Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-14393-2

1229: MULTIMEDIA DATA ANALYSIS FOR SMART CITY ENVIRONMENT SAFETY

SaReGAN: a salient regional generative adversarial network for visible and infrared image fusion

Mingliang Gao[1] · Yi'nan Zhou[2] · Wenzhe Zhai[1] · Shuai Zeng[3] · Qilei Li[4]

Received: 29 September 2022 / Revised: 17 November 2022 / Accepted: 6 January 2023

Abstract
Multispectral image fusion plays a crucial role in smart city environment safety. In the domain of visible and infrared image fusion, object vanishment after fusion is a key problem which restricts the fusion performance. To address this problem, a novel Salient Regional Generative Adversarial Network GAN (SaReGAN) is presented for infrared and VIS image fusion. The SaReGAN consists of three parts. In the first part, the salient regions of infrared image are extracted by visual saliency map and the information of these regions is preserved. In the second part, the VIS image, infrared image and salient information are merged thoroughly in the generator to gain a pre-fused image. In the third part, the discriminator attempts to differentiate the pre-fused image and VIS image, in order to learn details from VIS image based on the adversarial mechanism. Experimental results verify that the SaReGAN outperforms other state-of-the-art methods in quantitative and qualitative evaluations.

Keywords Smart city · Image fusion · Visible and infrared image · Generative adversarial network · Salient region

1 Introduction

Image fusion is a post-procedure that exploits information of multiple sensors and fuses their captured images to obtain the final enhanced results [7, 8, 12, 30]. Visible (VIS) and infrared (IR) image fusion plays a crucial rule in smart city environment safety. In the domain of VIR and IR image fusion, the information is usually relatively either complemented or collided.

✉ Yi'nan Zhou
littlemuji@163.com

[1] College of Electrical and Electronics Engineering, Shandong University of Technology, Zibo, Shandong, 255000, China

[2] Genesis AI Lab, Futong Technology, Chengdu, 610054, China

[3] Department of Obstetrics and Gynaecology, West China Second University Hospital, Sichuan University, Chengdu, Sichuan, China

[4] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, UK

Published online: 25 January 2023

The IR images are captured by infrared sensors, and they are characterized by different intensities of thermal information in all weather conditions [15]. Compared with IR, VIS images are captured by optical camera cameras and highly rely on the lightness condition. The fusion of IR and VIS images attempts to maintain more informative features.

In recent years, IR and VIS image fusion has been developed from conventional methods to learning-based methods [30]. Generally, steps of the conventional methods can be summarized as follows. Firstly, a certain filter/transform is applied to decompose source images [1, 14, 17]. Secondly, a corresponding fusion strategy is adopted to fuse features at different levels [4]. Lastly, the images are reconstructed, and the fused results are generated. A critical issue of IR and VIS image fusion is the design of the fusion rules [11]. For example, the average rule [29] regards the information of IR and VIS as the same important. It is widely used to fuse the detail and texture. However, this rule is sensitive to the intensity of pixels and fails to deal with objects region [24]. For detective tasks, the object is of great importance, and it is salient in the IR image. The defects of these kinds of methods are two-folds. Firstly, the filter/transform and fusion rules must be predefined and closely related to the prior knowledge of the designer [13]. Secondly, these methods are helpless to handle the salient regions in IR image individually. Thus, methods to extract salient regions of IR image and preserve such crucial information are urgently needed [27].

Subsequently, the learning-based fusion methods are proposed and have become the mainstream. The weights of convolution kernels and the fusion strategies in these methods are learned by a large amount of data instead of manually designed. Among these learning-based fusion methods, Generative Adversarial Network (GAN) is proven to be effective and efficient for image fusion [6, 32]. Ma et al. [16] proposed the FusionGAN for IR and VIS image fusion. In the process of training, it firstly concatenates the IR and VIS images and feeds them into a generator to generate a pre-fused image. Then, the pre-fused image is restricted to the texture and details of the VIS image under the supervision of the discriminator. As the discriminator regards the images are the same under the constraint of loss function, then it outputs the fused result. Therefore, the function of discriminator can be considered as a certain fusion rule. The fused result includes VIS information twice, once in the generator and once in the discriminator. While the IR information is included once in the generator. Thus, the fused result counts more on VIS while the weight of IR information is reduced. In other words, the FusionGAN excels in preserving texture and detail of VIS image, while neglecting to preserve the salient regions in IR image.

Taking all the aforementioned factors into consideration, a novel salient regional GAN (SaReGAN) is put forward to preserve essential salient regions in IR image during fusion and generate an overwhelming fused result. Detailed comparison between FusionGAN and the proposed SaReGAN will be provided in Section 4.2. The main contributions of the paper are three folds.

1. To obtain a fused result where the salient objects have clear edges and high intensities, the salient regions are extracted in the IR image. The salient regions contain the critical information of the target which ought to be demonstrated in the fused result. This idea will enable image fusion to make full use of essential information captured by different sensors.
2. To deal with the salient information of IR image, the visual saliency map (VSM) is introduced to extract crucial salient regions. We transfer the use of region extraction method from the VIS image to IR image. The VSM technique is efficient and effective to address the difficulties in maintaining the whole salient regions compared with other learning-based techniques like U-Net.

3. To integrate the novelties above, the SaReGAN is built in thie work. The SaReGAN is capable of dealing salient information in the generator, from which a better pre-fused image with salient information can be obtained. To the best of our knowledge, it is the first time that the fusion method is able to extract salient regions and fuse them simultaneously.

The rest of the paper is structured as follows. In Section 2, the necessity of salient region extraction is presented. Meanwhile, the comparison of the visual saliency map (VSM) and U-Net in extracting crucial salient regions is elaborated. In Section 3, the SaReGAN is introduced in detail. In Section 4, the experimental data of visual perception and objective measurements are provided. The conclusion is drawn in Section 5.

2 Salient region extraction in IR image

2.1 Effect of Salient regions in IR image

The salient regions have high intensities and intense contrast compared with the background. This is based on the fact that human cortical cells are sensitively responded to high contrast stimulus in their receptive fields [21]. The VIS images reflect much detail and texture, which contributes a lot to detective tasks. The salient regions are important in IR image since they reflect the location and the shape of objects. On the contrast, the texture and the details are crucial in VIS image. The texture and the details depict the environment where the objects are. This kind of visible information helps in decision-making. In our method, texture and the details of visible information are learned in GAN, while the salient information of IR information is obtained from salient region extraction. In this situation, extracting the salient objects in the IR image and preserving them in the process of fusion will benefit the fused result to a large extent. In those salient regions, the IR image is far more informative and crucial than the VIS image. In this scenario, the proposed method can deal with such regions separately (as shown in Fig. 1). In our method, texture and the details of

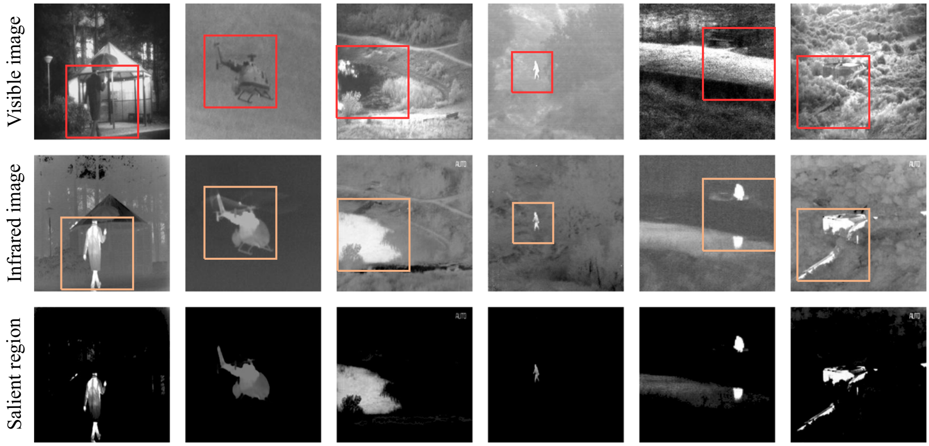

Fig. 1 Different presentation effects of salient regions in IR and VIS images. The salient regions in the IR image are ought to be preserved in purpose. Regions with high intensity in orange blocks refer to the salient regions

visible information are learned in GAN, while the salient information of infrared information is obtained from salient region extraction. Thus, it is of great necessity to utilize these different kinds of source images and generate an informative fused image for easier decision-making.

Cheng et al. [2] proposed to generate the visual saliency map (VSM) using histogram-based contrast. It calculates the global contrast differences and spatial coherence of VIS image and obtains a pretty good saliency result. To a large extent, the saliency result of VIS image and the saliency result of IR image share a lot. This can be attributed to the fact that the salient regions of two kinds of source images have high intensities in a certain channe. Meanwhile, they have high contrast to the background. However, there is a difference between VIS and IR. In original VSM dealing with VIS image, they use smooth and average operations to refine the saliency result.In our view, these operations are not suitable to deal with salient regions in IR image, because these operations do great damage to edges and weaken the intensities. We further remove these operations and make the uncertain pixels only sensitive to their original intensities. Thus, it is reasonable and applicable to employ the VSM to extract the salient regions of IR images. The extracted salient regions of IR images are shown in Fig. 1.

2.2 Superiority of VSM over U-Net in salient region extraction

VSM [2, 31] is a technique to generate an image where the visual interests of humans can be preserved and enhanced. It is not a learning-based technique in essence. U-Net [19, 23] is a learning-based technique that excels in classification and detection.The IR image dataset with labels is obtained from the Pascal-VOC dataset. There are 5829 VIS images and their labels used for training U- Net. The training results are shown in Fig. 2. The superiority of VSM over U-Net in salient region extraction can be concluded as follows.

1. The VSM is more efficient than U-Net. The running time of VSM is shorter than U-Net, because the parameters of VSM are smaller than U-Net. This is a common advantage of conventional techniques compared with the learning-based techniques.

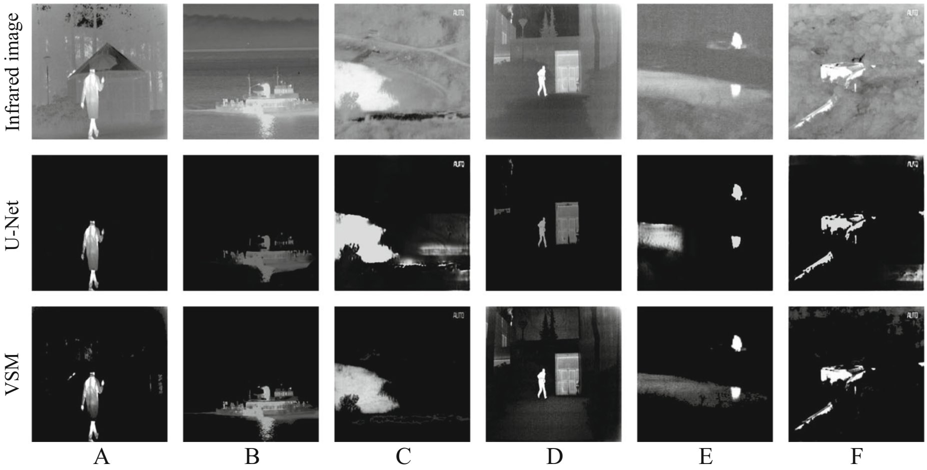

Fig. 2 Results of salient region extraction. The first, second and third row represents the original IR images, salient regions extracted by U-Net and salient regions extracted by VSM

2. The VSM is more effective than U-Net. VSM is sensitive to each pixel, and it is able to maintain the whole structure and the edges of salient regions. By contrast, U-Net is sensitive to objects, where the edges are not satisfying. Take the column B in Fig. 2 for instance, the U-Net fails to handle the sophisticated edges of the warship, while VSM maintains the crucial thermal information of salient regions. Moreover, in column F, the VSM extracted structural information of the plane is preciser while the holes in the U-Net extracted plane are larger than the source image. The structure and edges of extracted salient regions are of vital importance during the process of fusion and have an intense impact on the fused results.
3. The U-Net needs to be trained using the dataset with labels, while VSM does not rely on massive data nor training process. There are plenty of kinds of VIS image datasets with labels, but IR image dataset with labels is relatively rare. Thus, the training process of U-Net tends to be more complex. However, VSM is not a learning-based technique and does not need to be trained, which offers many conveniences to the whole process of fusion.

3 Proposed method

3.1 Pipeline of SaReGAN

The pipeline of SaReGAN is shown in Fig. 3. The SaReGAN consists of three parts, namely generator, discriminator and VSM processor. In SaReGAN, the IR and VIS image fusion can be regarded as an adversarial game. The VSM outputs the salient information image, which makes the SaReGAN capable of dealing to salient regions separately. The generator outputs a pre-fused image according to the source images and satisfies the discriminator based on given criterion, while the discriminator tries to differentiate whether the given image

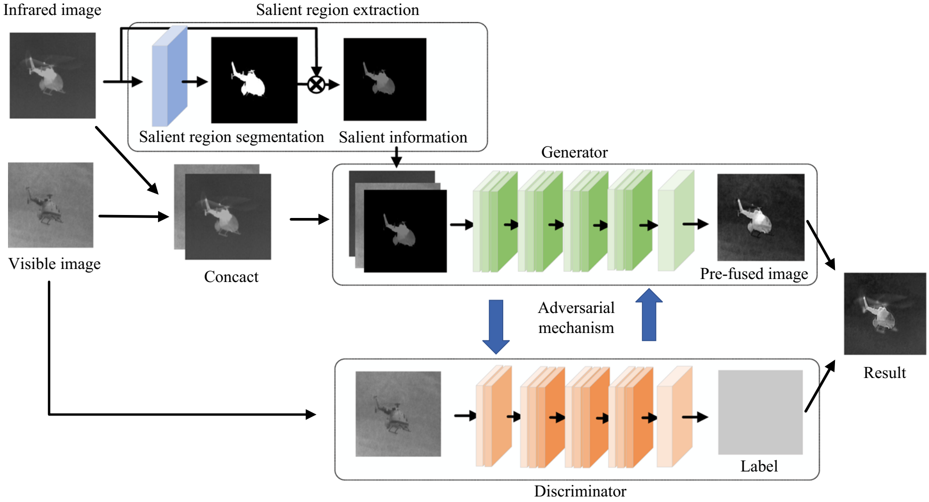

Fig. 3 Architecture of the SaReGAN. Each arrow controls the direction of data. Each block represents a certain process with a different function

is a pre-fused image or a VIS image. Moreover, the generator gathers the entire information of source images and the discriminator assists with fusing more details by adversarial processor.

3.2 Network design

As shown in Fig. 3, the IR image is initially converted into a saliency map by evaluating global contrast differences and pixel-level weighted continuity scores [2]. Then, saliency map is used to extract the salient information of IR image based on its intensity accordingly. Lastly, salient information is fed to the generator to maintain the salient regions in the fused image.

Specifically, the VSM is adopted to extract the salient information of IR image. The saliency and the saliency of an IR image share a lot in common. This kind of saliency or salient regions refers to the same objects which have high intensities and attract the perception of human eyes. Thus, it is applicable to use VSM to extract the salient regions in IR image and utilize the salient regions to improve the quality of fused results.

For the generator, it generates the pre-fused result based on the IR, VIS and salient information. This process can be regarded as a process of fusion, as it combines the multi-sources images and generates a rudimentary fused image. The pre-fuesed result is not satisfying and needs to be decorated in discriminator. The generator of SaReGAN consists of five groups of bunching blocks. In the first and second groups of bunching blocks, a 5 × 5 convolutional layer, BatchNorm operation and a Relu function are applied. In the third and fourth groups of bunching blocks, a 3 × 3 convolutional layer, BatchNorm operation and a Relu function are utilized. In the last group of bunching blocks, a 1 × 1 convolutional layer and a tanh activation layer are assembled to output a pre-fused image. The pre-fused image is a simple fused image, which is a basic information integration. In the discriminator, the pre-fused image is compared with VIS image. If the output label of the discriminator is false, more visible information will be added to the pre-fused image. If the output label is true, then the GAN process ends and outputs the decorated pre-fused image as the final fused result.

For discriminator, it outputs the predicted label and computes the loss between pre-fused image and VIS image. This process can be regarded as a process of fusion, as it tries to minimize the difference between the pre-fused image and the VIS image. The least square loss function is employed to restrict pre-fused image and VIS image, thus abundant visible information is fused to the final result. In other words, the discriminator is set to distinguish the difference between pre-fused image and VIS image, until the difference is ignored.

The discriminator of SaReGAN consists of five groups of bunching blocks. In the first group of bunching blocks, a 3 × 3 convolutional layer and a Relu function are employed. From the second to fourth groups of bunching blocks, a 3 × 3 Convolutional layer, BatchNorm operation and a Relu function are assembled. In the last group of blocks, a linear layer is adopted to output the predicted label.

In the learning-based networks, down-sampling and up-sampling operations are widely used to facilitate the training process. However, in the field of image fusion, these operations are not satisfying as they discard crucial information of source images [28]. Thus, in this work, the stride is set to 1 and no down-sampling or up-sampling operation is applied. Furthermore, to avoid the critical issue of vanishing gradient, the BatchNorm operations are employed in generator and discriminator.

3.3 Lose function

The loss function of SaReGAN consists of two parts, namely $L_{\text{generator}}$ and $L_{\text{discriminator}}$. The $L_{\text{generator}}$ represents the loss function of the generator and is calculated by (1):

$$L_{\text{generator}} = \frac{1}{HW}\left(\left\|(I_{salient}, I_{fused})_{\max} - I_{IR}\right\|_F^2 + \gamma \left\|\nabla I_{fused} - \nabla(I_{salient}, I_{VIS})_{\max}\right\|_F^2\right) + L_{\text{GAN}}, \tag{1}$$

where H and W denote the width and height of the image, $\|\cdot\|_F$ represents the matrix Frobenius norm, and ∇ denotes the gradient calculating operator. $(\cdot)_{\max}$ outputs the maximum value, and salient information can be preserved as its value is higher. γ represents the coefficient that balances the whole equation. $I_{salient}$, I_{IR}, I_{VIS} and I_{fused} represent salient information image, IR image, VIS image and pre-fused image, respectively. L_{GAN} denotes the adversarial loss of generator and discriminator, and it restricts the direction how the generator is supposed to generate a pre-fused image.

The $L_{\text{discriminator}}$ represents the loss function of the discriminator. It is formulated as:

$$L_{\text{discriminator}} = \frac{1}{N}\sum_{n=1}^{N}(D(I_{VIS}) - b)^2 + \frac{1}{N}\sum_{n=1}^{N}\left(D\left(I_{fused}\right) - a\right)^2, \tag{2}$$

where $D(\cdot)$ denotes the output of the discriminator. N represents the amount of images. a and b represent the labels of pre-fused image I_{fused} and VIS image I_{VIS}. The discriminator is set to differentiate the pre-fused image and VIS image. In an ideal condition, the output label of the discriminator ought to be the same, meaning that pre-fused image and VIS image share a lot of features in common.

4 Experiments

4.1 Details

To demonstrate the efficiency and effectiveness of the proposed SaReGAN, twenty pairs of IR and VIS images are employed in the experiment. The image pairs are previously registered and up-sampled to the same resolution. These image pairs are from TNO dataset [25] and are widely used in image fusion. The resolutions of raw IR images and raw VIS images are different. IR images are in low resolutions while the VIS images are in high resolutions, due to the attributes of the different sensors. To this aim, the IR images are up-sampled. Subsequently, we resize both IR and VIS image to the scale of 512 × 512 as refined image pairs. Lastly, we set the stride to 14 for each image to crop enough data. By this mean, we can generate plenty of IR and visible patches which are adequate for training the model.

4.2 Superiority of SaReGAN over FusionGAN

The proposed SaReGAN is more competitive than FusionGAN as it preserves essential IR information. The FusionGAN can not deal with such information separately. The FusionGAN also uses the pre-fusion mechanism, and it tried to strike a balance of visible and IR information. However, the infrared information more informative than visible information for salient objects. the salient region is extracted to preserve such informative information.

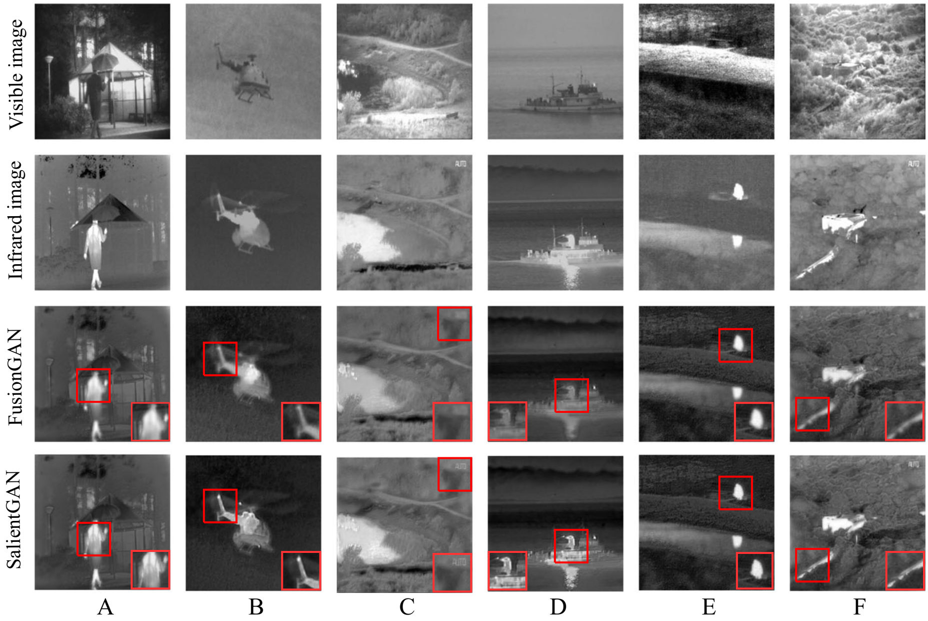

Fig. 4 Comparative results of FusionGAN and SaReGAN

In FusionGAN, visible information is learned twice and infrared information only contributes once. Thus, an issue is that after each iteration, the information of IR image is restrained because visible information weighs more than infrared information. As a result, the salient information in FusionGAN fused images is dimmed and blurred, as shown in Fig. 4.

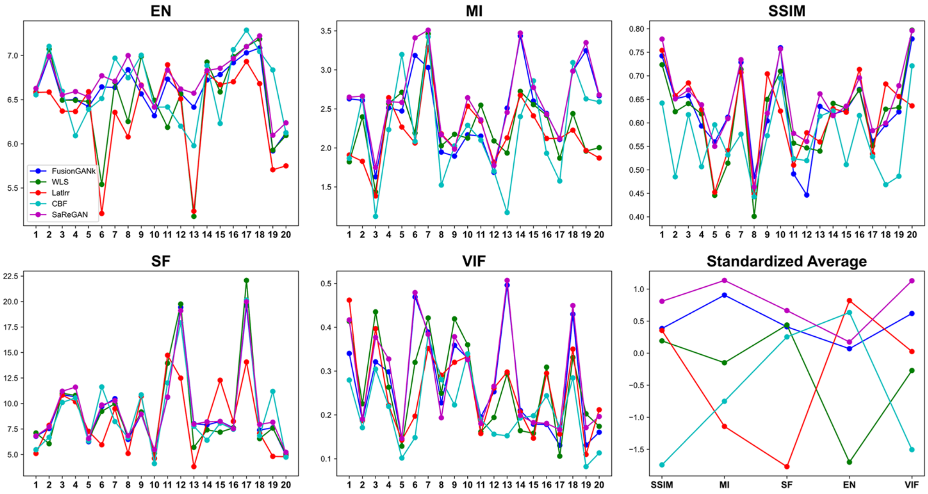

Fig. 5 Objective comparisons on 20 pairs of images. The sixth sub-figure is the standardized average of the five evaluation metrics

We settle this issue by introducing the process of VSM [2], where salient information takes part in fusing and can be dealt with separately in the mechanism of SaReGAN. In SaReGAN, the IR image and VIS image are both fused twice, and salient regions are successfully reserved in the fused image. Moreover, we employ a more suitable loss function to deal with information in salient regions. From the fused results in Fig. 4, one can see that the SaReGAN retains more infrared image features. For example, the proposed method has more texture on coat in column A and sharper edges on helicopter in column B.

4.3 Objective evaluation

Five evaluation metrics, i.e., structural similarity, mutual information, spatial frequency, entropy and visual information fidelity, are adopted to measure the objective results based on 20 pairs of images from TNO Dataset [25]. The results are shown in Fig. 5 and the average values of these evaluation indicators are shown in Table 1.

The structural similarity ($SSIM$) metric [26] is an indicator to evaluate the structural similarity among fused images and source images. The value of $SSIM$ varies from -1 to 1. A larger SSIM value indicates a more similar structure between the fused image and the source images. When the value equals to 1, the structure of fused image is the same as the structures of source images. From Fig. 5 and Table 1, one can see that the FusionGAN, WLS and Latlrr have high values which prove that these methods are better at preserving structures of source images. The performance of CBF is unsatisfying as it introduces artifacts which makes the structure of fused image difficult to distinguish. The CBF takes intensity likeness and spatial similarities of the surrounding pixels into considerations to fuse the detailed images. The fusion results are not satisfying because this method is not able to deal with conflict regions and causes lots of artifacts. The proposed SaReGAN scores the highest values, which proves that the proposed method can successfully maintain the structures of source images in detail.

Mutual information metric (MI) [20] measures how much the information of the source images is contained in the fused image. A higher MI value represents a more informative preservation. It can be observed from Table 1 and Fig. 5 that the FusionGAN and WLS have high values, proving that these methods succeed in preserving information of source images. The Latlrr and CBF have lower values as they introduce artifacts which are collided with source images. The proposed SaReGAN reaches the highest values, which illustrates that our method successfully maintains much information and the improvements are strictly based on source images.

Spatial frequency (SF) [3] is a metric that measures the statistical distribution of gradient in fused images. The larger the value of SF is, the more abundant edges and textures are

Table 1 Objective comparison results

Metrics	FusionGAN [16]	WLS [18]	Latlrr [10]	CBF [9]	SaReGAN
$SSIM$	0.6221	0.6168	0.6213	0.5639	**0.6337**
MI	2.4757	2.3282	2.2189	2.2441	**2.5078**
SF	9.2219	**9.2332**	8.3835	9.1616	9.2319
EN	6.7044	6.5170	**6.7255**	6.7179	6.7143
VIF	0.2715	0.2465	0.2548	0.2118	**0.2859**

The best results are highlighted in bold reserved in the fused image. The results indicate that, except for the Latlrr, all the methods have high values on this indicator. The Latlrr can not deal with salient information separately and hard to extract the edges and texture of IR image.

Entropy (EN) [22] is a metric that calculates the scale of uncertainty. Except for results of WLS, the values of all the methods are nearly the same. It proves that introducing the operation of salient region extraction barely changes the value of entropy.

Visual information fidelity (VIF) [5] is a metric based on natural scene statistics theory, which calculates the distortion information in the image fusion. Higher values of VIF represent better visual perception of fused image. Comparative results prove that the proposed

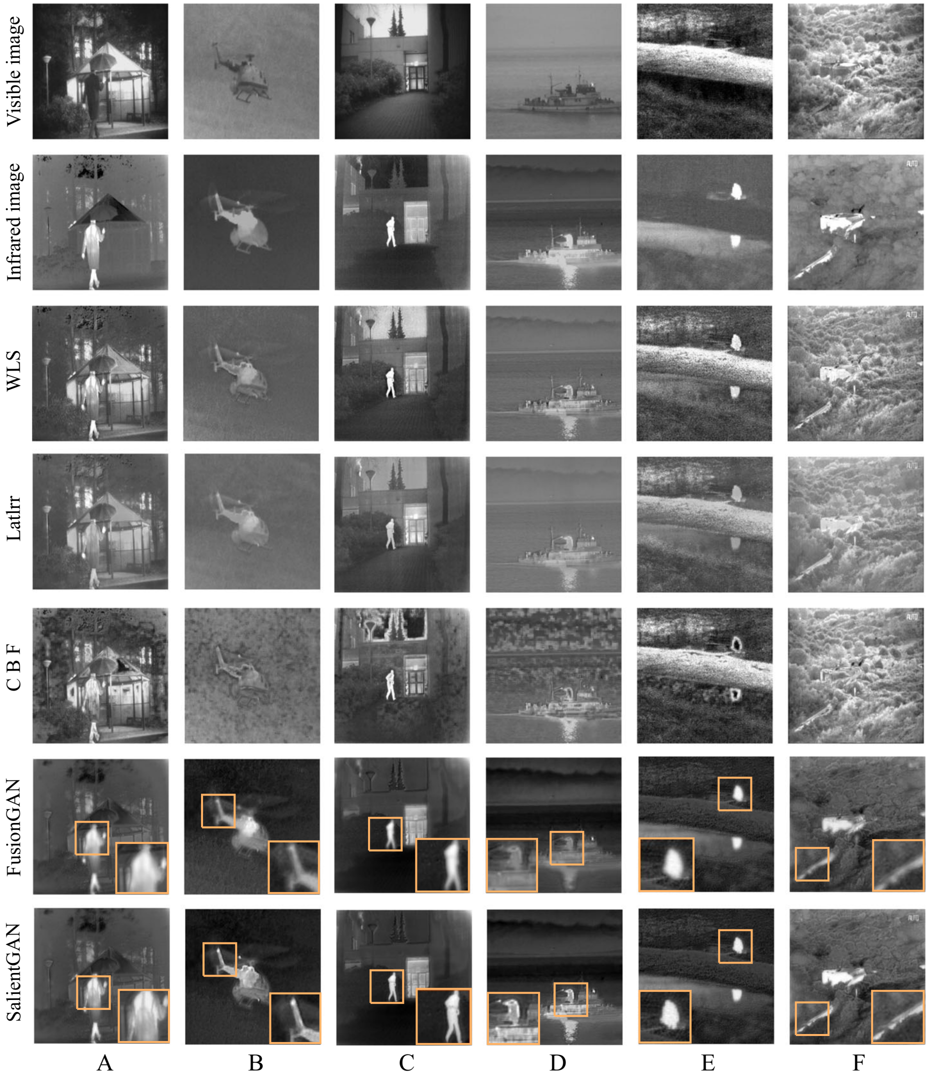

Fig. 6 Subjective comparison results on six representative IR and VIS image pairs from the TNO database [25]

SaReGAN reaches the highest values of VIF, which proves the superiority of SaReGAN in visual perception.

4.4 Subjective evaluation

Six typical image pairs are selected for subjective evaluation, and the comparative results are depicted in Fig. 6. The front two rows represent the source images captured by infrared sensors and visible cameras. From the third row to the sixth row are the fusion results of WLS, Lalrr, CBF and proposed SaReGAN, respectively.

It shows that WLS is capable of dealing with complementary regions where infrared and VIS images have similarities. However, in those conflict regions (where IR and VIS images have sheer contrast) of source images, WLS causes artifacts due to its simplicity (e.g., the artifacts above umbrella in A column of Fig. 6). In the conflict regions, the most common way is to sample information in both source images, regardless the specific attributes of each kind of source images. Compared with WLS, the SaReGAN is capable of dealing with these conflict regions and not create artefacts as it can protect the salient information of IR image and reduced the redundant visible information.

The Latlrr [10] decomposes images into low-rank parts and salient parts. However, from the fusion results(e.g., the man who sits beside the river in column E and the plane in column F in Fig. 6), the Latlrr fails to keep the salient parts as obvious as they are in the IR image. Compared with the Latlrr, the SaReGAN is capable to extract and preserve those salient parts. Thus, a more salient fusion result can be generated by the proposed SaReGAN.

The CBF [9] takes intensity likeness and spatial similarities of the surrounding pixels into considerations to fuse the detailed images. The fusion result is not satisfying because this method is not able to deal with conflict regions and causes lots of artifacts. It is obvious that the SaReGAN is superior to CBF to a large extent.

5 Conclusion

In this work, a novel SaReGAN is built to solve the problem of object vanishment in the visible and IR image fusion. The salient region extraction operation can maintain the most significant information of IR image. Moreover, in the experiment, the proposed VSM processor is testified superior to learning-based method (U-Net), since VSM is faster and its results are more natural. Comparative experiments are carried out between the SaReGAN and the SOTA competitors, and the objective and subjective results demonstrate the superiority of the SalienGAN in mechanism design and effectiveness. We consider that the idea of introducing salient region extraction to image fusion and the design of SaReGAN will provide some references in other domain of computer vision.

Funding This work is supported in part by the National Natural Science Foundation of Shandong Province (Nos. ZR2021QD041 and ZR2020MF127).

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Ethics approval We declare that there is no ethics issue.

Conflict of Interests We declare that we have no conflict of interest.

References

1. Chen J, Zhang L, Lu L, Li Q, Hu M, Yang X (2021) A novel medical image fusion method based on rolling guidance filtering. Internet Things 14:100172
2. Cheng M-M, Mitra NJ, Huang X, Torr PH, Hu S-M (2014) Global contrast based salient region detection. IEEE Trans Pattern Anal Mach Intell 37(3):569–582
3. Eskicioglu AM, Fisher PS (1995) Image quality measures and their performance. IEEE Trans Commun 43(12):2959–2965
4. Gao M, Jiang J, Zou G, John V, Liu Z (2019) Rgb-d-based object recognition using multimodal convolutional neural networks: a survey. IEEE Access 7:43110–43136
5. Han Y, Cai Y, Cao Y, Xu X (2013) A new image fusion performance metric based on visual information fidelity. Inf Fusion 14(2):127–135
6. Hermessi H, Mourali O, Zagrouba E (2021) Multimodal medical image fusion review: theoretical background and recent advances. Signal Process 183:108036
7. Hermessi H, Mourali O, Zagrouba E (2021) Multimodal medical image fusion review: theoretical background and recent advances. Signal Process 183:108036
8. Kaur H, Koundal D, Kadyan V (2021) Image fusion techniques: a survey. Arch Comput Meth Eng 28(7):4425–4447
9. Kumar BS (2015) Image fusion based on pixel significance using cross bilateral filter. SIViP 9(5):1193–1204
10. Li H, Wu X-J (2018) Infrared and visible image fusion using latent low-rank representation, arXiv:1804.08992
11. Li Q, Wu W, Lu L, Li Z, Ahmad A, Jeon G (2020) Infrared and visible images fusion by using sparse representation and guided filter. J Intell Transp Syst 24(3):254–263
12. Li B, Xian Y, Zhang D, Su J, Hu X, Guo W (2021) Multi-sensor image fusion: a survey of the state of the art. J Comput Commun 9(6):73–108
13. Li Q, Yang X, Wu W, Liu K, Jeon G (2021) Pansharpening multispectral remote-sensing images with guided filter for monitoring impact of human behavior on environment. Concurrency and Computation: Practice and Experience
14. Liu F, Chen L, Lu L, Ahmad A, Jeon G, Yang X (2020) Medical image fusion method by using laplacian pyramid and convolutional sparse representation. Concurrency and Computation: Practice and Experience
15. Liu S, Gao M, John V, Liu Z, Blasch E (2020) Deep learning thermal image translation for night vision perception. ACM Trans Intell Syst Technol (TIST) 12(1):1–18
16. Ma J, Yu W, Liang P, Li C, Jiang J (2018) Fusiongan: a generative adversarial network for infrared and visible image fusion. Information Fusion
17. Ma J, Zhou Y (2020) Infrared and visible image fusion via gradientlet filter. Comput Vis Image Underst 103016
18. 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 Phys Technol 82:8–17
19. Pan T, Jiang J, Yao J, Wang B, Tan B (2020) A novel multi-focus image fusion network with u-shape structure. Sensors 20(14):3901
20. Qu G, Zhang D, Yan P (2002) Information measure for performance of image fusion. Electron Lett 38(7):313–315
21. Reynolds JH, Desimone R (2003) Interacting roles of attention and visual salience in v4. Neuron 37(5):853–863
22. Roberts JW, Van Aardt JA, Ahmed FB (2008) Assessment of image fusion procedures using entropy, image quality, and multispectral classification. J Appl Remote Sens 2(1):023522
23. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 234–241
24. Tian J, Chen L (2012) Adaptive multi-focus image fusion using a wavelet-based statistical sharpness measure. Signal Process 92(9):2137–2146
25. Toet A (2014) TNO image fusion Dataset. https://doi.org/10.6084/m9.figshare.1008029.v1, https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029. Accessed 04 June 2021
26. Yang C, Zhang J-Q, Wang X-R, Liu X (2008) A novel similarity based quality metric for image fusion. Inf Fusion 9(2):156–160
27. Yu N, Qiu T, Bi F, Wang A (2011) Image features extraction and fusion based on joint sparse representation. IEEE J Sel Top Signal Process 5(5):1074–1082

28. Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions. arXiv: Computer Vision and Pattern Recognition
29. Zhang Q, Guo B (2009) Multifocus image fusion using the nonsubsampled contourlet transform. Signal Process 89(7):1334–1346
30. Zhang H, Xu H, Tian X, Jiang J, Ma J (2021) Image fusion meets deep learning: a survey and perspective. Inf Fusion 76:323–336
31. Zhao J, Feng H, Xu Z, Li Q, Liu T (2013) Detail enhanced multi-source fusion using visual weight map extraction based on multi scale edge preserving decomposition. Opt Commun 287:45–52
32. Zhou T, Li Q, Lu H, Cheng Q, Zhang X (2022) Gan review: models and medical image fusion applications. Information Fusion

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.