LarTap: A Luminance-aware Framework with Text-correlation Priors for Multi-Exposure Image Fusion

Enlong Wang, Jiawei Li, Tiantian Yan, *Member, IEEE,* Jia Lei, Shihua Zhou, *Member, IEEE,* Bin Wang, *Member, IEEE,* Jinyuan Liu, *Member, IEEE,* and Nikola K. Kasabov, *Life Fellow, IEEE*

***Abstract*—Conventional imaging devices often struggle to produce high-dynamic-range (HDR) images that accurately represent natural scenes. To overcome this limitation, multi-exposure image fusion (MEF) techniques have been introduced as a viable solution. Existing MEF approaches aim to enhance performance by optimizing or searching architectures. However, they face challenges in precise feature extraction and scene reconstruction, leading to distortion in the fused images. Additionally, most methods do not adequately address luminance variations across different image regions, which may result in the loss of essential details. To address these challenges, we present a novel luminance-aware MEF framework that integrates text-correlation priors (LarTap). By embedding textual information into fusion process, the proposed framework enhances content extraction and comprehension. Specifically, it consist of two key components: the text-image correlation network (N1) and the multi-exposure fusion network (N2). First, N1 performs correlation training to achieve a holistic alignment between text and image pairs. Its iterative vision encoders (VEs) generate text-correlated prior knowledge to facilitate the fusion process in N2. Second, N2 leverages these priors for scene reconstruction and dynamically adjusts luminance based on comparative perception. Extensive experiments on multiple datasets demonstrate that LarTap outperforms state-of-the-art methods. The source code is available at https://github.com/EnLong-wang/LarTap.**

***Index Terms*—Multi-exposure image fusion, luminance-aware framework, text-correlation priors.**

This work is supported by 111 Project (No. D23006), the National Natural Science Foundation of China (No. 62272079), the National Foreign Expert Project of China (No.D20240244), Natural Science Foundation of Liaoning Province (No. 2024-MS-212), Scientific Research Project of Liaoning Provincial Department of Education (No. LJ222411258005), the Artificial Intelligence Innovation Development Plan Project of Liaoning Province (No. 2023JH26/10300025), the Dalian Outstanding Young Science and Technology Talent Support Program (No. 2022RJ08), Dalian Major Projects of Basic Research (No. 2023JJ11CG002), the Interdisciplinary Project of Dalian University (No. DLUXK-2024-YB-001), Joint plan of Liaoning Province science and technology plan (Nos.2024JH2/102600064, 2024-MSLH-009). (Corresponding authors: Shihua Zhou; Bin Wang.) (Enlong Wang and Jiawei Li contributed equally to this work.)

Enlong Wang, Tiantian Yan, Shihua Zhou, and Bin Wang are with the Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, School of Software Engineering, Dalian University, Dalian 116622, China (e-mail: wangenlong@s.dlu.edu.cn; yantiantian@dlu.edu.cn; zhoushihua@dlu.edu.cn; wangbin@dlu.edu.cn).

Jiawei Li is with School of Computer and Communication Engineering, University of the Science and Technology Beijing, Beijing 100083, China (e-mail: ljw19970218@163.com).

Jia Lei is with the School of Software Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: leijia03@stu.xjtu.edu.cn).

Jinyuan Liu is with the School of Mechanical Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: atlantis918@hotmail.com)

Nikola K. Kasabov is with the Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Auckland 1061, New Zealand (e-mail: nkasabov@aut.ac.nz).

I. Introduction

THE human visual system (HVS) demonstrates remarkable adaptability to the high-dynamic-range (HDR) of natural scenes, facilitated by photoreceptor perception and pupil regulation [1, 2]. In contrast, conventional digital imaging devices often exhibit results with suboptimal brightness and detail, deviating from the way the HVS perceives scenes [3, 4]. To overcome this limitation, researchers have developed HDR imaging techniques from a hardware perspective [5, 6, 7]. Although advanced photographic equipment can achieve satisfactory results, its high cost and limited applicability remain significant drawbacks. To address these problems, multi-exposure image fusion (MEF) has been introduced as a promising alternative, offering notable advantages in fields, such as photography [8] and remote sensing [9].

MEF aims to merge multiple images captured at different exposure levels into a single output that integrates the most visually significant features from each source image [10, 11, 12, 13]. The fusion network preserves the structural integrity and details of the scene, producing results with greater visual expressiveness than any single exposure. Early MEF research primarily focused on mathematical models that employed spatial or transform domain fusion, relying on handcrafted rules to generate well-exposed images [14, 15]. They primarily focus on spatial or transform domain fusion, leveraging experience and handcrafted rules to create a well-exposed output. However, such fixed-rules approaches struggle with complex and unpredictable scenes, leading to issues, such as contrast degradation and texture distortion.

In recent years, deep learning has significantly advanced image fusion task [16, 17, 18, 19]. Techniques incorporating contrastive learning [20] and architecture search [21] have been developed to improve the MEF performance. While these approaches effectively extract visual features, they often struggle to achieve reliable fusion in complex scenarios. Given the potential of semantic textual information, researchers have started exploring the integration of text in image fusion [22, 23]. However, neglecting the inherent differences between heterogeneous images and texts can lead to suboptimal results. Since MEF datasets typically lack textual descriptions, most existing text-based models have limited practical applicability. Furthermore, another challenge is the significant luminance variations across different regions of source images [24], which many existing methods fail to address, potentially leading to the loss of crucial details in extremely exposed areas.

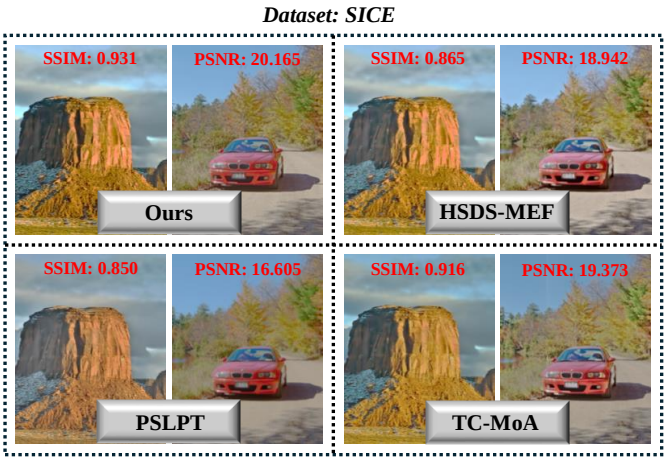


Fig. 1. A partial presentation of experiments with three SOTA methods on the SICE dataset, demonstrating that our approach outperforms existing state-of-the-art methods both qualitatively and quantitatively.

To address these challenges, we propose a luminance-aware fusion framework incorporating text-correlation priors, referred to as LarTap. Our framework leverages text captions to assist in extracting richer structural features [25]. Specifically, textual information serves as auxiliary prior knowledge for MEF by providing precise object concepts and contextual relationships [26]. To ensure practical applicability, textual inputs are only utilized during the training phase. Additionally, our method integrates luminance perception and adjustment within the fusion process. The proposed framework consists of two main components: the text-image correlation network (N1) and the multi-exposure fusion network (N2). In N1, caption and VEs project text and image inputs into a shared embedding space. Given the inherent heterogeneity between text and images, a correlation loss function is introduced to enforce holistic-level alignment. This correlation training strategy encourages the vision encoder to learn text-correlated priors, enhancing feature extraction in N2. The extracted image features are then refined through the prior guidance module (PGM), to align with the true distribution. Finally, the comparison perception module (CPM) and pseudo-label constraints determine the luminance properties of the source images. These properties are subsequently used in the luminance refinement module (LRM) to enhance the coarse fusion output. The dual modulation operations in LRM ensure balanced exposure and high contrast in the final result.

As shown in the Fig. 1, our method achieves higher contrast, rendering the light and dark sides of the hills more realistically. Additionally, the license plate number of the red car is clearly visible. The red font indicates the quantitative results on the SICE dataset, highlighting the advantages of our approach. The key contributions of this work are summarized as follows.

- We propose a novel prior-guided, luminance-aware framework that leverages extra prior knowledge to enhance fusion quality. The integration of text-correlation priors enables precise feature extraction and realistic scene reconstruction.
- We present an image-text correlation training strategy that allows the network to learn unified cross-modal representations, ensuring that textual information contributes harmoniously to the fusion task.
- We develop a luminance module group consisting of CPM and LRM. CPM estimates the luminance characteristics of the source images based on pseudo-label constraints, while LRM refines the initial fusion result through dual modulation operations.

The rest of this paper is structured as follows. Section II reviews existing MEF methods, categorizing them into traditional and deep learning-based approaches, along with a brief discussion of text-driven image processing. Section III details the implementation of the proposed method. Section IV presents qualitative and quantitative experimental results. Section V conducts ablation studies on modules, loss functions, and design choices to validate the proposed approach. In Section VI, we discuss the limitations of the proposed method. Finally, Section VII concludes the paper.

II. Related Works

Our proposed framework incorporates textual information into the MEF task. This section reviews relevant studies related to LarTap, encompassing both MEF approaches and text-driven image processing techniques.

A. Multi-exposure Image Fusion

MEF methodologies can be broadly classified into traditional approaches [14, 27, 28, 29] and deep learning-based methods [21, 30, 31, 32]. While traditional methods effectively integrate information from source images, their predetermined fixed fusion rules often struggle with complex natural scenes [33]. Conversely, deep learning-based techniques utilize advanced feature extraction capabilities, leading to improved fusion outcomes [34, 35, 36, 37].

1) Traditional methods: Paul et al. [38] proposed a method that differentiates between luminance and chrominance channels of color images, utilizing the gradient domain and weighted summation for fusion. Unlike conventional pixel-wise methods, Ma et al. [39] decomposed source images into three elements—signal strength, signal structure, and mean intensity—based on image patches. SPD-MEF [40] mitigates ghosting issues in dynamic scenes through patch decomposition, though it remains computationally demanding. To enhance efficiency, Li et al. [41] introduced a two-scale processing approach. Addressing ghosting artifacts, DSIFT [42] employed a dense scale-invariant feature transform, while Ulucan et al. [41] proposed robust weight map characterization. Recognizing that maximizing quality measure does not always enhance visual perception, Ma et al. [43] developed a fusion algorithm optimizing color MEF structural similarity. Later, Hayat et al. [44] introduced a weight map estimation technique leveraging local contrast, brightness, and color dissimilarity. Furthermore, PESPD-MEF [45] integrated perceptual components, enhancing SPD-MEF with perceptual gain and refinement rules for improved fusion quality.

2) Deep learning-based methods: Prabhakar et al. [46] pioneered the applications of deep learning to MEF, overcoming limitations inherent to hand-crafted fusion rules. Expanding on this, Qu et al. [47] incorporated transformers, employing auxiliary reconstruction tasks to enhance feature generalization. Furthermore, Zhang et al. [48] divided the source image into reflectance, shading, and color components, designing specialized sub-networks for each. Liu et al. [49] developed an attention-guided model within a coarse-to-fine framework, integrating edge loss and global-local learning to restore texture and preserve color fidelity. Similarly, Li et al. [50] proposed a cooperative learning network where self-attention and edge correction modules retained intricate image details. However, prioritizing detail and color preservation can sometimes compromise overall fusion performance. To address this, DMEF [51], leveraging Retinex theory, decomposed images into luminance and reflectance maps for separate fusion processing. Most methods overlook local alignment and fail to fully extract global features. To tackle this, Luo et al. [52] designed a bidirectional network utilizing deformable self-attention and achieved one of the best fusion results. Additionally, gamma correction techniques have been explored for MEF [24, 53]. Recently, Wu et al. [54] introduced HSDS-MEF, addressing the rigidity of manually designed loss functions and fixed network architectures.

Recognizing the synergy among various image fusion tasks, researchers have explored general networks capable of handling multiple fusion applications. PMGI [55] frames image fusion as a proportional maintenance of gradient and intensity. Building on this, Zhang et al. [56] introduced a squeeze and decomposition network to enhance information retention from source images. A key challenge with generalized fusion networks is catastrophic forgetting, complicating their training. To address this issue, Xu et al. [57] proposed U2Fusion, addressing universal fusion difficulties through structural similarity constraints. Cheng et al. [58] introduced a novel training paradigm using memory units, where intermediate results function as supervisory signals to enhance human visual perception. Regarding training architectures, PSLPT [59] employs a semi-supervised approach to enhance data complementarity. To minimize task interference, Zhu et al. [60] proposed the task-customized hybrid adapter, enabling adaptive fusion across multiple tasks within a unified framework.

B. Text-driven Image Processing

Text, as an independent modality from images, has gained traction for its ability to enhance image processing tasks. Providing rich contextual information, the text proves valuable in various vision-related applications. For instance, Zeng et al. [61] employed textual scene descriptions to predict dense depth maps from single images. Similarly, Guo et al. [62] demonstrated that using text-based prompts improved image recognition accuracy through joint visual-text feature learning. Additionally, Li et al. [63] addressed text-image alignment challenges, significantly enhancing performance by minimizing the cross-modality gap. In image fusion, Zhao et al. [22] and Yi et al. [23] integrated text guidance to extract deep semantic information, improving fusion quality. Nonetheless, challenges persist in bridging the gap between textual and visual embeddings, necessitating further exploration of text-guided image fusion. Even when both modalities describe the same scene, their feature representations may differ significantly, leaving room for improvement.

III. Proposed Method

This section begins with a problem formulation to elucidate our motivation. We then provide a comprehensive description of our framework and the composition of loss functions.

A. Problem Formulation

Textual descriptions offer high-level semantic cues that assist the model in comprehend the image[64]. However, paragraph-based texts often include redundant information. To address this limitation, we use concise, one-sentence captions that highlight the salient content of images. For instance, a caption, such as "a person on the beach at sunset" helps the model focus on preserving details of people while reconstructing the surrounding beach and ocean [62]. Many existing approaches that directly integrate text embeddings with image embeddings may cause fusion distortions. To mitigate this, we introduce a correlation network to align image-text pairs. Given that standard fusion datasets lack textual captions, we propose a novel strategy that reduces reliance on textual inputs during inference.

Given the variability in appearance, incorporating precise textual priors (e.g., flowers, sky, clouds) provides an advantage over exclusively relying on neural networks for feature extraction. Unlike purely visual feature-based fusion, a textual prior-driven model offers high-level object semantics and explicit contextual relationships, enhancing structural integrity. In addition, to address luminance variations across different scenes, a luminance module group comprising a comparison perception module and a luminance refinement module are introduced for dynamic luminance adaptation. Ultimately, we propose a luminance-aware framework integrating textual priors for MEF, illustrated in Fig. 2. The framework consists of a text-image correlation network (N1) and a multi-exposure fusion network (N2). During training, the four inputs (text-image pairs for under-exposed and over-exposed images) are fed into the fusion network. Notably, we manually assigned captions based on the SICE dataset [65]. After training, the vision encoder captures text-based priors, guiding fusion during inference. Our results can be generated by:

$$\begin{aligned} \mathbf{I}_f &= \mathcal{N}_{infer}(\mathbf{I}_{ue}, \mathbf{I}_{oe}; \omega^*), \\ \omega^* &\in \arg\min_{\omega} \mathcal{L}_{total}(\mathcal{N}_{train}(\mathbf{I}_{ue}, \mathbf{T}_{ue}, \mathbf{I}_{oe}, \mathbf{T}_{oe}; \omega)) \end{aligned} \tag{1}$$

where $\mathbf{I}$ and $\mathbf{T}$ represent images and text, respectively. The subscripts *ue* and oe correspond to under-exposure and over-exposure, respectively, while f denotes the fusion results. The subscript *infer* signifies the inference process. The entire network is denoted by $\mathcal{N}$, with $\mathcal{L}_{total}$ representing the total loss and ω indicating the learnable parameters of $\mathcal{N}$.

As denoted by N1 and N2, N1 is intended to mitigate the adverse effects of modality heterogeneity, whereas N2 focuses

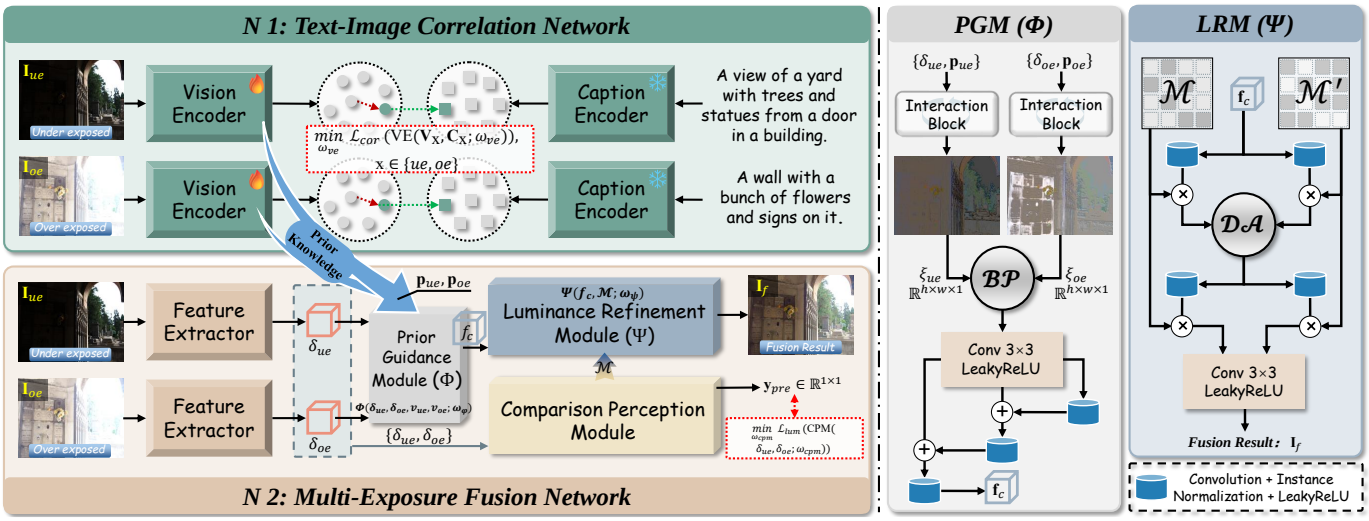


Fig. 2. The detailed diagram of the developed architecture. On the left, the overall workflow is depicted: N1 aligns the image-text pairs and generates text-correlation priors, while N2 integrates these priors and outputs the final fusion result after luminance refinement. On the right, the internal structures of the Prior Guidance Module (PGM) and the Luminance Refinement Module (LRM) are illustrated.

Fig. 3. A partial presentation of captions we used on the SICE dataset. The figure displays two sets of image-text pairs and the fusion results. Since the overexposed and underexposed images emphasize different aspects of the scene, their corresponding captions are also distinct.

on extracting and aggregating features from the source images. Their optimization is formulated as follows:

$$\min_{\omega_{N1}} \mathcal{L}_{cor}\left(\mathrm{N1}(\mathbf{I}_{ue}, \mathbf{T}_{ue}, \mathbf{I}_{oe}, \mathbf{T}_{oe}; \omega_{N1})\right) \tag{2}$$

$$\min_{\omega_{N2}} \mathcal{L}_{fus}\left(\mathrm{N2}(\mathbf{I}_{ue}, \mathbf{I}_{oe}; \omega_{N2})\right) \tag{3}$$

where $\mathcal{L}_{cor}$ and $\mathcal{L}_{fus}$ represent the loss constraints for N1 and N2, respectively. A detailed explanation will be provided in the following sections.

B. Network Architecture

1) Text-Image Correlation Network (N1): To encode textual inputs, we employ CLIP with frozen parameters. The VEs incorporate residual connections and convolution blocks (CBs), compromising a 3×3 convolution, instance normalization, and a LeakyReLU activation layer. During training, the parameters of VEs are continuously updated, allowing them to generate prior knowledge that aligns with textual meanings. To facilitate this process, we introduce correlation loss for MEF, where matching image-text pairs serve as positive samples, while other images and texts within the batch function as negative samples. Holistic-level alignment is achieved by bringing positive samples closer and distancing the negative ones. Detailed formula derivation is provided in the loss function section.

2) Multi-Exposure Fusion Network (N2): As illustrated in Fig. 2, in N2, the under-exposed $\mathbf{I}_{ue}$ and over-exposed $\mathbf{I}_{oe}$ images are first processed by feature extractors to obtain δ_{ue} and δ_{oe}. The extractors comprise five CBs and skip connections.

Next, text-correlation prior knowledge $\mathbf{p}_{ue}$ and $\mathbf{p}_{oe}$ is integrated into PGM alongside extracted features to enhance convergence. To ensure consistency, the VE in N1 is trained to conform to specific sentence patterns, eliminating the need for textual captions during inference. This approach enables N2 to maintain image structure while interpreting textual context. The framework ensures the generated fusion outcomes align with text-image distributions while retaining generality. The fusion process produces a coarse result $\mathbf{f}_c$ is obtained:

$$\mathbf{f}_c = \Phi(\delta_{ue}, \delta_{oe}, \mathbf{p}_{ue}, \mathbf{p}_{oe}) \tag{4}$$

where Φ represents the PGM.

Regions within the same scene may exhibit varying luminance levels, and a simple fusion pattern can lead to information loss in extremely exposed regions. To enhance dynamic range and texture richness, we design CPM to assess source image intensity, generating a luminance perception map ($\mathcal{M}$). Finally, LRM utilizes $\mathcal{M}$to refine the modulation result:

$$\mathbf{I}_f = \Psi(\mathbf{f}_c, \mathcal{M}) \tag{5}$$

where Ψ denotes the LRM, and $\mathbf{I}_f$ represents the final result.

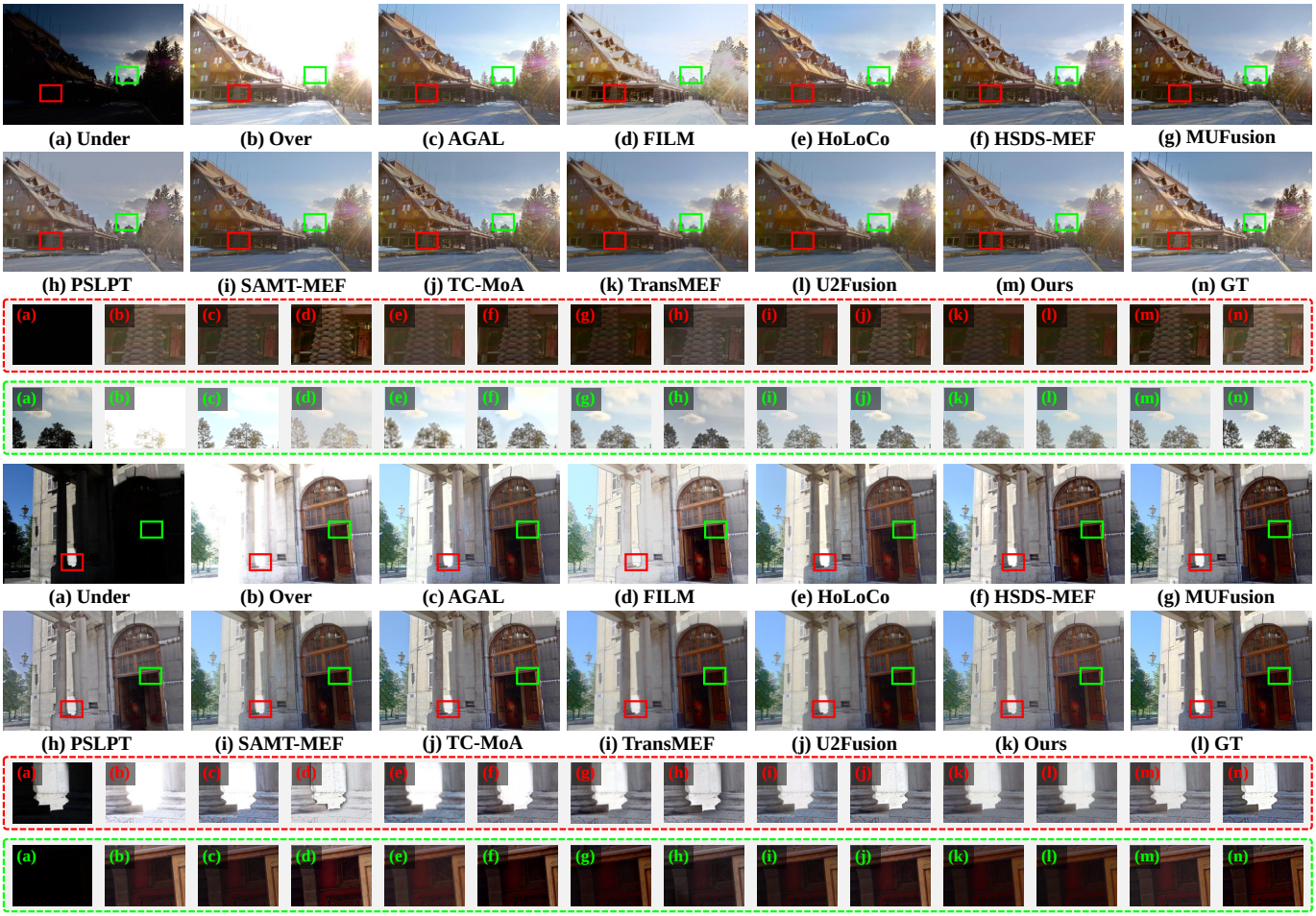


Fig. 4. Qualitative experiments with ten SOTA methods on the SICE dataset. The dashed box below the fusion result contains the magnified content. Our LarTap exhibits the richest details and appropriate exposure.

3) Prior Guidance Module: In PGM, four inputs are processed in pairs through the interaction block (IB). Atrous spatial pyramid pooling (ASPP) is employed within IB to improve the representational ability of prior knowledge, leveraging multi-scale convolution operations to expand the receptive field and manage contextual relationships.

As illustrated in Fig. 2, a bi-pooling block enhances integration following the concatenation of ξ_{ue} and ξ_{ue}. Maxpooling and Avgpooling layers capture key attributes and contextual information within local regions. Multiplication then activates beneficial expressions, described as:

$$\xi = \mathrm{Concat}(\xi_{ue}, \xi_{oe}) \tag{6}$$

$$\mathbf{f}_c^{'} = \mathcal{S}(\mathrm{MAX}(\xi) + \mathrm{AVG}(\xi)) \cdot \xi \tag{7}$$

where $\mathcal{S}$ represents the sigmoid function, and Concat denotes the concatenation operation. MAX and AVG refer to Maxpooling and Avgpooling, respectively.

Then we conduct dense element-wise Addition to connect the information preceding and succeeding the convolutions, preventing the unexpected degradation and generating the coarse fusion result $\mathbf{f}_c$.

4) Comparison Perception Module: Inspired by [66], CPM is designed using linear layers and residual connections to determine input luminance perception. Extracted features (δ_{ue} and δ_{oe}) are processed to predict luminance values of size $\mathbb{R}^{1\times1}$. Before backpropagation, the mean intensity of source images to compared to derive the luminance label:

$$\mathbf{y}_{lum} = \begin{cases} 1 & \text{if } \bar{\mathbf{I}}_{ue} < \bar{\mathbf{I}}_{oe} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $\bar{\mathbf{I}}$ represents mean intensity, and $\mathbf{y}_{lum}$ denotes the luminance label.

Binary cross-entropy loss between predicted and actual labels is employed to train CPM in learning luminance-related representations. The corresponding equations are detailed in the loss function section. Intermediate linear layer outputs form the perception outcome ($\mathcal{M}$).

5) Luminance Refinement Module: To enhance output quality, LRM refines results through $\mathcal{M}$ and $\mathcal{M}^{'}$ (where $\mathcal{M}^{'} = 1 - \mathcal{M}$). Following initial element-wise multiplication, dual-domain attention (spatial and channel) highlights key features. As depicted in Fig. 2, this is represented as:

$$\mathbf{I}_f^{'} = \mathrm{DA}(\mathrm{CB}(\mathbf{f}_c \cdot \mathcal{M}), \mathrm{CB}(\mathbf{f}_c \cdot \mathcal{M}^{'})) \tag{9}$$

where $\mathbf{I}_f^{'}$ denotes the intermediate result, DA refers to dual attention, and CB signifies CB processing.

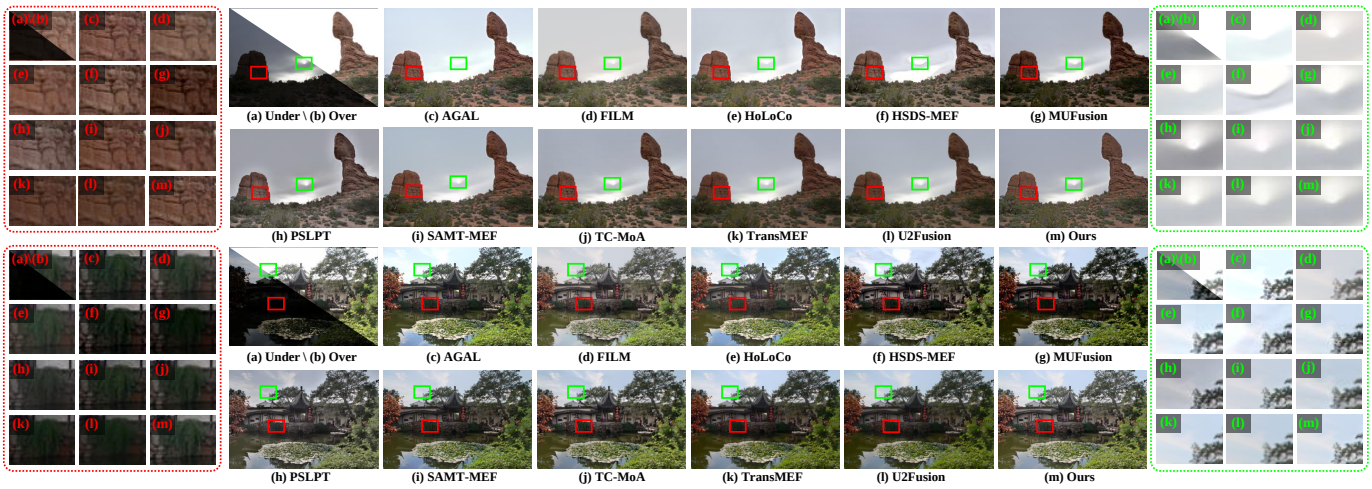


Fig. 5. Qualitative experiments with ten SOTA methods on the MEFB dataset. The middle of the figure presents the source image and the fusion results of each method. The left side displays the red enlarged patches, while the right side shows the green enlarged patches. Our results have remarkable advantages in luminance and details.

Subsequently, $\mathbf{I}_f^{'}$ undergoes further processing with the perception map to refine luminance modulation, ensuring a structurally faithful and detail-rich final result.

C. Loss Function

The total loss function $\mathcal{L}_{total}$ comprises fusion loss $\mathcal{L}_{fus}$ and correlation loss $\mathcal{L}_{cor}$, expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{fus} + \alpha\mathcal{L}_{cor} \tag{10}$$

where α is a hyperparameter balancing both components.

Specifically, $\mathcal{L}_{cor}$ enhances the alignment of image-text pairs in embedding space while mitigating modality constraints, given by:

$$\begin{aligned}\mathcal{L}_{cor} = \frac{1}{2N}(\textstyle\sum_{i=1}^{N}\mathrm{CE}(\hat{\mathbf{V}}_i \cdot \hat{\mathbf{C}}, \mathbf{y}_i)+ \\ \textstyle\sum_{j=1}^{N}\mathrm{CE}(\hat{\mathbf{C}}_j \cdot \hat{\mathbf{V}}, \mathbf{y}_j))\end{aligned} \tag{11}$$

where N is the batch size, CE denotes cross entropy, and $\hat{\mathbf{V}}_i \cdot \hat{\mathbf{C}}$ represents cosine similarity vector between the i-th image embedding and all text embeddings, while $\hat{\mathbf{C}}_j \cdot \hat{\mathbf{V}}$ represent opposite implication. $\mathbf{y}$ is the matching label of image-text pairs.

The fusion loss $\mathcal{L}_{fus}$ comprises the structure loss $\mathcal{L}_{str}$ and the luminance loss $\mathcal{L}_{lum}$, with their balance controlled by β. The structure loss $\mathcal{L}_{str}$ is formulated using three components: mean square error (MSE), structural similarity (SSIM), and gradient. The corresponding equation is given by:

$$\mathcal{L}_{str} = \lambda_1\mathcal{L}_{MSE} + \lambda_2\mathcal{L}_{SSIM} + \lambda_3\mathcal{L}_{grad} \tag{12}$$

where λ_x is determined through experimental validation to optimize performance. The mean square error loss $\mathcal{L}_{MSE}$ regulates pixel intensity, while the structural similarity loss $\mathcal{L}_{SSIM}$ [67] enhances alignment with human visual perception. Additionally, the gradient loss $\mathcal{L}_{grad}$ is employed to preserve fine details. It is defined as:

$$\mathcal{L}_{grad} = \frac{1}{HW}\left\|\,|\nabla\mathbf{I}_f| - |\nabla\mathbf{I}_g|\,\right\|_1 \tag{13}$$

where ∇ represents the Soble gradient operator, and $|\cdot|$ denotes the absolute operation.

Additionally, $\mathcal{L}_{lum}$ is employed to enable adaptive perception of image luminance conditions:

$$\begin{aligned}\mathcal{L}_{lum} = \mathbf{y}_{lum} \cdot \log(\mathbf{y}_{pre})+ \\ (1-\mathbf{y}_{lum}) \cdot \log(1-\mathbf{y}_{pre}))\end{aligned} \tag{14}$$

where $\mathbf{y}_{lum}$ and $\mathbf{y}_{pre}$ represent the luminance label and the predicted luminance value, respectively.

IV. Experiments

This section outlines the preliminary data preparation and detailed experimental configurations. Following this, we conduct both subjective and objective evaluations on MEF tasks using ten different methods, highlighting the superior performance of the proposed LarTap.

A. Data Preparation

The initial step of our study involves incorporating textual captions into each image in the SICE dataset. These captions are generated using a pre-trained model [68] and subsequently refined through manual adjustments. As illustrated in Fig. 3 each image is described using a single-sentence caption. Given the overexposed and underexposed images often exhibit content variations across scenes, their respective captions differ accordingly. Our scene interpretation considers both objects and their contextual relationships.

For each image, we employ text with a consistent grammatical structure to explicitly convey object details and spatial relationships. Additionally, we maintain standardized terminology for objects of the same type to ensure consistency. Variations in object appearance due to factors, such as viewing angles or lighting conditions pose challenges for the network in accurately interpreting images. Text, being a concise and structured representation, facilitates the use of uniform language to categorize objects. By enforcing standardized terminology, we enhance the network's ability to identify similarities among

TABLE I
QUANTITATIVE EXPERIMENTS WITH TEN SOTA METHODS ON THE SICE AND THE MEFB DATASETS. THE PINK BOXES IN THE TABLE HIGHLIGHT THE OPTIMAL VALUES, AND THE LIGHT BLUE BOXES INDICATE THE SUBOPTIMAL VALUES. ACROSS MOST METRICS, OUR PROPOSED METHOD OUTPERFORMS THE OTHERS.

Method	SICE						MEFB (*w/o* reference)				
	MEF-SSIM	PSNR	SD	EN	AG	SF	MEF-SSIM	SD	EN	AG	SF
AGAL	0.907	18.302	10.215	7.202	7.231	0.087	0.949	10.446	7.155	5.671	0.074
FILM	0.847	12.784	10.622	6.867	7.031	0.086	0.960	10.688	7.248	5.527	0.074
HoLoCo	0.929	19.542	10.139	7.155	5.105	0.050	0.904	10.301	7.148	4.437	0.049
HSDS-MEF	0.865	18.942	10.333	7.110	7.111	0.083	0.943	10.342	7.064	5.416	0.068
MUFusion	0.767	15.887	10.295	6.890	5.635	0.066	0.961	10.227	6.987	4.274	0.055
PSLPT	0.850	16.605	10.281	6.733	6.123	0.072	0.677	10.176	6.869	4.735	0.061
SAMT-MEF	0.887	19.023	10.157	6.971	5.943	0.072	0.962	10.413	7.054	4.921	0.065
TC-MoA	0.916	19.373	10.072	6.951	6.775	0.083	0.965	10.201	7.043	4.953	0.062
TransMEF	0.793	15.832	9.947	6.823	4.192	0.051	0.942	10.161	7.009	3.539	0.047
U2Fusion	0.848	17.311	9.923	6.788	3.883	0.043	0.904	10.056	6.872	3.006	0.037
Ours	0.931	20.165	10.369	7.295	7.344	0.088	0.963	10.473	7.281	5.673	0.076

homogeneous objects, thereby enhancing feature extraction and scene reconstruction.

While longer sentences may provide additional details, they increase the computational burden on the model [22]. Complex paragraphs not only complicate processing but may also introduce redundant information. Therefore, we hypothesize that single-sentence captions are more effective in guiding the network's fusion process. The comparative experiments presented in this section validate the effectiveness of our approach.

B. Network Implementation

The SICE dataset, along with its corresponding text captions, serves as the training set, compromising 374 groups. The descriptive text is structured to differentiate overexposed from underexposed scenes, ensuring the network focuses on crucial areas. The batch size is set to 4. The hyper-parameters α, β, λ_1, λ_2, and λ_3 are determined through ablation studies, with values of 0.1, 0.1, 50, 50, and 20, respectively. The training process uses the Adam optimizer with a learning rate of 1e-4. The framework is implemented in Python, and all experiments are conducted on an NVIDIA GeForce 3070ti GPU.

C. Comparison Methods and Evaluation Metrics

Qualitative and quantitative comparison are performed against ten state-of-the-art methods, including AGAL [49], FILM [22], HoLoCo [69], HSDS-MEF [54], MUFusion [58], PSLPT [59], SAMT-MEF [70], TC-MoA [60], TransMEF [47], and U2Fusion [57]. The evaluation metrics used to validate the effectiveness of the proposed approach include SSIM, PSNR, SD, EN, AG, and SF. Specifically, SSIM measures the structural similarity between the fused output and the reference image by assessing luminance, contrast, and structural details. PSNR quantifies the fidelity of the reconstructed image by calculating the peak signal-to-noise ratio. SD represents the variation in pixel intensity, serving as an indicator of the metric of image contrast. EN measures the amount of information contained in the image, while AG reflects sharpness and fine details. Finally, SF captures image activity and texture by analyzing frequency components along horizontal and vertical axes, providing insights into detail preservation.

D. Qualitative Experiments

As shown in Fig. 4, two challenging scenes are selected to comprehensively assess the fusion performance. In the first example, the proposed method successfully preserves comprehensive information and achieves appropriate exposure. AGAL exhibits noticeable overexposure near the sun, while FLIM struggles with boundary handling, leading to suboptimal visual perception. PSLPT introduces abnormal halos around the buildings, the HSDS-MEF produces artifacts in the sky region. MUFusion, TransMEF, and U2Fusion generate results with insufficient brightness in building areas. The results of SAMT-MEF and TC-MoA are closest to ours. In enlarged red patch, the proposed approach maintains high contrast, its reliance on overexposed inputs results in severe distortions affecting surfaces, such as roads and trees. Therefore, our fusion effect is the most realistic. HoLoCo displays blurred textures on windows. In green patch, unlike AGAL and HSDS-MEF, the proposed method effectively preserves cloud details. MUFusion, TransMEF, and U2Fusion show competitive performance but are less effective in capturing intricate details of branches.

The second example features a high dynamic range scene. AGAL fails to retain wall textures, FILM disrupts light and shadow relationships, and PSLPT introduces unrealistic light spots within the door frame, negatively affecting human visual perception. U2Fusion generates overly smooth results. While other methods achieve reasonable fusion, LarTap excels in the magnified region. AGAL, HoLoCo, and HSDS-MEF lose details due to strong light interference in red patch. MUFusion,

Fig. 6. Visual results about the source image sequences with different exposure ratios.(1) and (2) are under-exposed images, (3), (4) and (5) are over-exposed images. (a)-(c) are fused by inputs of (1) and (3)-(5). (d)-(f) are fused by inputs of (2) and (3)-(5).

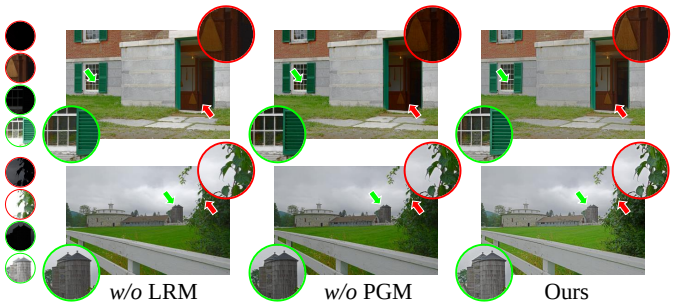


Fig. 7. Ablation experiment of the developed modules on the SICE dataset. The circular patches on the left show the underexposed and overexposed source images, and the right side displays the fusion results. After removing the module, the fusion effect is obviously degraded, which proves the effectiveness of the modules we designed.

Fig. 8. Ablation experiment of the loss functions on the SICE dataset. The circular patches on the left show the underexposed and overexposed source images, and the right side displays the fusion results. The model trained with the full loss (i.e., LarTap) performs best compared to the results with different loss components removed.

TransMEF, and U2Fusion lack sufficient clarity. In green patch, Our LarTap produces the closest to ground truth, accurately balancing exposure and detail preservation. The results of SAMT-MEF and TC-MoA both have invisible details due to insufficient luminance. With enhanced luminance perception, fine lines on windows remain distinguishable. Overall, on the SICE dataset with reference images, the proposed approach achieves the broadest dynamic range and preserves the most intricate details.

To further evaluate generalization, additional experiments are conducted on the MEFB dataset, which lacks reference images [71]. The first row of Fig. 5 shows that AGAL, FILM, HoLoCo, and the proposed method yield the best visual outcomes, whereas other approaches produce dim screens. MUFusion exhibits low contrast, obscuring grass details. PLSPT fails to maintain contextual consistency, causing a bright band around the hill and grass. The text-guided network enhances feature capture, preserving original image content. Although the exposure of the SAMT-MEF picture is appropriate, we still have a clear advantage in the stone details. The TC-MoA obviously lacks resolution, which seriously affects visual perception. In the red patches, HoLoCo, TransMEF, and U2Fusion display blurred object representations, while in the green sky patches, AGAL, FILM, and HoLoCo distort cloud structures. HSDS-MEF struggles with light source.

In the second row, all methods perform satisfactorily, but only the proposed approach distinctly renders plants in the enlarged red patch. Additionally, it retains the richest sky information. Other approaches exhibit similar limitations as observed previously. For instance, AGAL tends to overexpose inputs, FILM distorts clouds, HoLoCo and TC-MoA loses scene details, and HSDS-MEF is ineffective in complex lighting conditions. SAMT-MEF achieves effective fusion in the sky area, but shows low contrast around the red patches.The remaining methods exhibit low dynamic range. Overall, the proposed approach surpasses existing techniques, achieving superior dynamic range and detail preservation. The qualitative experiments validate the efficacy of text-based guidance and luminance refinement.

TABLE II
QUANTITATIVE ABLATION EXPERIMENTS ON THE SICE DATASET. THE PINK BOXES IN THE TABLE HIGHLIGHT THE OPTIMAL VALUES, AND THE LIGHT BLUE BOXES INDICATE THE SUBOPTIMAL VALUES. THE TABLE INCLUDES TWO SETS OF ABLATION EXPERIMENTS ON THE DESIGNED MODULES AND THE LOSS COMPONENTS.

Method	SICE				
	MEF-SSIM	PSNR	SD	EN	AG
w/o LRM	0.920	20.031	10.123	7.232	6.953
w/o PGM	0.913	19.627	10.184	7.289	6.956
w/o Grad	0.920	19.553	9.871	7.107	6.866
w/o Mse	0.930	20.136	10.246	7.292	7.056
w/o SSIM	0.817	17.242	10.141	7.192	7.099
Ours	0.931	20.165	10.369	7.295	7.344

E. Quantitative Experiments

The quantitative results in Table I demonstrate that the proposed method achieves optimal performance across most evaluation metrics. The high SSIM values indicate minimal structural distortion and strong similarity to original inputs. The highest PSNR value signifies reduced noise, ensuring superior image quality. Additionally, SD values suggest that LarTap enhances contrast and visual appeal. For MEF, a higher entropy value signifies increased image complexity and improved fusion effects. A higher AG value indicates sharper edges and improved detail preservation. The SF metric reflects higher spatial frequency, denoting richer textures and details. Overall, the proposed approach effectively preserves image textures and details, ensuring high-quality scene reconstruction.

F. Supplementary Experiments

To evaluate the generalization capability of the proposed method under varying exposure values (EVs), two representative sequences are analyzed, as shown in Fig. 6. Each sequence consists of five source images: three overexposed and two underexposed. Fusion results are generated using different combinations of these source images, with specific configurations detailed in the figure caption. Since images (2) and (4) were used as reference values during training, fusion result (e) closely matches the ground truth. Notably, the other fusion results also exhibit excellent luminance retention and rich detail preservation. Complex regions, such as rocks, cloudy skies, and trees are reconstructed with minimal distortion. Overall, the network produces visually appealing results even when source image EVs vary significantly.

V. ABLATION EXPERIMENTS

This section presents visual and numerical evaluations to validate the effectiveness of the designed modules, loss functions, and hyperparameters.

A. Ablation Experiment of Modules and Loss

To evaluate the impact of the proposed modules, an ablation experiment is conducted. As illustrated in Fig. 7, removing the CPM and LRM modules leads to suboptimal results, with objects, such as grass and buildings appearing dimmer. In the enlarged region, the full model retains details of glass windows, door frames, leaves, and buildings, demonstrating the importance of prior knowledge in structure retention.

Ablation studies on the three fusion loss components are shown in Fig. 8. Removing the gradient component degrades cloud and branch texture preservation. The version without MSE loss produces results closest to the final version but introduces noticeable artifacts. SSIM ablation results in cloud structure distortion, affecting visual perception. These findings confirm that all loss components collectively enhance network learning.

Quantitative ablation results in Table II show that the complete model achieves the best performance. Removing grad leads to lower AG values, while SSIM removal significantly impacts structure preservation. Overall, both the designed modules and loss functions positively contribute to the proposed method's effectiveness.

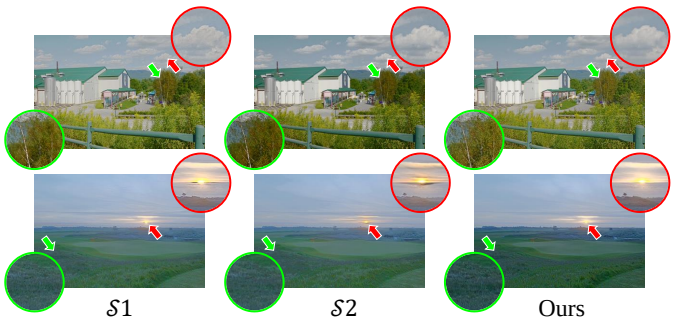


Fig. 9. Ablation experiment about the different schemes. The priors are output by the pre-trained image encoders from CLIP in $\mathcal{S}1$ and by the text encoders in $\mathcal{S}2$. Both their results contain artifacts, which prove the rationality for a trainable vision encoder to output text-related priors (our scheme).

TABLE III
QUANTITATIVE ABLATION EXPERIMENTS OF DIFFERENT SCHEMES. THE PINK BOXES IN THE TABLE HIGHLIGHT THE OPTIMAL VALUES, AND THE LIGHT BLUE BOXES INDICATE THE SUBOPTIMAL VALUES.

Scheme	SICE				
	MEF-SSIM	PSNR	SD	EN	AG
$\mathcal{S}1$	0.912	19.594	10.233	7.209	7.265
$\mathcal{S}2$	0.913	19.547	10.186	7.235	6.902
Ours	0.931	20.165	10.369	7.295	7.344

B. Ablation Experiment of Schemes

To further evaluate our approach, we examine two alternative schemes for text-guided MEF and compare their performance with our proposed method. In Scheme 1 ($\mathcal{S}1$), CLIP's image encoder is utilized to process the source image and generate output for the corresponding priors. Scheme 2 ($\mathcal{S}2$) disregards the modality gap by directly employing CLIP's text encoder. The experimental results are illustrated in Fig. 9, demonstrating that our approach yields the best outcomes. Specifically, S1 produces significant cloud distortions and

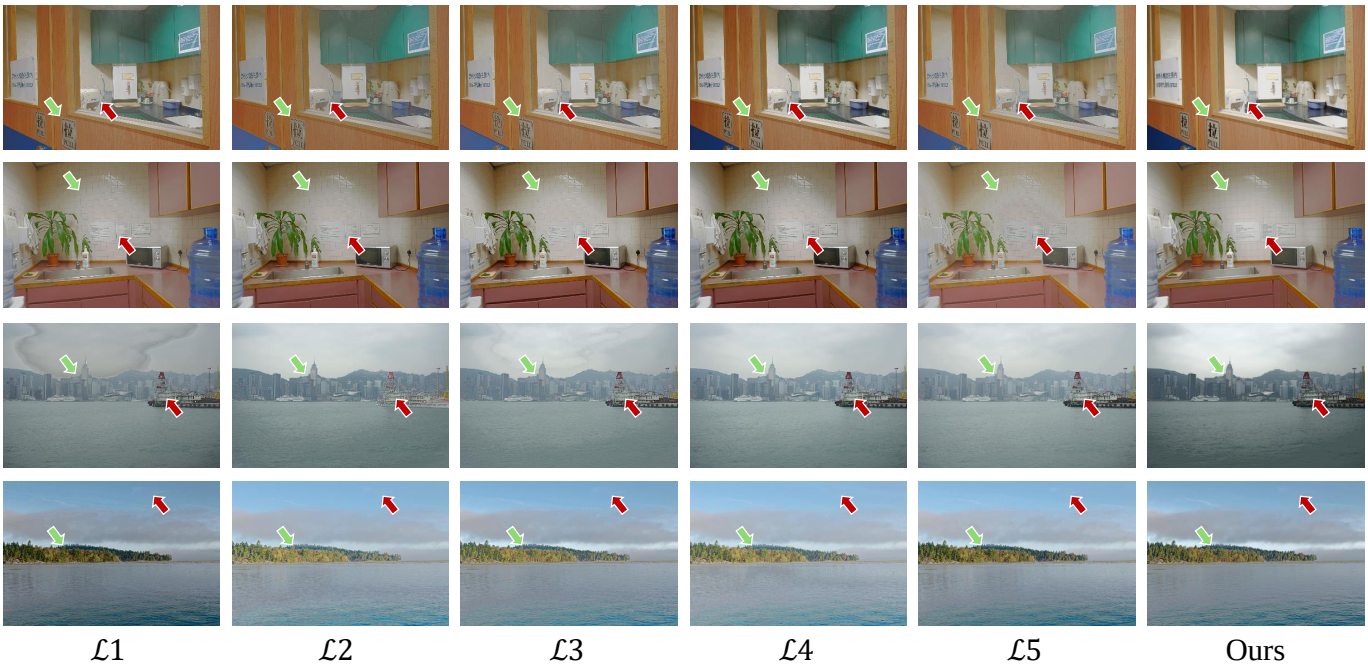


Fig. 10. Ablation experiments of hyper-parameters on the SICE dataset. The identifiers $\mathcal{L}1$, $\mathcal{L}2$, $\mathcal{L}3$, $\mathcal{L}4$, and $\mathcal{L}5$ represent different hyper-parameter settings, respectively. For specific settings, refer to Section IV and Table 4. From the fusion results, our LarTap ensures appropriate luminance and rich details.

TABLE IV
QUANTITATIVE ABLATION EXPERIMENTS OF HYPER-PARAMETERS. THE PINK BOXES IN THE TABLE HIGHLIGHT THE OPTIMAL VALUES, AND THE LIGHT BLUE BOXES INDICATE THE SUBOPTIMAL VALUES. DURING THE EXPERIMENT, WE TRIED DIFFERENT PARAMETER COMBINATIONS AND FINALLY CHOSE THE ONE WITH THE BEST PERFORMANCE.

Loss	SICE				
	MEF-SSIM	PSNR	SD	EN	AG
$\mathcal{L}1$	0.901	19.256	10.142	7.198	7.082
$\mathcal{L}2$	0.918	19.652	10.113	7.254	7.031
$\mathcal{L}3$	0.924	20.055	10.159	7.201	7.140
$\mathcal{L}4$	0.905	19.257	10.024	7.248	6.812
$\mathcal{L}5$	0.907	19.358	10.281	7.201	6.857
Ours	0.931	20.165	10.369	7.295	7.344

TABLE V
SPECIFIC NUMERICAL SETTINGS FOR ABLATION EXPERIMENT ON HYPER-PARAMETERS. THE LEFT COLUMN IS THE IDENTIFIER AND THE RIGHT COLUMN IS THE VALUE.

Loss	Hyper-parameter Settings
$\mathcal{L}1$	$\alpha = 0.1$, $\beta = 1$, $\lambda_1 = 50$, $\lambda_2 = 50$, $\lambda_3 = 20$
$\mathcal{L}2$	$\alpha = 1$, $\beta = 0.1$, $\lambda_1 = 50$, $\lambda_2 = 50$, $\lambda_3 = 20$
$\mathcal{L}3$	$\alpha = 1$, $\beta = 1$, $\lambda_1 = 50$, $\lambda_2 = 50$, $\lambda_3 = 20$
$\mathcal{L}4$	$\alpha = 0.1$, $\beta = 0.1$, $\lambda_1 = 20$, $\lambda_2 = 50$, $\lambda_3 = 50$
$\mathcal{L}5$	$\alpha = 0.1$, $\beta = 0.1$, $\lambda_1 = 50$, $\lambda_2 = 20$, $\lambda_3 = 50$

artifacts around the sun, while S2 lacks detail in branches and grass textures, resulting in an unrealistic appearance.

TABLE VI
EFFICIENCY COMPARISONS OF LARTAP WITH TEN STATE-OF-THE-ART APPROACHES. THE PINK BOXES IN THE TABLE HIGHLIGHT THE OPTIMAL VALUES, AND THE LIGHT BLUE BOXES INDICATE THE SUBOPTIMAL VALUES.

Method	Time/s	Parameters/M	FLOPs/G
AGAL	0.153	1.59	65.23
FILM	0.449	0.49	14.75
HoLoCo	0.384	17.40	29.94
HSDS-MEF	13.858	1.17	43.92
MUFusion	0.435	0.55	9.21
PSLPT	0.558	1.26	28.86
SAMT-MEF	0.133	1.23	47.45
TC-MoA	1.194	340.35	524.28
TransMEF	0.489	19.05	10.21
U2Fusion	0.159	0.66	43.17
Ours	0.508	51.434	3.34

The quantitative analysis of each scheme is presented in Table III. In $\mathcal{S}1$, CLIP's image encoder is leveraged to extract prior knowledge. Given that CLIP is trained using contrastive loss, it effectively associates image and text features. However, due to the unique nature of the MEF task, relying exclusively on the pre-trained encoder for prior generation may hinder performance. Experimental findings validate that our approach, incorporating an adaptive encoder with correlation loss, is more effective. In $\mathcal{S}2$, the text encoder of CLIP is directly applied, necessitating the inclusion of a text title during inference. The substantial modality gap is expected to impact

the results, which is confirmed by the metrics in Table III, where S2 exhibits the weakest performance.

C. Ablation Experiment of hyper-parameters

To identify the optimal hyper-parameters for the loss function, we initialized values based on empirical knowledge and conducted ablation experiments for refinement. The qualitative and quantitative analyses are shown in Fig. 10 and Table IV, respectively. The model configuration yielding the best performance was selected as the final setup. The hyperparameters settings are detailed in Table V. From the analysis of extensive experimental results, we draw two key conclusions. First, in fusion and related tasks, the weight assigned to fusion should be larger. This aligns with our original intention—ensuring that related training tasks assist the fusion task rather than dominate it. Notably, although the final values of α and β are both set to 0.1, the three λ values in the fusion task are significantly larger. This design ensures that the fusion component is artificially emphasized in the overall loss function. Second, among the three components of the fusion loss—MSE, SSIM, and gradient—the weight of the gradient term should be smaller than the other two to achieve the best performance. However, removing it entirely would lead to blurred results. Therefore, λ_3 is set to a value smaller than the first two to maintain optimal fusion quality.

VI. Limitation

This section discusses the limitations of the proposed approach in detail. As presented in Table VI, we assess the efficiency of comparative methods based on runtime, FLOPs, and parameter count. Although our model has better peformance, its inference efficiency is suboptimal. This may stem from the linear network design, which lacks structural optimization. Additionally, while the prior guidance module, comparison perception module, and luminance adjustment module have demonstrated effectiveness in ablation experiments, they inevitably introduce computational overhead.

During training, we employ the preprocessed SICE dataset, which is noise-free and well-aligned. However, real-world images often contain noise from environmental factors, such as rain, snow, and fog, leading to reduced image quality and diminished scene perception. Directly fusing such images significantly degrades the fusion results. Our proposed method focuses on aggregating complementary information from multi-exposure images and does not inherently address noise robustness. In denoising research, numerous algorithms have been proposed. For instance, Ju et al. [72] argue that spatial domain algorithms alone cannot effectively distinguish real structures from noise, such as snow artifacts, and introduce a snow removal network integrating Fourier frequency and spatial information. This dual-domain approach enhances image clarity. Given the advantages of robust fusion algorithms in improving usability, future research could explore integrating fusion and denoising techniques.

VII. Conclusion

This paper investigates the role of textual information as an auxiliary component in MEF. We introduce a luminance-aware fusion framework that incorporates text-correlation priors. By aligning images with their corresponding textual descriptions through text-image correlation training, VEs generate prior knowledge that facilitates effective fusion. Notably, our training strategy eliminates the need for textual data during inference. Furthermore, the CPM, regulated by pseudo-label constraints, ensures that the resulting images maintain optimal brightness and contrast. Our approach achieves an extended dynamic range while preserving intricate details and structural integrity, demonstrating its effectiveness in MEF tasks.

References

[1] J. Lei, J. Li, J. Liu, S. Zhou, Q. Zhang, and N. K. Kasabov, "Galfusion: Multi-exposure image fusion via a global–local aggregation learning network," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–15, 2023.

[2] C. She, F. Han, L. Wang, S. Duan, and T. Huang, "Mpc-net: Multi-prior collaborative network for low-light image enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 10 385–10 398, 2024.

[3] L. Wang and K.-J. Yoon, "Deep learning for hdr imaging: State-of-the-art and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8874–8895, 2022.

[4] C. Liu, F. Wu, and X. Wang, "Efinet: Restoration for low-light images via enhancement-fusion iterative network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8486–8499, 2022.

[5] W. Hong, H. Zhang, and J. Ma, "Merf: A practical hdr-like image generator via mutual-guided learning between multi-exposure registration and fusion," *IEEE Transactions on Image Processing*, vol. 33, pp. 2361–2376, 2024.

[6] S. Liu and Y. Zhang, "Detail-preserving underexposed image enhancement via optimal weighted multi-exposure fusion," *IEEE Transactions on Consumer Electronics*, vol. 65, no. 3, pp. 303–311, 2019.

[7] O. Ulucan, D. Karakaya, and M. Turkan, "Multi-exposure image fusion based on linear embeddings and watershed masking," *Signal Processing*, vol. 178, p. 107791, 2021.

[8] M. Inanici, "Evaluation of high dynamic range photography as a luminance data acquisition system," *Lighting Research Technology - LIGHTING RES TECHNOL*, vol. 38, pp. 123–136, 06 2006.

[9] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 639–643, 2017.

[10] H. Li, T. N. Chan, X. Qi, and W. Xie, "Detail-preserving multi-exposure fusion with edge-preserving structural patch decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 11, pp. 4293–4304, 2021.

[11] Y. Yang, D. Zhang, W. Wan, and S. Huang, "Multi-scale exposure fusion based on multi-visual feature measurement and detail enhancement representation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[12] Q. Wang, W. Chen, X. Wu, and Z. Li, "Detail-enhanced multi-scale exposure fusion in yuv color space," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2418–2429, 2020.

[13] X. Deng, J. Xu, F. Gao, X. Sun, and M. Xu, "DeepM^2m2cdl: Deep multi-scale multi-modal convolutional dictionary learning network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2770–2787, 2024.

[14] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2864–2875, 2013.

[15] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, and C. Zhang, "Probabilistic exposure fusion," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 341–357, 2012.

[16] E. Wang, J. Li, J. Lei, J. Liu, S. Zhou, B. Wang, and N. K.Kasabov, "Sdfuse: Semantic-injected dual-flow learning for infrared and visible image fusion," *Expert Systems with Applications*, vol. 252, p. 124188, 2024.

[17] J. Li, J. Chen, J. Liu, and H. Ma, "Learning a graph neural network with cross modality interaction for image fusion," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4471–4479.

[18] J. Li, J. Liu, S. Zhou, Q. Zhang, and N. K. Kasabov, "Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023.

[19] J. Liu, G. Wu, Z. Liu, D. Wang, Z. Jiang, L. Ma, W. Zhong, and X. Fan, "Infrared and visible image fusion: From data compatibility to task adaption," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2024.

[20] Y. Zhang, R. Xie, J. Chen, X. Sun, Z. Kang, and Y. Wang, "Enhancing contrastive learning inspired by the philosophy of "the blind men and the elephant"," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 21, 2025, pp. 22 659–22 667.

[21] Z. Liu, J. Liu, G. Wu, Z. Chen, X. Fan, and R. Liu, "Searching a compact architecture for robust multi-exposure image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6224–6237, 2024.

[22] Z. Zhao, L. Deng, H. Bai, Y. Cui, Z. Zhang, Y. Zhang, H. Qin, D. Chen, J. Zhang, P. Wang, and L. V. Gool, "Image fusion via vision-language model," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[23] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 27 026–27 035.

[24] K. Wu, J. Chen, Y. Yu, and J. Ma, "Ace-mef: Adaptive clarity evaluation-guided network with illumination correction for multi-exposure image fusion," *IEEE Transactions on Multimedia*, vol. 25, pp. 8103–8118, 2023.

[25] J. Wang, B. Zhang, J. Pang, H. Chen, and W. Liu, "Rethinking prior information generation with clip for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 3941–3951.

[26] M. Lee and J. Choi, "Text-guided variational image generation for industrial anomaly detection and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 26 519–26 528.

[27] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," in *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, 2007, pp. 382–390.

[28] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel-and region-based image fusion with complex wavelets," *Information fusion*, vol. 8, no. 2, pp. 119–130, 2007.

[29] S. Li and X. Kang, "Fast multi-exposure image fusion with median filter and recursive filter," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 626–632, 2012.

[30] J. Li, H. Yu, J. Chen, X. Ding, J. Wang, J. Liu, B. Zou, and H. Ma, "A^2rnet: Adversarial attack resilient network for robust infrared and visible image fusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 5, 2025, pp. 4770–4778.

[31] J.-L. Yin, B.-H. Chen, and Y.-T. Peng, "Two exposure fusion using prior-aware generative adversarial network," *IEEE Transactions on Multimedia*, vol. 24, pp. 2841–2851, 2022.

[32] R. Liu, Z. Liu, J. Liu, X. Fan, and Z. Luo, "A task-guided, implicitly-searched and meta-initialized deep model for image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 10, pp. 6594–6609, 2024.

[33] 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, pp. 1748–1775, 2022.

[34] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[35] Z. Zheng, W. Ren, X. Cao, X. Hu, T. Wang, F. Song, and X. Jia, "Ultra-high-definition image dehazing via multi-guided bilateral learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 180–16 189.

[36] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 105–119, 2022.

[37] Z. Jiang, Z. Zhang, J. Liu, X. Fan, and R. Liu, "Multispectral image stitching via global-aware quadrature pyramid regression," *IEEE Transactions on Image Processing*, vol. 33, pp. 4288–4302, 2024.

[38] S. Paul, I. Sevcenco, and P. Agathoklis, "Multi-exposure and multi-focus image fusion in gradient domain," *Journal of Circuits, Systems and Computers*, vol. 25, p. 1650123, 06 2016.

[39] K. Ma and Z. Wang, "Multi-exposure image fusion: A patch-wise approach," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 1717–1721.

[40] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang, "Robust multi-exposure image fusion: A structural patch decomposition approach," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2519–2532, 2017.

[41] H. Li, K. Ma, H. Yong, and L. Zhang, "Fast multi-scale structural patch decomposition for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 5805–5816, 2020.

[42] Y. Liu and Z. Wang, "Dense sift for ghost-free multi-exposure fusion," *Journal of Visual Communication and Image Representation*, vol. 31, pp. 208–224, 2015.

[43] K. Ma, Z. Duanmu, H. Yeganeh, and Z. Wang, "Multi-exposure image fusion by optimizing a structural similarity index," *IEEE Transactions on Computational Imaging*, vol. 4, no. 1, pp. 60–72, 2018.

[44] H. Naila and I. Muhammad, "Ghost-free multi exposure image fusion technique using dense sift descriptor and guided filter," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 295–308, 2019.

[45] J. Zhang, Y. Luo, J. Huang, Y. Liu, and J. Ma, "Multi-exposure image fusion via perception enhanced structural patch decomposition," *Information Fusion*, vol. 99, p. 101895, 2023.

[46] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4724–4732.

[47] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, Jun. 2022, pp. 2126–2134.

[48] H. Zhang and J. Ma, "Iid-mef: A multi-exposure fusion network based on intrinsic image decomposition," *Information Fusion*, vol. 95, pp. 326–340, 2023.

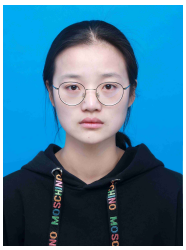
[49] J. Liu, J. Shang, R. Liu, and X. Fan, "Attention-guided global-

local adversarial learning for detail-preserving multi-exposure image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5026–5040, 2022.
[50] J. Li, J. Liu, S. Zhou, Q. Zhang, and N. K. Kasabov, "Learning a coordinated network for detail-refinement multiexposure image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 2, pp. 713–727, 2023.
[51] K. Wu, J. Chen, and J. Ma, "Dmef: Multi-exposure image fusion based on a novel deep decomposition method," *IEEE Transactions on Multimedia*, vol. 25, pp. 5690–5703, 2023.
[52] J. Luo, W. Ren, X. Gao, and X. Cao, "Multi-exposure image fusion via deformable self-attention," *IEEE Transactions on Image Processing*, vol. 32, pp. 1529–1540, 2023.
[53] A. Kumar, R. K. Jha, and N. K. Nishchal, "An improved gamma correction model for image dehazing in a multi-exposure fusion framework," *Journal of Visual Communication and Image Representation*, vol. 78, p. 103122, 2021.
[54] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, Mar. 2024, pp. 5985–5993.
[55] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, Apr. 2020, pp. 12 797–12 804.
[56] H. Zhang and J. Ma, "Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vision*, vol. 129, no. 10, p. 2761–2785, oct 2021.
[57] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2022.
[58] C. Cheng, T. Xu, and X.-J. Wu, "Mufusion: A general unsupervised image fusion network based on memory unit," *Information Fusion*, vol. 92, pp. 80–92, 2023.
[59] W. Wang, L.-J. Deng, and V. Vivone, "A general image fusion framework using multi-task semi-supervised learning," *Information Fusion*, vol. 108, p. 102414, 2024.
[60] P. Zhu, Y. Sun, B. Cao, and Q. Hu, "Task-customized mixture of adapters for general image fusion," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 7099–7108.
[61] Z. Zeng, D. Wang, F. Yang, H. Park, S. Soatto, D. Lao, and A. Wong, "Wordepth: Variational language prior for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 9708–9719.
[62] Z. Guo, B. Dong, Z. Ji, J. Bai, Y. Guo, and W. Zuo, "Texts as images in prompt tuning for multi-label image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 2808–2817.
[63] X. Li, Y. Huang, Z. He, Y. Wang, H. Lu, and M.-H. Yang, "Citetracker: Correlating image and text for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 10 2023, pp. 9940–9949.
[64] J. Guo, H. Manukyan, C. Yang, C. Wang, L. Khachatryan, S. Navasardyan, S. Song, H. Shi, and G. Huang, "Faceclip: Facial image-to-video translation via a brief text description," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 4270–4284, 2024.
[65] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.
[66] J.-H. Baek, D. Kim, S.-M. Choi, H.-J. Lee, H. Kim, and Y. J. Koh, "Luminance-aware color transform for multiple exposure correction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 6133–6142.
[67] K. Ma, Z. Duanmu, H. Zhu, Y. Fang, and Z. Wang, "Deep guided learning for fast multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 2808–2819, 2020.
[68] J. C. Hu, R. Cavicchioli, and A. Capotondi, "Exploiting multiple sequence lengths in fast end to end training for image captioning," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE Computer Society, 2023, pp. 2173–2182.
[69] J. Liu, G. Wu, J. Luan, Z. Jiang, R. Liu, and X. Fan, "Holoco: Holistic and local contrastive learning network for multi-exposure image fusion," *Information Fusion*, vol. 95, pp. 237–249, 2023.
[70] 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," *Information Fusion*, p. 102534, 2024.
[71] X. Zhang, "Benchmarking and comparing multi-exposure image fusion algorithms," *Information Fusion*, pp. 111–131, 2021.
[72] Y. Ju, J. Xiao, C. Zhang, H. Xie, A. Luo, H. Zhou, J. Dong, and A. C. Kot, "Towards marine snow removal with fusing fourier information," *Information Fusion*, vol. 117, p. 102810, 2025.

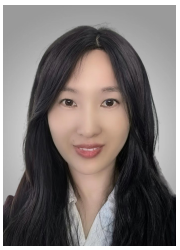
Enlong Wang received the B.S. degree in software engineering from Tiangong University, Tianjin, China, in 2022. He is currently working toward the M.S. degree in software engineering with the Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, School of Software Engineering, Dalian University, Dalian, China. His research interests include image fusion and image enhancement.

Jiawei Li received the M.S. degree in software engineering from Dalian University, Dalian, China, in 2023. He is currently pursuing the Ph.D. degree in computer science and technology with the School of Computer and Communication Engineering, University of the Science and Technology Beijing, Beijing, China. His research interests include image processing and adversarial robustness.

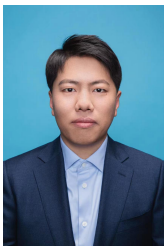
Tiantian Yan (Member, IEEE) received the Ph.D. degree in software technology from Dalian University of Technology in 2022. She is currently a lecturer with the National and Local Joint Engineering Laboratory of Computer Aided Design, School of Software Engineering, Dalian University, Dalian Liaoning. Her research interests include computer vision and deep learning.



Jia Lei received the M.S. degree in software engineering at Dalian University, Dalian, China, in 2024. She is currently pursuing the Ph.D. degree in computer science and technology at the School of Softerware Engineering, Xi'an Jiaotong University, China. Her current research interests include image processing and text-to-video generation.



Shihua Zhou (Member, IEEE) was born in Dalian, China, in 1982. She received the Ph.D. degree in mechanical design and theory from Dalian University of Technology, Dalian, China, in 2013. Since 2013, she has been working with Dalian University, where she is currently a Professor with the Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, School of Software Engineering. She is the author of more than 50 articles. Her research interests include deoxyribonucleic acid (DNA) computing, DNA self assembly, image encryption, and image fusion.



Bin Wang (Member, IEEE) received the B.S. degree in computer science and technology from Dalian University in June 2006 and the Ph.D. degree in mechanical design and theory from the Dalian University of Technology in October 2013. He is a professor with Dalian University. He has coauthored about 61 papers published. His research interests include intelligence computing, dna sequence design. DNA cryptography, and biological network.

Jinyuan Liu (Member, IEEE) received the Ph.D. degree in software engineering from Dalian University of Technology, Dalian, China, in 2022. He is currently an Assistant Research Fellow with the School of Mechanical Engineering, Dalian University of Technology. His research interests include computer vision, image processing, and deep learning.

Nikola K. Kasabov (Life Fellow, IEEE) received the Ph.D. degree from the Technical University of Sofia, Sofia, Bulgaria, in 1975. He is the Founding Director of Knowledge Engineering and Discovery Research Institute (KEDRI)Auckland. New Zealand. and a Professor of Knowledge Engineering with the School of Engineering Computing and Mathematical Sciences. Auckland University of Technology, Auckland. He holds the Professorial Chair position with the University of Ulster, Londonderry, UK and a wisiting professorship with theIICT. Bulgarian Academy of Sciences, Sofia, and Dalian University, Dalian, China. He has authored more than 700 articles. His research areas are computational intelligence, neuro informatics, knowledge discovery, and spiking neural networks.