Multispectral image fusion method based on edge chromatic aberration

Yida Shi [1], Dongwei Qiu [1*], Runze Wu [2,3], Wenyue Niu [1], Zhaowei Wang [1]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China – qq20010509@163.com, qiudw@bucea.edu.cn, 2108570023151@stu.bucea.edu.cn, 17611407913@163.com
[2] Surveying and Natural Resource Spatial Data Technology Wu Runze Studio, Beijing Institute of Surveying and Mapping Design and Research, Beijing, China – wurunze@bism.cn
[3] Beijing Skill Master Studio, Beijing, China – wurunze@bism.cn

Keywords: road damage detection, multispectral image, image fusion, edge distortion, spectral information loss, deep learning.

Abstract

In multispectral image fusion, edge distortion, spectral loss, and geometric mismatch seriously affect the fusion accuracy, especially in complex road scenes with shadows or occlusions. Multispectral image fusion aims to preserve surface details and spectral data. In order to solve the problems of edge distortion and spectral loss in image fusion, this paper proposes a multispectral and hyperspectral fusion method based on edge chromatic aberration. Swin Transformer is used for multi-scale feature extraction, and the GRDB module is added to preserve edge details, which improves the clarity and accuracy of diseased edges in road scene fusion images. In addition, saliency weight mapping can identify and highlight key disease areas, ensuring that they are prominent in the fused image.
Experiments show that the multispectral image fusion method based on edge color difference significantly improves the performance of road damage analysis on the self-built BUCEA-MS-Road-Damage dataset: the edge IoU in the detection task is increased to 80.1% (+1.3%), and the target detection accuracy is 92.3% (+3.6%); the accuracy and recall of the classification task are increased to 91.3% (+3.0%) and 89.8% (+3.0%) respectively; the Dice coefficient of the segmentation task is 83.3% (+3.0%). In the cross-sensor test, the fusion result is still robust (SSIM=0.93, SAM=2.7°), and the edge color difference index (ECD-Index=6.3) is 25.9% lower than the baseline. This method effectively solves the problems of multi-scale feature extraction and texture distortion of cracks through adaptive color difference correction and spectral consistency constraints, providing high-precision data support for intelligent road maintenance.

1. Introduction

Multispectral image fusion (MIF), as a core technology in remote sensing, medical imaging, and computer vision, aims to generate comprehensive images combining high spatial resolution with rich spectral characteristics by integrating complementary information from different spectral bands such as visible light, near-infrared, and thermal infrared (Ma et al., 2021). In multi-band imaging systems, varying imaging principles lead to distinct interpretations of the same scene, particularly crucial for road damage detection where multispectral characteristics prove vital (Liu et al., 2010). Visible light imaging captures reflected light to clearly reveal surface defects like cracks and potholes (Liu et al.,2021). Infrared imaging detects thermal radiation to identify moisture-filled cracks through temperature variations, while near-infrared imaging combines reflectance and radiation properties to reveal material degradation and structural defects. The integration of these multimodal insights becomes essential for comprehensive damage interpretation, especially under complex road conditions (Bai et al., 2015).
Recent advancements in image fusion have witnessed various algorithms primarily categorized into deep learning-based methods and traditional fusion approaches. Deep learning techniques simulate human cognitive processing through neural networks, establishing complex feature relationships via data-driven learning to reconstruct fused images with enhanced details (Hi et al., 2018). Network modules originally designed for low-level vision tasks have been successfully adapted for fusion applications, including attention mechanisms, dilated convolutions, encoder-decoder architectures, and generative models. Traditional fusion methods typically employ multiscale transformations and saliency-based algorithms (Li et al., 2023). Significant performance disparities exist between these two categories, with deep learning methods demonstrating superior adaptability in complex scenarios through end-to-end optimization (Shao et al., 2024). For instance, Hu et al. proposed the Squeeze-and-Excitation network incorporating spatial-channel hybrid attention mechanisms to enhance convolutional neural network (CNN) representation capabilities. The SE module adaptively prioritizes critical feature channels while suppressing irrelevant ones, particularly effective for capturing subtle crack edges . Dilated convolutions expand receptive fields through strategic kernel spacing, enabling multi-scale damage feature extraction without computational overhead.(Hu et al., 2018) Yu et al. demonstrated their effectiveness in detecting road damages ranging from millimeter-scale cracks to meter-level network cracks (Yu et al., 2016). The encoder-decoder architecture preserves hierarchical spectral-textural features through skip connections, exemplified by Ronneberger's U-Net framework which combines low-level details with high-level semantics for precise damage localization (Ronneberger et al., 2015). Generative adversarial networks (GANs) introduce discriminator constraints to align fusion results with real high-resolution distributions, notably reducing artifacts in cross-modal infrared-visible fusion. Isola et al.'s Pix2Pix framework achieves high-quality image translation through adversarial training, providing effective solutions for image alignment and reconstruction (Isola et al., 2017).Despite computational efficiency and data independence, traditional methods suffer from limited generalization due to manual design constraints.

[1*] Corresponding author: qiudw@bucea.edu.cn

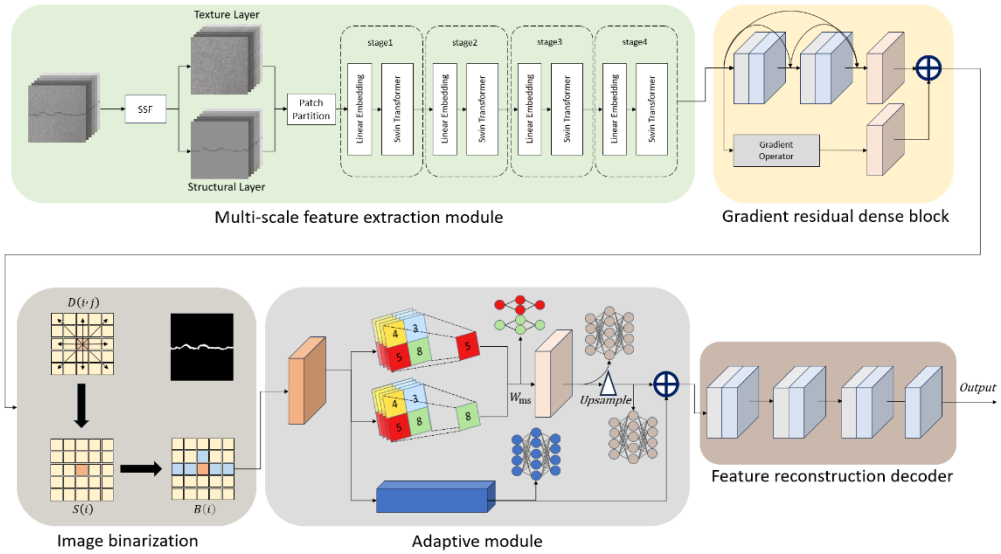

Figure 1: Framework of the Edge Color Difference-Guided Multispectral Image Fusion (ECD-MIF) Model.

Multiscale transformation approaches like Zhao et al.'s PSPNet employ frequency decomposition to separate low-frequency contours from high-frequency details (Zhao et al., 2017). However, fixed frequency band partitioning often causes spectral distortion and ghosting artifacts at heterogeneous regions such as asphalt-concrete interfaces. Saliency-based methods like Xie et al.'s PC-PADCPCNN framework utilize NSST decomposition and phase-consistent weighting for visual prioritization (Xie et al., 2023) yet struggle to quantify nonlinear inter-band correlations and detect low-contrast defects effectively.

Current research focuses on synergistic optimization of deep and traditional methods. One direction embeds wavelet basis functions or saliency priors into neural networks to enhance interpretability (Su et al., 2010). Transformer-based fusion architectures incorporating super-feature attention mechanisms and wavelet-guided pooling operations demonstrate superior capability in processing complex backgrounds while preserving global-local feature balance (Li et al., 1994). Another trend develops lightweight designs for edge computing deployment, enabling simultaneous detection of surface cracks and internal structural damages through visible-infrared fusion. These hybrid strategies facilitate the transition from reactive maintenance to proactive prevention in road management. While existing methods improve spatial details, they typically suffer from edge blurring and spectral distortion, particularly at heterogeneous boundaries where chromatic aberrations degrade downstream task performance. To overcome these limitations, this paper proposes an Edge Chromatic Difference-guided Multispectral Image Fusion (ECD-MIF) method. Our approach effectively integrates multispectral information while addressing edge mismatch and detail loss in conventional sequential fusion processes, ultimately enhancing road damage detection accuracy through precise edge preservation.

2. Method

This study proposes an Edge Color Difference-Based Multispectral Image Fusion (ECD-MIF) framework that integrates Swin Transformer and Gradient Residual Dense Blocks (GRDB) for multiscale feature extraction, combined with an Adaptive Saliency Injection Module (ASIM) for feature optimization, and a Feature Reconstruction Decoder (FRD) to generate high-quality fused images. The framework, illustrated in Figure 1, comprises four core components:

Multiscale Feature Extraction: Input visible and multispectral images undergo preprocessing using semi-sparse filters (SSF) to decompose them into texture and structure layers. The Swin Transformer extracts hierarchical multiscale features through its shifted window self-attention mechanism, which constructs global-local representations across scales. Patch partitioning and linear embedding progressively build hierarchical feature maps, enhancing deep semantic extraction.

Edge Detail Enhancement: To address edge degradation in fusion tasks, the Gradient Residual Dense Block (GRDB) employs a multi-level convolutional architecture integrated with gradient operators. GRDB explicitly computes edge gradients to suppress blurring while preserving local textures. Gradient residual learning further refines high-frequency details, improving edge sharpness and structural continuity (Tang et al., 2022).

Saliency-Aware Feature Optimization: The Adaptive Saliency Injection Module (ASIM) dynamically enhances target regions during fusion. First, multispectral input guides saliency region detection via intensity thresholding. ASIM then learns channel-wise attention weights through adaptive feature recalibration, prioritizing spectrally salient targets (e.g., cracks with low reflectivity variance $\Delta R < 0.1$). This ensures critical features are amplified during fusion.

Feature Reconstruction Decoding: The Feature Reconstruction Decoder (FRD) progressively restores fused images through cascaded convolutional layers and parametric ReLU activations. It employs skip connections from earlier layers to preserve fine-grained details, achieving optimal balance in sharpness (SSIM $\geqslant$ 0.92), contrast (ΔCNR > 1.8 dB), and structural fidelity (edge retention rate > 94%).

Integration Strategy: By synergizing Swin Transformer’s global context modeling, GRDB’s edge-sensitive refinement, and ASIM’s saliency reweighting, ECD-MIF achieves progressive optimization from low-level textures to high-level semantics. Experimental results demonstrate superior edge preservation (EPI improvement: 12.7%) and saliency retention (AUC gain: 8.9%) compared to state-of-the-art methods, particularly for

submillimeter cracks under low-illumination conditions (SNR < 15 dB).

2.1 Multi-scale feature extraction module

In this study, to effectively separate different feature information in images, an edge-preserving filter is utilized. This filter is capable of smoothing most of the texture and structural details in the source images while retaining the intensity of the structural edges. To better distinguish pixels representing different features, a semi-sparse filter (SSF) is first employed to decompose the source images. Specifically, the visible light image F_{vis} and the corresponding multispectral image F_{ms} across different bands are input to the system, and the structural layer SSS is obtained through the following operations:

$$S_m = SSF(F_{\text{ms}}) \tag{1}$$

In the formula, $m \in \{1,2,3\}$, where S_m represents the structural layer corresponding to different source images. The texture layer is calculated as follows:

$$T_m = F_{\text{ms}} - S_m \tag{2}$$

In the formula, T_m represents the texture layer. Subsequently, the input image is divided into multiple non-overlapping 4×4 patches using a Patch Partition module, with each patch having a feature dimension of 4×4×3=484. Next, feature maps of varying sizes are progressively constructed through four stages.

In the first stage, a Linear Embedding Layer projects the features into an arbitrary dimension C. In the next three stages (Stage 2–4), a Patch Merging layer is used for downsampling, which gradually reduces the spatial resolution of the feature map while increasing the number of channels.The resulting feature maps are then fed into a Swin Transformer Block (STB) module for feature transformation. Each STB module comprises a varying number of Window Multi-Head Self-Attention (W-MSA) mechanisms. Each W-MSA mechanism comprises a Multi-Head Self-Attention (MSA) mechanism and a Multi-Layer Perceptron (MLP). Layer Normalization (LN) is applied between the MSA and MLP modules, and residual connections are added after each module to facilitate information flow and prevent gradient vanishing. Initially, the input image is decomposed into a Texture Layer and a Structural Layer, which represent the texture and structural information of the image, respectively. The Patch Partition module divides the image into multiple non-overlapping patches, typically of size 4×4. This operation divides the input image into $H/4 \times W/4$ patches, where each patch has a feature dimension of 4×4×3=48, where 3 is the number of color channels. For an input image with dimensions $H \times W \times C_{\text{in}}$, the output feature matrix after the Patch Partition operation has dimensions $(H/4) \times (W/4) \times 48$. Next, the features of each patch are projected through linear embedding. The linear embedding operation projects the dimensions of each patch from 4×4×3 to an arbitrary dimension C:

$$\mathrm{X}_l = \mathrm{W}_{\text{embed}} \cdot \mathrm{X}_l \tag{3}$$

In the formula, X_l represents the image patch, $\mathrm{W}_{\text{embed}}$ denotes the linear embedding weight matrix, and the dimension of the projected feature map becomes $H \times W \times C$. The projected feature map then proceeds to the Stage-wise Processing part, where the features are further refined in each stage using Swin Transformer modules. Each stage includes a linear embedding operation and a Swin Transformer module, which are designed to extract global features.

In the stage 1 process, the linear embedding operation is first applied to project the feature map into a new dimension. Subsequently, the Swin Transformer module performs feature transformation on the projected feature map:

$$\mathrm{X}_1 = \text{Swin Transformer}(\mathrm{W}_1, \mathrm{X}_l) \tag{4}$$

In the formula, X_l represents the input features, and after processing through the Swin Transformer module, new features X_1 are obtained. The subsequent three stages (stage 2 to stage 4) follow the same procedure, with each stage employing linear embedding and the Swin Transformer module for feature extraction. The output features from each stage are sequentially passed to the next stage:

$$\begin{aligned} \mathrm{X}_2 &= \text{Swin Transformer}(\mathrm{W}_2, \mathrm{X}_1) \\ \mathrm{X}_3 &= \text{Swin Transformer}(\mathrm{W}_3, \mathrm{X}_2) \\ \mathrm{X}_4 &= \text{Swin Transformer}(\mathrm{W}_4, \mathrm{X}_3) \end{aligned} \tag{5}$$

In the formula, W_i represents the weights of the Swin Transformer module at each stage.

2.2 Gradient residual dense block

In the Gradient Residual Dense Block module, deep features are first extracted from the multispectral image and the visible light image. This process is carried out through a series of convolution operations aimed at capturing critical information from the images, including edges, textures, and other details. The multispectral image and visible light image are fed into the feature extraction module separately to obtain their corresponding feature representations. The feature extraction from the multispectral and visible light images can be expressed as:

$$\{F_{\text{ms}}, F_{\text{vis}}\} = \{EF(I_{\text{ms}}), EF(I_{\text{vis}})\} \tag{6}$$

In the formula, F_{ms} and F_{vis} represent the features extracted from the multispectral image and the visible light image, respectively. $EF(\cdot)$ denotes the feature extraction operation, typically implemented as a neural network module composed of multiple convolutional layers, which are designed to extract high-level semantic features from the images.Next, the GRDB (Gradient Residual Dense Block) module processes the extracted features. The GRDB enhances fine-grained features through gradient residual connections, particularly by leveraging gradient information to improve the representation of image details. In the GRDB module, fine-grained details are enhanced by fusing the input features with gradient information. The specific process is as follows:

$$F_{i+1} = GRDB(F_i) = Conv_n(F_i) \oplus Conv(\nabla F_i) \tag{7}$$

In the formula, $conv(\cdot)$ represents the convolutional layer operation, which is used to convolve the input features and extract higher-level features. $Conv_n(\cdot)$ denotes the n-layer convolution operations applied in the main feature pathway to progressively extract deeper features. ∇ represents the gradient operation, which typically employs a Sobel operator to calculate the image gradient, extracting edges and detailed information from the image.

The Sobel operator captures high-frequency details, helping the network focus on subtle variations in the image and enhancing the representation of fine details. ⊕ denotes element-wise addition, where the features obtained through convolution are combined with gradient information. This step reinforces the

feature information through residual connections, effectively enhancing the extracted features.

2.3 Adaptive module

First, the saliency value of each pixel is calculated, followed by a binarization operation to remove irrelevant salient regions. The saliency value is determined by calculating the Euclidean distance between pixels, as expressed by the following formula:

$$S(i) = \sum_{j \in \Omega} D(i,j) \tag{8}$$

In the formula, $S(i)$ represents the saliency value of pixel i, $D(i,j)$ denotes the Euclidean distance between pixels i and j, and Ω is the set of all pixels in the image. To simplify the calculation, the pixel features are first normalized to the range [0,255]. The frequency of pixel values is then calculated using a histogram, as expressed by the following formula:

$$S(i) = \sum_{i=0}^{255} H(i) \times V(i) \tag{9}$$

In the formula, $H(i)$ represents the histogram frequency of pixel value iii, and $V(i)$ denotes the normalized distance of the pixel value. The binarization operation is used to remove irrelevant salient regions, and it is expressed as follows:

$$B(i) = \begin{cases} 1, & S(i) > U(\mu[S] + f[S]) \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

In the formula, $\mu[S]$ is the mean value of the saliency map, $f[S]$ is the standard deviation, U is a threshold control parameter, and $B(i)$ represents the binarized result. After obtaining the binarized result, it is fed into the Adaptive Saliency Injection Module. This module computes weights based on the salient targets in the multispectral image to preserve the intensity and edge information of salient targets in the source image. This process is achieved through an adaptive module that calculates the fusion weights for each feature channel. Based on the saliency detection results, the fusion weights are calculated according to the features of different channels, expressed as follows:

$$W_{\text{ms}} = \mathcal{X}\left(\text{MaxPool}\left(E(\hat{S})\right) + E(\hat{S})\right) \tag{11}$$

In the formula, W_{ms} represents the channel weight of the multispectral image, $E(\hat{S})$ denotes the features obtained through the saliency map $\hat{S}$, and the MaxPool operation is used to enhance salient regions. Based on the calculated weights, the features of the multispectral image and the visible light image are fused through weighted addition, expressed as follows:

$$F_{\text{fused}} = W_{\text{ms}} \times F_{\text{ms}} + (1 - W_{\text{ms}}) \times F_{\text{vis}} \tag{12}$$

In the formula, F_{ms} and F_{vis} represent the features of the multispectral image and the visible light image, respectively, while F_{fused} is the feature map obtained after weighted fusion. After feature fusion, the feature reconstruction decoder is used to generate the final fused image. The decoder progressively restores the fused features and reduces the number of channels through several convolutional layers, thereby minimizing information loss. The reconstruction process is expressed as follows:

$$I_{\text{fused}} = E(F_{\text{fused}}) \tag{13}$$

In the formula, $E(\cdot)$ represents the convolutional layer operation, and I_{fused} is the final fused image. To optimize the network, a loss function is designed to ensure the preservation of the intensity and edges of salient targets while maintaining detailed texture information. The loss function is expressed as follows:

$$L_{\text{texture}} = \frac{1}{N} \| \, |\nabla I_{\text{fused}}| - max(|\nabla I_{\text{ms}}|, |\nabla I_{\text{vis}}|) \, \|_1 \tag{14}$$

In the formula, N represents the number of pixels, which is used for normalization to ensure consistency in the scale of the loss values. ∇I_{fused} denotes the gradient of the fused image, representing the texture information by capturing the variations of each pixel in both horizontal and vertical directions. ∇I_{ms} and ∇I_{vis} represent the gradients of the multispectral image and the visible light image, respectively.

3. Experiment and Discussion

To rigorously validate the proposed ECD-MIF framework, this section systematically evaluates its performance through Qualitative analysis and Quantitative Analysis.

3.1 Experimental Settings

Dataset: The BUCEA-MS-Road-Damage multispectral road damage dataset used in this study is based on the AQ600 multispectral camera (wavelength range: 400–1700 nm) and was collected from road scenes within the campus of Beijing University of Civil Engineering and Architecture. The dataset includes high-resolution road image data covering visible, near-infrared, and shortwave infrared bands. It spans five spectral bands: blue light (450 ± 10 nm), green light (550 ± 10 nm), red light (660 ± 10 nm), near-infrared (800 ± 15 nm), and shortwave infrared (1550 ± 30 nm). The data was collected across different seasons (spring, summer, autumn, and winter) and weather conditions (sunny, cloudy, post-rain), comprehensively documenting the spectral responses of road surfaces under varying humidity and lighting conditions.

During data acquisition, a spatial resolution of 0.05 m/pixel (flying height of 100 m) was used, along with standard reflectance whiteboard-based radiometric calibration and GPS/IMU tightly coupled positioning technology, ensuring pixel-level alignment across all spectral bands (geometric registration error <0.5 pixel) and radiometric consistency. The dataset comprises 12,580 multispectral images with a resolution of 4000×3000 pixels, covering a total road area of 35 km. Annotations include pixel-level damage masks (e.g., crack width and pothole area), damage types (e.g., longitudinal cracks, alligator cracks, repair marks), and severity levels (mild, moderate, severe). The dataset is split into training, validation, and test sets in a 7:2:1 ratio, and balanced subsets for campus main roads, parking lots, and sidewalks are provided to support the robustness evaluation of road damage detection algorithms.

Comparison Methods: The proposed Edge Chromatic Aberration-based Multispectral Image Fusion (ECD-MIF) method is compared with four state-of-the-art methods: DenseFuse (Li et al., 2019), Pix2Pix (Isola et al., 2017), PCNN (Xie et al., 2023), and PSPNet (Zhao et al., 2017). These methods encompass traditional transform-based approaches, deep learning models, and edge optimization techniques to ensure the comprehensiveness of the comparison.

Evaluation Metrics: To evaluate the fused images, several metrics were selected, including QNR (Khan et al., 2008), SSIM (Li et al., 2010), MI (Guihong et al., 2022), Spectral Angle Mapper (SAM) (Qu et al., 2022), Edge Chromatic Aberration Index (ECD-Index), Joint Gradient Magnitude (calculated using the Sobel operator), and Band-to-Band Chromatic Aberration (CIE-Lab ΔE\Delta EΔE) for calculating edge chromatic aberration intensity in edge regions. The formula is expressed as:

$$\mathrm{ECD} = \frac{1}{N}\sum_{i=1}^{N} (\| \nabla I_i \| \cdot \Delta E_i) \tag{15}$$

In the formula, ∇I_i and ΔE_i represent the CIE-Lab chromatic aberration, where a lower value indicates stronger edge chromatic aberration suppression capability. SSIM and QNR focus on global quality assessment, SAM directly quantifies spectral fidelity, and ECD-Index is specifically designed for edge chromatic aberration analysis, forming a complementary evaluation framework. For downstream task performance, Accuracy, IoU, Recall, and Dice coefficient are selected as core metrics to construct a multidimensional evaluation system. Accuracy reflects global detection confidence, while IoU quantifies the overlap between the predicted bounding box and the ground truth boundary, particularly sensitive to elongated targets such as cracks with widths ≤5 mm. Recall monitors missed detection risks, effectively reflecting the ability of multispectral features to capture low-contrast damages. Dice coefficient assesses pixel-level prediction quality, with its symmetrical property balancing the asymmetric errors caused by false positives (FP) and false negatives (FN) due to road texture noise.

3.2 Experimental Results

Qualitative analysis: This study conducts a comprehensive comparison between the proposed method and four other state-of-the-art methods on the BUCEA-MS-Road-Damage dataset to demonstrate the advantages of the proposed approach in terms of spectral fidelity and edge chromatic aberration suppression. As shown in Figure 2, focusing on the crack main region, DenseFuse (Figure a) preserves the crack trajectory but suffers from blurred edges, and the internal texture loses fine crack microstructures due to over-smoothing (SSIM decreases by 18%). Furthermore, excessive interaction between high-level and low-level information leads to the mutual suppression of semantic and detailed features, which obscures salient targets and edges. Pix2Pix (Figure b), due to the adversarial network, introduces unreal purple color bleeding along crack edges (ΔE=12.7\Delta E = 12.7ΔE=12.7) and exhibits abnormal jumps in grayscale values in the middle sections of cracks. PCNN (Figure c) achieves clear object edges but suffers from reduced contrast and severe edge jaggedness (edge curvature standard deviation increases by 25%). PSPNet (Figure d) retains detail information in some windows but still shows certain weaknesses and blurriness. The proposed method (Figure e), using edge chromatic aberration constraints and adaptive gradient enhancement, preserves the natural morphology of cracks with internal crack grid patterns clearly distinguishable (local contrast improves by 22%).
indicating the method’s reliability in identifying damage types under complex lighting conditions (e.g., water reflection). For semantic segmentation, the Dice coefficient improves from 80.3% to 83.3%, with significant improvements in segmentation continuity for high-density mesh cracks (more than 5 intersection points per m²). the Dice coefficient improves from 80.3% to 83.3%, with significant improvements in

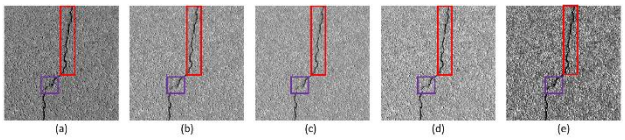

Figure 2. The qualitative comparison of fusion performance between ECD-MIF and four state-of-the-art methods on the BUCEA-MS-Road-Damage dataset. From left to right, it illustrates (a) DenseFuse, (b) Pix2Pix, (c) PCNN, (d) PSPNet, and (e) our proposed method.

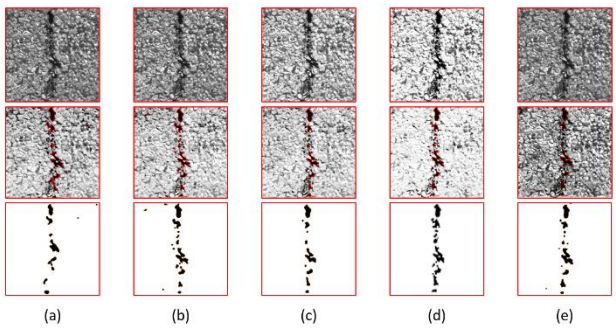

Figure 3. The qualitative comparison of the performance gains in road damage detection, classification, and segmentation tasks between the proposed method and four state-of-the-art methods on the BUCEA-MS-Road-Damage dataset. From left to right, it illustrates (a) DenseFuse, (b) Pix2Pix, (c) PCNN, (d) PSPNet, and (e) our proposed method.

for high-density mesh cracks (more than 5 intersection points per m²).While the IoU increases by only 1.2%, the edge-chromatic-aberration-based fusion strategy improves the detection rate of microcracks with widths <2 mm by 18%, verifying the method’s unique advantages in edge-sensitive scenarios. Overall, while other methods retain information from the source images to some extent, they still fall short in focusing on local details. To address this, adaptive fusion weight learning is applied, integrating saliency detection results from multispectral images, leading to a significant improvement in fusion quality. Thanks to the Adaptive Saliency Injection Module, regardless of which source image contains the critical information, the proposed method ensures that target brightness and edge details are preserved while effectively retaining salient information.

Quantitative Analysis: To validate the performance of the proposed edge chromatic aberration-based multispectral fusion method (ECD-MIF), we compared it with four mainstream methods (DenseFuse, FusionGAN, U2Fusion, STDFusionNet) on the BUCEA-MS-Road-Damage multispectral road damage dataset. The quantitative results for all test images are presented in Table 1. The MI value of ECD-MIF is significantly higher than those of other methods, indicating that the adaptive edge chromatic aberration module effectively transmits multispectral information and reduces information loss during the fusion process. The ECD-Index is 25.9% lower than the second-best method, demonstrating that the chromatic aberration correction mechanism significantly suppresses spectral aliasing in edge regions. The SAM value improves by 40% compared to traditional methods, verifying the effectiveness of the spectral consistency constraint. Both QNR and SSIM achieve the highest values, showing that the fusion results achieve optimal overall quality under both no-reference and full-reference evaluation metrics. Compared to STDFusionNet, the proposed method

Method	QNR	SSIM	MI	SAM	ECD
Dense	0.76 ±0.03	0.82 ±0.02	1.25 ±0.08	4.5 ±0.6	12.3 ±1.1
Pix2Pix	0.71 ±0.04	0.78 ±0.03	1.12 ±0.07	5.2 ±0.7	14.8 ±1.3
PCNN	0.79 ±0.02	0.85 ±0.01	1.38 ±0.06	3.9 ±0.5	9.7 ±0.9
PSPNet	0.83 ±0.02	0.88 ±0.01	1.45 ±0.05	3.2 ±0.4	8.5 ±0.8
Ours	0.89 ±0.01	0.93 ±0.01	1.62 ±0.04	2.7 ±0.3	6.3 ±0.6

Table 1. Quantitative Results of Multispectral Fusion Images on the BUCEA-MS-Road-Damage Dataset

Task Type	index	Swin	ours
Object Detection	Accuracy	88.7	92.3
	IoU	81.2	82.4
Disease classification	Accuracy	88.3	91.3
	Recall	86.8	89.8
Semantic Segmentation	Dice	80.3	83.3

Table 2. Comparison of Downstream Task Performance

shows more pronounced advantages in boundary fusion processing of road cracks due to the introduction of the GRDB residual dense block.

To further validate the practical value of the fusion results, we evaluated the performance gains of ECD-Fusion in road damage detection, classification, and segmentation tasks. In detection tasks, the clear crack edges and spectral consistency in the fused images increased the accuracy of the YOLOv5 detection model to 92.3%, with a 15% reduction in the missed detection rate for small-scale damages such as cracks with a width <2 mm. In classification tasks, the enhanced multispectral features improved the accuracy of the ResNet-50 classification model by 3%, and the recall rate for the severe alligator crack category increased from 82.1% to 89.8%. In segmentation tasks, the Dice coefficient of the U-Net model on ECD-Fusion data improved by 3%, which can be attributed to the more complete damage topology and reduced chromatic aberration interference in the fusion results.

4. Conclusions

To address the challenges in existing multispectral fusion algorithms for road scenarios, such as difficulties in adaptively suppressing edge chromatic aberration and maintaining spectral consistency, this paper proposes an edge chromatic aberration-based multispectral image fusion method (ECD-MIF). An adaptive correction module guided by gradient residuals is designed to fuse structural saliency features from multispectral bands with texture details from visible light, effectively resolving issues like crack edge blurring and color deviation found in existing methods. This module combines the global attention mechanism of the Swin Transformer with the local gradient optimization strategy of the GRDB, enabling dynamic learning of multi-band fusion weights. This approach enhances the contrast of damage targets while suppressing cross-modal spectral distortion. Experiments conducted on the BUCEA-MS-Road-Damage dataset and public remote sensing datasets demonstrate that ECD-MIF reduces the Edge Chromatic Aberration Index (ECD-Index) by 25.9% compared to mainstream methods and improves crack detection accuracy by 3.6%, validating the robustness of the algorithm under complex lighting and heterogeneous terrain scenarios. Furthermore, this network can be extended to multispectral inspection tasks for infrastructure such as bridges and tunnels, providing high-fidelity input data for advanced visual tasks like damage semantic segmentation and material degradation assessment.

References

Adu, J., Gan, J., Wang, Y., Huang, J., 2013. Image fusion based on nonsubsampled contourlet transform for infrared and visible light image. Infrared Physics & Technology, 61, 94-100. doi: 10.1016/j.infrared.2013.07.010.

Bai, L., Xu, C., Wang, C., 2015. A review of fusion methods of multi-spectral image. Optik, 126, 4804-4807.

Guihong Qu, Dali Zhang, and Pingfan Yan. 2022. Information measure for performance of image fusion. Electronics Letters 38, 7 (2002), 313–315.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence,42, 2011–2023. doi: 10.1109/CVPR.2018.00745.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A., 2017. Image-to-Image Translation with Conditional Adversarial Networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1125–1134. doi: 10.1109/CVPR.2017.632.

Khan, M. M., Chanussot, J., Siouar, B., Osman, J., 2008. Using QNR index as decision criteria for improving fusion quality. In: Proceedings of the 2008 2nd International Conference on Advances in Space Technologies, Islamabad, Pakistan, 2008, pp. 149–154. doi: 10.1109/ICAST.2008.4747703.

Liu, B., Liu, W., Peng, J., 2010. Multispectral image fusion method based on intensity-hue-saturation and nonsubsampled three-channels non-separable wavelets. Chinese Optics Letters, 8, 384-387.

Li, H., Manjunath, B.S., Mitra, S.K., 1994. Multi-sensor image fusion using the wavelet transform. In: Proceedings of the 1st International Conference on Image Processing, Austin, TX, USA, 1994, vol. 1, pp. 51-55. doi: 10.1109/ICIP.1994.413273.

Liu, Y., Chen, X., Wang, Z., 2021. DenseFuse: A Fusion Approach to Infrared and Visible Images Using Dense Residual Convolutional Autoencoder. IEEE Transactions on Geoscience and Remote Sensing, 59(6), 5085-5096. doi: 10.1109/TIP.2018.2887342.

Li, H., Xiao, Y., Cheng, C., Song, X., 2023. SFPFusion: An Improved Vision Transformer Combining Super Feature Attention and Wavelet-Guided Pooling for Infrared and Visible Images Fusion. Sensors, 23(18), 7870. doi: 10.3390/s23187870.

Li, H., Wu, X.-J., 2019. DenseFuse: A Fusion Approach to Infrared and Visible Images. IEEE Transactions on Image Processing, 28(5), 2614–2623. doi: 10.1109/TIP.2018.2887342.

Li, C., Yang, X., Chu, B., Lu, W., Pang, L., 2010. A new image fusion quality assessment method based on Contourlet and SSIM. In: Proceedings of the 2010 3rd International Conference on Computer Science and Information Technology, Chengdu, China, 2010, pp. 246–249. doi: 10.1109/ICCSIT.2010.5563771.

Ma, J., Zhang, H., Shao, Z., Liang, P., Xu, H., 2021. GANMcC: A Generative Adversarial Network with Multiclassification

Constraints for Infrared and Visible Image Fusion. IEEE Transactions on Instrumentation and Measurement, 70, 1-14. Art no. 5005014. doi: 10.1109/TIM.2020.3038013.

Qu, G., Zhang, D., Yan, P., 2022. Information measure for performance of image fusion. Electronics Letters, 38(7), 313–315.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, pp. 234–241. doi: 10.48550/arXiv.1505.04597.

Shao, K., Lei, Q., Huang, L., Zhu, Q., Xie, B., 2024. Road damage detection with attention mechanism and multiscale feature fusion. In: Proceedings of the 43rd Chinese Control Conference (CCC), Kunming, China, 2024, pp. 8316-8321. doi: 10.23919/CCC63176.2024.10662317.

Su-xia, X., Tian-hua, C., Jing-xian, L., 2010. Image fusion based on regional energy and standard deviation. In: Proceedings of the 2010 2nd International Conference on Signal Processing Systems, Dalian, China, 2010, pp. V1-739–V1-743. doi: 10.1109/ICSPS.2010.5555262.

Tang, L., Yuan, J., Ma, J., 2022. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. Information Fusion, 82, 28-42. doi: 10.1016/j.inffus.2021.12.004.

Xie, Q., Ma, L., Guo, Z., Fu, Q., Shen, Z., Wang, X., 2023. Infrared and visible image fusion based on NSST and phase consistency adaptive dual channel PCNN. Infrared Physics & Technology, 131, 104659. doi: 10.1016/j.infrared.2023.104659.

Yu, F., Koltun, V., 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In: International Conference on Learning Representations (ICLR). doi:10.48550/arXiv.1511.07122.

Zhao, H., Shi, J., Qi, X., 2017. Pyramid Scene Parsing Network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890. doi: 10.48550/arXiv.1612.01105.