Dr. Yuting Wan
State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University
129 Luoyu Road, Wuhan, Hubei, 430079
P. R. China
E-mail: wanyuting@whu.edu.cn

Dear Prof. Zhao-Liang LI,

Revision of our manuscript TGRS-2024-05984

Thank you for your comments and suggestions. We have carefully revised our paper (TGRS-2024-05984) in light of the associate editor's and reviewers' comments. A point-by-point response to the comments is attached to this letter. The major changes have been highlighted by colored text and are summarized as follows:

1. The Section II has been adjusted for better understanding with more comprehensive details added to lead to our contributions. (see the response to reviewer 1).
2. For the contributions of the proposed fusion method, they have been reorganized to correspond to the stated problem. In addition, the details of our proposed fusion methods have been added for a comprehensive understanding. (see the response to reviewer 1, reviewer 2 and reviewer 3).
3. The variables in the formulas have been explained clearly with specific characters representing specific meaning in the revised paper (see the response to reviewer 1, reviewer 2 and reviewer 3).
4. In the experimental section, more state-of-the-art deep-learning-based fusion methods have been added for comparison. The parameter settings have been introduced and adjusted to be consistent with the references. (see the response to reviewer 1).
5. The heat maps of the detection results have been added, with more detailed explanations and descriptions of the corresponding results. (see the response to reviewer 1 and reviewer 2).
6. All the figures have been redrawn for better readability. A unified paradigm has be established for presenting the figures with the same style. (see the response to reviewer 1, reviewer 2 and reviewer 3).
7. All the references are corrected and formatted properly. (see the response to reviewer 3).

For detailed changes, please refer to the response letters attached.
Thank you again for your comments and time.
Best regards,

Dr. Yuting Wan
Post-doctoral of remote sensing

Response to the Comments of the Associate Editor

The manuscript presents an infrared and low-light multimodal anomaly detection dataset and a benchmark for SOTA anomaly detection methods. In overall, the work is of some value, the structure is good, and experiments are extensive. Still, some important issues related to the writting, descriptions, and experiments were pointed out, which need to be addressed before considering its publication.

Thank you for your comments and suggestions. We have carefully revised our paper (TGRS-2024-05984) in light of the associate editor's and reviewers' comments. A point-by-point response to the comments is attached to this letter. The major changes have been highlighted by colored text and are summarized as follows:

1. The Section II has been adjusted for better understanding with more comprehensive details added to lead to our contributions. (see the response to reviewer 1).
2. For the contributions of the proposed fusion method, they have been reorganized to correspond to the stated problem. In addition, the details of our proposed fusion methods have been added for a comprehensive understanding. (see the response to reviewer 1, reviewer 2 and reviewer 3).
3. The variables in the formulas have been explained clearly with specific characters representing specific meaning in the revised paper (see the response to and reviewer 3 reviewer 2 and reviewer 3).
4. In the experimental section, more state-of-the-art deep-learning-based fusion methods have been added for comparison. The parameter settings have been introduced and adjusted to be consistent with the references. (see the response to reviewer 1).
5. The heat maps of the detection results have been added, with more detailed explanations and descriptions of the corresponding results. (see the response to reviewer 1 and reviewer 2).
6. All the figures have been redrawn for better readability. A unified paradigm has be established for presenting the figures with the same style. (see the response to reviewer 1, reviewer 2 and reviewer 3).
7. All the references are corrected and formatted properly. (see the response to reviewer 3).

Finally, thank you again for your very helpful suggestions.

Response to the Comments of the Reviewer #1

This paper focuses on nighttime rescue missions, establishing a low-light infrared multimodal remote sensing image dataset, and proposing a rank decomposition-based fusion method. Based on this dataset and fusion technique, several typical unsupervised anomaly detection methods were tested, addressing the deficiencies in nighttime rescue scene datasets while introducing and evaluating different types of anomaly detection algorithms. However, there are several problems with this paper.

1. ***Overall, the content of this paper is extensive. However, it is necessary to define whether the primary contribution of the paper is to serve as a comprehensive review or to introduce a new dataset and benchmark. If the paper aims to be a review, a significant portion of the second section should be dedicated to detailing the research trajectory of this field, culminating in the introduction of new methods and their comparison with existing ones.***

R. Thank you for your suggestion.

We sincerely appreciate your feedback regarding our paper. We would like to clarify that the main contribution of our work is to introduce a new dataset and benchmark, as opposed to being a comprehensive review.

In the revised manuscript, we have made the following adjustments to better emphasize this aspect. In the second section, while still providing an overview of the research trajectory to offer context, we have streamlined the content to focus more on the gaps and challenges that our new dataset and benchmark are designed to address. We have highlighted how the existing datasets and benchmarks in the field were insufficient for certain aspects of nighttime rescue mission anomaly detection, and thus motivated the creation of our new resources. (Page 2, right column)

"Research in the field of night rescue using multimodal low-light and infrared remote sensing has undergone a remarkable evolution. Initially, the focus was primarily on the stand-alone application of low-light and infrared remote sensing technologies. Low-light remote sensing was explored for its ability to capture spatial and textural detail, although it is highly sensitive to lighting conditions. Infrared remote sensing, on the other hand, was recognised for its ability to detect prominent targets with less dependence on ambient light, although it suffered from issues such as lower resolution and blurred backgrounds."

"As the field progressed, the concept of fusing these two modalities emerged. Early attempts involved traditional fusion algorithms based on spatial or transform domains. These aimed to combine the complementary features of low-light and infrared images to enhance the overall information content. However, with the advent of deep learning, more advanced fusion algorithms such as autoencoder, convolutional neural network and adversarial neural network algorithms were proposed. These deep learning-based methods showed promise in improving the quality of fused images and extracting more meaningful features."

2. ***The proposed low-light and infrared fusion method based on rank decomposition primarily extracts high-frequency and low-frequency features using two existing feature extractors. Is it appropriate to name it rank decomposition? Additionally, the author claims that transformers can extract long-range low-frequency information while CNNs effectively provide high-frequency information, I would advise doing some validation about this in the context.***

R. Thank you for your question and suggestion.

Firstly, we are sorry for the incorrect and misleading naming of our method. We have thoroughly revised the manuscript to make sure that all references to the method are based on "frequency domain feature decomposition" rather than "rank decomposition”. This includes updating the titles of relevant sections, figure captions, and any other instances where the incorrect naming was used.

Regarding the capabilities of transformers and CNNs in extracting specific frequency information: The “long-range information” in the context actually refers to global features while the “short-range information” indicates local features [1]. Transformers-based models can effectively exploit the global structures as they split an image to a sequence of patches and model their dependencies with the spatial self-attention mechanism [2,3]. CNN-based models can effectively extract local features within the receptive fields [4,5] with convolution operation. From some previous studies [6,7,8], the global structures (e.g., background) are more associated with the low-frequency information while the local structures (e.g., edges and lines) are more related to the high-frequency components of the images.

In order to provide further validation, we randomly selected and illustrated 9 channels of the output feature maps from the Lite-transformer-based block (Transformer-based) and the Invertible Residual Neural Network block (CNN-based). The specific analysis is as follows.

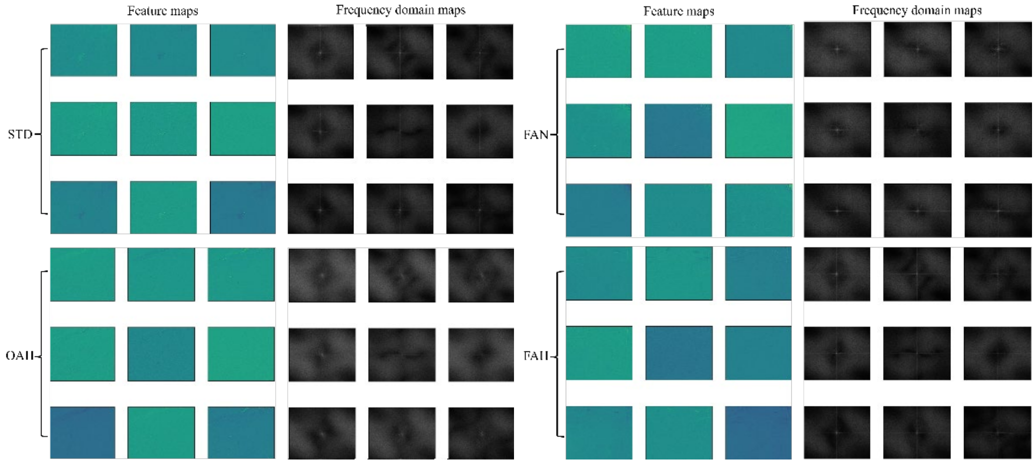

Fig. 1. The visualization of the output feature maps and their frequency domain maps of Lite-transformer-based block.

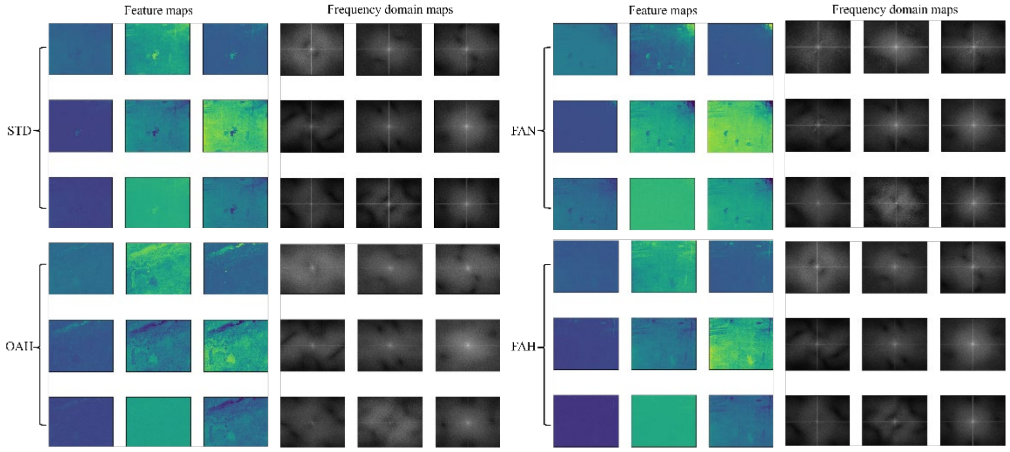

Fig. 2. The visualization of the output feature maps and their frequency domain maps of Invertible Residual Neural Network block.

As illustrated in Fig. 1, the feature map of Lite-transformer-based block exhibits obvious details and texture, along with a more substantial high-frequency information in the frequency domain. Conversely, the feature map of the IRNN block demonstrates similar color, containing rich background information, and preserves more low frequencies in the frequency domain.

The related reference is listed as follows.

[1] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. In *ICLR*, 2020. 3, 4

[2] Park, N., Kim, S.: How do vision transformers work? In: *ICLR* (2022)

[3] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: *ICCV* (2021).

[4] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: *ICLR* (2018).

[5] Brendel, W., Bethge, M.: Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In: *ICLR* (2019).

[6] Campbell, F.W., Robson, J.G.: Application of fourier analysis to the visibility of gratings. *The Journal of physiology* 197(3), 551 (1968)

[7] De Valois, R.L., De Valois, K.K.: Spatial vision. *Annual review of psychology* 31(1), 309–341 (1980)

[8] Sweldens, W.: The lifting scheme: A construction of second generation wavelets. *SIAM journal on mathematical analysis* 29(2), 511–546 (1998).

3. In validating the effectiveness of the proposed fusion method, the comparison is made only with wavelet transforms, lacking comparisons with other deep learning-based fusion methods, this is not convincing enough and diminishes the argument's strength. Furthermore, the implementation process and experimental setup for the wavelet transform are not detailed in the paper.

R. Thank you for your suggestions.

In the experimental section, more state-of-the-art deep-learning-based fusion methods, such as U2Fusion [107], SDNet [108], TarDAL [109] and DeFusion [110], have been added for comparison. The parameter settings have been introduced and adjusted to be consistent with the references. Subsequently, six metrics are employed to quantitatively compare the above results, which are displayed in Table. II. The specific analysis is as follows. (Page 13, left column)

> “The experimental results demonstrate that the proposed method has excellent performance on almost all metrics, proving that our method is suitable for various kinds of scenarios. Specifically, the method proposed loses some MI (e.g. in STD and FAH scenarios) due to independence loss, but it still achieves better results compared to the baseline. Furthermore, the SD and SSIM of the proposed fusion method are less effective in the OAH scenario, which is related to the inherent contrast and environmental gradients in the OAH images.”

The related reference is listed as follows. (Page 18, right column)

[107] H. Qin, R. Gong, X. Liu, M. Shen, Z. Wei, F. Yu, and J. Song, “Forward and backward information retention for accurate binary neural networks,” in *CVPR*, pp. 2250–2259, 2020.

[108] H. Zhang and J. Ma. Sdnet, “A versatile squeeze-and-decomposition network for real-time image fusion,” Int. J. Comput. Vis., 129(10):2761–2785, 2021.

[109] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, “Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection,” in CVPR, pp. 5792–5801, 2022.

[110] P. Liang, J. Jiang, X. Liu, and J. Ma, “Fusion from decomposition: A self-supervised decomposition approach for image fusion,” in ECCV, 2022.

In addition, we added the implementation process and experimental setup for wavelet transform in the revised manuscript. The specific details are as follows. (Page 11, left column)

"For the wavelet transform, the wavelet basis function was the second-order Daubechies wavelet, with the decomposition level set to 1, where the fusion strategy was using the average method for low-frequency and the maximum method for high-frequency. For the deep-learning-based fusion methods, the parameter settings were adjusted to be consistent with the references."

4. ***Additionally, there are issues with writing and formatting. In Fig. 1, the small diagram in the preprocess section lacks necessary annotations, making the preprocessing steps unclear. In Figure 2 which illustrates the pipeline of the proposed fusion model, there is only text and a few pictures in this drawing, and no clear visual distinction is made between the different components, especially between the "Low-frequency Extractor" module and the "High-frequency Extractor" module. It is recommended to add a formula or simple model structure diagram and use color to distinguish different modules. Besides, the font size of in Figure 2 is too small to read, so it is recommended to add some other content and make the entire figure take up more space. In the last paragraph on page 3, the author seems to confuse high-frequency and low-frequency when discussing the low-frequency feature extractor, please check the words and make it clearer. On page 6, the constraint condition: α1 + α2 + α3 = 1, appears abruptly after the sentences without explanation, and this may be integrated into the formula like st.α1 + α2 + α3 = 1. The function F in formula (4) is not introduced elsewhere in the paper.***

R. Thank you for your careful review and suggestions.

All the figures in the revised manuscript have been redrawn and adjusted for better readability. In addition, we have checked all the formulas to make sure that the specific variables have specific meanings.

5. ***This paper evaluates the detection performance of various methods on low-light, infrared, and fused images. The detection outcomes post-fusion do not appear to surpass those of the single-modality images, questioning the necessity of fusion. In the detection performance based on fusion, many indicators are used but lack specific instructions. It includes how indicators are calculated and compared, as well as numerical analysis of model differences. It is not recommended to merely list tables without adding detailed explanations. At the same time, for example, Figure 10 represents the performance comparison of traditional and deep learning detection algorithms on the fusion data set, but the text and figure note do not include specific analysis.***

R. Thank you for your question and suggestions.

We have meticulously compared the detection results based on the fused and single-modality images.

In qualitative comparison, we have added the heat maps to better analyze the differences in detection. The illustration of heat maps demonstrates that, in comparison with the single-modality, high anomaly scores (i.e. the hotter regions) are more concentrated on the person to be rescued, with diminished distribution over the background. Furthermore, in the binary maps, the anomaly regions cover the ground truth more comprehensively, demonstrating the improvements with fusion. The specific illustration is as follows.

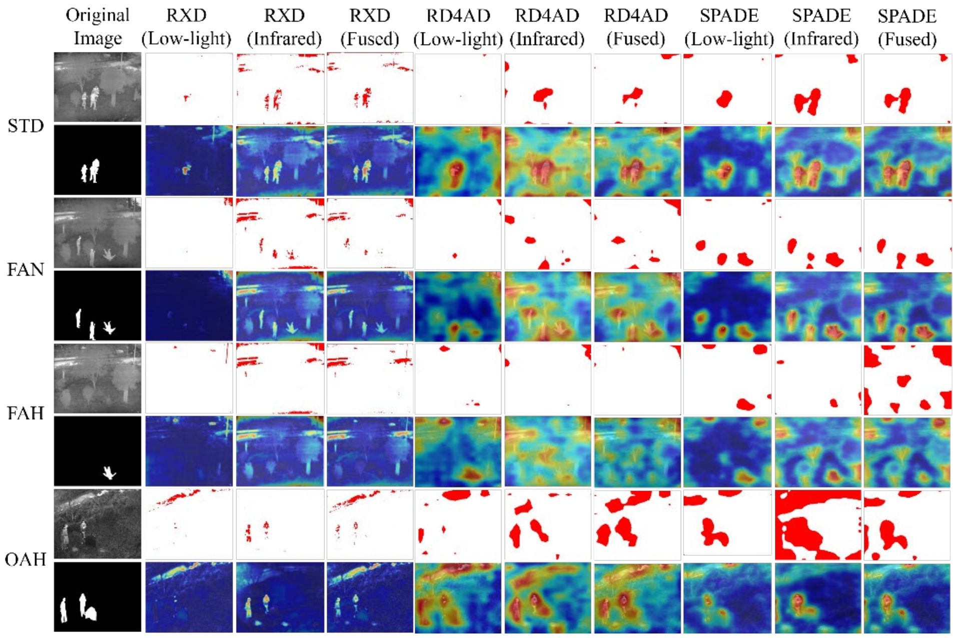

Fig. 3. The binary maps and heat maps of the three detection methods based on the fused and single-modality images.

In quantitative comparison, the detection results of post-fusion have more superior AUC and TPR performance on almost all scenarios than single-modality. However, the performance of fused images in FPR does not consistently exceed that of the single-modal images, where the misplaced regions are more localized in Fig. 10. Alleviating this problem requires associating more features with contextual relationships, which is the limitation of the proposed fusion approach for anomaly detection tasks in night-time rescue missions. The specific and detailed analysis is as follows. (Page 13, right column)

"The quantitative comparison reveals that the detection results of fusion outperform single-modality in almost all scenarios. The SPADE based on fused images demonstrated improving AUC scores in all four scenarios, with higher TPR scores in STD, FAN, and OAH scenarios. Similarly, the RD4AD method, based on the fusion, exhibited enhanced AUC and TPR values in nearly four scenarios (e.g. AUC values in the STD, FAN, and OAH). In the case of the RXD, the AUC of fusion surpassed the single-modal in the STD and OAH scenarios, and TPR scores were superior in the STD and FAH. However, the performance of fused images in FPR exceed that of the single-modal images only in certain instances (e.g. RXD in STD and OAH scenarios, and RD4AD in STD scenario), where the misplaced regions are more localized in Fig. 10."

In the revised manuscript, the calculation of indicators used in fusion and detection performance has been added, with thorough descriptions. (Page 9, right column and Page 10). Furthermore, we have provided detailed explanations in all tables and figures to make better understanding.

6. ***Regarding the illustrative diagrams of various methods in the review section, a unified paradigm should be established for presenting these diagrams. Additionally, the style of illustrations throughout the paper should be consistent.***

R. Thank you for your suggestion.

All the figures in the revised manuscript have been redrawn and adjusted for better readability. Furthermore, we have also established a unified paradigm, which includes using the same color scheme, line thickness, and notation style to ensure consistency.

Finally, thank you again for your very helpful suggestions.

Response to the Comments of the Reviewer #2

In this paper, the author proposed a low-light and thermal infrared dataset for nighttime rescue mission, a benchmark for SOTA anomaly detection methods and a multi-modal fusion method based on rank decomposition. Using low-light and thermal infrared data for nighttime rescue is novel and practical, some comments are provided as follows.

1. ***In Fig. 1, for the construction of the MRSI-NERD dataset, the author should provide corresponding explanations for the methods or operations used in denoising and registration, and attach the references.***

R. Thank you for your comment and helpful suggestion.

The Fig.1 in the revised manuscript has been redrawn and adjusted with detailed explanations of the denoising and registration methods. Moreover, a thorough description of the MRSI-NERD dataset construction has been provided both in the text and beneath the figure, with appropriate references. (Page 3)

"Fig. 1. The construction of the MRSI-NERD dataset. The overall process is comprised of three distinct steps: data fetching, data preprocessing, and annotation. Initially, the original low-light and infrared images were obtained from the device in four different real scenarios. For the preprocessing step, the denoised images were then obtained by subtracting the infrared dark image (solely containing thermal noise) from the original infrared image [46]. Subsequently, the infrared images were subjected to a process of cropping, whereby the dark edges were removed. The RIFT algorithm [47] was then employed to facilitate the alignment of the low-light and infrared images during the registration process. Finally, the synchronized low-light and infrared images, which contained targets, were fully annotated."

The related reference is listed as follows. (Page 17, left column)

[46] H. Santosa, M. Jiyoun Hong, S.-P. Kim, and K.-S. Hong, "Noise reduction in functional near-infrared spectroscopy signals by independent component analysis," Rev. Instrum., vol. 84, no. 7, 2013.

[47] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," IEEE Trans Process., vol. 29, pp. 3296–3310, 2019.

2. ***Mistake in the first paragraph of section II. "The collected data was divided into five categories of detection tasks", it should be four categories.***

R. Thank you for your careful review.

We are sorry for the incorrect descriptions of the categories of detection tasks and a misunderstanding may have been caused. Four categories instead of five categories were used to consist of the detection dataset. (Page 2, right column)

"The collected data was divided into four categories of detection tasks: 'Fallen And Hypothermic', 'Fallen And Normothermic', 'Obstructed And Hypothermic', and 'Standing', abbreviated as 'FAH', 'FAN', 'OAH', and 'STD' respectively."

3. ***The Section III is the fusion method proposed by the author, but it is not prominent enough in the contribution explanation. In addition, the title "multi-modal feature encoder" is unclear to summarize this section and is similar with the title of subsection B.***

R. Thank you for your careful review and suggestion.

To make the contribution of the proposed fusion method more prominent, we have revised the contribution to provide a more comprehensive and explicit overview. The specific details are as follows. (Page 2)

"The dual-stream fusion network is predicated on frequency domain feature decomposition with a Transformer-CNN framework. This enables the extraction and fusion of global and local features, ensuring a more accurate reflection of modality-specific and modality-shared features. Specifically, the work introduces the innovative utilization of IRNN and Lite-transformer blocks for optimizing the balance between fusion quality and computational cost. The superiority of the proposed fusion method compared to existing algorithms was comprehensively validated through metrics such as entropy (EN), standard deviation (SD), spatial frequency (SF), mutual information (MI), sum of correlation differences (SCD), and structural similarity (SSIM). Improvements in the AUC (area under the curve) metric are reported across multiple downstream anomaly detection algorithms, validating the effectiveness of the approach in the nighttime emergency rescue mission."

Regarding the title "multi-modal feature encoder", we are sorry for the unclear and misleading naming of the title of section III. It is inappropriate to summarize the content thus we have changed it to "DUAL-STREAM MULTI-MODAL FUSION MODEL" instead in the revised paper.

4. ***I suggest that the authors should add the heat maps to show experimental results, and analyze the advantages and disadvantages of the results together with the detection binary maps.***

R. Thank you for your suggestion.

In order to show further experimental results, the heat maps have been added in single-modality and fusion detection results. The further analysis is as follows. (Page 11)

"As shown in Fig. 8, in the low-light modality, the traditional methods performed poorly in all four scenarios. The binary and heat maps of POFT and PCA were complex and lacked the ability to be recognized. The DEC resulted in almost blank binary maps, with fewer high values (i.e. hotter regions) being detected. However, RXD was effective in detecting anomalies, particularly in STD scenarios, and showed suboptimal performance in other scenes compared to the other methods."

"As shown in Fig. 9, in the infrared modality, the four methods had a superior performance than the low-light modality in all four scenarios. It is obvious that the person was easier to recognize in the heat maps of DEC and POFT, but the binary maps of DEC were still empty. The PCA heat maps had less noise. In addition, the RXD exhibited superior performance, especially in the STD, FAN, and OAH scenes, with obvious objects detected in both binary and heat maps, where high anomaly scores (i.e. the hotter regions) were more concentrated on the person to be rescued."

"As shown in Fig. 8, in the low-light modality, the anomaly regions detected by the SPADE algorithm were able to encompass the person to be rescued, particularly in the STD, FAN, and FAH scenes. However, the binary maps of OCR-GAN were complex, with fewer high values (i.e. hotter regions) detected on the ground truth. The binary maps of RIAD and RD4AD did not cover the person, instead the heat maps of RD4AD covered the regions of the anomaly."

"As shown in Figure 9, the deep-learning-based methods achieved better performance in the infrared modality compared to the traditional algorithms proposed above. Compared to the low-light modality, RIAD and OCR-GAN detected more edges of the ground truth in the binary maps. Specifically, SPADE performed poorly in the OAH scenario, detecting misplaced regions (i.e. background) in both binary and heat maps."

5. ***The experimental results could be further analysis and discussed based on the principle of methods. Why is the best performing method better than others? Why do some methods perform poorly? Moreover, the author should analyze the experimental results from the perspective of different types of methods.***

R. Thank you for your suggestions.

We have conducted a more in-depth analysis to explain why the best-performing method outperforms others. This will involve a detailed examination of its unique features and mechanisms that contribute to its superiority. Moreover, we also analyzed the experimental results from the perspective of different types of methods. The specific and detailed analysis is as follows. (Page 11, right column)

"The results presented in Table I demonstrate that, among all algorithms, the RXD algorithm achieved the best performance. PCA and DEC struggled to cope with complex nighttime rescue anomaly detection, resulting in poorer performance. This may be PCA's inability to capture the complex non-linear relationships and DEC's difficulties due to its low-rank distribution assumptions. The POFT algorithm performed well in a few scenarios, as indicated by the AUC metric, where the frequency domain may contain more information and structure to detect. However, traditional detection algorithms generally face challenges of high false positive rates, making it difficult for direct deployment."

"The results in Table I show that the SPADE algorithm based on memory banks outperformed others in most of the recognition tasks, followed by the state-of-the-art knowledge distillation method RD4AD. On the other hand, methods based on GAN, such as OCR-GAN, and those using E-D structures, such as RIAD, showed relatively poorer performance on this dataset."

6. ***For some methods perform noneffective, please show more hyperparameters to prove the algorithm implementation is correct, such as the PCA, DEC, POFT, GAN methods.***

R. Thank you for your suggestion.

All the hyperparameters in fusion methods and anomaly detection methods have listed in the revised paper. The relevant details are as follows. (Page 11, left column)

"For the wavelet transform, the wavelet basis function was the second-order Daubechies wavelet, with the decomposition level set to 1, where the fusion strategy was using the average method for low-frequency and the maximum method for high-frequency. For the deep-learning-based fusion methods, the parameter settings were adjusted to be consistent with the references. For the fusion method proposed, the number of epochs for training was set to 150 with 60 and 90 epochs in the first and second stages, respectively. The batch size was set to 32. For the model hyperparameters settings, the number of GRL blocks in multi-modal feature encoder was 6, with 8 attention heads and 64 channels. The channel of the Lite-transform-based block was also 64 with 8 attention heads. The configuration of the decoder is the same as that of the encoder. As for loss functions Eq. (6), α1 to α3 were set to 0.3, 0.3, and 0.4. As for loss functions Eq. (7), μ was set to 0.5. As for loss functions in Eq. (9), λ_1 to λ_5 were all set to 0.2. For the anomaly detection algorithms, the summary of the percentage of larger eigenvalues is over 99%, the tuning hyperparameter was set to 1 in DEC, and the hyperparameters in RIAD, RD4AD, OCR-GAN and SPADE were consistent with the references."

7. ***Please check the definitions of variables and letters in all formulas, such as ev in (16), X' in (19) and (20).***

R. Thank you for your suggestions.

We are sorry for the undefined symbol in the paper. All the variables and letters have been checked to make sure that the specific variables and letters have specific meanings. 'ev(p)' refers to the weight of p-th possible k values. In the revised manuscript, 'ev(p)' in Eq. (15)-(16) was replaced with 'w_p' to be consistent. The relevant descriptions are as follows.

"For each pixel x_i of the test image, where $i = 1,2,...,n$, the PCA anomaly score can be computed in a typical approach, namely weighted summation of all possible k values. Suppose there are q possible k values in total, the anomaly score c is calculated using the formula in Eq. (15)-(16)"

$$c = \sum_{p=1}^{q} | x_i - x'_i | \cdot w_p \tag{15}$$

$$w_p = \frac{\sum_{j=1}^{p} \lambda_j}{\sum_{j=1}^{n} \lambda_j} \tag{16}$$

"Where w_p is the weight of the p-th possible k values, with p principal eigenvectors selected."

In the revised manuscript, $\mathcal{X}'$ in Eq. (19)-(20) is the normalized image in the frequency domain. The relevant descriptions are as follows.

$$\mathcal{X}'(u,v)=\frac{\mathcal{X}(u,v)}{\mathbf{M}(u,v)},\forall u\in[0,r),\forall v\in[0,c) \qquad (19)$$

“Where Eq. (19) represents the extraction of the phase spectrum, $\mathcal{X}(u,v)$ is the value at position (u,v) after Fourier transform, while $\mathbf{M}(u,v)$ represents the amplitude value at that point and $\mathcal{X}'(u,v)$ is the normalized result. This process can remove the majority of periodic textures from the original image, highlighting the anomalous regions [77].”

$$\mathbf{I}'=F^{-1}(\mathcal{X}') \qquad (20)$$

“Where Eq. (20) is the IDCT process, $F^{-1}(*)$ represents the inverse Fourier transform, $\mathcal{X}'$ is the normalized image in the frequency domain, and $\mathbf{I}'$ refers the output image through phase-only-based methods.”

Finally, thank you again for your very helpful suggestions.

Response to the Comments of the Reviewer #3

This manuscript presents an infrared and low-light multimodal anomaly detection dataset with application to nighttime emergency rescue, and unsupervised anomaly detection algorithms are outlined and experimented with. In addition, they propose a multimodal fusion network based on frequency domain feature decomposition for use as a preprocessing step for multiple anomaly detection algorithms to enhance their detection performance. The article is well-narrated and flows smoothly. However, there are still a few issues that need to be addressed:

1. ***The decoupled high-frequency features and low-frequency features fusion method is not described, please complete it.***

R. Thank you for your suggestions.

We are sorry for the lack of clarity in the initial description of the decoupled high/low-frequency features fusion method. To better describe it, we have split the original subsection "*Fused Feature Decoder*" into two subsections, namely "*Frequency-based Fusion Layer*" and "*Fused Feature Decoder*". The aforementioned fusion method is described in the "*Fused Feature Decoder*" subsection, with detailed fusion procedures and module constructions. Furthermore, the architecture of the proposed dual-stream multi-modal fusion model has been re-illustrated in Fig. 2 for a more profound comprehension. (Page 4, right column)

"The fused feature decoder is comprised of the GRL blocks. The frequency-based fused features $\mathcal{X}_{low}$ and $\mathcal{X}_{high}$ were concatenated and reduced in dimension. They were then fed into the fused feature decoder to generate the final fused image, as given by Eq. (5).

$$\mathbf{I}_{fused} = D(CAT(\mathcal{X}_{low}, \mathcal{X}_{high})) \qquad (5)$$

Where $D(*)$ is the decoder and $\mathbf{I}_{fused}$ is the final fused image. $CAT(*,*)$ is the channel concatenation operation."

2. ***Please specify the meaning of the variables related to equations (10) and (11) in the paper.***

R. Thank you for your suggestions.

All the variables have been checked to make sure that the specific variables have specific meanings. In the revised paper, $L_{VIS_{int}}$ and $L_{VIS_{grad}}$ represent the intensity loss and gradient loss of low-light inputs respectively (similarly for $L_{IR_{int}}$ and $L_{IR_{grad}}$). ∇ represents the Sobel gradient operator, where H and W are the height and width of the image, respectively. (Page 5)

3. ***Is there any connection between the variables between formula (12) and formula (13) in the paper, please explain in detail. And λ is not represented in equation (13).***

R. Thank you for your questions and suggestion.

The connection between the variables in original formula (12) and formula (13) has been expressed in the revised manuscript. The specific details are as follows. (Page 6 and Page 7, left column)

"The RX algorithm is based on two hypotheses H_0 and H_1. Assuming the background and the anomaly target have the same covariance matrix $\mathbf{C}_b$, if the target is present, the mean of the background becomes the feature vector $\mathbf{s}$ of the target. H_0 represents the absence of the target, and the background follows a $N(0,\mathbf{C}_b)$ distribution. Conversely, H_1 represents the presence of a target, and the background follows a $N(\mathbf{s},\mathbf{C}_b)$ distribution."

With regard to the parameter λ, its representation in the text is as follows.

"Given a threshold λ, if a pixel's RX statistic r is larger than the threshold, then the pixel is determined to be an anomalous one, which conforms more to the H_1 hypothesis (the target is present); otherwise, it is a background pixel under H_0 hypothesis (the target is absent)."

4. ***In the subsection Sparse coding reconstruction-based methods, please add the specific meaning of n in "(k << n)".***

R. Thank you for your suggestion.

We are sorry for the undefined variable in the paper. For an abnormal/normal image input, n refers to the number of pixels of the image and "(k<<n)" represents that k is far smaller than n. (Page 7, left column)

5. ***In the statistical model summary "Utilizing statistical models to describe distribution of pixels ofr feature vectors in normal images." in Fig. 3, the word ofr is misspelled, so please check the words in the text carefully.***

R. Thank you for your careful review.

We sincerely apologize for the spelling error in the statistical model summary of Fig. 3. The misspelled word "ofr" has been corrected to "or". In addition, we have also checked all the words in the text to make sure there are no other similar mistakes.

6. ***Please standardize the writing of formulas in the paper, e.g. differentiate between matrices, vectors, scalars, etc.***

R. Thank you for professional suggestion.

In the revised manuscript, we have used bold uppercase letters to represent matrices (e.g., $\mathbf{X}$), bold lowercase letters for vectors (e.g., $\mathbf{x}$), and regular lowercase letters for scalars (e.g., a). This convention is used consistently throughout the paper to make sure the clear distinction between different types of mathematical objects.

7. ***Some references are not formatted properly, e.g. journal abbreviations, please correct them.***

R. Thank you for your careful review and suggestion.

All the references have been thoroughly corrected and properly formatted.

Finally, thank you again for your very helpful suggestions.