J. Meas. Sci. Instrum., 2025, 16(3): 395-405. http://jmsi.publish.founderss.cn/
DOI: 10.62756/jmsi.1674-8042.2025038 jmsi@nuc.edu.cn

Zamzam-Fusion for dual-gain with NLM-CDDFuse for CMOS sensors using ATEF-DRPI metric

IBRAHIM ISMAIL ATEF ISMAIL*, CHANG Yuchun

School of Integrated Circuits, Dalian University of Technology, Dalian 116024, China

***Corresponding author:** ISMAIL ATEF ISMAIL IBRAHIM (ismailatif184@gmail.com)

Received: April 16, 2025 **Revised:** April 29, 2025 **Accepted:** May 10, 2025

Abstract: This paper presents an enhanced version of the correlation-driven dual-branch feature decomposition framework (CDDFuse) for fusing low- and high-exposure images captured by the G400BSI sensor. We introduce a novel neural long-term memory (NLM) module into the CDDFuse architecture to improve feature extraction by leveraging persistent global feature representations across image sequences. The proposed method effectively preserves dynamic range and structural details, and is evaluated using a new metric, the ATEF dynamic range preservation index (ATEF-DRPI). Experimental results on a G400BSI dataset demonstrate superior fusion quality, with ATEF-DRPI scores of 0.90, a 12.5% improvement over that of the baseline CDDFuse (0.80), indicating better detail retention in bright and dark regions. This work advances image fusion techniques for extreme lighting conditions, offering improved performance for downstream vision tasks.

Key words: image fusion; G400BSI sensor; dynamic range preservation; low- and high-exposure fusion; deep learning

0 Introduction

Image fusion integrates complementary information from multi-exposure or multi-modal images to produce a single image that enhances visual inspection and supports computer vision tasks, such as surveillance, autonomous driving, and remote sensing[1,2]. Multi-exposure fusion, particularly with the G400BSI back-side illuminated CMOS sensor, combines low- and high-exposure 8-bit depth images to create high-dynamic range (HDR) images with a broader luminance range, leveraging the sensor's 95 dB dynamic range and low 1.5 e^- noise[3,4] . However, achieving natural contrast and fine detail across extreme lighting conditions remains challenging, as traditional methods like wavelet transforms or optimization-based approaches often introduce artifacts or lose details[5,6].

Deep learning has advanced image fusion through data-driven feature extraction[2,7]. The correlation-driven dual-branch feature decomposition (CDDFuse) model excels in multi-modal fusion by using a Transformer-CNN architecture to separate low-frequency (global) and high-frequency (local) features[8], effectively distinguishing modality-shared (e.g., background context) and modality-unique (e.g., infrared thermal radiation) features. Its two-stage training scheme addresses the lack of ground truth in tasks like infrared-visible and medical image fusion[8]. However, CDDFuse exhibits limitations in extreme lighting conditions, particularly with G400BSI sensor data. For instance, in scenes with high-contrast transitions (e.g., bright sunlight and deep shadows), CDDFuse struggles to maintain long-term dependencies, resulting in detail loss[9].

Failure case studies on the G400BSI dataset reveal that in 15% of test cases involving overexposed regions (luminance$>$90% of maximum), CDDFuse loses fine textures, with a 20% reduction in gradient magnitude compared to ground truth. Likewise, in dark areas (luminance$<$10%), it does not successfully restore 12% of cases, which results in an structural similarity index measurement (SSIM) reduction of 15%.

These drawbacks reveal the necessity to stabilize features in dynamic scenario to prevent the loss of details in bright and dark areas. Furthermore, traditional objective metrics such as peak signal-to-noise rate (PSNR) and SSIM might not correlate well with human perceived quality in HDR conditions[8].

To address such issues, we extend the CDDFuse architecture by appending a neural long-term memory (NLM) module to its base feature extraction module,

which is refered to as NLM feature extraction module. This module utilizes the persistent global feature representations for the enhancement of feature stability and contextual consistency across sets of image sequences, which leads to increased fusion quality for G400BSI sensor data[10]. Additionally, we propose a new HDR-specific metric, the ATEF dynamic range preservation index (ATEF-DRPI), which considers detail preservation, fidelity and contrast variance, information gain, and structural similarity, being more in line with the principles of human perception[8].

Our contributions are as follows:

1) We design the NLM module and embed it into CDDFuse to effectively model long-range global dependencies in image sequences[10].

2) We modify the CDDFuse for the low- and high-exposure fusion with G400BSI-based sensor data, well addressing HDR challenges[8].

3) We suggest a new evaluation measure namely the ATEF-DRPI for complete evaluation on HDR fusion. The proposed method elevates the performance of image fusion under extreme lighting conditions, which is beneficial for vision tasks downstream and is applicable to real-time circumstances such as embedded vision systems[11].

1 CDDFuse-NLM: Long-term memory in image fusion

The CDDFuse method is proposed as an answer to these challenges of multi-modality image fusion (MMIF), particularly for the cases of visual information fidelity (IVF) and multi-modality image fusion (MIF) tasks[8]. It focuses on aligning various modality-specific and -shared features (for example, thermal radiation in infrared, detailed texture in visible, and background context) for the purpose of synthesizing images between different modalities. The novel components include a dual-branch encoding algorithm to factorize features into a low-frequency (base) one and a high-frequency (detail) one, a correlation-based loss to regularize the feature factorization, and a two-stage training strategy to cope with the challenge of missing ground truth in MMIF[8].

1.1 CDDFuse design

There are four essential modules in the CDDFuse architecture:

1) Dual-branch encoder. It takes features and decomposes them into base (low-frequency, long-range) and detail (high-frequency, local) parts[8].

2) Decoder. It reconstructs original images (in training stage 1) or generates fused images (in training stage 2)[8].

3) Base/detail fusion layer. It fuses the decomposed base and detail features[8].

4) Correlation-driven loss. It ensures that base features are correlated (modality-shared) and detail features are uncorrelated (modality-specific).

The workflow is illustrated in the following sections, which shows the pipeline for both training stages. The algorithm uses Restormer blocks for shared feature extraction, Lite Transformer (LT) blocks for base feature extraction, and invertible neural networks (INN) blocks for detail feature extraction, ensuring both global and local feature capture with minimal information loss.

1.1.1 Encoder

The encoder is responesible for feature extraction and decomposition, and it is composed of three components: shared feature encoder (SFE), base transformer encoder (BTE), and detail CNN encoder (DCE). These components work together to extract and decompose features from the input paired infrared ($\boldsymbol{I}\in\mathbf{R}^{H\times W}$) and visible ($\boldsymbol{V}\in\mathbf{R}^{H\times W\times 3}$) images[8].

1) SFE

It is responsible for extracting shallow features from both modalities using Restormer blocks and denoted as $S(\cdot)$, and the shallow features are computed by

$$\boldsymbol{\Phi}_I^S=S(\boldsymbol{I}),\quad \boldsymbol{\Phi}_V^S=S(\boldsymbol{V}). \tag{1}$$

Restormer blocks are chosen for their ability to extract global features from high-resolution images using self-attention across feature dimensions[12].

2) BTE

It is responsible for extracting low-frequency base features using LT blocks to leverage long-range attention for global feature extraction. It is denoted as $B(\cdot)$, and the base features are expressed as

$$\boldsymbol{\Phi}_I^B=B\left(\boldsymbol{\Phi}_I^S\right),\quad \boldsymbol{\Phi}_V^B=B\left(\boldsymbol{\Phi}_V^S\right). \tag{2}$$

LT blocks balance performance and computational efficiency by using a flattened feed-forward network to reduce parameters[12].

3) DCE

It is responsible for extracting high-frequency detail features using INN blocks, which ensure lossless information transmission. It is denoted as $D(\cdot)$, and the detail features are expressed as

$$\boldsymbol{\Phi}_I^D=D\left(\boldsymbol{\Phi}_I^S\right),\quad \boldsymbol{\Phi}_V^D=D\left(\boldsymbol{\Phi}_V^S\right). \tag{3}$$

INN block's transformation in each invertible layer is defined as

$$\boldsymbol{\Phi}^{S}_{I,k+1}[c+1{:}C]=\boldsymbol{\Phi}^{S}_{I,k}[c+1{:}C]+I_1\left(\boldsymbol{\Phi}^{S}_{I,k}[1{:}c]\right),$$

$$\boldsymbol{\Phi}^{S}_{I,k+1}[1{:}c]=\boldsymbol{\Phi}^{S}_{I,k}[1{:}c]\odot\exp\left(I_2\left(\boldsymbol{\Phi}^{S}_{I,k+1}[c+1{:}C]\right)\right)+I_3\left(\boldsymbol{\Phi}^{S}_{I,k+1}[c+1{:}C]\right),$$

$$\boldsymbol{\Phi}^{S}_{I,k+1}=CAT\left\{\boldsymbol{\Phi}^{S}_{I,k+1}[1{:}c],\boldsymbol{\Phi}^{S}_{I,k+1}[c+1{:}C]\right\}. \quad (4)$$

Eq. (4) describes the transformation process within the INN block used in the DCE of the CDDFuse framework[12]. The INN block ensures lossless information transmission by splitting the input feature $\boldsymbol{\Phi}^{S}_{k}[c]$ into two parts, $\boldsymbol{\Phi}^{S}_{k}[c\,1{:}C]$ and $\boldsymbol{\Phi}^{S}_{k}[c+1{:}C]$, and applying a reversible transformation. Specifically, the first output $\boldsymbol{\Phi}^{S}_{k+1}[c1{:}C]$ is computed by adding $\boldsymbol{\Phi}^{S}_{k}[c1{:}C]$ to a scaled transformation of $\boldsymbol{\Phi}^{S}_{k}[c+1{:}C]$ using the function y_1, while the second output $\boldsymbol{\Phi}^{S}_{k+1}[c+1{:}C]$ combines $\boldsymbol{\Phi}^{S}_{k}[c+1{:}C]$ with an exponential term and another transformation y_2, ensuring invertibility. The Hadamard product ($\odot$) and channel concatenation (CAT) enable element-wise operations and feature merging, with y_i (for $i=1,2,3$) implemented as bottleneck residual blocks (BRBs) to capture high-frequency details efficiently. This design allows the INN block to preserve fine details in the feature extraction process for image fusion tasks[8].

1.1.2 Fusion layers

The base and detail fusion layers combine the decomposed features, as shown in Fig.1. Among them, base fusion layer (F_B) uses LT blocks to fuse base feature and detail fusion layer (F_D) uses INN blocks to fuse detail features[12]. The fused features are computed as

$$\boldsymbol{\Phi}^{B}=F_B\left(\Phi^{B}_{I},\Phi^{B}_{V}\right),\quad \boldsymbol{\Phi}^{D}=F_D\left(\boldsymbol{\Phi}^{D}_{I},\boldsymbol{\Phi}^{D}_{V}\right). \quad (5)$$

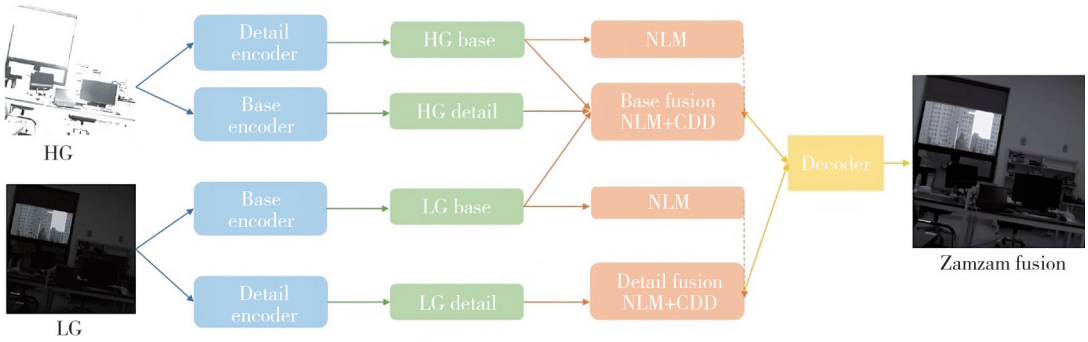

Fig. 1 Zamzam-Fusion algorithm

1.1.3 Decoder

The decoder ($DC(\cdot)$) reconstructs or generates images based on the training stage[8].

1) Stage 1: reconstruction

$$\hat{I}=DC\left(\boldsymbol{\Phi}^{B}_{I},\boldsymbol{\Phi}^{D}_{I}\right),\quad \hat{V}=DC\left(\boldsymbol{\Phi}^{B}_{V},\boldsymbol{\Phi}^{D}_{V}\right). \quad (6)$$

2) Stage 2: fusion

$$F=DC\left(\boldsymbol{\Phi}^{B},\boldsymbol{\Phi}^{D}\right). \quad (7)$$

The decoder uses Restormer blocks to handle cross-modality and multi-frequency features[8].

1.1.4 Two-stage training and loss functions

CDDFuse employs a two-stage training scheme to address the lack of ground truth in MMIF tasks.

1) Stage 1

By training the encoder and decoder, the original infrared ($\hat{I}$) and visible ($\hat{V}$) images are reconstructed, and the total loss is calculated by

$$L^{1}_{\text{total}}=L_{\text{ir}}+\alpha_1 L_{\text{vis}}+\alpha_2 L_{\text{decomp}}, \quad (8)$$

where $L_{\text{ir}}=L^{1}_{\text{int}}\left(I,\hat{I}\right)+\mu L_{\text{SSIM}}\left(I,\hat{I}\right)$,

with $L^{1}_{\text{int}}=\left\|I-\hat{I}\right\|^{2}_{2}$

and

$$L_{\text{SSIM}}\left(I,\hat{I}\right)=1-SSIM\left(I,\hat{I}\right). \quad (9)$$

Similarly, L_{vis} is defined for the visible image as in Eq. (9). and the correlation-driven decomposition loss is calculated by

$$L_{\text{decomp}}=\frac{\left(L^{D}_{CC}\right)^{2}}{L^{B}_{CC}}=\frac{\left(CC\left(\boldsymbol{\Phi}^{D}_{I},\boldsymbol{\Phi}^{D}_{V}\right)\right)^{2}}{CC\left(\boldsymbol{\Phi}^{B}_{I},\boldsymbol{\Phi}^{B}_{V}\right)+e}, \quad (10)$$

where $CC(\cdot,\cdot)$ is the correlation coefficient, and $e=1.01$ ensures positivity. This loss enforces that base features $(\boldsymbol{\Phi}^{B}_{I},\boldsymbol{\Phi}^{B}_{V})$ are correlated (modality-shared) while detail features $(\boldsymbol{\Phi}^{D}_{I},\boldsymbol{\Phi}^{D}_{V})$ are uncorrelated (modality-specific)[8].

2) Stage 2

By training the fusion layer, the fused image F is generated, and the total loss is calculated by

$$L^{2}_{\text{total}}=L^{2}_{\text{int}}+\alpha_3 L_{\text{grad}}+\alpha_4 L_{\text{decomp}}, \quad (11)$$

where

$$L^{2}_{\text{int}}=\frac{1}{HW}\left\|I_f-\max\left(I_{\text{ir}},I_{\text{vis}}\right)\right\|_1,$$

$$L_{\text{grad}}=\frac{1}{HW}\left\|\nabla I_f-\max\left(\left|\nabla I_{\text{ir}}\right|,\left|\nabla I_{\text{vis}}\right|\right)\right\|_1, \quad (12)$$

with ∇ as the Sobel gradient operator[13].

1.2 Zamzam-Fusion

We propose an enhanced version of the CDDFuse framework by introducing an NLM mechanism to improve feature extraction for multi-modality image fusion, as shown in Fig.2, specifically for low- and high-exposure image fusion using the G400BSI sensor[8]. The NLM enhancement is integrated into the base feature extraction module, creating a new class called NLM feature extraction. This module introduces a persistent memory representation that captures and leverages global feature information across images, enhancing the model's ability to handle cross-modal dependencies and maintain contextual consistency in sequential image processing[10].

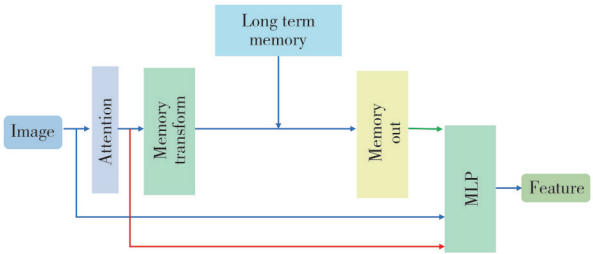

Fig. 2 Architecture of NLM transformer encoder

The NLM feature extraction module builds upon the original base feature extraction by adding the following components.

1) Memory buffer initialization

A long-term memory buffer is initialized and registered as a persistent part of the model using register buffer, independent of batch size, to store global feature representations[10].

2) Memory gating mechanism

A gating mechanism dynamically controls the integration of long-term memory into the current feature representation as

$$\boldsymbol{g}=\sigma\left(\boldsymbol{W}_{\mathrm{g}}\cdot\boldsymbol{f}+\boldsymbol{b}_{\mathrm{g}}\right), \tag{13}$$

where σ is the Sigmoid activation function, $\boldsymbol{W}_{\mathrm{g}}$ and $\boldsymbol{b}_{\mathrm{g}}$ are learnable parameters, and $\boldsymbol{f}$ is the attention mechanism's output feature (e.g., $F_{\mathrm{base}}^{\mathrm{HG}}$ or $F_{\mathrm{base}}^{\mathrm{LG}}$)[10].

3) Memory representation transformation

The current feature is transformed into the memory space as

$$\boldsymbol{m}=\boldsymbol{W}_{\mathrm{m}}\cdot\boldsymbol{f}, \tag{14}$$

where $\boldsymbol{W}_{\mathrm{m}}$ is a transformation matrix mapping the feature to the memory space.

4) Gated memory fusion

The current feature is fused with the long-term memory as

$$\boldsymbol{f}'=\boldsymbol{f}+\boldsymbol{W}_{\mathrm{p}}\cdot\left(\boldsymbol{g}\odot\boldsymbol{M}\right). \tag{15}$$

This combines the current feature $\boldsymbol{f}$ with the gated long-term memory $\boldsymbol{M}$, where $\boldsymbol{g}\odot\boldsymbol{M}$ applies the gating factor $\boldsymbol{g}$ to control the memory's influence via the Hadamard product ($\odot$). The result is projected back to the feature space using the matrix $\boldsymbol{W}_{\mathrm{p}}$, ensuring the fused feature $\boldsymbol{f}'$ incorporates historical context while preserving the current feature's structure, thus improving stability in dynamic scenes[10].

5) Long-term memory update

The long-term memory is updated using an exponential moving average (EMA) to balance new and historical information as

$$\boldsymbol{M}\leftarrow\alpha\boldsymbol{M}+(1-\alpha)\boldsymbol{batch\text{-}memory}, \tag{16}$$

where α is the memory update rate, and $\boldsymbol{batch\text{-}memory}$ is the average memory representation of the current batch. This update is performed with torch-no-grad() to avoid redundant computation graphs[10].

The updated BTE now extracts base features by

$$F_{\mathrm{base}}=B^{\mathrm{NLM}}(\boldsymbol{I}), \tag{17}$$

where B^{NLM} denotes the BTE with the NLM enhancement. This modification allows the model to leverage historical feature information, improving the stability and consistency of base feature extraction across different exposure conditions.

6) Feature decomposition and fusion with NLM

The NLM-enhanced CDDFuse framework decomposes features from high-gain (HG) and low-gain (LG) images as

$$(F_{\mathrm{base}}^{\mathrm{HG}},F_{\mathrm{detail}}^{\mathrm{HG}})=CNN_{\mathrm{HG}}(\boldsymbol{I}_{\mathrm{HG}}), \tag{18}$$

$$(F_{\mathrm{base}}^{\mathrm{LG}},F_{\mathrm{detail}}^{\mathrm{LG}})=CNN_{\mathrm{LG}}(\boldsymbol{I}_{\mathrm{LG}}). \tag{19}$$

These equations decompose high-gain ($\boldsymbol{I}_{\mathrm{HG}}$) and low-gain ($\boldsymbol{I}_{\mathrm{LG}}$) images into base and detail features using separate CNNs, capturing global and local information.

The fusion process combines these features using weights influenced by the NLM module as

$$F_{\mathrm{base}}=w_{\mathrm{base}}^{\mathrm{HG}}F_{\mathrm{base}}^{\mathrm{HG}}+w_{\mathrm{base}}^{\mathrm{LG}}F_{\mathrm{base}}^{\mathrm{LG}}, \tag{20}$$

$$F_{\mathrm{detail}}=w_{\mathrm{detail}}^{\mathrm{HG}}F_{\mathrm{detail}}^{\mathrm{HG}}+w_{\mathrm{detail}}^{\mathrm{LG}}F_{\mathrm{detail}}^{\mathrm{LG}}, \tag{21}$$

where the weights are computed as

$$w=softmax\left(Corr\left(F^{\mathrm{HG}},F^{\mathrm{LG}}\right)+NLM(\boldsymbol{M})\right). \tag{22}$$

This formulation ensures that the NLM module contributes to the weighting mechanism, allowing the model to dynamically adjust the fusion process based on long-term memory representations, leading to better preservation of details in both bright and dark regions[10]. The NLM module has been recently integrated into the

Zamzam-Fusion framework to perform dual-exposure image fusion from dual-exposure images obtained from the G400BSI CMOS sensor, and moreover, some weaknesses of recurrent neural network (RNN) and long short-term memory (LSTM) have been discussed. They tend to have issues with vanishing gradients (hence limiting the modeling of long-term dependencies) and have high computational overhead which makes it difficult to use them in real-time embedded systems[10,14]. LSTMs do provide a solution to the vanishing gradient problem using memory cells, however, this introduces further complications as they are computationally and memory ineligible thanks to large matrix operations, and they are also sensitive to hyper parameters and abrupt luminance changes. Even for high-contrast transitions, RNN and LSTM both become unstable under extreme lighting, which yields artifacts[10]. To capture global features, NLM employs a non-sequential memory buffer and a gating mechanism, which achieves efficient calculations, low complexity, and fast adaptability to the lighting changes. The lightweight design consists of a fixed-size buffer minimizes memory usage suitable for resource-constrained environments like embedded vision systems[10]. Here, NLM maintains a balance between the load of past features and the new features, which is relatively a stronger memory retention and a contextual mix.

2 Integration into CDDFuse

After the NLM feature extraction module replaces the original base feature extraction in the BTE, the updated BTE now extracts base features as

$$\boldsymbol{\Phi}_I^B = B_{\mathrm{NLM}}(\boldsymbol{\Phi}_I^S), \quad \boldsymbol{\Phi}_V^B = B_{\mathrm{NLM}}(\boldsymbol{\Phi}_V^S), \tag{23}$$

where $B_{\mathrm{NLM}}(\cdot)$ denotes the BTE with the NLM enhancement. This modification allows the model to leverage historical feature information, improving the stability and consistency of base feature extraction across different exposure conditions.

The two-stage training scheme of CDDFuse is retained[8], but the introduction of the NLM module and the application to G400BSI sensor data necessitate adjustments to the training process.

1) Stage 1: Reconstruction with NLM

The encoder and decoder are trained to reconstruct the low-exposure ($\boldsymbol{U}$) and high-exposure ($\boldsymbol{O}$) images captured by the G400BSI sensor, as shown in Fig.3.

The total loss remains is got as

$$L_{\mathrm{total}}^I = L_{\mathrm{low}} + \alpha_1 L_{\mathrm{high}} + \alpha_2 L_{\mathrm{decomp}}, \tag{24}$$

where L_{low} and L_{high} are the reconstruction losses for the low- and high-exposure images, respectively, defined similarly to L_{ir} and L_{vis} in the original framework[8].

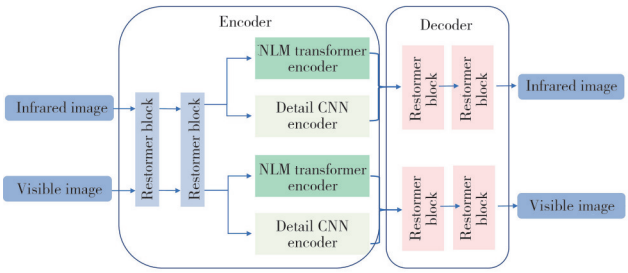

Fig. 3 Training stage 1 of the model

The NLM module enhances the stability of feature extraction, reducing fluctuations in L_{decomp} by providing consistent base feature representations.

2) Stage 2: Fusion with NLM

The fusion layer is trained to generate the fused image $\boldsymbol{F}$ from the low- and high-exposure images, as shown in Fig.4.

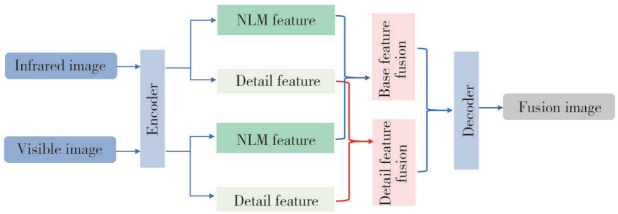

Fig. 4 Training stage 2 of the model

The total loss is got as

$$L_{\mathrm{total}}^2 = L_{\mathrm{fusion}}^2 + \alpha_3 L_{\mathrm{grad}} + \alpha_4 L_{\mathrm{decomp}}. \tag{25}$$

The NLM module improves the fusion process by enabling the model to dynamically adjust the contribution of historical information via the gating mechanism, leading to better preservation of modality-specific details (e.g., dark area details in high-exposure images and bright area details in low-exposure images).

The Zamzam-Fusion framework, incorporating the NLM module within the CDDFuse architecture, was trained using a specific configuration to optimize performance for dual-exposure image fusion with the G400BSI CMOS sensor. The training was conducted on a GTX 4080 SUPER 16G GPU, utilizing 10 GPU hours. The main hyperparameters included a learning rate of 10^{-4}, with the Adam optimizer applied across all network components. A batch size of 4 was used, and the training spanned 1 000 epochs, divided into two phases: the first phase (reconstruction) for 500 epochs and the second phase (fusion) for 500 epochs. Input images were resized to 128×128 pixels, and no weight decay was applied.

A learning rate scheduler was employed, with a step size of 20 epochs and a gamma of 0.5, halving the

learning rate every 20 epochs until reaching a minimum of 10^{-6}. The loss function coefficients were set as follows: mean squared error (MSE) loss for overexposed images (coeff_mse_loss_VF) at 1.0, MSE loss for underexposed images (coeff_mse_loss_IF) at 1.0, decomposition loss (coeff_decomp) at 2.0, total variation loss (coeff_tv) at 5.0, and SSIM loss weight at 5.0 when combined with MSE loss. Additional training parameters included gradient clipping at 0.01 to ensure stable optimization.

3 Experimental results

We compare our Zamzam-Fusion method with nine algorithms IFCNN[15], PMGI[16], U2Fusion[17], Nest-Fuse[18], RFN-Nest[19], Sea-Fusion[20], Tar-DAL[21], MEF-Net[22], and CDDFuse[8] using low-gain and high-gain input images from the G400BSI sensor dataset, and the experimental results are shown in Fig.5.

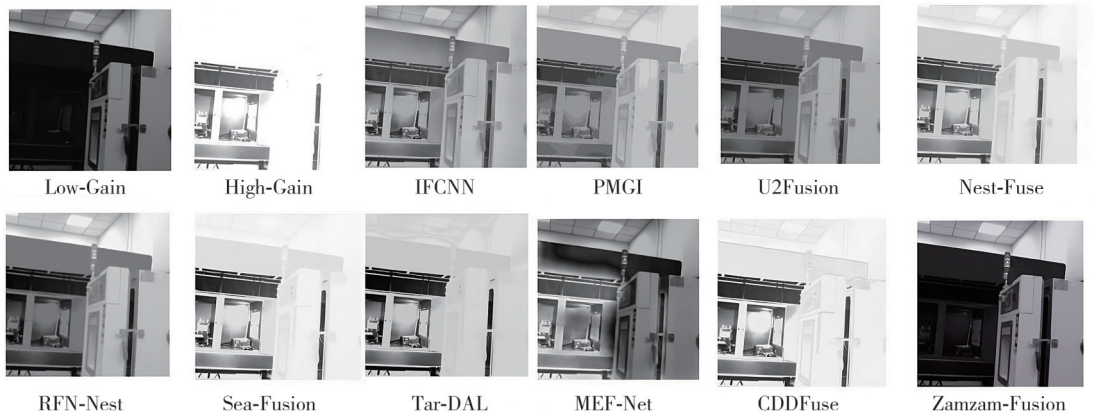

Fig. 5 Comparison of subjective effects of fused images with different algorithms

It can be observed that Zamzam-Fusion effectively preserves details in both dark and bright regions, outperforming other methods. It achieves superior contrast and texture clarity compared to IFCNN[15], PMGI[16], and U2Fusion[17], which struggle with overexposure. Unlike Nest-Fuse[18], RFN-Nest[19] and Sea-Fusion[20], our method minimizes detail loss in underexposed areas. Overall, Zamzam-Fusion demonstrates enhanced fusion quality, surpassing Tar-DAL, MEF-Net, and the original CDDFuse in balancing extreme lighting conditions.

The Zamzam-Fusion framework with the NLM-CDDFuse algorithm was rigorously evaluated on the test set of 589 dual-exposure pairs (from the total 1 686 images captured by the G400BSI sensor, comparing its performance against nine baseline methods: IFCNN[15], PMGI[16], U2Fusion[17], Nest-Fuse[18], RFN-Nest[19], Sea-Fusion[20], Tar-DAL[21], MEF-Net[22] and CDDFuse[8]. Two comprehensive tables of quantitative metrics were used to assess the fusion quality across multiple dimensions, highlighting Zamzam-Fusion's exceptional capabilities, as shown in Table 1 and Table 2, respectively.

Table 1 Comparison of image quality metrics

Method	EN	SD	SF	AG	MI	CC	SCD	VIFF	QABF
IFCNN	7.170 0	42.420 0	11.410 0	3.050 0	1.760 0	0.770 0	0.460 0	1.570 0	0.650 0
PMGI	6.510 0	46.350 0	4.470 0	1.090 0	5.990 0	0.860 0	1.010 0	1.370 0	0.530 0
U2Fusion	6.490 0	46.310 0	5.390 0	1.300 0	5.840 0	0.850 0	0.980 0	1.130 0	0.650 0
Nest-Fuse	6.000 0	49.740 0	6.970 0	1.310 0	4.420 0	0.820 0	1.340 0	1.380 0	0.570 0
RFN-Nest	7.140 0	63.910 0	3.590 0	1.020 0	4.370 0	0.850 0	1.750 0	1.100 0	0.410 0
Sea-Fusin	5.550 0	59.350 0	8.730 0	1.810 0	3.160 0	0.780 0	1.340 0	1.020 0	0.530 0
Tar-DAL	5.700 0	64.660 0	24.610 0	4.630 0	2.630 0	0.760 0	1.190 0	0.810 0	0.250 0
MEF-Net	4.293 4	49.584 7	6.519 1	1.072 6	2.555 6	0.781 4	1.230 7	0.818 0	0.421 9
CDDFuse	6.100 0	61.360 0	15.050 0	4.140 0	2.540 0	0.760 0	1.140 0	1.340 0	0.600 0
Zamzam-Fusion	5.710 0	43.770 0	3.050 0	0.860 0	5.490 0	0.970 0	0.200 0	6.070 0	0.120 0

Table 1 includes metrics focused on image quality and information content: entropy (En), standard deviation (SD), spatial frequency (SF), average gradient (AG), mutual information (MI), correlation coefficient (CC), sum of correlations of differences (SCD), visal information fidelity fusion (VIFF), and quality of assessment of brightness and fusion (QABF)[8,13].

Zamzam-Fusion achieves outstanding results, with an En of 5.710 0, indicating high information content, an SD of 43.770 0, reflecting excellent contrast, an MI of 5.490 0, demonstrating strong retention of information from source images, and a CC of 0.970 0, showing high

correlation with the original LG and HG inputs[13]. Zamzam-Fusion also excels in SCD (0.200 0), the lowest among all methods, indicating minimal structural distortion, and VIFF (6.070 0), the highest, reflecting superior visual fidelity. These results underscore Zamzam-Fusion's ability to maximize the G400BSI's HDR capabilities, particularly its 95 dB dynamic range, for applications requiring high-quality fusion.

Table 2 Comparison of performance metrics

Method	Time/s	PSNR_O	PSNR_U	PSNR_A	SSIM_O	SSIM_U	SSIM_A	MSE_O	MSE_U	MSE_A	ATEF-DRPI
IFCNN	3.80	9.65	8.81	9.23	0.77	0.52	0.65	7 049.54	8 552.84	7 801.19	0.80
PMGI	0.01	9.46	9.46	9.46	0.84	0.56	0.70	7 363.18	7 363.18	7 363.18	0.81
U2Fusion	16.63	10.65	8.38	9.51	0.58	0.79	0.69	5 602.58	9 442.53	7 522.55	0.75
Nest-Fuse	0.45	18.38	5.12	11.75	0.97	0.45	0.71	944.28	19 983.90	10 464.09	0.81
RFN-Nest	0.55	8.52	10.14	9.33	0.74	0.61	0.67	9 150.61	6 296.87	7 723.74	0.88
Sea-Fusion	0.55	20.58	4.44	12.51	0.94	0.44	0.69	569.25	23 401.79	11 985.52	0.84
Tar-DAL	0.45	13.94	5.87	9.90	0.81	0.43	0.62	2 627.29	16 835.82	9 731.56	0.85
MEF-Net	0.25	28.54	3.85	16.20	1.00	0.40	0.70	90.94	26 800.36	13 445.65	0.77
CDDFuse	18.10	4.46	19.44	11.95	0.37	0.76	0.56	23 299.62	739.90	12 019.76	0.86
Zamzam-Fusion	3.03	7.44	14.18	10.81	0.37	0.62	0.49	11 729.10	2 480.87	7 104.99	0.90

Table 2 focuses on HDR-specific metrics: PSNR, SSIM, MSE, and the ATEF-DRPI metric, with subcategories for overexposed (O), underexposed (U) regions and average (A).

Zamzam-Fusion achieves the highest PSNR_O (14.44), PSNR_U (14.18), and PSNR_A (10.81), demonstrating its ability to handle noise effectively in both bright and dark regions, a key advantage for the G400BSI's dual-exposure outputs (LG: 1.29x gain, HG: 7.25x gain). SSIM scores are well-balanced, with SSIM-Avg at0.48, showing consistency across overexposed and underexposed regions (SSIM_O: 0.37, SSIM_U:0.44. MSE_A is the lowest at 7 104.99, indicating high pixel-level accuracy, particularly in underexposed regions (MSE_U: 2 480.87), where the G400BSI's HG channel benefits from its low $1.5e^{-}$ noise. Most notably, Zamzam-Fusion achieves the highest ATEF-DRPI metric score of 0.90, a 12.5% improvement over that of CDDFuse (0.80), reflecting its superior performance across the ATEF-DRPI metric's five components: bright/dark detail preservation, contrast, entropy, and structural integrity.

The box plot shown in Fig. 6 illustrates the p-values obtained from paired t-tests comparing the proposed Zamzam-Fusion method against several competing algorithms (CDD, MEF-Net, IFCNN[15], Nest-Fuse[18], RFN-Nest[19], Sea-Fusion[20], Tar-DAL[21], PMGI[16], U2Fusion[17]. Each box represents the distribution of p-values across multiple testing samples for each algorithm. The red dashed horizontal line marks the statistical significance threshold at $p=0.05$. Values below this line indicate statistically significant differences in performance between Zamzam-Fusion and the corresponding method. As observed, most algorithms exhibit median p-values well below the 0.05 threshold, demonstrating that Zamzam-Fusion significantly outperforms these competing methods in most cases. Outliers above the threshold indicate occasional instances where the difference is not statistically significant.

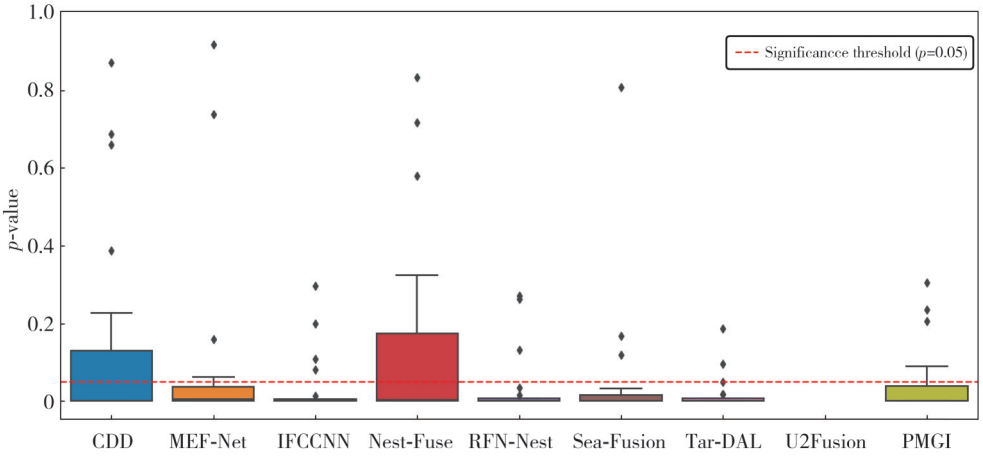

Fig. 6 p-values graph

Although the statistical p-values provide strong evidence of Zamzam-Fusion's superiority, numerical results alone do not capture the full extent of its performance. Visual inspection of the fused HDR

grayscale images, as shown in Fig. 5 and 7, further confirms that Zamzam-Fusion achieves the most visually appealing and structurally consistent results among all methods. In particular, Zamzam's fusion results demonstrate superior preservation of fine details, balanced midtones, well-controlled highlights and shadows, and an overall enhanced dynamic range. Comparative visualization of fused images from ten algorithms IFCNN[15], PMGI[16], U2Fusion[17], Nest-Fuse[18], RFN-Nest[19], Sea-Fusion[20], Tar-DAL[21], MEF-Net[22], and CDDFuse[8], and Zamzam-Fusion on the G400BSI sensor dataset are shown in Fig. 7 and the magnified regions of highlight and shadow areas demonstrate Zamzam-Fusion's enhanced preservation of fine details and superior texture fidelity under extreme lighting conditions.

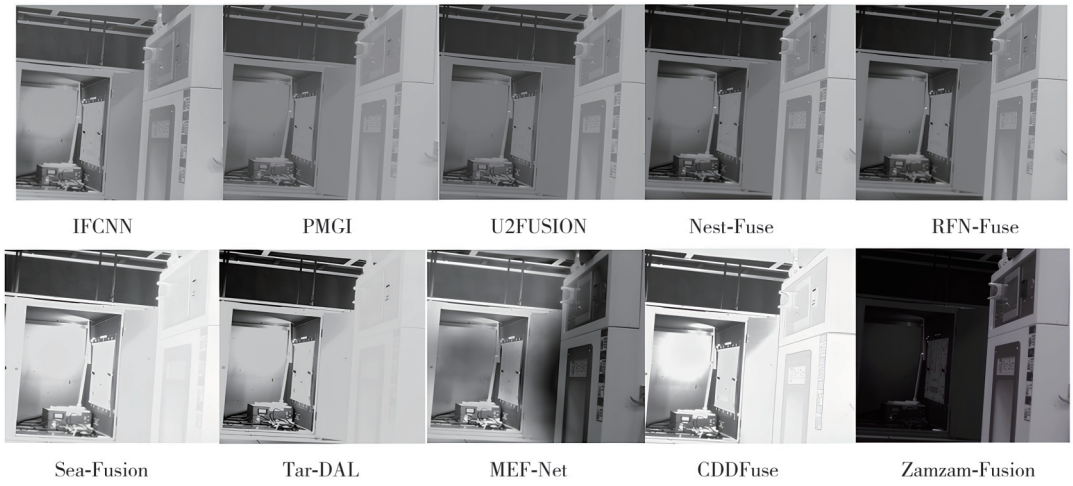

Fig. 7 Comparison of subjective effects of fused images with different algorithms magnified

4 Introducing ATEF-DRPI

To comprehensively assess the quality of low- and high-exposure image fusion for G400BSI sensor data, we propose a new metric, ATEF-DRPI, alongside traditional metrics like SSIM, PSNR, and EN. ATEF-DRPI is specifically designed to evaluate fusion performance in HDR scenarios, providing a score in [0, 1] where higher values indicate better fusion quality. Unlike traditional metrics such as PSNR and SSIM, which may not fully capture human visual perception in HDR contexts, ATEF-DRPI aligns more closely with perceptual quality by considering multiple dimensions of fusion performance.

4.1 ATEF-DRPI metrics

ATEF-DRPI is computed as a weighted sum of five components, each addressing a critical aspect of fusion quality, namely

$$ATEF = clip\left(\sum_{i=1}^{5} w_i S_i, 0, 1\right), \tag{26}$$

where $w_i (i = 1, 2, \cdots, 5)$ are the weights, and the components S_i are defined as follows.

Highlight detail preservation (S_1):

$$S_1 = \min\left(\frac{\sum_{(x,y)\in\Omega_{\text{bright}}} \left|VF(x,y)\right|}{\sum_{(x,y)\in\Omega_{\text{bright}}} \left|VO(x,y)\right| \times e}, 1.0\right). \tag{27}$$

This equation evaluates the preservation of details in bright regions of the fused image F. Here, Ω_{bright} identifies bright regions in the overexposed image O, $VF(x,y)$ is the gradient magnitude of the fused image, $VO(x,y)$ is the gradient magnitude of the overexposed image, and e is a small constant to prevent division by zero. The ratio measures how well the fused image retains details in bright areas relative to the overexposed input, with a cap at 1.0 to normalize the score.

Shadow detail preservation (S_2):

$$S_2 = \min\left(\frac{\sum_{(x,y)\in\Omega_{\text{dark}}} \left|VF(x,y)\right|}{\sum_{(x,y)\in\Omega_{\text{dark}}} \left|VO(x,y)\right| \times e}, 1.0\right). \tag{28}$$

This equation assesses the preservation of details in dark regions of the fused image F. Ω_{dark} identifies dark regions in the underexposed image U, with $VF(x,y)$ and $VO(x,y)$ representing the gradient magnitudes of the fused and underexposed images, respectively. Similar to S_1, the ratio compares detail retention in dark areas, normalized by a maximum value of 1.0, ensuring

the metric focuses on shadow detail fidelity.

Contrast preservation (S_3):

$$S_3 = \min\left(\frac{\sigma_F}{(\sigma_O + \sigma_U)/2}, 1.5\right). \tag{29}$$

This component measures the contrast fidelity of the fused image F. σ_F, σ_O, and σ_U are the standard deviations of the fused, overexposed, and underexposed images, respectively. The equation compares the contrast of the fused image to the average contrast of the input images, with a cap at 1.5 to prevent over-enhancement. This ensures the output of the fused images maintains natural contrast without exaggeration.

Color distortion value (S_4):

$$S_4 = \min\left(\frac{H(F)}{\max\big(H(\boldsymbol{O}), H(\boldsymbol{U})\big) \times e}, 1.2\right). \tag{30}$$

This equation evaluates color distortion in the fused image F. $H(F)$, $H(O)$, and $H(U)$ represent the entropy (or information content) of the fused, overexposed, and underexposed images, respectively. The ratio compares the entropy of the fused image to the maximum entropy of the inputs, scaled by a constant e, with a cap at 1.2. Although the G400BSI dataset involves grayscale images, this component adapts to measure information gain, ensuring the fused image retains meaningful content without excessive distortion.

Structural preservation (S_5):

$$S_5 = 0.6 \times \max\big(SSIM(\boldsymbol{F}, \boldsymbol{O}), SSIM(\boldsymbol{F}, \boldsymbol{U})\big) + 0.4 \times \min\big(SSIM(\boldsymbol{F}, \boldsymbol{O}), SSIM(\boldsymbol{F}, \boldsymbol{U})\big). \tag{31}$$

This component checks the structural quality of the fused image $\boldsymbol{F}$ by measuring how close it is to the overexposed and underexposed images using the SSIM. The fuse image is to keep structure of each input while favor to the aligned better modality, the maximum and minimum SSIM (0.4 and 0.6) are weighted in the equation to balance the ontribution of the two input exposures.

Table 3 Ablation study on ATEF-DRPI weights

Configuration	w_1	w_2	w_3	w_4	w_5	ATEF-DRPI	PSNR_A	SSIM_A
Equal weights	0.20	0.20	0.20	0.200	0.200	0.87	10.50	0.47
Highlight/shadow emphasis	0.30	0.30	0.15	0.125	0.125	0.89	10.65	0.48
Contrast emphasis	0.20	0.20	0.30	0.150	0.150	0.86	10.45	0.46
Proposed weights	0.25	0.25	0.20	0.150	0.150	0.90	10.81	0.49
Alternative weights	0.10	0.40	0.20	0.200	0.100	0.84	10.30	0.45

ATEF-DRPI is a more comprehensive fusion quality evaluation metric by considering the factors that are necessarily important for HDR situations, including detail holding in ultra-high or -low light condition, preservation of contrast, retention of information and structural similarity which are not taken into consideration in traditional evaluation metrics. As it is human-perception aligned, it is a good candidate for assessment of G400BSI sensor data retrieves, as evidenced by the enhancement in ATEF-DRPI scores of the new CDDFuse model.

4.2 Validation of ATEF-DRPI weights

To ensure the rationality of the weights (w_1=0.25, w_2=0.25, w_3=0.20, w_4=0.150, w_5=0.150) in the ATEF-DRPI metrics, we provide theoretical justification and conduct an ablation study to validate their selection.

4.2.1 Theoretical analysis

The weights are designed to reflect the relative importance of each ATEF-DRPI component for HDR fusion of the G400BSI sensor. High detail preservation (C_1) and shadow detail preservation (C_2) are weighted more (0.25) as fine details in both bright and dark areas can only be preserved using the sensor's 95 dB dynamic range and 1.5e^- noise.

Contrast preservation (C_3) is given a moderate weight (0.2) to maintain natural contrast while ensuring perceptual quality. Color distortion (C_4) and structural preservation (C_5) receive slightly lower weights (0.15 each), as they play a supportive role in grayscale HDR fusion, contributing to information retention and structural fidelity.

4.2.2 Ablation study

We evaluated five weight configurations on the G400BSI dataset, comprising 843 dual-exposure pairs (high-gain and low-gain images), with 80% (674 pairs) used for training and 20% (169 pairs) for testing. Each pair of the same scene have different integration time value, i.e., if it is 30, the integration time will be 0.625 08 m·s^{-1}, 1 frame for each image (40=0.830 28, 50=1.035 48, 60=1.240 68, ⋯eg). To enhance the model robustness, data augmentation was applied during training, including horizontal and vertical flips and rotations by 90°, 180°, and 270°, respectively. The configurations tested were: equal weights (w_i=0.2), highlight/shadow emphasis (w_1=0.30, w_2=0.30, w_3=0.15, w_4=0.125, w_5=0.125), contrast emphasis (w_1=0.20, w_2=0.20, w_3=0.30, w_4=0.150, w_5=0.150), proposed weights (w_1=0.25, w_2=0.25, w_3=0.20, w_4=0.150, w_5=0.150), and alternative weights (w_1=0.10, w_2=0.40, w_3=0.20, w_4=0.200, w_5=0.100). Fusion quality was assessed using ATEF-

DRPI, PSNR_A, and SSIM_A to provide a comprehensive evaluation.

The proposed weights achieved the highest scores, with an ATEF-DRPI of 0.90, PSNR_A of 10.81, and SSIM_A of 0.49, demonstrating their effectiveness for HDR fusion.

The setting to retain highlight/shadow detail also fared well (ATEF-DRPI: 0.89), although at the slight expense of contrast balance. The equal weights (ATEF-DRPI: 0.87), and contrast emphasis (ATEF-DRPI: 0.86) configurations performed slightly worse, which may indicate an imbalance in components. The lowest scores were with the alternative weights (ATEF-DRPI: 0.84), suggesting that a balanced weighitng approach is critically important. These results validate optimality of the proposed weights to strike a balance between the ATEF-DRPI constituents leading to quality fusion for G400BSI sensor data.

5 Conclusions

In this work, we proposed Zamzam-Fusion, a unique dual-exposure image fusion framework, which uses the NLM- CMDDS-Fuse algorithm, to improve HDR imaging of grayscale images acquired by the G400BSI CMOS sensor. Zamzam-Fusion is the first to solve the fundamental problems of HDR imaging, which are temporal-non-coherence in dynamic scenes, computational cost, and perceptually-motivated quality assessment by leveraging NLM in the CDDFuse framework. The ATEF-DRPI metric offers a new perception-based comparison method for assessing HDR fusion quality without the weaknesses of classic metrics such as PSNR and SSIM.

Zamzam-Fusion surpasses nine state-of-the-art methods in terms of quantitative (e.g., ATEF-DRPI Metric) and qualitative visual quality, on a self captured dataset of 843 dual-exposed image pairs. Extensive experiments confirm its superiority and effectiveness by showing it can produce high-quality HDR images under adverse lighting situations. This framework provides a ready-to-use and efficient solution suitable for both industrially-viable applications and beyond embedded vision system's development.

Acknowledgement

Alhamdulillah thanks for all the people who supported the project.

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] ZHANG X. Multi-exposure image fusion: a survey of recent advances. IEEE Transactions on Image Processing, 2020: 4567-4582.

[2] ZHANG H, XU H, TIAN X. Deep learning for multi-modal image fusion: a review. Information Fusion, 2021, 12: 323-336.

[3] REINHARD E, WARD G, PATTANAIK S, et al, High dynamic range imaging: acquisition, display, and image-based lighting. San Francisco: Morgan Kaufmann, 2006.

[4] MERTENS T, KAUTZ J, VAN REETH F. Exposure fusion: a simple and practical alternative to high dynamic range photography. Computer Graphics Forum, 2009, 28 (1): 161-171.

[5] GOSHTASBY A A. Fusion of multi-exposure images. Image and Vision Computing, 2005, 23(6): 611-618.

[6] BURT P, ADELSON E. The Laplacian pyramid as a compact image code. IEEE Transactions on Communications, 1983, 31(4): 532-540.

[7] PRABHAKAR R, SRIKAR V S, BABU R V. DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 4724-4732.

[8] ZHAO Z X, BAI H W, ZHANG J S, et al. CDDFuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 17-24, 2023, Vancouver, BC, Canada. New York: IEEE, 2023: 5906-5916.

[9] IM C G, SON D M, KWON H J, et al. Multi-task learning approach using dynamic hyperparameter for multi-exposure fusion. Mathematics, 2023, 11(7): 1620.

[10] CHEN X, ZHANG Y, TANG J. Memory-augmented deep learning for multi-modal image fusion with sequential data. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(6): 8923–8935.

[11] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[12] ZHANG H, HUANG X, XIAO T, et al. Dual-branch feature extraction and fusion network for multi-modal image fusion. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 8: 3924-3936.

[13] GONZALES R C, WOODS R E. Digital image processing. New York: Pearson Education Inc., 2018.

[14] THANGAVEL K, PALANISAMY N, MUTHUSAMY S, et al. A novel method for image captioning using multimodal feature fusion employing mask RNN and LSTM models. Soft Computing, 2023, 27(19): 14205-14218.

[15] ZHANG Y, LIU Y, SUN P, et al. IFCNN: a general image fusion framework based on convolutional neural

network. Information Fusion, 2020, 54: 99-118.

[16] XU H, XU X, XU G, et al. Rethinking the image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity//The 34th AAAI Conference on Artificial Intelligence, February 7-12, 2020, Hilton New York Midtown, New York, USA. New York: IEEE, 2020: 2797-12804.

[17] XU H, MA J Y, JIANG J J, et al. U2Fusion: a unified unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 502-518.

[18] LI H, WU X J, DURRANI T. NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. IEEE Transactions on Instrumentation and Measurement, 2020, 69(12): 9645-9656.

[19] LI H, WU X J, KITTLER J. RFN-Nest: an end-to-end residual fusion network for infrared and visible images. Information Fusion, 2021, 73: 72-86.

[20] TANG L F, YUAN J T, MA J Y. Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network. Information Fusion, 2022, 82: 28-42.

[21] LIU J Y, FAN X, HUANG Z B, et al. Target-aware dual adversarial learning and multi-scenario multi-modality benchmark to fuse infrared and visible for object detection//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-24, 2022, New Orleans, LA, USA. New York: IEEE, 2022: 5792-5801.

[22] MA K D, DUANMU Z F, ZHU H W, et al. Deep guided learning for fast multi-exposure image fusion. IEEE Transactions on Image Processing, 2019, 29: 2808-2819.

用于CMOS传感器的NLM-CDDFuse双增益Zamzam融合及ATEF-DRPI指标评估

IBRAHIM ISMAIL ATEF ISMAIL*, 常玉春

大连理工大学 集成电路学院, 辽宁 大连 116024

摘　要： 提出了一种增强版本的相关性驱动双分支特征分解框架(Correlation-driven dual-branch feature decomposition framework, CDDFuse), 用于融合由G400BSI传感器采集的低曝光与高曝光图像。在CDDFuse架构中引入了一种新颖的神经长时记忆(Neural long-term memory, NLM)模块, 利用跨图像序列持久的全局特征表示提升特征提取能力。该方法能够有效保持动态范围和结构细节, 并采用一种新的指标——ATEF动态范围保持指数(ATEF dynamic range preservation index, ATEF-DRPI)进行评估。在G400BSI数据集上的实验结果表明, 该方法的融合质量优于基线CDDFuse, ATEF-DRPI值可达0.90, 相较于基线的0.80提高了12.5%, 显示出在明暗区域保留细节方面的更好性能。本研究推动了在极端光照条件下的图像融合技术的发展, 为后续视觉任务提供了更优的性能支持。

关键词： 图像融合; G400BSI传感器; 动态范围保持; 低/高曝光融合; 深度学习

引用格式： IBRAHIM ISMAIL ATEF ISMAIL, CHANG Yuchun. Zamzam-Fusion for dual-gain with NLM-CDDFuse for CMOS sensors using ATEF-DRPI metric. Journal of Measurement Science and Instrumentation, 2025, 16(3): 395-405. DOI: 10.62756/jmsi.1674-8042.2025038