TextMEF: Text-guided Prompt Learning for Multi-exposure Image Fusion

Jinyuan Liu[1], **Qianjun Huang**[2], **Guanyao Wu**[2], **Di Wang**[2], **Zhiying Jiang**[3], **Long Ma**[2], **Risheng Liu**[2] and **Xin Fan**[2*]

[1]School of Mechanical Engineering, Dalian University of Technology, China
[2]School of Software Technology, Dalian University of Technology, China
[3]College of Information Science and Technology, Dalian Maritime University,China
atlantis918@hotmail.com, hqj9994ever@gmail.com, xin.fan@dlut.edu.cn

Abstract

Multi-exposure image fusion (MEF) aims to integrate a set of low dynamic range images, producing a single image with a higher dynamic range than either one. Despite significant advancements, current MEF approaches still struggle to handle extremely over- or under-exposed conditions, resulting in unsatisfactory visual effects such as hallucinated details and distorted color tones. With this regard, we propose TextMEF, a prompt-driven fusion method enhanced by prompt learning, for multi-exposure image fusion. Specifically, we learn a set of prompts based on text-image similarity among negative and positive samples (over-exposed, under-exposed images, and well-exposed ones). These learned prompts are seamlessly integrated into the loss function, providing high-level guidance for constraining non-uniform exposure regions. Furthermore, we develop a attention Mamba module effectively translates over-/under- exposed regional features into exposure invariant space and ensure them to build efficient long-range dependency to high dynamic range image. Extensive experimental results on three publicly available benchmarks demonstrate that our TextMEF significantly outperforms state-of-the-art approaches in both visual inspection and objective analysis.

1 Introduction

Natural scenes exhibit a vast range of light intensities, ranging from intense sunlight to faint starlight, differing by up to 100 million orders of magnitude [McCann and Rizzi, 2011]. However, conventional photography equipment, such as mobile phones and SLR cameras, captures only a fraction of this dynamic range. This limitation results in low dynamic range (LDR) images that often exhibit over- or under-exposed areas, thereby failing to faithfully reproduce details visible to the human eyes under extreme lighting conditions. High dynamic range (HDR) imaging has emerged as a solution to this challenge, offering well-exposed images that enhance visual perception and support various computer vision tasks including super-resolution [Park *et al.*, 2003], panoramic photography [Inanici, 2006], object detection [Biswas and Milanfar, 2017], and remote sensing [Palsson *et al.*, 2017].

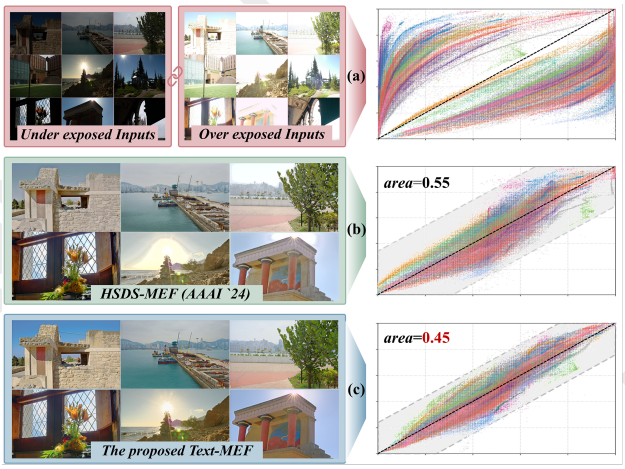

Figure 1: (a) is the input–ground truth luminance mapping curve on the SICEdataset, (b) and (c) are the generated results-ground truth luminance mapping curve of HSDS-MEF and our proposed TextMEF. A smaller area represents better results with fewer correction errors.

While specific hardware devices can directly produce HDR images [Nayar and Mitsunaga, 2000], their high production costs restrict widespread adoption in civil applications. Consequently, MEF methods have garnered significant attention. MEF aims to integrate multiple LDR images captured at different exposures into a unified HDR image. Depending on the number of input LDR images used, MEF methods are broadly categorized into non-extreme [Li *et al.*, 2013; Shen *et al.*, 2011; Li and Kang, 2012] and extreme exposure fusion [Ma *et al.*, 2017; Li *et al.*, 2020] techniques. Non-extreme methods rely on a larger set of input images to achieve optimal fusion performance, albeit at the cost of increased storage requirements and computational complexity. In contrast, our work focuses on extreme fusion techniques, which streamline the process by utilizing only a pair of extreme exposure images to achieve efficient fusion without compromising quality.

Recently, drawing on non-liner fitting ability of deep neural networks, a large numbers of learning-based MEF meth-

ods [Xu *et al.*, 2020b; Xu *et al.*, 2020a; Zhang and Ma, 2021; Xu *et al.*, 2023; Wu *et al.*, 2024] have been proposed and applied well. Despite great achievement, there are still come across several challenges that hinder further improvement of producing high-quality fused results. First, most of existing approaches cannot precisely perceive the bright and dark regions, causing over-enhancement in well-exposed regions or under-enhancement in low-light ones. Second, when designing neural networks, existing approaches often overlook subtle details. As a result, the integrated outputs frequently fail to accurately reproduce realistic textures, especially in excessively dark or bright regions.

To address these challenges, we propose TextMEF, a prompt guided multi-exposure fusion framework enhanced by learning-based prompts, which effectively merges multiple LDR images into an HDR result. Specifically, to mitigate color discrepancies caused by uneven exposures, we develop a new pipeline that adapts the Contrastive Language-Image Pre-Training (CLIP) [Radford *et al.*, 2021] model to our task. This approach refines the prompts by learning distinct feature distributions among over, under, and properly exposed images. The optimized prompts then direct the training of our multi-exposure fusion network through text-image similarity constraints in the CLIP embedding space. Given that network interference often results in the loss of critical information, we incorporate the advanced Mamba [Gu and Dao, 2023; Zhu *et al.*, 2024] block along with newly introduced attention Mamba and mutual guided Mamba, achieving enhanced contrast and detail in the resultant fused image.

Thanks to the meticulously designed modules and the effective prompt learning strategy outlined above, our TextMEF framework demonstrates superior performance across several widely used datasets. A straightforward example is given in Figure 1, where the results of our method not only enhance visual appeal but also deliver fused images with distinct targets and realistic details. The main contributions of this paper are summarized as follows:

- We leverage a pretrained CLIP model to provide cross-modal contrastive constraints from text to image, benefiting from its robust visual-language priors. Our method achieves a trade-off between metric-oriented and perceptual-oriented objectives.
- To fully unleash the potential of the CLIP priors, we employ prompt learning to discover more precise cues, thus providing richer and more detailed semantic guidance.
- We design the attention Mamba module and the mutual-guided Mamba module. The former helps capture prominent features and protect them under varying exposure levels, the latter promotes efficient feature fusion, enabling the model to autonomously learn the complementary information between different exposure inputs.
- Evaluations on three prevailing benchmark datasets and against seven state-of-the-art multi-exposure fusion approaches demonstrate that our proposed TextMEF can significantly enhance color representation while accurately recovering texture details.

2 Related Work

2.1 Learning-based MEF Approaches

In recent years, deep learning [Liu *et al.*, 2024a] has made significant advancements in the field of multi-exposure fusion. DeepFuse [Ram Prabhakar *et al.*, 2017], a pioneering approach, set a precedent by applying the feature processing capabilities of neural networks to MEF. Subsequent methods have predominantly focused on innovations in network architecture and improvements in loss functions. Architecturally, MEF-GAN [Xu *et al.*, 2020b] and AGAL [Liu *et al.*, 2022] introduced Generative Adversarial Networks (GANs), with the former leveraging GANs to effectively extract latent representations from reference images, while the latter employed global-local hierarchical constraints to enhance network learning. More recently, TransMEF [Qu *et al.*, 2022] combined CNN and Transformer modules to extract richer features through self-supervised multi-task learning, and SwinFusion [Ma *et al.*, 2022] utilized the advanced Swin-Transformer to achieve global information integration and detail preservation. Differently, CRMEF [Liu *et al.*, 2024c] employed neural architecture search to adaptively derive a super-network. In terms of loss functions, following the pioneering MEF-SSIM [Ma *et al.*, 2015], MEF-CL [Xu *et al.*, 2023] and HoLoCo [Liu *et al.*, 2023] introduced contrastive learning-based loss constraints, guiding the fusion process through distinct positive and negative sample settings. HSDS [Wu *et al.*, 2024], on the other hand, used an automated search method to construct loss functions without human intervention. Unfortunately, most current structural improvements tend to focus on network stacking rather than simplification [Liu *et al.*, 2024b], and the potential of cross-modal constraints in loss function improvements remains largely untapped.

2.2 Prompt Learning in Vision

Contrastive Language-Image Pre-Training (CLIP)[Radford *et al.*, 2021] achieved significant success in zero-shot prediction by leveraging large-scale image-text pairs. It has since been widely adapted for a range of visual tasks. By incorporating learnable prompt tokens, CLIP can internalize dataset-specific biases, enhancing its recognition ability for tasks such as object detection[Vidit *et al.*, 2023], style transformation [Kwon and Ye, 2022], and image enhancement [Liang *et al.*, 2023]. CLIP-based prompt learning typically adopts templates from natural language processing, which are fed into the text encoder, while image features are aligned with the textual prompts through the image encoder. Techniques such as StyleCLIP [Patashnik *et al.*, 2021] and StyleGAN [Gal *et al.*, 2022] combine the generative power of GANs with CLIP's semantic guidance, allowing images to be optimized toward target directions while being steered away from undesired features. However, existing prompt learning approaches have rarely addressed low-level vision tasks. In this work, we explore the potential of prompt learning for extracting more accurate exposure representations, enabling more effective multi-exposure image fusion.

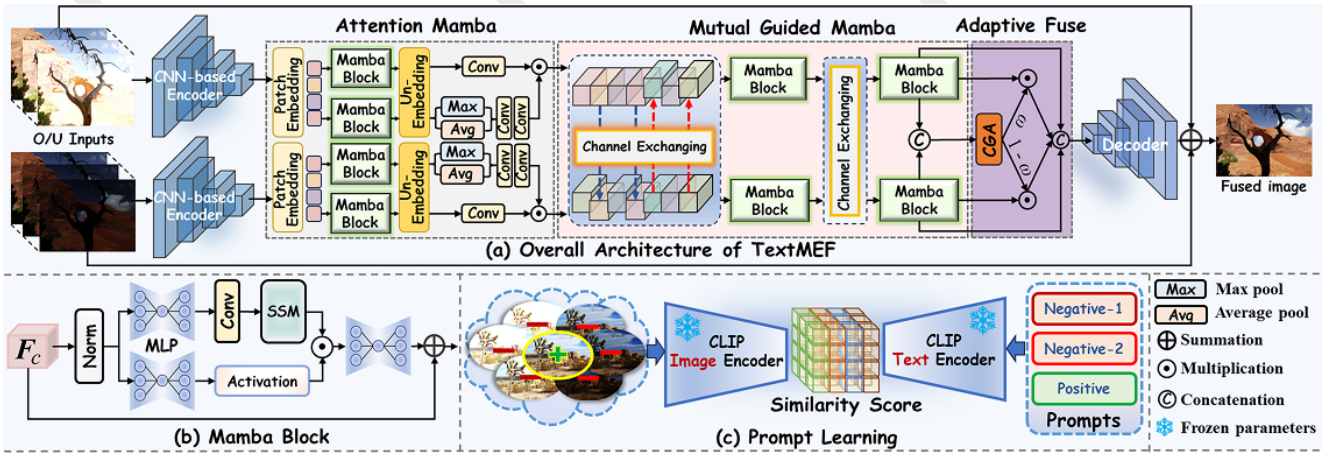

Figure 2: The workflow for our proposed TextMEF.

3 The Proposed Method

Our TextMEF is developed in two distinct phases. In the first phase, using CLIP priors, we initialize a set of image prompts derived from under-/over-exposed, and correctly exposed images. These prompts are then refined and fine-tuned to enable better semantic alignment. In the second phase, the learned prompts are fixed and used in conjunction with a contrastive CLIP perceptual loss, which guides the multi-exposure image fusion network based on Mamba architecture.

3.1 Prompt Learning

We start with a set of images, $\{\mathbf{I}_o, \mathbf{I}_u, \mathbf{I}_w\}_{i=1}^{M}$, where $\mathbf{I}_o$ represents over-exposed images, $\mathbf{I}_u$ represents under-exposed images, and $\mathbf{I}_w$ represents well-exposed images. The variable M denotes the total number of images in the dataset.

For these images, we initialize one positive prompt $\mathbf{T}_p \in \mathbb{R}^{N\times 512}$ and two negative prompts $\mathbf{T}_{n_1}$ and $\mathbf{T}_{n_2} \in \mathbb{R}^{N\times 512}$, where N refers to the number of embedded tokens in each prompt. These prompts are intended to represent the different lighting conditions of the images. The initialization is done by feeding the images into the image encoder of a pre-trained CLIP model to extract image features Φ_{image}, and the prompts are processed through the text encoder to generate their respective text features Φ_{text}.

For effective prompt learning, we leverage a multi-class cross-entropy loss function to fine-tune these prompts so that they properly differentiate between under-exposed, over-exposed, and well-exposed images inspired by [Liang *et al.*, 2023]. The multi-class cross-entropy loss function is formulated as:

$$\mathcal{L}_{prompt} = -\sum_{i=1}^{M}\sum_{c=1}^{3} y_{i,c}\log(\hat{y}_{i,c}), \tag{1}$$

where $\hat{y}_{i,c}$ is the predicted probability of the i-th sample belonging to class c, calculated as:

$$\hat{y}_{i,c} = \frac{e^{\cos(\Phi_{\text{image}}(\mathbf{I}_i),\Phi_{\text{text}}(\mathbf{T}_c))}}{\sum_{c'\in\{n_1,n_2,p\}} e^{\cos(\Phi_{\text{image}}(\mathbf{I}_i),\Phi_{\text{text}}(\mathbf{T}_{c'}))}}, \tag{2}$$

where the cosine similarity is used to measure the alignment between image features $\Phi_{\text{image}}(\mathbf{I}_i)$ and prompt features $\Phi_{\text{text}}(\mathbf{T}_c)$.

The labels $y_{i,c}$ are one-hot encoded, indicating which class the image belongs to. Specifically, $[1,0,0]$ corresponds to over-exposed images $\mathbf{I}_o$, $[0,1,0]$ for under-exposed images $\mathbf{I}_u$, and $[0,0,1]$ for well-exposed images $\mathbf{I}_w$.

3.2 Loss Function

Once the prompt learning phase is complete, the learned prompts can be used to guide the fusion network. The loss function for the multi-exposure image fusion task consists of two major components: Mean Squared Error (MSE) Loss and CLIP Perceptual Loss. These losses are designed to ensure that the fused image both retains pixel-level accuracy and aligns well with the learned semantic cues from the prompts.

CLIP Perceptual Loss. The purpose of the CLIP perceptual loss is to constrain the fused image to be semantically closer to the positive prompt $\mathbf{T}_p$ and farther from the negative prompts $\mathbf{T}_{n_1}$ and $\mathbf{T}_{n_2}$ in the CLIP feature space. The formula for this loss is as follows:

$$\mathcal{L}_{clip} = \frac{\sum_{c\in\{n_1,n_2\}} e^{\cos(\Phi_{\text{image}}(\mathbf{I}_i),\Phi_{\text{text}}(\mathbf{T}_c))}}{\sum_{c\in\{n_1,n_2,p\}} e^{\cos(\Phi_{\text{image}}(\mathbf{I}_i),\Phi_{\text{text}}(\mathbf{T}_c))}}. \tag{3}$$

This function uses cosine similarity to measure the alignment between the image features $\Phi_{\text{image}}(\mathbf{I}_i)$ and the text features of the prompts. The numerator calculates the distance between the image and the negative prompts, and the denominator normalizes the result by considering all prompts, including the positive one.

Mean Squared Error Loss. To minimize pixel-wise differences between the ground truth and the fused image, we employ the Mean Squared Error (MSE) Loss, calculated as:

$$\mathcal{L}_{MSE} = \|\mathbf{I}_{gt} - \mathbf{I}_f\|_2^2, \tag{4}$$

where $\mathbf{I}_{gt}$ is the ground truth image, and $\mathbf{I}_f$ is the fused image. The MSE loss ensures that the final fused image retains similar pixel intensities to the ground truth image, improving the quality of the fusion process.

Figure 3: Visual comparison of our method with seven state-of-the-art methods on SICE dataset. The signal maps demonstrate accurate pixel intensity between the ground truth and our method.

Total Loss. The total loss function combines the CLIP Perceptual Loss and the MSE Loss, weighted by a hyperparameter λ:

$$\mathcal{L}_{total} = \mathcal{L}_{MSE} + \lambda \mathcal{L}_{clip}, \tag{5}$$

where λ is a hyperparameter that controls the trade-off between the pixel-level similarity and the semantic alignment.

3.3 Network Architecture

To fully leverage the advantages of the State-Space Model [Gu and Dao, 2023], we have integrated the Mamba into our network. Specifically, our network architecture includes the Attention Mamba, Mutual-guided Mamba, and Adaptive Deep Fuse modules. The source image is first processed through a CNN to convert to feature space, then the Attention Mamba module is used to extract important features. Next, the Mutual-guided Mamba module performs preliminary feature fusion, and then, the Adaptive Deep Fuse module achieves adaptive information fusion. Finally, the fused features are passed through a convolutional layer and a skip connection from the source images to produce the final fused image. Each module will be introduced below.

Attention Mamba Block. Given the feature maps $\mathbf{F}_i^{\mathbf{I}_u}$ and $\mathbf{F}_i^{\mathbf{I}_o} \in \mathbb{R}^{B\times C\times H\times W}$ for under-exposed and over-exposed images, respectively, the Attention Mamba module contributes to compute attention weight maps to guide the extraction of more prominent features, thereby laying a foundation for more complementary and complete information fusion. The feature extraction process can be defined as follows:

$$\begin{aligned} \mathbf{F}_{i+1}^{\mathbf{I}_u} &= \mathcal{A}_u(\mathbf{F}_i^{\mathbf{I}_u}) \odot \mathbf{F}_i^{\mathbf{I}_u}, \\ \mathbf{F}_{i+1}^{\mathbf{I}_o} &= \mathcal{A}_o(\mathbf{F}_i^{\mathbf{I}_o}) \odot \mathbf{F}_i^{\mathbf{I}_o}, \end{aligned} \tag{6}$$

where $\mathcal{A}_u(\cdot)$ and $\mathcal{A}_o(\cdot)$ are the Attention Mamba Modules, and $\odot$ indicates point-wise multiplication.

Mutual-Guided Mamba Block. The mutual-guided Mamba block achieves lightweight information complementarity through a simple channel exchange method. Given the image features $\mathbf{F}_i^{\mathbf{I}_u} \in \mathbb{R}^{B\times N\times C}$ and $\mathbf{F}_i^{\mathbf{I}_o} \in \mathbb{R}^{B\times N\times C}$, the Mutual-Guided Mamba aims to achieve lightweight information complementarity and interaction through channel swapping. Specifically, Mutual-Guided Mamba splits the features along the channel dimension into two halves. The first half of the channels from $\mathbf{F}_i^{\mathbf{I}_o}$ is concatenated with the second half of the channels from $\mathbf{F}_i^{\mathbf{I}_u}$ to obtain $\mathbf{F}_{i+1}^{\mathbf{I}_u}$ and vice versa.

Adaptive Deep Fuse Block. The common element-wise addition fusion method overlooks the weights of the fused objects, which is detrimental to preserving prominent features. Inspired by [Chen *et al.*, 2024], we adopt a Content Guided Attention (CGA)-based Mixup Fusion scheme to form our adaptive deep fusion module. The CGA module provides self-learned channel and spatial attention weights to adaptively modulate the feature fusion. The detailed feature fusion process is defined as follows:

$$\omega = CGA(\mathbf{F}_i^{\mathbf{I}_u} + \mathbf{F}_i^{\mathbf{I}_o}), \tag{7}$$

$$\mathbf{F}_{fuse} = \mathcal{C}_{1\times 1}\left(\mathbf{F}_i^{\mathbf{I}_u} * \omega + \mathbf{F}_i^{\mathbf{I}_o} * (1-\omega) + \mathbf{F}_i^{\mathbf{I}_u} + \mathbf{F}_i^{\mathbf{I}_o}\right). \tag{8}$$

This final fused feature map $\mathbf{F}_{fuse}$ integrates the adaptive fusion process, where the weights ω control the contributions of the under-exposed and over-exposed images in the fusion process. The convolutional layer $\mathcal{C}_{1\times 1}$ is used to process the weighted fusion output.

In this formulation, we ensure that the network learns optimal fusion strategies by adapting the weight allocation dynamically, considering both channel and spatial attention.

4 Experiments

4.1 Implementation Details

We conduct experiments on the SICE [Cai *et al.*, 2018] dataset. The training process consists of prompt learning and

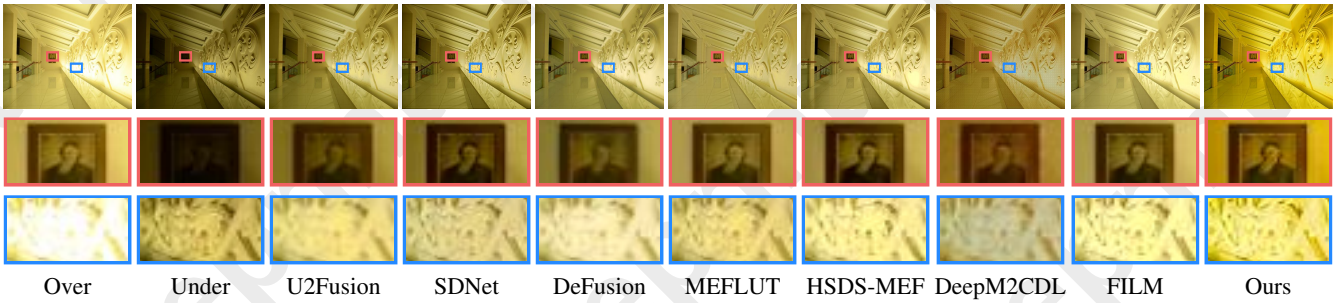

Figure 4: Visual comparison of our method with seven state-of-the-art methods on MEF dataset.

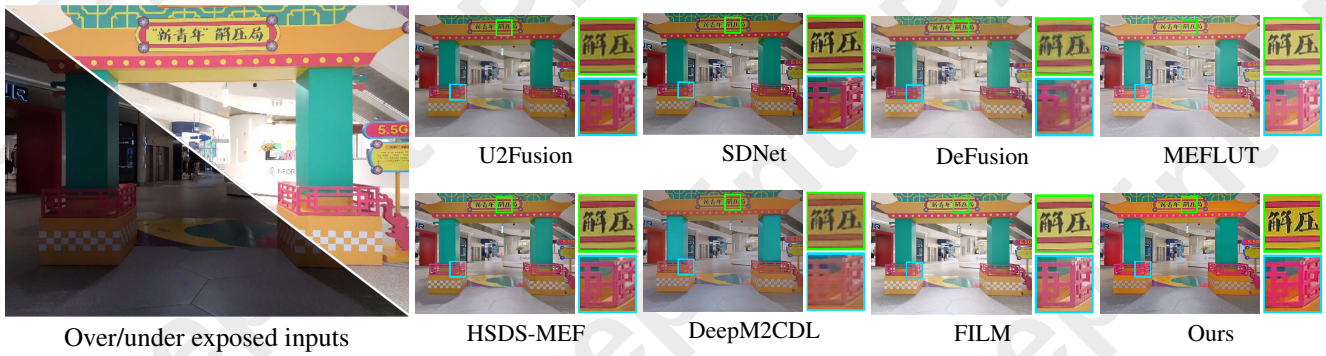

Figure 5: Visual comparison of our method with seven state-of-the-art methods on Mobile dataset.

image fusion stages. In the first stage, we employ 315 pairs of over/under-exposure images and 383 well-exposed images. In the second stage, 367 over/under-exposed image sequences with reference images are collected. To evaluate the performance, 100/26/30 image sequences from SICE [Cai *et al.*, 2018], Mobile [Jiang *et al.*, 2023], and MEF [Ma *et al.*, 2017] datasets are adopted for evaluation.

During the training process, all images are cropped to a size of 224×224, and data augmentations including flip, rotate, and zoom are used. The batch size and epochs for the prompt learning and image fusion are set to 16/8 and 160/200, respectively. We employ the Adam optimizer to guide parameter optimization. The learning rate for the prompt learning is set to $5e^{-5}$, while the initial learning rate for the image fusion is set to $2e^{-4}$, with a learning rate decay of 0.1 at epochs 100 and 135. The overall framework is implemented on Pytorch with an NVIDIA Tesla V100 GPU.

4.2 Comparison Methods & Evaluation Metrics

We compare our method with seven state-of-the-art deep learning-based competitors, *i.e.,* U2Fusion [Xu *et al.*, 2020a], SDNet [Zhang and Ma, 2021], DeFusion [Liang *et al.*, 2022], MEFLUT [Jiang *et al.*, 2023], HSDS-MEF [Wu *et al.*, 2024], DeepM2CDL [Deng *et al.*, 2023], and FILM [Zhao *et al.*, 2024]. All methods are tested with their official codes.

For quantitative analysis, a total of nine metrics are adopted, including three reference-based metrics, *i.e.,* PSNR [Huynh-Thu and Ghanbari, 2008], SSIM [WANGZ *et al.*, 2004], and TMQI [Yeganeh and Wang, 2012], and six no-reference metrics, *i.e.,* EN [Roberts *et al.*, 2008], AG [Eskicioglu and Fisher, 1995], EI [Rajalingam and Priya, 2018], SF [Eskicioglu and Fisher, 1995], MUSIQ [Ke *et al.*, 2021], and PaQ-2-PiQ [Ying *et al.*, 2020]. Among them, PSNR, SSIM, TMQI, EN, MUSIQ, and PaQ-2-PiQ are employed for quantitative comparisons on the SICE and Mobile datasets. For the MEF dataset, which lacks reference images, we utilize the aforementioned six no-reference metrics.

4.3 Qualitative Comparisons

We present a visual comparison on the SICE [Cai *et al.*, 2018] dataset in Figure 3. Certain methods, *e.g.,* U2Fusion and HSDS-MEF, suffer from significant detail loss in extremely under-exposed and over-exposed areas (see the paper in the red region of the first image sequence and the building in the blue region of the second image sequence). Other methods, *e.g.,* DeFusion, MEFLUT, DeepM2CDL, and FILM, exhibit noticeable artifacts (particularly evident in the unnatural colors of the clouds in the second image sequence). Thanks to the CLIP Perceptual Loss, our method excellently balances details and luminance, producing fused results that realistically simulate natural illumination. Moreover, in the RGB curves shown below, our method exhibits a distribution closest to the ground truth, demonstrating the outstanding color balance capability of our method.

Furthermore, we evaluate our method on the MEF [Ma *et al.*, 2017] and Mobile [Jiang *et al.*, 2023] datasets, with the comparison results shown in Figures 4 and 5. As illustrated in Figure 4, most methods suffer from blurred artifacts, leading to the loss of high-frequency information. Additionally, as seen in the second image sequence, many competing meth-

Dataset		SICE						Mobile					
Method	Source	EN	PSNR	SSIM	TMQI	MUSIQ	PaQ-2-PiQ	EN	PSNR	SSIM	TMQI	MUSIQ	PaQ-2-PiQ
U2Fusion	TPAMI'20	6.603	19.08	0.774	0.872	61.88	71.55	6.955	20.24	0.803	0.884	62.90	72.27
SDNet	IJCV'21	6.638	19.27	0.808	0.888	65.91	73.66	7.142	20.10	0.836	0.893	68.15	74.21
DeFusion	ECCV'22	6.648	12.79	0.718	0.736	59.78	71.96	7.076	17.07	0.802	0.838	64.95	73.09
MEFLUT	ICCV'23	6.983	14.32	0.754	0.766	60.71	72.06	7.231	17.73	0.824	0.836	66.56	73.49
HSDS-MEF	AAAI'24	7.039	20.20	0.862	0.908	67.09	73.66	7.232	20.96	0.833	0.893	66.83	73.71
DeepM2CDL	TPAMI'24	6.698	18.91	0.779	0.892	61.64	71.96	6.969	20.65	0.802	0.895	61.89	72.34
FILM	ICML'24	7.048	13.68	0.815	0.805	63.89	73.68	7.323	16.53	0.847	0.844	66.93	73.90
Ours	Proposed	7.268	22.15	0.895	0.926	68.64	75.10	7.290	21.97	0.848	0.910	67.28	74.18
Improvement		**0.220**	**1.950**	**0.033**	**0.018**	**1.550**	**1.420**	**-0.033**	**1.010**	**0.001**	**0.015**	**-0.870**	**-0.030**

Table 1: Qualitative comparison with other state-of-the-art methods. Red: the best; Blue: the 2nd best.

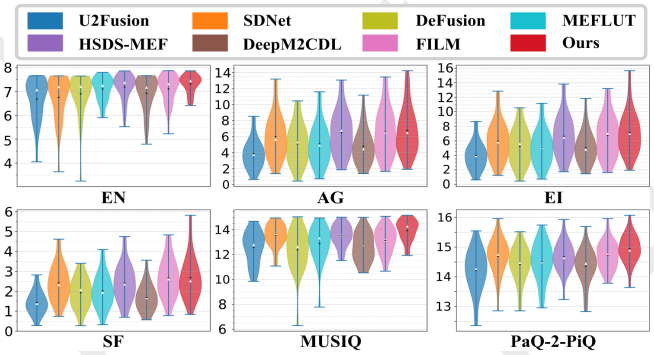

Figure 6: Quantitative comparison with seven state-of-the-art methods on the MEF dataset.

ods display dull colors. Empowered by the specially designed Mamba modules, our method excels at preserving edge information while maintaining good contrast. Figure 5 presents a similar situation to Figure 4, where our method demonstrates superior ability to restore texture details and vibrant colors compared to the competitors.

4.4 Quantitative Comparisons

Table 1 presents a qualitative comparison on the SICE and Mobile datasets. Our method achieved significantly higher performance on reference-based metrics compared to the competitors, demonstrating that the proposed method delivers image content structure and visual perception closest to the ground truth. Moreover, on the learning-based no-reference metrics, MUSIQ and PaQ-2-PiQ, our method achieves the best performance on the SICE dataset and came in second only to SDNet on the Mobile dataset, indicating that our method strongly aligns with the priors of high perceptual quality. Additionally, the quantitative comparison on the MEF dataset is shown in Figure 6. It is notable that our method significantly outperforms all other competitors across all metrics, demonstrating its ability to produce images with rich details, high edge strength and contrast, and alignment with human visual perception.

4.5 Ablation Studies

Study on Prompt Learning. We investigated the impact of prompt learning on the fusion stage using three variants for constructing $\mathcal{L}_{clip}$: fixed handcrafted prompts, a learnable pair of one positive and one negative prompt, and our method. Figure 7 and Table 2 shows that our learned prompts achieved the best performance qualitatively and quantitatively, demonstrating superior distinguishing capability.

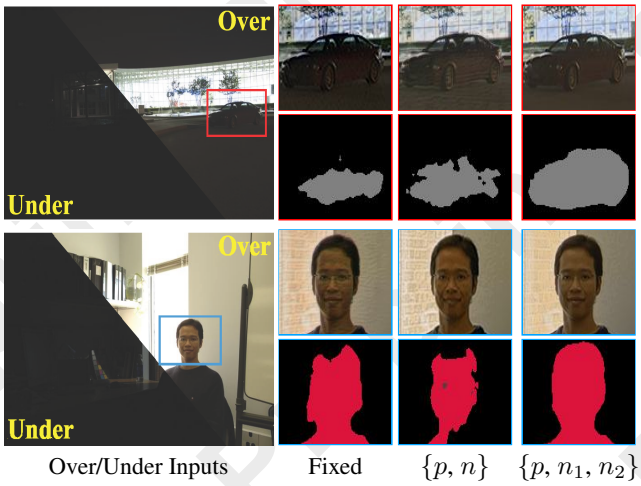

Figure 7: Visual ablation results of prompt learning.

Study on CLIP Perceptual Loss Weight. We conducted experiments on the weight of the CLIP perceptual loss. As shown in Figure 8, $\lambda = 0$ (*i.e., w/o* $\mathcal{L}_{clip}$) maintains overall proximity to the reference but struggles with varying brightness, leading to local detail loss (prominently shown on the building in the red rigion of the third image sequence). Higher weights (*i.e.,* $\lambda = 5e^{-1}$) enhance texture details but introduce unnatural colors and severe noise due to weakened pixel constraints (see the first and third sequences). At a weight of $1e^{-2}$, the loss function strikes an optimal balance between pixel-level and semantic constraints, resulting in visually appealing outcomes. The quantitative results in Table 2 further validate this fact.

Condition	Prompt Learning Ablation			CLIP Loss Weight Ablation				Network Architecture Ablation			
	Fixed	$\{p, n\}$	$\{p, n_1, n_2\}$	0	2e-3	1e-2	5e-1	*w/o* $\mathcal{A}$	*w/o* $\mathcal{M}$	*w/o* $\mathcal{D}$	Full Set
EN	**7.290**	7.286	7.268	7.254	7.261	**7.268**	7.207	7.237	7.185	7.252	**7.268**
PSNR	21.94	21.84	**22.15**	**22.29**	22.24	22.15	21.17	22.10	21.81	**22.21**	22.15
SSIM	0.862	0.861	**0.895**	**0.897**	**0.897**	0.895	0.797	0.888	0.876	0.890	**0.895**
TMQI	0.923	0.924	**0.926**	0.924	0.925	**0.926**	0.921	0.924	0.925	0.924	**0.926**
MUSIQ	66.32	**68.81**	68.64	67.60	68.54	**68.64**	66.03	**68.95**	68.35	68.24	68.64
PaQ-2-PiQ	73.19	75.08	**75.09**	74.74	74.99	**75.10**	73.89	**75.14**	74.46	75.08	75.09

Table 2: Quantitative results of ablation experiments on the SICE dataset. **Bold**: the best; underline: the 2nd best.

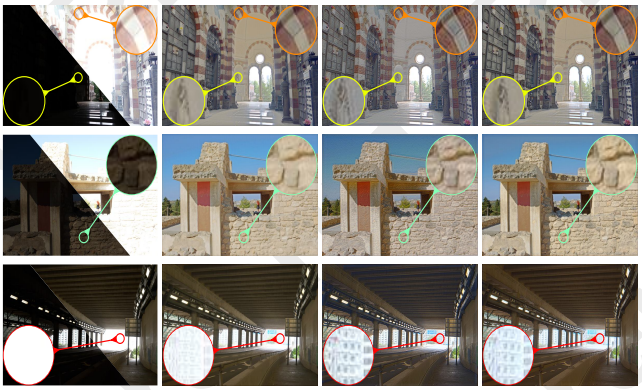

Figure 8: Visual ablation results of CLIP perceptual loss weight.

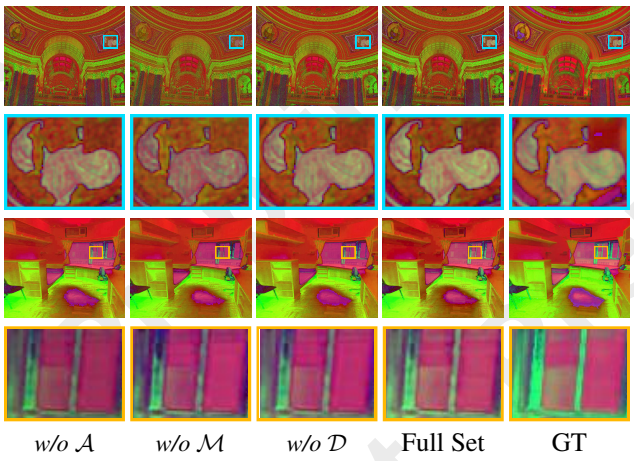

Figure 9: Visual ablation results of the network architecture. The images are converted to the HSV color space.

Study on Network Architecture. To validate the effectiveness of each network component, we constructed three variants: eliminating the Attention Mamba Block (*w/o* $\mathcal{A}$), the Mutual-guided Mamba Block (*w/o* $\mathcal{M}$), and the Adaptive Deep Fuse Block (*w/o* $\mathcal{D}$). As shown in the HSV images in Figure 9 and Table 2, *w/o* $\mathcal{M}$ significantly reduces the opportunity for image feature exchange, resulting in dull brightness with low contrast. *w/o* $\mathcal{A}$ and *w/o* $\mathcal{D}$ weakens the ability to extract important features and adaptively fuse them, leading to a drop in performance.

Extended Study. Figure 10 showcases experiments on fusing images with varying exposure ratios. The proposed method consistently produces robust and coherent results across diverse exposure conditions, highlighting the adaptability of the model.

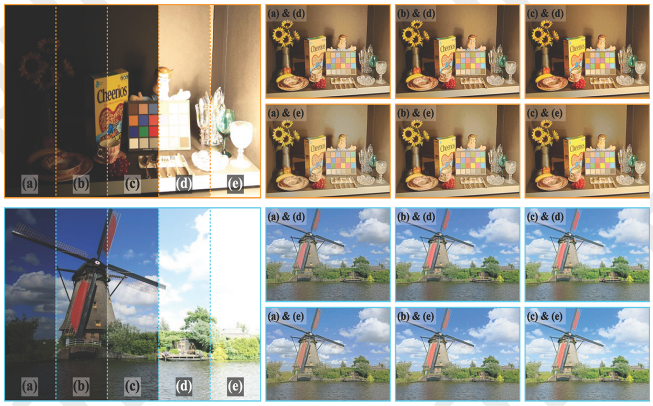

Figure 10: Visual results of our method to fuse images under different exposure ratios.

5 Conclusion

This paper propose a prompt-driven method enhanced by prompt learning for multi-exposure image fusion. By leveraging CLIP's rich priors and a prompt learning strategy, we generate precise prompts that effectively characterize different exposure images. Besides, we introduce two Mamba-based modules to further enhance the vivid color and accurate details of the fused images. This is the first study to use CLIP to investigate image fusion, and we anticipate this methodology will have broader applications in the future.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 62450072, U22B2052, 62302078), Central Guidance for Local Science and Technology Development Fund (Youth Science Fund Project, Category A, No. 2025JH6/101100001), the Distinguished Young Scholars Funds of Dalian (No. 2024RJ002), the China Postdoctoral Science Foundation (No. 2023M730741) and the Fundamental Research Funds for the Central Universities.

References

[Biswas and Milanfar, 2017] Sujoy Kumar Biswas and Peyman Milanfar. Linear support tensor machine with lsk channels: Pedestrian detection in thermal infrared images. *IEEE transactions on image processing*, 26(9):4229–4242, 2017.

[Cai *et al.*, 2018] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018.

[Chen *et al.*, 2024] Zixuan Chen, Zewei He, and Zhe-Ming Lu. Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*, 2024.

[Deng *et al.*, 2023] Xin Deng, Jingyi Xu, Fangyuan Gao, Xiancheng Sun, and Mai Xu. Deepm 2 cdl: Deep multi-scale multi-modal convolutional dictionary learning network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Eskicioglu and Fisher, 1995] Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on communications*, 43(12):2959–2965, 1995.

[Gal *et al.*, 2022] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[Huynh-Thu and Ghanbari, 2008] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[Inanici, 2006] Mehlika N Inanici. Evaluation of high dynamic range photography as a luminance data acquisition system. *Lighting Research & Technology*, 38(2):123–134, 2006.

[Jiang *et al.*, 2023] Ting Jiang, Chuan Wang, Xinpeng Li, Ru Li, Haoqiang Fan, and Shuaicheng Liu. Meflut: Unsupervised 1d lookup tables for multi-exposure image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10542–10551, 2023.

[Ke *et al.*, 2021] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

[Kwon and Ye, 2022] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022.

[Li and Kang, 2012] Shutao Li and Xudong Kang. Fast multi-exposure image fusion with median filter and recursive filter. *IEEE Transactions on Consumer Electronics*, 58(2):626–632, 2012.

[Li *et al.*, 2013] Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *IEEE Transactions on Image processing*, 22(7):2864–2875, 2013.

[Li *et al.*, 2020] Hui Li, Kede Ma, Hongwei Yong, and Lei Zhang. Fast multi-scale structural patch decomposition for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29:5805–5816, 2020.

[Liang *et al.*, 2022] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European Conference on Computer Vision*, pages 719–735. Springer, 2022.

[Liang *et al.*, 2023] Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8094–8103, 2023.

[Liu *et al.*, 2022] Jinyuan Liu, Jingjie Shang, Risheng Liu, and Xin Fan. Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5026–5040, 2022.

[Liu *et al.*, 2023] Jinyuan Liu, Guanyao Wu, Junsheng Luan, Zhiying Jiang, Risheng Liu, and Xin Fan. Holoco: Holistic and local contrastive learning network for multi-exposure image fusion. *Information Fusion*, 95:237–249, 2023.

[Liu *et al.*, 2024a] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132(5):1748–1775, 2024.

[Liu *et al.*, 2024b] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, and Xin Fan. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Liu *et al.*, 2024c] Zhu Liu, Jinyuan Liu, Guanyao Wu, Zihang Chen, Xin Fan, and Risheng Liu. Searching a compact architecture for robust multi-exposure image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Ma *et al.*, 2015] Kede Ma, Kai Zeng, and Zhou Wang. Perceptual quality assessment for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 24(11):3345–3356, 2015.

[Ma *et al.*, 2017] Kede Ma, Hui Li, Hongwei Yong, Zhou Wang, Deyu Meng, and Lei Zhang. Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Transactions on Image Processing*, 26(5):2519–2532, 2017.

[Ma *et al.*, 2022] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022.

[McCann and Rizzi, 2011] John J McCann and Alessandro Rizzi. *The art and science of HDR imaging*. John Wiley & Sons, 2011.

[Nayar and Mitsunaga, 2000] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 472–479. IEEE, 2000.

[Palsson *et al.*, 2017] Frosti Palsson, Johannes R Sveinsson, and Magnus O Ulfarsson. Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 14(5):639–643, 2017.

[Park *et al.*, 2003] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE signal processing magazine*, 20(3):21–36, 2003.

[Patashnik *et al.*, 2021] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021.

[Qu *et al.*, 2022] Linhao Qu, Shaolei Liu, Manning Wang, and Zhijian 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*, volume 36, pages 2126–2134, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Rajalingam and Priya, 2018] B Rajalingam and R Priya. Hybrid multimodality medical image fusion technique for feature enhancement in medical diagnosis. *International Journal of Engineering Science Invention*, 2(Special issue):52–60, 2018.

[Ram Prabhakar *et al.*, 2017] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proceedings of the IEEE international conference on computer vision*, pages 4714–4722, 2017.

[Roberts *et al.*, 2008] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008.

[Shen *et al.*, 2011] Rui Shen, Irene Cheng, Jianbo Shi, and Anup Basu. Generalized random walks for fusion of multi-exposure images. *IEEE Transactions on Image Processing*, 20(12):3634–3646, 2011.

[Vidit *et al.*, 2023] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3219–3229, 2023.

[WANGZ *et al.*, 2004] BOVIK AC WANGZ, HR Sheikh, et al. Image qualityassessment: Fromerrorvisibility-tostructural similarity. *IEEETransactionsonImageProcessing*, 13(4):600G612, 2004.

[Wu *et al.*, 2024] Guanyao Wu, Hongming Fu, Jinyuan Liu, Long Ma, Xin Fan, and Risheng Liu. Hybrid-supervised dual-search: Leveraging automatic learning for loss-free multi-exposure image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5985–5993, 2024.

[Xu *et al.*, 2020a] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2020.

[Xu *et al.*, 2020b] Han Xu, Jiayi Ma, and Xiao-Ping Zhang. Mef-gan: Multi-exposure image fusion via generative adversarial networks. *IEEE Transactions on Image Processing*, 29:7203–7216, 2020.

[Xu *et al.*, 2023] Han Xu, Liang Haochen, and Jiayi Ma. Unsupervised multi-exposure image fusion breaking exposure limits via contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3010–3017, 2023.

[Yeganeh and Wang, 2012] Hojatollah Yeganeh and Zhou Wang. Objective quality assessment of tone-mapped images. *IEEE Transactions on Image processing*, 22(2):657–667, 2012.

[Ying *et al.*, 2020] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3585, 2020.

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

[Zhao *et al.*, 2024] Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jiangshe Zhang, Peng Wang, et al. Image fusion via vision-language model. *arXiv preprint arXiv:2402.02235*, 2024.

[Zhu *et al.*, 2024] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.