An Unsupervised SAR and Optical Image Fusion Network Based on Structure-Texture Decomposition

Yuanxin Ye, *Member IEEE*, Wanchun Liu, Liang Zhou, Tao Peng, Qizhi Xu

***Abstract*—Although the unique advantages of optical and synthetic aperture radar (SAR) images promote their fusion, the integration of complementary features from the two types of data and their effective fusion remains a vital problem. To address that, a novel framework is designed based on the observation that the structure of SAR images and the texture of optical images look complementary. The proposed framework, named SOSTF, is an unsupervised end-to-end fusion network that aims to integrate structural features from SAR images and detailed texture features from optical images into the fusion results. The proposed method adopts the nest connect-based architecture, including an encoder network, a fusion part, and a decoder network. To maintain the structure and texture information of input images, the encoder architecture is utilized to extract multi-scale features from images. Then, we use the densely connected convolutional network (DenseNet) to perform feature fusion. Finally, we reconstruct the fusion image using a decoder network. In the training stage, we introduce a structure-texture decomposition model. In addition, a novel texture-preserving and structure-enhancing loss function are designed to train the DenseNet to enhance the structure and texture features of fusion results. Qualitative and quantitative comparisons of the fusion results with nine advanced methods demonstrate that the proposed method can fuse the complementary features of SAR and optical images more effectively.**

***Index Terms*—SAR and optical images, image fusion, SOSTF, unsupervised**

I. Introduction

Remote sensing images from different sensors in the same scene reflect the complementary content of ground objects. These multi-sensor remote sensing images are organized, associated, and synthesized according to certain rules to achieve information complementation [1]. High-quality fusion of remote sensing images can meet different application requirements, such as object recognition, change detection, and image classification [2]. Optical images can obtain information about surface characteristics, such as true texture and greyscale information, but are susceptible to clouds or other extreme weather conditions. On the contrary, SAR has a strong penetrating ability and is not affected by climate and environment. Nevertheless, SAR images generally are corrupted by speckle noise. Thus, we focus on reasonably integrating their complementary information and effectively reducing speckle noise of SAR images to serve in more scenes.

Currently, optical and SAR image fusion methods mainly draw on the fusion idea of multispectral images and panchromatic images (*i.e.*, Pan-sharpening fusion approach). For instance, some academics integrated the spatial and spectral information from SAR images and optical images [3, 4]. Pal *et al*. [5] combined the principal component analysis (PCA) method and feature-oriented principal component selection technology to fuse the ERS-2 SAR and IRS-1C LISS III data for generating false-color composite images. Chen *et al*. [6] introduced a wavelet transform and empirical mode decomposition into a generalized IHS transform to propose a spectral preserve fusion method. Furthermore, related researchers have proposed the area-based image fusion algorithm [7], SAR and optical image fusion based on fast sparse representation of low-frequency images [8], and image fusion based on the curvelet transform [9]. These methods perform image fusion at the pixel-level and inevitably require complex image transformations or the design of fusion rules.

As a result of the quick development of artificial intelligence, deep learning (DL) has assumed a more critical role in remote sensing image fusion in recent years. The advantage of DL is that a large amount of high-level semantic information can be automatically extracted from images. As expected, current DL techniques have been applied to multi-focus or multi-exposure image fusion, as well as panchromatic and multispectral image fusion. But few studies aim to fuse SAR and optical images using DL at the pixel-level. To our best knowledge, only Kong *et al*. [10] proposed a SAR-optical fuse method using the Dense-UGAN and Gram-Schmidt transform. However, it is still a semi-supervised method and requires a certain number of labels for training. The other scholars focus on optical-SAR fusion applications such as the identification and extraction of clouds, glaciers, and sea ice [11–13]. These DL-based fusion methods avoid manual image transformation and designed rules so that the fusion process is simpler and more adaptable. Accordingly, an unsupervised optical-SAR fusion architecture using DL is worth exploring.

Optimizing the integration of their complementary features is crucial for the fusion of optical and SAR images. Taking infrared and visible image fusion as an example [14], the radiation intensity of infrared images and the detail of visible images are widely accepted as the complementary features of

This paper was supported by the National Natural Science Foundation of China under Grant 41971281, 61972021, and 42271446. (*Corresponding author:Qizhi Xu*)

Y. Ye, W. Liu, L. Zhou and T. Peng are with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 610031, China, and also with the State-province Joint Engineering Laboratory of Spatial Information Technology for High-speed Railway Safety, Southwest Jiaotong University, 611756, China (e-mail: yeyuanxin@home.swjtu.edu.cn).

Qizhi Xu is with the School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China. (e-mail: qizhi@bit.edu.cn)

This article has been accepted for publication in IEEE Geoscience and Remote Sensing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LGRS.2022.3219341

them, respectively. Motivated by that, we try to find the significant complementary features between optical and SAR images. Fig.1 illustrates their characteristics in detail. It can be clearly observed that optical images can accurately reflect the real texture, fine edges, and other details of ground objects, with rich texture and powerful visual interpretation capability. In contrast, the main structure and large edges of the SAR image are clearer, such as road directions and mountain contours (Fig. 1b). Based on such observations, we can speculate that the texture of the optical image and the structure of the SAR image are a pair of complementary features. Besides fusing complementary features, we also consider how to simultaneously denoise SAR images because it is difficult to balance speckle reduction and preserve features in the post-processing of SAR images.

Focusing on the fusion of complementary features of SAR and optical images, we attempt to develop a fusion framework employing the DL technique. In this way, we construct a nest connection network to carry out image fusion of SAR-based structures and optical-based textures (named SOSTF), which can automatically synthesize the input images into the fused images without any handcrafted labels.

(a) Optical image (b) SAR image

Fig. 1. Examples of optical and SAR images

II. Methodology

The innovations of the proposed SOSTF method consists of two aspects: 1) A novel fusion framework, which is shown in Fig. 2, includes encoder and decoder networks for feature extraction and reconstruction as well as the DenseNet fusion network for feature fusion; 2) We devise a loss function to constrain structure and texture for the output result and effectively reduce speckle noise. The framework is described in detail as below.

A. Architecture Overview

The nest connection-based network utilized for feature extraction and image reconstruction includes encoder and decoder modules. Each of the four encoder blocks inspired by RFN-Nest [15] is composed of three layers, including convolution, ReLU activation, and max pooling. This architecture has the ability to extract various deeper semantic information. By convoluting and up-sampling, the decoder network eventually reconstructs the fusion image. The DenseNet blocks for fusing multi-scale information are a critical part between the encoder and decoder. Four individual DenseNet blocks are applied to fuse features from SAR and optical images (*i.e.*, the output of the encoder) at four different scales.

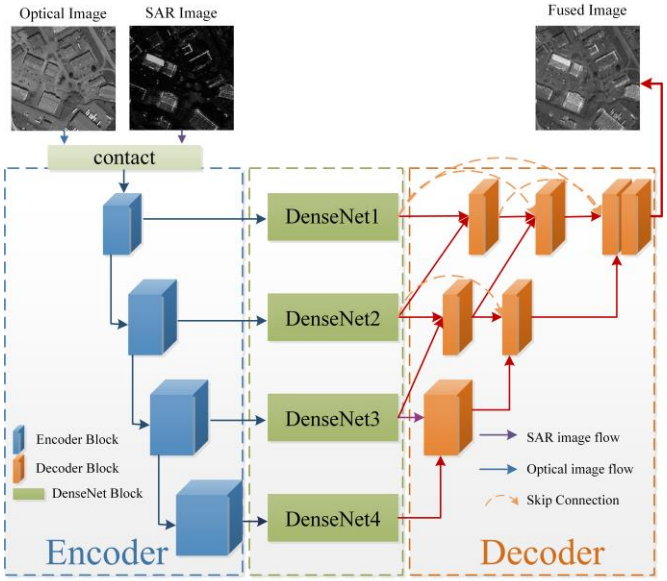

Fig. 2. Overview of the proposed SOSTF.

B. Extraction of Structure and Texture

To provide important constraints for the network model, especially structure and texture information, we introduce an image decomposition method referred to as a cartoon-texture decomposition [16] in order to accomplish effective segmentation and enhancement for images. This approach decomposes an image into the cartoon part (*i.e.*, structure) and the textural part, which is similar to classical signal processing (*i.e.*, low-pass and high-pass filter decomposition). The extraction process of structure and texture is formulated as follows.

$$\lambda = \frac{LTV(I_{raw}) - LTV(I_{con})}{LTV(I_{raw})} \tag{1}$$

$$W(\lambda) = \begin{cases} 0, & \lambda < 0.25 \\ \dfrac{\lambda - 0.25}{0.25}, & 0.25 \le \lambda \le 0.5 \\ 1, & \lambda > 0.5 \end{cases} \tag{2}$$

$$I_{str} = W(\lambda) \cdot (I_{con} - I_{raw}) + I_{raw} \tag{3}$$

Where I_{raw} represents the original image, I_{con} denotes the image after Gaussian convolution, and I_{str} is the structure image. I_{tex} denotes the texture image that is produced by the simple subtraction of the original image and the structure image. Local total variation (LTV (·)) represents the LTV process of performing Gaussian convolution on image gradients. W (·) is used for normalization, which is shown in Equation 3. The parameter λ is an indicator illustrating the LTV reduction rate. Fig. 3 shows an example of image decomposition.

C. Loss Function

Generally, the loss function is used to measure the differences between the predicted result and ground truth during the network training stage. The loss function in this study is designed for the source and reconstructed images. We directly constrain the distinctions between features of input and

This article has been accepted for publication in IEEE Geoscience and Remote Sensing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LGRS.2022.3219341

output to achieve unsupervised fusion. As shown in Fig. 3,

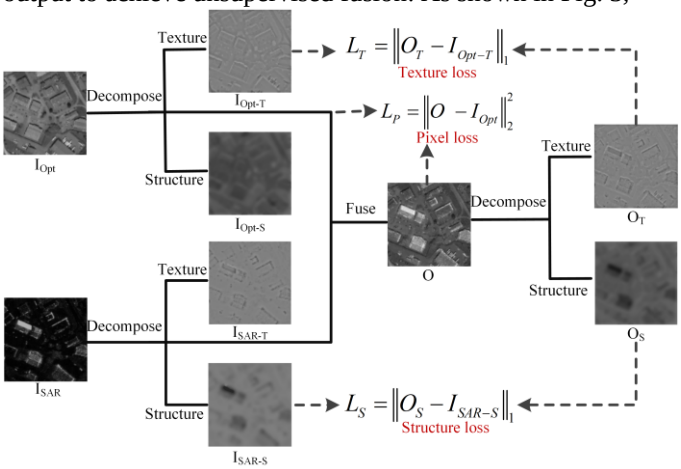

Fig. 3. Constraint strategy of structure and texture in loss function

considering the speckle noise of the input SAR image, we introduce three key constraints (*i.e.*, pixels, textures, and structures) to gradually capture the complementary features of the optical and SAR image and contain as little noise as possible. This is primarily due to the pixel loss in the loss function accounting for a larger proportion, and the structure loss acts on the structural part with detail inhibition of the SAR image, which makes the background of the fusion result closer to the optical image and filters out part of the noise of the SAR image. Note that, the design of the loss function only applies to the different DenseNet modules because the used nest framework has been trained in the first training stage. That is, the reverse adjustment of the DenseNet module parameters can make these prediction results better. Specifically, the designed loss function is defined as:

$$L_{total} = L_P + \alpha_1 L_S + \alpha_2 L_T \tag{4}$$

Where L_{total} is the total loss in the fusion process. L_S and L_T represent structure loss and texture loss, respectively. α_1 and α_2 represent the weights of the two loss functions, respectively. In order to maintain the real information from the optical image to make the background of the fusion result more similar to the original optical image, we design a pixel loss called L_P:

$$L_P = \left\| O - I_{Opt} \right\|_2^2 \tag{5}$$

Where O is the fused output image and I_{Opt} is the input optical image. The pixel loss is determined by the L2-norm of the pixel difference between O and I_{Opt}.

After the structure and texture information of images are extracted by the process described in Section II-B, they are introduced into the objective function to constrain the structure and texture information of the fusion image. Therefore, we define the loss of structure and texture as follows:

$$L_T = \left\| O_T - I_{Opt-T} \right\|_1 \tag{6}$$

$$L_S = \left\| O_S - I_{SAR-S} \right\|_1 \tag{7}$$

Where O_T stands for the fused image's texture, $I_{Opt\text{-}T}$ for the input optical image's texture, O_S for the fused image's structure, and $I_{SAR\text{-}S}$ for the input SAR image's structure. The structure loss L_S and texture loss L_T are determined by the L1-norm of the difference between the structure and texture components, which are from the fused image and the original input image, respectively.

III. Experiments

To our best knowledge, there are currently few unsupervised DL-based fusion methods for optical and SAR images. For this reason, we choose some state-of-the-art fusion methods to compare with the proposed method. Among these, traditional methods include Intensity-Hue-Saturation (IHS), High-Pass Filter (HPF), Laplacian Pyramid (LP), Discrete Wavelet Transform (DWT), Curvelet Transform (CVT), Dual-Tree Complex Wavelet Transform (DTCWT), and Nonsubsampled Contourlet Transform (NSCT), and DL-based fusion methods include NestFuse [17] and U2Fusion [18]. To more precisely evaluate the fusion result, we perform both quantitative and qualitative evaluations. The quantitative evaluation is carried out by using some classic objective metrics. These metrics include entropy (EN), peak signal-noise ratio (PSNR), mean squared error (MSE), correlation coefficient (CC), the sum of the correlations of differences (SCD), and structure similarity (SSIM). The larger values of these metrics indicate better fusion quality, except for the MSE. The qualitative evaluation mainly depends on human subjective visual evaluation in our study.

A. Datasets

We use a high-resolution optical and SAR image dataset, which was built by Xiang *et al.*, [19] and have been further registered precisely by the method present in [20]. This dataset is declared to be suitable for DL-based image fusion tasks. Table I gives the details of these images in the dataset, including the number of image pairs, image size, resolution, etc.

TABLE I
DETAILS OF THE TEST DATA SETS

SENSOR	Quantity	Size	Resolution	Source
OPTICAL	9740 pairs	256×256 pixels	Resampled to 1-m	Google Earth
SAR			1-m	GaoFen-3

In the dataset, we only select the images of size 256×256. The training data set contains 8044 pairs of optical and SAR images, while the test data set contains 1696 pairs.

B. Implementation Details

Before the fusion training phase, a log function is used to stretch the SAR images to alleviate the effects of strong exposure. The encoder and decoder networks are first trained as the auto-encoders in our two-stage training method. In the second training phase, the auto-encoders' parameters won't be adjusted. Next, we train the DenseNet network through the designed loss function. During this training, the two images are concatenated together and fed into the model. If processing RGB inputs, users can first convert them to the YCbCr color space and choose the Y (luminance) channel for fusion (Compared with the chrominance channel, this channel has

This article has been accepted for publication in IEEE Geoscience and Remote Sensing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LGRS.2022.3219341

more significant luminance changes, and contains major structure and details) [15]. After obtaining the single-channel fusion result, it is inversely transformed, and the fusion image can be converted into RGB space. Therefore, all issues can be categorized as a single-channel image fusion problem. The following details the parameters of the model training.

During the second training step, we set the batch size and epoch to 4 and 10, respectively. We encode two source images at four different scales, which is same as the nest's scale parameters [64, 112, 160, 208]. The probability of an element being zeros is set to 0.5, and the size of padding in all boundaries in each convolutional layer is 1. We modify the front and back ends of the DenseNet network so that it can adjust to the encoder output and decoder input in order to guarantee consistency of the model structure. That is, the DenseNet's input in this letter is [4, 1, 256, 256], and its output is also [4, 1, 256, 256]. Moreover, we use the Adam algorithm for stochastic optimization and set the learning rate to 0.0001. After multiple rounds of testing, we set both $\alpha 1$ and $\alpha 2$ to 0.015 as the weighting factors.

C. *Result analysis*

For comparison, we use ten sets of optical and SAR images covering different scenes. Metrics for fusion results of various methods are shown in Fig. 4. Table II provides their average values. It is evident that the IHS method performs the worst in terms of the four metrics (*i.e.*, EN, CC, SCD, and SSIM). The HPF method performs the worst on two metrics (*i.e.*, PSNR and MSE). However, HPF obtains the highest EN, which denotes the amount of information contained. This may be because the result of HPF includes much noise. For the proposed SOSTF，the SOSTF outperforms the other methods in five metrics (except EN) as described in Table II. Moreover, the SOSTF-based SSIM, the metric we care about most, is higher than others and is above 0.5, indicating that the fusion image derived from SOSTF contains a significant amount of structural information. The excellent performance of the metrics CC and MSE demonstrates that SOSTF preserves more source image features. In other words, SOSTF complies with the constraints we anticipate, which keeps major optical textures together with additional SAR structure information. Overall, our fusion method outperforms the other fusion approaches.

In the qualitative evaluation, Fig. 5 illustrates the fusion results of optical and SAR images obtained by the proposed method and the other fusion methods under the same conditions. To examine the fusion details of different methods in different scenarios, we selected three sets of images for comparison (*i.e.*, rows a, b, and c in Fig. 5). It can be seen from Fig. 5 that in the three scenes a, b, and c, the fusion results obtained by SOSTF have more realistic grayscale information than the fusion images obtained by IHS, HPF and U2Fusion. The fusion results of IHS lose part of the texture information of the optical image, such as the road in the yellow box in row a, the vehicle in the yellow box in row b, and the details in the blue box in row c. The results indicates that HPF and NestFuse lose part of the structure edge information of the SAR image, such as the structures in the blue box in row a, the edges of buildings in the yellow box in row b, and the road in the yellow box in row c. The large red box in the upper left corner of the fusion results in Fig. 5 is a partially enlarged view of the corresponding small red box. We can clearly see that IHS, HPF and NestFuse methods all lose some features of the original image, whereas the fusion result obtained by the SOSTF method suppresses the speckle noise better than the DWT method and the LP method.

To sum up, the fused image obtained by SOSTF more completely maintains the original image's complementary properties, strengthening the structural edge information from the SAR image while containing the majority of the texture information from the optical image. The grayscale of the fused image appears better and contains less noise.

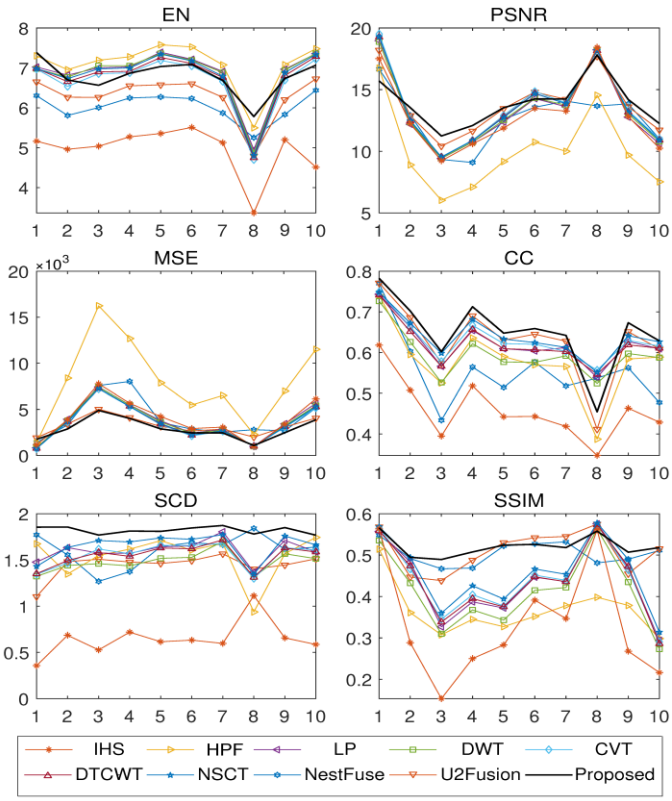

Fig. 4. Evaluation results of multiple methods

TABLE II

METRICS RESULTS FOR VARIOUS FUSION METHODS

Methods	EN	PSNR	MSE	CC	SCD	SSIM
IHS	4.9509	12.9683	3916.2859	0.4583	0.6489	0.3322
HPF	7.0961	10.0420	7933.4565	0.5782	1.5431	0.3660
LP	6.8657	13.3201	3674.8508	0.6227	1.6022	0.4349
DWT	6.8462	13.4016	3593.8972	0.5954	1.4801	0.4096
CVT	6.6738	13.6567	3417.7844	0.6304	1.5531	0.4378
DTCWT	6.7431	13.5700	3475.4061	0.6209	1.5460	0.4356
NSCT	6.8020	13.5699	3479.0105	0.6389	1.6472	0.4544
NestFuse	6.0265	12.6758	3993.8958	0.5556	1.6058	0.5040
U2Fusion	6.2889	13.8667	3226.6168	0.6296	1.4452	0.5104
Proposed	6.7899	13.8830	2874.5536	0.6504	1.8219	0.5209

This article has been accepted for publication in IEEE Geoscience and Remote Sensing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LGRS.2022.3219341

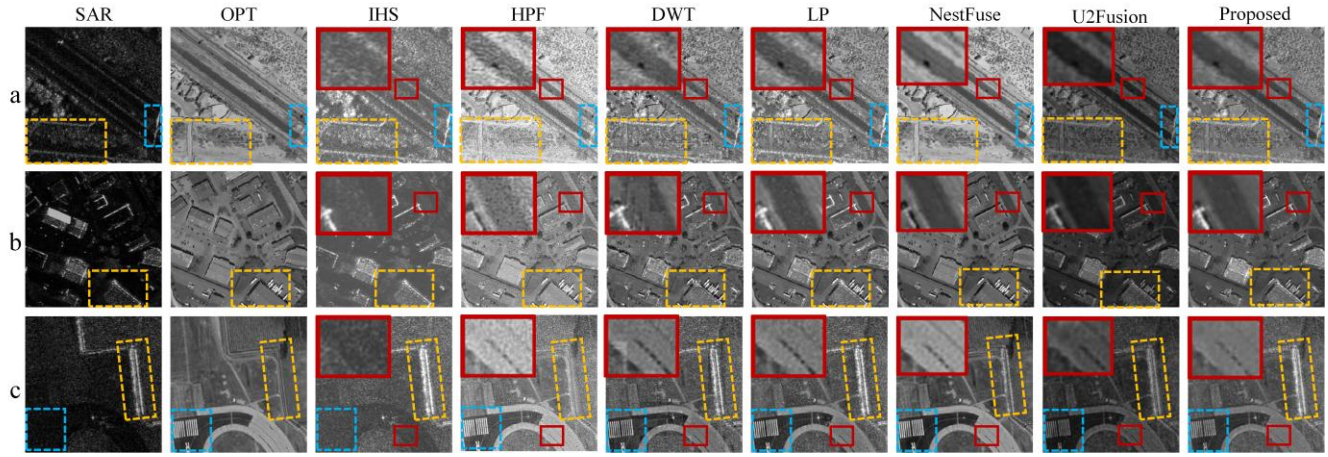

Fig. 5. Experimental results. The images sorted by column are SAR, Optical, HIS, HPF, DWT, LP, NestFuse, U2fusion, and proposed method.

IV. Conclusions

We propose an unsupervised end-to-end framework (named SOSTF) to fuse complementary features of optical and SAR images. The nest connection and the DenseNet network are combined to form the main fusion architecture in the proposed framework. The structure and texture features are extracted from the input images using the cartoon-texture method. A new loss function is then developed, and these features are gradually fused into the fusion results as constraints. The experimental results demonstrate that the SOSTF method can fuse optical and SAR images with high quality without relying on labels. Additionally, it qualitatively displays a better fusion effect in terms of the six metrics when compared to the other state-of-the-art fusion methods. The fusion results can also be applied to some future image analysis tasks, such as change detection, object detection, and image classification.

References

[1] Y. Ye, L. Shen, M. Hao, J. Wang, and Z. Xu, "Robust optical-to-SAR image matching based on shape properties," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 564-568, Apr. 2017.

[2] S. Hao, W. Wang, Y. Ye, T. Nie and L. Bruzzone, "Two-stream deep architecture for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2349-2361, Apr. 2018.

[3] Samadhan C. Kulkarni and Priti P. Rege, "Pixel level fusion techniques for SAR and optical images: A review," *Inf. Fusion*, vol. 59, pp. 13-29, Jul. 2020.

[4] C.M. Chen, G.F. Hepner, and R.R. Forster, "Fusion of hyperspectral and radar data using the IHS transformation to enhance urban surface features," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, no.1-2, pp. 19-30, Jun. 2003.

[5] S. K. Pal, T. J. Majumdar, and Amit K. Bhattacharya, "ERS-2 SAR and IRS-1C LISS III data fusion: A PCA approach to improve remote sensing based geological interpretation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 61, no.5, pp. 281-297, 2007.

[6] S. Chen, R. Zhang, H. Su, J. Tian, and J. Xia, "SAR and Multispectral Image Fusion Using Generalized IHS Transform Based on à Trous Wavelet and EMD Decompositions," *IEEE Sensors Journal*, vol. 10, no. 3, pp. 737-745, Mar. 2010.

[7] Y. Byun, J. Choi, and Y. Han, "An Area-Based Image Fusion Scheme for the Integration of SAR and Optical Satellite Imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 5, pp. 2212-2220, Oct. 2013.

[8] Z. Yin, "Fusion algorithm of optical images and SAR with SVT and sparse representation," *International Journal on Smart Sensing and Intelligent Systems*, vol. 8, no. 2, pp. 1123-1141, Jun. 2015.

[9] B. Yan and Y. Kong, "A Fusion Method of SAR Image and Optical Image Based on NSCT and Gram-Schmidt Transform," IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, pp. 2332-2335, 2020.

[10] Y. Kong, F. Hong, H. Leung, and X. Peng, "A Fusion Method of Optical Image and SAR Image Based on Dense-UGAN and Gram–Schmidt Transformation," *Remote Sensing*, vol. 13, no. 21, pp. 4274, Oct. 2021.

[11] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, "Cloud Removal with Fusion of High Resolution Optical and SAR Images Using Generative Adversarial Networks" *Remote Sensing*, vol. 12, no. 1, pp. 191, Jan. 2021.

[12] J. Wang, F. Chen, M. Zhang, and B. Yu, "ACFNet: A Feature Fusion Network for Glacial Lake Extraction Based on Optical and Synthetic Aperture Radar Images," *Remote Sensing*, vol. 13, no. 24, pp. 5091, Dec. 2021.

[13] W. Li, L. Liu, and J. Zhang, "Fusion of SAR and Optical Image for Sea Ice Extraction," *Journal of Ocean University of China*, vol. 20, pp. 1440-1450, Mar. 2021.

[14] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11-26, Aug. 2019.

[15] H. Li, X. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72-86, Sept. 2021.

[16] J. Zhang, R. Lai, and C. J. Kuo, "Adaptive Directional Total-Variation Model for Latent Fingerprint Segmentation," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 8, pp. 1261-1273, Aug. 2013.

[17] H. Li, X. Wu, and T. Durrani, "NestFuse: An Infrared and Visible Image Fusion Architecture Based on Nest Connection and Spatial/Channel Attention Models," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9645-9656, Dec. 2020.

[18] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A Unified Unsupervised Image Fusion Network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502-518, Jan. 2022.

[19] Y. Xiang, R. Tao, F. Wang, H. You, and B. Han, "Automatic Registration of Optical and SAR Images Via Improved Phase Congruency Model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5847-5861, Sept. 2020.

[20] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li and S. Hao, "A multiscale framework with unsupervised learning for remote sensing image registration", *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022.