Low-Light and Infrared Multimodal Remote Sensing in Nighttime Rescue Mission: A Review of Anomaly Detection Methods

Yuting Wan[#], *Member, IEEE*, Haoyu Yao[#], Jiawei Liu, Chen Sun, Ailong Ma, *Member, IEEE*, Yanfei Zhong*, *Senior Member, IEEE*

***Abstract*—When facing natural disasters like sudden floods at night, due to sudden nature disasters, high timeliness of rescue information and complexity and diversity of rescue environment, it is difficult for ground personnel to perform rescue operations in disaster area in time. Remote sensing UAV technology plays an escalating role in disaster relief due to fast response and high flexibility advantages. However, with low nighttime visibility level, complex post-disaster environment, and numerous obstructions, traditional UAV's visible light remote sensing struggles to achieve accurate rescue detection at night. Therefore, the multimodal detection methods are investigated using low-light and infrared modalities, exploring the integration of data fusion and detection in night rescue applications, and examine the advantages and disadvantages of different anomaly detection methods here. This paper provides the following contributions: 1) a fully-annotated low-light infrared co-observation multimodal remote sensing image dataset for nighttime emergency rescue, termed MRSI-NERD; 2) a benchmark test for most state-of-the-art unsupervised anomaly detection methods to thoroughly explore their capability in extracting useful information from normal samples; and 3) a low-light infrared bimodal fusion method based on frequency domain feature decomposition, which enhances the performance of detectors. The performance of eight types of traditional or deep learning-based detection methods on the MRSI-NERD dataset is reported, including metrics such as ROC-AUC, FPR, TPR, etc. Additionally, a comprehensive analysis of the principles and performance of each category of detection methods is provided.**

***Index Terms*—Low-light and infrared images, multimodal remote sensing, image fusion, anomaly detection, nighttime emergency rescue.**

I. Introduction

NIGHT emergency rescue scenarios pose great challenges for first responders and rescue teams due to the suddenness of natural disasters [1], the high timeliness of rescue information [2], and the complex diversity of rescue environments [3]. Efficient and accurate rescue operations during the night require advanced technologies that can provide real-time situational awareness, aid in the detection and localization of the disaster targets in low-light and complicated environments [4].

With the continuous development of computer vision, object recognition and detection algorithms based on deep neural networks are being applied in emergency rescue scenarios [5], providing a viable solution to address issues such as the limitation of timeliness in disaster rescue [6]. The role of remote sensing unmanned aerial vehicle (UAV) technology equipped with deep convolutional neural networks in rescue and disaster relief operations is increasingly significant [7].

In the context of remote sensing and emergency rescue and search scenarios, low-light remote sensing technology and thermal infrared remote sensing technology are two widely applied and mature image recognition methods in the field of traditional image recognition [8]. They exhibit good complementarity and integration [9]. The fusion of low-light and thermal infrared images has become the forefront of the current discipline's development [10]. However, traditional offline remote sensing image recognition methods face challenges in achieving precise detection of rescue targets under extreme conditions such as low nighttime light intensity and complex post-disaster terrain [11].

Low-light and infrared images fusion is a branch of image fusion that has significant applications in remote sensing image recognition [12]. In the context of nighttime emergency rescue, low-light imagery is highly sensitive to lighting conditions [13], but can capture spatial information, texture details more effectively [14]; Infrared imagery exhibits prominent targets and is less affected by lighting conditions [15], but it typically has lower image resolution and blurry target backgrounds [16]. The fusion of the above two can effectively combine the advantages of both [17], increase the amount of information, and thus improve nighttime rescue operations [18]. The fusion algorithms include traditional fusion algorithms based on spatial or transform domains [19], as well as autoencoder algorithms [20], convolutional neural network algorithms [21], and adversarial neural network algorithms [22] based on deep learning. Although some fusion methods have achieved remarkable performance in recent decades [23], the existing

Y. Wan, C. Sun, A. Ma, and Y. Zhong are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China, and also with the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (E-mail: {wanyuting, sunchen, maailong007, zhongyanfei}@whu.edu.cn).

H. Yao is with the School of Remote Sensing and Information Engineering, Wuhan University, Hubei Province 430072, China (E-mail: yaohaoyu@whu.edu.cn).

J. Liu is with the School of Cyber Science and Engineering, Wuhan University, Hubei Province 430072, China (E-mail: 2021302181179@whu.edu.cn)

([#]Y. Wan and H. Yao contributed equally to this work, *Corresponding author: Y. Zhong, e-mail: zhongyanfei@whu.edu.cn).

algorithms still have limitations when it comes to the specific scenario of nighttime emergency rescue [24].

Anomaly detection refers to distinguishing the identification and localization of anomalies in limited or even absence of prior knowledge [25]. Anomaly detection and target detection based on high-resolution remote sensing images can effectively utilize the existing spectral information in the images to accurately locate rescue targets without the need for a large amount of prior information specifically tailored to the targets [26], [27]. This approach better meets the requirements of high timeliness and accuracy in emergency rescue missions [28]. Unsupervised anomaly detection mainly includes methods such as sample reconstruction, pseudo anomaly augmentation, and knowledge distillation [29]-[31]. Specifically in nighttime emergency rescue scenarios, the complexity of background information [32], the diversity of anomalous samples [33], and the presence of weak target information pose challenges for the deployment and application of existing anomaly detection algorithms [34].

Considering the aforementioned issues, this article established a synchronized low-light and infrared coaxial imaging system and collects a dataset that simulates night-time emergency rescue scenarios. Furthermore, some advanced unsupervised anomaly detection methods were evaluated on the proposed low-light and infrared datasets. This contributes to the development of a night-time remote sensing emergency rescue system designed for low visible light conditions. The main contributions of this paper are summarized as follows:

1) A dataset of multimodal infrared and low-light remote sensing images simulating emergency rescue scenarios. To address the lack of remote sensing image datasets for nighttime emergency rescue scenarios, a highly realistic low-light infrared multimodal nighttime dataset is proposed, termed MRSRI-NERD. It includes various defects such as hypothermia [35], occlusion [36], fallen rescue targets [37], etc., and is fully annotated to simulate real-world scenarios accurately [39]. To the knowledge, this is currently the only available remote sensing multimodal dataset specifically designed for nighttime emergency rescue scenarios.

2) A benchmark for unsupervised anomaly detection models in emergency rescue datasets. In this paper, a survey of unsupervised image segmentation methods in the existing literature was conducted, identifying challenges and outlining future research directions. In the MRSI-NERD dataset, four categories of traditional detectors and four categories of deep detectors were comprehensively evaluated, reporting metrics such as AUC score, FPR, FNR, etc., to provide a benchmark for extensive research on unsupervised nighttime multimodal detection.

3) A dual-stream low-light and infrared modality fusion network based on frequency domain feature decomposition. 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.

II. Construction of Multimodal Dataset for Nighttime Scenarios

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.

Nevertheless, the development of deep and statistical-based image recognition algorithm relies on training on a large amount of valid data [38], but traditional single-modal visible light data is limited by daylight conditions and cannot meet the requirements of nighttime rescue [39]. There is a lack of image data for nighttime emergency rescue scenarios. Thus, in this article, a fully annotated **M**ultimodal **R**emote **S**ensing **I**mage Dataset for **N**ighttime **E**mergency **R**escue **D**etection (**MRSI-NERD**) was constructed, to address the issue of data scarcity, as shown in Fig.1. Basically, a total of 4087 frames of video stream data was collected, where 4044 of them are images containing targets, and the remaining 43 frames are background reference images without anomalies, used for training the anomaly detection model. 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.

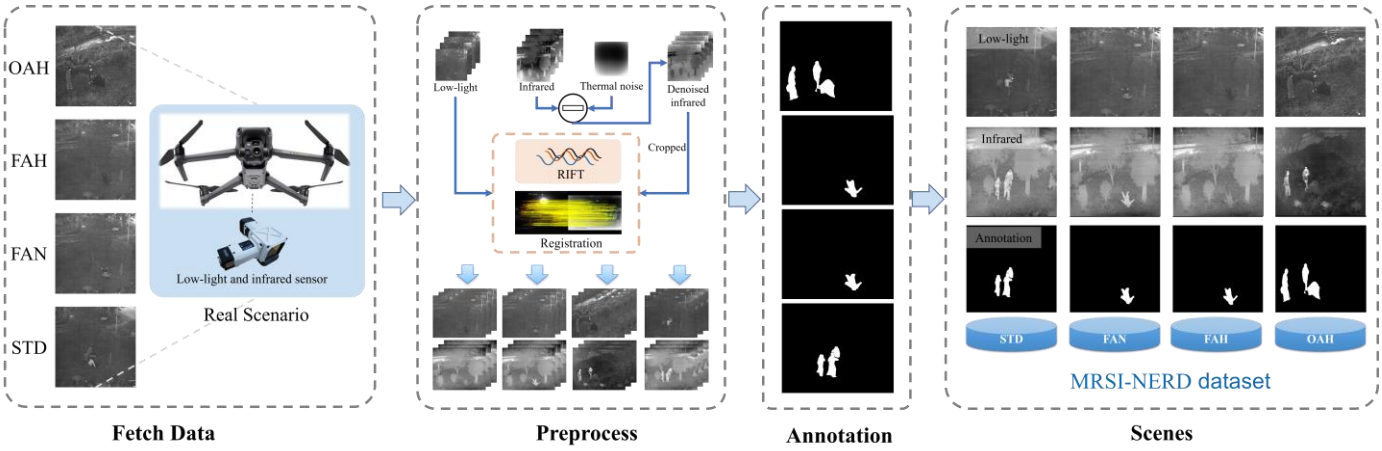

Fig. 1. The construction of the MRSI-NERD dataset. The overall process is comprised of three distinct steps: data fetching, data preprocessing, and annotation. Initially, the original low-light and infrared images were obtained from the device in four different real scenarios. For the preprocessing step, 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.

1) FAH represents the scenario where the rescue target has fallen and is hypothermic. This situation is common in emergency rescue scenarios, where the target is in an urgent state due to hypothermia [40], but at the same time, it interferes with the detection of infrared detectors, posing a certain level of difficulty. Obtaining images of hypothermic individuals is not a viable option, so this condition was simulated using insulating foil boards, which were cropped into the shape of a person [41].

2) FAN refers to the scenario where the rescue target has fallen and cannot stand up but still maintains a normal body temperature. In comparison to FAH, in this scenario, the infrared detector can detect the target, but the low-light detector may not effectively capture information about the target [42]. The FAN rescue scenario is designed to simulate real-life conditions [43].

3) OAH scenario refers to the scenario where the target is obstructed and cannot be detected. In this situation, the infrared detector becomes ineffective due to the thick obstructive material providing thermal insulation. This scenario is commonly encountered in emergency rescue situations, where detecting obstructed rescue targets beyond obstacles is a persistent challenge in the field of emergency rescue. OAH is also widely present in actual emergency rescue scenarios [44].

4) STD is a more general rescue scenario where the target can still stand and has a normal body temperature, showing a significant contrast with the surrounding environment. In this situation, the detector should exhibit high performance, serving as a criterion to verify the effectiveness of the detector.

A low-light and infrared co-calibrated imaging device [45] was applied to capture data from the aforementioned scenarios, which were processed and annotated appropriately. The data processing workflow is illustrated in Fig. 1. Firstly, raw data was captured from the device. The resolution of low-light images are 2048×2048, while the infrared images are 640×512. The original infrared images have thermal noise, which must be eliminated through a process of denoising. By deactivating the infrared sensor in an environment devoid of light exposure, an infrared dark image could be obtained with solely thermal noise. The denoised images were then obtained by subtracting the infrared dark image (thermal noise) from the original infrared image [46]. Afterwards, the resulting infrared images have dark edges that needs to be cropped out. Additionally, although the low-light and infrared images are aligned at the center, they have different fields of view, resulting in varying degrees of distortion at the image edges. Therefore, the RIFT algorithm [47] was employed for registration. Finally, the synchronized low-light and infrared images were obtained after the processing.

III. DUAL-STREAM MULTI-MODAL FUSION MODEL

In this section, a dual-stream low-light and infrared modality fusion network is introduced, based on frequency domain feature decomposition, which leverages the complementary information from visible light and infrared modalities to enhance detection performance [48].

A. Method Overview

The model contains four components, namely:

1) An Encoder-Decoder module for high-fidelity reconstruct -ion.

2) A Low-frequency extractor for extracting global features.

3) A High-frequency extractor for extracting local detailed features.

4) A dual-stream feature fusion module.

The overall workflow is illustrated in Fig. 2, the network achieves excellent performance on the low-light and infrared fusion task.

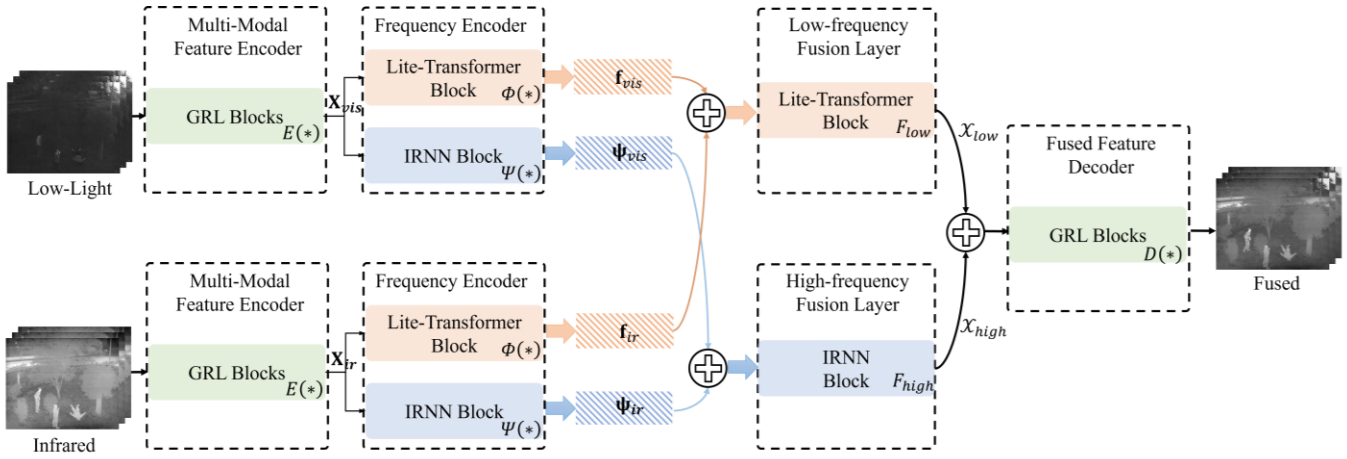

Fig. 2. The architecture of the proposed dual-stream multi-modal fusion model based on frequency domain feature decomposition

B. *Multi-Modal Feature Encoder*

For the low-light and infrared inputs, a shared feature backbone encoder was applied to help the subsequent frequency domain feature extractor for further processing. Initially, the input pair of low-light and infrared images $\mathbf{I}_{vis}$ and $\mathbf{I}_{ir}$ were fed into the encoder respectively, as represented by the Eq. (1).

$$\mathbf{X}_{vis} = E(\mathbf{I}_{vis}), \mathbf{X}_{ir} = E(\mathbf{I}_{ir}) \tag{1}$$

Where $E(*)$ is the encoder, $\mathbf{X}_{vis}$ and $\mathbf{X}_{ir}$ are the shared features extracted from low-light images and infrared images, respectively.

For the encoder, the GRL [49] block, a state-of-the-art method adopted in the field of image restoration, is selected. The reason is that GRL can efficiently and explicitly extract high-resolution image features and perform restoration, which is beneficial for improving the quality of image fusion [50].

C. *Frequency-based Feature Extractor*

Then, the aforementioned shallow features $\mathbf{X}_{vis}$, $\mathbf{X}_{ir}$ were fed into a low-frequency extractor and a high-frequency extractor respectively, obtaining the corresponding frequency domain features, represented as Eq. (2)-(3).

$$\mathbf{f}_{vis} = \Phi(\mathbf{X}_{vis}), \boldsymbol{\psi}_{vis} = \Psi(\mathbf{X}_{vis}) \tag{2}$$

$$\mathbf{f}_{ir} = \Phi(\mathbf{X}_{ir}), \boldsymbol{\psi}_{ir} = \Psi(\mathbf{X}_{ir}) \tag{3}$$

Where $\Phi(*)$ represents the low-frequency extractor and $\Psi(*)$ represents the high-frequency extractor. $\mathbf{f}_{vis}$ and $\mathbf{f}_{ir}$ are the extracted low-frequency features of low-light and infrared inputs, respectively. Similarly, $\boldsymbol{\psi}_{vis}$ and $\boldsymbol{\psi}_{ir}$ are the extracted high-frequency features of low-light and infrared inputs, respectively.

A lite-transformer-based low-frequency extractor was applied because it can improve the speed of inference [51]. The lightweight transformer module allows it to obtain more long-range information [52], corresponding to high-frequency information [53], while its ability to extract short-range information is reduced. This symbolizes weaker low-frequency information extraction capability. For the high-frequency extractor, an Invertible Residual Neural Network [54] was used. The convolutional structure of CNN is effective in extracting short-range information [55], corresponding to high-frequency components [56]. Invertible neural networks have the theoretical capability of lossless information compression [57], enabling the preservation of high-frequency information

D. *Frequency-based Fusion Layer*

The low-frequency features $\mathbf{f}_{vis}$ and $\mathbf{f}_{ir}$ were firstly concatenated and then fed into the low-frequency fusion layer, while the high-frequency features $\boldsymbol{\psi}_{vis}$ and $\boldsymbol{\psi}_{ir}$ underwent the same operation but were fed into the high-frequency fusion layer. This procession can be described with Eq. (4).

$$\begin{aligned} \mathcal{X}_{low} &= F_{low}(CAT(\mathbf{f}_{vis}, \mathbf{f}_{ir})) \\ \mathcal{X}_{high} &= F_{high}(CAT(\boldsymbol{\psi}_{vis}, \boldsymbol{\psi}_{ir})) \end{aligned} \tag{4}$$

Where $F_{low}(*)$ is the low-frequency fusion layer and $F_{high}(*)$ is the high-frequency fusion layer. $\mathcal{X}_{low}$ is the fused low-frequency feature and $\mathcal{X}_{high}$ is the fused high-frequency feature. $CAT(*,*)$ is the channel concatenation operation.

The low-frequency fusion layer consists of the Lite-transformer-based block, while the high-frequency fusion layer consists of the IRNN block, which aims to fuse low/high frequency features, respectively.

E. *Fused Feature Decoder*

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).

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

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

F. Loss Function

To achieve a high-fidelity output that combines features from both modalities, the network was trained by two stages.

In the first stage, the network's ability was trained to reconstruct images for each modality. This is attained by concatenating high-frequency and low-frequency features and then feeding them into the GRL decoder. The loss function is described by Eq. (6)

$$L_1 = \alpha_1 L_{vis} + \alpha_2 L_{ir} + \alpha_3 L_{indep} \quad \text{st. } \alpha_1 + \alpha_2 + \alpha_3 = 1 \tag{6}$$

Where the L_{vis} and L_{ir} are the reconstruction errors for low-light and infrared images, respectively. L_{indep} represents the loss that quantifies the independence of low-frequency representations from high-frequency features. α_1, α_2 and α_3 are the tuning parameters, with the constraint condition of $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

For the reconstruction loss L_{vis}, which is similar to L_{ir} in calculation, it is conducted by MSE [59], namely Eq. (7). Among them, $\bar{\mathbf{I}}_{vis}$ is the reconstructed low-light image and $\mathbf{I}_{vis}$ is the original image, consistent with the previous definition mentioned above, where μ is the tuning parameter.

$$L_{vis} = \| \bar{\mathbf{I}}_{vis} - \mathbf{I}_{vis} \|_2^2 + (1-\mu) MSE(\bar{\mathbf{I}}_{vis}, \mathbf{I}_{vis}) \tag{7}$$

For the independence loss, assuming each pixel of image $\mathbf{I}_{vis}$ is denoted by x_i, and each pixel of infrared image $\mathbf{I}_{ir}$ is denoted by y_i, the calculation of L_{indep} can be given in Eq. (8). It specifies that the independence loss is computed using the cosine distance [58].

$$L_{indep} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2}} \tag{8}$$

In the second stage, the network was trained for multi-modal image fusion and feature extraction. The paired low-light and infrared images $\{\mathbf{I}_{vis}, \mathbf{I}_{ir}\}$ were fed into a well-trained GRL encoder to obtain the shallow features $\{\mathbf{X}_{vis}, \mathbf{X}_{ir}\}$. Then the low-frequency extractor and the high-frequency extractor were applied to obtain the decoupled features $\{\mathbf{f}_{vis}, \boldsymbol{\psi}_{vis}, \mathbf{f}_{ir}, \boldsymbol{\psi}_{ir}\}$ respectively. Through the fusion layer, $\{\mathbf{f}_{mi}, \mathbf{f}_{ir}\}$ and $\{\boldsymbol{\psi}_{mi}, \boldsymbol{\psi}_{ir}\}$ were combined, obtaining the aggregated low-frequency feature $\mathcal{X}_{low}$ and high-frequency feature $\mathcal{X}_{high}$, these combined features were then fed into the GRL decoder to produce the final fused image $\mathbf{I}_{fused}$.

During this stage, the loss function is defined as Eq. (9), which is comprised of intensity loss, gradient loss and independent loss.

$$L_2 = \lambda_1 L_{VIS_{int}} + \lambda_2 L_{IR_{int}} + \lambda_3 L_{VIS_{grad}} + \lambda_4 L_{IR_{grad}} + \lambda_5 L_{indep} \tag{9}$$

The intensity loss $L_{VIS_{int}}$, which is similar to $L_{IR_{int}}$ in calculation, is described in Eq. (10).

$$L_{VIS_{int}} = \frac{1}{HW} \| \mathbf{I}_{fused} - \mathbf{I}_{vis} \|_2^2 \tag{10}$$

Where H and W are the height and width of the image, respectively.

The gradient loss $L_{VIS_{grad}}$, which is similar to $L_{IR_{grad}}$ in calculation, is described in Eq. (11).

$$L_{VIS_{grad}} = \frac{1}{HW} \| \nabla \mathbf{I}_{fused} - \nabla \mathbf{I}_{vis} \|_2^2 \tag{11}$$

Where ∇ represents the Sobel gradient operator. H and W are the height and width of the image, respectively.

IV. Anomaly Detection Methods in Emergency Rescue Situations

The typical goal of anomaly detection is to identify anomalies that differ from normal samples through unsupervised or semi-supervised methods, performing binary classification or localizing tasks. In recent years, there has been significant progress in image anomaly detection techniques, yet their application in the field of emergency rescue remains unexplored. In this section, unsupervised image anomaly detection methods are reviewed and categorized into traditional methods and deep-learning-based methods, distinguishing whether they incorporate Deep Neural Networks (DNNs). The overview of the image anomaly detection methods is shown in Fig. 3.

A. Anomaly Detection Methods Based on Traditional Approaches

In this section, traditional image anomaly detection methods are categorized into the following types: statistical model-based, sparse coding reconstruction-based, decomposition-based, and frequency domain analysis-based.

1) Statistical model-based methods

Statistical model-based methods typically utilize statistical models to describe the distribution of pixels or feature vectors in normal images. Regions in images that deviate from this dist ribution are considered anomalies. These methods typically rep resent the normal parts of images using Gaussian distributions in the spatial or frequency domain, detecting abnormal regions based on outliers. Representative works include RX (Reed-xia o) [60], Gaussian Expectation-Maximization [61], and Markov Random Field Gaussian Model [62] and so on.

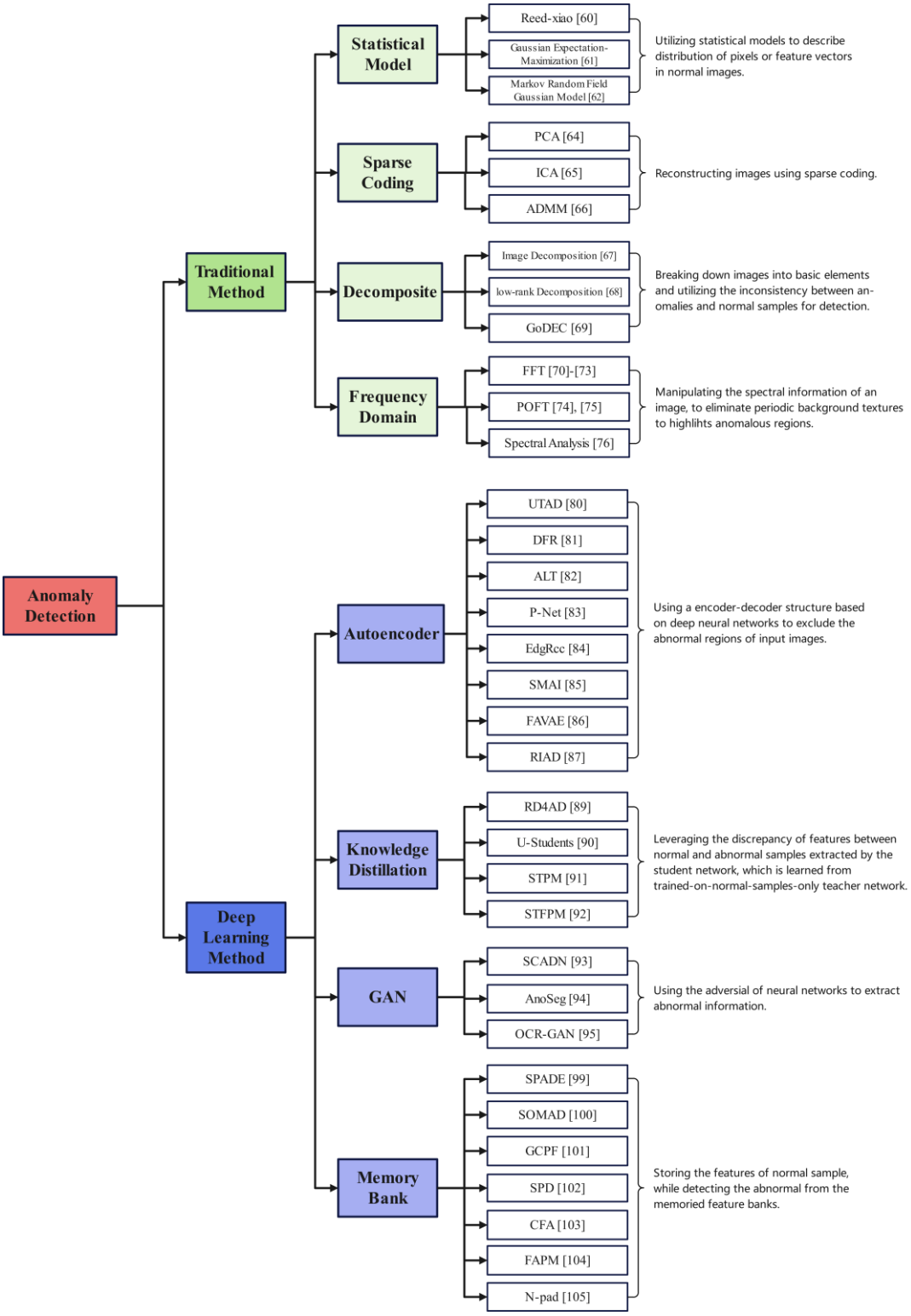

Fig. 3. Anomaly detection methods overview.

RX (Reed-Xiao) algorithm was selected as the representative method for detailed analyses. 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 a target, and at this time, the background follows a $N(0,\mathbf{C}_b)$ distribution; H_1 represents the presence of a target, and the background follows a $N(\mathbf{s},\mathbf{C}_b)$ distribution.

For the input image that consist of P spectral bands, the background containing N pixels can be represented as a $P*N$ matrix $\mathbf{X}_b=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N]$,where $\mathbf{x}_i=[x_1,x_2,...,x_P]^T$. Under the H_0 hypothesis, the mean vector $\boldsymbol{\mu}_b$ and covariance matrix $\mathbf{C}_b$ of the background is estimated by Eq. (12).

$$\boldsymbol{\mu}_b = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i,\ \mathbf{C}_b = \frac{1}{N}\mathbf{X}_b\mathbf{X}_b^T \tag{12}$$

For each pixel $\mathbf{x}_i$ with P spectral bands, its RX statistic r is calculated in Eq. (13), which is equivalent to Mahalanobis Distance [63].

$$r = (\mathbf{x}_i - \boldsymbol{\mu}_b)^T \mathbf{C}_b^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_b) \tag{13}$$

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).

2) Sparse coding reconstruction-based methods

Methods of this kind typically involve reconstructing images using sparse coding. In this process, a dictionary is learned to represent normal images. Subsequently, during the testing phase, anomaly detection is performed based on aspects such as reconstruction differences and sparsity. This includes methods like PCA [64], ICA [65], and Alternating Direction Method of Multipliers (ADMM) [66] etc.

Analyzing with the classical PCA algorithm as a representative. Suppose there are m normal images, and each image is flattened into a vector $\mathbf{x} = [x_1, x_2, ..., x_n]^T$ with n pixels, then a $m*n$ matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m]$ can be conducted.

For the matrix $\mathbf{X}$, the covariance matrix $\mathbf{C}$ can be computed firstly. The eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$ and corresponding eigenvectors $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n$ of $\mathbf{C}$ are then computed. The top k principal eigenvectors are selected with the larger eigenvalues and are rearranged as the projection matrix $\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_k]$, where k is far smaller than n.

In the testing phase, the flattened vector $\mathbf{x}$ of test image can be mapped to a k-dimensional vector $\mathbf{y}$ in the principal component space, which is calculated by $\mathbf{y} = \mathbf{P}^T\mathbf{x}$. Then the reconstruction vector $\mathbf{x}'$ is computed by Eq. (14).

$$\mathbf{x}' = \mathbf{P}\mathbf{y} \tag{14}$$

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.

3) Image decomposition-based methods

Methods based on image decomposition are mostly designed for the detection of small anomalous areas on surfaces with periodic textures. Since anomalous regions typically appear randomly and have weaker periodicity, this characteristic allows them to be distinguished from backgrounds with strong periodic textures. Classical methods include hyperspectral image decomposition [67], low-rank decomposition [68], GoDec [69], etc.

DEC represents classical low-rank decomposition methods. For the input image, low-rank decomposition is employed to decompose the original test image into a low-rank matrix representing the background and a sparse matrix representing the anomalous regions. It is represented by Eq. (17).

$$\min_{\mathbf{L},\mathbf{S}} \| \mathbf{L} \|_* + \lambda \| \mathbf{S} \|_1,\ s.t.\ \mathbf{F} = \mathbf{L} + \mathbf{S} \tag{17}$$

Where $\mathbf{F}$ represents original image matrix, $\mathbf{L}$ and $\mathbf{S}$ represent low-rank matrix and sparse matrix, respectively. $\| \cdot \|_*$ represents the rank of matrix, and $\| \cdot \|_1$ represents the L1 norm to approximately describe the sparsity of the matrix. λ is the tuning hyperparameter.

4) Frequency domain-based methods

Frequency domain-based methods mainly manipulate the spectral information of an image, attempting to eliminate periodic background textures to highlight anomalous regions. This method can be categorized into two domains: background spectrum elimination and Phase-Only Fourier Transform. The former aims to eliminate periodic background spectral information to highlight anomalous regions, while the latter separates repetitive background and anomalous regions by extracting only the phase spectrum.

For the former, most existing works [70]-[73] first utilize Fourier transform to convert the original image into the frequency domain. After removing the spectral components corresponding to periodic background textures in the magnitude spectrum, the inverse Fourier transform is applied to obtain the spatial information of anomalous regions.

For the phase-only-based methods [74]-[76], the amplitude spectrum is discarded, with only the phase spectrum being subjected to IDCT, as shown in Eq (18)-(20).

$$\boldsymbol{\mathcal{X}} = F(\mathbf{I}) \tag{18}$$

Where Eq. (18) denotes the DCT processing, $F(*)$ is the Fourier transform, $\mathbf{I}$ refers the original image, and $\boldsymbol{\mathcal{X}}$ is the image in the frequency domain with r rows and c columns.

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

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

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

Where Eq. (20) is the IDCT process, $F^{-1}(*)$ represents the inverse Fourier transform, $\boldsymbol{\mathcal{X}}'$ is the normalized image in the

frequency domain, and $\mathbf{I}'$ refers the output image through phase-only-based methods.

B. Anomaly Detection Methods Based on Deep Learning

Deep learning-based methods do not rely on manually desig ned features; instead, they leverage deep neural networks to aut omatically learn extracting discriminative features, providing t he algorithm with higher generalization capability. The existin g methods can be broadly categorized into four types, namely Autoencoder-based, Knowledge Distillation-based, GAN-base d, and Memory bank-based approaches.

1) Autoencoder-based methods

These methods typically utilize the architecture of Autoencoder (AE) [78] or Variational Autoencoder (VAE) [79] to extract features. For an input image $\mathbf{X}$, the process begins by obtaining the latent code $\mathbf{X}'$ of $\mathbf{X}$ through an encoder. Subsequently, $\mathbf{X}'$ is decoded by a decoder to produce $\mathbf{Y}$. Due to the anomaly-free nature of the training process for the encoder-decoder, it can only learn features from normal samples and cannot reconstruct anomalous regions. Therefore, the difference between $\mathbf{Y}$ and $\mathbf{X}$ can serve as an indicator for detecting anomalies. Representative works include: UTAD [80], DFR [81], ALT [82], P-Net [83], EdgRec [84], SMAI [85], FAVAE [86], RIAD [87].

For detailed analysis, RIAD was selected, as shown in Fig. 4. The input image is uniformly divided into n patches of size $k*k$. n patches are randomly masked out, and this process is repeated n times to obtain n masked images. It is ensured that the intersection of masked patches in each image is empty. During the model processing, the covered portions of the n masked images are reconstructed, and the reconstructed components are then assembled to obtain a complete reconstructed image.

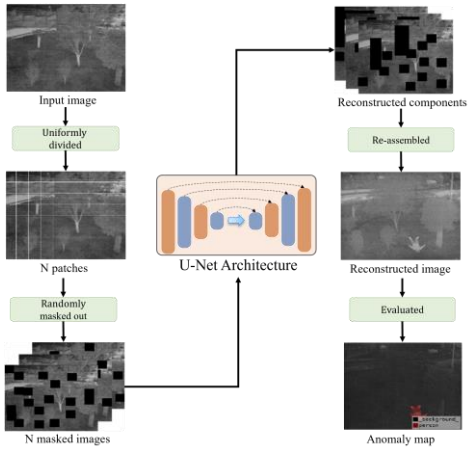

Fig. 4. Pipeline of RIAD. Initially, the input image is divided into n patches of equal size. These patches are then subjected to random masking, a process that is repeated n times to yield n masked images. It is imperative to ensure that the intersection of masked patches in each image is devoid of any content. Subsequently, the covered portions of the n masked images are reconstructed by a U-Net architecture model, and the reconstructed components are then assembled to obtain a complete reconstructed image. Finally, following evaluation of the results, an anomaly map is generated.

2) Knowledge Distillation-based methods

Knowledge distillation aims to transfer the knowledge contained in the pre-trained Teacher-model to the smaller Student-model, thereby achieving model compression [88]. In anomaly detection, knowledge distillation is applied to constrain the generalization capability of the encoder-decoder, disrupting the reconstruction of anomalous samples, thereby magnifying the distance between normal and anomalous samples in evaluation metrics [89]-[92].

To replicate the experiments on the MRSI-NERD dataset, the current state-of-the-art reverse knowledge distillation method, namely RD4AD, was selected. The process is illustrated in Fig. 5. For the input image, channel normalization was firstly performed and then fed into the model. The model consists of a fixed, extensively pre-trained teacher encoder (T-Encoder), a trainable student decoder (D-Decoder), and a one-class embedding module (OCBE). The input image is encoded by T-Encoder, undergoes redundancy removal through the OCBE module, and is then decoded by S-Decoder with the expectation of restoring the normal image information. The similarity loss between Encoder and Decoder is calculated, and the optimization is performed by backpropagating gradients to update Decoder and OCBE.

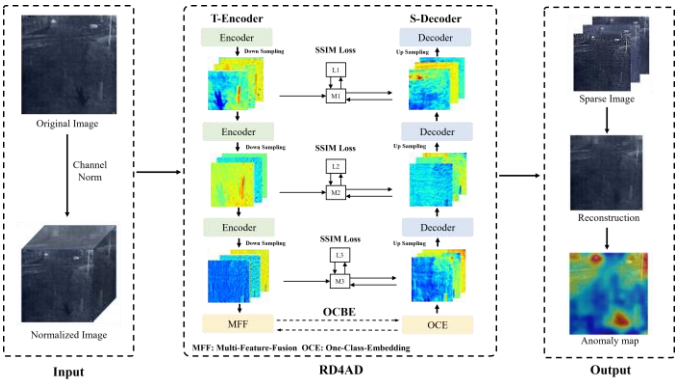

Fig. 5. Pipeline of RD4AD. The model itself consists of three components: a pre-trained teacher encoder (T-Encoder), a trainable student decoder (D-Decoder), and a one-class embedding (OCBE) module. The OCBE module includes Multi-Feature-Fusion module (MFF) and the One-class-embedding module (OCE). And the SSIM Loss in the illustration refers to the similarity loss.

3) GAN-based methods

The application of Generative Adversarial Networks (GANs) in image generation is widespread. Some efforts have been made to constrain the generator to learn features only from normal samples, preventing it from generalizing to anomalous samples. There are various works on GAN-based anomaly detection [93]-[95], and here OCR-GAN was selected as a representative example [95], as shown in Fig. 6.

For the generator, frequency decoupling is performed to decompose input images at different frequencies. This step is achieved through pre-designed Gaussian blurring. The decoupled inputs are sequentially fed into generators corresponding to different frequencies through a Channel Selection module to reconstruct the respective images. The

final output is obtained by accumulating these reconstructions.

For the discriminator, it takes the original image $\mathbf{X}_1$ and the reconstructed image $\mathbf{X}_2$ as input $\mathbf{X}_1$ serves as negative inputs, representing normal samples, while $\mathbf{X}_2$ and the augmented samples of $\mathbf{X}_1$ serve as positive inputs, representing abnormal samples. During the training phase, only $\mathbf{X}_1$ and its augmented version (referred to as CutPaste [96] and CutOut [97] in the original text) are used as positive and negative samples, respectively. This is done to learn the distribution of normal images and train the discriminator. During the testing phase, the discriminator produces binary classification labels (normal/abnormal), and the location of anomalies is identified through the reconstruction error.

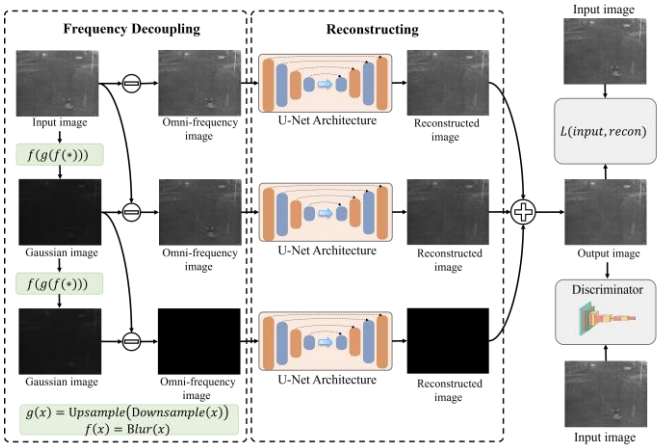

Fig. 6. Pipeline of OCR-GAN. OCR-GAN comprises two distinct steps: frequency decoupling and reconstruction. In the frequency decoupling step, the input image is decomposed at different frequencies through pre-designed Gaussian blurring. In the reconstruction step, through a Channel Selection module, the decoupled inputs with different frequencies are then fed sequentially into a corresponding U-Net architecture generator. The generators are used to reconstruct the respective images and the final output is obtained by accumulating these reconstructions. In the training phase, the discriminator processes the original image and its augmented version to learn the distribution of normal images. In the testing phase, the discriminator produces binary classification labels (normal/abnormal), and the location of anomalies is identified through the reconstruction loss.

4) Memory bank-based methods

To address the issue where neural networks can reconstruct unseen anomalous regions, Memory bank-based methods store features of normal samples as prototypes. During decoding, the decoder indexes the stored features for reconstruction, thereby limiting the CNN's generalization ability to unseen features. These field includes many classic works, such as PatchCore [98], SPADE [99], SOMAD [100], GCPF [101], SPD [102], CFA [103], FAPM [104], and N-pad [105].

Selecting SPADE as a general representative, as shown in Fig.7, this method extends the KNN anomaly detection algorithm from image-level to pixel-level. During the training phase, a ResNet feature extractor pre-trained on the ImageNet dataset is used to extract features from normal images, and a feature bank is maintained. While during the inference phase, firstly, image-level KNN matching is performed to distinguish whether an image is normal or abnormal. For abnormal images, multi-scale Image Alignment is introduced for pixel-level anomaly localization. For a normal sample that is misclassified as abnormal, the pixel-level localization should output a mask of all zeros.

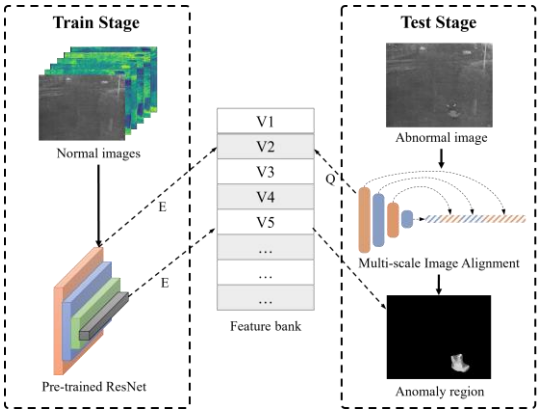

Fig. 7. Pipeline of SPADE. In the training stage, SPADE employs a pre-trained ResNet to extract features from normal images. The features are subsequently stored in a feature bank. During the test stage, a multi-scale Image Alignment is utilized for anomaly localization. This involves querying the features in the feature bank to generate an anomaly map. In the illustration, "E" denotes "Extract features" and "Q" signifies "Query".

V. Experiments and Analysis

A. Experimental Settings

Datasets. In this paper, the dataset was organized following the format of MVTec, where each learning-based algorithm was trained on the corresponding target-free sample scene classification. There are a total of four different scenes for each modal, each containing 1 to 3 targets. To validate the effectiveness of the fusion method, the fused data was synthesized using the fusion algorithm proposed and conducted the same validation analysis. The results are presented in the following section.

Metrics. For the evaluation of fusion, EN, SD, SF, MI, SCD and SSIM were chosen as comprehensive metrics.

Suppose the fused image is $\mathbf{F}$, and the source images are $\mathbf{A}$ and $\mathbf{B}$ respectively. And for an image $\mathbf{I}$ with size $M \times N$, assuming there are L gray levels, μ_I presents the mean value of the image.

Entropy (EN) is a metric used to measure the amount of information in an image. The formula for EN can be presented by the Eq. (21), where $p(i)$ represents the probability of a pixel with gray value i appearing.

$$EN = -\sum_{i=1}^{L} p(i)\log_2 p(i) \tag{21}$$

Standard deviation (SD) reflects the dispersion degree of pixel gray values in an image relative to the mean value. The formula for SD can be expressed by the Eq. (22).

$$SD = \sqrt{\frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}(\mathbf{I}(i,j) - \mu_I)^2} \tag{22}$$

Spatial frequency (SF) refers to the rate of change of the greyscale of an image. In general, the greater the spatial frequency, the higher quality of the fused image.

SF includes two further elements, namely row frequency (RF) and column frequency (CF). These are described by the following equation Eq. (23), where RF and CF are calculated by the following equations Eq. (24) and Eq. (25) respectively.

$$SF = \sqrt{RF^2 + CF^2} \tag{23}$$

$$RF = \sqrt{\frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}(\mathbf{I}(i,j) - \mathbf{I}(i,j-1))^2} \tag{24}$$

$$CF = \sqrt{\frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}(\mathbf{I}(i,j) - \mathbf{I}(i-1,j))^2} \tag{25}$$

Mutual information (MI) is used to measure the amount of information transferred from the source image to the fused image. The higher the MI of the fused image, the more information is transferred from the source to the fused.

The total MI can be presented in Eq. (26). For the source image $\mathbf{A}$, the MI between the source $\mathbf{A}$ and the fused image $\mathbf{F}$ is described as Eq. (27), where h_A and h_F denotes the edge histogram of them, and $h_{A,F}$ denotes the joint histogram of the source and fused images.

$$MI = MI_{A,F} + MI_{B,F} \tag{26}$$

$$MI(A,F) = \sum_i \sum_j h_{A,F}(i,j)\log_2 \frac{h_{A,F}(i,j)}{h_A(i)h_F(j)} \tag{27}$$

Sum of correlation differences (SCD) is a measure of the discrepancy between the fused image and the source image, which is used to evaluate the efficacy of the fusion algorithm. It can be observed that an elevated SCD value indicates a greater degree of information content present in the source image.

The formula for SCD can be described by the Eq. (28), where $D_{A,F}$ denotes the difference between the fused image $\mathbf{F}$ and the source image $\mathbf{A}$ (similarly for $D_{B,F}$). For any two images $\mathbf{X}$ and $\mathbf{Y}$, $r(X,Y)$ is presented in Eq. (29).

$$SCD = r(A, D_{A,F}) + r(B, D_{B,F}) \tag{28}$$

$$r(X,Y) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(\mathbf{X}(i,j) - \mu_X)(\mathbf{Y}(i,j) - \mu_Y)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(\mathbf{X}(i,j) - \mu_X)^2 \sum_{i=1}^{M}\sum_{j=1}^{N}(\mathbf{Y}(i,j) - \mu_Y)^2}} \tag{29}$$

Structural similarity index measure (SSIM) is employed to quantify the information loss and distortion during the fusion process. A higher SSIM value indicates a greater structural similarity between the two images, implying a lower information loss and distortion during fusion.

The total SSIM can be presented in Eq. (30), and the SSIM between two images $\mathbf{X}$ and $\mathbf{Y}$ is described as Eq. (31). σ_X presents the standard deviation of image X (similarly for σ_F), where $C_1 = (0.01 \times L)^2$ and $C_2 = (0.03 \times L)^2$.

$$SSIM = SSIM_{A,F} + SSIM_{B,F} \tag{30}$$

$$SSIM_{X,F} = \frac{(2\mu_X\mu_F + C_1)(2\sigma_{XF} + C_1)}{(\mu_X{}^2 + \mu_F{}^2 + C_1)(\sigma_X{}^2 + \sigma_F^2 + C_2)} \tag{31}$$

For the testing of anomaly detection, the settings of SPADE [106] were followed, adopting three popular metrics to get a comprehensive performance evaluation of all benchmarks, including traditional and deep learning-based methods, namely FPR (false positive rate), TPR (true positive rate) and AUC (area under the curve). Negative denotes the background pixels in the image without target objects.

FPR in anomaly detection measures the proportion of normal data points that the model incorrectly identifies as anomalies, representing the rate of false alarms in a model.

It is described in Eq. (32), where FP (False Positive) is the number of negative cases that are incorrectly predicted as positive, and TP (True Negative) is the number of negative cases that are correctly predicted as negative.

$$FPR = \frac{FP}{FP + TN} \tag{32}$$

TPR is a crucial metric that measures the proportion of actual anomalies that the detection algorithm correctly identifies, reflecting the ability of a model to identify the positive instances accurately.

It is presented in Eq. (32), where TP (True Positive) is the number of positive cases that are correctly predicted as positive, and FP (False Negative) is the number of positive cases that are incorrectly predicted as negative.

$$TPR = \frac{TP}{TP + FN} \tag{33}$$

AUC provides a comprehensive evaluation of the performance of an anomaly detection method across different decision thresholds. The Receiver Operating Characteristic (ROC) curve is plotted with TPR on the y - axis and FPR on the x-axis for various threshold settings. The AUC is the area under this ROC curve.

Tested approaches. For image fusion, the proposed method was compared with a wavelet transform and the state-of-the-art deep-learning-based methods including U2Fusion [107], SDNet [108], TarDAL [109] and DeFusion [110]. For traditional anomaly detection algorithms, the following were tested (1) RXD as a Statistical model-based method; (2) PCA as a Sparse coding reconstruction-based method; (3) DEC as an Image decomposition-based method; (4) POFT as a Frequency domain-based method. For deep-learning-based anomaly detection algorithms, the following were tested (1) RIAD as Autoencoder-based methods; (2) RD4AD as a Knowledge Distillation-based method; (3) OCR-GAN as a GAN-based method; (4) SPADE as a Memory bank-based method. All parameters were carefully tuned to achieve the best results.

Parameter settings. 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.

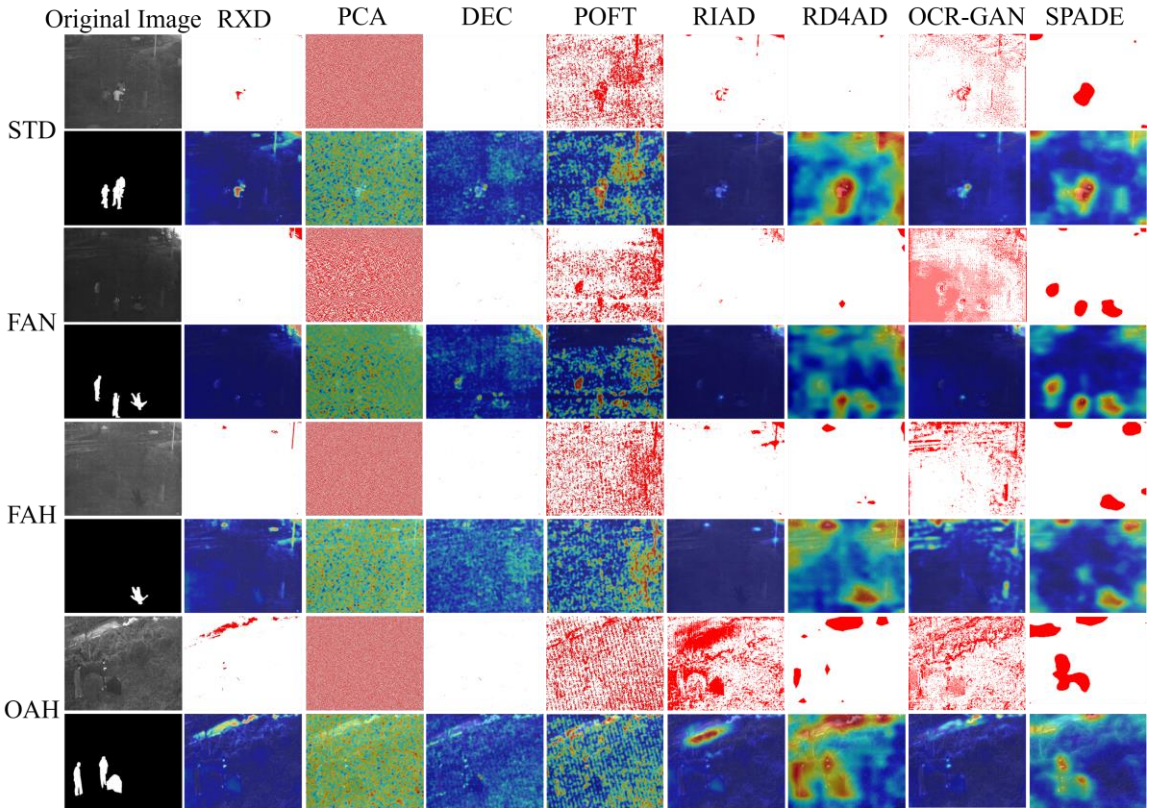

Fig. 8. Performances of the traditional and deep-learning detection algorithms on the Low-Light dataset.

B. Comparison of the Traditional Methods

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.

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.

C. Comparison of the Deep-Learning-Based Methods

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.

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.

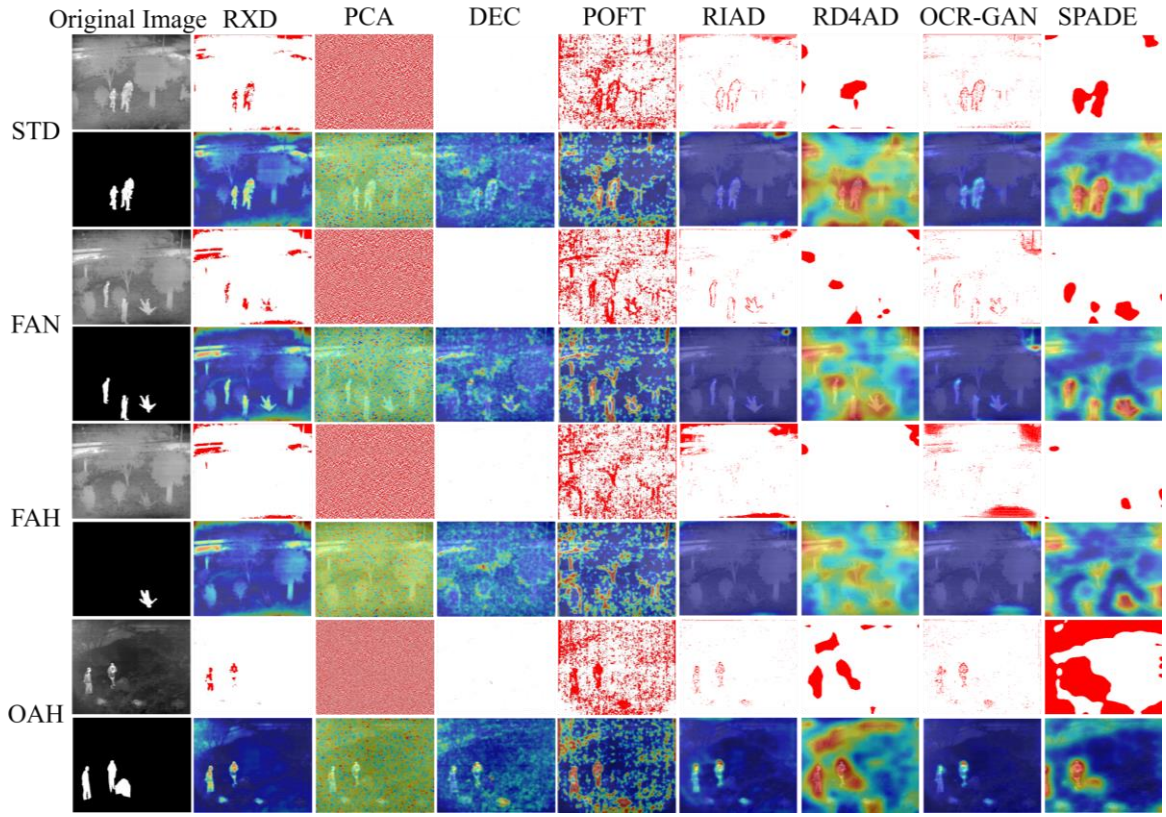

Fig. 9. Performances of the traditional and deep-learning detection algorithms on the Infrared dataset.

D. Performance of Fusion-Based Detections

The fusion results of the low-light and the infrared images are as shown in Table II. 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.

To further evaluate the performance of the fused images, downstream detection tasks were conducted on the fused dataset. RXD, which performed best among traditional algorithms, and RD4AD and SPADE, which exhibited excellent performance among deep learning algorithms, were selected. The results of the experiment are shown in Fig.10 and Table III.

As illustrated in Fig.10, the heat maps demonstrates that, in comparison with the single-modality-based, 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.

As demonstrated in Table III, the RXD and RD4AD algorithms, respectively, exhibit the most optimal performance under the STD and OAH scenarios on the fused dataset, whereas SPADE demonstrates superior performance under the FAN and FAH scenarios.

TABLE I
RESULTS OF TRADITIONAL METHODS AND DEEP-LEARNING METHODS IN DETECTION

Metric	Modality	Scene	RXD	PCA	DEC	POFT	RIAD	RD4AD	OCR-GAN	SPADE
FPR↓	Low-Light	STD	0.981	**0.499**	0.990	0.630	0.813	0.881	0.871	0.742
	Low-Light	FAN	0.978	0.500	0.990	0.487	0.983	0.987	0.342	**0.150**
	Low-Light	FAH	1.000	0.497	1.000	0.569	1.000	0.701	1.000	**0.101**
	Low-Light	OAH	0.997	0.499	0.990	0.703	**0.492**	0.801	0.937	0.980
	Infrared	STD	0.895	0.895	1.000	0.630	0.652	0.881	0.920	**0.200**
	Infrared	FAN	0.610	0.517	1.000	0.388	0.849	0.824	0.840	**0.132**
	Infrared	FAH	1.000	0.498	1.000	0.598	0.997	0.999	0.990	**0.144**
	Infrared	OAH	0.980	**0.501**	0.990	0.700	0.970	0.815	0.959	0.845
TPR↑	Low-Light	STD	**0.997**	0.500	0.990	0.710	0.975	0.979	0.995	0.954
	Low-Light	FAN	**0.995**	0.499	0.990	0.759	0.990	0.571	0.624	0.891
	Low-Light	FAH	**0.995**	0.500	0.990	0.683	0.939	0.897	0.988	0.903
	Low-Light	OAH	0.983	0.499	0.990	0.661	0.789	0.421	0.929	**0.993**
	Infrared	STD	0.964	0.499	**0.990**	0.753	0.980	0.437	0.987	0.831
	Infrared	FAN	0.940	0.498	**0.990**	0.757	0.948	0.398	0.971	0.912
	Infrared	FAH	0.945	0.498	**0.990**	0.713	0.938	0.937	0.965	0.843
	Infrared	OAH	**0.990**	0.500	**0.990**	0.698	0.985	0.976	0.959	0.947
AUC↑	Low-Light	STD	0.454	0.500	0.493	0.541	0.660	0.643	**0.684**	0.680
	Low-Light	FAN	0.837	0.500	0.559	0.636	0.539	0.408	0.666	**0.931**
	Low-Light	FAH	0.826	0.502	0.491	0.557	0.526	0.744	0.368	**0.931**
	Low-Light	OAH	0.575	0.499	0.486	0.478	0.687	**0.867**	0.527	0.552
	Infrared	STD	0.562	0.499	0.509	0.562	0.699	**0.894**	0.569	0.715
	Infrared	FAN	0.829	0.489	0.555	0.683	0.640	**0.893**	0.673	**0.968**
	Infrared	FAH	0.285	0.503	0.490	0.557	0.442	0.477	0.399	**0.919**
	Infrared	OAH	0.447	0.499	0.504	0.499	0.483	**0.650**	0.486	0.607

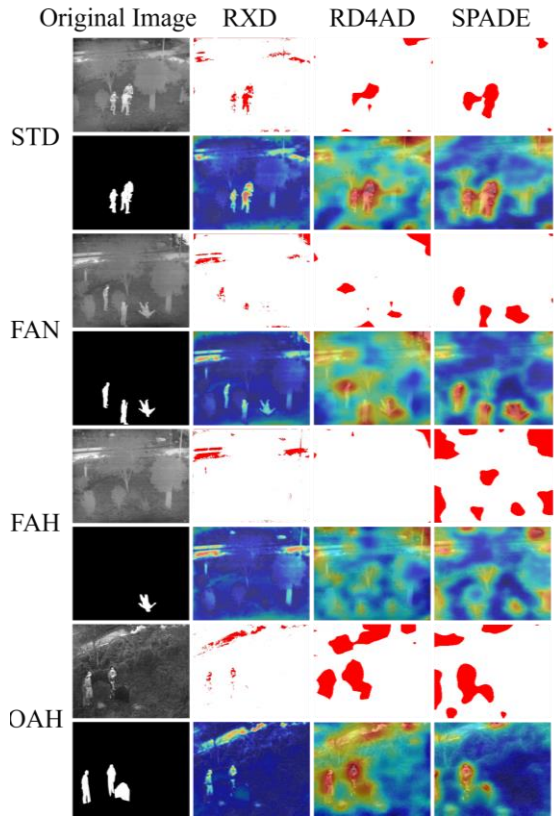

Fig. 10. Performances of the traditional and deep-learning detection algorithms on the fused dataset. In comparison with the single-modality, high anomaly scores (i.e. the hotter regions) in the heat maps are more concentrated on the person to be rescued, with diminished distribution over the background. Furthermore, in the binary maps, the anomaly regions cover the ground truth more comprehensively, demonstrating the improvements with fusion.

TABLE II
RESULTS OF FUSION METHODS

Scene	Method	EN↑	SD↑	SF↑	MI↑	SCD↓	SSIM↑
STD	Wavelet	6.52	26.46	5.1	2.3	1.98	0.87
	U2Fusion	6.59	26.74	7.4	**2.45**	1.91	0.94
	SDNet	6.61	27.05	8.42	1.99	1.89	1.13
	TarDAL	6.67	27.36	7.89	2.08	1.84	1.06
	DeFusion	6.64	27.89	7.15	2.12	1.79	1.19
	Ours	**6.7**	**27.95**	**8.83**	2.02	**1.77**	**1.22**
FAN	Wavelet	6.5	24.54	6.43	2.44	2.1	0.88
	U2Fusion	6.57	25.93	6.66	2.83	1.99	0.92
	SDNet	6.64	28.64	7.98	2.67	1.94	0.95
	TarDAL	6.73	27.12	8.12	2.56	1.88	0.94
	DeFusion	6.71	28.72	6.7	2.99	1.92	0.99
	Ours	**6.75**	**29.84**	**8.3**	**3.09**	**1.84**	**1.02**
FAH	Wavelet	6.43	27.93	8.13	1.98	2.05	0.97
	U2Fusion	6.50	28.1	8.45	2.04	1.95	1.09
	SDNet	6.46	28.67	8.71	1.41	1.96	1.22
	TarDAL	6.74	29.46	9.86	1.37	1.89	1.17
	DeFusion	6.72	29.63	8.42	**2.3**	1.83	1.25
	Ours	**6.81**	**29.94**	**9.89**	1.57	**1.77**	**1.27**
OAH	Wavelet	6.57	33.1	12.15	1.46	1.9	0.99
	U2Fusion	6.62	33.92	13.89	1.71	1.86	0.99
	SDNet	6.59	34.97	16.05	1.62	1.81	1.06
	TarDAL	6.67	34.6	15.51	1.5	1.75	1.03
	DeFusion	6.65	**35.43**	12.86	1.84	1.79	**1.12**
	Ours	**6.7**	31.22	**16.57**	**1.97**	**1.7**	1.07

Boldface and underline show the best and second-best values, respectively. 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.

TABLE III
RESULTS OF DETECTION ON FUSED IMAGES

Metric	Scene	RXD	RD4AD	SPADE
FPR↓	STD	0.686	0.850	**0.667**
	FAN	0.889	0.853	**0.143**
	FAH	1.00	0.879	**0.139**
	OAH	0.973	**0.943**	0.962
TPR↑	STD	**0.982**	0.978	0.962
	FAN	0.924	0.614	**0.944**
	FAH	**0.964**	0.944	0.906
	OAH	0.976	**0.979**	0.967
AUC↑	STD	0.906	**0.914**	0.740
	FAN	0.809	0.898	**0.974**
	FAH	0.318	0.731	**0.936**
	OAH	0.704	**0.869**	0.633

Boldface and underline show the best and second-best values, respectively. The SPADE outperforms in FAN and FAH scenarios, while the RXD and RD4AD algorithms, respectively, demonstrate superior performance in the FAN and FAH scenarios.

To facilitate a comparison of the performance of RXD, RD4AD and SPADE in different scenarios and multimodal datasets, a further illustration is provided in Fig. 11. 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. Addressing this problem requires associating more features with contextual relationships, which is the limitation of our fusion approach for anomaly detection tasks in night-time rescue missions.

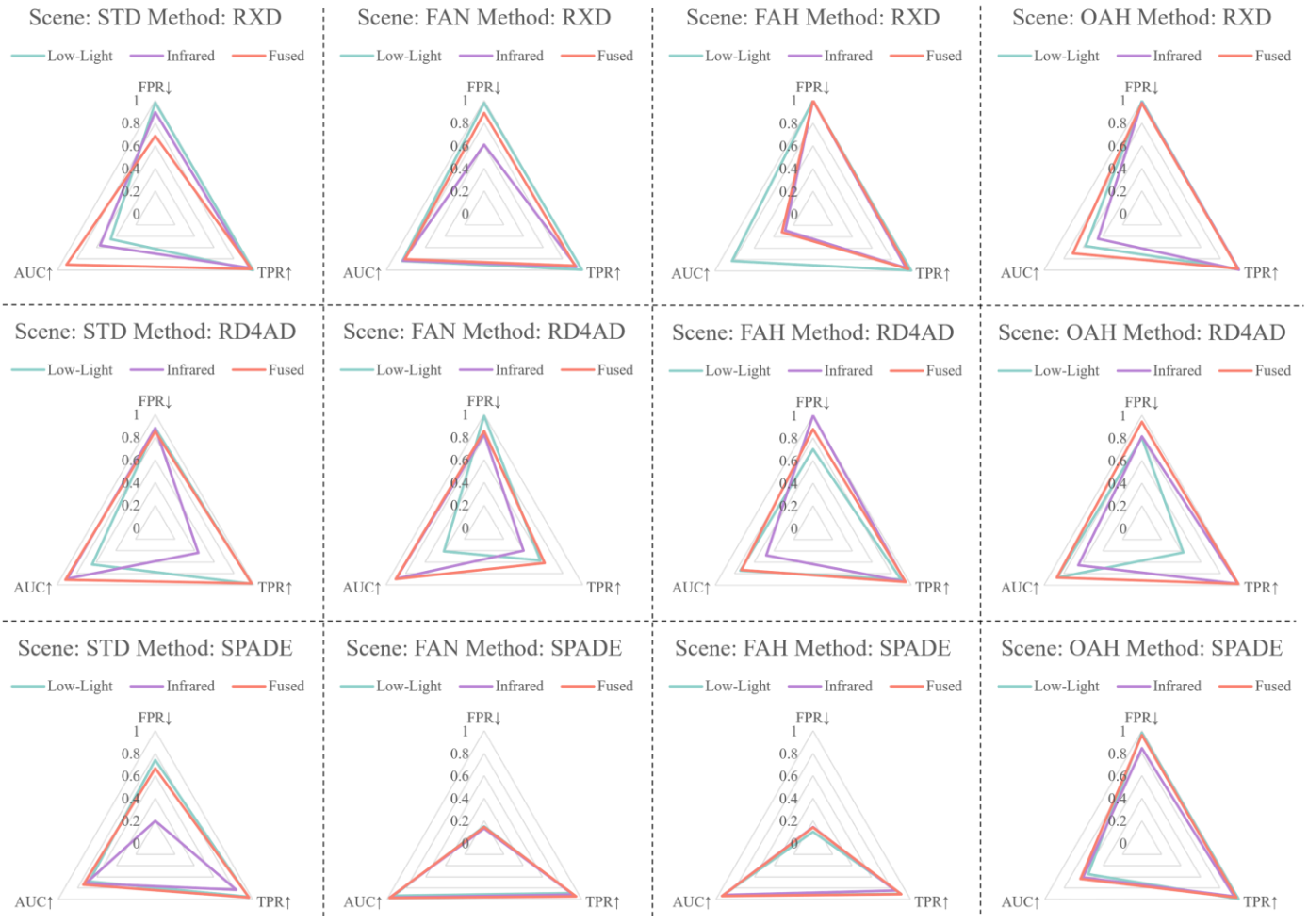

Fig. 11. Comparison of RXD, RD4AD and SPADE in different scenes and multimodal datasets. The detection results of fusion have more superior AUC and TPR performance on almost all scenarios than single-modality. The SPADE based on fused images demonstrated superior AUC scores in all four scenarios, with 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. In addition, 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).

VI. Summary and Discussion

In this paper, a fair evaluation of mainstream unsupervised anomaly detection methods in the field of nighttime emergency rescue was conducted, and a fusion algorithm based on frequency domain feature decomposition was proposed, which has been shown to enhance the performance of multiple detectors. To address the issue of insufficient data for target recognition and detection in nighttime emergency rescue situations, a well-annotated and augmented dataset (MRSI-NERD) was proposed. Considering the labor-intensive nature of constructing task-specific annotated datasets in real-world scenarios, which may not meet the timely requirements of emergency rescue, pixel-wise anomaly detection algorithms were researched. These algorithms treat a small number of annotated samples as anomaly information and extensively extract meaningful features from normal images. Thanks to the powerful feature extraction capabilities of deep neural networks, deep learning-based detection methods go beyond traditional algorithms. However, traditional algorithms based on RX also demonstrate decent performance in complex detection tasks. Among the deep learning methods, approaches based on reconstruction and Memory Bank have achieved superior performance. To further enhance the performance of unsupervised anomaly detection, a deep multimodal fusion algorithm based on frequency domain feature decomposition was further proposed. Experimental results demonstrate that this fusion strategy effectively improves the performance of the detectors.

In addition, due to the complexity of natural backgrounds, it has been observed that some anomaly detection methods struggle to meet the precision requirements for nighttime emergency rescue environments. Therefore, it is hoped that future research directions can focus on the following aspects:

1) Developing unsupervised domain adaptation methods

Anomaly detection methods rely on comprehensive normal datasets that are thoroughly annotated and representative of the operational environment. However, collecting such datasets can be challenging in nighttime emergency rescue scenarios due to the variability and unpredictability inherent to these situations. Unsupervised domain adaptation methods can effectively leverage data from both domains to learn an efficient anomaly detection model for the target domain [111]. Therefore, more unsupervised domain adaptation methods should be developed to promote the detection performance.

2) Developing active learning-based methods

Current process of labelling data is typically time-consuming and, as a consequence, a hindrance to the adoption of machine learning methods for the automated anomaly detection [112]. Employing expert feedback, active learning is an effective tool for building anomaly detection models. The model queries samples that need to be labelled by experts and then retrains the model using the samples. This approach can reduce the burden of acquiring labelled datasets while improving the anomaly detection performance [113]. Therefore, active learning-based Low-light and infrared remote sensing images anomaly detection methods should be further developed.

3) Constructing different operators based on low-light and infrared data

Night-time operations are typified by low-light conditions, which can significantly restrict the visibility and efficacy of conventional imaging systems. While infrared data offers distinct advantages in low-light settings, it necessitates the utilization of specialized processing techniques. The characteristics of low-light and infrared data can be leveraged to construct operators that make full use of prior knowledge [114]. Thus, how best to construct the different operators based on low-light and infrared data, preserving texture details, performance, and operational efficiency, is an interesting topic.

4) Coupling physical mechanisms and data-driven methods in low-light and infrared multimodal detection system for nighttime rescue mission.

Due to the inability of capturing physical mechanisms present in the images, data-driven deep learning algorithms to image data frequently results in suboptimal performance in complex scenarios [115]. Some traditional anomaly detection algorithms, designed based on statistical or physical features of the image, such as RXD, demonstrate performance on par with some deep learning algorithms on the fused dataset. This suggests the potential to enhance the efficacy of low-light and infrared multimodal detection systems by coupling data-driven methods with physical mechanisms. Consequently, the combination of deep learning anomaly detection methods with physical mechanisms in low-light and infrared multimodal data could be promising for the further investigation.

References

[1] H. Guo, “Understanding global natural disasters and the role of earth observation,” *International Journal of Digital Earth*, vol. 3, no. 3, pp. 221–230, 2010.

[2] K. Sha, W. Shi, and O. Watkins, “Using wireless sensor networks for fire rescue applications: Requirements and challenges,” *IEEE International Conference on Electro/Information Technology*, 2006: 239-244.

[3] J. Bravo-Arrabal, M. Toscano-Moreno, J. J. Fernández-Lozano, A. Mandow, J. A. Gomez-Ruiz, and A. García-Cerezo, “The internet of cooperative agents architecture (X-ioca) for robots, hybrid sensor networks, and mec centers in complex environments: A search and rescue case study,” *Sensors*, vol. 21, no. 23, p. 7843, 2021.

[4] T. Oron-Gilad, J. Y. C. Chen, and P. A. Hancock, “Remotely operated vehicles (ROVs) from the top-down and the bottom-up,” *Human factors of remotely operated vehicles*, Emerald Group Publishing Limited, 2006, pp. 37–47.

[5] R. Bahmanyar and N. Merkle, “Saving lives from above: Person detection in disaster response using deep neural networks,” *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 10, pp. 343–350, 2023.

[6] M. Zhao, P. Wang, Q. Zhao, X. Mao, J. Nie, and Y. Huang, “Research on the route selection of emergency rescue materials based on fairness and timeliness in the early post earthquake period,” *International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, 2020: 1-4.

[7] R. Geraldes et al., ‘UAV-based situational awareness system using deep learning’, *IEEE access*, vol. 7, pp. 122583–122594, 2019.

[8] A. Villringer and B. Chance, “Non-invasive optical spectroscopy and imaging of human brain function,” *Trends in neurosciences*, vol. 20, no. 10, pp. 435–442, 1997.

[9] J. Xing, “Research on detection probability of infrared and low light detection systems,” vol. 2073. *AIP Publishing*, 2019.

[10] J. Zhao and W. Liu, "A fusion method of infrared and low light level images based on improved NSCT," *Ninth Symposium on Novel Photoelectronic Detection Technology and Applications*, vol. 12617. SPIE, 2023.

[11] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2207–2219, 2016.

[12] J. Xing, "Research on detection probability of infrared and low light detection systems," *AIP Conference Proceedings*, vol. 2073. AIP Publishing, 2019.

[13] N. Levin et al., "Remote sensing of night lights: A review and an outlook for the future," *Remote Sensing of Environment*, vol. 237, p. 111443, 2020.

[14] J. Hai et al., "R2rnet: Low-light image enhancement via real-low to real-normal network," *Journal of Visual Communication and Image Representation*, vol. 90, p. 103712, 2023.

[15] Z. Zhu, K. Fujimura, and Q. Ji, "Real-time eye detection and tracking under various light conditions," *Proceedings of the 2002 symposium on Eye tracking research & applications*. 2002: 139-144.

[16] A. Jara et al., "Joint de-blurring and nonuniformity correction method for infrared lowscopy imaging," *Infrared Physics & Technology*, vol. 90, pp. 199–206, 2018.

[17] A. M. Waxman, M. Aguilar, D. A. Fay, D. B. Ireland, and J. P. Racamato, "Solid-state color night vision: fusion of low-light visible and thermal infrared imagery," *Lincoln Laboratory Journal*, vol. 11, no. 1, pp. 41–60, 1998.

[18] Y. Liu, L. Dong, and W. Xu, "Infrared and visible image fusion via salient object extraction and low-light region enhancement", *Infrared Physics & Technology*, vol. 124, p. 104223, 2022.

[19] K. Amolins, Y. Zhang, and P. Dare, "Wavelet based image fusion techniques—An introduction, review and comparison," *ISPRS Journal of photogrammetry and Remote Sensing*, vol. 62, no. 4, pp. 249–263, 2007.

[20] X. Zhang, and Y. Demiris, "Visible and Infrared Image Fusion Using Deep Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10535–10554, 2023.

[21] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Information Fusion*, vol. 76, pp. 323–336, 2021.

[22] P. Shamsolmoali et al., "Image synthesis with adversarial networks: A comprehensive survey and case studies," *Information Fusion*, vol. 72, pp. 126–146, 2021.

[23] Q. Tang, J. Liang, and F. Zhu, "A Comparative Review on Multi-Modal Sensors Fusion Based on Deep Learning," *Signal Processing*, vol. 213, pp. 109165-109165, 2023.

[24] A. Agarwal et al., "Comparison of the theoretical and statistical effects of the PCA and CNN image fusion approaches," *Handbook of Research on Thrust Technologies' Effect on Image Processing*, IGI Global, 2023, pp. 193–205.

[25] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[26] H. E. Egilmez and A. Ortega, "Spectral anomaly detection using graph-based filtering for wireless sensor networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014: 1085-1089.

[27] L. Bruzzone and S. B. Serpico, "Detection of changes in remotely-sensed images by the selective use of multi-spectral information," *Information Fusion*, vol. 18, no. 18, pp. 3883–3888, 1997.

[28] M. Zhao, P. Wang, Q. Zhao, X. Mao, J. Nie, and Y. Huang, "Research on the route selection of emergency rescue materials based on fairness and timeliness in the early post earthquake period," *2020 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, 2020: 1-4.

[29] J. Yang, R. Xu, Z. Qi, and Y. Shi, "Visual anomaly detection for images: A systematic survey," *Procedia computer science*, vol. 199, pp. 471–478, 2022.

[30] L. Ruff et al., "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, 2021.

[31] J. Yang, R. Xu, Z. Qi, and Y. Shi, "Visual anomaly detection for images: A survey," *arXiv preprint arXiv*:2109.13157, 2021.

[32] U. Pietsch et al., "Efficacy and efficiency of indoor nighttime human external cargo mission simulation in a high-fidelity training Centre," *Scandinavian journal of trauma, resuscitation and emergency medicine*, vol. 28, pp. 1–7, 2020.

[33] A. M. R. Bernal, W. Scheirer, and J. Cleland-Huang, "NOMAD: A Natural, Occluded, Multi-scale Aerial Dataset, for Emergency Response Scenarios," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024: 8584-8595.

[34] S. Leroux, B. Li, and P. Simoens, "Multi-branch neural networks for video anomaly detection in adverse lighting and weather conditions," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022: 2358-2366.

[35] D. J. A. Brown, H. Brugger, J. Boyd, and P. Paal, "Accidental hypothermia," *New England Journal of Medicine*, vol. 367, no. 20, pp. 1930–1938, 2012.

[36] S. Davies and R. M. J. Gray, "What is occlusion," *British dental journal*, vol. 191, no. 5, pp. 235–245, 2001.

[37] O. Taisei et al. "Detection of Fallen Persons and Person Shadows from Drone Images," *Proceedings of International Conference on Artificial Life and Robotics*, vol. 28, pp. 890–894, 2024.

[38] M. Zaabi, N. Smaoui, W. Hariri, and H. Derbel, "Deep and statistical-based methods for alzheimer's disease detection: A survey," *Journal of Computing Science and Engineering*, vol. 16, no. 1, pp. 1–13, 2022.

[39] D. Broyles, C. R. Hayner, and K. Leung, "Wisard: A labeled visual and thermal image dataset for wilderness search and rescue," *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022: 9467-9474.

[40] C. Frei et al., "Clinical characteristics and outcomes of witnessed hypothermic cardiac arrest: a systematic review on rescue collapse," *Resuscitation*, vol. 137, pp. 41–48, 2019.

[41] K. Wang, L. Yang, and M. Kucharek, "Investigation of the effect of thermal insulation materials on packaging performance," *Packaging Technology and Science*, vol. 33, no. 6, pp. 227–236, 2020.

[42] N. Bustos et al, "A Systematic Literature Review on Object Detection Using near Infrared and Thermal Images," *Neurocomputing*, vol. 560, pp. 126804–126804, 2023.

[43] E. L. Lloyd, "The cause of death after rescue," *International journal of sports medicine*, vol. 13, no. S 1, pp. S196–S199, 1992.

[44] L. Jia, Z. Qian, X. Feng, and J. Lijun, "Obstacle detection method for underwater mine emergency rescue auv," *2017 2nd International Conference on Frontiers of Sensors Technologies (ICFST)*. IEEE, 2017

[45] S. U. N. Aiping, G. Yangyun, and Z. H. U. Youpan, "Optical system design of low-light-level and infrared image fusion hand-held viewer," *Infrared Technology*, vol. 35, no. 11, pp. 712–717, 2013.

[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," *Review of Scientific Instruments*, 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 Transactions on Image Processing*, vol. 29, pp. 3296–3310, 2019.

[48] D. A. Kalashnikov, A. V. Paterova, S. P. Kulik, and L. A. Krivitsky, "Infrared spectroscopy with visible light," *Nature Photonics*, vol. 10, no. 2, pp. 98–101, 2016.

[49] Y. Li et al., "Efficient and explicit modelling of image hierarchies for image restoration," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023: 18278-18289.

[50] Z. Zhao et al., "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023: 5906-5916

[51] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite transformer with long-short range attention," arXiv preprint arXiv:2004.11886, 2020.

[52] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[53] G. G. Gregoriou, S. J. Gotts, H. Zhou, and R. Desimone, "High-frequency, long-range coupling between prefrontal and visual cortex during attention," *science*, vol. 324, no. 5931, pp. 1207–1210, 2009.

[54] J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, and J.-H. Jacobsen, "Invertible residual networks," *International conference on machine learning*. PMLR,2019:573-582

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: 770-778.
[56] M. Ladegaard, F. H. Jensen, M. De Freitas, V. M. Ferreira da Silva, and P. T. Madsen, "Amazon river dolphins (Inia geoffrensis) use a high-frequency short-range biosonar," *Journal of Experimental Biology*, vol. 218, no. 19, pp. 3091–3101, 2015.
[57] L. Liu, L. Tang, and W. Zheng, "Lossless image steganography based on invertible neural networks," Entropy, vol. 24, no. 12, p. 1762, 2022.
[58] M. Kryszkiewicz, "The cosine similarity in terms of the euclidean distance," *Encyclopedia of Business Analytics and Optimization*, IGI Global, 2014, pp. 2498–2508.
[59] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE) –Arguments against avoiding RMSE in the literature," *Geoscientific model development*, vol. 7, no. 3, pp. 1247–1250, 2014.
[60] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE transactions on acoustics, speech, and signal processing*, vol. 38, no. 10, pp. 1760–1770, 1990.
[61] T. Veracini, S. Matteoli, M. Diani, and G. Corsini, "Fully unsupervised learning of Gaussian mixtures for anomaly detection in hyperspectral imagery," *2009 Ninth International Conference on Intelligent Systems Design and Applications*. IEEE, 2009: 596-601.
[62] H. Zhang, X. Jin, Q. M. J. Wu, Y. Wang, Z. He, and Y. Yang, "Automatic visual detection system of railway surface defects with curvature filter and improved Gaussian mixture model," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 7, pp. 1593–1608, 2018.
[63] G. J. McLachlan, "Mahalanobis distance," *Resonance*, vol. 4, no. 6, pp. 20–26, 1999.
[64] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.
[65] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4–5, pp. 411–430, 2000.
[66] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," now, 2011, doi: 10.1561/2200000016.
[67] P. Rizk, R. Younes, A. Ilinca, and J. Khoder, "Wind turbine blade defect detection using hyperspectral imaging," *Remote Sensing Applications: Society and Environment*, vol. 22, p. 100522, 2021.
[68] B. Shi, J. Liang, L. Di, C. Chen, and Z. Hou, "Fabric defect detection via low-rank decomposition with gradient information," *IEEE Access*, vol. 7, pp. 130423–130437, 2019.
[69] T. Zhou and D. Tao, "Godec: Randomized low-rank & sparse matrix decomposition in noisy case," in Int. Conf., ICML 2011.
[70] H. Zhang, Z. Guo, Z. Qi, and J. Wang, "Research of glass defects detection based on DFT and optimal threshold method," *2012 International Conference on Computer Science and Information Processing (CSIP)*. IEEE, 2012: 1044-1047.
[71] H. Liu, W. Zhou, Q. Kuang, L. Cao, and B. Gao, "Defect detection of IC wafer based on spectral subtraction," *IEEE transactions on semiconductor manufacturing*, vol. 23, no. 1, pp. 141–147, 2010.
[72] D.-M. Tsai and T.-Y. Huang, "Automated surface inspection for statistical textures," Comput., vol. 21, no. 4, pp. 307–323, 2003.
[73] D.-M. Tsai and C.-Y. Hsieh, 'Automated surface inspection for directional textures', *Image and Vision computing*, vol. 18, no. 1, pp. 49–62, 1999.
[74] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in Conf. Comput. Recognit., Anchorage, AK, 2008, pp. 1-8, doi: 10.1109/CVPR.2008.4587715.
[75] D. Aiger and H. Talbot, "The phase only transform for unsupervised surface defect detection," *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010: 295-302.
[76] D.-M. Tsai and C.-C. Kuo, "Defect detection in inhomogeneously textured sputtered surfaces using 3D Fourier image reconstruction," *Machine Vision and Applications*, vol. 18, pp. 383–400, 2007.
[77] X. Bai, Y. Fang, W. Lin, L. Wang, and B.-F. Ju, "Saliency-based defect detection in industrial images by using phase spectrum," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2135–2145, 2014.
[78] J. A. Dabin, A. M. Haimovich, J. Mauger and A. Dong, "Blind Source Separation with L1 Regularized Sparse Autoencoder," *2020 29th Wireless and Optical Communications Conference (WOCC)*. IEEE, 2020.
[79] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," IE, vol. 2, no. 1, pp. 1–18, 2015.
[80] Y. Liu, C. Zhuang, and F. Lu, "Unsupervised two-stage anomaly detection," *arXiv preprint arXiv*:2103.11671, 2021.
[81] J. Yang, Y. Shi, and Z. Qi, "Dfr: Deep feature reconstruction for unsupervised anomaly segmentation," *arXiv preprint arXiv*:2012.07122, 2020.
[82] Y. Yan, D. Wang, G. Zhou, and Q. Chen, "Unsupervised anomaly segmentation via multilevel image reconstruction and adaptive attention-level transition," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.
[83] K. Zhou et al., "Encoding structure-texture relation with p-net for anomaly detection in retinal images," *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer International Publishing, 2020: 360-377.
[84] T. Liu, B. Li, Z. Zhao, X. Du, B. Jiang, and L. Geng, "Reconstruction from edge image combined with color and gradient difference for industrial surface anomaly detection," *arXiv preprint arXiv*:2210.14485, 2022.
[85] Z. Li et al., "Superpixel Masking and Inpainting for Self-Supervised Anomaly Detection," *Bmvc*. 2020.
[86] D. Dehaene and P. Eline, "Anomaly localization by modeling perceptual features," *arXiv preprint arXiv*:2008.05369, 2020.
[87] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, p. 107706, 2021.
[88] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv*:1503.02531, 2015.
[89] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022: 9737-9746.
[90] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020: 4183-4192.
[91] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021: 14902-14912.
[92] G. Wang, S. Han, E. Ding, and D. Huang, "Student-teacher feature pyramid matching for anomaly detection," *arXiv preprint arXiv*:2103.04257, 2021.
[93] X. Yan, H. Zhang, X. Xu, X. Hu, and P.-A. Heng, "Learning semantic context from normal samples for unsupervised anomaly detection," *Proceedings of the AAAI conference on artificial intelligence*. 2021, 35(4): 3110-3118.
[94] J. Song, K. Kong, Y.-I. Park, S.-G. Kim, and S.-J. Kang, "AnoSeg: Anomaly segmentation network using self-supervised learning," arXiv preprint arXiv:2110.03396, 2021.
[95] Y. Liang, J. Zhang, S. Zhao, R. Wu, Y. Liu, and S. Pan, "Omni-frequency channel-selection representations for unsupervised anomaly detection," *IEEE Transactions on Image Processing*, 2023.
[96] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *Conf. Comput. Recognit.* IEEE/CVF 2021: 9664-9674.
[97] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv*:1708.04552, 2017.
[98] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Conf. Comput. Recognit.* 2022: 14318-14328.
[99] J. Yoon et al, "SPADE: Semi-Supervised Anomaly Detection under Distribution Mismatch," *arXiv preprint arXiv*:2212.00173, 2022.
[100] N. Li et al, "Anomaly Detection via Self-Organizing Map," *arXiv preprint* arXiv:2107.09903, 2021.
[101] Q. Wan, L. Gao, X. Li, and L. Wen, "Industrial image anomaly localization based on Gaussian clustering of pretrained feature," *IEEE*

Transactions on Industrial Electronics, vol. 69, no. 6, pp. 6182–6192, 2021.

[102] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022: 392-408.

[103] S. Lee, S. Lee, and B. C. Song, "Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization," *IEEE Access*, vol. 10, pp. 78446–78454, 2022.

[104] D. Kim, C. Park, S. Cho, and S. Lee, Fapm: "Fast adaptive patch memory for real-time industrial anomaly detection," *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023: 1-5.

[105] J. Jang, E. Hwang, and S.-H. Park, "N-pad: Neighboring pixel-based industrial anomaly detection," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023: 4364-4373.

[106] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD--A comprehensive real-world dataset for unsupervised anomaly detection," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE/CVF 2019: 9592-9600.

[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.

[111] Z. Yang, I. Soltani, and E. Darve, "Anomaly detection with domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2958-2967, 2023.

[112] S. Russo et al, "Active Learning for Anomaly Detection in Environmental Data," in *Environmental Modelling & Software*, vol. 134, p. 104869, 2020.

[113] M. Kim, J. Kim, and J. Yu, "Active Anomaly Detection Based on Deep One-Class Classification," in *Pattern Recognition Letters*, vol. 167, pp. 18–24, 2023.

[114] W. Ma et al, "Infrared and Visible Image Fusion Technology and Application: A Review." *Sensors*, vol. 23, no. 2, pp. 599–599, 2023.

[115] Li J, Wang X, Wang S, et al, "One Step Detection Paradigm for Hyperspectral Anomaly Detection via Spectral Deviation Relationship Learning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.