OPTICAL AND SAR IMAGE FUSION BASED ON VISUAL SALIENCY FEATURES

Jiacheng Zhang [1], Xiaoyue Ren [1], Jinjin Li [1], Lei Wang [2], Yuanxin Ye [1 *]

[1] Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China - swjtu_zjc@163.com, rxy1203mm@163.com, swjtuljj@163.com, yeyuanxin@home.swjtu.edu.cn
[2] Third Engineering Surveying and Mapping Academy in Sichuan Province, Chengdu, China - qileigirl@163.com

KEY WORDS: Optical and SAR, Visual Saliency Features, Image Fusion, Main Structure, Detail Texture.

ABSTRACT:

With the expansion of optical and SAR image fusion application scenarios, it is necessary to integrate their information in land classification, feature recognition, and target tracking. Current methods focus excessively on integrating multimodal feature information to enhance the information richness of the fused images, while neglecting the highly corrupted visual perception of the fused results by modal differences and SAR speckle noise. To address this problem, in this paper we propose a novel optical and SAR image fusion framework named Visual Saliency Features Fusion (VSFF). We improved the decomposition algorithm of complementary feature to reduce most of the speckle noise in the initial features, and divide the image into main structure features and detail texture features. For the fusion of main structure features, we reconstruct a visual saliency features map that contains significant information from optical and SAR images, and input it together with the optical image into a total variation constraint model to compute the fusion result and achieve the optimal information transfer. Meanwhile, we construct a new feature descriptor based on Gabor wavelet, which separates meaningful detail texture features from residual noise and selectively preserves features that can improve the interpretability of fusion result. Further a fast IHS transform fusion is used to supplement the fused image with realistic color information. In a comparative analysis with five state-of-the-art fusion algorithms, VSFF achieved better results in qualitative and quantitative evaluations, and our fused images have a clear and appropriate visual perception.

1. INTRODUCTION

With the remote sensing image application requirement increasing and a single image offering limited information, it is necessary to integrate multimodal image data into one image to form a more abundant and meaningful fused image. Due to the fact that the multimodal image data contains redundant information, we need to extract salient and complementary feature from pairs of images to enhance the interpretability of fusion result. Image features of different types are usually interlaced in the spatial domain, and a single fusion model can result in misrepresented information and confusing visual perception. The ideal fusion process is to separate complementary feature in different scale spaces, and then establish a corresponding fusion model based on the characteristics of the features.

Recently, many researchers have focused on optical and Synthetic Aperture Radar (SAR) image fusion, which has already been used to offer distinctive information for all-weather land classification, target recognition and object detection. Optical sensors passively receive information about the reflection of solar illumination from earth objects, so it can provide rich spectral information and sharp detailed features that are consistent with the observation of the human visual system, but can be easily influenced by adverse weather and poor illumination. In contrast, SAR is an active microwave sensor that receives backscattered energy and can acquire information under almost all weather and environmental conditions, which can capture prominent reflective targets and salient structure features (Moreira et al., 2013). However, the coherent imaging mechanism of the SAR sensor generates speckle noise that severely corrupts the image, thus all fusion algorithms should try to reduce the noise as much as possible. On the other hand, due to the imaging mode of the optical sensor, two different structural objects may appear identical spectral response information which cannot be effectively distinguished in optical imagery, but can be clearly differentiated in SAR imagery. The respective superior information of optical and SAR images can complement each other to generate rich structural and spectral information of a region (Kulkarni and Rege, 2020). Therefore, in most fusion application scenarios, the complementary information of optical and SAR images is combined to achieve high-quality image interpretability and simultaneously reduce useless information and speckle noise to make the fusion result suitable for human visual perception.

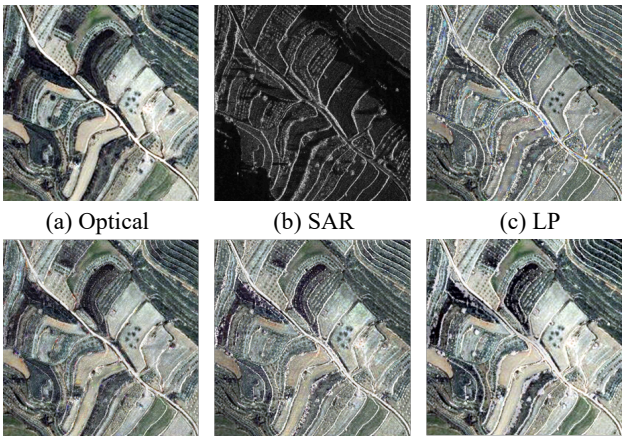
Generally, although there is spatial interlacing of image features, the detail and texture information can be extracted in the small-scale space while the main structure objects are distinguished in the large-scale space when the images are observed in different scale spaces. Thus, optical and SAR imagery can be decomposed into a set of complementary features such as main structure features (MSF) and detail texture features (DTF) by multi-scale filters. In last decades, much attention has been paid to the decomposition of structure and texture features of images. (Buades et al., 2010) proposed fast cartoon and texture image filtering algorithm that use nonlinear filter to achieve a simplified fast approximation of the total variation minimization problem. In particular, it uses only one parameter to achieve the control of the decomposition effect. However, when processing SAR images, it is difficult to separate a large amount of random speckle noise from other features, resulting in decomposed features that all contain noise information similar to their properties. The fusion results are highly corrupted by noise and suffer from spectral distortion and loss of local structure. In general algorithms, Gaussian filtering is used to implement multi-scale decomposition of images, which treats all pixel information of an image equally and cannot distinguish the

* Corresponding author: Yuanxin Ye (yeyuanxin@home.swjtu.edu.cn)

difference between features and noise. To separate the original features from the noise, the Wiener filtering is introduced into our feature decomposition process due to the adaptive ability to adjust the filter values.

After obtaining the required complementary features, further constructing the corresponding fusion model is the key issue. In the past decades, many methods for optical and SAR image fusion have been proposed, which are divided into four main categories: component substitution methods, multiscale decomposition methods, hybrid methods and model-based methods (Kulkarni and Rege, 2020). Due to the better feature extraction and data representation capabilities of deep learning. (Kong et al., 2021) proposes a Dense-UGAN method to extract spectral and texture information from source images. (Zhang et al., 2020) proposes a generalized CNN-based fusion network to achieve multimodal image fusion of optical and infrared and medical images. Although these methods have established better fusion models, the modal differences between optical and SAR can still lead to extremely poor visual perception of the fusion results, and current deep learning methods lack research on this problem. As an example, a large area of land appears as bright areas with rich texture in the optical image and as flat and smooth dark areas (lacking the backscattered signal) in the SAR image. However, the conventional "averaging" fusion rule simply superimposes different feature information from the two images, and the fusion result appears as dark gray blurred areas with a loss of real spectral and structural information. To solve this problem, we propose a fusion model based on visual saliency features (VSF). VSF is a relatively prominent area of an image that attracts human visual attention in a bottom-up way (Toet, 2011). It reflects the human visual behavior when freely observing images, where the observer is first attracted to areas of rich color or prominent brightness, followed by the large contours and fine edge structures, and finally by the regularly arranged textural information. During the fusion process, the VSF of optical and SAR imagery must be preserved and integrated to eliminate non-significant information (e.g., noise and redundant information) and enhance the visual perception effect and information interpretation of fused images.

In the fusion model of MSF, VSF are spectral information and fine edge structures from optical images and large contour structures and distinctive target regions from SAR images. Therefore, our requirements for the fusion results are to strike an appropriate balance in preserving the VSF from the optical and SAR images while having the similar pixel intensity distribution with the optical image, which keeps the best overall visual perception. The variational model can satisfy the above requirements. Among them, the gradient transfer fusion (GTF) (Ma et al., 2016) is a recent representative algorithm to formulate the constraint functions of pixel intensity distribution and pixel gradient variation, and then use total variational minimization to achieve information transfer fusion. However, the GTF algorithm assumes that the gradient variation information of the fusion result only come from a single image, which is inconsistent with the actual situation of optical and SAR fusion. Because the gradient variation information is an important representation of VSF, if the gradient variation constraint is considered only for SAR images, it will result in severe local structure loss and spectral distortion. To generate the required constrained images, we reconstructed a visual saliency feature map (VSFM) based on the priority of the contribution of the VSF of the optical and SAR images to the fusion results. For DTF fusion, VSF are meaningful fine targets and abundant texture information. However, a part of the speckle noise with high backscattered signal is easily retained in



(a) Optical (b) SAR (c) LP

(d) Hybrid-MSD (e) WLS (f) Our VSFF

Figure 1. Our fusion method (VSFF) is compared with some other state-of-the-art fusion methods.

DTF as texture information due to the relatively aggregated distribution. Therefore, the requirement of fusion processing is to separate the VSF from the noise-containing image and to obtain meaningful information in the DTF while eliminating the residual noise. On the one hand, it can be achieved by constructing a novel feature descriptor based on Gabor wavelet to further abstract the representation of the initial DTF. On the other hand, the VSF information from the two images is redundant and conflicting, and the VSF with more detail is selected to be preserved to the fusion result, thus bringing higher interpretability to the fused image.

In this paper we propose a novel fusion framework for optical and SAR images, named visual saliency features fusion (VSFF). By integrating and emphasizing VSF in the image, it eliminates noise interference and enhances the visual perception quality of the fusion result. Figure 1 shows the fusion results obtained by our method and some other state-of-the-art fusion methods, including Laplacian pyramid (LP) (Burt and Adelson, 1987), hybrid multi-scale decomposition (Hybrid-MSD) (Zhou et al., 2016) and weighted least square (WLS) (Ma et al., 2017). It can be seen that due to the modal differences between source optical and SAR images lead to severe spectral distortion in the classical multi-scale decomposition method LP. Even though the state-of-the-art method Hybrid-MSD overcomes the spectral distortion, it loses more detailed information. For the visual saliency map-based method WLS, it also suffers from the structural corruption of the image by SAR noise. In contrast, our method can obtain the best visual perception and clear detail presentation while reducing a large amount of speckle noise to make the fusion result look more refreshing.

The main contributions of this paper are as follows:

1. We propose a novel optical and SAR image fusion method that integrates information based on visual saliency features and eliminates the corruption of fusion results by SAR image speckle noise.

2. A new VSFM to optimize the total variation model of the fused image, which enhances the visual perception effect of the fusion result and emphasizes the important structural feature information in the images.

3.To construct a texture feature descriptor to further extract meaningful feature information and enhance the interpretability of the fusion results.

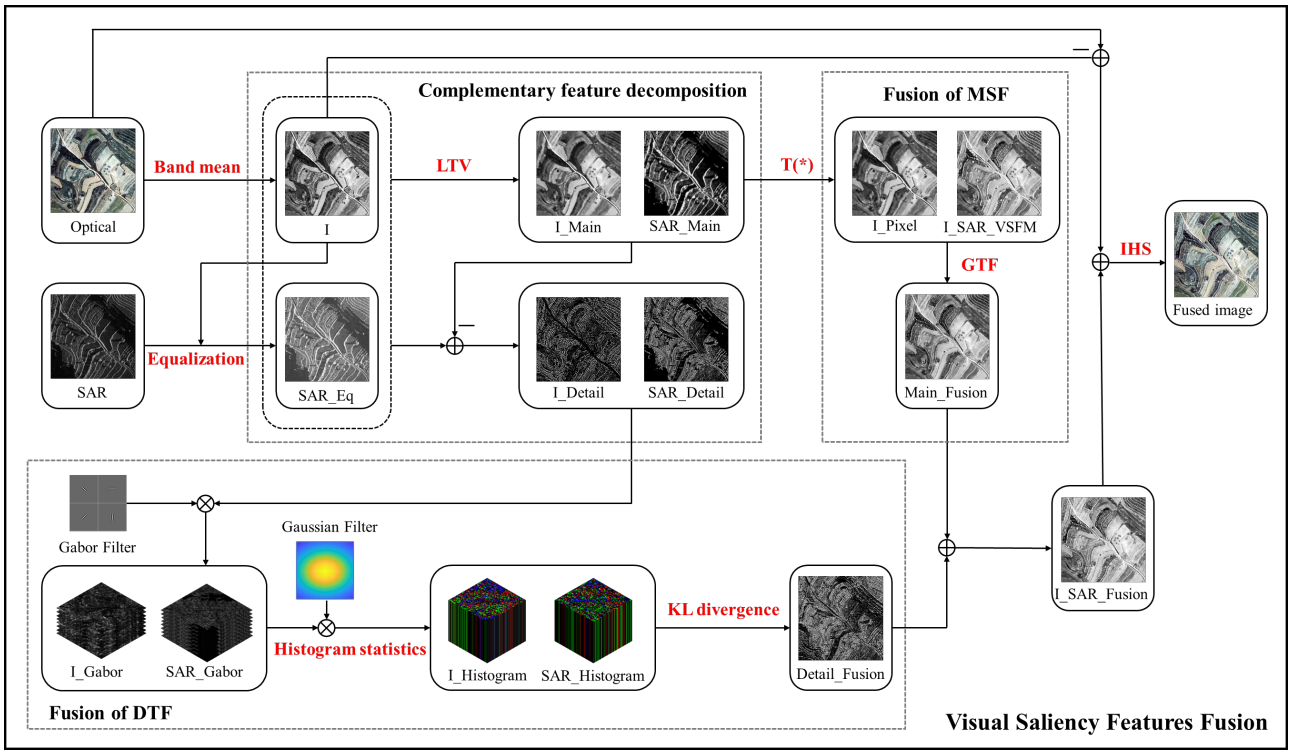


Figure 2. A fusion framework for optical and SAR images based on visual saliency features fusion.

2. A FUSION FRAMEWORK BASED ON VISUAL SALIENCY FEATURES

In this section, we describe the VSFF fusion framework in detail. Figure 2 shows the whole fusion framework in detail, which consists of four critical parts and contributions: 1) an improved complementary feature decomposition algorithm can effectively suppress speckle noise, 2) a total variation fusion algorithm that introduces visual saliency features can enhance the overall visual perception, 3) a novel texture feature descriptor can preserve richer detail information. 4) a fast IHS transform fusion can supplement the realistic color information.

2.1 Image Complementary Feature Decomposition

Any remote sensing image can be decomposed into a set of complementary features: MSF and DTF. The source image and the decomposed parts are defined as follows:

$$f = u + v \tag{1}$$

where f = optical or SAR remote sensing image
u = MSF
v = DTF

In the first step, we need to build a local indicator to divide whether each pixel belongs to MSF or DTF. MSF is the part of the image that has relatively stable local variation at different scales, while DTF is the part that tends to have large local variation after filtering. The local total variation (LTV) of the image can effectively respond to the relative degree of variation under low-pass filtering to distinguish MSF from DTF, and we define the LTV of the image and its relative reduction rate as follows:

$$LTV_\sigma(f)(x) = L_\sigma \times |\nabla f|(x) \tag{2}$$

$$\lambda(x) = \frac{LTV_\sigma(f)(x) - LTV_\sigma(L_\sigma \times f)(x)}{LTV_\sigma(f)(x)} \tag{3}$$

where L_σ = nonlinear filter
λ = relative reduction rate
∇f = gradient image
x = pixel position

As can be seen, LTV is obtained by low-pass filtering the image gradient map, and the relative reduction rate gives us the oscillatory behavior of the image in the local area. Further, it is necessary to consider that there are many fragmented feature edges and speckle noise in SAR images, which will show bright spots on the image and can be easily taken as MSF. To suppress speckle noise, the Wiener wavelet is selected for smoothing the image, and this filter can adaptively adjust the filter effect based on local gray information.

In the second step, the decomposition of image complementary features is achieved by a set of fast low-pass and high-pass filter pairs, which are calculated by weighting the relative reduction rate of the LTV of the image over the original and filtered images. The specific calculations for this step of the operation are as follows:

$$u(x) = w(\lambda(x))L_\sigma \times f + (1 - w(\lambda(x)))f \tag{4}$$

$$v(x) = f(x) - u(x) \tag{5}$$

$$w(x) = \begin{cases} 0 & x \le a_1 \\ (x - a_1)/(a_2 - a_1) & a_1 \le x \le a_2 \\ 1 & x \ge a_2 \end{cases} \tag{6}$$

where w = pixel weight
a_1 = 0.25
a_2 = 0.5

2.2 Fusion Strategy of MSF

In the fusion of MSF, structural information is presented by a combination of pixel grayscale distribution and gradient variation. Among them, the pixel grayscale distribution is the essential information to distinguish the area and type of terrestrial scene, which directly determines the overall visual effect of the fusion result. We expect the grayscale distribution of the fusion result to be similar to the optical image in order to have a more natural visual perception. Meanwhile, gradient variation is an important feature information in the image that can easily attract human attention, and it is also the expression of VSF in the image. The VSF information from optical and SAR images should be integrated to offer more informative interpretability for the fusion results. Therefore, the fusion strategy follows the principles: the pixel grayscale distribution of the fused image is similar to that of the optical image, and the gradient variation information of the fused image is similar to that of the optical and SAR images. According to the above principles and conditions, the mathematical constraint model can be constructed as follows:

$$E_1(x) = \frac{1}{p}\left\|x - u_o\right\|_p^p \tag{7}$$

$$E_2(x) = \frac{1}{q}\left\|\nabla x - \nabla u_{os}\right\|_q^q \tag{8}$$

where x = fusion result of MSF
u_o = MSF of optical image
u_{os} = VSFM
p, q = norm

Then, we introduce the visual saliency feature-based method to generate a new VSFM as the input information in the constraint Eq. (8). On the one hand, we compare the feature values of the equalized SAR image and the optical image, and treat the images with prominent features as significant VSF and retain them to the new VSGM. On the other hand, we perform gain processing on the VSF from SAR images to better present the contribution of large contour structure information to the fusion results. The u_{os} calculation process is as follows:

$$u_{os} = T(u_o, u_s) = k_1 u_o + (1 - k_1) k_2 u_{s\_Eq} \tag{9}$$

$$u_{s\_Eq} = \frac{(u_s - \mu(u_s))\sigma(u_o)}{\sigma(u_s)} + \mu(u_o) \tag{10}$$

$$k_1 = \begin{cases} 1 & u_o(m,n) > u_{s\_Eq}(m,n) \\ 0 & u_{s\_Eq}(m,n) \geq u_o(m,n) \end{cases} \tag{11}$$

Where u_{s_Eq} = equalized SAR image
σ = variance
μ = mean
k_1 = feature weight
k_2 = feature gain
m, n= pixel coordinates

Next, the objective function of the fused image is generated by combining the above two constraint terms Eq. (7) (8). The λ are positive parameters that control the trade-off between the two terms.

$$E(x) = E_1(x) + \lambda E_2(x) = \frac{1}{p}\left\|x - u_o\right\|_p^p + \lambda\frac{1}{q}\left\|\nabla x - \nabla u_{os}\right\|_q^q \tag{12}$$

where λ = positive parameter

Now we need to consider the specific p and q norm. In the constraint term (7), the best expected result is 0, so p=1. As the gradient of the image is sparsely distributed, an approximate solution to handle the problem that q=0 is NP-hard is to replace l^0 norm by l^1 norm. Thus, the gradient difference minimization problem is converted to a total variational problem. Let $y = x - u_{os}$, the optimization problem (12) can be rewritten as:

$$y^* = \arg\min_y \left\{\sum_{i=1}^{mn}\left|y_i - (u_{oi} - u_{osi})\right| + \lambda J(y)\right\} \tag{13}$$

$$J(y) = \sum_{i=1}^{mn}\left|\nabla_i y\right| = \sum_{i=1}^{mn}\sqrt{(\nabla_i^h y)^2 + (\nabla_i^v y)^2} \tag{14}$$

where argmin = minimization solution
J = first derivative of the image

The Eq. (13) is a standard l^1 total variation minimization problem. (Rodr´ıguez and Wohlberg, 2008) offer an algorithm for solving generalized total variation minimization models using the iterative weighted norm (IRN) algorithm. The algorithm can efficiently compute y and then generate the final MSF fused image.

2.3 Fusion Strategy of DTF

Generally, the DTF of optical images contain rich and fine edge information, while the DTF of SAR images contain some valuable information of tiny terrain radiation. In addition, it is inevitable that a part of SAR image noise is blended into the DTF components in the complementary feature decomposition, which needs to be separated out before fusion. Therefore, the main purpose of DTF fusion is to selectively retain richer and more meaningful features information and further remove interference noise information. For the specific fusion strategy, it is necessary to first describe the input DTF more effectively, then select the feature information with high interpretability, and finally preserve the more informative features through feature similarity measure processing.

DTF are locally oscillating distribution in the image with strong repetition and orientation. Even though DTF may vary at different scales, the highly interpretable texture features in them always have a stable information representation. For that reason, a new feature descriptor is designed that can capture multi-directional and multi-scale texture information within a local region of the image. It is known that Gabor wavelet is a kernel function similar to the response to simple stimuli in the human visual system. Meanwhile, Gabor wavelet is also sensitive to the image edge information, which is an excellent texture feature filter. The following is the mathematical form of Gabor wavelet filter:

$$\begin{aligned} g_{\lambda,\theta,\varphi,\sigma,\gamma}(x,y) &= e^{-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}}\cos(2\pi\frac{x'}{\lambda} + \varphi)\gamma \\ x' &= (x\cos\theta + y\sin\theta) \\ y' &= (-x\cos\theta + y\sin\theta) \end{aligned} \tag{15}$$

where λ = wavelength parameter of the cosine function
θ = strip direction
φ = phase parameter of the cosine function
σ = standard deviation of the Gaussian
γ = spatial aspect ratio
x, y = coordinates inside the filter

After obtaining new DTF at different scales and orientations using the Gabor wavelet function, Gaussian filtering is used to obtain the most stable DTF representation of the main scale and orientation. Subsequently, the noise that has not been eliminated needs to continue to be processed. Due to the fact that speckle noise appears as cluttered bright spots in the DTF, a direct feature selection operation would result in retaining a large amount of noise in the fusion result. In this case, considering that the speckle noise presents an irregular distribution, while the DTF are essentially regularly gathered in a local area. Therefore, local histogram statistics can be used to improve the reliability of the DTF and eliminate the random speckle noise. After a series of operational steps, the obtained feature information has high interpretability and relative stability. The specific processing steps are shown in Figure 3.

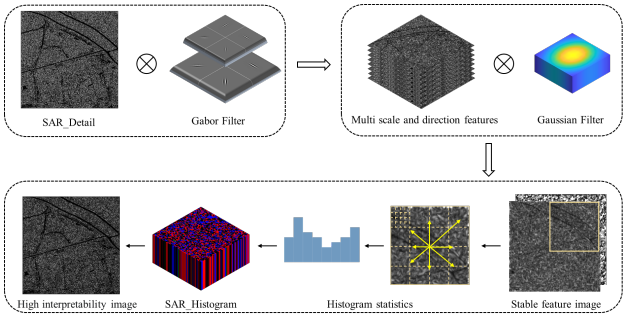


Figure 3. The fusion process of DTF includes Gabor wavelet-based feature description and Gaussian filtering to select stable feature information, and histogram statistics to retain high interpretability features.

Finally, the high interpretability features from optical and SAR images are integrated and the richer features are selected to be preserved in the fusion results. We first perform a similarity measure on the obtained features, and when the features from optical and SAR images are similar, the average of the feature values from the two images is directly taken as the fusion result. In the case where the features are not similar, the features with larger local gradient values are considered as enriched information to be retained in the fusion result. Next, we consider determining the similarity measure algorithm that normalizes the feature description vector to form a statistical vector of feature probability distributions. KL divergence is the best choice for the similarity measure, which provides an asymmetric measure of the difference of two probability distributions, and it is defined as follows:

$$KL[P \| Q] = \sum_{x \in X} [P(x) \log \frac{P(x)}{Q(x)}] \tag{16}$$

where P, Q = probability distribution vector
x = coordinates inside the vector

Specifically, KL divergence is not satisfying symmetry, i.e., $KL[P \| Q] \neq KL[Q \| P]$. We construct a new similarity measure by computing the average value of $KL[P \| Q]$ and $KL[Q \| P]$. If the value is smaller, it means that the features at that pixel are more similar. The new similarity metric value (SMV) is calculated as follows:

$$SMV = \frac{KL[P \| Q] + KL[Q \| P]}{2} \tag{17}$$

Next, a suitable threshold is chosen to determine whether the feature information is similar, which is defined by calculating the mean of the SMV of all pixels in the image. In summary, we give the computing steps for the fusion of DTF as follows:

$$v_f(m,n) = \begin{cases} \dfrac{v_O(m,n) + v_S(m,n)}{2} & SMV_{(m,n)} < SMV_{mean} \\ H(m,n) & SMV_{(m,n)} \geq SMV_{mean} \end{cases} \tag{18}$$

$$H(m,n) = \begin{cases} v_O(m,n) & G_O(m,n) \geq G_S(m,n) \\ v_S(m,n) & G_O(m,n) < G_S(m,n) \end{cases} \tag{19}$$

where SMV_{mean} = the mean of the SMV of all pixels
v_O , v_S = feature values
G_O , G_S = gradient values
m,n = image coordinates
v_f = fusion result of DTF

2.4 Fast IHS Transform Method

The complementary feature decomposition inevitably results in the loss of some spectral information in the image. Therefore, in order to recover the realistic image color information, we need to process the fused results again. The intensity–hue–saturation (IHS) fusion is a classical image fusion method, which can realize the transfer of spectral information through simple calculation. In brief, the IHS transform divides the optical image into I, H and S components, where H and S contain the spectral color information of the optical image. It is possible to supplement optical color information to the final fusion result by replacing the original I component image with the fusion result of the optical image intensity information and SAR image in the fusion process. (Tu et al., 2004) introduces a fast IHS transform method with the following main steps:

$$\begin{cases} R_f = R + I_f - I \\ G_f = G + I_f - I \\ B_f = B + I_f - I \end{cases} \tag{20}$$

where R_f , G_f , B_f = fusion result of RGB bands
R , G , B = RGB bands of optical image
I_f = fusion results of the intensity components
I = intensity component of optical image

3. EXPERIMENTAL RESULTS

In this section, we proposed method is compared with five state-of-the-art fusion methods: LP (Burt and Adelson, 1987), DTCWT (Selesnick et al., 2005), NSCT (Da Cunha et al., 2006), Hybrid-MSD (Zhou et al., 2016), WLS (Ma et al., 2017). Among them, the first three are classical multiscale decomposition methods, the fourth is the latest hybrid multiscale decomposition method, and the fifth is a fusion method based on visual saliency map.

In our experiments, the quality of image fusion is evaluated in qualitative and quantitative terms. Qualitative evaluation is a visual perception analysis of the overall image and local details of the fusion result. Of course, there are some differences in the visual perceptual focus of different source images. A large number of index theories have emerged in image fusion for quantitative evaluation, including the measurement of image characteristics such as image information, gradient and structural similarity. Each of the image evaluation indexes has its advantages and disadvantages, so it is necessary to synthesize multiple indexes. Six evaluation indexes EN, MI, SF, SD, Qabf and Qo are adopted in this paper (Zhang et al., 2020). According to different types of information description, these indexes can be divided into information theory based, image feature based, structural similarity based, original image and fused image correlation.

3.1 Datasets and Parameter Settings

To verify the fusion effect on optical and SAR salient features and noise removal capability of the experimental algorithm. All algorithms are tested on a high-resolution (sub-meter level) SAR and optical dataset and a publicly available WHU-OPT-SAR dataset. Here is a detailed description of datasets.

High-resolution SAR and optical dataset: The dataset is a high-resolution SAR image (0.5m) acquired from the surrounding areas of Baicheng City in Jilin Province and Weinan City in Shaanxi Province, including three typical ground object scenes of houses, farmland and mountains. After downloading the Google-Earth optical images of the corresponding areas and performing high-precision matching (Ye et al., 2019), they were further cropped into 1000*1000-pixel image pairs, and then 60 pairs of images with abundant information of scenes were selected as the test dataset.

WHU-OPT-SAR dataset: (Li et al., 2022) open sources a set of optical and SAR image dataset collected in Hubei province. Optical images are from GF-1 satellite (2m resolution), and SAR images are from GF-3 satellite (5m resolution). WHU-OPT-SAR dataset covers a wide range of area, including diverse terrains such as mountains, woodlands, hills, plains and vegetation. In order to better show the fusion details, the image of the dataset was cropped to the size of 1000*1000 pixels in the experiment.

Parameter settings: In order to suppress the effect of SAR image speckle noise, the size of the Wiener filter is set to 3 in the complementary feature decomposition module. To enhance the fused visual effect of VSF from SAR images, the feature gain k_2 was set to 1.2 and the positive parameter λ is set to 20. For more detailed texture description, the scale of the Gabor wavelet is set to [4, 8], the direction is set to [0°, 45°, 90°, 135°], and other parameters are defaulted.

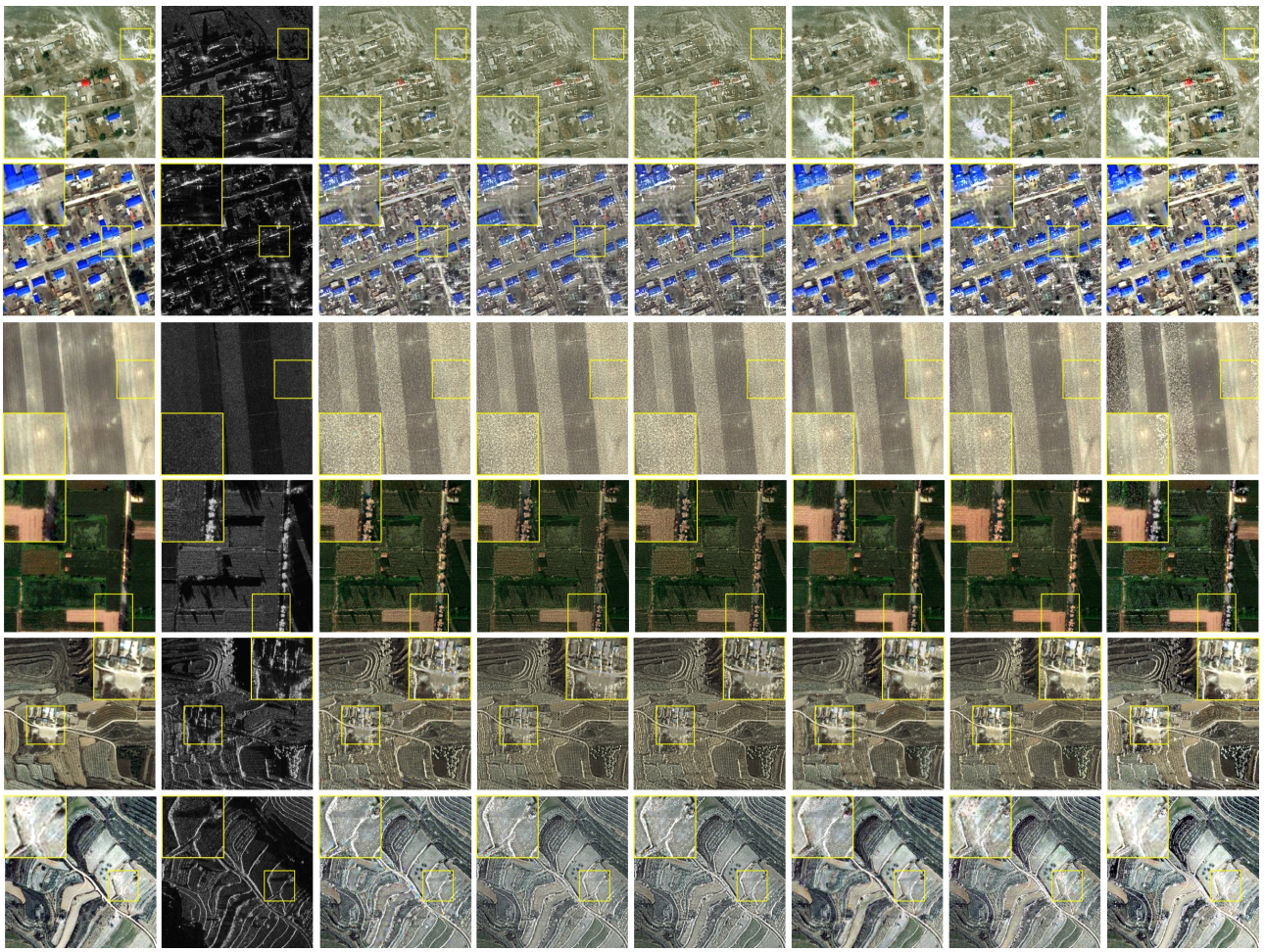


Figure 4. Qualitative evaluation of fusion results from six different scenarios (from the high-resolution SAR and optical dataset). From left to right are optical and SAR images, LP, DTCWT, NSCT, Hybrid-MSD, WLS and our VSFF fusion results.

3.2 Result Analysis of High-Resolution SAR And Optical Dataset

The fusion results are shown in Figure 4, all the methods have a good fusion of the house, road, and field contour boundaries in SAR images, as well as SAR unique information such as tree shadows and field crop textures. However, in the three classical multi-scale decomposition methods, LP, DTCWT, and NSCT, the spectral information of optical images is seriously damaged. It can be obviously seen that the low gray level information of the SAR image leads to the overall darkening of the fusion result, and some important ground object scenes are covered by shadows. In farmland and mountain images, color information is an important condition to judge the species of covered plants, which must be guaranteed to be completely consistent with optical images. In contrast, while Hybrid-MSD retains better optical color information, it lacks some key salient information. For example, in the close-up scene in the fifth row, it can be seen that Hybrid-MSD is missing the field edge information provided by SAR in the lower left. Similarly, in the close-up scene in the sixth row, the SAR tree shadow interference causes it to lose the important field path information provided by the optical image. The overall vision of the WSL fusion results is still disturbed by the hue and noise of the SAR images, and there is some spectral distortion. As a result, in some WSL scenes, the visually salient information of optical and SAR images cannot be better balanced. For instance, the excessive focus on SAR in the first two images makes the noise information on houses and roads serious and prevents the correct observation of contour and demarcation. On the contrary, the excessive focus on optical in the latter four images results in inconspicuous trees and their shadows and unclear field outlines. From the all-close-up views, the VSFF method achieves the best optical color fusion and highlights the visual perception of SAR main contour information while eliminating SAR image noise interference. Benefiting from the advantage of the selection of salient features, the unique details of each of optical and SAR are perfectly preserved without interfering with each other. Table 1 shows the quantitative analysis of the fusion results from different methods on the high-resolution SAR and optical dataset. We can see that VSFF outperforms other methods on most indexes, which indicates that VSFF has a great advantage in integrating structure and detail information.

Methods	EN	MI	SF	SD	Qabf	Qo
LP	7.21	1.22	29.61	43.83	<u>0.47</u>	0.38
DTCWT	7.12	1.12	29.89	40.52	0.41	0.36
NSCT	7.15	1.19	29.94	41.51	0.46	0.39
Hybrid-MSD	<u>7.33</u>	1.28	28.27	<u>45.69</u>	0.45	0.40
WLS	7.21	<u>1.46</u>	<u>30.33</u>	45.39	0.45	**0.42**
VSFF	**7.51**	**2.87**	**39.12**	**56.55**	**0.62**	<u>0.41</u>

Table 1. Quantitative evaluation of fusion results (from the high-resolution SAR and optical dataset), with the best results highlighted in bold and the second-best results in underlined.

3.3 Result Analysis of WHU-OPT-SAR Dataset

When tested on a medium resolution dataset, as shown in Figure 5, the VSFF fusion results have a more detailed representation of the salient feature information and excellent denoising effect for different images. On the one hand, the close-up areas of the fusion result map look cleaner and more comfortable, and the optical and SAR image features are clearly distinguished. Especially in the first, second, and fifth scenes, the VSFF method care about the real texture and color information of the land and houses than other methods and avoids having them obscured by the cluttered speckle noise. On the other hand, the

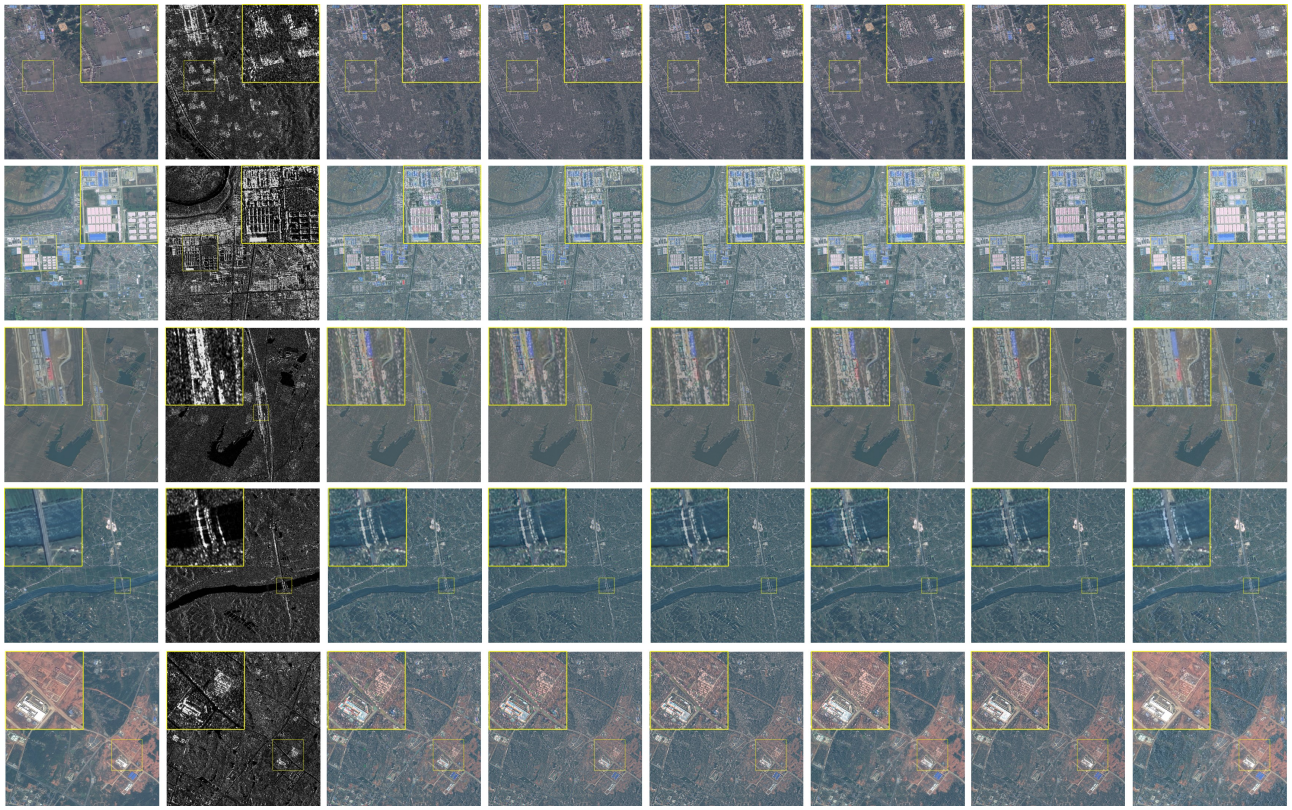


Figure 5. Qualitative evaluation of fusion results from five different scenarios (from the WHU-OPT-SAR dataset). From left to right are optical and SAR images, LP, DTCWT, NSCT, Hybrid-MSD, WLS and our VSFF fusion results.

VSFF method can accurately retain salient and important target information. For example, in the third and fourth scenes, the red trains and bridges in the optical images are highly valuable features. However, it is obvious that the fusion results of other methods show blurred or even non-existent, only our method reconstructs the target completely. In conclusion, the VSFF fusion method achieves effective removal of SAR image noise without losing important feature information, and restores the real color of ground objects to the maximum extent. Table 2 shows the quantitative analysis of the fusion results from different methods on the WHU-OPT-SAR dataset. Among them, the EN and MI indexes show the high information content of the VSFF fusion results, which intuitively indicates that the salient features extracted are the high interpretability features of the images. In particular, the higher SF index indicates that the fusion results have better clarity. Thus, the fusion results of VSFF are more suitable for visual interpretation. As only some of the structural features of SAR images are selected for the saliency features in this paper, the global calculation of the Qo metrics may appear to be low.

Methods	EN	MI	SF	SD	Qabf	Qo
LP	5.89	1.39	18.03	16.72	0.26	0.38
DTCWT	5.73	1.16	17.28	15.49	0.21	0.37
NSCT	5.76	1.32	17.48	16.00	0.24	0.40
Hybrid-MSD	5.93	1.35	19.22	17.28	0.24	0.41
WLS	5.86	1.38	19.29	16.59	0.25	**0.42**
VSFF	**6.34**	**1.89**	**29.62**	**24.76**	**0.29**	0.35

Table 2. Quantitative evaluation of fusion results (from the WHU-OPT-SAR dataset), with the best results highlighted in bold and the second-best results in underlined.

4. CONCLUSION

In this paper, we propose a novel optical and SAR image fusion framework named VSFF based on visual saliency features. It extracts the salient and complementary information of the image and then achieves the purpose in different fusion methods and rules. From the fusion results, it is obvious that our method eliminates more noise and retains the salient and important feature targets in both optical and SAR images. Surely, our method achieves good results in several quantitative evaluation indexes when compared with five state-of-the-art fusion methods. It proves that our fusion results have richer spectral information and clearer visual perception.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 41971281, Grant 42271446, and in part by the Natural Science Foundation of Sichuan Province under Grant 2022NSFSC0537.

REFERENCES

Buades, A., Le, T. M., Morel, J.-M., Vese, L. A., 2010. Fast cartoon+ texture image filters. *IEEE Transactions on Image Processing*, 19(8), 1978–1986.

Burt, P. J., Adelson, E. H., 1987. The Laplacian pyramid as a compact image code. 671–679.

Da Cunha, A. L., Zhou, J., Do, M. N., 2006. The nonsubsampled contourlet transform: theory, design, and applications. *IEEE transactions on image processing*, 15(10), 3089–3101.

Kong, Y., Hong, F., Leung, H., Peng, X., 2021. A Fusion Method of Optical Image and SAR Image Based on Dense-UGAN and Gram–Schmidt Transformation. *Remote Sensing*, 13(21), 4274.

Kulkarni, S. C., Rege, P. P., 2020. Pixel level fusion techniques for SAR and optical images: A review. *Information Fusion*, 59, 13–29.

Li, X., Zhang, G., Cui, H., Hou, S., Wang, S., Li, X., Chen, Y., Li, Z., Zhang, L., 2022. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106, 102638.

Ma, J., Chen, C., Li, C., Huang, J., 2016. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31, 100–109.

Ma, J., Zhou, Z., Wang, B., Zong, H., 2017. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Physics & Technology*, 82, 8–17.

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., Papathanassiou, K. P., 2013. A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, 1(1), 6-43.

Rodriguez, P., Wohlberg, B., 2008. Efficient minimization method for a generalized total variation functional. *IEEE Transactions on Image Processing*, 18(2), 322–332.

Selesnick, I. W., Baraniuk, R. G., Kingsbury, N. C., 2005. The dual-tree complex wavelet transform. *IEEE signal processing magazine*, 22(6), 123–151.

Toet, A., 2011. Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *IEEE transactions on pattern analysis and machine intelligence*, 33(11), 2131–2146.

Tu, T.-M., Huang, P. S., Hung, C.-L., Chang, C.-P., 2004. A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery. *IEEE Geoscience and Remote sensing letters*, 1(4), 309–312.

Ye, Y., Bruzzone, L., Shan, J., Bovolo, F., Zhu, Q., 2019. Fast and Robust Matching for Multimodal Remote Sensing Image Registration. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11), 9059–9070.

Zhang, X., Ye, P., Xiao, G., 2020. Vifb: A visible and infrared image fusion benchmark. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 104–105.

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L., 2020. IFCNN: A General Image Fusion Framework Based on Convolutional Neural Network. *Information Fusion*, 54, 99–118.

Zhou, Z., Wang, B., Li, S., Dong, M., 2016. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters. *Information Fusion*, 30, 15–26