Explainable analysis of infrared and visible light image fusion based on deep learning

Abstract

Explainability is a very active area of research in machine learning and image processing. This paper aims to investigate the explainability of visible light and infrared image fusion technology in order to enhance the credibility of model understanding and application. Firstly, a multimodal image fusion model was proposed based on the advantages of convolutional neural networks (CNN) for local context extraction and Transformer global attention mechanism. Secondly, to enhance the explainability of the model, the Delta Debugging Fuse Image (DDFImage) algorithm was employed for generating local explanatory information. Finally, we gain deeper insights into the internal workings of the model through feature importance analysis of the generated explanatory fusion images. Comparative analysis with other explainability algorithms demonstrates the superior performance of our algorithm. This comprehensive approach not only improves the explainability of the model but also provides more reference for practical application of the model.

Introduction

Currently, machine learning technology has been extensively used in various real-world fields, including image processing[1,2], natural language processing[3,4], autonomous driving[5], and malware detection[6]. These applications demonstrate the high accuracy of machine learning models. Although machine learning models have achieved significant successes, their internal mechanisms and behavioural characteristics can be challenging for people to understand due to their black-box nature and opaque learning processes. Recent studies have shown that machine learning models may exhibit unexpected behaviours[7,8]. However, when inputting data into a machine learning model and receiving output, it can be challenging for individuals to discern the correlation between the input and output. This can result in an inability to confirm the reliability of the output results provided by the model, which can directly impact the promotion and real-world application of machine learning models. As a result, various explainability methods have been proposed[4,8]. These explanations can be either global or local[9,10]. Global explanations aim to clarify the overall decision-making process of the model, while local explanations focus on explaining the predictions for specific inputs to the model. Local explanations typically use information about the target input, such as feature values, to aid in understanding the relationship between model inputs and outputs, i.e. predictions. Thus, assessing the reliability of individual model predictions is crucial for their practical application[5,11].

Infrared–visible light image fusion, as an important subclass of multimodal image fusion, is of utmost significance in computer vision. By fully extracting and integrating the information of the same scene collected by different sensors, infrared–visible light image fusion can maximize the exploitation of the most valuable and meaningful information in single-modal images while eliminating redundant information, generating fused images of higher quality[12]. In recent years, many methods have been developed for multimodal image fusion research, such as convolutional neural networks (CNNs), generative adversarial networks (GANs), VGG networks, and other neural network models. However, due to the opacity of their algorithms in multimodal image fusion, our understanding and application of these network models are limited.

The Transformer model has recently shown excellent performance in computer vision. Its network model demonstrates good generalization capabilities, and its self-attention mechanism and global feature extraction enhance fusion results. In this paper, we propose a network model that combines CNNs with Transformers to accomplish multimodal image fusion tasks and enhance their explainability using our algorithm.

Essentially, a machine learning model represents a complex nonlinear function, where image fusion models take raw images as input, perform complex calculations through multiple layers, and ultimately output images containing important features of both modalities as fusion results. However, the opaque learning process of machine learning models impedes people's understanding of their internal workings[11].

The article describes a method for multimodal image fusion that combines the advantages of CNNs and Transformers. The article avoids biased language and employs a formal register. The text is grammatically correct and adheres to conventional academic structure and formatting. The model is designed to be explainable, and the authors introduce an algorithm called DDFImage for generating local explanations. Finally, we conducted a feature importance analysis on the generated explanatory fusion images. This approach helps us to better understand the internal workings of the model and provides more references for the practical application of the model.

The major contributions of this study are summarized as follows:

1. 1.

A multimodal image fusion model integrating CNN and Transformer is proposed, enhancing the capability of local and global feature extraction.

2. 2.

The DDFImage algorithm is introduced to generate local explanatory information, significantly improving the model's explainability.

3. 3.

Feature importance analysis reveals the internal workings of the fusion model, providing deeper insights into model interpretation.

4. 4.

Comparative analysis with other explainability algorithms demonstrates the superior performance and interpretability of our approach, offering valuable references for practical applications.

To provide a comprehensive overview, the structure of this paper is outlined as follows: "Introduction" section introduces the study's background and motivation. "Related work" section reviews relevant work in the field of image fusion and explainability. "Methods" section outlines the proposed methods, including the multimodal fusion model and the DDFImage algorithm. "Results" section presents the results of the experiments, and "Discussion" section discusses the comparative performance evaluation with other explainability algorithms. Finally, "Conclusion" section concludes the paper, highlighting the contributions and suggesting directions for future work.

Related work

This section provides a brief overview of relevant work on image fusion using deep learning and explainability analysis.

Multimodal image fusion based on deep learning

Mainstream fusion algorithms based on deep learning can be classified into four types: autoencoders (AEs), convolutional neural networks (CNNs), generative adversarial networks (GANs), and Transformers. Li et al.13 proposed the DenseFuse fusion architecture based on autoencoder (AE) for feature extraction. They used traditional addition and L1-norm strategies for fusion rules in the feature extraction stage. However, this resulted in limited fusion performance and independence of fusion rules from AE, making it unable to achieve adaptive linkage. Zhang et al.14 utilised channel-wise attention to model cross-modal interactions and weight each feature map extracted from different spectra during the multispectral feature fusion stage. This approach achieved fully adaptive fusion of visible light image features and infrared image thermal features. Liu et al.15 were the first to apply CNNs to feature fusion and proposed a medical image fusion algorithm. However, feature extraction and reconstruction still rely on manually designed traditional fusion rules, resulting in limited performance and poor adaptive capability. To address these difficulties, Xu et al. proposed an end-to-end EMFusion16. This method establishes consistency measurement loss between fusion images and source images, which solves the problems of measuring activity levels and weight allocation in fusion problems. However, the network structure is relatively simple and does not fully utilise deep feature information. In 2019, Ma et al. proposed the FusionGan algorithm17, which was the first to apply GAN networks to image fusion. This algorithm offers a more concise and efficient end-to-end fusion approach. However, it only establishes an adversarial game between fusion images and visible light images in the discriminator, which can lead to imbalances in information extraction. To achieve information balance, Ma and his team proposed an algorithm in 2020 that uses multiple classifiers to estimate both distributions18, balancing fusion while improving the common instability issue in GAN network training.

Multimodal image fusion based on visual transformer models

The fusion algorithms based on Transformers have attracted the attention of many researchers in the past two years. Vaswani et al.19 first proposed the Transformer model in 2017, initially aimed at addressing two problems in the field of NLP: the inherent sequence order requirement of recurrent neural networks (RNNs) and the lack of global information understanding ability in sequences. Vs et al.[20] first applied it to the field of image fusion, with the core of the algorithm being the Spatio-Transformer fusion module, effectively compensating for the insufficient global feature extraction capability of CNN models by learning local information and long-distance information at multiple scales. SwinFuse proposed by Wang et al.[21] was the first to use pure Transformers as backbone networks for feature extraction without relying on CNNs. It not only achieved surprising fusion performance but also demonstrated strong generalization capabilities and computational efficiency.

Local explainability methods

Local explainability methods mainly explain the prediction results of the model based on single inputs and outputs of the model. In local explainability methods for image fusion, a pair of images is taken as input, and through fusion of the images, the influence of each part of the input image on the model's prediction fusion result is displayed, indicating which parts of the image have a greater impact on the fusion result.

Sufficiency sufficiency means that retaining only the features in the explanation allows the model to obtain the same fusion result as the original input.

Necessity necessity means that when removing the features in the explanation from the original image and inputting it into the model for fusion, a fusion result different from the original input will be obtained.

A very popular explainability method LIME[9] uses image segmentation techniques to segment the image into several regions and then calculates the contribution of each region to the classification result, while the explainability method RISE[22] segments the image into several regions and randomly obscures each region to calculate the change in the classification result, thereby obtaining the importance weight of each region and visually representing the contribution of each region to the image classification.

Methods

In this chapter, we introduce the local explainability algorithm and the detailed structure of each module in our proposed model.

Local explainability algorithm

Given the importance of sufficiency and necessity for image fusion models, we propose the Delta Debugging Fuse Image (DDFImage) method in this paper. Inspired by Zeller's delta debugging algorithm[23], DDFImage is designed and implemented to meet the need for local explanations.

Let I be an image composed of a set of components $\{i1,i2,\ldots,im\}(m\geq 1)$, with the fused image F output by the image fusion operator f. Local explanation correctly answers why the model assigns features to the fused image. Therefore, the DDFImage method reduces I to I', as follows:

$$I'\subset I\vee Suf(I')\wedge Nec(I')\wedge minimal(I').$$

(1)

I' contains only segments of I. I' should maintain sufficiency and necessity for the fused image I' output by the image fusion model F. Specifically, sufficiency of I' implies that the fused image F retains the important information of I', and necessity of I' means that excluding the changes in I' (denoted as $I\backslash I'$), will cause changes in the fused image F. Ultimately, the expected properties retained by I' are $Suf(I')\wedge Nec(I')$, where:

$$Suf(I')\Leftrightarrow f(I')=F.$$

(2)

$$Nec(I')\Leftrightarrow f(I\backslash I')\neq F.$$

(3)

Removing any segment W from I' will not retain the desired properties. That is,

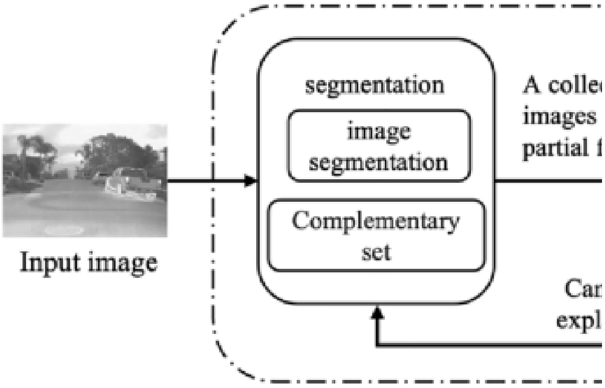
$$\forall W\subset I',(Suf(W)\wedge Nec(W))\neq True.$$

(4)

DDFImage algorithm

Algorithm 1 takes as input the target image fusion operator f and the input image I as inputs and outputs the explanation I'. The overall process is as follows: starting from the granularity n, with the initial granularity initialized to 2, the original image I is partitioned using an image segmentation method to obtain segmented images S (lines 1–3 of the algorithm). As long as the partition operation on the image is allowed, the image is divided into multiple images at granularity n, resulting in multiple images, and then candidate explanations Ecandidate are obtained by checking against the desired properties (line 4). The successful identification of Ecandidate (which is the result of the partition and is therefore smaller in scale than the original image) will affect the update of the granularity (i.e., generate a new granularity n_{new}), and further processing of the candidate explanations is performed with n_{new} (lines 6–8). However, if no candidate image retains the expected properties, granularity is increased to further divide the image (line 10). After undergoing continuous partition operations and updates in granularity, we obtain the final candidate explanation, which is the explanation of the minimum scale that satisfies sufficiency and necessity. Figure 1 illustrates the overall framework of our algorithm.

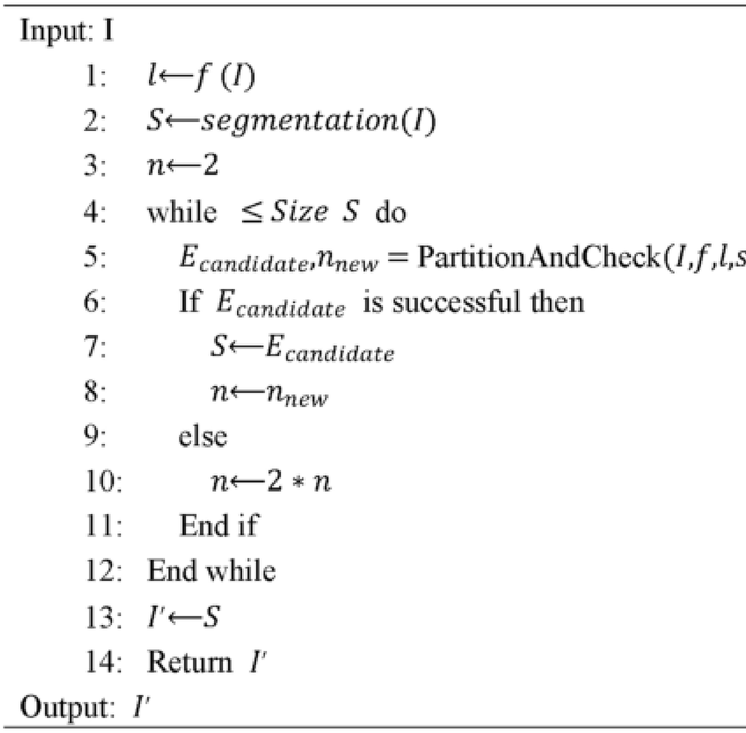
Fig. 1



DDFImage algorithm structure diagram.

Full size image

Algorithm 1



```
Input: I
1:  l←f (I)
2:  S←segmentation(I)
3:  n←2
4:  while  ≤ Size S do
5:      E_candidate,n_new = PartitionAndCheck(I,f,l,s
6:      If E_candidate is successful then
7:          S←E_candidate
8:          n←n_new
9:      else
10:         n←2 * n
11:     End if
12: End while
13: I′←S
14: Return I′
Output: I′
```

DDFImage

Full size image

The core of the algorithm lies in partitioning the given image at granularity n. Operations such as partitioning the image at granularity n and checking whether the desired properties are retained in the partitioned images are performed by the function PartitionAndCheck.

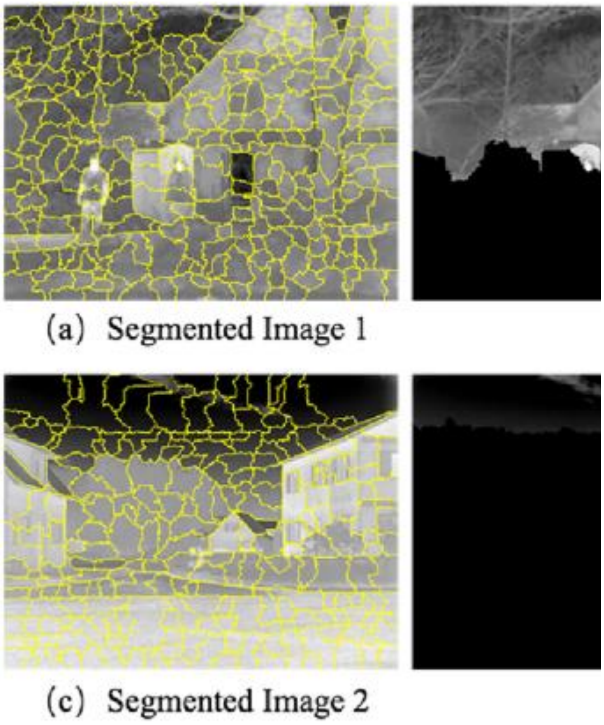
Algorithm 1 shows the details of the PartitionAndCheck function. For a given image, this function obtains a set of sub-images by applying partitioning operations, and then checks each sub-image. The function returns the first sub-image that passes the check as a candidate explanation.

The PartitionAndCheck function mainly uses a series of rules to obtain candidate explanations that satisfy the desired properties. Moreover, the explanation consists of a subset of fragments of the original input image, and the process of reducing the original input image to the explanation involves a series of image partitioning operations. Since partitioned images can be processed by masking their parts, we adjust the operations of cutting and complementing (applied through Delta Debugging24) to handle partitioned images.

In Algorithm 1, the cutting operation is performed first by dividing the image into several subsets S1,S2,…,Sn by applying the segmentation method to the image S and granularity n. A subset S1 is obtained by masking a specific subset of fragments from S. Specifically, for all Si and Sj (i≠j), it is required that S1∪S2∪S3∪...∪Sn=S and S1∩S2=∅.

Figure 2 illustrates the cutting operation. Figure 2a shows an image after applying the image segmentation technique, and Fig. 2b shows the image after cutting it into two images at granularity 2. Additionally, in this study, masked subset fragments are filled with black pixels. If all images generated after cutting operations can successfully identify candidate explanations, further partitioning of G (lines 4–5) is performed. If images Si generated by segmenting G cannot satisfy the desired properties and cannot serve as candidate explanations, further processing (i.e., complementing operation) is required. For each image Si, its complement with respect to S, the image (I\Si) generated from I when Si does not exist, is constructed. Figure 2a and c serve as S, and Fig. 2b and d are the complementary images relative to S, each complementing the other.

Fig. 2



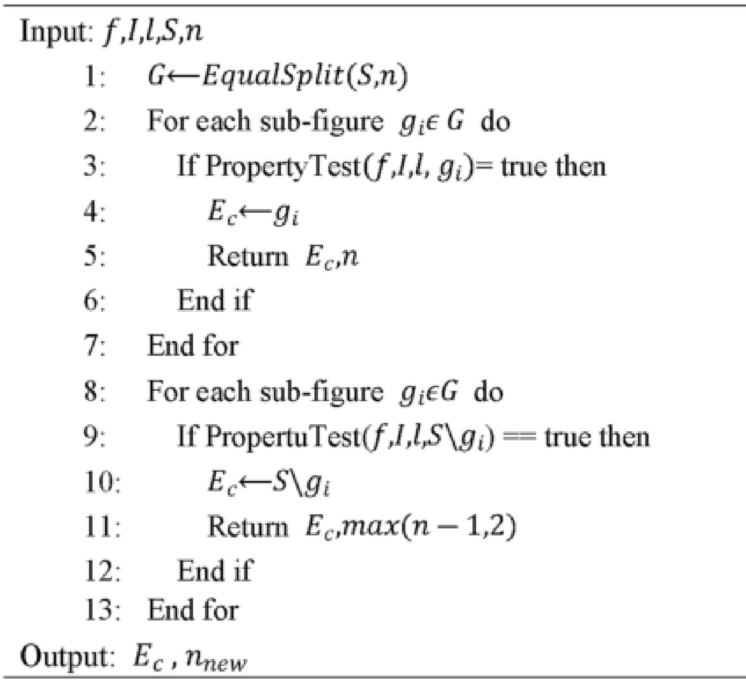
DDFImage image partitioning algorithm.

Full size image

The PartitionAndCheck function not only obtains a series of sub-images containing subsets through image partitioning but also checks these images against the desired properties to obtain candidate explanations. Lines 3 and 9 of the algorithm perform property checks, where sufficiency and necessity are the expected properties. To verify whether any image (Si) is sufficient to explain f(I), the PropertyTest function implements the conditions specified in Eqs. (2 and 3). The first step of this function is to examine the fusion results by comparing the result f(Si) with the original fusion result. If the two results are consistent, it indicates that Si satisfies sufficiency. The second step involves generating the complement of Si with respect to I, that is, generating the image (I\Si) when Si does not exist. Similar to the previous step, the result f(I\Si) needs to be compared with the original fusion image F. If the two results are inconsistent, it indicates that Si satisfies necessity.

The PartitionAndCheck function obtains a set of sub-images containing fragments by continuously partitioning the image at granularity n and determines whether the current sub-image satisfies sufficiency and necessity, thereby obtaining candidate explanations.

Algorithm 2



```
Input: $f,I,l,S,n$
  1:  $G \leftarrow EqualSplit(S,n)$
  2:  For each sub-figure $g_i \epsilon G$ do
  3:      If PropertyTest($f,I,l, g_i$)= true then
  4:          $E_c \leftarrow g_i$
  5:          Return $E_c,n$
  6:      End if
  7:  End for
  8:  For each sub-figure $g_i \epsilon G$ do
  9:      If PropertuTest($f,I,l,S\backslash g_i$) == true then
  10:         $E_c \leftarrow S\backslash g_i$
  11:         Return $E_c,max(n-1,2)$
  12:     End if
  13: End for
Output: $E_c$ , $n_{new}$
```

PartitionAndCheck

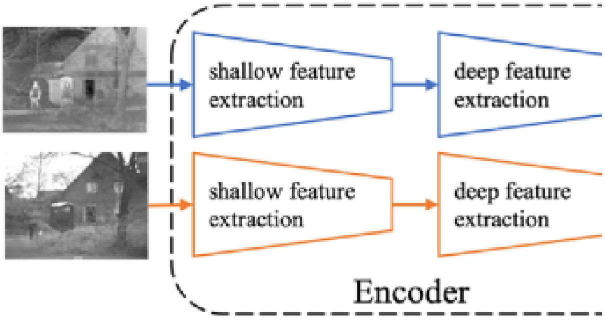
Full size image

CMCFFuse image fusion model

Encoder

Considering both shallow and deep feature extraction, the encoder consists of a CNN encoder and a Transformer encoder. The fusion model is illustrated in Fig. 3.

Fig. 3



CMCFFuse image fusion model structure.

Full size image

Simple convolutional layers can learn basic image features like background edges and textures. The CNN encoder comprises two 3 × 3 convolutional kernels with LeakyReLU activation, a stride of 1, and padding of 1. The input infrared and visible light images are denoted as IϵRH×W and IϵRH×W, respectively. The CNN encoder extracts shallow features FI1 and FV1 from the infrared and visible inputs {I,V}. Equation (5) represents the shallow feature extraction process.

FI1=H1(I),FV1=H1(V).

(5)

Here, FI1 and FV1 represent the shallow feature information of the infrared and visible light images, respectively. I and V denote the infrared and visible light images, respectively, and H1 () represents the 3 × 3 convolution + LeakyReLU activation operation.

Next, FI1 and FV1 undergo deep feature extraction to extract detailed features, as shown in Eq. (6). The deep feature extraction module consists of MDTA25 and SE26.

FI2=H2(FI1),FV2=H2(V).

(6)

Here, FV1 and FV2 represent the deep feature information of the infrared and visible light images, respectively, and H2 () represents the deep feature extraction operation.

Cross-modal feature fusion and decoder

After extracting sufficient deep-level information, a cross-modal feature fusion module is designed. It combines the shallow feature information of infrared images with the deep feature information of visible light images and vice versa to achieve cross-modal fusion. LT and INN (invertible neural network) blocks are used in the shallow and deep fusion layers. The LT block flattens the structure of the feedforward network, flattens the bottleneck part of the Transformer block and reduces the embedding dimension, thereby reducing the number of parameters. The INN module is designed for reversibility, allowing mutual generation of input and output features to prevent information loss, in line with the goal of preserving high frequency features in the fused image.

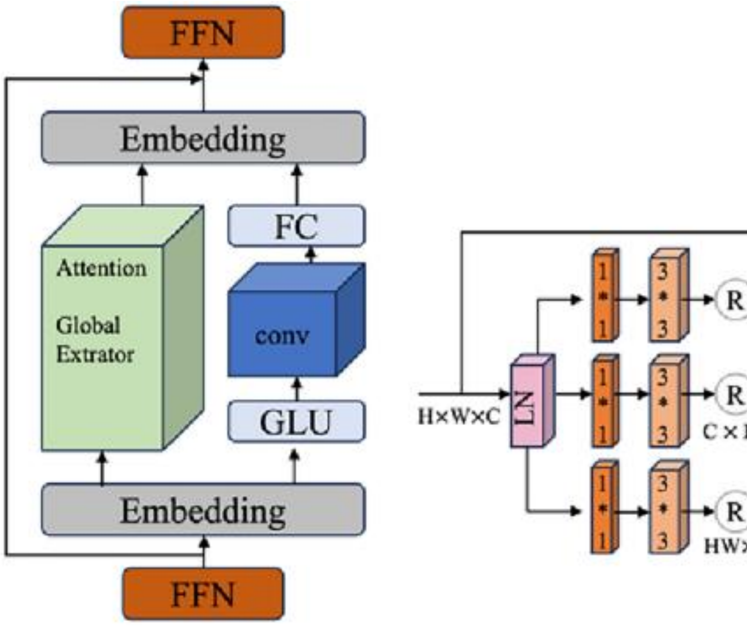
FI=concat(FI1,FV1),FV=concat(FI2,FV2).

(7)

Here, FI and FV represent the results of cross-modal fusion.

As image fusion involves cross-modal and multi-frequency features, our decoder uses the same structure as the encoder, consisting of a CNN encoder, MDTA, and SE. The structures of these modules are shown in Figs. 4 and 5.

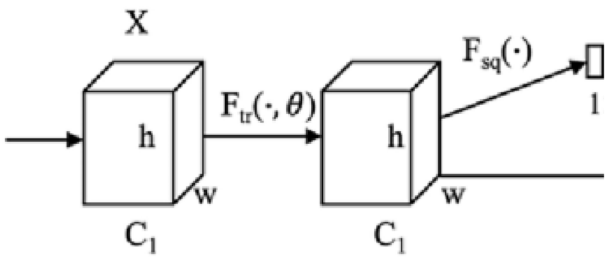
Fig. 4



LT Block and MDTA block structure.

Full size image

Fig. 5



Network of channel attention squeezing excitation module (SE) block structure.

Full size image

Other explainability methods and data construction

1. 1.

BayLIME BayLIME27 is an attribution-based method that uses Bayesian inference and prior knowledge to generate model-agnostic explanations. It extends the popular explanation method LIME by enhancing the consistency of individual prediction explanations and other aspects through prior knowledge and Bayesian inference. In our experiments, we use the same configuration as Zhao et al.27.

2. 2.

SEDC SEDC is a counterfactual-based method that applies heuristic search to explore model-agnostic explanations. Initially proposed for document classification28, SEDC has recently been applied to image classification29. In this study, SEDC outputs a set of explanations to explain the decision-making of the final fused image obtained from the input images. We follow existing research and select the smallest scale explanation as the output.

3. 3.

Feature importance analysis Image fusion involves merging multiple images or different features within an image to produce an output with comprehensive information. During this process, it is essential to extract the most representative and informative features for each input image to ensure the quality and effectiveness of the fusion result. We use Gradient-weighted Class Activation Mapping (Grad-CAM) to recursively remove features and evaluate accuracy to determine which features are more important during the fusion process.

4. 4.

Model selection and datasets We select the CMCFFuse model proposed in our study, which achieves satisfactory accuracy in fusing infrared and visible light images. In the experiments, we use a pre-trained CMCFFuse model. We select two widely used datasets for image fusion, namely the TNO dataset and the RoadSence dataset.

Results

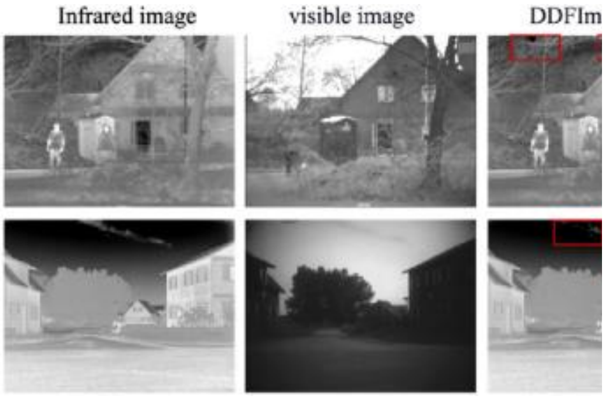
We will demonstrate the importance of providing sufficient and necessary explanations for image fusion models through specific experiments. For a given image I and image fusion model F, different explanations may be obtained when applying different explainability methods. Based on the generated explanations, the original infrared image is fused with the visible light image of the explanation result to generate a more explainable image.

The experiments were conducted using the PyTorch framework, leveraging its capabilities for deep learning and image processing. The testing machine was equipped with an Intel Core i5-12400F processor, a GeForce RTX 4060 graphics card, and 32 GB of RAM. This configuration provided sufficient computational power to efficiently execute the proposed multimodal image fusion model and the DDFImage algorithm, allowing for effective handling of large datasets and facilitating rapid model training and evaluation. We trained the model for 100 epochs using the publicly available TNO and Roadscene datasets, ensuring the robustness and generalizability of our approach.

Ability to generate sufficient and necessary local explanations

The DDFImage algorithm generates both sufficient and necessary explanations for input images. For example, as illustrated in Fig. 6, the explanations provided by DDFImage effectively highlight critical features that contribute to the image fusion process. In contrast, the explanations generated by the BayLIME method only fulfill the criterion of sufficiency, lacking essential information needed for comprehensive understanding. Similarly, the SEDC method offers necessary explanations, yet it fails to provide sufficient context, leading to a deficit in important feature representation in the fused images. This disparity is evidenced in Table 1, where the fused images generated using BayLIME and SEDC methods are shown to lack significant features, thereby diminishing the overall effectiveness of the fusion process.

Fig. 6



Example of DDFImage generation with sufficient necessary explanation.

Full size image

Table 1 The percentage of explanations generated by different methods (retaining sufficiency, necessity, or both).

Full size table

Influence of information preservation rate on the explainability of image fusion models

The information preservation rate is a critical factor influencing the explainability of image fusion models. To quantify this, we utilized the Structural Similarity Index (SSIM) metric to evaluate the similarity between the fused images generated by the DDFImage method and the original images. The SSIM metric ranges from − 1 to 1, where values closer to 1 indicate higher structural similarity. Higher SSIM scores suggest that the generated images effectively preserve the essential information of the original images, thus enhancing the model’s explainability.

In our experiments, we calculated the SSIM values for both the generated occluded images and the fused images in comparison to the original images. Additionally, we performed experiments using the SwinFuse model for comparative analysis. The results, summarized in Table 2, demonstrate that employing SSIM as a metric for information preservation effectively evaluates the explainability of image fusion models. Notably, higher SSIM values reflect the model’s capacity to retain crucial structures and features during image generation, thereby enhancing the overall explainability of the model.

Table 2 Percentage of information retention after explanation generated by different methods.

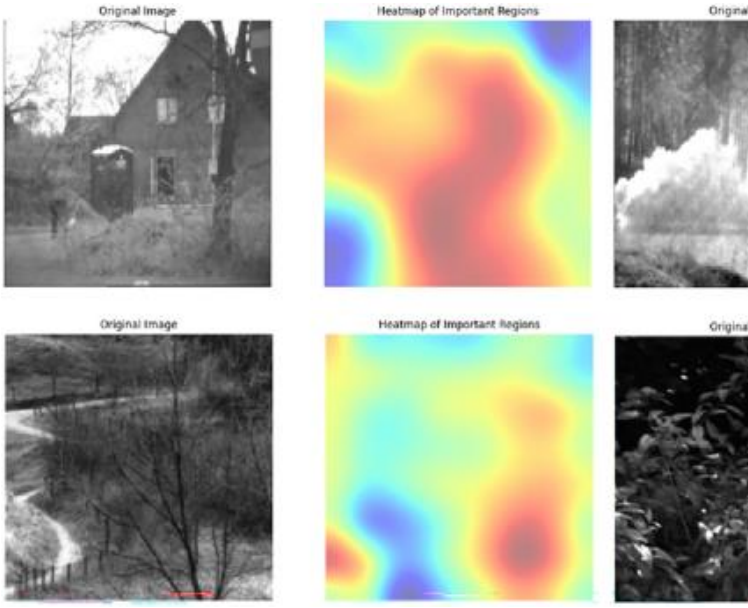
Full size table

Our experimental results demonstrate that using the SSIM values as a metric for the information preservation rate enables effective evaluation of the explainability of image fusion models. Higher SSIM values reflect the model’s preservation of important structures and features of the original images during image generation, thereby improving the model’s explainability.

Enhancing the explainability of image fusion models through feature importance analysis

Leveraging the DDFImage algorithm, we gained insights into the local explanations provided, revealing that the model assigns significant importance to certain features within the fused image. We conducted feature importance analysis using Grad-CAM, which integrates gradient information with feature map data, specifically for the CMCFFuse model. Important features are visualized as heatmaps, as shown in Fig. 7, illustrating the areas the CMCFFuse model prioritizes when generating fused images. The red regions indicate critical areas of focus for the model, whereas blue regions represent areas with less emphasis. This visualization aids in understanding how the model makes fusion decisions based on the salient features of the input images.

Fig. 7



Feature importance analysis heat map.

Full size image

Discussion

This study, through the explainability analysis of visible light and infrared image fusion technology, reveals key insights into the internal mechanisms of fusion models.

Firstly, we found that local explainability algorithms significantly enhance the interpretability of image fusion models. This indicates that focusing on local information in multimodal image fusion aids in understanding the model's decision-making process and results more effectively.

Secondly, the application of the DDFImage algorithm has a marked impact on the explainability of fusion models. Through its unique data processing and feature fusion methods, the DDFImage algorithm not only boosts model performance but also enhances its credibility, making the model's decisions more trustworthy. For instance, our analysis showed that the inclusion of certain features led to a noticeable increase in performance metrics, such as accuracy and precision. In our study, the DDFImage algorithm has demonstrated significant improvements in the performance of fusion models across various datasets. However, we acknowledge certain limitations that warrant further investigation. Specifically, there are scenarios where the integration of DDFImage with fusion models did not yield the anticipated advantages. For instance, we observed that when applied to images with high noise levels or extreme blur, the performance of DDFImage could be adversely affected.

Through our analysis, we noted that images characterized by complex textures and low contrast may lead to suboptimal feature extraction, ultimately diminishing the effectiveness of the DDFImage algorithm. Such characteristics highlight the need for careful consideration of input image quality when employing this algorithm. These findings underscore the importance of exploring these edge cases in future research. Further experiments are necessary to investigate the conditions under which DDFImage may not perform optimally and to develop potential modifications to enhance its robustness against such challenges.

Additionally, we observed a strong correlation between the explainability of image fusion models and their performance. Feature importance analysis revealed specific key features that are critical for model performance and explainability. These findings suggest that integrating explainability factors into the design of image fusion models can lead to improved performance and credibility.

Furthermore, practical applications of this research, such as in medical imaging or autonomous driving, can benefit from enhanced explainability, thereby improving user trust and model adoption.

Conclusion

This study presents a comprehensive analysis of the explainability of visible light and infrared image fusion technology, yielding several significant findings. Firstly, through the application of local explainability algorithms and specific fusion models, we demonstrated that focusing on local information in multimodal image fusion substantially enhances both the explainability and performance of the models. The findings indicate that incorporating local context is crucial for a deeper understanding of the model's decision-making processes.

Secondly, we established that the DDFImage algorithm not only improves the performance of fusion models but also significantly enhances their explainability. By generating sufficient and necessary local explanations, DDFImage offers a novel approach to image fusion, thus providing valuable insights and methodologies for advancing the field.

In the Results and Analysis section, we conducted an in-depth evaluation of model performance, feature importance, and information preservation rates, revealing the intricate relationship between explainability and model efficacy. Our analyses confirmed that higher information preservation rates, as indicated by SSIM metrics, correlate positively with the explainability of image fusion models, highlighting the necessity of incorporating explainability factors in model design and evaluation.

Despite achieving meaningful results, this study acknowledges certain limitations, including the need for validation across more diverse datasets and real-world scenarios. Future research should explore various explainability methods and their applicability in broader contexts, aiming to further enhance the explainability and performance of image fusion technologies.

In summary, this study contributes new perspectives and methodologies for the explainability analysis of visible light and infrared image fusion technology.

The implications of our findings are both theoretical and practical, underscoring the importance of explainability in improving the performance and credibility of image fusion models.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

1. He, K., Zhang, X., Ren, S. et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).

2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90 (2017).

Article MATH Google Scholar

3. Bahdanau, D., Cho, K., & Bengio, Y. Neural machine translation by jointly learning to align and translate. Print at arXiv:1409.0473 (2014).

4. Sutskever, I., Vinyals, O., & Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* 27–35 (2014).

5. Geiger, A., Lenz, P., & Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition* 3354–3361 (IEEE, 2012).

6. Tobiyama, S., Yamaguchi, Y., Shimada, H. et al. Malware detection with deep neural network using process behavior. In *2016 IEEE 40th annual computer software and applications conference (COMPSAC)*, vol. 2, 577–582 (IEEE, 2016).

7. Nguyen, A., Yosinski, J., & Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 427–436 (2015).

8. Papernot, N., McDaniel, P., Jha, S. et al. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, 372–387 (IEEE, 2016).

9. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* 1135–1144 (2016).

10. Lundberg, S. M., Lee, S. I. A unified approach to interpreting model predictions. In *Advances in neural information processing systems* 30–39 (2017).

11. Guidotti, R. et al. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **51**(5), 1–42 (2018).

Article MATH Google Scholar

12. Ma, J., Ma, Y. & Li, C. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* **45**, 153–178 (2019).

Article MATH Google Scholar

13. Li, H. & Wu, X. J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **28**(5), 2614–2623 (2018).

Article MathSciNet MATH Google Scholar

14. Zhang, L. et al. Cross-modality interactive attention network for multispectral pedestrian detection. *Inf. Fusion* **50**, 20–29 (2019).

Article MATH Google Scholar

15. Liu, Y., Chen, X., Cheng, J. et al. A medical image fusion method based on convolutional neural networks. In *2017 20th international conference on information fusion (Fusion)* 1–7 (IEEE, 2017).

16. Xu, H. & Ma, J. EMFusion: An unsupervised enhanced medical image fusion network. *Inf. Fusion* **76**, 177–186 (2021).

Article MATH Google Scholar

17. Ma, J. et al. FusionGAN: A generative adversarial network for infrared and visible

image fusion. *Inf. Fusion* **48**, 11–26 (2019).

Article CAS MATH Google Scholar

18. Ma, J. et al. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **70**, 1–14 (2020).

MATH Google Scholar

19. Vaswani, A., Shazeer, N., Parmar, N. et al. Attention is all you need. Advances in neural information processing systems 30–40 (2017).
20. Vs, V., Valanarasu, J. M. J., Oza, P. et al. Image fusion transformer. In *2022 IEEE International conference on image processing (ICIP)*, 3566–3570 (IEEE, 2022).
21. Wang, Z. et al. SwinFuse: A residual swin transformer fusion network for infrared and visible images. *IEEE Trans. Instrum. Meas.* **71**, 1–12 (2022).

Article MATH Google Scholar

22. Petsiuk, V., Das, A., & Saenko, K. Rise: Randomized input sampling for explanation of black-box models. Print at arXiv:1806.07421 (2018).
23. Zeller, A. & Hildebrandt, R. Simplifying and isolating failure-inducing input. *IEEE Trans. Softw. Eng.* **28**(2), 183–200 (2002).

Article MATH Google Scholar

24. Wang, G., Shen, R., Chen, J. et al. Probabilistic delta debugging. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* 881–892 (2021).
25. Zamir, S. W., Arora, A., Khan, S. et al. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 5728–5739 (2022).
26. Hu, J., Shen, L., & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 7132–7141 (2018).
27. Zhao, X., Huang, W., Huang, X. et al. Baylime: Bayesian local interpretable model-agnostic explanations. In *Uncertainty in artificial intelligence*, 887–896 (PMLR, 2021).
28. Martens, D. & Provost, F. Explaining data-driven document classifications. *MIS Q.* **38**(1), 73–100 (2014).

Article MATH Google Scholar

29. Vermeire, T. et al. Explainable image classification with evidence counterfactual. *Pattern Anal. Appl.* **25**(2), 315–335 (2022).