URFusion: Unsupervised Unified Degradation-Robust Image Fusion Network

Han Xu, *Member, IEEE*, Xunpeng Yi, Chen Lu, Guangcan Liu, *Senior Member, IEEE*, and Jiayi Ma, *Senior Member, IEEE*

***Abstract*—When dealing with low-quality source images, existing image fusion methods either fail to handle degradations or are restricted to specific degradations. This study proposes an unsupervised unified degradation-robust image fusion network, termed as URFusion, in which various types of degradations can be uniformly eliminated during the fusion process, leading to high-quality fused images. URFusion is composed of three core modules: intrinsic content extraction, intrinsic content fusion, and appearance representation learning and assignment. It first extracts degradation-free intrinsic content features from images affected by various degradations. These content features then provide feature-level rather than image-level fusion constraints for optimizing the fusion network, effectively eliminating degradation residues and reliance on ground truth. Finally, URFusion learns the appearance representation of images and assigns the statistical appearance representation of high-quality images to the content-fused result, producing the final high-quality fused image. Extensive experiments on multi-exposure image fusion and multi-modal image fusion tasks demonstrate the advantages of URFusion in fusion performance and suppression of multiple types of degradations. The code is available at https://github.com/hanna-xu/URFusion**

***Index Terms*—Image fusion, degradation, intrinsic content, appearance representation.**

I. Introduction

With the advancement of sensor technologies, different imaging modalities or shooting settings help provide unique advantages but suffer from modality- or setting-related limitations. To compensate for the limitations of single modality or setting, image fusion merges the complementary information in multi-modal source images (*e.g.*, visible and infrared images) or multiple digital photographic source images (*e.g.*, multi-exposure images). The fused image can exhibit more comprehensive and accurate scene representation, promoting human visual system and machine interpretation. It is essential in various applications, *e.g.*, semantic segmentation [1], remote sensing [2], medical diagnosis [3], vehicle detection [4].

In recent years, growing amount and diversity image fusion approaches step into a deep-learning era, encompassing introduced a broad spectrum of networks or loss functions for improving visual enhancement [5]. According to network architecture, these methods can be categorized into methods based on convolutional neural network (CNN) [6], [7], generative adversarial networks (GANs) [8], [9], Transformer [10], [11], Mamba [12], and diffusion model [13], [14]. Despite continuous evolution of network architecture, the loss functions (critical factor guiding network optimization) are all image-level constraints. Specifically, due to the lack or scarcity of ground truth in image fusion, existing loss functions constrain the similarity between the fused image and source images, *e.g.*, similarities based on intensity, gradient, or structures. However, these image-level critical information are based on human priors. *When source images suffer from degradations (e.g., inappropriate illumination, contrast, noise), the degradations will be mistakenly identified as critical information and retained in fused images.* As illustrated in Fig. 1 (a), based on the image-level constraint, the fused image a pair of low-light visible and infrared images still suffers from low illumination.

To avoid degradation residues, some methods couple the suppression for specific types of degradations during the fusion process. As illustrated in Fig. 1 (b), some methods incorporate additional degradation-related constraints with the basic image-level constraints. For instance, some methods additionally define illumination-related loss functions to alleviate degradation of low light [15], [16], [17]. *However, these constraints are customized for specific types of degradation and therefore unsuitable for handling multiple types of degradations in real conditions.* Other methods suppress degradation residues through pseudo high-quality supervision. They introduce artificial degradations into source images and construct pseudo high-quality source images for image-level constraint. The artificial paired data prompts the fusion network to learn the mapping from low-quality to high-quality data [18]. However, *due to the limitations posed by artificially constructed limited degradations and high-quality data, the applicability to diverse degradations remains considerably restricted.*

To address these issues, we propose an unsupervised unified degradation-robust image fusion network, termed as URFusion. We assume that various degradations can be categorized into two types: content-related and appearance representation-related degradations. Based on the assumption, a high-quality

Received 25 December 2024; revised 23 May 2025 and 8 August 2025; accepted 31 August 2025. Date of publication 16 September 2025; date of current version 19 September 2025. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U21B2027, in part by the Natural Science Foundation of Jiangsu Province of China under Grant BK20241280, and in part by the Start-Up Research Fund of Southeast University under Grant RF1028623006 and Grant RF1028624061. The associate editor coordinating the review of this article and approving it for publication was Prof. Yipeng Liu. *(Corresponding authors: Guangcan Liu; Jiayi Ma.)*

Han Xu, Chen Lu, and Guangcan Liu are with the School of Automation, Southeast University, Nanjing 210096, China (e-mail: xu_han@seu.edu.cn; 230239505@seu.edu.cn; guangcanliu@seu.edu.cn).

Xunpeng Yi and Jiayi Ma are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: yixunpeng@whu.edu.cn; jyma2010@gmail.com).

Digital Object Identifier 10.1109/TIP.2025.3607628

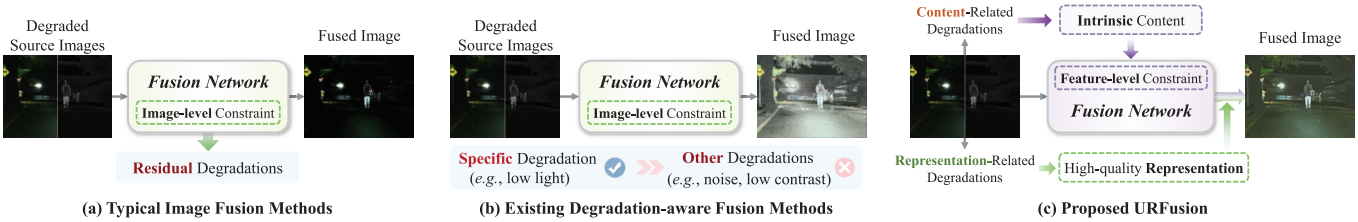


Fig. 1. Schematic illustration of existing typical image fusion methods, degradation-aware image fusion methods and the proposed unsupervised unified degradation-robust image fusion network (URFusion) in dealing with multiple types of degraded source images. The fused images in (a) and (b) are generated by SwinFusion [10] and DIVFusion [15], repsectively.

fused image should be composed of two key components: i) intrinsic scene content that is independent of degradations and representations, and ii) visually satisfactory appearance representation. To achieve this, URFusion consists of three core modules: intrinsic content extraction, intrinsic content fusion, and appearance representation learning and assignment. The intrinsic content extraction extracts degradation-free intrinsic content features from source images affected by various degradations. Then, the feature-level constraints are defined based on the content features to optimize the fusion network. As the content features are in feature level and independent of degradations, the fusion network guided by the features can effectively eliminates degradation residues and the reliance on ground truth. It results in a unified content fusion model robust to multiple types of degradations. Finally, we learn the appearance representation of images and assign the statistical appearance representation of high-quality images to the content-fused result, producing a high-quality fused image. The contributions of URFusion are summarized as:

1) We propose, for the first time, an unsupervised unified degradation-robust image fusion network URFusion. It leverages a unified degradation-robust fusion network for different fusion tasks, which is capable of fusing images affected by various degradations is designed to generate high-quality fused images.
2) We design a feature-level fusion constraint by learning degradation-free intrinsic scene features from degraded source images. The feature-level constraint can effectively eliminate the degradation residues or avoid the reliance on ground truth in existing image-level constraints.
3) We design an appearance representation learning method that, in conjunction with scene content, forms the visual representation of fused images. By leveraging the characteristics from a set of high-quality images, we assign the statistical representation to content-fused results, consistently generating high-quality fused images.
4) We conduct the experiments on several datasets of multi-exposure and multi-modal images with various degradations including inappropriate brightness, contrast, saturation, and noise. Qualitative and quantitative results validate the effectiveness and generalization of URFusion.

II. Related Work

A. Typical Image Fusion Methods

Typical fusion methods include traditional and learning-based methods.In traditional types, methods based on multi-scale transform separate source images into different frequency bands and employ hierarchical fusion strategies for recombination [19]. In sparse representation-based methods [20], [21], image blocks are sparsely decoded by an overcomplete dictionary. The fusion criterion then preserves salient features and suppress redundant components. Finally, coefficients or representations are reconstructed to produce fused images.

Learning-based methods can be divided into CNN, GAN, Transformer, Mamba, and diffusion model-based methods. CNN-based methods apply CNN for feature extraction and reconstruction [22], [23] or end-to-end networks [24]. GAN-based methods use adversarial relationship to push fused images closer to source images from probability distribution [25], [26]. Transformer devises a self-attention mechanism to capture global interactions, compensating for the limited receptive field of CNN [27], [28]. As Transformers are hindered by quadratic complexity, some methods introduce Mamba to enable global awareness with linear complexity [12], [29]. Some methods integrate the advantages of these network structures. For instance, CDDFuse [30] decomposes modal-specific and shared features where Restormer, Transformer-CNN, and invertible neural networks blocks handle cross-modal shallow, low-frequency global, and high-frequency local features. Some methods introduce diffusion models to leverage the superior feature extraction [13] and generation capabilities [14], [31]. To address the absence of ground truth in diffusion-based frameworks, Diff-IF [14] introduces priors with targeted distribution search for specific fusion task. VDMUFusion [31] reformulates image fusion as a weighted averaging paradigm under multi-task learning, establishing noise assumptions and enabling simultaneous prediction of noise and fusion weights. Current methods predominantly address aligned image fusion while practical challenges arise from image mismatches [32]. Some approaches [33], [34], [35], [36], [37] pioneer joint registration-fusion frameworks to tackle unaligned scenarios, eliminating preprocessing dependencies while enhancing robustness.

Although the aforementioned methods introduce various innovations in the design of network architectures, the progress in designing optimization objectives, *i.e.*, loss functions, remains relatively limited. Existing methods generally rely on image-level constraints to preserve source information, such as intensity distribution, gradient, structures, *etc*. The motivation of image-level constraints lies in the prior assumption that critical source information can be represented through aforementioned forms. However, source images in real scenarios may be degraded due to environmental factors or equipment limitations. *The degradations are also manifest in these image-level forms as critical information. Thus, image-level similarity*

constraints cannot effectively distinguish complex degradations, leading to residual degradations in fused images.

B. Image Restoration Methods

Recent advances in image restoration witness a paradigm shift toward model-driven learning frameworks that explicitly integrate priors and mathematical interpretability. By splitting modal-shared common and modal-specific unique information with convolutional sparse coding blocks, CU-Net [38] proposes a novel multi-modal convolutional sparse coding model for multi-modal image restoration or fusion. Yao et al. [39] proposes an all-in-one image restoration network that handles multiple degradations through a neural degradation representation capturing underlying degradation characteristics and degradation query and injection modules. For flash guided non-flash image denoising, LGCNet [40] reveals the distribution of flash and non-flash modality gaps and designs a Laplacian gradient consistency model. Each network component strictly adheres to mathematical formulation. For blind image restoration, DeepSN-Net [41] formulates an improved second-order semi-smooth Newton algorithm to enable network implementation, and for the first time designs an innovative second-order deep unfolding network aligned with this algorithm.

C. Degradation-Aware Image Fusion Methods

Some methods notice the degradations in source images and inject suppression for specific types of degradations into the method. These methods mainly consider the illumination degradation, and design illumination-related auxiliary loss functions. PIAFusion [16] designs the image-level constraint, which are weighted through illumination-aware weights computed by illumination probability of source images. Also aimed at illumination awareness, DIVFusion [15] defines enhancement-fusion loss functions. The pre-enhancement of low-light visible images is based on Retinex theory to provide optimization guidance. DDBF [17] constructs adversarial relationship between the visible and infrared intensity distributions. *In summary, these methods aim to suppress illumination degradation with auxiliary illumination-related constraints, which cannot cope with other types of degradations.*

Some methods solve degradations via pseudo supervision, *i.e.*, artificially constructing degraded or high-quality source images, and then constraining similarity between fused image and high-quality source images. Text-IF [18] artificially constructs degraded or high-quality source images and uses text to preserve high-quality information. Also with high-quality source images, DRMF [42] leverages diffusion models where degradation-robust conditional diffusion models for different modalities are trained by paired degraded and high-quality data. Then, a priori module integrates uni-modal priors to enable fusion. OmniFuse [43] addresses composite degradations via a language-guided latent diffusion model. It integrates a semantics-aware fusion strategy to aggregate multi-modal features, leveraging visual-semantic constraints to refine localization priors. *Due to the fixed limitations of artificially constructed mapping and high-quality data or degradations, the performances for various degradations are still limited.*

The proposed method extracts degradation-free intrinsic content features from images affected by various degradations. The content features provide feature-level fusion constraints for optimizing the fusion network, effectively eliminating degradation residues and reliance on ground truth. Then, we learn the appearance representation of images and assign a statistical appearance representation of high-quality images to content-fused results, producing high-quality fused images.

III. Proposed Method

A. Problem Formulation

URFusion aims to remove various degradations in source images during the fusion process and generate a high-quality fused image. We first conduct an investigation into the manifestation of various degradations, identifying that they can be classified into two distinct categories. One pertains to scene content, represented as various types of noise. The and the other one pertains to appearance representation, *e.g.*, illumination, contrast, and saturation. Thus, the fusion process in URFusion consists of two aspects: the fusion of intrinsic scene contents and the assignment of appropriate appearance representation to fused content. By learning the intrinsic content and appearance representation, an unsupervised unified fusion method robust to various degradations can be designed. The overall framework of URFusion is shown in Fig. 2.

First, an intrinsic content encoder $\mathcal{C}$ extracts intrinsic content features from source images, which are free from both content and representation degradations. Mathematically, this process can be represented as $C_x = \mathcal{C}(I_x), C_y = \mathcal{C}(I_y)$, where I_x and I_y are source images, C_x and C_y are intrinsic content features.

Second, a content fusion network $\mathcal{F}$ takes source images as input and the output $I_f^C = \mathcal{F}(I_x, I_y)$ contains information in intrinsic content features C_x and C_y. As C_x and C_y are independent of appearance representation, $\mathcal{F}$ maps source images with various representations into a fixed domain.

On this basis, an appearance representation network $\mathcal{A}$ learns to extract the appearance representation from an image, *i.e.*, $A = \mathcal{A}(I)$, where A is the representation of the image I. As $\mathcal{A}$ has been trained, we can obtain a statistical satisfactory representation v_F with a set of high-quality images. v_F is then assigned to the fused scene content I_f^C, and generate the final high-quality fused image as $I_F = \mathcal{M}(I_f^C, v_f)$, where $\mathcal{M}$ denotes the representation modulation or assignment.

B. Intrinsic Content Encoder

For a low-quality image I_d which may suffer from various degradations, the corresponding high-quality image is not always available. To avoid the reliance on supervision, we artificially apply two different random degradations to I_d. The degradations are randomly selected from the degradation sets including various random noise, illumination adjustment, contrast adjustment, and also the original low-quality image. Then, two degraded images of I_d, denoted as I_d^1 and I_d^2, can be obtained. Due to the random and diverse nature of these degradations, the essential scene content in I_d^1 and I_d^2 is the invariant intrinsic content information.

An intrinsic content encoder $\mathcal{C}$ in Fig. 3 is designed to extract intrinsic contents in I_d^1 and I_d^2, *i.e.*, $C_d^1 = \mathcal{C}(I_d^1), C_d^2 = \mathcal{C}(I_d^2)$. To get rid of the influence of image-level degradations on the characterization of content, $\mathcal{C}$ characterizes content in feature space. To equip the encoder with aforementioned capabilities, the loss function of $\mathcal{C}$ is composed of three terms.

1) Similarity Loss: The intrinsic content features extracted from I_d^1 and I_d^2 should be similar. According to their degradations, we consider three conditions: i) I_d^1 and I_d^2 both suffer

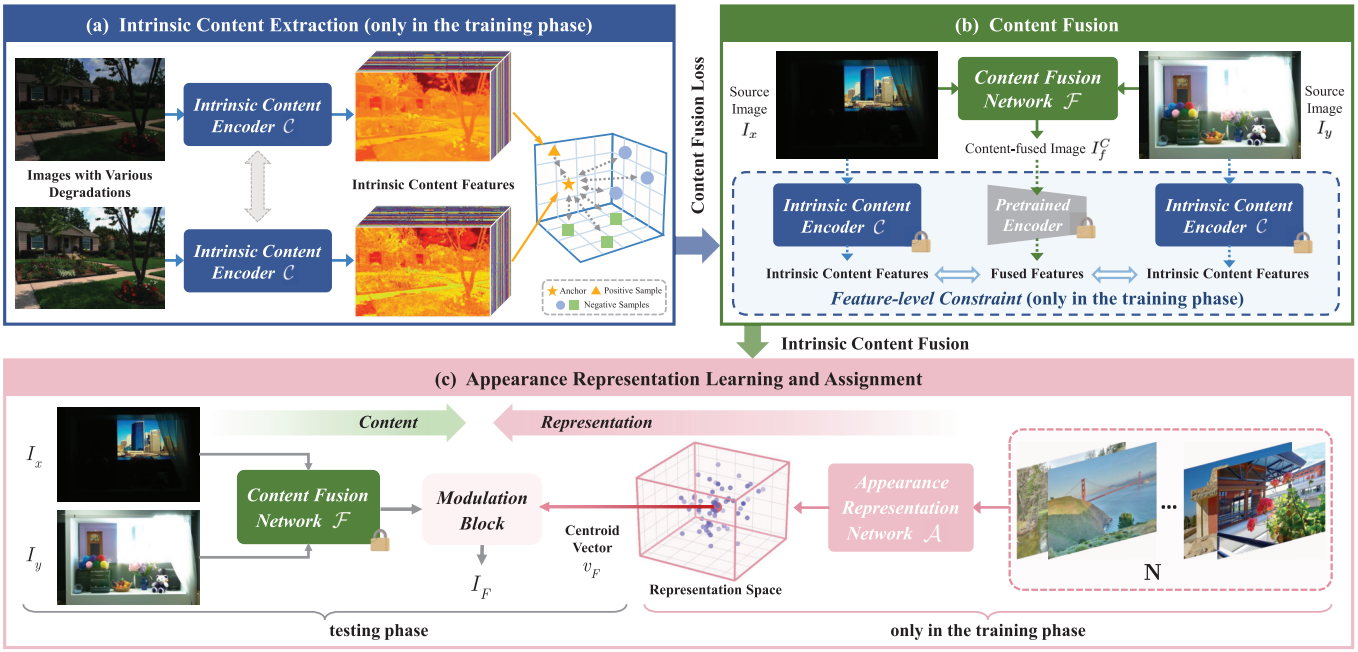


Fig. 2. Overall framework of the proposed unsupervised unified degradation-robust image fusion method.

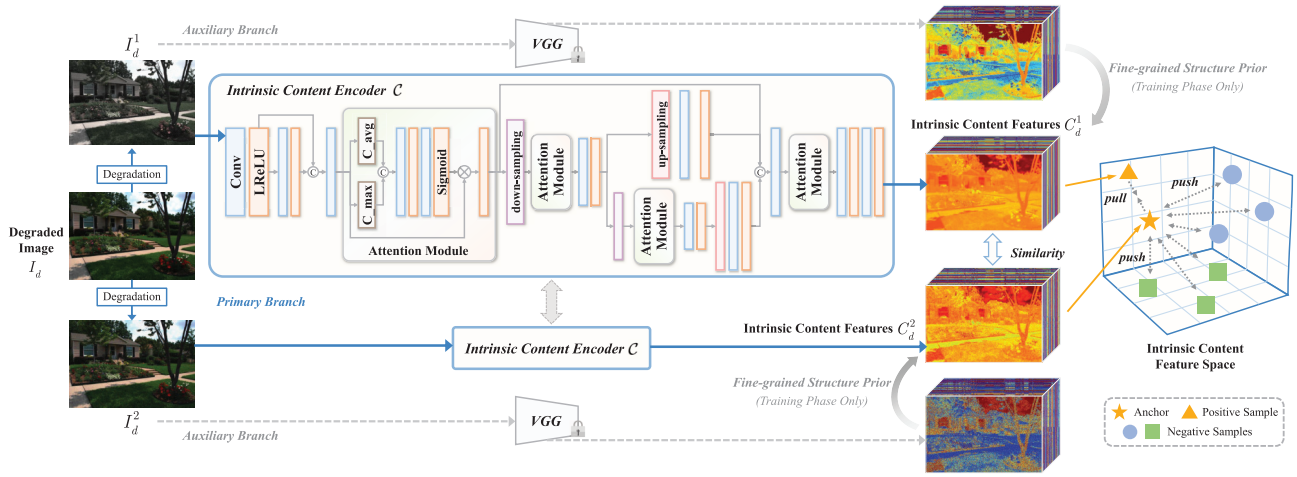


Fig. 3. Framework of intrinsic content extraction. The degraded image I_d suffers from slight noise and low light degradations. The two degradations of creating I_d^1 and I_d^2 are randomly set as low saturation and more serious noise. For all the training data, the degradations are randomly selected from more complex and comprehensive degradation sets. Conv: convolutional layer, C_max and C_avg: channel-wise max and average pooling, c: concatenation.

from degradations related to appearance representations, *e.g.*, low saturation, dark illumination, *etc.*. The mapping from them to similar content features is intuitive and comprehensible. ii) I_d^1 and I_d^2 suffer from noise and inappropriate representation, respectively. The similarity constraint between noisy and noise-independent features is beneficial for noise suppression during content extraction. iii) I_d^1 and I_d^2 both suffer from noise. Training the encoder to map a noisy content to another noisy content is equivalent to mapping it to a clean intrinsic content [44]. Thus, a similarity loss is defined as:

$$\mathcal{L}_{sim} = \|C_d^1 - C_d^2\|_2 + \omega\|\psi_2(C_d^1) - \psi_2(C_d^2)\|_2, \tag{1}$$

where $\psi_2(\cdot)$ denotes the 1/2 downsampling operator. This operation primarily serves the purpose of low-dimensional consistency enforcement through hierarchical representation, prioritizing the structure preservation and extraction robustness while attenuating the impact of detailed degradations. With explicit constraint on the low-dimensional feature consistency, it allows the encoder to enforce the geometric coherence of main scene structures in the low-dimension feature space by suppressing the high-dimensional interference. ω is the hyper-parameter to control the trade-off between two scales.

2) Contrastive Loss: As a single similarity loss will lead to trivial solutions, *i.e.*, features contain few scene-relevant information, a contrastive loss is defined. That is, degraded images of the same scene should correspond to close content features while content features of different scenes should be far apart. For an anchor feature C_d^1 extracted from I_d^1, the feature C_d^2 extracted from the same scene is the positive sample. C_d^1 and C_d^2 should be pulled in the feature space. The content

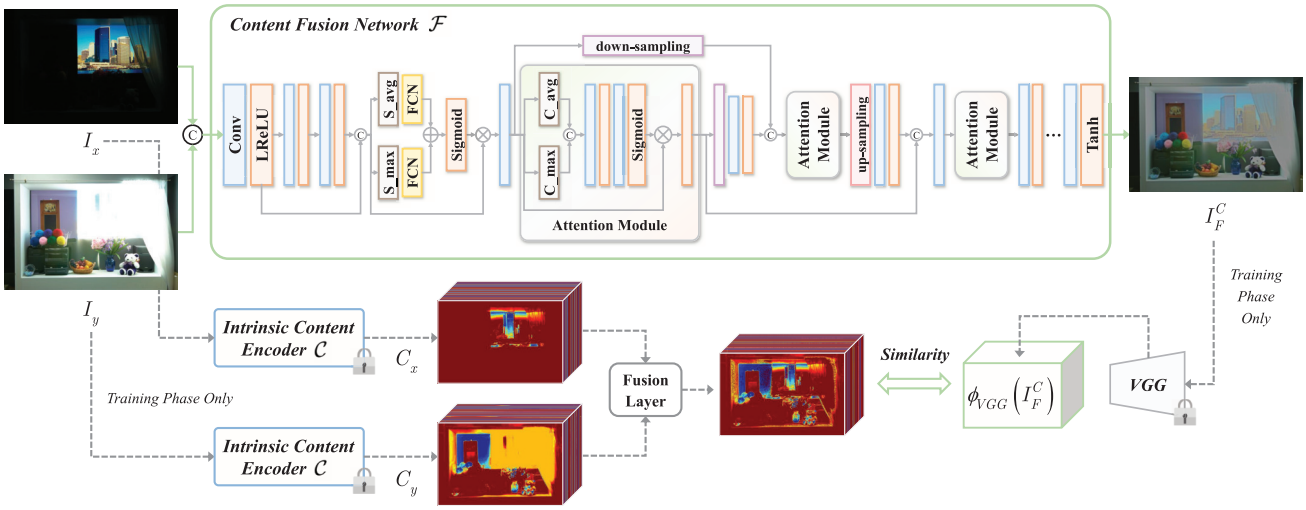


Fig. 4. Framework of the intrinsic content fusion. Conv: convolutional layer, S_max and S_avg: spatial-wise global max and average pooling, C_max and C_avg: channel-wise max and average pooling, FCN: fully connected layer, c: concatenation.

features extracted from other scenes are negative samples as:

$$C_z = \mathcal{C}(I_z),\ z \in \{1, \cdots, Z\}, \tag{2}$$

where $\{I_z\}_{z=1}^{Z}$ are images of different scenes. Z is the number of negative samples. C_z should be pushed away from C_d^1. Therefore, the contrastive loss is defined as:

$$\mathcal{L}_{contrast} = -\log \frac{\exp(\|C_d^1 - C_d^2\|_2/\tau)}{\sum_{z=1}^{Z} \exp(\|C_d^1 - C_z\|_2/\tau)}, \tag{3}$$

where τ is the temperature coefficient and set to 0.4.

3) Dual-Branch Loss: Pure contrastive training often pushes representations toward higher-level abstraction. To improve the fidelity of fine-grained structures in content features, we incorporate prior structure knowledge from a pre-trained model through a dual-branch architecture. The primary branch consists of trainable siamese encoders $\mathcal{C}$. The auxiliary branch includes fixed feature extraction in VGG for lightweight guidance. By incorporating pre-trained models, the encoders can also benefit from the prior knowledge of pretrained model [45]. The dual-branch loss is defined as:

$$\mathcal{L}_{dual} = \|\mathcal{C}_d^1 - \phi_{VGG}(I_d^1)\|_2 + \|\mathcal{C}_d^2 - \phi_{VGG}(I_d^2)\|_2, \tag{4}$$

where $\phi_{VGG}(\cdot)$ represents the feature map extracted by VGG-16 [46]. To avoid blurring and artifacts of details, the feature map of a shallow (*i.e.*, 2^{nd}) layer of VGG is used for guidance.

Finally, with hyper-parameters η_1 and η_2 to control the trade-off, the loss function of content encoder is defined as:

$$\mathcal{L}_{\mathcal{C}} = \eta_1 \mathcal{L}_{sim} + \eta_2 \mathcal{L}_{contrast} + \mathcal{L}_{dual}. \tag{5}$$

C. Content Fusion Network

The content fusion network $\mathcal{F}$ takes the source images $\{I_x, I_y\}$ as input and directly outputs a content-fused image $I_F^C = \mathcal{F}(I_x, I_y)$ without various appearance representation.

1) Network Architecture: The attention module in Fig. 4 selectively emphasizes informative features and suppresses useless ones by computing attention weights. Then, $\mathcal{F}$ can focus on the most content-relevant features, improving the content fusion performance. Short connections facilitate feature reuse and texture retention, mitigate vanishing gradients, enhance propagation, and promote efficient fusion.

2) Loss Function: The intrinsic content encoder $\mathcal{C}$ extracts crucial contents $\{C_x, C_y\}$. $\mathcal{F}$ then aims to integrate the information in C_x and C_y into I_F^C. To this end, C_x and C_y are firstly integrated in a fusion layer, which outputs a guidance for I_F^C in content feature space. As the ultimate goal is to present as much scene content as possible, the fusion layer measures the feature activity with gradients and perform fusion accordingly. Thus, the loss function of $\mathcal{F}$ is defined as:

$$\mathcal{L}_{\mathcal{F}} = \|\phi_{VGG}(I_F^C) - \left(M \odot C_x + (1 - M) \odot C_y\right)\|_1, \tag{6}$$

where $\odot$ denotes Hadamard product. The mapping from I_F^C to the content feature space is achieved through the original VGG rather than the encoder $\mathcal{C}$. The reason is that the encoder $\mathcal{C}$ strips away degradations, leaving residual degradations in the fused image unconstrained. In contrast, VGG preserves degradation traces in the features, enabling the alignment with the intrinsic features of source images to suppress residual degradations effectively. M is a mask reflecting the gradient relationship between features, denoted as:

$$M(i,j) = \begin{cases} 1, & \text{if } |\nabla C_x(i,j)| > \left|\nabla C_y(i,j)\right| \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where ∇ is the Laplacian operator. i, j denote pixel position.

D. Appearance Representation Network

The content fusion network preserves the intrinsic contents in source images, and outputs a content-fused image in a fixed domain. In this section, we aim to map the content-fused image to diverse image domains by leveraging different appearance representations. For this purpose, we design an appearance representation network $\mathcal{A}$ to extract the content-independent representation vector from an image. This vector is used to alter the appearance representation of a content-fused image while maintaining scene contents. The combination of the content-fused image and representation vector is mapped back to the complete image domain through a modulation block.

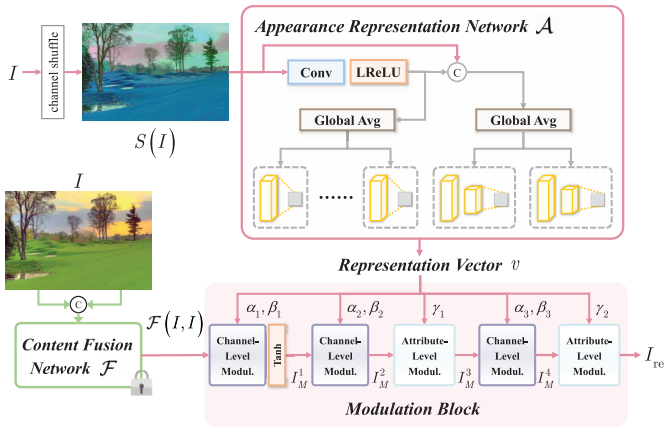


Fig. 5. Framework of appearance representation modulation.

1) Network Architecture and Modulation Block: In Fig. 5, to extract content-independent representation from a given image I_A, $\mathcal{A}$ generates the channel-level modulation parameters ($\alpha_1,\beta_1,\alpha_2,\beta_2\,\alpha_3,\beta_3 \in \mathbb{R}$) and attribute-level modulation parameters ($\gamma_1 \in \mathbb{R}$ for contrast and γ_2 for saturation).

For I_F^C which contains predetermined scene content and awaits representation modulation, the first-part modulation is:

$$I_M^1 = \text{Tanh}(r(\alpha_1)\cdot I_F^C + r(\beta_1)),\ I_M^2 = r(\alpha_2)\cdot I_M^1 + r(\beta_2), \quad (8)$$

where $r(\cdot)$ is channel replication. Contrast adjustment is on the illuminance channel (Y_M) of I_M^2 to avoid color deviation:

$$Y'_M = (Y_M - \mu_M)\cdot\gamma_1 + \mu_M, \quad (9)$$

where μ_M denotes the mean value of Y_M. Y'_M and the chrominance of I_M are concatenated and transformed to RGB space as I_M^3, which is further modulated as:

$$I_M^4 = r(\alpha_3)\cdot I_M^3 + r(\beta_3). \quad (10)$$

γ_2 for saturation adjustment is performed on three channels. A chrominance-related mask m is first generated as:

$$I_M^{ch} = c(R_M^4 - g_M^4, G_M^4 - g_M^4, B_M^4 - g_M^4),\ m = 1 - g_M^{ch}, \quad (11)$$

where g_M^4 and g_M^{ch} are gray versions of I_M^4 and I_M^{ch}. R_M^4,G_M^4,B_M^4 denote the R, G, B channels of I_M^4. $c(\cdot)$ denotes channel concatenation. The result can be obtained as:

$$R_M^5 = R_M^4 \odot (1 + \gamma_2 \cdot m). \quad (12)$$

The other channels can be obtained in the same way. Concatenation of R_M^5,G_M^5,B_M^5 is the final modulated image I_M. The overall modulation process can be summarized as:

$$I_M = \mathcal{M}\left(I_F^C, \mathcal{A}(I_A)\right), \quad (13)$$

where $\mathcal{M}$ is the modulation block and $\mathcal{A}(I_A)$ is the representation vector $v = [\alpha_1,\beta_1,\alpha_2,\beta_2,\alpha_3,\beta_3,\gamma_1,\gamma_2] \in \mathbb{R}^8$.

2) Loss Function: When combining the content of source images and the appearance representation of another image, there is no ground truth for supervision. Thus, we employ a self-supervised approach to train $\mathcal{A}$. The combination of the content-fused image and representation vector extracted of this image is expected to reconstruct the original one.

Specifically, for an image I, it is firstly concatenated with itself and fed into the content fusion network. When extracting the appearance representation, we first perform channel shuffle on I as $S(I)$, to avoid the influence of scene color on the representation vector. $S(I)$ is fed into $\mathcal{A}$ to generate the vector. Thus, the loss function of $\mathcal{A}$ constrains the similarity between the reconstructed image and the original one as:

$$\mathcal{L}_{\mathcal{A}} = \|I - I_{\text{re}}\|_2, \quad (14)$$

where $I_{\text{re}} = \mathcal{M}(\mathcal{F}(I,I), \mathcal{A}(S(I)))$.

E. Robust Fusion With Statistical Representation

This section finds an appropriate representation and assign it to the content-fused image generated by $\mathcal{F}$, and generate the final fused image I_F. The appearance network $\mathcal{A}$ can extract an appropriate representation vector from a high-quality image. However, the representation of a single image may have randomness and specificity. Thus, we extract the representations of a set of high-quality images $\mathbf{N}$. Subsequently, the statistical centroid of this set of representations are assigned to the content-fused image. This process is illustrated in Fig. 2 (c) and the centroid v_F can be represented as:

$$v_F = \frac{1}{|\mathbf{N}|}\sum_{I_n \in \mathbf{N}} (\mathcal{A}(I_n)), \quad (15)$$

where I_n is an image in $\mathbf{N}$ and $|\mathbf{N}|$ is the number of images in $\mathbf{N}$. The final high-quality fused image I_f is obtained as:

$$I_F = \mathcal{M}(\mathcal{F}(I_x, I_y), v_F), \quad (16)$$

where $\mathcal{F}(I_x, I_y)$ is the fused content. The modulation block $\mathcal{M}$ combines the content and representation information.

F. Dealing With Multi-Modal Image Fusion

This section applies URFusion to a typical multi-modal image fusion task, namely visible and infrared image fusion. Different from multi-exposure images, we need to train different intrinsic content encoders to extract intrinsic scene information related to specific modalities. On this basis, a content fusion network is also retrained to dealing with multi-modal information. More detailed, a paired of visible and infrared images are defined as I_{vis} and I_{ir}, respectively.

1) Intrinsic Content Encoder: The intrinsic content information of I_{vis} can be extracted through the abovementioned $\mathcal{C}$, redefined as $\mathcal{C}_{vis}$. For infrared images, we retrain a pseudo-siamese encoder $\mathcal{C}_{ir}$ through the design in Sec. III-B.

2) Content Fusion Network: For the extracted content information $C_{vis} = \mathcal{C}_{vis}(I_{vis})$ and $C_{ir} = \mathcal{C}_{ir}(I_{ir})$, we train a content fusion network $\mathcal{F}_m$. It follows the design in Sec. III-C to fuse the information extracted from two source images as $I_F^C = \mathcal{F}_m(I_{vis}, I_{ir})$, which uses C_{vis} and C_{ir} for constraint.

3) Appearance Representation Network: In multi-modal image fusion, we retrain an appearance representation network $\mathcal{A}_m$. Based on the fact that the human visual system is naturally attuned to interpreting visible information, we display the fused content in a form that closely resembles visible imaging. Thus, we learn a mapping from the content-fused domain to the visible domain. It leverages the strengths of infrared imaging for content enhancement while maintaining a familiar and intuitive visual representation. When training $\mathcal{A}_m$, only visible images instead of fused images are available. In this condition, I_{ir} is set to globally consistent numerical value. Then, we can obtain a statistical representation of high-quality visible images. The assigned representation and modulation are obtained or processed as Secs. III-D and III-E.

TABLE I
EXPERIMENT SETTINGS OF DIFFERENT NETWORKS

Networks	Patch Size	Batch Size	Epoch
Intrinsic Content Encoder $\mathcal{C}$	160×160	12	50
Content Fusion Network $\mathcal{F}$	160×160	12	80
Appearance Representation Network $\mathcal{A}$	248×248	8	20

IV. EXPERIMENT

A. Implementation Details

Training and testing multi-exposure data come from *SICE*.[1] Visible and infrared images are from *RoadScene*[2] and *MSRS*.[3] To validate effectiveness and robustness, we test on original and degraded images. The source images in these datasets are already aligned. Degraded images include some original low-quality images in the datasets and some manually created ones. For normalized images, the degradations are randomly introduced in the form of Gaussian noise ($\sigma \in [0.04, 0.2]$), gamma transformation ($\gamma \in [1.2, 2]$ for darkening or $[0.75, 0.95]$ for brightening), contrast adjustment (scaling factor $\alpha \in [0.5, 0.9]$ for lower contrast or $[1.2, 2]$ for higher contrast), and saturation adjustment (scaling factor $c \in [0.5, 0.9]$). To simulate complex real-world degradations, we randomly combine these degradations to model mixed degradations.

The test data includes real degraded images in datasets and artificially degraded images. Specifically, multi-exposure images in SICE inherently suffer from brightness degradations. The noisy multi-exposure images are obtained by adding Gaussian noises. The low-saturation and low-contrast multi-exposure images are created by adjusting saturation and contrast. For multi-modal images, the visible images captured at night in MSRS exhibit low light. For noisy images, we similarly add Gaussian noise these images. For low-contrast image, images in RoadScene have a certain degree of low contrast, so we partially use the original images in RoadScene, and partially perform more severe contrast degradations.

Algorithm 1 URFusion for Multi-Exposure Image Fusio

Training phase:
1. Initialize the intrinsic content encoder $\mathcal{C}$, content fusion network $\mathcal{F}$, and appearance representation network $\mathcal{A}$;
2. Update parameters of $\mathcal{C}$ by minimizing $\mathcal{L}_{\mathcal{C}}$ in Eq. (5);
3. Fix $\mathcal{C}$, and update the parameters of $\mathcal{F}$ by minimizing $\mathcal{L}_{\mathcal{F}}$ defined in Eq. (6);
4. Fix $\mathcal{F}$, and update the parameters of $\mathcal{A}$ by minimizing $\mathcal{L}_{\mathcal{A}}$ defined in Eq. (14);
5. Fix $\mathcal{A}$, and obtain the centroid representation vector v_F of a set of high-quality images according to Eq. (15).

Testing phase:
1. Feed multi-exposure images I_x, I_y into $\mathcal{F}$ to generate the content-fused image $I_F^C = \mathcal{F}(I_x, I_y)$;
2. Assign v_F to the content-fused image I_F^C to generate the final fused image I_F according to Eq. (16).

Experiment settings are reported in Tab. I. For infrared-related networks, the settings of $\mathcal{C}_{ir}$ and $\mathcal{F}_m$ are the same to Tab. I. Epoch of $\mathcal{A}_m$ is 30. The intrinsic content encoders and content fusion network face higher optimization complexity due to spatially-variant tasks. By comparison, the global representation inferred by the appearance representation network enables simpler optimization, requiring larger patch sizes and allowing fewer epochs. All networks are optimized through the Adam optimizer. Learning rate is 0.0001 with exponential decay. Hyper-parameters are set as: $\omega = 4, \eta_1 = 0.06, \eta_2 = 1e-4$. Experiments are performed on NVIDIA Geforce RTX 3090 GPU and 2.4 GHz Intel Core i5-1135G7 CPU. The overall description of URFusion is summarized as Algs. 1–2.

Algorithm 2 URFusion for Visible and Infrared Image Fusio

Training phase:
1. Initialize the pseudo-siamese intrinsic content encoders $\mathcal{C}_{vis}, \mathcal{C}_{ir}$, content fusion network $\mathcal{F}$, and appearance representation network $\mathcal{A}$;
2. Update parameters of $\mathcal{C}_{vis}$ by minimizing $\mathcal{L}_{\mathcal{C}}$ in Eq. (5) with visible images, *i.e.*, $\mathcal{C}$ in Alg. 1;
3. Update parameters of $\mathcal{C}_{ir}$ by minimizing $\mathcal{L}_{\mathcal{C}}$ in Eq. (5) with infrared images;
4. Fix $\mathcal{C}_{vis}, \mathcal{C}_{ir}$, and update the parameters of $\mathcal{F}$ by minimizing $\mathcal{L}_{\mathcal{F}}$ in Eq. (6) with paired visible and infrared images;
5. Follow steps 4-5 in Alg. 1, obtain the centroid representation vector v_F of a set of high-quality visible images.

Testing phase:
1. Feed visible and infrared images I_x, I_y into $\mathcal{F}$ to generate the content-fused image $I_F^C = \mathcal{F}(I_x, I_y)$;
2. Assign v_F to the content-fused image I_F^C to generate the final fused image I_F according to Eq. (16).

B. Multi-Exposure Image Fusion

We validate the effectiveness on typical under- and over-exposed image fusion, and multi-exposure images with various degradations (*i.e.*, degraded multi-exposure image fusion).

*1) **Typical Multi-Exposure Image Fusion**:* SOTA comparison multi-exposure image fusion methods include SDNet [47], TransMEF [48], U2Fusion [49], SwinFusion [10], MUFusion [50], SAMT-MEF [51] and HSDS_MEF [52].

Qualitative Results. We compare and analyze the fusion performances from both local and global perspectives in Fig. 6. *From the local perspective*, in the first scenario, URFusion can overcome the drawback of low brightness in some competitors by better balancing the information retention of two source images and appropriate representation assignment. *From the global perspective*, as shown in the second scenario, the results of some competitors suffer from inappropriate brightness, halo artifacts, or distorted textures. By comparison, URFusion exhibits more uniform brightness, higher contrast, and more prominent structural information.

Quantitative Results. In SICE, fused images of 13 fusion methods are generated and the best one is selected as ground truth subjectively. Thus, in typical fusion, we use metrics that measure the similarity between fused and source images. In degradation scenes, we use the artificial ground truth for evaluation. Peak signal-to-noise ratio (PSNR) quantifies the ratio of peak signal power to noise power, indicating introduced distortion. Structural similarity index measure (SSIM) evaluates the similarity from correlation, illuminance and contrast. Feature mutual information (FMI) measures transferred information.

[1]https://github.com/csjcai/SICE
[2]https://github.com/hanna-xu/RoadScene
[3]https://github.com/Linfeng-Tang/MSRS

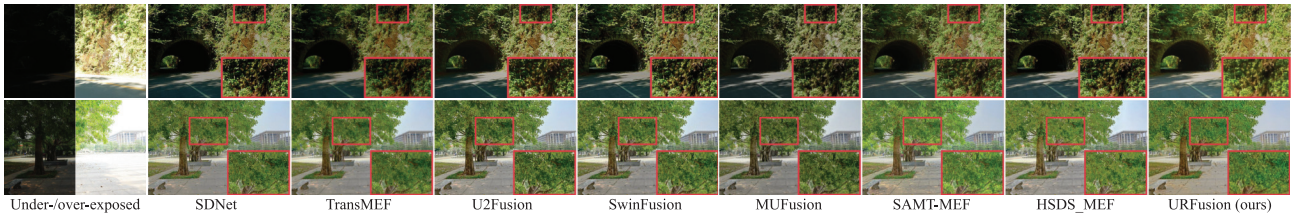


Fig. 6. Qualitative comparison results of the proposed URFusion with the state-of-the-art image fusion methods on typical multi-exposure image fusion.

TABLE II

QUANTITATIVE RESULTS OF URFUSION WITH SOTA FUSION METHODS ON TYPICAL MULTI-EXPOSURE IMAGE FUSION. MEAN AND STANDARD DEVIATION ARE REPORTED (**BOLD**: OPTIMAL, UNDERLINE: SUBOPTIMAL)

Methods	SDNet	TransMEF	U2Fusion	SwinFusion	MUFusion	SAMT-MEF	HSDS_MEF	URFusion (ours)
PSNR↑	12.281 ± 3.663	12.544 ± 3.416	11.126 ± 3.973	12.044 ± 3.368	12.574 ± 3.173	11.555 ± 3.391	10.585 ± 2.113	**13.291 ± 3.779**
SSIM↑	0.600 ± 0.201	0.643 ± 0.218	0.581 ± 0.218	0.602 ± 0.210	0.644 ± 0.221	0.596 ± 0.204	0.569 ± 0.188	**0.644 ± 0.221**
FMI↑	0.872 ± 0.062	**0.881 ± 0.058**	0.870 ± 0.057	0.867 ± 0.061	0.875 ± 0.057	0.868 ± 0.057	0.859 ± 0.049	0.876 ± 0.054

Quantitative comparison is performed on 30 pairs of multi-exposure images and the results are reported in Tab. II. The advanced performance of URFusion is evidenced by the optimal PSNR and SSIM values, and the suboptimal FMI value. The optimal PSNR and SSIM values indicate the minimal distortion and the best structure preservation in our results, preserving the fidelity and visual quality of the fused images. The suboptimal FMI value reflects that URFusion integrates comparable important features into the fused image.

*2) **Degraded Multi-Exposure Image Fusion**:* Over exposure often leads to washed-out appearance with diminished contrast and saturation while under exposure leads to noise and detail distortion. Thus, we consider three degradation scenes: i) noisy multi-exposure images; ii) low-saturation over-exposed images; and iii) low-contrast over-exposed images. We compare URFusion with both image fusion methods and the combination of restoration and fusion methods. Fusion methods fuse restored source images enhanced by restoration methods, including MaskedDenoising [53] for denoising or AirNet [54] for contrast/saturation adjustment.

Qualitative Results. Results on three degraded multi-exposure image fusion tasks are shown in Fig. 7.

Compared with SOTA image fusion methods, URFusion can effectively address the degradations in source images. Existing fusion methods are only dedicated to preserving the contents in source images and fail to consider the information quality. When source images suffer from noise, low saturation or contrast, the degradations are also preserved or even enhanced during fusion. URFusion can extract the intrinsic content from low-quality images and assign appropriate representations to generate clean results with better colors and contrast.

Compared with the combination of restoration and fusion methods, our method generates more appropriate brightness distribution and clearer scene presentation. Due to the lack of coupling between restoration and fusion methods, the restoration methods can alleviate degradations to some degree, but may pose other challenges to subsequent fusion methods. In Figs. 7 (a) and (c), MaskedDenoising or AirNet can denoise or adjust the contrast, while the final results exhibit inappropriate brightness due to pre-processing. In Fig. 7(b), the restoration method may not always be effective. In contrast, our method does not need to consider the coupling between restoration and fusion, so the results are more in line with visual perception.

Quantitative Results. As the source images contain degradations, we quantitatively validate the performances with the similarity between fused image and the ground truth (generated by multiple methods and manually selected in [55]). The metrics still include PSNR, SSIM, and FMI. Quantitative results performed on 90 pairs of multi-exposure images containing three types of degradation scenes are reported in Tab. III. URFusion achieves the optimal or suboptimal performances on PSNR, SSIM, and FMI for all the types of degradation scenes. It indicates that our results are similar to the ground truth with high image quality. Besides, URFusion achieves advantageous performances for different degradations and comparable standard deviation, indicating its generalization for both different scenes and different degradations.

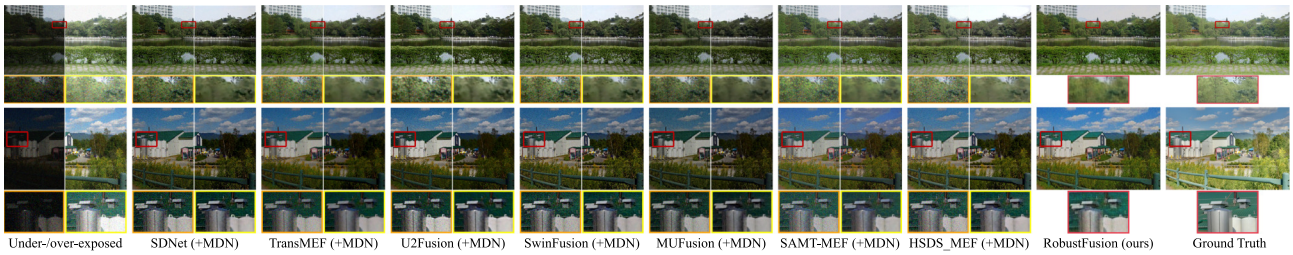
C. Visible and Infrared Image Fusion

We validate the effectiveness of URFusion on typical visible and infrared image fusion and degraded image fusion.

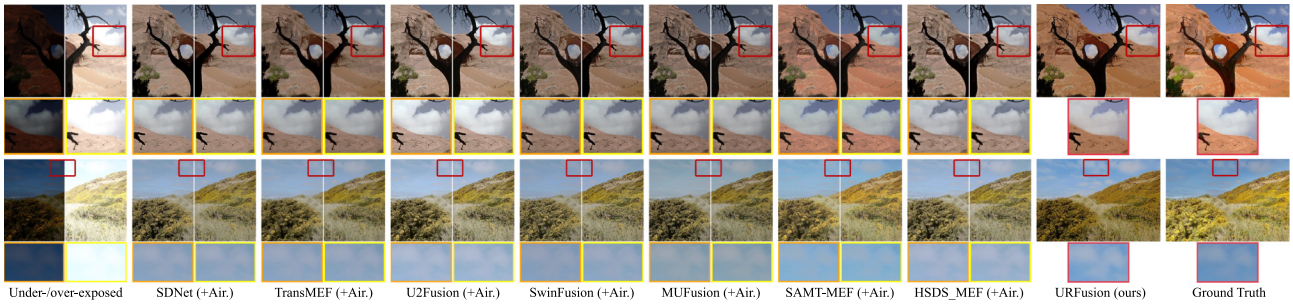
*1) **Typical Visible and Infrared Image Fusion**:* Competitors are RFN-Nest [56], U2Fusion [49], SwinFusion [10], MUFusion [50], MetaFusion [57], EgeFusion [58], Text-IF [18].

Qualitative Results. As shown in Fig. 8, our results exhibit three typical advantages. First, URFusion can better integrate the information in source images to present the scene information to a greater extent, as in the first group of results. Second, URFusion can enhance the low-quality information during fusion. As shown in the second scenario, our method can enhance the hidden details in the visible image, leading to a clearer presentation in fused image. Finally, our fusion strategy offers a more effective assessment of source information, reducing the interference of irrelevant information. In the second scenario, the infrared image contains little information. The competitors are influenced and tend to diminish the retention of visible information. In contrast, URFusion preserves the valuable visible information in the fused image.

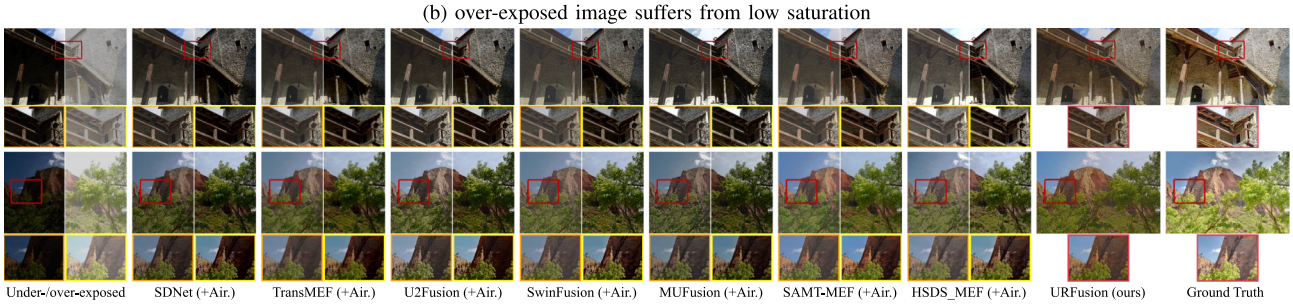
Quantitative Results. Evaluation metrics are those in Sec. IV-B. As Tab. IV reports, our method achieves the optimal PSNR, attributed to its capability of preserving source information. It ensures fused image maintain high similarity to visible image without introducing extraneous artifacts. The lower SSIM than Text-IF is because our method does not force



(a) both source images suffer from noise



(b) over-exposed image suffers from low saturation



(c) over-exposed image suffers from low contrast

Fig. 7. Qualitative comparison of URFusion with SOTA competitors on degraded multi-exposure image fusion. In the comparison results, the left parts show the results of image fusion methods, and the right parts show the results of combination of image restoration and fusion methods. The orange and yellow boxes show the highlighted regions of left and right parts, respectively. "MDN" represents MaskedDenoising and "Air." represents AirNet.

TABLE III

QUANTITATIVE RESULTS OF URFUSION WITH SOTA IMAGE FUSION METHODS ON DEGRADED MULTI-EXPOSURE IMAGE FUSION

Degradation Scenes		Methods	MDN+SDNet	MDN+Trans.	MDN+U2Fus.	MDN+SwinFus.	MDN+MUFus.	MDN+SAMT.	MDN+HSDS.	URFusion (ours)
	Noise	PSNR↑	15.508 ± 3.102	15.654 ± 3.938	14.993 ± 2.608	16.417 ± 3.267	15.868 ± 3.816	17.170 ± 3.165	16.936 ± 2.874	**17.588 ± 2.982**
		SSIM↑	0.617 ± 0.124	0.613 ± 0.139	0.620 ± 0.112	0.662 ± 0.113	0.592 ± 0.117	0.678 ± 0.114	0.657 ± 0.112	**0.696 ± 0.096**
		FMI↑	0.831 ± 0.052	0.836 ± 0.048	0.834 ± 0.053	0.835 ± 0.053	0.834 ± 0.051	0.826 ± 0.049	0.825 ± 0.045	**0.838 ± 0.050**
		Methods	**Air.+SDNet**	**Air.+Trans.**	**Air.+U2Fus.**	**Air.+SwinFus.**	**Air.+MUFus.**	**Air.+SAMT.**	**Air.+HSDS.**	**URFusion (ours)**
	Low Saturation	PSNR↑	16.316 ± 3.163	16.973 ± 3.293	16.580 ± 3.018	17.231 ± 2.835	16.917 ± 2.855	**19.384 ± 3.650**	18.263 ± 2.948	18.793 ± 2.598
		SSIM↑	0.626 ± 0.126	0.668 ± 0.133	0.677 ± 0.125	0.684 ± 0.113	0.655 ± 0.130	**0.779 ± 0.109**	0.726 ± 0.109	0.733 ± 0.090
		FMI↑	0.861 ± 0.050	0.873 ± 0.047	0.863 ± 0.044	0.862 ± 0.053	0.868 ± 0.047	0.870 ± 0.045	0.864 ± 0.049	**0.874 ± 0.042**
		Methods	**Air.+SDNet**	**Air.+Trans.**	**Air.+U2Fus.**	**Air.+SwinFus.**	**Air.+MUFus.**	**Air.+SAMT.**	**Air.+HSDS.**	**URFusion (ours)**
	Low Contrast	PSNR↑	12.796 ± 3.554	13.687 ± 3.375	13.920 ± 2.714	13.470 ± 3.465	14.032 ± 2.796	15.442 ± 3.166	15.816 ± 2.257	**17.064 ± 2.198**
		SSIM↑	0.485 ± 0.172	0.542 ± 0.142	0.591 ± 0.114	0.533 ± 0.158	0.547 ± 0.112	0.673 ± 0.105	0.641 ± 0.105	**0.675 ± 0.139**
		FMI↑	0.828 ± 0.043	0.854 ± 0.045	0.841 ± 0.042	0.840 ± 0.050	0.845 ± 0.041	0.854 ± 0.043	0.841 ± 0.046	**0.858 ± 0.041**

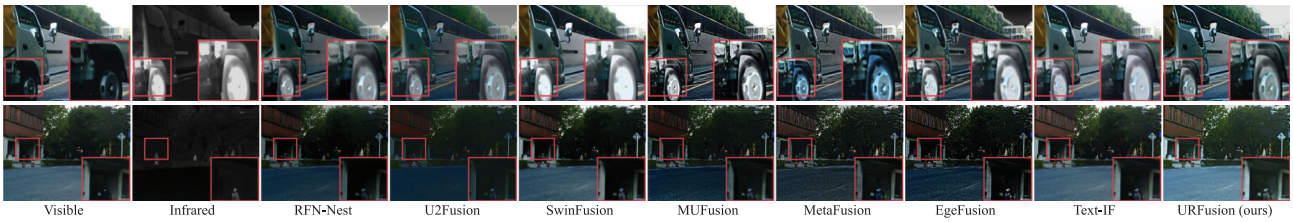


Fig. 8. Qualitative results of URFusion with SOTA image fusion methods on typical visible and infrared image fusion.

structural mimicry of source images. As a supervised method, Text-IF explicitly enforces structure-level fidelity with the attributes of source images as direct supervision. By contrast, our method applies unsupervised feature-level denoising and

TABLE IV

QUANTITATIVE RESULTS OF URFUSION WITH SOTA IMAGE FUSION METHODS ON TYPICAL VISIBLE AND INFRARED IMAGE FUSION

Methods	RFN-Nest	U2Fusion	SwinFusion	EgeFusion	MetaFusion	MUFusion	Text-IF	URFusion (ours)
PSNR↑	14.107 ± 1.377	14.106 ± 1.330	15.641 ± 1.019	14.037 ± 0.787	15.269 ± 1.134	13.466 ± 1.655	15.007 ± 1.245	**15.910 ± 0.974**
SSIM↑	0.588 ± 0.089	0.698 ± 0.047	0.732 ± 0.052	0.439 ± 0.049	0.666 ± 0.036	0.657 ± 0.063	**0.791 ± 0.026**	0.761 ± 0.036
FMI↑	0.927 ± 0.011	0.905 ± 0.014	0.925 ± 0.015	0.876 ± 0.017	0.910 ± 0.012	0.900 ± 0.014	**0.943 ± 0.010**	0.919 ± 0.014

TABLE V

QUANTITATIVE RESULTS OF URFUSION WITH SOTA IMAGE RESTORATION AND FUSION METHODS ON DEGRADED VISIBLE AND INFRARED IMAGES

Degradation Scenes		Methods	Bread+RFN.	Bread+U2Fus.	Bread+SwinFus.	Bread+EgeFus.	Bread+MetaFus.	Bread+MUFus.	Text-IF	URFusion (ours)
	Low Light	IL-NIQE↓	42.947 ± 3.933	33.310 ± 3.612	31.747 ± 3.244	36.943 ± 5.097	**27.804 ± 3.149**	39.957 ± 5.161	30.048 ± 1.799	32.436 ± 2.350
		CLIP-IQA↑	0.094 ± 0.020	0.105 ± 0.022	0.081 ± 0.017	0.134 ± 0.039	0.122 ± 0.022	0.121 ± 0.018	0.103 ± 0.024	**0.138 ± 0.024**
		NRQM↑	4.546 ± 0.412	**6.933 ± 0.659**	5.470 ± 0.799	6.086 ± 0.322	5.827 ± 0.898	3.586 ± 0.522	6.118 ± 0.844	6.132 ± 0.404
		Methods	**MDN+RFN.**	**MDN+U2Fus.**	**MDN+SwinFus.**	**MDN+EgeFus.**	**MDN+MetaFus.**	**MDN+MUFus.**	**Text-IF**	**URFusion (ours)**
	Noise	IL-NIQE↓	45.189 ± 8.959	43.070 ± 7.130	37.789 ± 8.998	118.588 ± 27.905	47.384 ± 7.360	42.288 ± 5.505	64.809 ± 11.698	**30.719 ± 3.566**
		CLIP-IQA↑	0.098 ± 0.017	0.187 ± 0.077	0.141 ± 0.053	0.229 ± 0.056	0.172 ± 0.060	0.110 ± 0.022	0.145 ± 0.026	**0.239 ± 0.066**
		NRQM↑	7.801 ± 0.528	7.868 ± 0.260	8.313 ± 0.392	6.046 ± 0.279	6.638 ± 0.683	6.703 ± 0.774	7.760 ± 0.233	**8.409 ± 0.251**
		Methods	**Air.+RFN.**	**Air.+U2Fus.**	**Air.+SwinFus.**	**Air.+EgeFus.**	**Air.+MetaFus.**	**AirNet+MUFus.**	**Text-IF**	**URFusion (ours)**
	Low Contrast	IL-NIQE↓	47.397 ± 6.601	34.347 ± 4.255	33.068 ± 4.120	38.354 ± 4.830	**29.416 ± 3.618**	44.380 ± 4.942	29.548 ± 3.272	32.583 ± 3.050
		CLIP-IQA↑	0.122 ± 0.059	0.125 ± 0.056	0.145 ± 0.052	0.132 ± 0.049	0.116 ± 0.050	0.113 ± 0.030	**0.195 ± 0.074**	0.178 ± 0.060
		PaQ-2-PiQ↑	55.506 ± 2.749	64.105 ± 2.185	63.902 ± 2.327	**67.603 ± 1.895**	64.477 ± 2.344	59.184 ± 2.142	64.849 ± 2.092	65.205 ± 2.445

leans universal natural representation priors from large-scale natural image datasets. The constraints are more relaxed than direct supervision but can decouple fusion from source-specific degradations. It ranks third in FMI due to the denoising process slightly subtle suppressing features. However, this trade-off enhances overall robustness by filtering noise-induced false correlations. The comprehensive results imply that our results effectively combine the relevant information from source images while maintaining high visual quality.

*2) **Degraded Visible and Infrared Image Fusion**:* Degradation scenes include: i) low-light visible images; ii) noisy infrared and visible images; and iii) low-contrast visible images. These degradation scenes are considered and set based on the image characteristics in several datasets and realistic conditions. In degradation cases, image restoration methods are MaskDenoising [53] for denoising, Bread [59] for low-light enhancement, and AirNet [54] for contrast adjustment. As Text-IF handles degradations according to input text description, we no longer add image restoration methods for it.

Qualitative Results. Results on three types of degradations are shown in Fig. 9. In (a), the competitors are dark to see. When the low-light image is pre-enhanced, colored patch noise is introduced and some methods suffer from detail distortion. URFusion realizes enhancement during fusion, achieving high-fidelity and high-quality scenes. In (b), existing methods exacerbate the noise. The denoising method blurs some structures and the noise is reduced but still exists. URFusion preserve the structures when removing noise. In (c), competitors tend to preserve infrared information, resulting in distortion and noise remnant. AirNet improves the results with higher contrast, but introduces color distortion and intensifies noise. Our method can present scene content with appropriate contrast enhancement while suppressing noise. Moreover, compare with degradation-aware fusion method Text-IF, our method demonstrates generalization and collaborative processing capabilities for multiple types of degradations.

Quantitative Results. As there is no ground truth in degraded scenes, we evaluate with no-reference quality assessment metrics, including integrated local natural image quality evaluator (IL-NIQE) [60], CLIP-IQA [61], non-reference quality metric (NRQM) [62], and PaQ-2-PiQ [63]. As reported in Tab. V, URFusion achieves the optimal/suboptimal results of CLIP-IQA on three types of degradations, indicating clear content expression and semantic consistency in generated images. The comparable results on IL-NIQE and NRQM indicate that our results are consistent with the visual perception of natural images. The advantage of URFusion on PaQ-2-PiQ proves the advantage in visual perception quality, detail preservation, and distortion reduction. The comprehensive results indicate the generalization of our method to different types of degradations.

Our performance shows a significant improvement in the noise scenario while relatively insignificant improvement in low-light/contrast scenarios. The divergences are from distinct mechanisms for two types of degradations. The noise is content-correlated degradation, which is addressed by intrinsic content extraction while the representation-correlated degradations (including low-light and low-contrast degradations) are solved via appearance representation assignment. Compared with intrinsic content extraction, it uses channel-wise transformations, which is of low degree of freedom to ensure structural invariance of fused content.

D. Validation of Content and Representation Learning

The validation is performed through three experiments. From the perspective of comprehensive coverage of fusion tasks, they are conducted on multi-exposure fusion and infrared and visible image fusion, respectively.

*1) **Intrinsic Content Features**:* We visualize the features of a high-quality image and low-quality images of different degradations of the same scene, and the features extracted by VGG-16 for comparison. As shown in Fig. 10, when using VGG, the features are related to degradations and significantly different, particularly in the residual noise and the missing textures. By comparison, the features extracted by intrinsic content encoder show higher consistency in intensity and structures. The noise in the features can also be suppressed.

We also perform the quantitative validation by measuring the similarity of features extracted from images in Fig. 10 by VGG and the intrinsic content extractor. The similarity

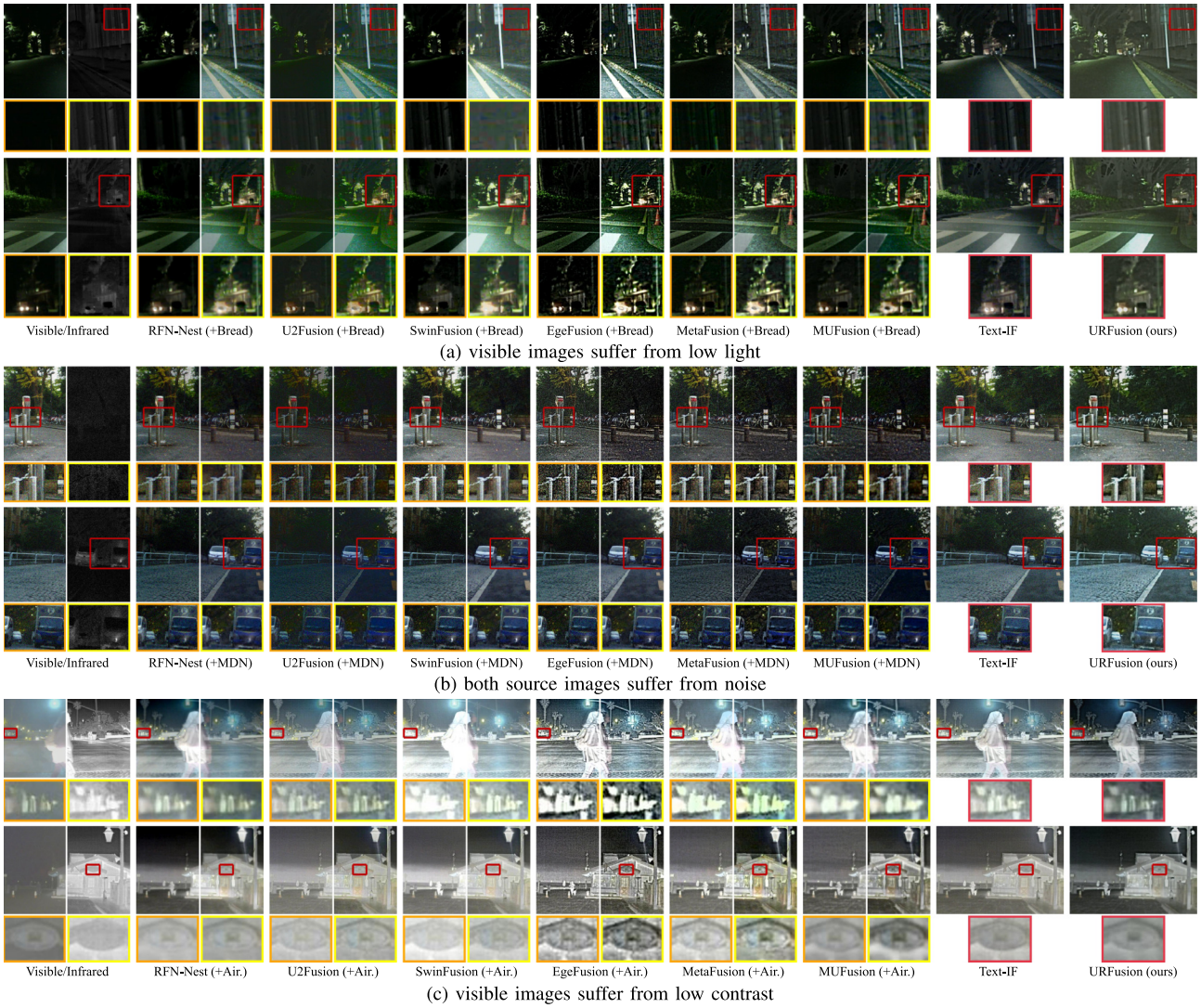


Fig. 9. Qualitative results on degraded visible and infrared image fusion. In competitors, the left parts show results of image fusion methods, and the right parts show results of combination of restoration and fusion methods. Orange and yellow boxes show highlighted regions of left and right parts, respectively.

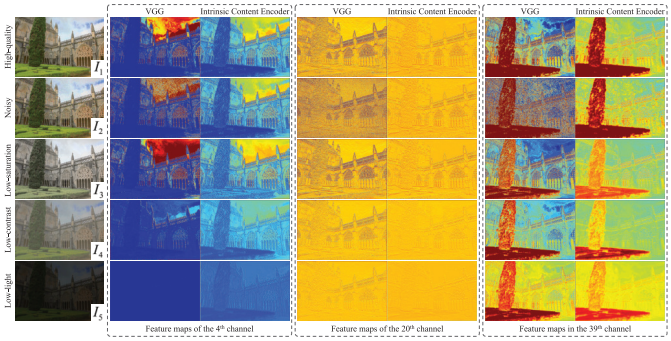


Fig. 10. Qualitative visualization of features extracted from a high-quality image and low-quality images with different degradations by VGG or intrinsic content encoder in URFusion (normalized and shown in pseudo color).

is measure by mean square error (MSE). As reported in Tab. VI, our intrinsic content extractor demonstrates enhanced consistency in feature extraction across varied representations of the same scene, indicating that these features inherently characterize the intrinsic content of the scene.

*2) **Breaking Exposure Limitation**:* As the representation is disentangled from intrinsic content, URFusion is expected to break exposure limitation. Thus, source images are not traditionally a pair of under- and over-exposed images, but both under/over-exposed images. In Fig. 11, the brightness of fused images generated by competitors is highly related with source images. URFusion can enhance dark regions and present more scene contents hidden in competitors or adjust the brightness of over-exposed regions by assigning appropriate representation. The exposure breakthrough demonstrates the effectiveness of intrinsic content and representation.

*3) **Different Representation Guidance**:* Visible and infrared images are firstly fused by content fusion network. Then, we assign the same content with appearance representations of different guidance images with various illumination, contrast, and saturation. The fused images in Fig. 12 present similar representation characteristics as guidance images. It validates that the appearance representation network extracts relevant characteristics not affected by contents and the content fusion network can map source images to a fixed domain.

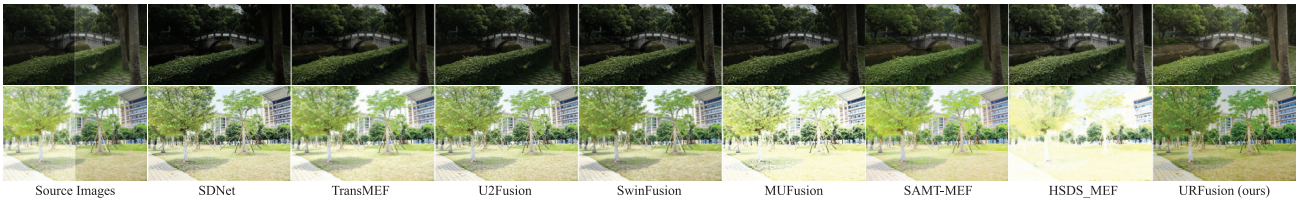


Fig. 11. Qualitative comparison results on multi-exposure image fusion where both source images are under-exposed or over-exposed images.

TABLE VI

QUANTITATIVE SIMILARITY RESULTS (MEASURE BY MEAN SQUARE ERROR, I.E., MSE) OF FEATURES EXTRACTED FROM IMAGES IN FIG. 10 BY VGG OR INTRINSIC CONTENT ENCODER. $fsim(I_m, I_n)$ REPRESENTS THE FEATURE SIMILARITY BETWEEN I_m AND I_n

Feature MSE (1e-3)	$fsim(I_1, I_2)$	$fsim(I_1, I_3)$	$fsim(I_1, I_4)$	$fsim(I_1, I_5)$	$fsim(I_2, I_3)$	$fsim(I_2, I_4)$	$fsim(I_2, I_5)$	$fsim(I_3, I_4)$	$fsim(I_3, I_5)$	$fsim(I_4, I_5)$
VGG	5.021	1.850	3.958	8.502	6.756	8.637	13.012	3.525	7.616	1.599
Intrinsic Content Encoder	0.177	0.383	0.739	1.478	0.528	0.836	1.548	0.607	1.345	0.287

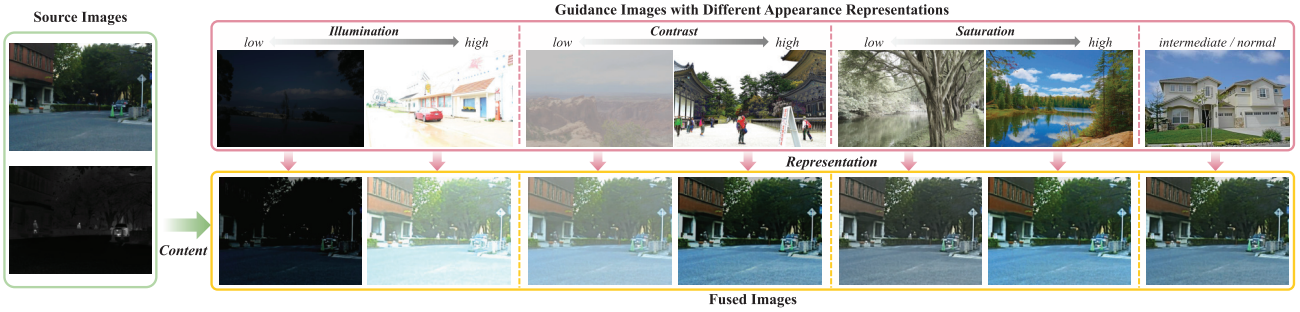


Fig. 12. Qualitative results with different appearance representations which are from different guidance images with variable illumination, contrast or saturation. As an image can simultaneously exhibit normal illumination, contrast, and saturation, the result of normal situation is uniformly shown in the last column.

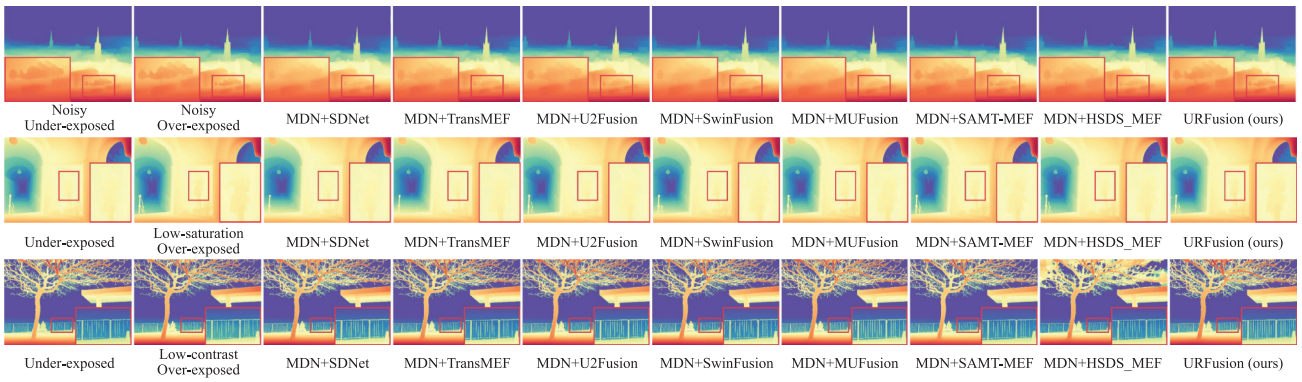


Fig. 13. External validation results on degraded multi-exposure image fusion tasks with Depth Anything V2 for depth estimation.

E. External Verification of High-Level Tasks

To externally validate the impact of fusion, we conduct experiments on downstream high-level vision tasks. For multi-exposure images, detectable targets are typically scarce. We employ a SOTA monocular depth estimation method Depth Anything V2 [64] to assess depth perception in fused images. Results in Fig. 13 reveal that the estimation results on our fusion results outperforms those on the competitors, particularly in reconstructing fine geometric details. This improvement demonstrates the capacity of our method for preserving and restoring critical spatial information during the fusion process.

For degraded visible-infrared image fusion, we conduct external validation on object detection. Detection performance serves as a critical metric for assessing fusion outputs. Detection results on degradation scenarios are shown in Fig. 14. In source images, due to the drawback of degraded unimodal images, some targets are difficult to be detected. By enhancing and fusing multi-modal images, comprehensive fused information improves detection accuracy. In fusion results, URFusion shows more complete detection categories and higher confidence scores. It validates the advantages of our method.

F. Ablation Study

1) ***Intrinsic Content Encoder****:* We replace content fusion loss with features extracted by VGG-16, and compare the content fusion results. To visual image distribution, we use t-SNE to reduce the dimension, as in Fig. 15. In the first two

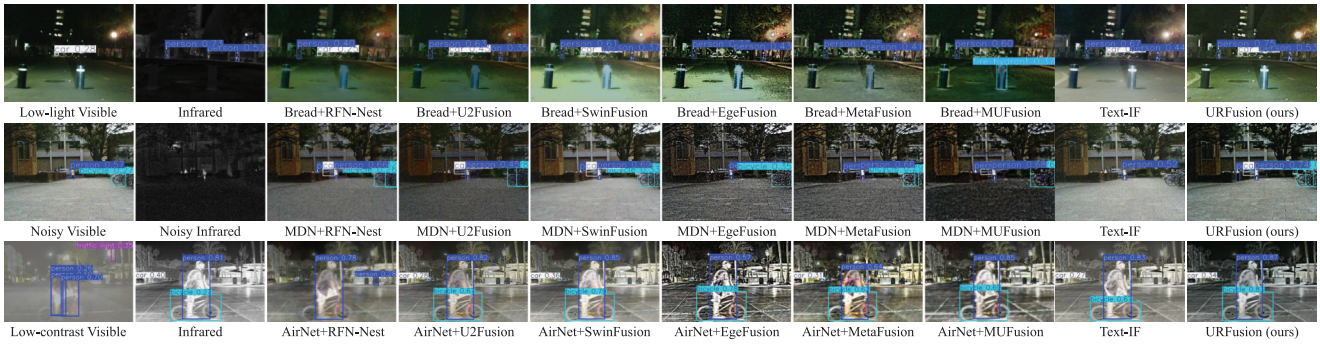


Fig. 14. External validation results on degraded visible and infrared image fusion tasks with YOLOv8 for object detection.

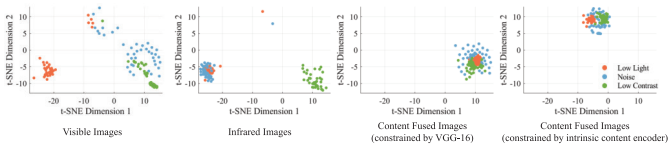


Fig. 15. Low-dimension visualization of content fusion results I_F^C with features extracted by intrinsic encoders and VGG-16. 40 groups of source images and content fusion results are dimensionality reduced by t-SNE.

Fig. 16. Qualitative results of URFusion and different fusion strategies.

TABLE VII
QUANTITATIVE RESULTS OF URFUSION WITH DIFFERENT STRATEGIES

Strategy	mean strategy	max strategy	URFusion
PSNR↑	13.418 ± 0.638	14.686 ± 0.984	**15.910 ± 0.974**
SSIM↑	0.725 ± 0.021	0.703 ± 0.034	**0.761 ± 0.036**
FMI↑	0.910 ± 0.014	0.911 ± 0.013	**0.919 ± 0.014**

sub-figures, the distributions of visible and infrared images in different degradation scenes show significant differences. In the third sub-figure, the distribution of content-fused images tends to be more concentrated than source images and similar to source features extracted by VGG-16. When we build constrain based on intrinsic content encoder in URFusion, the outputs are further aggregated, as in the last sub-figure. It validates the effectiveness of intrinsic content encoder.

*2) **Content Fusion Network**:* We compare the gradient-related strategy with mean and max strategies and retrain the fusion network. As shown in Fig. 16, mean and max strategies lead to blur results, particularly in results of mean strategy. The max strategy tends to preserve the higher-value features while fails to consider the actual physical significance, resulting in severe color and content distortion and decreased brightness. By comparison, our strategy can preserve the information to a greater extent and present clear and sharp details. Quantitative results in Tab. VII also validates the superiority.

*3) **Appearance Representation Modulation**:* We remove the channel- or attribute-level modulation, and *retrain* appearance representation network. Results are shown in Fig. 17 and Tab. VIII. The single attribute-level modulation results in high contrast but relatively low saturation. The single channel-level modulation shows more color information while some apparent representation, *e.g.*, worse contrast. With assistance of attribute-level modulation, it can generate better results.

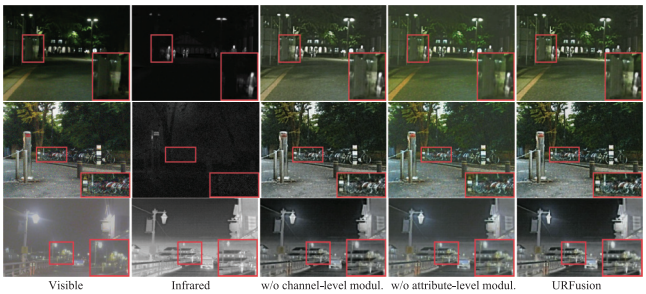


Fig. 17. Qualitative results of URFusion w/and w/o channel- or attribute-level modulation. Three scenes suffer from low light, noise, and low contrast.

TABLE VIII
QUANTITATIVE RESULTS OF URFUSION W/AND W/O CHANNEL- OR ATTRIBUTE-LEVEL MODULATION ON DEGRADATION SCENES

Degradation Scenes		Methods	w/o channel	w/o attribute	URFusion
	Low Light	IL-NIQE↓	32.553 ± 2.527	34.049 ± 3.220	**32.436 ± 2.350**
		CLIP-IQA↑	0.131 ± 0.021	0.136 ± 0.030	**0.138 ± 0.024**
		NRQM↑	5.327 ± 0.621	4.878 ± 0.514	**6.132 ± 0.404**
	Noise	**Methods**	**w/o channel**	**w/o attribute**	**URFusion**
		IL-NIQE↓	36.199 ± 2.768	32.259 ± 4.231	**30.719 ± 3.566**
		CLIP-IQA↑	0.235 ± 0.065	0.238 ± 0.062	**0.239 ± 0.066**
		NRQM↑	8.348 ± 0.252	**8.454 ± 0.250**	8.409 ± 0.251
	Low Contrast	**Methods**	**w/o channel**	**w/o attribute**	**URFusion**
		IL-NIQE↓	34.578 ± 3.837	35.071 ± 3.511	**32.583 ± 3.050**
		CLIP-IQA↑	0.162 ± 0.064	0.172 ± 0.070	**0.178 ± 0.060**
		PaQ-2-PiQ↑	64.962 ± 2.384	63.537 ± 2.624	**65.205 ± 2.445**

TABLE IX
QUANTITATIVE RESULTS ON REAL-WORLD IMAGES WITH UNSEEN AND COMPLEX DEGRADATIONS

Methods	RFN-Nest	U2Fusion	SwinFusion	EgeFusion
CLIP-IQA↑	0.091 ± 0.018	0.102 ± 0.020	0.114 ± 0.020	0.109 ± 0.015
PaQ-2-PiQ↑	57.001 ± 3.105	59.994 ± 2.278	60.913 ± 2.517	64.326 ± 1.940
Methods	**MetaFusion**	**MUFusion**	**Text-IF**	**URFusion (ours)**
CLIP-IQA↑	0.132 ± 0.033	0.104 ± 0.014	0.111 ± 0.023	**0.144 ± 0.022**
PaQ-2-PiQ↑	62.655 ± 2.628	58.524 ± 2.213	61.465 ± 2.604	**64.357 ± 2.157**

G. Generalization to Real-World Complex Degradations

To validate the generalization, we perform experiments on real-world images with unseen, mixed degradations. As not

TABLE X

PARAMETER COMPARISON OF URFUSION WITH SOTA IMAGE FUSION METHODS OR THE COMBINATION OF IMAGE RESTORATION AND FUSION METHODS ON TYPICAL OR DEGRADED IMAGE FUSION (UNIT: M)

	Degradations	Restoration \ Fusion	SDNet	TransMEF	U2Fusion	SwinFusion	MUFusion	SAMT_MEF	HSDS_MEF	URFusion (ours)
Multi-exposure	None	None	**0.067**	19.053	0.659	0.974	0.555	1.233	1.166	0.243
	Noise	MDN	0.891	19.877	1.483	1.798	1.379	2.057	1.990	
	Low Saturation	AirNet	7.677	26.663	8.270	8.584	8.165	8.844	8.777	
	Low Contrast	AirNet	7.677	26.663	8.270	8.584	8.165	8.844	8.777	
	Degradations	**Restoration \ Fusion**	**RFN-Nest**	**U2Fusion**	**SwinFusion**	**EgeFusion**	**MetaFusion**	**MUFusion**	**Text-IF**	**URFusion (ours)**
Vis. and IR	None	None	7.524	0.659	0.974	–	0.812	0.555	215.117	**0.243**
	Low Light	Bread	9.544	2.679	2.994	–	2.832	2.575		
	Noise	MDN	8.348	1.483	1.798	–	1.636	1.379		
	Low Contrast	AirNet	15.135	8.270	8.584	–	8.422	8.165		

TABLE XI

INFERENCE TIME OF URFUSION WITH SOTA IMAGE FUSION METHODS OR THE COMBINATION OF IMAGE RESTORATION AND FUSION METHODS ON TYPICAL OR DEGRADED IMAGE FUSION (UNIT: SECOND)

	Degradations	Restoration \ Fusion	SDNet	TransMEF	U2Fusion	SwinFusion	MUFusion	SAMT_MEF	HSDS_MEF	URFusion (ours)
Multi-exposure	None	None	0.176	0.631	0.758	5.180	8.393	**0.004**	10.834	0.012
	Noise	MDN	14.872	15.474	15.445	21.527	22.966	14.701	14.697	**0.011**
	Low Saturation	AirNet	3.154	4.002	4.587	12.131	20.657	2.915	2.905	**0.018**
	Low Contrast	AirNet	2.930	3.916	4.210	7.435	20.313	2.688	2.682	**0.020**
	Degradations	**Restoration \ Fusion**	**RFN-Nest**	**U2Fusion**	**SwinFusion**	**EgeFusion**	**MetaFusion**	**MUFusion**	**Text-IF**	**URFusion (ours)**
Vis. and IR	None	None	0.607	0.668	3.935	2.257	0.230	2.007	0.875	**0.040**
	Low Light	Bread	1.060	1.222	4.404	2.949	1.010	2.771	0.738	**0.037**
	Noise	MDN	11.289	11.268	14.080	13.019	11.002	12.547	0.850	**0.040**
	Low Contrast	AirNet	0.714	1.043	5.840	2.611	0.600	1.421	0.426	**0.039**

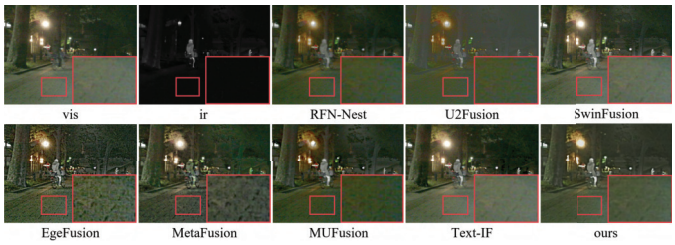


Fig. 18. Results on real-world images with unseen and complex degradations. For visual clarity, the visible image and fusion results of comparative image fusion methods are enhanced by gamma correction with $\gamma = 0.5$.

all data in existing publicly available datasets exhibit genuine degradations, we manually select 10 image pairs with real-world degradations from the MSRS dataset for validation. As the degradations are complex, it is difficult to determine optimal pre-restoration methods before fusion. Thus, we adopt SOTA image fusion for direct comparison.

In Fig. 18, the visible image contains mixed degradations, *e.g.*, darkness, chroma noise, *etc*. In the competitors, the degradations are persisted in the fused images, especially chroma noise. By comparison, these degradations are alleviated in our results. For quantitative evaluation, owing to the absence of high-quality source images for full-reference assessment, no-reference metrics are employed for evaluation. As reported in Tab. IX, our method demonstrates more superior performance, reflecting a substantial reduction in degradation artifacts within the fused outputs, validating its generalization.

H. Complexity Comparison

URFusion only uses the content fusion network in the testing phased. For typical fusion tasks, the comparison is performed with SOTA fusion methods. In degradation cases, we compare URFusion with the combination of image restoration and fusion methods. As reported in Tabs. X–XI, in typical image fusion tasks, the parameter number and inference time of URFusion are the suboptimal in fusing multi-exposure images and optimal in fusing visible and infrared images. The restoration methods usually contain much parameters and cost much inference time to process degradations. By comparison, URFusion comprehensively handles multiple degradations and image fusion with fewer parameters and less inference time.

V. CONCLUSION

This work proposes an unsupervised unified degradation-robust image fusion network URFusion. It overcomes limitations of existing methods, which either fail to handle degradations or are restricted to specific degradations. It consists of intrinsic content extraction, intrinsic content fusion, and appearance representation learning and assignment. The intrinsic content features extracted from source images provide feature-level fusion constraint. The appearance representation learned from high-quality images is assigned to content-fused result for high-quality fused image. Experiments on multi-exposure and multi-modal image fusion validate its advantages.

REFERENCES

[1] H. Zhang, X. Zuo, J. Jiang, C. Guo, and J. Ma, "MRFS: Mutually reinforcing image fusion and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26964–26973.

[2] H. Wang, M. Gong, X. Mei, H. Zhang, and J. Ma, "Deep unfolded network with intrinsic supervision for pan-sharpening," in *Proc. AAAI*, 2024, vol. 38, no. 6, pp. 5419–5426.

[3] H. Zhang, X. Zuo, H. Zhou, T. Lü, and J. Ma, "A robust mutual-reinforcing framework for 3D multi-modal medical image fusion based on visual-semantic consistency," in *Proc. AAAI*, 2024, vol. 38, no. 7, pp. 7087–7095.

[4] B. Du, C. Du, and L. Yu, "MEGF-Net: Multi-exposure generation and fusion network for vehicle detection under dim light conditions," *Vis. Intell.*, vol. 1, no. 1, p. 28, Nov. 2023.

[5] J. Liu et al., "Infrared and visible image fusion: From data compatibility to task adaption," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2349–2369, Apr. 2025.

[6] H. Xu and J. Ma, "EMFusion: An unsupervised enhanced medical image fusion network," *Inf. Fusion*, vol. 76, pp. 177–186, Dec. 2021.

[7] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[8] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[9] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[10] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.

[11] W. Tang, F. He, Y. Liu, Y. Duan, and T. Si, "DATFuse: Infrared and visible image fusion via dual attention transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3159–3172, Jul. 2023.

[12] S. Peng, X. Zhu, H. Deng, L.-J. Deng, and Z. Lei, "FusionMamba: Efficient remote sensing image fusion with state space model," 2024, *arXiv:2404.07932*.

[13] J. Yue, L. Fang, S. Xia, Y. Deng, and J. Ma, "Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models," *IEEE Trans. Image Process.*, vol. 32, pp. 5705–5720, 2023.

[14] X. Yi, L. Tang, H. Zhang, H. Xu, and J. Ma, "Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior," *Inf. Fusion*, vol. 110, Oct. 2024, Art. no. 102450.

[15] L. Tang, X. Xiang, H. Zhang, M. Gong, and J. Ma, "DIVFusion: Darkness-free infrared and visible image fusion," *Inf. Fusion*, vol. 91, pp. 477–493, Mar. 2023.

[16] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vols. 83–84, pp. 79–92, Jul. 2022.

[17] H. Zhang, L. Tang, X. Xiang, X. Zuo, and J. Ma, "Dispel darkness for better fusion: A controllable visual enhancer based on cross-modal conditional adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26477–26486.

[18] X. Yi, X. Han, H. Zhang, L. Tang, and J. Ma, "Text-IF: Leveraging semantic text guidance for degradation-aware and interactive image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 27016–27025.

[19] G. Li, Y. Lin, and X. Qu, "An infrared and visible image fusion method based on multi-scale transformation and norm optimization," *Inf. Fusion*, vol. 71, pp. 109–129, Jul. 2021.

[20] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, "A novel multi-modality image fusion method based on image decomposition and sparse representation," *Inf. Sci.*, vol. 432, pp. 516–529, Mar. 2018.

[21] Y. Chen, A. Liu, Y. Liu, Z. He, C. Liu, and X. Chen, "Multi-dimensional medical image fusion with complex sparse representation," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 9, pp. 2728–2739, Sep. 2024.

[22] H. Xu, H. Zhang, and J. Ma, "Classification saliency-based rule for visible and infrared image fusion," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 824–836, 2021.

[23] H. Xu, J. Ma, J. Yuan, Z. Le, and W. Liu, "RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 19679–19688.

[24] J. Liu, R. Lin, G. Wu, R. Liu, Z. Luo, and X. Fan, "CoCoNet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion," *Int. J. Comput. Vis.*, vol. 132, no. 5, pp. 1748–1775, May 2024.

[25] H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Inf. Fusion*, vol. 66, pp. 40–53, Feb. 2021.

[26] J. Zhang et al., "Transformer based conditional GAN for multimodal image fusion," *IEEE Trans. Multimedia*, vol. 25, pp. 8988–9001, 2023.

[27] X. Shang, G. Li, Z. Jiang, S. Zhang, N. Ding, and J. Liu, "Holistic dynamic frequency transformer for image fusion and exposure correction," *Inf. Fusion*, vol. 102, Feb. 2023, Art. no. 102073.

[28] L. Tang, Z. Yin, H. Su, W. Lyu, and B. Luo, "WFSS: Weighted fusion of spectral transformer and spatial self-attention for robust hyperspectral image classification against adversarial attacks," *Vis. Intell.*, vol. 2, no. 1, p. 5, Feb. 2024.

[29] X. Xie, Y. Cui, T. Tan, X. Zheng, and Z. Yu, "FusionMamba: Dynamic feature enhancement for multimodal image fusion with mamba," *Vis. Intell.*, vol. 2, no. 1, p. 37, Dec. 2024.

[30] Z. Zhao et al., "CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5906–5916.

[31] Y. Shi, Y. Liu, J. Cheng, Z. J. Wang, and X. Chen, "VDMUFusion: A versatile diffusion model-based unsupervised framework for image fusion," *IEEE Trans. Image Process.*, vol. 34, pp. 441–454, 2025.

[32] X. Yi, Y. Ma, Y. Li, H. Xu, and J. Ma, "Artificial intelligence facilitates information fusion for perception in complex environments," *Innovation*, vol. 6, no. 4, Apr. 2025, Art. no. 100814.

[33] H. Xu, J. Yuan, and J. Ma, "MURF: Mutually reinforcing multi-modal image registration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12148–12166, Oct. 2023.

[34] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "SuperFusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 12, pp. 2121–2137, Dec. 2022.

[35] H. Li, J. Liu, Y. Zhang, and Y. Liu, "A deep learning framework for infrared and visible image fusion without strict registration," *Int. J. Comput. Vis.*, vol. 132, no. 5, pp. 1625–1644, May 2024.

[36] Z. Liu, J. Liu, G. Wu, Z. Chen, X. Fan, and R. Liu, "Searching a compact architecture for robust multi-exposure image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 6224–6237, Jul. 2024.

[37] H. Li, Z. Yang, Y. Zhang, W. Jia, Z. Yu, and Y. Liu, "MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 5, pp. 3673–3690, May 2025.

[38] X. Deng and P. L. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3333–3348, Oct. 2021.

[39] M. Yao, R. Xu, Y. Guan, J. Huang, and Z. Xiong, "Neural degradation representation learning for all-in-one image restoration," *IEEE Trans. Image Process.*, vol. 33, pp. 5408–5423, 2024.

[40] J. Xu, X. Deng, C. Zhang, S. Li, and M. Xu, "Laplacian gradient consistency prior for flash guided non-flash image denoising," *IEEE Trans. Image Process.*, vol. 33, pp. 6380–6392, 2024.

[41] X. Deng, C. Zhang, L. Jiang, J. Xia, and M. Xu, "DeepSN-Net: Deep semi-smooth Newton driven network for blind image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2632–2646, Apr. 2025.

[42] L. Tang, Y. Deng, X. Yi, Q. Yan, Y. Yuan, and J. Ma, "DRMF: Degradation-robust multi-modal image fusion via composable diffusion prior," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 8546–8555.

[43] H. Zhang, L. Cao, X. Zuo, Z. Shao, and J. Ma, "OmniFuse: Composite degradation-robust image fusion with language-driven semantics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 9, pp. 7577–7595, Sep. 2025.

[44] Y. Mansour and R. Heckel, "Zero-shot noise2noise: Efficient image denoising without any data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 14018–14027.

[45] J. Kossen et al., "Three towers: Flexible contrastive learning with pretrained image models," in *Proc. Adv. Neural Inform. Process. Syst.*, 2023, pp. 31340–31371.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[47] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2761–2785, Oct. 2021.

[48] L. Qu, S. Liu, M. Wang, and Z. Song, "Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 2126–2134.

[49] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.
[50] C. Cheng, T. Xu, and X.-J. Wu, "MUFusion: A general unsupervised image fusion network based on memory unit," *Inf. Fusion*, vol. 92, pp. 80–92, Apr. 2023.
[51] Q. Huang, G. Wu, Z. Jiang, W. Fan, B. Xu, and J. Liu, "Leveraging a self-adaptive mean teacher model for semi-supervised multi-exposure image fusion," *Inf. Fusion*, vol. 112, Dec. 2024, Art. no. 102534.
[52] G. Wu, H. Fu, J. Liu, L. Ma, X. Fan, and R. Liu, "Hybrid-supervised dual-search: Leveraging automatic learning for loss-free multi-exposure image fusion," in *Proc. AAAI*, 2023, pp. 5985–5993.
[53] H. Chen et al., "Masked image training for generalizable deep image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1692–1703.
[54] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-in-one image restoration for unknown corruption," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17431–17441.
[55] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.
[56] H. Li, X.-J. Wu, and J. Kittler, "RFN-nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, Sep. 2021.
[57] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "MetaFusion: Infrared and visible image fusion via meta-feature embedding from object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13955–13965.
[58] H. Tang, G. Liu, Y. Qian, J. Wang, and J. Xiong, "EgeFusion: Towards edge gradient enhancement in infrared and visible image fusion with multi-scale transform," *IEEE Trans. Comput. Imag.*, vol. 10, pp. 385–398, 2024.
[59] X. Guo and Q. Hu, "Low-light image enhancement via breaking down the darkness," *Int. J. Comput. Vis.*, vol. 131, no. 1, pp. 48–66, Jan. 2023.
[60] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.
[61] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, 2023, pp. 2555–2563.
[62] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Understand.*, vol. 158, pp. 1–16, May 2017.
[63] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3575–3585.
[64] L. Yang et al., "Depth anything V2," in *Proc. Adv. Neural Inform. Process. Syst.*, 2024, pp. 21875–21911.

Han Xu (Member, IEEE) received the B.S. and Ph.D. degrees from the Electronic Information School, Wuhan University, Wuhan, China, in 2018 and 2023, respectively. She is currently an Associate Researcher with the School of Automation, Southeast University, Nanjing. She has first-authored several refereed journal and conference papers, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IJCV, CVPR, ICCV, AAAI, and IJCAI. Her current research interests include computer vision and image processing.

Xunpeng Yi received the B.E. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2023, where he is currently pursuing the master's degree. He has first-authored several refereed journal and conference papers, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Information Fusion*, CVPR, and ICCV. His research interests include computer vision and image processing.

Chen Lu received the master's degree from Nanjing University of Information Science and Technology, Nanjing, China, in 2023. He is currently pursuing the Ph.D. degree with Southeast University, Nanjing. He has first-authored several refereed journal, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His current research interests include computer vision, image processing, and deep learning and its application in remote sensing image analysis.

Guangcan Liu (Senior Member, IEEE) received the bachelor's degree in mathematics and the Ph.D. degree in computer science and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2004 and 2010, respectively.

He was a Postdoctoral Researcher with the National University of Singapore, Singapore, from 2011 to 2012; the University of Illinois at Urbana–Champaign, Champaign, IL, USA, from 2012 to 2013; Cornell University, Ithaca, NY, USA, from 2013 to 2014; and Rutgers University, Piscataway, NJ, USA, in 2014. He was a Professor with the School of Automation, Nanjing University of Information Science and Technology, Nanjing, China, from 2014 to 2021. He is currently a Professor with the School of Automation, Southeast University, Nanjing. His research interests include the areas of machine learning, computer vision, and signal processing.

Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University, Wuhan. He has co-authored more than 400 refereed journal and conference papers, including *Cell*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and IJCV. He was a recipient of the Information Fusion Best Paper Award in 2024 and the Hsue-Shen Tsien Paper Award in 2023. He is an Area Editor of *Information Fusion*, an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE/CAA JOURNAL OF AUTOMATICA SINICA, *Neurocomputing*, *Geospatial Information Science*, and *Image and Vision Computing*, and a Youth Editor of *The Innovation* and *Fundamental Research*.