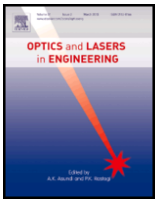
Contents lists available at ScienceDirect

Optics and Lasers in Engineering

journal homepage: www.elsevier.com/locate/optlaseng



# End-to-end deep guided learning for reconstruction of all-in-focus ferrograph image

Xinliang Liu, Jingqiu Wang*, Xiaolei Wang

*National Key Laboratory of Helicopter Aeromechanics, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu 210016, PR China*

ARTICLE INFO

*Keywords:*
Ferrograph image reconstruction
Multi-focus image fusion
Convolutional neural networks
Guided filter

ABSTRACT

Due to the limited depth of field (DoF) of optical microscopes, wear particles of varying thicknesses and sizes cannot be simultaneously presented in sharp focus within a single image, leading to potential misidentification of defocused particles in ferrograph analysis. To address this issue, an end-to-end unsupervised multi-focus ferrograph image fusion model, WearIF, is proposed, which takes a sequence of images as input and outputs an all-in-focus image. First, low-resolution focus weight maps are obtained using a bilinear downsampling operation and a multi-scale dense focus feature extraction network (MDFFEN). These maps are then refined through a convolutional guided filter network to generate high-resolution focus weight maps. Finally, the maps are weighted and summated with the source images to generate an all-in-focus ferrograph image. Moreover, a joint content and gradient based unsupervised loss function is designed to train WearIF, with attention to image structure, texture details, and brightness balance. Experimental results show that WearIF retains more information from the source images and produces fusion results that are more natural and realistic compared to current deep learning-based fusion methods. The proposed model effectively reconstructs the morphology of defocused wear particles in ferrograph images, providing a solid foundation for ferrograph image analysis.

## 1. Introduction

Wear particle analysis is a methodology that uses a high-gradient magnetic field to separate wear particles from lubricating oil, deposits them on a glass substrate, and provides qualitative and quantitative examination and analysis of ferrograph images taken through microscopes [1]. It has been applied to wear monitoring and fault diagnosis in equipment such as aerospace systems, mining equipment, and petrochemical machinery.

Ferrograph images not only provide general characteristics of wear particles, such as type, number, and concentration, but also offer detailed information on individual particles, including size, shape, and morphology, all of which are critical for assessing the wear condition of equipment. However, due to the limited DoF of optical microscopes, particles of varying thicknesses cannot be simultaneously captured and clearly presented in sharp focus within a single image, which may lead to potential omission or misidentification of defocused particles. As shown in Fig. 1, three sequential images are captured during vertical movement of the microscope platform. It can be observed that Particles 1 and 2 are clearest in Images 1 and 2 respectively while the region containing Particle 3 in Image 3 includes a sphere particle, which appears blurred in Images 1 and 2. This blurring could lead to misidentification of particles, as sphere particles may be mistakenly recognized as part of an oxide particle.

Achieving all-in-focus ferrograph images is therefore the primary and crucial step in ferrograph analysis. Several studies have focused on methods for reconstructing the morphology of defocused particles from a single image. For instance, Xi [2] proposed a ferrograph image restoration algorithm that extracted the edges of particles using the Laplace operator, then generated a gain curve to magnify and adjust the edges based on their distance from the center. Wu [3] developed a defocus degradation model for particles using a large convolution kernel CNN model, which minimized the error between the restored and true images, mapping defocused images to focused ones. Although these methods reconstructed ferrograph images by inferring the true morphology of defocused particles from a single image, the three-dimensional characteristics of the particles may be neglected, potentially leading to inaccuracies. The fundamental principle in ferrograph image reconstruction is to restore the defocused particle morphology as accurately and comprehensively as possible.

To address this issue, we developed an automatic ferrograph image acquisition platform, and proposed a microscope autofocus algorithm

[*] Corresponding author.
*E-mail address:* meejqwang@nuaa.edu.cn (J. Wang).

https://doi.org/10.1016/j.optlaseng.2025.109155
Received 25 February 2025; Received in revised form 22 May 2025; Accepted 4 June 2025
0143-8166/© 20XX

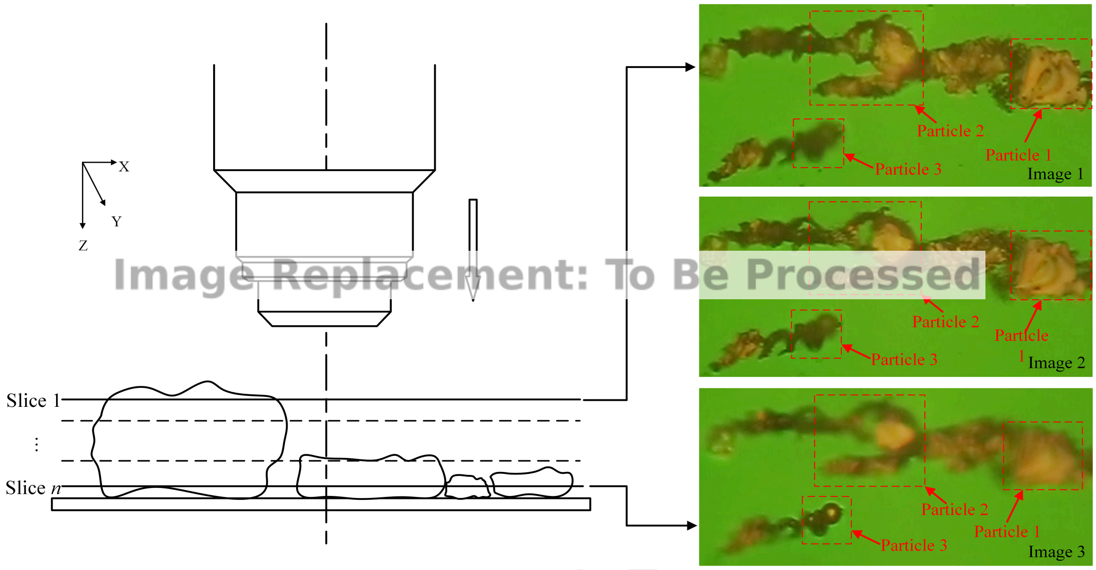


**Fig. 1.** An example of 3 frames of ferrograph images captured by moving the microscope platform in the vertical direction.

[4] in our previous work. A sequence of images is captured while the microscope platform moves at a constant speed along the thickness direction of the wear particles, comprehensively representing the three-dimensional morphology of the particles. Based on the acquired sequence of images, multi-focus image fusion methodology is introduced, which stitches the sharp regions from a sequence of images captured at different DoFs within the same scene to generate a single all-in-focus image. In contrast to single-image restoration methods, this approach directly integrates the focused regions from the source images, effectively restoring the true morphology of defocused particles.

Current multi-focus image fusion methods can be broadly categorized into those based on transform-domains, spatial-domains, and deep learning approaches. Transform-domain methods utilize domain transformation theories, such as discrete wavelet transform [5] and Laplacian pyramids [6], to design clarity evaluation metrics for fusion. Spatial-domain methods assess clarity by analyzing pixel values, image gradients, or image patch features, such as spatial frequency [7], weighted gradients [8], and dense feature transformations [9].

In recent years, deep learning has been applied to multi-focus image fusion owing to its end-to-end processing, automatic feature extraction, and self-learning capabilities [10]. Ma [11] proposed an autoencoder-based model, SESF-Fuse, for unsupervised image fusion, which integrated deep features from the encoder using Active Level Set to generate initial decision maps, and applied consistency verification methods to refine these maps for the final fusion results. Li [12] proposed the DRPL model, which leveraged CNN networks to extract binary focus masks from a pair of images and employed gradient and structure loss functions to enhance texture features of the fused image. Ma [13] proposed GACN, a model designed for the fusion of arbitrary numbers of images, which incorporates CNN, guided filter, and boundary constraints to generate and refine decision maps, along with a decision calibration module to facilitate fusion of multiple images. Zhang [14] proposed MFF-GAN, which employed adaptive and gradient joint constraints in a generative adversarial network framework, of which the generator predicted the fused results and the discriminator evaluated whether they constitute true all-in-focus images. Ma [15] proposed the SwinFusion model, which utilized the Swin Transformer for feature extraction and global interaction, and was trained in an unsupervised manner through joint structure, texture, and intensity loss functions.

Although these multi-focus fusion methods have demonstrated promising results in natural image reconstruction tasks, several challenges remain when applying them directly to ferrograph image fusion. In natural images, defocused regions tend to be well-defined with clear boundaries as shown in Fig. 2; In contrast, ferrograph images involve complex and diverse particle morphologies leading to scattered and isolated defocused regions. Additionally, due to defocus diffusion effects [16], the edges of defocused particles may extend outward, complicating the fusion process as shown in Fig. 3. Furthermore, most existing

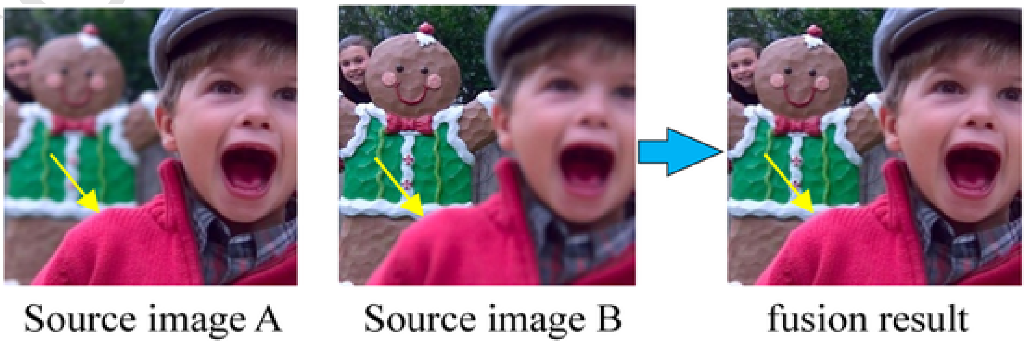

Source image A
Source image B
fusion result

**Fig. 2.** An example of natural images.

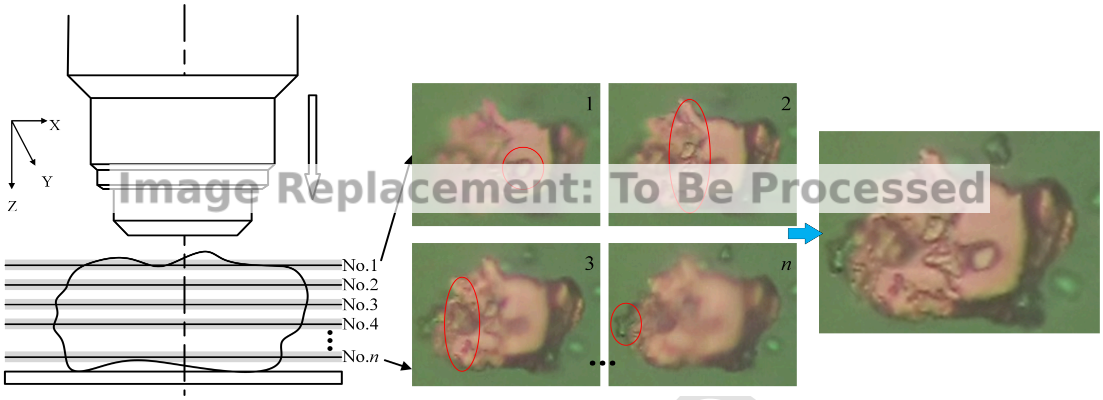


**Fig. 3.** An example of ferrograph sequence images.

models are primarily designed for the fusion of paired image, while the three-dimensional morphology of a wear particle is best captured through a sequence of images acquired by vertically moving the microscope platform. Consequently, these models often rely on pairwise fusion strategies when fusing a sequence of images, which compromises the preservation of the particles' three-dimensional characteristics and reduces fusion efficiency.

To address these challenges, we propose an end-to-end unsupervised multi-focus ferrograph image fusion model, WearIF, which is capable of fusing an arbitrary number of images collected by vertically moving the microscope platform into a single all-in-focus image. The proposed model utilizes a multi-scale dense focus feature extraction network (MDFFEN) and a convolutional guided filter network to extract and refine particle focus weight maps. These maps are then fused through weighted summation to generate the final fused image. Additionally, an unsupervised loss function that combines content and gradient information is designed to train WearIF, optimizing image structure, texture details, and brightness balance.

## 2. End-to-end unsupervised multi-focus ferrograph image fusion network

As shown in Fig. 4, the proposed model consists of a bilinear downsampler, an MDFFEN, a convolutional guided filter network, and a weighted summation fusion module. First, the input sequence of high-resolution images $\boldsymbol{Xh} = \{\boldsymbol{x}_1{}^h, \boldsymbol{x}_2{}^h, \ldots, \boldsymbol{x}_n{}^h\}$ is bilinearly downsampled to obtain a sequence of low-resolution images $\boldsymbol{Xl} = \{\boldsymbol{x}_1{}^l, \boldsymbol{x}_2{}^l, \ldots, \boldsymbol{x}_n{}^l\}$. These low-resolution images are then passed through the MDFFEN to predict the low-resolution focus weight map $\boldsymbol{Al} = \{\boldsymbol{a}_1{}^l, \boldsymbol{a}_2{}^l, \ldots, \boldsymbol{a}_n{}^l\}$. Subsequently, $\boldsymbol{Xh}$, $\boldsymbol{Xl}$, $\boldsymbol{Al}$ are fed into the convolutional guided filter network to obtain the high-resolution weight map $\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$. Finally, the fused result $Y$ is generated through a weighted summation, as calculated below:

$$Y = \sum_{k=1}^{n} \mathbf{a}_k \otimes \mathbf{x}_k^h \tag{1}$$

where ⊗ represents the Hadamard product, and $n$ denotes the total number of image frames in the sequence.

### 2.1. Multi-scale dense focus feature extraction network MDFFEN

The MDFFEN is designed to extract features from the low-resolution input sequence $\boldsymbol{Xl}$ and convert them into corresponding weight maps $\boldsymbol{Al}$. The network takes the sequence images, concatenated along the batch dimension as input, and performs multi-scale semantic feature extraction and fusion through an encoder module and a cascaded co-prime atrous spatial pyramid pooling (CC-ASPP) module [17], thereby capturing global semantic focus features. Subsequently, a series of operations - including a 3×3 convolution, a 1×1 convolution, and an argmax operation - are applied to generate the low-resolution weight map $\boldsymbol{Al}$. The structure of this network is depicted in Fig. 5.

**The encoder module** is composed of three stages with a total of six encoding layers, as illustrated in Fig. 5(a). Stages 1 and 3 consist of convolutional layers (7×7, 3×3, where $\boldsymbol{r}$ represents dilation rate), batch normalization (BN) layers, and ReLU activation functions, with the output channel $c$ set to 24. These stages are responsible for extracting shallow image features and integrating deeper semantic information. Stage 2 comprises four ConvBlock layers (as shown in Fig. 5(b)), which utilize atrous convolutions to expand the receptive field without increasing the number of parameters or affecting spatial resolution. Additionally, dense connection [18] is incorporated in this module, enabling the concatenation of outputs from previous layers to serve as input for subsequent layers. This design promotes the reuse of shallow focus features and mitigates the vanishing gradient problem. The specific parameters of the encoder module are summarized in Table 1.

**The cascaded co-prime atrous spatial pyramid pooling (CC-ASPP) module** further performs multi-scale feature fusion on the output of the encoder module. Compared to the traditional ASPP [19], CC-ASPP introduces two key improvements: (1) The use of co-prime atrous rates to increase the participation of pixels in feature computation, reducing the Grid Effect; (2) The adoption of a cascaded structure to enlarge receptive field gaps between different branches, enabling better extraction and fusion of features across multiple scales. Table 2 provides a comparison of the receptive fields between ASPP and CC-ASPP. As observed, in the ASPP module, the receptive fields of the first four branches exhibit only minimal variation, which limits its ability to effectively extract features from larger particles. In contrast, the CC-ASPP module, through its cascaded structure, creates a substantial gap in receptive field sizes between branches, facilitating the extraction of both global and local features.

### 2.2. Convolutional guided filter network

We utilize the guided filter [20] to address the edge artifacts induced by defocusing, as it effectively reconstructs the structure and edge information from the input image using a guidance image, while simultaneously performing smoothing and denoising. However, two critical challenges emerge when directly applying the guided filter to the ferrograph image fusion task: First, using source images as guidance images may introduce redundant information, compromising the distinctiveness of particle edges and textures. Second, reliance on manually designed kernels limits the generalization capacity of the algo-

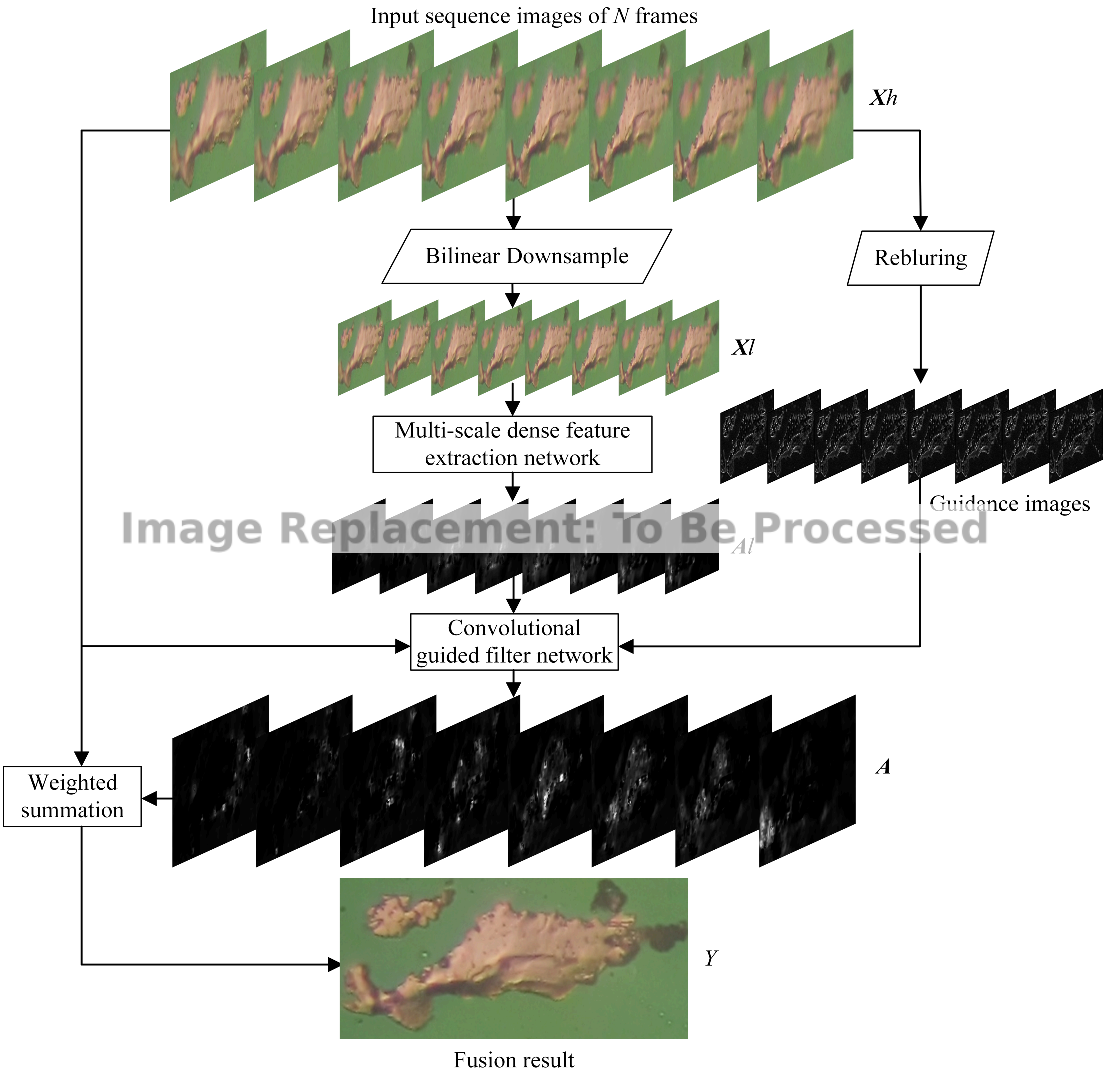

Input sequence images of N frames
Xh
Bilinear Downsample
Rebluring
Xl
Multi-scale dense feature extraction network
Guidance images
Al
Convolutional guided filter network
A
Weighted summation
Y
Fusion result

**Fig. 4.** Structure diagram of WearIF.

rithm. To overcome these limitations, we propose a convolutional guided filter network, incorporating the following key improvements:

(1) A guidance image is generated through a Reblur operation, as illustrated in Fig. 6. The source image is subjected to two iterations of Gaussian filtering to produce a blurred version, after which the absolute difference between the blurred image and source image is computed to create a difference map. Within this map, sharp regions exhibit prominent activation (highlighted by the red ellipse), while blurred regions demonstrate negligible activation (indicated by the blue ellipse). These maps are subsequently aggregated across all source images to form the guidance image for the convolutional guided filter network.

(2) Convolutional layers substitute manually designed kernels [21] with the guided filter network being integrated into the MDFFEN. Our WearIF model is optimized end-to-end at full resolution to learn nonlinear guided filter kernels, thereby enhancing the generalization ability of the model.

The guided filter process is detailed in Table 3, where $f_{bf}(\cdot)$ denotes a $r \times r$ convolutional layer coupled with a ReLU nonlinear activation, and $f_{conv}(\cdot)$ comprises two $1 \times 1$ convolutional layers interleaved withBN layers and ReLU activations. The input, intermediate, and output channels are configured as 2, 8, and 1, respectively. The $[\cdot]$ operation signifies channel-wise concatenation, and upsampling$(\cdot)$ represents to the upsampling operation.

### 2.3. *Unsupervised Loss Function for Joint Content and Gradient*

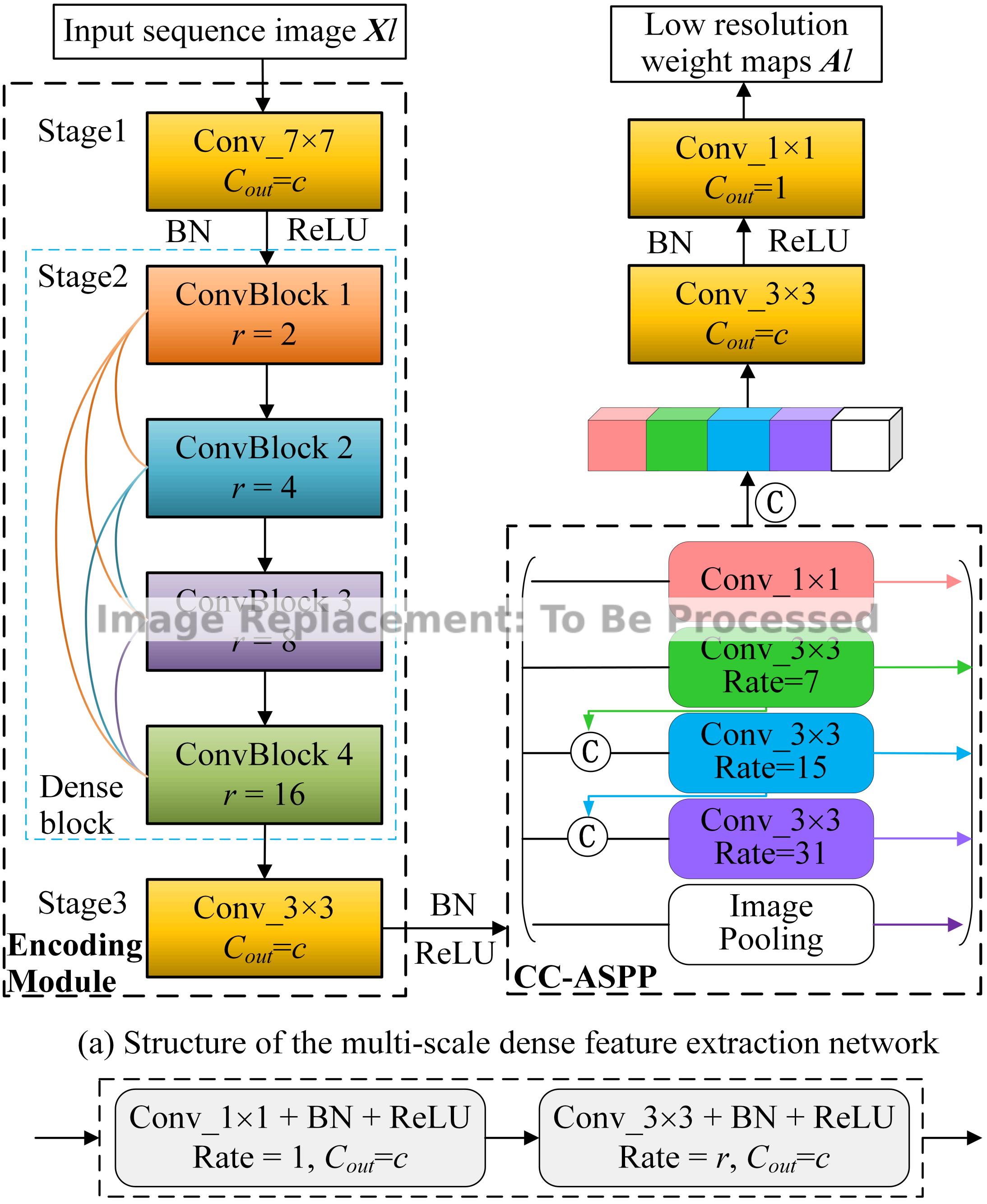
The proposed unsupervised joint content-and-gradient loss function is formulated with three terms: the content loss $L_{con}$, the gradient loss $L_{grad}$, and the constraint term $L_{mse}$. These terms collectively ensure the preservation of the overall image structure, the enhancement of texture details of particle, and the maintenance of the image's global brightness distribution, respectively. The loss function is defined as follows:

$$L = L_{con} + \lambda_1 \cdot L_{grad} + \lambda_2 \cdot L_{mse} \tag{2}$$

where $\lambda_1$ and $\lambda_2$ are set to 0.5 and 0.3, respectively after a series of contrast experiments.

#### 2.3.1. *Content Loss term $L_{con}$*

The primary goal of multi-focus image fusion is to preserve the optimally focused content from the sequence of images at each spatial loca-



(a) Structure of the multi-scale dense feature extraction network

(b) Structure of Conv Block

**Fig. 5.** Structure diagram of the MDFFEN.

**Table 1**
Encoding module parameter table.

| Encoding layer index | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Kernel size | 7×7 | 1×1<br>3×3 | 1×1<br>3×3 | 1×1<br>3×3 | 1×1<br>3×3 | 3×3 |
| Atrous rate | 1 | 2 | 4 | 8 | 16 | 1 |
| Number of output channels | 24 | 24 | 24 | 24 | 24 | 24 |
| Dense connection | × | √ | √ | √ | √ | × |
| Size of receptive field | 7×7 | 11×11 | 19×19 | 35×35 | 67×67 | 69×69 |

**Table 2**
Receptive fields of ASPP modules with different structures.

| | Conv_1×1 | Conv_3×3<br>r=7 | Conv_3×3<br>r=15 | Conv_3×3<br>r=31 | Image pooling |
|---|---|---|---|---|---|
| ASPP | 69 | 83 | 99 | 121 | 256 |
| CC-ASPP | 69 | 83 | 113 | 175 | 256 |

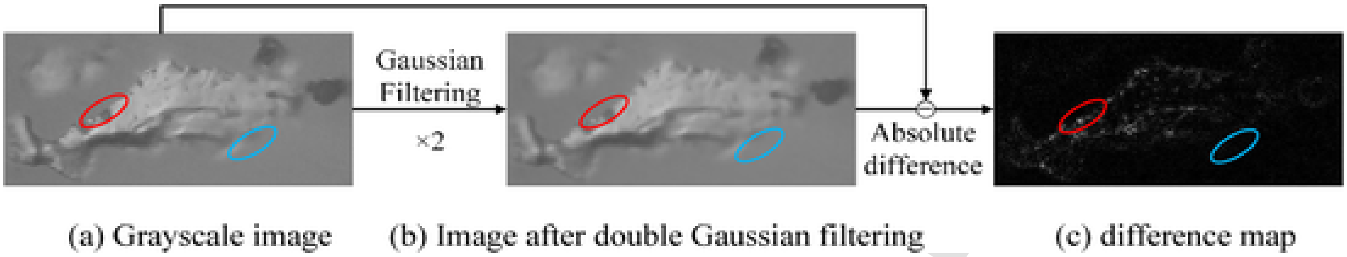


**Fig. 6.** The process of Reblur.

**Table 3**
Steps of Guided Filter.

| The process of Guided Filter |
|---|
| **Input:** weight map at low resolution $\boldsymbol{Al}_k$; |
| **Guide Filter:** difference map C generated by repeated blurring operations, and input sequence images at high resolution $\boldsymbol{X}_k$; |
| **Output**: weight map at high resolution $\boldsymbol{A}_k$; |
| 1: $mean_g = f_{bf}(\mathrm{C})$; $mean_i = f_{bf}(\boldsymbol{Al}_k)$; $corr_g = f_{bf}(\mathrm{C}\otimes\mathrm{C})$; $corr_{gi} = f_{\mathrm{bf}}(\mathrm{C}\otimes\boldsymbol{Al}_k)$ |
| 2: $var_g = corr_g - mean_g\otimes mean_g$; $cov_{gi} = corr_{gi} - mean_g\otimes mean_i$ |
| 3: $\boldsymbol{Al}_k = f_{\mathrm{conv}}([cov_{gi}, var_g])$; $\boldsymbol{Bl}_k = mean_i - \boldsymbol{Al}_k\otimes mean_g$ |
| 4: $\boldsymbol{A}_k = \mathrm{upsampling}(\boldsymbol{Al}_k)$; $\boldsymbol{B}_k = \mathrm{upsampling}(\boldsymbol{Bl}_k)$ |
| 5: $\boldsymbol{W}_k = \boldsymbol{A}_k\otimes\boldsymbol{X}_k + \boldsymbol{B}_k$ |

tion in the fused image. To achieve this, the content loss function is designed based on the Structural Similarity Index (SSIM) [22] and its enhanced variants [23; 24], as follows:

Let $\{P_k(i, j) \mid 1\le k \le K\}$ denote the set of image patches extracted from the spatial position $(i, j)$ across multi-focus sequence images $\boldsymbol{Xh}$. For each patch $\boldsymbol{P}_k$, it is projected into an $N^2$-dimensional space to obtain a column vector $\boldsymbol{P}_k$, where $N$ is the height/width of the patch. Following the standard SSIM formulation, $\boldsymbol{P}_k$ is decomposed into three independent components: contrast component, texture structure component, and luminance component, expressed as:

$$\begin{aligned}\mathbf{P}_k &= \left\|\mathbf{P}_k - \mu_{p_k}\right\| \times \frac{\mathbf{P}_k - \mu_{p_k}}{\left\|\mathbf{P}_k - \mu_{p_k}\right\|} + \mu_{p_k} \\ &= \left\|\widehat{\mathbf{P}_k}\right\| \times \frac{\widehat{\mathbf{P}_k}}{\left\|\widehat{\mathbf{P}_k}\right\|} + \mu_{p_k} \\ &= c_k \times \mathbf{s}_k + l_k\end{aligned} \tag{3}$$

Where $||\cdot||$ denotes the $l_2$-norm, $\mu_{p_k}$ is the mean luminance of the patch, and $\widehat{\mathbf{P}_\mathbf{k}} = \mathbf{P}_k - \mu_{p_k}$ represents the patch vector after mean subtraction. Thus, the contrast value $c_k$, texture structure vector $\boldsymbol{s}_k$, and luminance value $l_k$ are defined as the scalar $\left\|\widehat{\mathbf{P}_k}\right\|$, the unit vector $\widehat{\mathbf{P}_k}\Big/\left\|\widehat{\mathbf{P}_k}\right\|$, and the scalar $\mu_{p_k}$, respectively,as derived above.

Let the target all-in-focus image be $\hat{\mathbf{Y}}$, with its corresponding patch centered at $(i,j)$ denoted as $\hat{\mathbf{p}}$. Its three components $\hat{c}$, $\hat{\mathbf{s}}$, and $\hat{l}$ are computed from the component sets $\{c_1, c_2, \ldots, c_K\}$,$\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_K\}$, and $\{l_1, l_2, \ldots, l_K\}$ of $\{\boldsymbol{P}_k \mid 1 \le k \le K\}$, as:

$$\widehat{\mathbf{P}} = f_c\left(c_1, c_2, \ldots, c_K\right) \times f_{\mathbf{s}}\left(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K\right) + f_l\left(l_1, l_2, \ldots, l_K\right) \tag{4}$$

Here, $f_c$, $f_s$, and $f_l$ represent the mapping functions for contrast, texture structure, and luminance components, respectively, from the multi-focus patch sequence to the target all-in-focus patch. The following criteria govern these mappings:

1) Contrast, which quantifies the intensity difference between the brightest and darkest regions in an image, directly affects sharpness, detail representation, and visual appeal. Higher contrast improves detail visibility, enhancing subject sharpness. Thus, the output contrast is selected as the maximum contrast among all source patches:

$$\hat{c} = f_c\left(c_1, c_2, \ldots, c_K\right) = \max_{1\le k\le K} c_k = \max_{1\le k\le K}\left\|\widehat{\mathbf{P}_k}\right\| \tag{5}$$

2) Texture structure vector reflects the orientation of the patch in the $N^2$-dimensional vector space, where distinct textures correspond to divergent orientations. Therefore, the output texture vector aligns with the statistically dominant direction of all source patches, computed via:

$$\hat{\mathbf{s}} = \frac{\bar{\mathbf{s}}}{\|\bar{\mathbf{s}}\|}, \quad \bar{\mathbf{s}} = \frac{\sum_{k=1}^{K} w_s\left(\widehat{\mathbf{P}_k}\right)\mathbf{s}_k}{\sum_{k=1}^{K} w_s\left(\widehat{\mathbf{P}_k}\right)} \tag{6}$$

Where $w_s$ $(\cdot)$ denotes the $l_\infty$-norm.

3) Luminance governs image brightness, critically influencing visual perception. Excessive brightness causes overexposure (highlights detail loss), while insufficient brightness induces shadow blurring. Hence, the output luminance is a weighted average of all source patches, defined as:

$$\hat{l} = \frac{\sum_{k=1}^{K} w_l\left(\mu_k, l_k\right) l_k}{\sum_{k=1}^{K} w_l\left(\mu_k, l_k\right)} \tag{7}$$

Here, $\mu_k$ is the global mean luminance of $\boldsymbol{x}_k^h$, $l_k$ is the local mean luminance of $P_k$, and $w_l$ $(\cdot)$ is a weight function combining global and local luminance via a 2D Gaussian:

$$w_l\left(\mu_k, l_k\right) = \exp\left(-\frac{\left(\mu_k - \tau\right)^2}{2\sigma_g^2} - \frac{\left(l_k - \tau\right)^2}{2\sigma_l^2}\right) \tag{8}$$

Where $\sigma_g$ and $\sigma_l$ are photometric spread coefficients, set to 0.2 and 0.5, respectively; and $\tau=128$ for 8-bit images. The Content Loss term $L_{con}$ is formulated as:

$$L_{con} = 1 - SSIM_{MFF}\left(\left\{\mathbf{P}_k\right\}, \mathbf{Y}\right) \tag{9}$$

$$SSIM_{MFF}\left(\left\{\mathbf{P}_k\right\}, \mathbf{Y}\right) = \frac{1}{M}\sum_{i=1}^{M} S\left(\left\{\mathbf{P}_k(i)\right\}, Y(i)\right) \tag{10}$$

$$S\left(\left\{\mathbf{p}_k\right\}, y\right) = \frac{\left(2\mu_{\hat{\mathbf{p}}}\mu_{\mathbf{y}} + C_1\right)\left(2\sigma_{\hat{\mathbf{p}}\mathbf{y}} + C_2\right)}{\left(\mu_{\hat{\mathbf{p}}}^2 + \mu_{\mathbf{y}}^2 + C_1\right)\left(\sigma_{\hat{\mathbf{p}}}^2 + \sigma_{\mathbf{y}}^2 + C_2\right)} \tag{11}$$

where $\mu_{\hat{\mathbf{p}}}$ and $\mu_{\mathbf{y}}$ represent the mean brightness values of the desired all-in-focus image patches $\{\boldsymbol{p}_k\}$ and the fused image patches $\boldsymbol{y}$, respectively. $\sigma_{\hat{\mathbf{p}}}$, $\sigma_{\mathbf{y}}$ and $\sigma_{\hat{\mathbf{p}}\mathbf{y}}$ represent the variances and covariances of $\hat{\mathbf{p}}$ and $\boldsymbol{y}$. Here, $\mathrm{C}_1$ and $\mathrm{C}_2$ are two small stabilizing constants. The $SSIM_{MFF}$ ranges within [0, 1], with higher values corresponding to superior fusion quality.

### 2.3.2. Gradient Loss term $L_{grad}$

To amplify texture details of wear particles, the gradient loss term $L_{grad}$ is formulated using Laplacian edge detection. For each image in the sequence $\boldsymbol{Xh}$, the Laplacian gradient operator is applied to generate a gradient image sequence $\{\Lambda k\}$. The maximum gradient at each spatial location is preserved to form the final gradient image $\boldsymbol{\Lambda}$, defined as:

$$\{\boldsymbol{\Lambda}_k\} = \nabla\left(\{\mathbf{X}_h^k\}\right) \tag{12}$$

$$\boldsymbol{\Lambda} = \max_{1\le i\le w, 1\le j\le h, 1\le k\le K}\left(\{\boldsymbol{\Lambda}_k(i,j)\}\right) \tag{13}$$

Where, $\nabla(\cdot)$ represents Laplacian operator. Thus, the gradient loss $L_{grad}$ is defined as:

$$L_{grad} = \frac{1}{w\times h}\cdot\sum_{i,j}(\mathbf{Y}(i,j) - \boldsymbol{\Lambda}(i,j))^2 \tag{14}$$

### 2.3.3. Constraint Term $L_{mse}$

Solely optimizing content and gradient losses may cause WearIF overemphasize regions with high clarity and rich detail, disregarding global pixel distribution characteristics. This can lead to artifacts such as amplified noise and unbalanced background brightness in the fused image. To enhance model robustness and alleviate overfitting, we introduce a mean-image constraint, derived by averaging the pixel values across the input sequence images. This constraint ensures WearIF selectively reconstruct relevant information while maintaining attention to the global brightness distribution of the image. The constraint term $L_{mse}$ is defined as:

$$L_{mse} = \frac{1}{w\times h}\cdot\sum\left(\mathbf{Y} - \overline{\mathbf{Y}}\right) \tag{15}$$

$$\overline{\mathbf{Y}} = \frac{1}{K}\sum_k \mathbf{X}_h^k \tag{16}$$

## 3. Model Training

### 3.1. Weighted Fusion Strategy

The YCbCr color space is adopted for ferrograph image fusion task, as it better aligns with the perceptual characteristics of the human visual system. Within this space, the luminance component is fused using our WearIF model, while the chrominance components are fused via a weighted averaging strategy defined as:

$$Cx = \frac{\sum_{k=1}^{K} w_x\cdot\left(Cx_k\right)\cdot Cx_k}{\sum_{k=1}^{K} w_x\cdot\left(Cx_k\right)},\quad x = b, r \tag{17}$$

$$w_x\left(Cx_k\right) = \left\|Cx_k - \tau\right\|_1 \tag{18}$$

where $Cx_k$ denotes the $C_b$ or $C_r$ component of the $k^{th}$ image in the sequence $\boldsymbol{Xh}$; $||\cdot||_1$ represents the $L_1$ norm. Finally, the fused luminance and chrominance components are converted from the YCbCr back to the RGB color space.

### 3.2. Model Training

As shown in Fig. 7, the experimental platform comprises a microscope (XJZ-6), an XYZ motorized platform, a digital camera and a computer. The XYZ motorized platform is connected to the computer, enabling software-controlled adjustment of displacement speed and positioning accuracy. A total of 579 sequences of multi-focus images with varying wear particle distributions were acquired, forming a ferrograph image fusion dataset partitioned into 300 training sequences and 279 testing sequences. For each sequence, only 32 frames with the highest Laplacian gradient values were retained. Input images were resized to 910 × 512 pixels (high-resolution) and 228 × 128 pixels (low-resolution) to mitigate GPU memory constraints during training while maintaining computational efficiency.

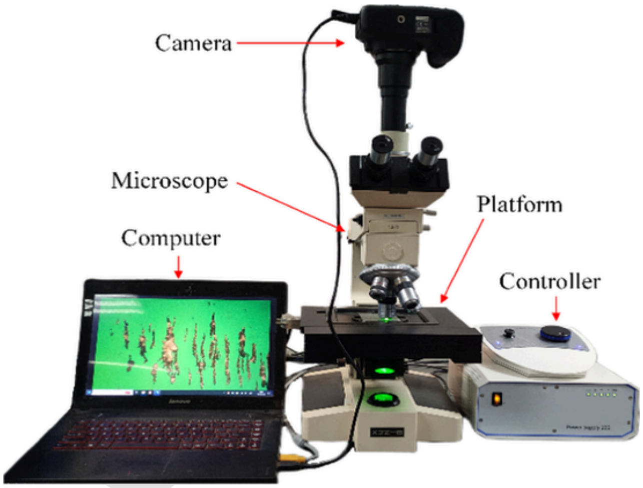


**Fig. 7.** Image acquisition platform.

The training hyperparameters for WearIF are configured as follows: a batch size of 32 for each iteration, an initial learning rate of 0.0001 using the Adam optimizer, and 30000 total iterations. The network was trained and validated on a workstation equipped with i7-8700 K CPU, 24 G RAM, and a NVIDIA RTX 2070 GPU. The total training duration was approximately 8 hours, with the loss curve versus iterations illustrated in Fig. 8. It can be observed that the loss decreases rapidly within the first 10000 iterations, demonstrating the model's rapid acquisition of global clarity features. Between 10000 and 20000 iterations, the loss declines gradually, indicating fine-grained refinement of texture detail representation. Beyond 20000 iterations, the loss plateaus, demonstrating model convergence.

## 4. Experimental Results

### 4.1. Evaluation Metrics

To accurately evaluate the performance of the proposed model, both subjective and objective evaluation metrics are adopted. Subjective evaluation relies on human visual perception to assess fusion quality with a focus on criteria including details preservation, sharpness, contrast fidelity, and perceptual naturalness. Objective metrics provide a quantitative assessment of the fusion results. Five objective metrics are adapted and extended for multi-focus fusion tasks.

Let the input sequence images be $\mathbf{S} = \{S1, S2, \ldots\}$. The extended formulations for multi-image fusion are defined below, and the origin equations are shown in the relevant references:

(1)The image feature-based evaluation metric $Q_g$ [25]

$$Q_g = \frac{\sum_{S_i\in S}\sum_{x=1}^{N}\sum_{y=1}^{M} Q^{S_iF}(x,y)\left[g_{S_i}(x,y)\right]^L}{\sum_{S_i\in S}\sum_{x=1}^{N}\sum_{y=1}^{M}\left[g_{S_i}(x,y)\right]^L} \tag{19}$$

(2)The structure similarity-based evaluation metric $Q_y$ [26]

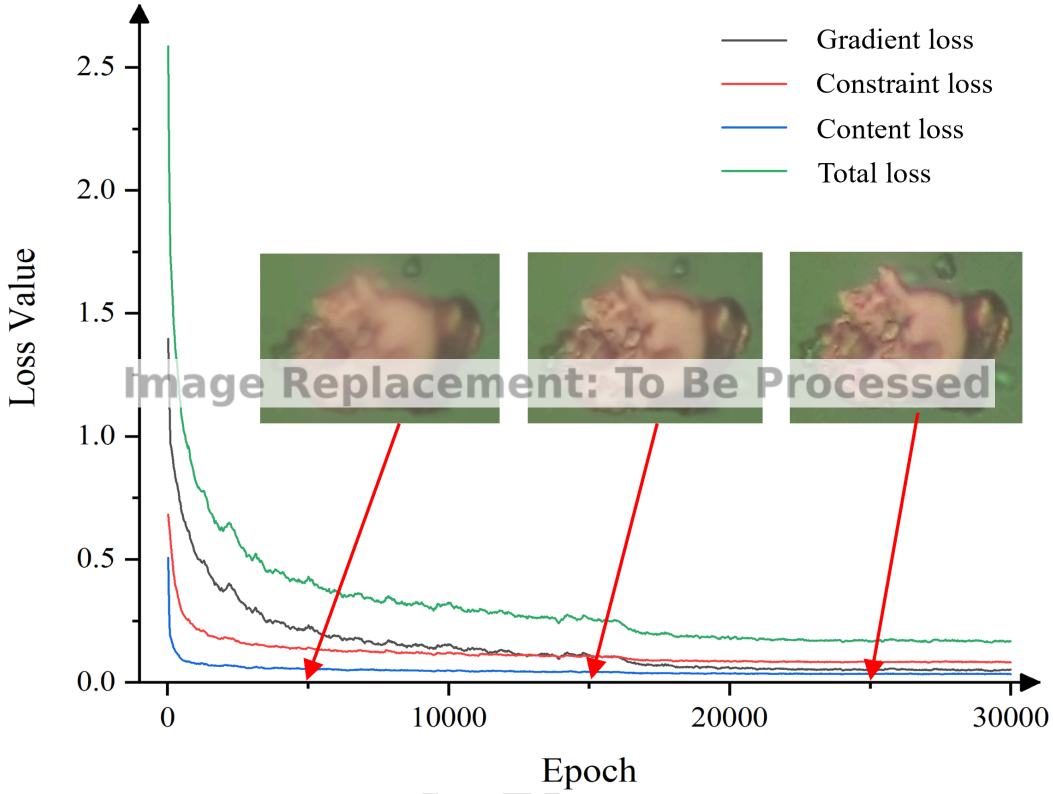


**Fig. 8.** Loss curve during the Training Process of WearIF.

$$Q_y = \frac{1}{|W|} \sum\nolimits_{w \in W} Q_y^w \tag{20}$$

$$Q_y = \begin{cases} \sum\limits_{i=1}^{n} \lambda_i(w)\, SSIM\left(S_i, F|w\right) & , SSIM_{mean}(w) \geq 0.75 \\ \max\left\{ SSIM\left(S_i, F|w\right) \middle| i = 1, 2, \ldots, n \right\}, SSIM_{mean}(w) < 0.75 \end{cases} \tag{21}$$

(3)The human visual perception-based evaluation metric $Q_{cb}$ [27]

$$Q_C(a,b) = \frac{\sum\limits_{S_i \in S} C_{S_i}{}'(a,b) \cdot Q_{S_i F}(a,b)}{\sum\limits_{S_i \in S} C_{S_i}{}'^2(a,b)} \tag{22}$$

(4)The information-theory-based evaluation metric FMI [28]

$$FMI_F^S = \sum_{S_i \in S} \frac{I_{FS_i}}{H_F + H_{S_i}} \tag{23}$$

(5)The correlation-based metric SCD [29]

$$SCD = \frac{1}{n} \sum_{i=1}^{n} r\left(D_i, S_i\right) \tag{24}$$

### 4.2. Evaluation Results

Fig. 9 depicts the fusion process for a sequence of ferrograph images. Panel (a) presents 5 representative frames from a sequence, and Panel (b) shows the corresponding weighted focus maps. In these maps, regions with higher intensity values (brighter pixels) indicate areas of superior focus in the respective frames, which dominate the fused result. For instance, from the 7th to the 17th frame, the focal quality of the Chain particles improves progressively. In the red bounding box region, as highlighted by the red ellipses, the right, middle, and left portions of a particle achieve optimal clarity in the 7th, 12th, and 17th frames, respectively. These regions exhibit maximum activation in the focus maps, confirming their dominant contribution to the particle's representation in the fused image.

From the 22nd to the 27th frame, the sphere particle (left) gains sharpness, peaking in the 26th frame, as shown in the red box areas. The weighted focus map reveals a centripetal activation pattern, where the region of interest is gradually highlighted from the periphery toward the center,culminating in peak activation at the 26th frame. Furthermore, the elevated activation at the lower-left edge of the particle in the 27th frame suggests this subregion predominantly defines the sphere particle's morphology in the final fusion.

The fused result is shown in Panel (c), which demonstrates visually coherent fusion with wear particles of diverse thicknesses retained simultaneously in a single image. Critically, the texture fidelity, morphology integrity, and diagnostic features of the particles are preserved, fulfilling the requirements for wear particle analysis.

Fig. 10 presents the testing results of WearIF applied to four ferrograph sequence images, each exhibiting distinct particle distributions, brightness levels, and contrast. From a subjective evaluation perspective, the fusion results achieve overall clarity, with contrast and brightness aligned with human visual perception. Moreover, texture details of wear particles are well-retained without noticeable distortion or artifacts.

A key strength of WearIF lies in its flexibility to adjust the number of input frames for more comprehensive fusion. Taking Fig. 11(a) as an example, when processing 32 input frames, the central region of the particle (arrow-marked) remains defocused across all frames, resulting in blurring in the fused output. In contrast, when the input is increased to

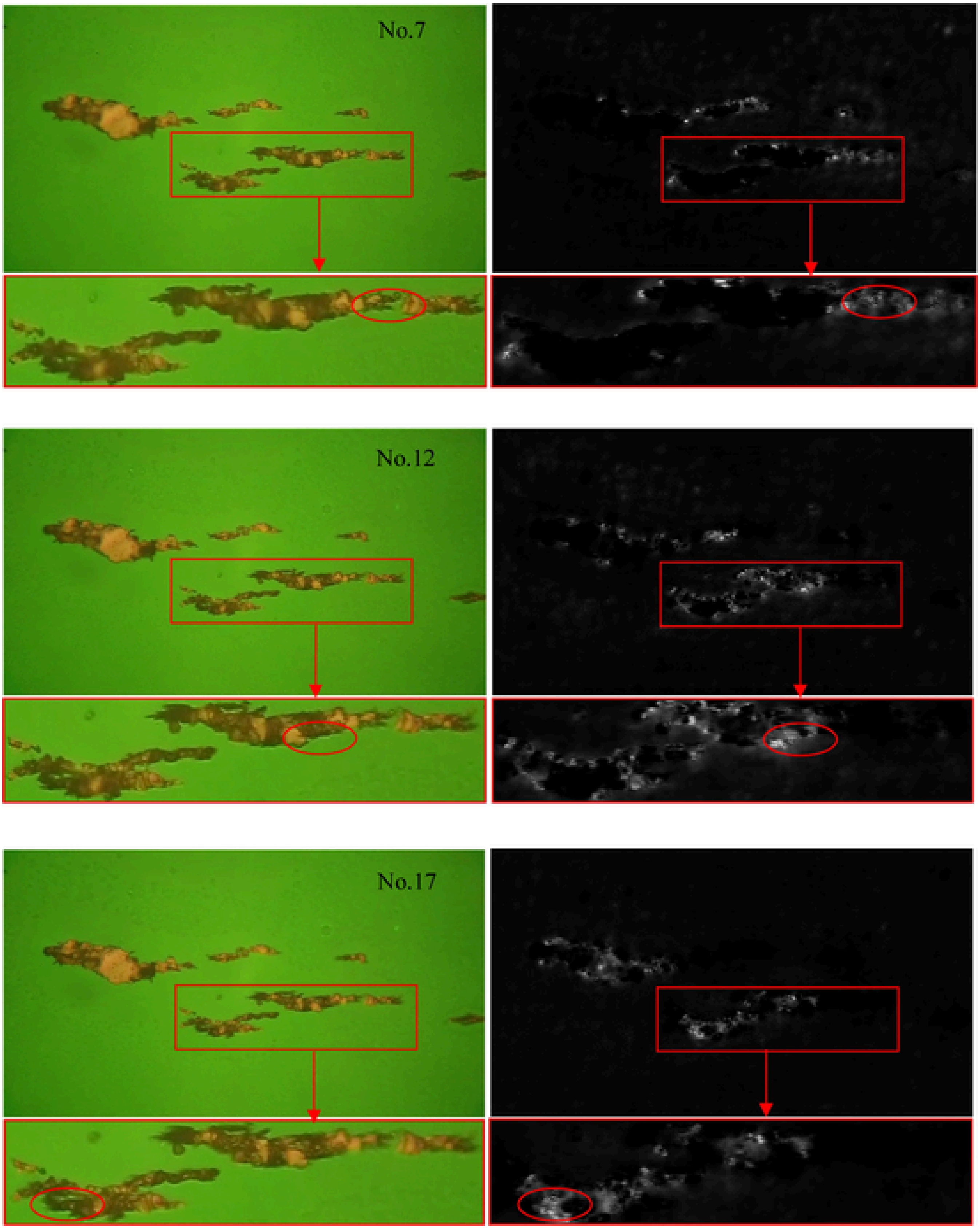
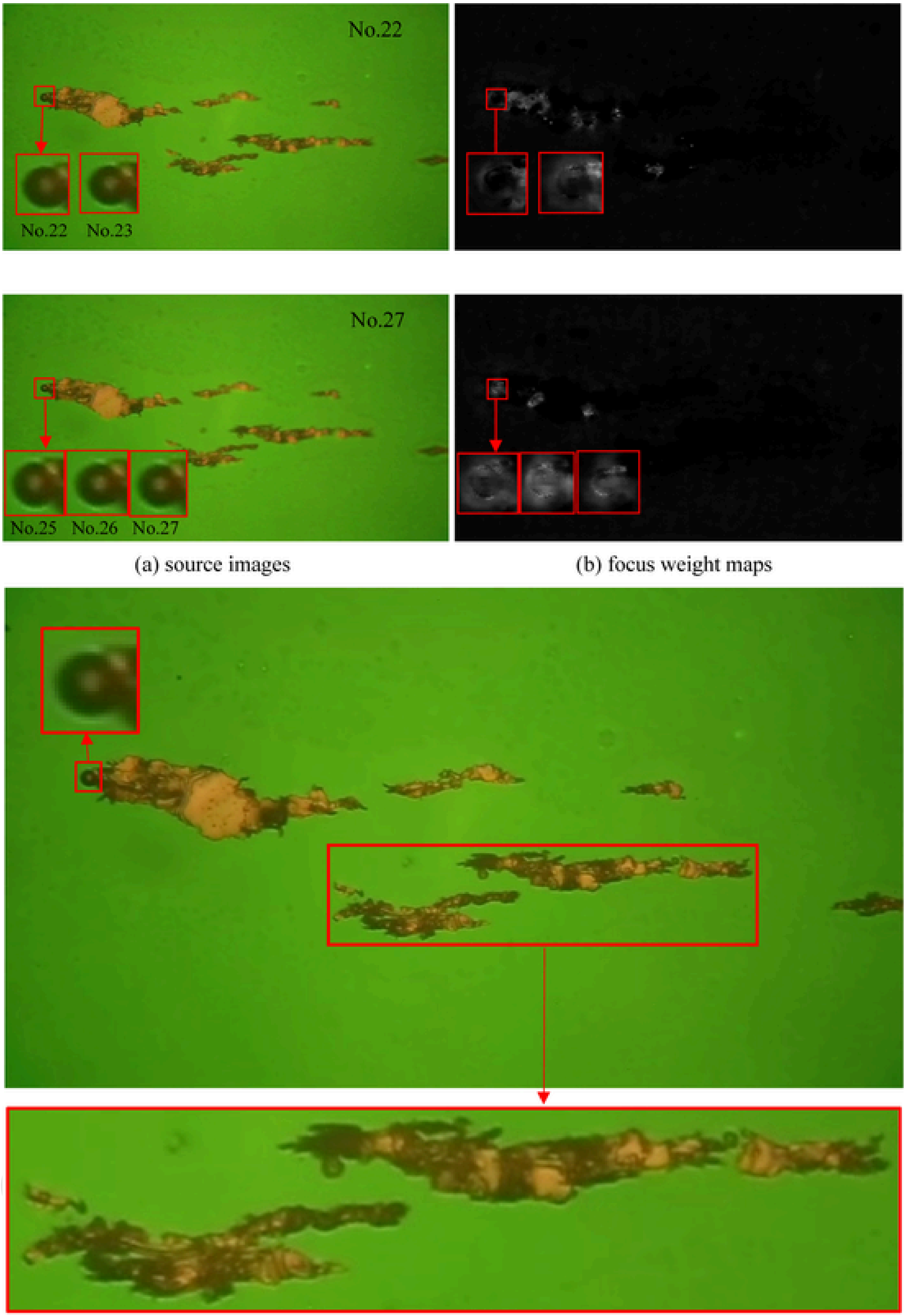


**Fig. 9.** the process of image fusion.



(a) source images

(b) focus weight maps

(c) fusion result

**Fig. 9.** (*continued*)

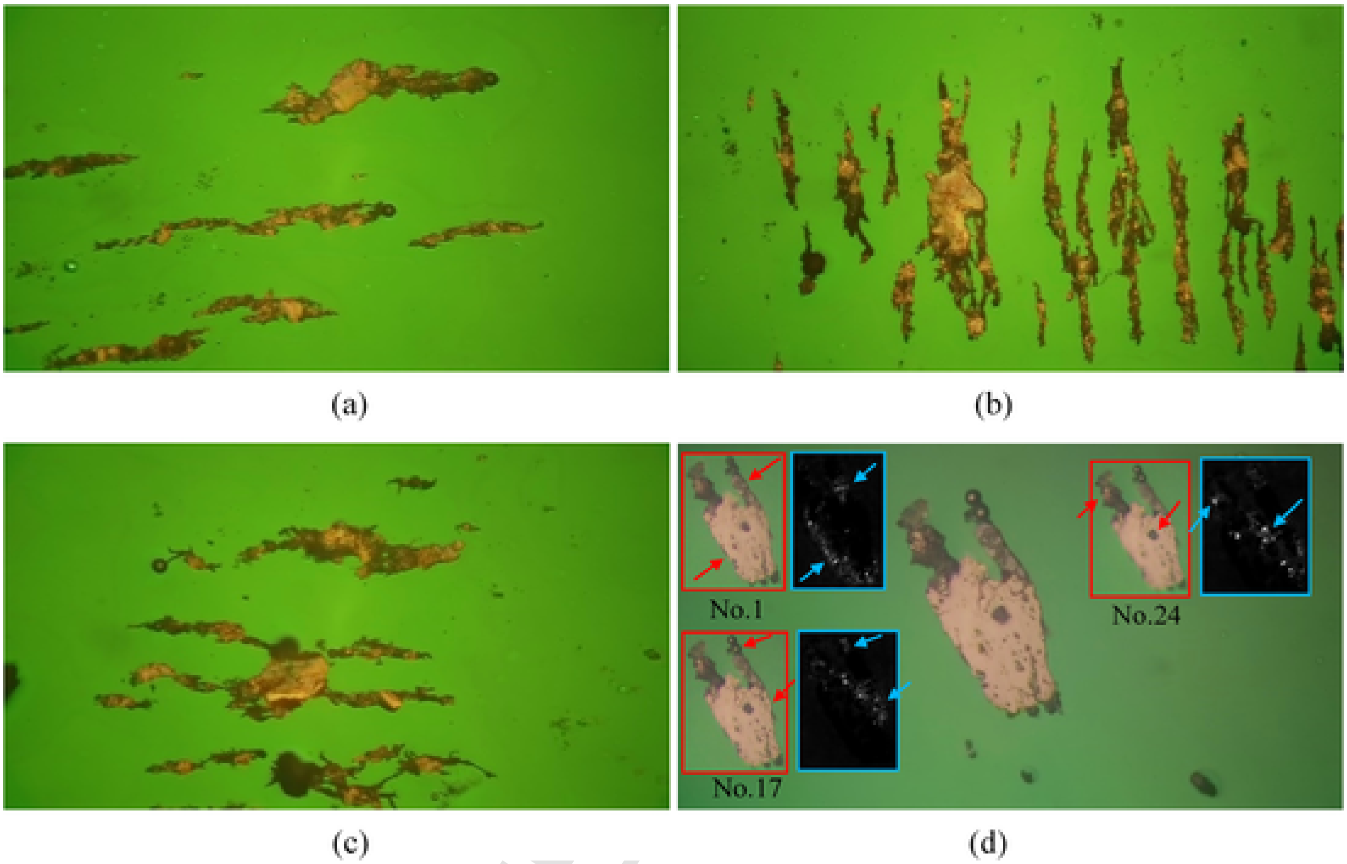


**Fig. 10.** fusion result of WearIF on four sequences of ferrograph images.

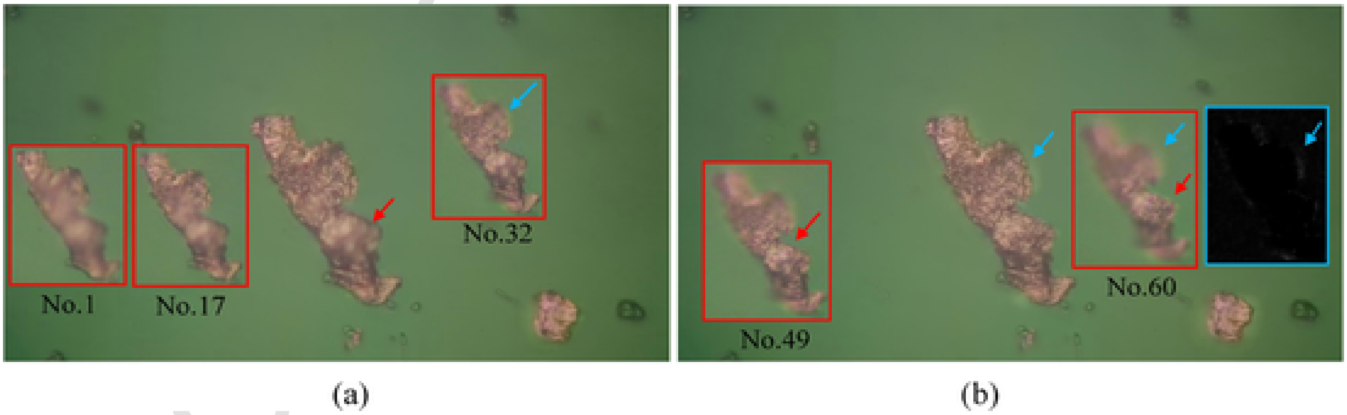


**Fig. 11.** Fusion results of WearIF with a flexible number of input ferrograph images.

60 frames (Fig. 11(b)), the entire Block particle is rendered with enhanced clarity, preserving texture, morphology, and critical features. This validates WearIF's capability to extract focused content from variable-length input sequences, highlighting robust generalization. However, comparison between Panels (a) and (b) shows that increasing input frames may induce subtle artifacts, manifested as annular pseudo-boundaries around particles due to defocus diffusion. Furthermore, accumulation of defocused gradients in guidance images compromises the guided filtering process, leading to performance degradation

The WearIF model was quantitatively evaluated using objective evaluation metrics, achieving scores of 0.4258 (FMI), 0.5319 ($Q_g$), 0.6707 ($Q_y$), 0.5953 ($Q_{cb}$),and 0.2573 (SCD). Additionally, it attained a processing speed of 0.53 s/frame, demonstrating efficient fusion performance.

## 5. Ablation study analysis

### *5.1. Comparison of different guided filter network structures*

To investigate the impact of different guided filter methods on the fusion performance, three widely-used guided filter variants are compared. Method 1 involves replacing the convolutional guided filter network with bilinear upsampling operation. Method 2 utilizes the input image as the guidance image for the proposed guided filter network. Method 3 applies the traditional guided filter algorithm while keeping the guidance image unchanged. The fusion results are shown in Fig. 12 and quantified in Table 4. Panels (a) - (c) are the result images corresponding to method 1 to 3, and panel (d) displays the result image of WearIF. Red boxes highlight the magnified regions of representative wear particles. Quantitative and qualitative evaluations demonstrate that WearIF achieves superior performance, delivering optimal image clarity, contrast, and edge/texture preservation. Objectively, it excels in extracting critical information (highest FMI), preserving more image features (peak $Q_g$), retaining structural fidelity (maximized $Q_y$), aligning better with human visual perception (best $Q_{cb}$), and maintaining source correlation (highest SCD). Method 1 produces the weakest fusion results, exhibiting severe texture loss on particles and significant artifacts around edges. Its objective evaluation results, particularly the near-zero SCD, highlight the model's inability to effectively extract and preserve structure and feature information due to the lack of a guided filter network. Method 2 yields a decline in overall clarity and contrast, with noticeable texture loss on particles and amplified edge artifacts, resulting in lower objective evaluation results. Method 3 performs marginally worse than WearIF but retains acceptable clarity, contrast and texture details. However, its evident artifacts around sphere particles and lower objective scores validate the advantage of convolutional layers over  manually defined mathematical operations  in learning nonlinear features via the guided filter network.

### 5.2. Comparison of different loss functions

To evaluate the impact of individual loss terms on the training results, three ablation studies are conducted: (1) Ablation 1, without the constraint term; (2) Ablation 2, without the gradient loss term; and (3) Ablation 3, without the content loss term. The results are presented in Fig. 13 and Table 5, where Panels (a) - (c) correspond to the results of Ablation 1 to 3, and Panel (d) displays the result of WearIF.

In Ablation 1, texture details retained on particles without edge artifacts, and the objective evaluation results of *FMI,* $Q_g$, $Q_y$ and SCD remain marginally lower than those of WearIF, indicating robust fusion capability. However, the global brightness distribution is imbalanced, with lower $Q_{cb}$ score, suggesting the MSE constraint is critical for brightness consistency.

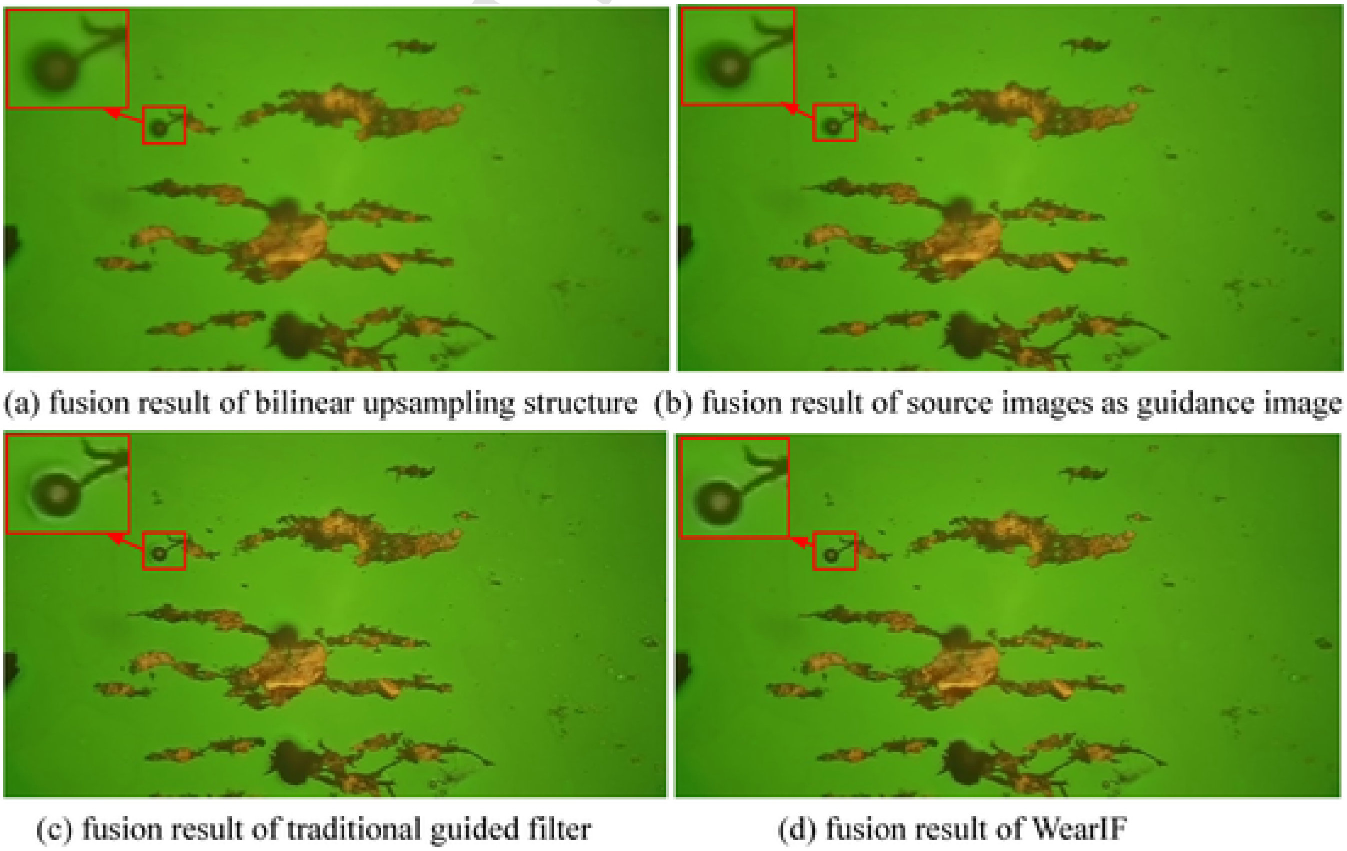
In Ablation 2, the global brightness is balanced with artifact-free edges. The result of $Q_{cb}$ is slightly inferior to WearIF’s, but texture degradation occurs, particularly in $Q_g$ performance. This underscores the gradient loss term as essential for learning gradient-driven texture features critical to detail preservation.

In Ablation 3, the fused result exhibits imbalanced brightness and contrast, substantial texture loss, and pronounced edge artifacts, yielding the poorest visual performance and objective performance (near-zero SCD). This confirms the content loss $L_{con}$ as indispensable for preserving structural integrity and perceptual features via SSIM-based brightness, gradient, and contrast modeling.

Collectively, these results demonstrate that the combined loss terms, particularly the SSIM-based content loss, play a crucial role in optimizing the model for balanced brightness, textural fidelity,and structural coherence.

### 5.3. Comparison of different feature extraction network structures

To investigate the impact of different MDFFEN architectures on fusion performance, five ablation experiments are conducted. The results are summarized in Table 6,where the second-best performance for each


(a) fusion result of bilinear upsampling structure (b) fusion result of source images as guidance image

(c) fusion result of traditional guided filter (d) fusion result of WearIF

**Fig. 12.** Fusion result images with different guided filter methods.

**Table 4**
Objective evaluation results of different guided filter methods.

| Index | Structure | *FMI* | $Q_g$ | $Q_y$ | $Q_{cb}$ | SCD |
|---|---|---|---|---|---|---|
| 1 | bilinear upsampling | 0.0745 | 0.1781 | 0.1165 | 0.1940 | 0.0804 |
| 2 | Source images as guidance image | 0.1531 | 0.2554 | 0.3853 | 0.3312 | 0.1031 |
| 3 | Traditional guided filter | 0.3255 | 0.4259 | 0.4945 | 0.5193 | 0.1759 |
| 4 | WearIF | **0.4258** | **0.5319** | **0.6707** | **0.5953** | **0.2573** |

metric is marked with an asterisk (*). The analysis reveals that the inclusion or exclusion of the three proposed structures significantly affects the fusion quality, yielding either optimal or suboptimal performance (see Experiments 1 and 5).

In Experiment 2 (only incorporating atrous convolution), the network achieves the second-best $Q_g$ and $Q_y$ scores, matching WearIF's performance. This validates that atrous convolution enhances global structural and semantic feature extraction via expanded receptive fields.

In Experiment 3 (solely using dense connections), the network achieves second-best FMI and $Q_{cb}$ scores, indicating that dense connection improves inter-layer information flow, thereby boosting human-aligned perceptual quality.

In Experiment 4 (solely using CC-ASPP module), the network ranks second in SCD, demonstrating that CC-ASPP mitigates grid effects while enlarging receptive fields, enabling joint global-local feature learning for multi-scale fusion.

## 6. Comparison with Other multi-focus image fusion algorithms

To validate the superiority of WearIF in multi-focus ferrograph image fusion, we compared it with five deep learning-based multi-focus image fusion models. All comparisons adopt a pairwise serial fusion strategy, with network architectures and parameters initialized from original references and fine-tuned on the aforementioned dataset. The fusion results for two ferrograph sequences are visualized in Fig. 14 and quantified in Table 7.

(1) **WearIF** achieves the best performance across multiple metrics, delivering optimally balanced contrast/brightness and sharp preservation of texture and morphology details on wear particles. It effectively suppresses defocus-induced artifacts while faithfully retaining source image information. However, non-particle regions exhibit slight texture suppression, attributable to the convolutional guided filter's prioritization of particle regions and background smoothing via the loss constraint term, which marginally degrades background detail.

(2) **Generative fusion models** such as MFF-GAN [14] and SwinFusion [15] implicitly learn the inherent fusion rules from multi-focus sequence images to directly synthesize all-in-focus images. While generally clear with well-preserved particle textures and no grid artifacts, the fusion results of these models suffer from structural distortions and spurious features. For instance, MFF-GAN introduces excessively high contrast, resulting in noise and overly enriched texture details that highlight unnecessary background boundaries (e.g., Fig. 14(a)). Similarly, SwinFusion exhibits content distortion, such as black shadows near the center of the sphere particle in Fig. 14(b). These flaws stem from implicit fusion rules in generative fusion models which makes the fusion process less controllable and more sensitive to noise, meanwhile serial fusion strategy further exacerbates these problems. Furthermore, they underperform

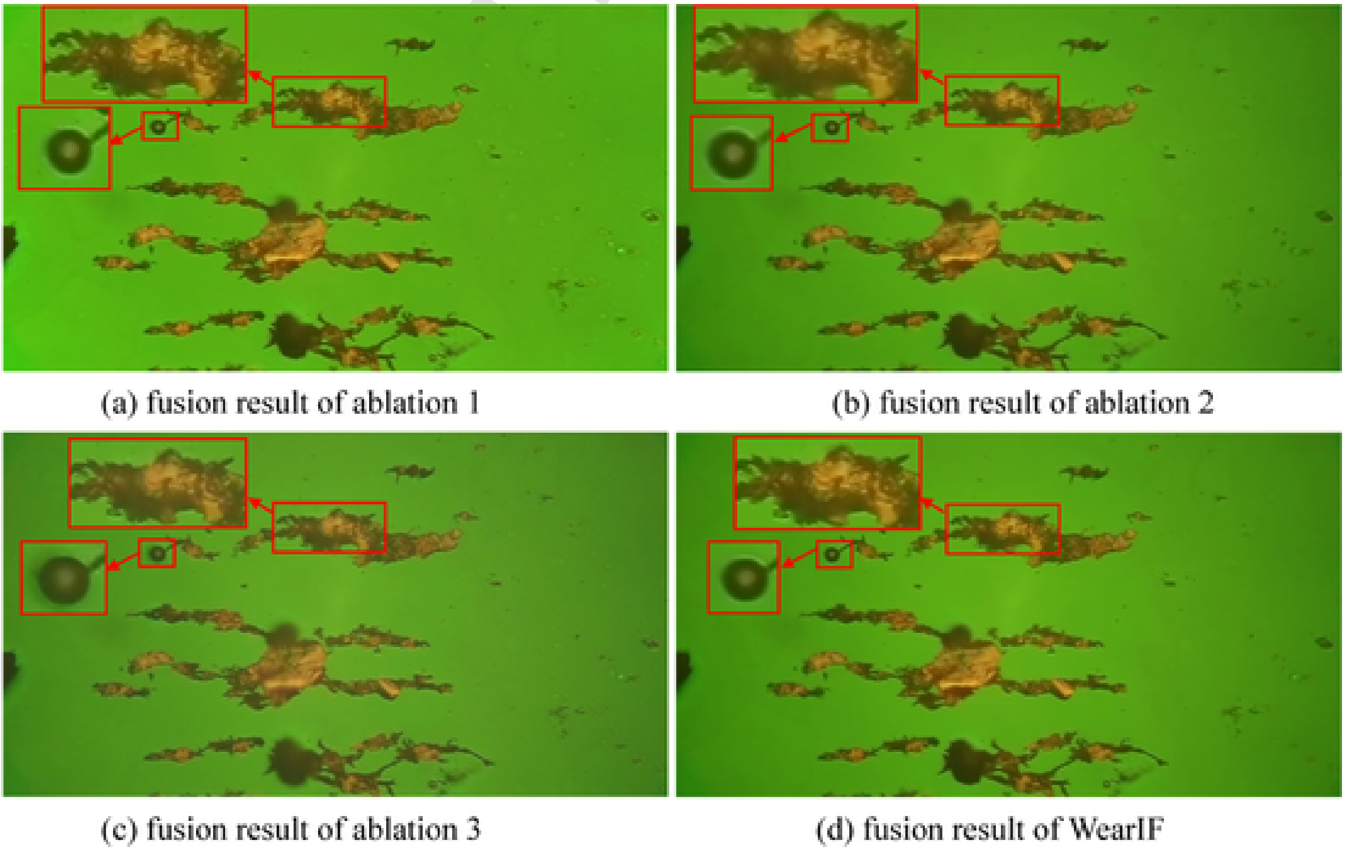

(a) fusion result of ablation 1
(b) fusion result of ablation 2
(c) fusion result of ablation 3
(d) fusion result of WearIF

**Fig. 13.** Fusion results after training with different loss functions.

**Table 5**
Objective evaluation results after training with different loss functions.

| Experiment | MSE | Grad | SSIM | *FMI* | $Q_g$ | $Q_y$ | $Q_{cb}$ | SCD |
|---|---|---|---|---|---|---|---|---|
| Ablation 1 | | √ | √ | 0.3937 | 0.4478 | 0.5269 | 0.4214 | 0.1658 |
| Ablation 2 | √ | | √ | 0.3724 | 0.3565 | 0.4876 | 0.4673 | 0.1462 |
| Ablation 3 | √ | √ | | 0.2332 | 0.3027 | 0.3134 | 0.3012 | 0.0127 |
| WearIF | √ | √ | √ | **0.4258** | **0.5319** | **0.6707** | **0.5953** | **0.2573** |

WearIF in objective evaluation metrics, processing speed and computational efficiency.

(3) **Decision Map-Based Fusion Models** like SESF-Fuse [11], DRPL [12], and GACN [13] generate binary decision maps to select in-focus pixels for fusion, and then combines them with the source images to generate the all-in-focus fusion results. They produce clear and natural fusion results, with well-preserved texture, structure and content of particles from source images. However, significant artifacts at boundaries of particles are observed in some fused results. For instance, SESF-Fuse and DRPL models introduce circular black artifacts around sphere and oxide particles (Fig. 14(c) and (d)), since their binary decision maps are highly sensitive to gradient edges of defocused particles. In contrast, GACN mitigates circular artifacts via a guided filtering algorithm and decision map calibration strategy (Fig. 14(e)). Table 7 shows that while SESF-Fuse and DRPL yield objective evaluation results and processing speeds comparable to WearIF, there remains a slight performance gap. GACN outperforms WearIF in the $Q_g$ and SCD metrics, indicating higher correlation with the source images and greater information retention. However, GACN fails to fully resolve the defocus-induced artifact issue, and slower processing limit its practicality.

WearIF outperforms current multi-focus image fusion models in terms of fusion quality, maintaining a balance between clarity, detail retention, and processing speed while mitigating defocus artifacts more effectively.

## 7. Conclusion

In this study, we proposed a novel end-to-end unsupervised weighted multi-focus image fusion model for ferrograph images, named WearIF. The model integrates a multi-scale dense focus feature extraction network MDFFEN with a convolutional guided filter network to effectively extract and refine the focus weight maps of wear particles from both low- and high-resolution sequence images. The model is optimized using a joint content and gradient loss function, which facilitates the preservation of the overall image structure, enhances texture details, and ensures balanced brightness in the fused result. Experimental results demonstrate that WearIF outperforms existing multi-focus fusion methods in terms of image quality, preserving a greater amount of detail and providing more accurate representations of particle morphology. The fusion results exhibit improved retention of texture, shape, and brightness balance, which is critical for precise wear particle analysis. Furthermore, WearIF demonstrates high computational efficiency, making it highly suitable for practical applications in wear monitoring and fault diagnosis.

## CRediT authorship contribution statement

**Xinliang Liu:** Writing – original draft, Software, Methodology, Investigation, Data curation. **Jingqiu Wang:** Writing – review & editing, Validation, Methodology, Investigation. **Xiaolei Wang:** Writing – review & editing, Validation, Methodology.

## Declaration of competing interest

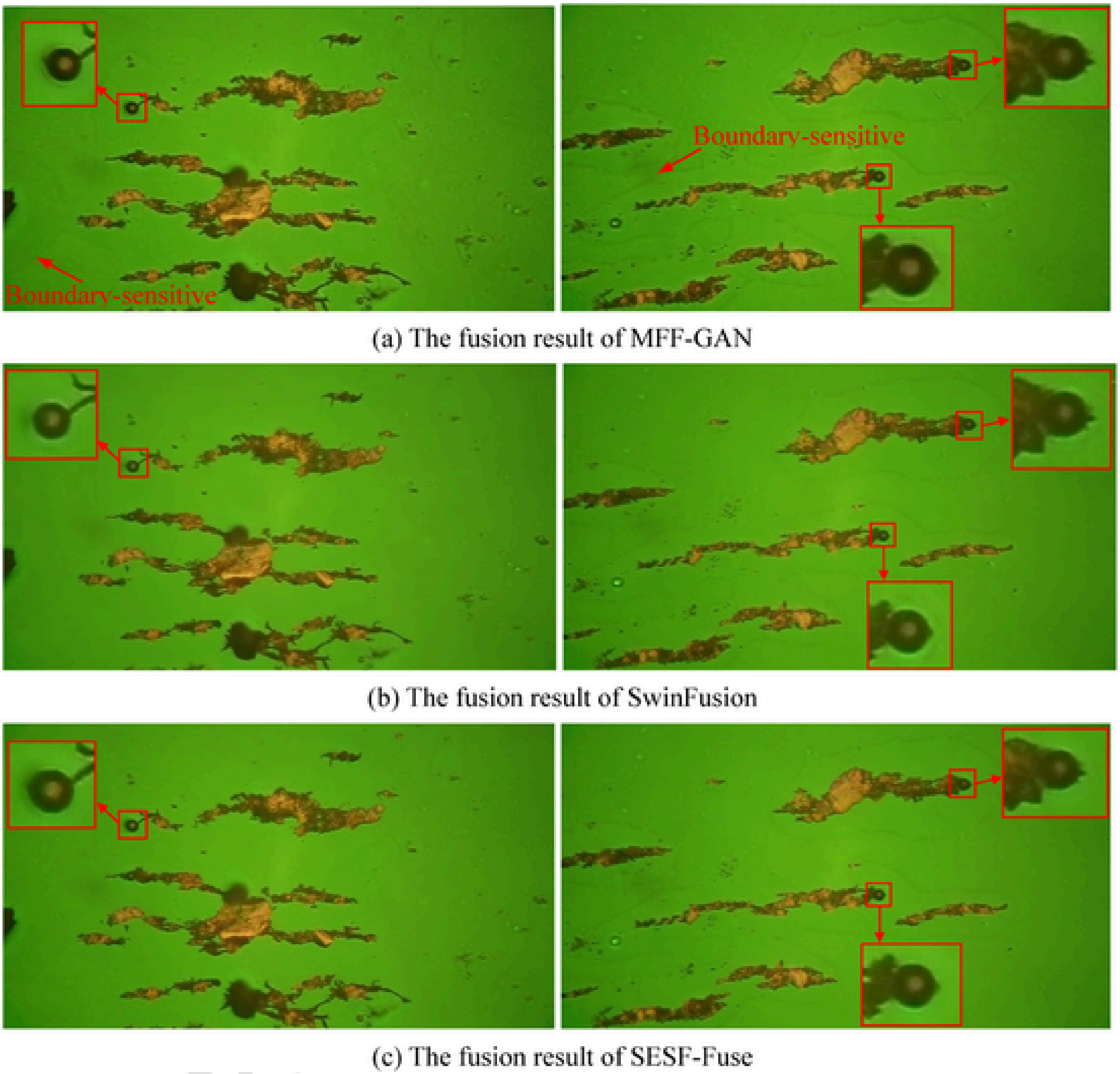
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The data that has been used is confidential.

**Table 6**
Objective evaluation results of different feature extraction network architectures.

| Index | Atrous convolution | Dense connection | CC-ASPP | *FMI* | $Q_g$ | $Q_y$ | $Q_{cb}$ | SCD |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | 0.3695 | 0.4596 | 0.5562 | 0.4741 | 0.2080 |
| 2 | √ | | | 0.4079 | 0.5204* | 0.6337* | 0.5398 | 0.2144 |
| 3 | | √ | | 0.4146* | 0.5017 | 0.6038 | 0.5503* | 0.2136 |
| 4 | | | √ | 0.3993 | 0.5013 | 0.5989 | 0.5374 | 0.2398* |
| 5 | √ | √ | √ | **0.4258** | **0.5319** | **0.6707** | **0.5953** | **0.2573** |

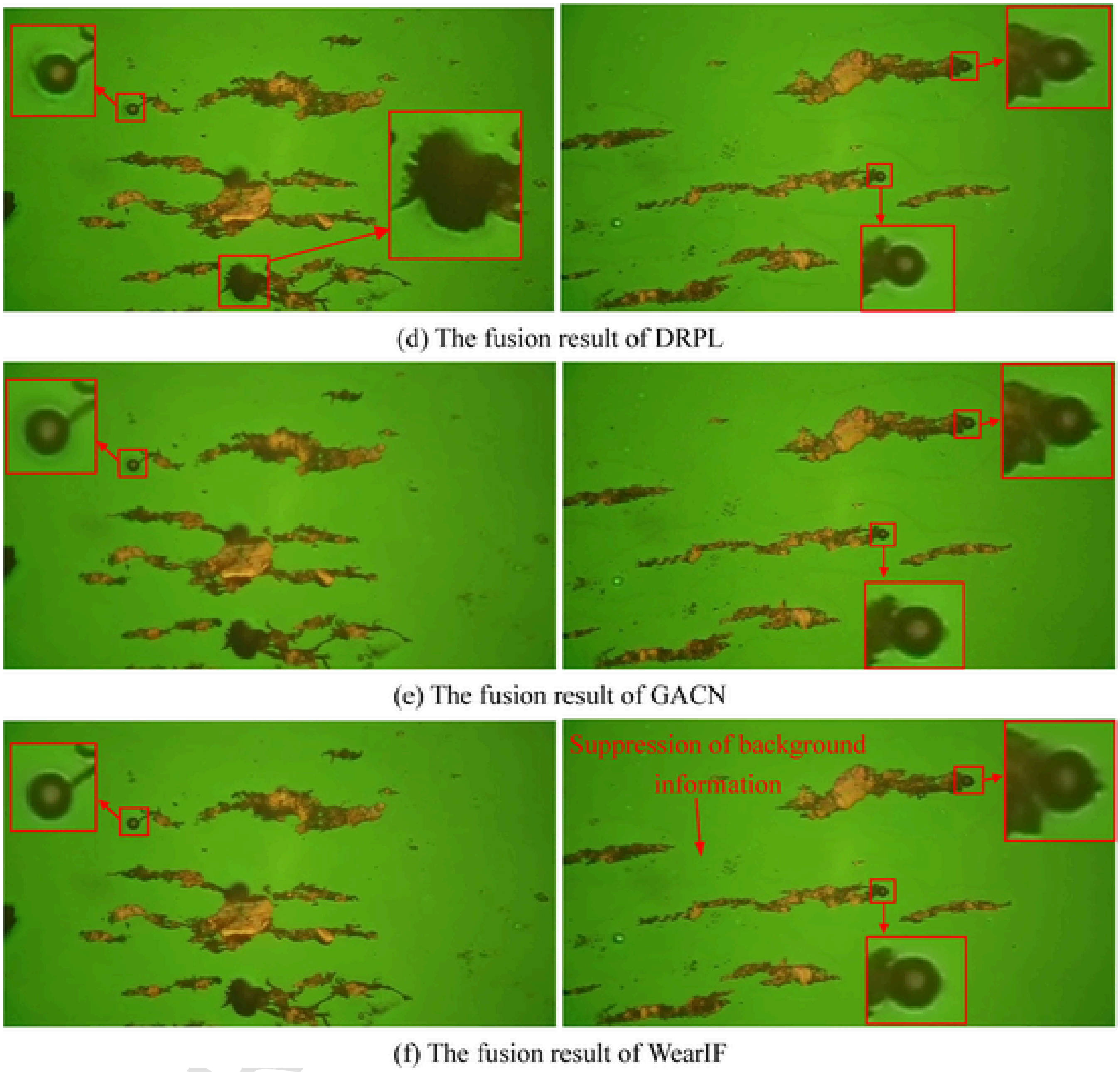


(a) The fusion result of MFF-GAN

(b) The fusion result of SwinFusion

(c) The fusion result of SESF-Fuse

**Fig. 14.** Results of different multi-focus image fusion algorithms.



(d) The fusion result of DRPL

(e) The fusion result of GACN

(f) The fusion result of WearIF

**Fig. 14.** (*continued*)

**Table 7**
Objective evaluation results of different multi-focus image fusion algorithms.

| Method | *FMI* | $Q_g$ | $Q_y$ | $Q_{cb}$ | SCD | Time(s) |
|---|---|---|---|---|---|---|
| MFF-GAN | 0.3157 | 0.4547 | 0.3968 | 0.3272 | 0.1782 | 2.45 |
| SwinFusion | 0.3018 | 0.4291 | 0.2718 | 0.4932 | 0.2055 | 2.76 |
| SESF-Fuse | 0.3309 | 0.4953 | 0.5640 | 0.4779 | 0.1987 | 0.62 |
| DRPL | 0.3840 | 0.3986 | 0.4386 | 0.5311 | 0.2420 | 1.52 |
| GACN | 0.4099 | **0.5371** | 0.6579 | 0.5716 | **0.2619** | 1.81 |
| WearIF | **0.4258** | 0.5319 | **0.6707** | **0.5953** | 0.2573 | **0.53** |

## References

[1] Roylance B J. Ferrography—Then and now. J Tribol Int 2005;38(10):857–62.

[2] Xi W, Wu T, Yan K, et al. Restoration of online video ferrography images for out-of-focus degradations. J Video Proc 2018;2018(1).

[3] Wu H, Kwok N M, Liu S, et al. Restoration of defocused ferrograph images using a large kernel convolutional neural network. Wear 2019;426:1740–7.

[4] Liu X, Zhang L, Leng S, et al. An autofocus algorithm for fusing global and local information in ferrographic images. Chin Opt. 2024;17(2):423–34.

[5] Li H, Manjunath B S, Mitra S K. Multisensor image fusion using the Wavelet transform. Graph Models Image Proc 1995;57(3):235–45.

[6] Burt P J, Adelson E H. The laplacian pyramid as a compact image code. IEEE Trans Commun 1983;COM-31(4):532–40.

[7] Shutao L, Kwok J T, Yaonan W. Combination of images with diverse focuses using the spatial frequency. Inf Fusion 2001;2(3):169–76.

[8] Zhou Z, Li S, Wang B. Multi-scale weighted gradient-based fusion for multi-focus images. Inf Fusion 2014;20:60–72.

[9] Liu Y, Liu S, Wang Z. Multi-focus image fusion with dense SIFT. Inf Fusion 2015;23:139–55.

[10] Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85–117.

[11] Ma B, Zhu Y, Yin X, et al. SESF-Fuse: an unsupervised deep model for multi-focus image fusion. Comput Sci Eng 2021;33(11):5793–804.

[12] Li J, Guo X, Lu G, et al. DRPL: deep regression pair learning for Multi-focus image fusion. Ieee T Image Proc 2020;29:4816–31.

[13] Ma B, Yin X, Wu D, et al. End-to-end learning for simultaneously generating decision map and multi-focus image fusion result. Neurocomputing 2022;470:204–16.

[14] Zhang H, Le Z, Shao Z, et al. MFF-GAN: an unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. Inf Fusion 2021;66:40–53.

[15] Ma J, Tang L, Fan F, et al. SwinFusion: cross-domain long-range learning for General image Fusion via Swin Transformer. Ieee-Caa J Autom Sin. 2022;9(7): 1200–17.

[16] Yin X, Ma B, Ban X, et al. Defocus spread effect elimination method in multiple multi-focus image fusion for microscopic images. Chin J Eng 2021;43(9):1174–81.

[17] Liu X, Cheng L, Chen G, et al. Recognition of fatigue and severe sliding wear particles using a CNN model with multi-scale feature extractor. Ind Lubr Tribol 2022;74(7):884–91.

[18] Huang G, Liu Z, Maaten L V D, et al. Densely connected convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 2261–9.

[19] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 2018;40(4):834–48.

[20] He K, Sun J, Tang X. Guided image filtering [C]. In: Proceedings of the 11th European Conference on Computer Vision; 2010. p. 1. +.

[21] Wu H, Zheng S, Zhang J, et al. Fast end-to-end trainable guided filter. In: Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 1838–47.

[22] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. Ieee T Image Proc 2004;13(4):600–12.

[23] Ma K, Zeng K, Wang Z. Perceptual quality assessment for Multi-Exposure image fusion. Ieee T Image Proc 2015;24(11):3345–56.

[24] Ma K, Duanm Z, Zhu H, et al. Deep guided learning for fast Multi-exposure image fusion. Ieee T Image Proc 2020;29:2808–19.

[25] Xydeas C S, Petrovic V. Objective image fusion performance measure. Comput Sci Eng 2000;36(4):308–9.

[26] Li S, Hong R, Wu X. A novel similarity based quality metric for image fusion. In: Proceedings of the International Conference on Audio, Language and Image Processing; 2008. p. 167–72.

[27] Chen Y, Blum R S. A new automated quality assessment algorithm for image fusion. Image Vision Comput 2009;27(10):1421–32.

[28] Haghighat M, Razian M A. Ieee. Fast-FMI: non-reference image fusion metric. In: Proceedings of the 8th IEEE International Conference on Application of Information and Communication Technologies (AICT); 2014. p. 424–6.

[29] Aslantas V, Bendes E. A new image quality metric for image fusion: the sum of the correlations of differences. Aeu-Int J Electron Commun 2015;69(12):160–6.