layers of feature sampling and 4 layers of desblock. There is no down-sampling process in the entire network. In the generator, the next layer of feature image of each channel will be added with the source image, which greatly retains the features of the original image and effectively utilizes each feature image. The generator assigns the gradient information of the source image as a
proportion to SSIM, and finally adds the L1 loss of the gradient image between the generated image and the source image as the content loss. Experiments prove this paper's method has obvious advantages over traditional methods in terms of image quality and visualization comparison. As for the deep learning method PFnet, each has its own advantages.

Acknowledgments
This article was supported by the Natural Science Foundation of Hunan Province, fund number: 2020JJ4670

References
[1] Lewis J J , RJ O’Callaghan, Nikolov S G , et al. Pixel- and region-based image fusion with complex wavelets[J]. Information Fusion, 2007, 8(2):119-130.
[2] F . Nencini, A. Garzelli, S. Baronti, and L. Alparone. Remote sensing image fusion using the curvelet transform[J]. Information Fusion, 2007, 8(2):143-156
[3] Naidu V . Image Fusion Technique using Multi-resolution Singular Value Decomposition[J]. Defence Science Journal, 2011, 61(5):479-484.
[4] Li S , Kang X , Hu J. Image Fusion With Guided Filtering[J]. IEEE Transactions on Image Processing, 2013, 22(7):2864-2875.
[5] Li S, Yang B, Hu J. Performance comparison of different multi-resolution transforms for imagefusion[J].InformationFusion,2011,12(2):74-84.
[6] J Ma and C Chen and C Li and J Huang. " Infrared and visible image fusion via gradient transfer and total variation minimization"[J]. Information Fusion, 2016, 31, 100-109.
[7] Zhang J C, Shao J B, Chen J L, Yang D G, Liang B G, and Liang R G, "PFNet: an unsupervised deep network for polarization image fusion,"[J] Opt. Lett. 2020,45, 1507-1510.
[8] Zhou W, Bovik A C , Sheikh H R , et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Trans Image Process, 2004, 13(4), 600-612.
[9] Zhang J C,Luo H B,Liang R G, Ashfaq Ahmed, Zhang X Y, Hui B, and Chang Z, "Sparse representation-based demosaicing method for microgrid polarimeter imagery," [J]Optics Letters, 2018, 43(14), 3265-3268.
[10] Toet A. Image fusion by a ratio of low-pass pyramid[J]. Pattern Recognition Letters, 1989,9(4):245-253.