Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ro-FusionGAN:An Adversarial Framework for High-Quality Multi-focus image fusion

Yongli Xian

xyl123@mail.ustc.edu.cn

Xihua University

Heng Zhou

Xihua University

Zhijie Gong

Xihua University

Congzheng Wang

Chinese Academy of Sciences

Research Article

Keywords: Multi-focus image fusion, Generative Adversarial Network, Result-Oriented Learning, Soft Decision Map, End-to-End.

Posted Date: March 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9113517/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Signal, Image and Video Processing on October 14th, 2026. See the published version at https://doi.org/10.1007/s11760-026-05762-z.

Ro-FusionGAN:An Adversarial Framework for High-Quality Multi-focus image fusion

Yongli Xian[1,2*], Heng Zhou[1], Zhijie Gong[1], Congzheng Wang[3,4*]

[1]School of Electrical Engineering and Electronic Information, Xihua University, Chengdu, 610039, China.
[2]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, 610054, China.
[3]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, 610209, China.
[4]University of Chinese Academy of Sciences, Beijing, 100049, China.

*Corresponding author(s). E-mail(s): xyl123@mail.ustc.edu.cn; wangcongzheng@ioe.ac.cn;

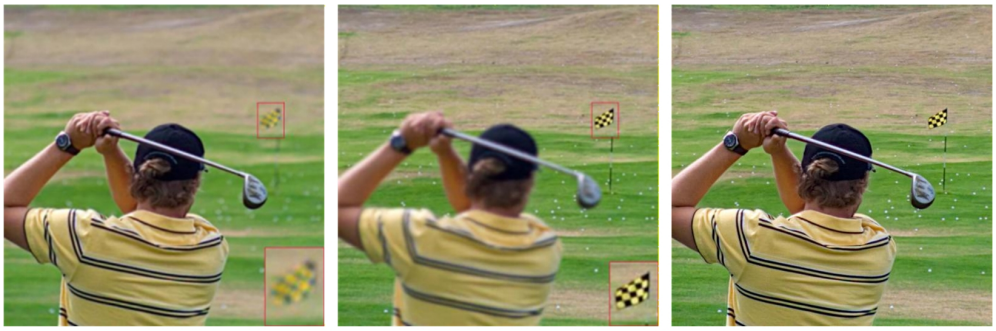
Abstract

Multi-focus image fusion (MFIF) aims to synthesize an all-in-focus image from source images with varying focal depths. While deep learning has advanced this field, existing "process-imitation" approaches often rely on binary mask supervision, leading to boundary artifacts and a dependency on heuristic post-processing. To address these limitations, we propose Result-Oriented Fusion Generative Adversarial Network (Ro-FusionGAN), a novel result-oriented adversarial framework. Unlike methods that mimic intermediate focus maps, our framework directly optimizes the perceptual quality of the final fused image via a composite fusion-aware loss function. Furthermore, we introduce a differentiable Total Variation (TV) regularizer to autonomously enforce spatial smoothness, enabling the generation of soft decision maps and eliminating the need for post-processing. Extensive experiments demonstrate that Ro-FusionGAN outperforms eleven state-of-the-art methods in visual quality, quantitative metrics, and computational efficiency, yielding artifact-free images with natural focus transitions.

Keywords: Multi-focus image fusion, Generative Adversarial Network, Result-Oriented Learning, Soft Decision Map, End-to-End.

1 Introduction

The finite depth-of-field (DOF) in optical systems often results in images where only partial regions are in focus, a phenomenon complicated by the defocus spread effect (DSE). Multi-focus image fusion addresses this by synthesizing a single all-in-focus image from a sequence of source images (Fig. 1), serving as a critical preprocessing step for downstream computer vision tasks [1, 2].



(a) (b) (c)

Fig. 1 The source of MFIF images and desired fusion result.

Historically, MFIF methodologies are categorized into transform domain and spatial domain approaches. Transform domain methods (e.g. NSCT [3], Sparse Representation [4]) efficiently separate features but often introduce artifacts near focus boundaries due to windowing effects [5]. Spatial domain methods operate directly on pixel intensities; however, they remain sensitive to misregistration and block size choices, often depending heavily on segmentation accuracy [6, 7].

In recent years, deep learning (DL) has dominated MFIF research [8–13]. However, a prevalent "process-imitation" paradigm—training networks to replicate binary ground-truth focus masks—suffers from a fundamental limitation. As illustrated in Fig.2(a), hard binary masks fail to model the gradual transition at the focus-defocus boundary (FDB)[14], causing unnatural edges and artifacts. Consequently, many state-of-the-art methods rely on non-differentiable post-processing to refine these masks, compromising the end-to-end integrity of the pipeline. Ideally, the fusion process should instead leverage a "soft" decision map with smooth transitions (Fig.2 (b)).

To overcome these limitations, we propose Ro-FusionGAN, a novel result-oriented adversarial framework. Departing from mask imitation, we task the generator with a direct objective: to produce a fused image that is perceptually indistinguishable from an ideal all-in-focus ground truth. By employing a composite fusion-aware loss and a TV regularizer, our network autonomously learns a spatially continuous soft decision map. This approach eliminates the need for post-processing and effectively suppresses boundary artifacts.

Our main contributions are summarized as follows:



Fig. 2 I_A without DSE and I_B suffering from DSE.

1. We propose Ro-FusionGAN, a result-oriented GAN for MFIF. It utilizes a novel fusion-aware loss function to directly optimize the perceptual quality of the final image rather than mimicking binary masks.
2. We establish a fully end-to-end differentiable pipeline. The integration of a TV loss encourages the autonomous generation of smooth soft decision maps, obviating heuristic post-processing.
3. Extensive experiments show that our method achieves state-of-the-art results, yielding fused images with significantly reduced artifacts and natural focus transitions.
4. We design a parameter-efficient, weight-shared encoder that effectively extracts robust features across the full RGB space, ensuring high performance with low computational cost.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 details the proposed method; Section 4 presents experimental results. Ablation experiments are also conducted in this section; and Section 5 concludes the paper.

2 Related Work

2.1 Defocus Spread Effect and Dataset Synthesis

A critical shortcoming in many MFIF datasets is the failure to accurately model the DSE, a physical phenomenon where out-of-focus light blurs across boundaries [15]. Models trained on datasets neglecting DSE often generalize poorly to real-world scenarios due to the gap between synthetic and authentic optical effects [16]. To address this, Ma et al. [5] introduced a data generation method based on the α-matte model,

which simulates realistic interactions at FDB. In this work, we adopt this strategy to construct a high-fidelity training dataset that robustly prepares our model for the complexities of real-world fusion tasks.

2.2 Attention Mechanisms in Focus Discrimination

Beyond standard CNN architectures [17], attention mechanisms have been widely integrated to enhance feature representation. In MFIF, the core challenge is to distinguish sharp regions from blurred ones. We employ the Squeeze-and-Excitation (SE) block [18], a computationally efficient channel attention module, to dynamically prioritize informative features. Unlike previous works that simply stack layers, we embed SE blocks within a residual backbone. This design enables our shared-weight encoder to explicitly model channel interdependencies, extracting highly discriminative features crucial for capturing subtle focus variations and texture details across the full RGB space.

2.3 From Mask-Imitation to Result-Oriented Fusion

Generative Adversarial Networks (GANs) [17, 19–23] have advanced image fusion tasks significantly [24, 25]. However, early adversarial MFIF approaches typically frame the problem as an "image-to-mask translation" task, where the generator is trained to mimic binary ground-truth masks. This paradigm suffers from three key limitations:

1. Loss of Color Information: Independent processing of RGB channels often dilutes color-specific focus cues.
2. Indirect Supervision: Optimizing for pixel-wise mask similarity rather than the final fusion quality often leads to boundary artifacts, as a perfect binary mask cannot represent gradual defocus transitions [26].
3. Dependency on Post-Processing: Generated masks often contain noise, necessitating heuristic, non-differentiable post-processing that disrupts the end-to-end training pipeline [27, 28].

Motivated by these issues, our Ro-FusionGAN redefines the adversarial objective. Instead of mask imitation, we propose a result-oriented framework where the generator is explicitly tasked with optimizing the perceptual quality of the final fused image, thereby eliminating the need for post-processing and ensuring artifact-free results.

3 Method

In this study, we propose Ro-FusionGAN, a supervised, end-to-end GAN for the MFIF task. Our approach systematically resolves the limitations of conventional mask-imitation models by introducing a novel result-oriented framework.

3.1 Workflow Overview

The overall architecture of Ro-FusionGAN is illustrated in Fig.3, comprising a generator configured as a symmetric Encoder-Decoder network and a discriminator designed as a deep convolutional classifier.

The generator features a weight-sharing encoder that simultaneously processes the 3-channel RGB source images (I_A, I_B). This strategy not only preserves critical cross-channel chromatic information but also ensures consistent feature extraction while maximizing parameter efficiency. The encoding pathway initiates with a 7×7 convolutional layer, followed by stride-2 convolutions for down-sampling. To capture highly discriminative features for focus detection, we employ a deep stack of SE-ResNet blocks, which adaptively recalibrate channel-wise feature responses.

Subsequently, the extracted deep feature maps, Feature A and Feature B, are concatenated along the channel dimension. This fused tensor is then propagated through the decoder, which progressively restores the spatial resolution via up-sampling operations. Finally, a terminal 7×7 convolutional layer with a Sigmoid activation function generates the single-channel soft decision map $\hat{F}$.

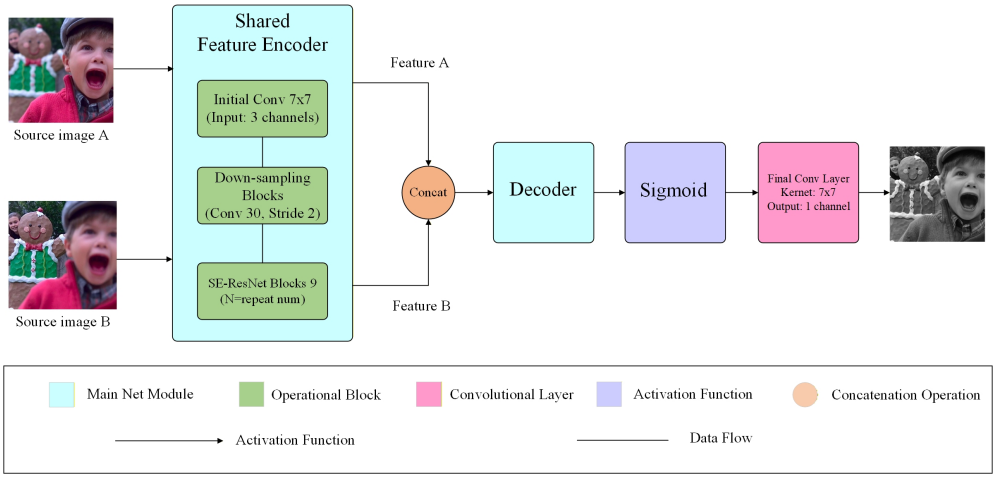


Fig. 3 Detailed architecture of the proposed Ro-FusionGAN generator.

3.2 Result-Oriented Objective Function

To guide the generator towards producing perceptually superior results, we introduce a composite, result-oriented objective function. This function shifts the supervision paradigm from merely reconstructing a binary mask to directly optimizing the quality of the final fused image. For stable adversarial training, we employ the WGAN-GP objective. The key innovation lies in our generator's loss, which is a weighted sum of four components:

- Adversarial Loss ($\mathcal{L}_{adv}$): Encourages the generator to produce outputs that are indistinguishable from real, perfectly fused images.
- Fusion L_1 Loss ($\mathcal{L}_{fusion_L1}$): Directly minimizes the pixel-wise absolute difference between the generator's fused image and an ideal ground-truth fusion.

- Fusion SSIM Loss ($\mathcal{L}_{fusion_SSIM}$): Preserves structural and textural integrity by maximizing the similarity between the generated and ideal fused results.
- Total Variation (TV) Loss ($\mathcal{L}_{tv}$): Acts as a differentiable regularizer on the decision map $\hat{F}$, promoting spatial smoothness and eliminating the need for any post-processing steps.

This multi-faceted loss function ensures that our RO-FusionGAN is trained end-to-end to produce high-quality, artifact-free fused images.

3.3 Detailed Network Architecture

As a variant of the GAN, our proposed Ro-FusionGAN comprises two main components: a generator (G) and a discriminator (D). The complete architecture, detailing both the high-level data flow and low-level module specifications, is illustrated in Fig.4.

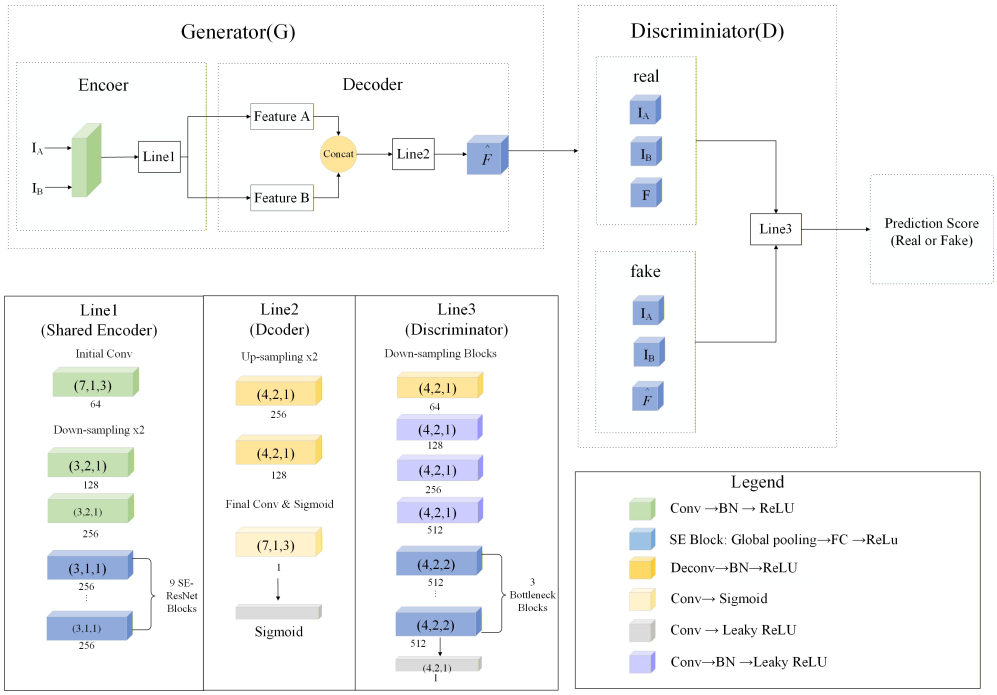


Fig. 4 Architecture of the proposed Ro-FusionGAN. The network consists of a G with a shared-weight Encoder-Decoder structure (Line1, Line2) and a D (Line3). The top row shows the overall data flow, while the bottom row provides detailed specifications for each module. Numbers indicate output channels, and tuples (k, s, p) define convolution parameters.

3.3.1 Generator

The generator's primary role is to translate a pair of source images (I_A, I_B) into a high-quality, soft decision map $\hat{F}$. It is designed as a symmetric Encoder-Decoder network. A key architectural choice is the use of a single, shared-weight encoder (Line1) to process both 3-channel RGB inputs. This approach ensures consistent feature extraction while maintaining parameter efficiency. As detailed in Fig.4, the encoder begins with an initial convolutional layer, followed by stride-2 convolutions for down-sampling. Its core is a deep stack of SE-ResNet blocks, which leverage channel attention to adaptively emphasize the most informative features for focus discrimination.

The resulting deep feature maps (Feature A, Feature B) are then concatenated and fed into the decoder (Line2). The decoder progressively up-samples this fused representation using transposed convolutions, restoring the original spatial resolution. Finally, a terminal convolutional layer with a Sigmoid activation function produces the final single-channel, soft decision map $\hat{F}$.

3.3.2 Discriminator

The D is tasked with distinguishing the generator's synthesized map $\hat{F}$ from the real ground-truth map F. As depicted in Fig. 4, its input is a 7-channel tensor formed by concatenating the source images (I_A, I_B) with either a real or a fake map. The architecture of the discriminator (Line3) is a deep convolutional classifier composed of a series of stride-2 convolutional layers, each followed by Batch Normalization and a Leaky ReLU activation. This design effectively down-samples the input to extract increasingly abstract features, ultimately outputting a single scalar score that represents the realness prediction.

3.4 Loss Function

The adversarial loss for the D and the G is formulated based on the *Wasserstein distance* with gradient penalty. The objective for the discriminator is to maximize the difference between its evaluation of real and fake samples, while the generator aims to minimize this difference. The adversarial losses are defined in Eq. (1) and (2):

$$\mathcal{L}_{adv}(D) = \mathbb{E}_{I_A, I_B, \hat{F} \sim \mathbb{P}_q} \left[D(I_A, I_B, \hat{F}) \right] - \mathbb{E}_{I_A, I_B, F \sim \mathbb{P}_g} \left[D(I_A, I_B, F) \right] \tag{1}$$

$$\mathcal{L}_{adv}(G) = -\mathbb{E}_{I_A, I_B, \hat{F} \sim \mathbb{P}_g} \left[D(I_A, I_B, \hat{F}) \right] \tag{2}$$

where $\hat{F} = G(I_A, I_B)$. $\mathbb{P}_g$ denotes the real data distribution, and $\mathbb{P}_q$ represents the distribution of the data generated by G. To enforce the *Lipschitz* constraint required by the *Wasserstein* distance, a gradient penalty term $\mathcal{L}_{gp}$ is added to the discriminator's loss, as shown in Eq. (3).

$$\mathcal{L}_{gp} = \mathbb{E}_{\tilde{F} \sim \mathbb{P}_{\tilde{F}}} \left[\left(\|\nabla_{\tilde{F}} D(I_A, I_B, \tilde{F})\|_2 - 1 \right)^2 \right] \tag{3}$$

Where $\tilde{F}$ is sampled uniformly along the straight line between points sampled from the real distribution F and the generated distribution $\hat{F}$.

3.4.1 Fusion-Aware Loss

As discussed, a core innovation of our work is the result-oriented supervision strategy. Instead of solely relying on the reconstruction of the decision map, we introduce a composite fusion-aware loss $\mathcal{L}_{fusion_L1}$ that directly measures the quality of the final fused image. This loss consists of two components: a pixel-wise L_1 loss and a structure-aware SSIM loss.

The fusion L_1 loss, $\mathcal{L}_{fusion_L1}$, measures the absolute difference between the ideally fused image and the image fused with the generator's map $\hat{F}$, as shown in Eq. (4).

$$\mathcal{L}_{fusion_L1} = \mathbb{E}[||(I_A \odot \hat{F} + I_B \odot (1 - \hat{F})) - (I_A \odot F + I_B \odot (1 - F))||] \tag{4}$$

where $\odot$ denotes element-wise multiplication.

The fusion SSIM loss, $\mathcal{L}_{fusion_SSIM}$, evaluates the structural similarity between the two fused images, which is crucial for preserving texture and perceptual quality, as shown in Eq. (5).

$$\mathcal{L}_{fusion_SSIM} = 1 - \text{SSIM}(I_A \odot \hat{F} + I_B \odot (1 - \hat{F}), I_A \odot F + I_B \odot (1 - F)) \tag{5}$$

3.5 Regularization Loss

To ensure the generated decision map $\hat{F}$ is spatially smooth and free of artifacts, we incorporate a TV loss $\mathcal{L}_{tv}$ as a regularization term. This term intrinsically encourages coherence between neighboring pixels in the decision map, effectively suppressing high-frequency noise and enforcing spatial continuity, as shown in Eq. (6).

$$\mathcal{L}_{tv} = \mathbb{E}[||\nabla_x \hat{F}||_2^2 + ||\nabla_y \hat{F}||_2^2] \tag{6}$$

3.5.1 Overall Objective Functions

By combining the aforementioned loss components, the final objective functions for the discriminator D and the generator G of our Ro-FusionGAN can be defined by Eq. (7) and (8), respectively:

$$\min \mathcal{L}(D) = \mathcal{L}_{adv}(D) + \lambda_{gp}\mathcal{L}_{gp} \tag{7}$$

$$\min \mathcal{L}(G) = \mathcal{L}_{adv}(G) + \lambda_{fusion_L1}\mathcal{L}_{fusion_L1} + \lambda_{fusion_SSIM}\mathcal{L}_{fusion_SSIM} + \lambda_{tv}\mathcal{L}_{tv} \tag{8}$$

where λ_{ap}, λ_{fusion_L1}, λ_{fusion_SSIM}, and λ_{tv} are hyperparameters that balance the contribution of each loss term. In our experiments, we set $\lambda_{ap} = 10$, following the standard practice in [29]. The other weights are determined empirically to achieve the best fusion performance.

3.6 End-to-End Training Strategy

Many deep learning-based MFIF methods rely on heuristic post-processing, such as Small Region Removal (SRR) [30], to mitigate noise in generated decision maps. However, these non-differentiable operations exist outside the optimization loop and often risk eroding legitimate fine-grained details. Ro-FusionGAN overcomes this limitation by integrating Total Variation regularization directly into the training objective. This constraint compels the generator to autonomously suppress high-frequency noise and produce spatially coherent soft decision maps. Consequently, our framework establishes a robust, fully end-to-end pipeline where the generated map $\hat{F}$ is used directly for fusion, obviating the need for any post-hoc refinement.

4 Experiments

In this section, we evaluate Ro-FusionGAN through comprehensive experiments, covering dataset details, metrics, and implementation setups. To benchmark performance, we compare our method against ten representative SOTA algorithms across three categories: (1) Traditional Spatial Domain methods: GFF [31], SESF [32], MUFusion [33]; (2) Transform Domain methods: BRW [34], CSR [35]; (3) Deep Learning-based methods: CCSR-Net [29], CNN [8], MCCSR-Net [36], MFFGAN [37], U2Fusion [38]. Furthermore, we conduct ablation studies to systematically verify the contributions of key components, specifically the fusion-aware loss terms ($\mathcal{L}_{fusion_L1}$, $\mathcal{L}_{fusion_SSIM}$, $\mathcal{L}_{tv}$) and the SE modules.

4.1 Experimental Setup

4.1.1 Datasets and Data Synthesis

Due to the lack of ground-truth focus maps in real-world datasets, we synthesized a high-fidelity training set using the PASCAL VOC 2012 dataset [39]. We adopted the α-matte generation method from Ma et al.[5] to accurately model the DSE. The synthesis process is formulated as:

$$I_A = F \odot I + (1 - F) \odot \{G_\sigma * [(1 - F) \odot I]\} \tag{9}$$

$$I_B = (1 - F) \odot I + F \odot \{G_\sigma * [F \odot I]\} \tag{10}$$

where G_σ is a Gaussian filter with standard deviation $\sigma = 2.0$. For evaluation, we utilized two standard benchmarks: the Lytro dataset (20 color image pairs) and the Grayscale dataset (10 grayscale image pairs).

4.1.2 Implementation Details

Ro-FusionGAN was implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU. We used the Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$) with an initial learning rate of 10^{-4}, which linearly decayed to zero over the latter half of training. Following WGAN-GP standards, the discriminator was updated 5 times for every generator update. The model was trained for 120,000 iterations with a batch size of 2.

During inference, the generator directly produces a soft decision map $\hat{F}$. The final fused image I_{fused} is reconstructed without post-processing via:

$$I_{fused} = I_A \odot \hat{F} + I_B \odot (1 - \hat{F}) \tag{11}$$

To ensure consistency, grayscale inputs were replicated across 3 channels to match the network's expected input dimensions.

4.1.3 Evaluation Metrics

To objectively assess fusion quality, we employ seven quantitative metrics categorized into four groups: (1) Information Theory-based metrics: Mutual Information (Q_{MI}) [40], gradient-based edge preservation ($Q_{AB/F}$) [41], and Nonlinear Correlation Information Entropy (Q_{NCIE}) [42]; (2) Structure Similarity metrics: Yang's metric (Q_Y) [43]; (3) Human Perception-inspired metrics: Chen-Blum metric (Q_{CB}) [44]; and (4) Feature-based metrics: Contrast-based quality (Q_C) [45] and Contrast Visibility (Q_{CV}) [46]. For all metrics except Q_{CV}, higher scores indicate better performance; for Q_{CV}, a lower score signifies less blur. Detailed formulations are available in the respective references.

4.2 Comparison with SOTA Methods

4.2.1 Quantitative Comparison

Table 1 Objective assessment results of different methods on the "Natural" dataset.

Methods	Q_{MI}	Q_{NCIE}	$Q_{AB/F}$	Q_Y	Q_C	Q_{BC}	Q_{CV}
CSR[35]	1.0596	0.8368	0.7406	0.9469	0.7965	0.7738	15.7332
BRW[34]	1.0944	0.8394	0.7440	0.9707	0.8096	0.7884	20.2566
GFF[31]	1.0662	0.8375	0.7503	0.9740	0.8153	0.7904	16.1960
CNN[8]	1.1097	0.8402	0.7520	0.9780	0.8170	0.7999	16.3189
SESF[32]	1.1577	0.7694	0.7694	0.9827	0.8207	0.8049	16.0159
U2Fusion[38]	0.7725	0.8221	0.6091	0.7915	0.7012	0.5682	97.5885
MFFGAN[37]	0.8047	0.8237	0.6593	0.8822	0.7659	0.6457	42.5640
CCSR-Net[29]	1.1912	0.8456	0.7943	0.9836	0.8203	0.8067	18.6377
MCCSR-Net[36]	1.1942	0.8459	0.7983	0.9846	0.8244	0.8084	16.8185
MUFusion[33]	0.8020	0.8235	0.3961	0.8907	0.7506	0.6780	32.9077
Ro-FusionGAN (ours)	**1.1988**	**0.8474**	**0.8017**	**0.9852**	0.8228	**0.8092**	16.0214

Based on the aforementioned test datasets and evaluation metrics, the quantitative results of the proposed Ro-FusionGAN and other SOTA approaches are compiled in Table 1 and Table 2.

Quantitative results on the Lytro dataset reveal a consistent performance advantage for Ro-FusionGAN. Our method outperforms competing algorithms, as shown in Fig.5, across the majority of metrics. This is particularly evident in feature-based metrics (Q_C) and structure similarity metrics (Q_Y), attributing to our result-oriented,

Table 2 Objective assessment results of different methods on the "Grayscale" dataset.

Methods	Q_{MI}	Q_{NCIE}	$Q_{AB/F}$	Q_Y	Q_C	Q_{BC}	Q_{CV}
CSR[35]	1.0579	0.8365	0.6662	0.9404	0.7957	0.7731	15.7857
BRW[34]	1.0877	0.8388	0.6888	0.9645	0.8081	0.7851	20.3788
GFF[31]	1.0632	0.8372	0.6928	0.9688	0.8140	0.7861	16.3148
CNN[8]	1.1030	0.8396	0.7006	0.9717	0.8155	0.7965	15.9145
SESF[32]	1.1577	0.8431	0.7693	0.9827	0.8217	0.7693	16.0158
U2Fusion[38]	0.7737	0.8219	0.3797	0.7878	0.7054	0.5678	97.1782
MFFGAN[37]	0.8054	0.8236	0.3972	0.8780	0.7410	0.6456	42.4635
CCSR-Net[29]	1.1911	0.8456	0.7943	0.9836	0.8219	0.8066	18.6378
MCCSR-Net[36]	1.1942	0.8459	0.7982	0.9846	0.8217	0.8084	16.8187
MUFusion[33]	0.8011	0.8235	0.3962	0.8905	0.7790	0.6781	32.2326
Ro-FusionGAN (ours)	**1.1973**	**0.8471**	**0.8013**	0.9831	**0.8221**	**0.8108**	16.1428

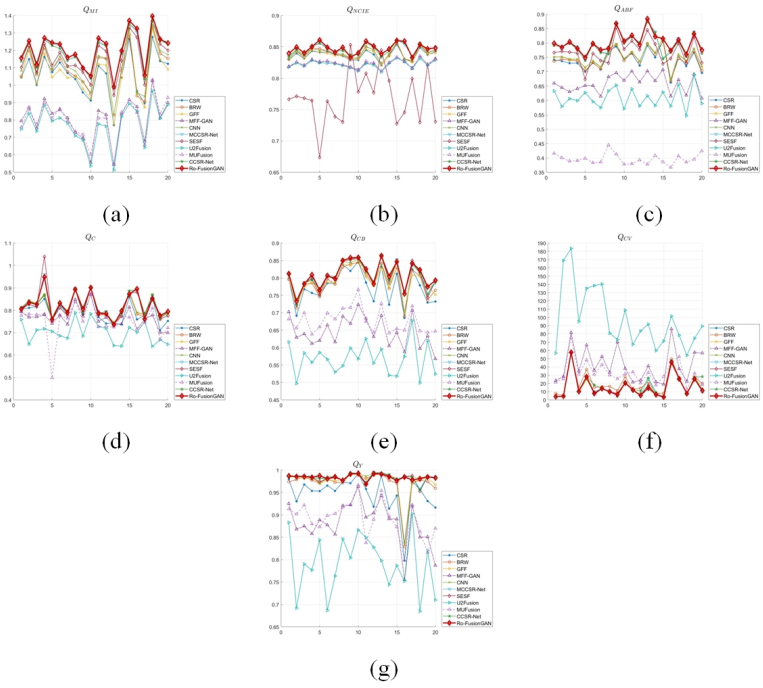


Fig. 5 A more detailed sample-wise comparison of objective performance among the proposed methods and other ten methods on the Lytro dataset.

fusion-aware loss function that directly optimizes the perceptual quality of the final image rather than imitating a binary mask.

Similarly,for the grayscale dataset, our method achieves highly competitive results compared to the top-performing traditional methods, as shown in Fig.6. The consistent high scores across diverse datasets validate the effectiveness and generalization capability of our proposed framework.

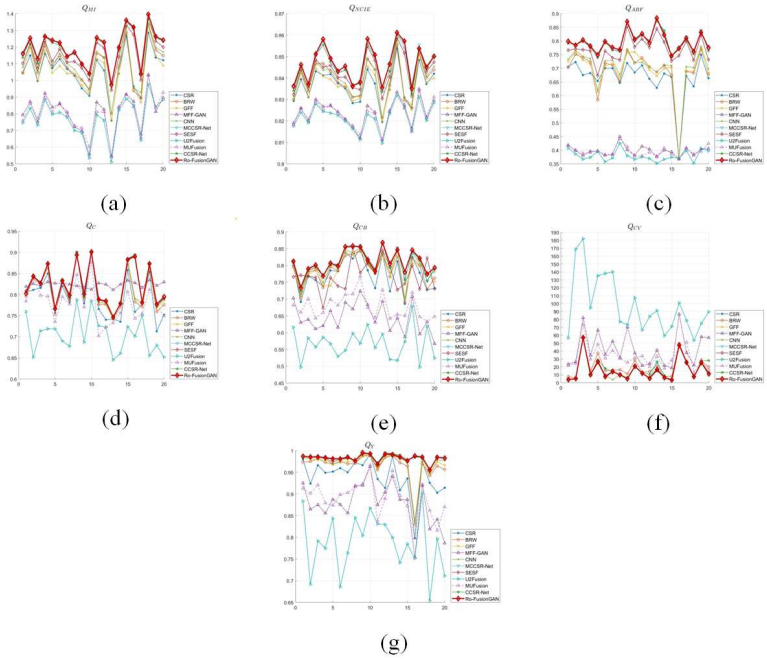


Fig. 6 A more detailed sample-wise comparison of objective performance among the proposed methods and other ten methods on the "Grayscal" dataset.

4.2.2 Qualitative Comparison

While quantitative metrics provide an objective measure of performance, qualitative visual inspection is crucial for assessing perceptual quality, especially regarding artifacts and the naturalness of fused boundaries. Fig.7 presents a detailed visual comparison between our Ro-FusionGAN and a representative baseline method, MFFGAN[37], on a challenging sample from the Lytro dataset.

As observed in the final fusion results Fig.7(h) and Fig.7(i), our method produces a visibly superior image. The baseline method Fig.7(h) exhibits noticeable artifacts and ringing effects along the high-contrast boundary of the subject's hat, an issue highlighted in the magnified inset. In contrast, our Ro-FusionGAN result Fig.7(i) displays clean, sharp edges with a seamless and natural transition between the focused subject and the blurred background.

Visual inspection confirms the superiority of our approach. As observed in Fig.7, the baseline method Fig.7(h) exhibits noticeable ringing effects along the high-contrast boundary of the subject. In contrast, Ro-FusionGAN Fig.7(i) displays clean edges with a seamless transition between the focused subject and the blurred background. This improvement is directly attributable to the spatial coherence of the generated decision map. Guided by the TV loss and fusion-aware objective, our network generates a map Fig.7(a) that is significantly smoother and more refined than the noisy baseline output Fig.7(b).

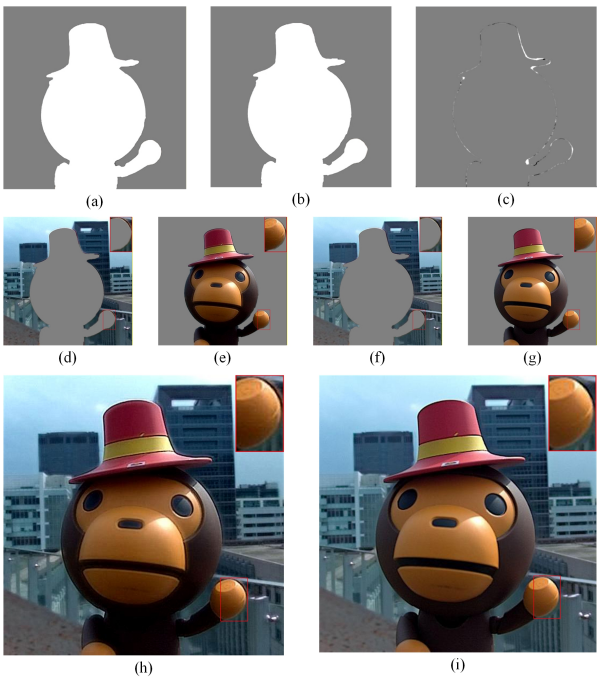


Fig. 7 Visual comparison on the "Monkey" image. (a, b) Decision maps from our Ro-FusionGAN and MFFGAN. (c) Their difference map. (d, e) Our extracted foreground using the soft map.(f, g) Counterparts from MFFGAN. (h, i) Final fusion results. Note the superior performance of our method (i) in eliminating artifacts, as shown in the magnified insets.

4.3 Qualitative Analysis and Visual Comparison

4.3.1 Detailed Comparison on Color Images

Beyond quantitative metrics, a qualitative assessment is indispensable for evaluating the perceptual fidelity of fused images, as the human visual system is highly sensitive to subtle artifacts and unnatural details. In this section, we present a detailed visual comparison of our Ro-FusionGAN against a wide array of SOTA methods.

Fig.8 demonstrates cases the fusion results from various algorithms on the challenging "Volleyball" image from the Lytro dataset. This scene is particularly difficult due to the fine, semi-transparent structure of the chain-link fence in the foreground against a complex background.

A close examination of the results reveals the distinct advantages of our approach. Many competing methods, including traditional algorithms like GFF[31] (a) and CSR[35] (e), as well as deep learning-based models like CNN[8] (f) and U2Fusion[38] (h), struggle with this scene. Their outputs exhibit significant artifacts, such as ghosting and defocus spread, where the blur from the background bleeds into the foreground fence. This is particularly evident in the magnified insets, where the clarity of the foreground object is compromised.

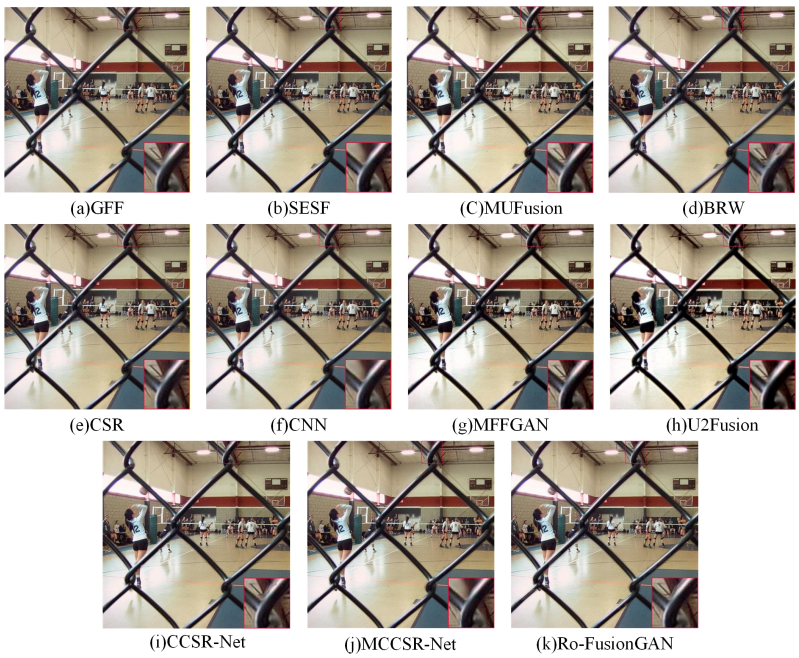


Fig. 8 Qualitative comparison on the "Volleyball" image. Fusion results from SOTA methods (a-j) are compared with our Ro-FusionGAN (k). As shown in the magnified insets, competing methods suffer from noticeable ghosting and blur artifacts around the fence structure. In contrast, our method yields a sharp, artifact-free result, validating the effectiveness of the proposed result-oriented training strategy.

In stark contrast, our Ro-FusionGAN (k) produces a visually superior result. The chain-link fence is rendered with exceptional sharpness and clarity, and the transition to the background is seamless and natural. There are no discernible ghosting or halo artifacts. This superior performance stems directly from our result-oriented training paradigm. Unlike methods that aim to produce a rigid binary mask, our fusion-aware loss function guides the generator to learn a nuanced, soft decision map that optimally preserves high-frequency details while ensuring a smooth integration of different focal planes.

4.4 Analysis of the Soft Fusion Boundary

Fig.9 displays the grayscale difference maps for the same "Volleyball" scene. This visualization provides compelling evidence of our model's superiority. The difference maps of most other methods (a-j) are plagued with strong residual structures corresponding to the foreground fence. These bright outlines clearly indicate that significant ghosting and artifacts are present in their final fused images.

The difference map generated by our Ro-FusionGAN (k), however, is remarkably clean. The residual energy from the foreground fence is almost entirely suppressed, resulting in a much darker and more uniform map. This powerfully demonstrates that our model has successfully extracted and integrated the focused regions from

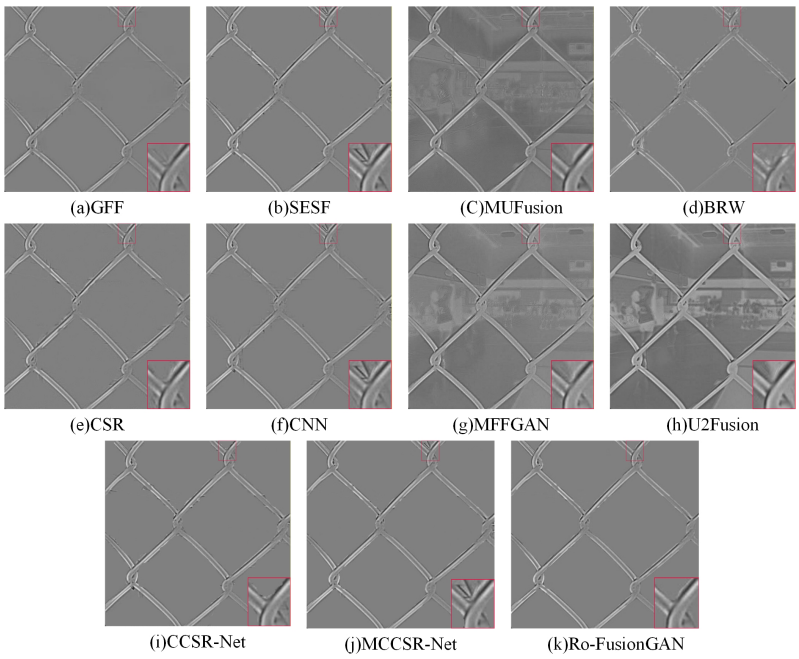


Fig. 9 Difference maps for the "Volleyball" fusion, visualizing error residuals. The prominent bright structures in (a-j) indicate ghosting artifacts, whereas the remarkably clean map from Ro-FusionGAN (k) confirms its superior capability in artifact suppression.



Fig. 10 (a)–(e) are focus maps with varying degrees of edge diffusion or contraction in which (c) is the original focus map generated by our method.

both source images with minimal interference and artifact generation. This innovative analysis technique provides a clear, quantitative-like visual tool to corroborate the perceptual quality observed in the color fusion results.

A key hypothesis in MFIF is that the optimal decision boundary for fusion is not a sharp line, but rather a transitional zone that smoothly blends focused and defocused regions to be imperceptible to the HVS. Traditional methods often approximate this by manually expanding or contracting a challenging binary mask. Our Ro-FusionGAN, however, is designed to learn this optimal boundary profile end-to-end. In this section, we conduct an experiment to validate the superiority of our learned soft boundary.

To establish a baseline for comparison, we take the binary ground-truth focus map of the "Monkey" image (Fig.10c, labeled "origin") and apply morphological erosion



Fig. 11 Final fusion results produced by Ro-FusionGAN using the autonomously learned soft decision map. The visual comparison demonstrates the superior performance of our end-to-end strategy over methods relying on heuristic boundary maps.

and dilation operations to it. Erosion contracts the foreground region by a specified number of pixels (Fig.10(a), -8 pixels; Fig.10(b), -2 pixels), simulating a decision map that under-segments the focused object. Dilation expands the foreground region (Fig.10(d), +2 pixels; Fig.10(e), +8 pixels), simulating a map that over-segments it. These five challenging-boundary maps represent different heuristic choices for the fusion boundary.To verify the generalization ability of the proposed method, we demonstrate the fusion effects of multiple sets of typical multi-focus source images, as illustrated in Fig.11.

Visual comparison of fusion results using these heuristic boundaries reveals a significant performance trade-off. Contracted maps (a, b) lead to "edge-eating" artifacts, where focused object details are truncated. Conversely, expanded maps (d, e) induce noticeable halos by erroneously incorporating defocused background content. In contrast, Ro-FusionGAN autonomously learns a nuanced transitional boundary that transcends simple erosion or dilation. Guided by the fusion-aware loss, our framework optimizes the boundary profile on a pixel-wise basis. This learned soft

transition preserves foreground integrity and ensures seamless, artifact-free integration with the background, demonstrating that direct boundary learning is superior to spatially-uniform morphological operations.

4.5 Computational Efficiency

Table 3 compares the average inference time per image pair on the Lytro dataset. We standardized the environment: traditional methods and CNN ran on an Intel Core i9 CPU, while other DL-based methods ran on an NVIDIA RTX 4090 GPU.

Table 3 The average running time of different algorithms

Method	**Time (s)**	**Method**	**Time (s)**	**Method**	**Time (s)**
BRW [34]	0.2694	CSR [35]	121.77	GFF [31]	0.3318
SESF [32]	0.5087	MUFusion [33]	1.2276	CNN [8]	57.86
MFFGAN [37]	0.1579	U2Fusion [38]	4.302	CCSR-Net [29]	0.2088
MCCSR-Net [36]	0.4336	**Ro-FusionGAN (ours)**	**0.4175**		

Results show a clear efficiency gap. Optimization-based methods like CSR are computationally expensive (121s). While lightweight models like MFFGAN are slightly faster, Ro-FusionGAN achieves a highly competitive speed of 0.4175s. Compared to CPU-bound architectures like CNN (57s), our method offers a vastly superior trade-off between perceptual quality and processing speed, making it suitable for real-time applications.

4.6 Ablation experiments

To verify the contribution of each loss component in our framework, we conducted an ablation study on the Lytro dataset. We designed three variants to compare with the full Ro-FusionGAN:

Model I: Trained solely with Adversarial loss ($\mathcal{L}_{adv}$).

Model II: Adds Fusion L_1 loss ($\mathcal{L}_{adv} + \mathcal{L}_{fusion_L1}$).

Model III: Further adds SSIM loss ($\mathcal{L}_{adv} + \mathcal{L}_{fusion_L1} + \mathcal{L}_{fusion_SSIM}$).

Ro-FusionGAN: The full model including Total Variation loss. The sample-wise quantitative comparisons are visualized in Fig.12

As shown in Fig.12, the baseline (Model I) exhibits the most unstable performance, particularly in structure-related metrics like Q_Y and Q_{ABF}, confirming that adversarial training alone is insufficient for pixel-level fidelity.

The progressive addition of $\mathcal{L}_{fusion_L1}$ (Model II) and $\mathcal{L}_{fusion_SSIM}$ (Model III) leads to consistent improvements. L_1 loss stabilizes the training by enforcing intensity consistency, while SSIM loss enhances structural preservation, as evidenced by the upward trend in Q_{NCIE} and Q_C.

However, the most significant gain is achieved by our full Ro-FusionGAN. By integrating $\mathcal{L}_{tv}$, the model effectively suppresses high-frequency noise in the decision map. This is highlighted in Fig.12(g) (Q_Y), where the full model maintains high stability

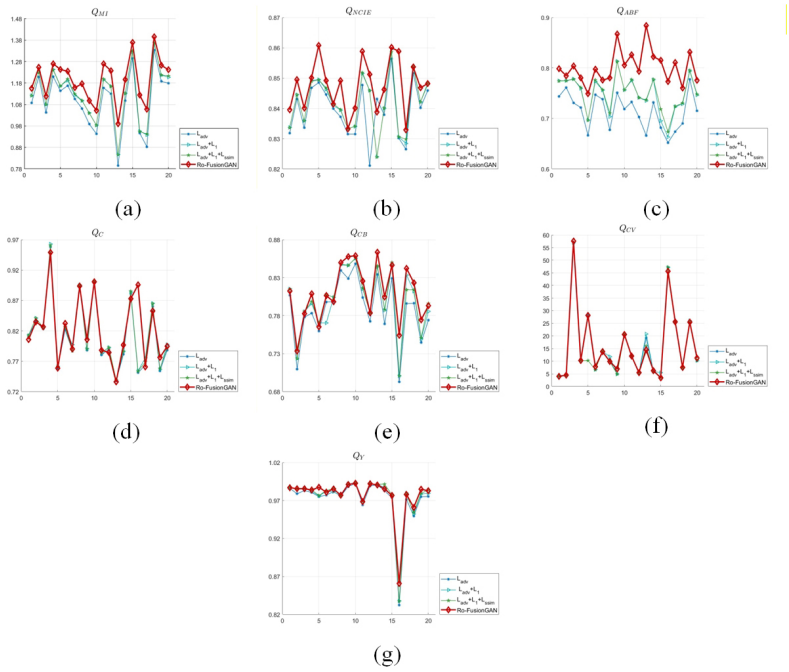


Fig. 12 Score lines of ablation experiments based on Lytro with regard to each metric.

even on challenging samples where other variants fail significantly. Overall, the red curves consistently demonstrate superior scores across the majority of metrics, validating that the combination of all loss terms is essential for generating artifact-free, high-quality fused images.

5 Conclusion

This paper introduces Ro-FusionGAN, an end-to-end adversarial framework for MFIF that shifts from the traditional "process-imitation" paradigm to a "result-oriented" approach. By employing a composite fusion-aware loss—integrating adversarial, L_1, SSIM, and TV terms—the generator autonomously learns a spatially coherent soft decision map. This eliminates the need for heuristic post-processing and ensures artifact-free fusion with natural focal transitions. Extensive experiments on public datasets demonstrate that Ro-FusionGAN achieves state-of-the-art performance in both visual and objective evaluations. Future work will focus on automated hyper-parameter optimization and extending this framework to multi-modal tasks, such as infrared-visible and medical image fusion.

Declarations

Funding. This research was funded by the National Natural Science Foundation of China under Grant No. 62205342, and Fund of Robot Technology Used for Special Environment Key Laboratory of Sichuan Province Grant No. 22kftk03.

Competing interests. The authors declare no competing interests.

Data Availability. No datasets were generated or analysed during the current study

References

[1] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4), 640–651 (2015)

[2] Li, Q., Zhuang, Y., Huai, J.: Multi-sensor fusion for robust localization with moving object segmentation in complex dynamic 3d scenes. International Journal of Applied Earth Observation and Geoinformation **124**, 103507 (2023) https://doi.org/10.1016/j.jag.2023.103507

[3] Lyu, F., Liu, X.-Y., Chen, L., Li, C., Zhou, J., Zhou, H.-L.: Low-amplitude structure recognition method based on non-subsampled contourlet transform. Petroleum Science **21**(5), 3062–3078 (2024) https://doi.org/10.1016/j.petsci.2024.03.024

[4] Hu, Y., Chen, Z., Zhang, B., Ma, L., Li, J.: A multi-focus image fusion method based on multi-source joint layering and convolutional sparse representation. IET Image Processing **16**(1), 216–228 (2022)

[5] Ma, H., Liao, Q., Zhang, J., Liu, S., Xue, J.-H.: An α-matte boundary defocus model-based cascaded network for multi-focus image fusion. IEEE Transactions on Image Processing **29**, 8668–8679 (2020) https://doi.org/10.1109/TIP.2020.3018261

[6] Dolinar, S., Divsalar, D., Pollara, F.: Code performance as a function of block size. Telecommunications and Mission Operations Progress Report **133** (1998)

[7] Wang, C., Yan, K., Zang, Y., Zhou, D., Nie, R.: Focus-aware and deep restoration network with transformer for multi-focus image fusion. Digital Signal Processing **149**, 104473 (2024) https://doi.org/10.1016/j.dsp.2024.104473

[8] Liu, Y., Chen, X., Peng, H., Wang, Z.: Multi-focus image fusion with a deep convolutional neural network. Information Fusion **36**, 191–207 (2017) https://doi.org/10.1016/j.inffus.2016.12.001

[9] Zhai, H., Zhang, G., Zeng, Z., Xu, Z., Fang, A.: Lskn-mfif: Large selective kernel network for multi-focus image fusion. Neurocomputing **635**, 129984 (2025) https://doi.org/10.1016/j.neucom.2025.129984

[10] Li, J., Li, X., Li, X., Han, D., Tan, H., Hou, Z., Yi, P.: Multi-focus image fusion based on multiscale fuzzy quality assessment. Digital Signal Processing **153**, 104592 (2024) https://doi.org/10.1016/j.dsp.2024.104592

[11] Pan, X., Zhai, H., Yang, Y., Chen, L., Li, A.: Improving multi-focus image fusion through noisy image and feature difference network. Image and Vision Computing **142**, 104891 (2024) https://doi.org/10.1016/j.imavis.2023.104891

[12] Zhai, H., Chen, Y., Wang, Y., Ouyang, Y., Zeng, Z.: W-shaped network combined with dual transformers and edge protection for multi-focus image fusion. Image and Vision Computing **150**, 105210 (2024) https://doi.org/10.1016/j.imavis.2024.105210

[13] Li, X., Chen, H., Li, Y., Peng, Y.: Multi-focus image fusion via adaptive fractional differential and guided filtering. Multimedia Tools & Applications **83**(11) (2024)

[14] Chen, P., Jiang, J., Li, L., Yao, J.: A defocus and similarity attention-based cascaded network for multi-focus and misaligned image fusion. Information Fusion **103**, 102125 (2024) https://doi.org/10.1016/j.inffus.2023.102125

[15] Liu, Y., Chen, X., Wang, Z., Wang, Z.J., Ward, R.K., Wang, X.: Deep learning for pixel-level image fusion: Recent advances and future prospects. Information Fusion **42**, 158–173 (2018) https://doi.org/10.1016/j.inffus.2017.10.007

[16] Zhao, Q.-J., Chen, X., Ma, Y.-Y., Zhao, G.-Q.: Investigations of synthetic jet control effects on helicopter rotor in forward flight based on the cfd method. The Aeronautical Journal **122**(1253), 1102–1122 (2018)

[17] Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L.: Ifcnn: A general image fusion framework based on convolutional neural network. Information Fusion **54**, 99–118 (2020) https://doi.org/10.1016/j.inffus.2019.07.011

[18] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745

[19] Krichen, M.: Generative adversarial networks. In: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–7 (2023). https://doi.org/10.1109/ICCCNT56998.2023.10306417

[20] Li, J., Guo, X., Lu, G., Zhang, B., Xu, Y., Wu, F., Zhang, D.: Drpl: Deep regression pair learning for multi-focus image fusion. IEEE Transactions on Image Processing **29**, 4816–4831 (2020) https://doi.org/10.1109/TIP.2020.2976190

[21] Zhang, H., Ma, J.: Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. International Journal of Computer Vision **129**(10), 2761–2785 (2021)

[22] Wang, Z., Li, X., Duan, H., Zhang, X.: A self-supervised residual feature learning model for multifocus image fusion. IEEE Transactions on Image Processing **31**, 4527–4542 (2022) https://doi.org/10.1109/TIP.2022.3184250

[23] Xiao, B., Xu, B., Bi, X., Li, W.: Global-feature encoding u-net (geu-net) for multi-focus image fusion. IEEE Transactions on Image Processing **30**, 163–175 (2021) https://doi.org/10.1109/TIP.2020.3033158

[24] Ma, J., Yu, W., Liang, P., Li, C., Jiang, J.: Fusiongan: A generative adversarial network for infrared and visible image fusion. Information Fusion **48**, 11–26 (2019) https://doi.org/10.1016/j.inffus.2018.09.004

[25] Ma, J., Xu, H., Jiang, J., Mei, X., Zhang, X.-P.: Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Transactions on Image Processing **29**, 4980–4995 (2020) https://doi.org/10.1109/TIP.2020.2977573

[26] Guo, X., Nie, R., Cao, J., Zhou, D., Mei, L., He, K.: Fusegan: Learning to fuse multi-focus image via conditional generative adversarial network. IEEE Transactions on Multimedia **21**(8), 1982–1996 (2019) https://doi.org/10.1109/TMM.2019.2895292

[27] Xu, H., Ma, J., Zhang, X.-P.: Mef-gan: Multi-exposure image fusion via generative adversarial networks. IEEE Transactions on Image Processing **29**, 7203–7216 (2020) https://doi.org/10.1109/TIP.2020.2999855

[28] Yang, Z., Chen, Y., Le, Z., Fan, F., Pan, E.: Multi-source medical image fusion based on wasserstein generative adversarial networks. IEEE Access **7**, 175947–175958 (2019) https://doi.org/10.1109/ACCESS.2019.2955382

[29] Zheng, K., Cheng, J., Liu, Y.: Ccsr-net: Unfolding coupled convolutional sparse representation for multi-focus image fusion. In: Liu, Q., Wang, H., Ma, Z., Zheng, W., Zha, H., Chen, X., Wang, L., Ji, R. (eds.) Pattern Recognition and Computer Vision, pp. 285–297. Springer, Singapore (2024)

[30] He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence **37**(9), 1904–1916 (2015)

[31] Li, S., Kang, X., Hu, J.: Image fusion with guided filtering. IEEE Transactions on Image processing **22**(7), 2864–2875 (2013)

[32] Ma, B., Zhu, Y., Yin, X., Ban, X., Huang, H., Mukeshimana, M.: Sesf-fuse: An

unsupervised deep model for multi-focus image fusion. Neural Computing and Applications **33**(11), 5793–5804 (2021)

[33] Cheng, C., Xu, T., Wu, X.-J.: Mufusion: A general unsupervised image fusion network based on memory unit. Information Fusion **92**, 80–92 (2023) https://doi.org/10.1016/j.inffus.2022.11.010

[34] Ma, J., Zhou, Z., Wang, B., Miao, L., Zong, H.: Multi-focus image fusion using boosted random walks-based algorithm with two-scale focus maps. Neurocomputing **335**, 9–20 (2019) https://doi.org/10.1016/j.neucom.2019.01.048

[35] Liu, Y., Chen, X., Ward, R.K., Wang, Z.J.: Image fusion with convolutional sparse representation. IEEE signal processing letters **23**(12), 1882–1886 (2016)

[36] Zheng, K., Cheng, J., Liu, Y.: Unfolding coupled convolutional sparse representation for multi-focus image fusion. Information Fusion **118**, 102974 (2025)

[37] Zhang, H., Le, Z., Shao, Z., Xu, H., Ma, J.: Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. Information Fusion **66**, 40–53 (2021) https://doi.org/10.1016/j.inffus.2020.08.022

[38] Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: A unified unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(1), 502–518 (2022) https://doi.org/10.1109/TPAMI.2020.3012548

[39] Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (2015)

[40] Qu, G., Zhang, D., Yan, P.: Information measure for performance of image fusion. Electronics Letters **38**(7), 313–315 (2002)

[41] Petrovic, V., Xydeas, C.: Objective image fusion performance characterisation. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, vol. 2, pp. 1866–18712 (2005). https://doi.org/10.1109/ICCV.2005.175

[42] Wang, Q., Shen, Y., Zhang, J.Q.: A nonlinear correlation measure for multivariable data set. Physica D: Nonlinear Phenomena **200**(3), 287–295 (2005) https://doi.org/10.1016/j.physd.2004.11.001

[43] Cui, G., Feng, H., Xu, Z., Li, Q., Chen, Y.: Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. Optics Communications **341**, 199–209 (2015) https://doi.org/10.1016/j.optcom.2014.12.032

[44] Chen, Y., Blum, R.S.: A new automated quality assessment algorithm for image

fusion. Image and Vision Computing **27**(10), 1421–1432 (2009) https://doi.org/10.1016/j.imavis.2007.12.002 . Special Section: Computer Vision Methods for Ambient Intelligence

[45] Cvejic, N., Loza, A., Bull, D., Canagarajah, N.: A similarity metric for assessment of image fusion algorithms. International journal of signal processing **2**(3), 178–182 (2005)

[46] Chen, H., Varshney, P.K.: A human perception inspired quality metric for image fusion based on regional information. Information Fusion **8**(2), 193–207 (2007) https://doi.org/10.1016/j.inffus.2005.10.001 . Special Issue on Image Fusion: Advances in the State of the Art