

There and Back Again: A Flexible-Frame Transformer for Multi-Exposure Fusion

Lishen Qu^{1,3,5}, Yao Liu^{1,3}, Shihao Zhou^{1,3}

Jie Liang⁵, Hui Zeng⁵, Lei Zhang^{4,5}, Jufeng Yang^{1,2,3,*}

¹Nankai International Advanced Research Institute (SHENZHEN-FUTIAN)

²Peng Cheng Laboratory ³College of Computer Science, Nankai University

⁴The Hong Kong Polytechnic University ⁵OPPO Research Institute

{qulishen, liuyao, zhousihao96}@mail.nankai.edu.cn,

liang27jie@163.com, cshzeng@gmail.com, cslzhang@comp.polyu.edu.hk,

yangjufeng@nankai.edu.cn

Abstract. Multi-exposure fusion (MEF) brings the dynamic range of conventional cameras closer to that of human vision, producing images with rich scene content. Given the large variability in scene luminance, exposure strategies often require different numbers of frames to capture the full radiance range faithfully. However, conventional MEF techniques are typically designed for a fixed number of inputs, forcing deployment systems to maintain separate models for different frame-count requirements, which undermines deployment efficiency. To address this limitation, we propose FreeMEF, the first flexible-frame transformer for MEF that seamlessly accommodates varying numbers of input exposures without retraining or architectural changes. The proposed approach consists of two key modules. First, we introduce a recurrent state space module (RSSM) that sequentially fuses features from arbitrary sequences via adaptive alignment and state-space recurrent modeling, thereby providing global information guidance for the subsequent restoration. Second, we devise a global feature guided block (GFGB) incorporating an extremity-aware hybrid attention (EAHA) and an affine-injection feed-forward network (AFFN), which effectively resolves the similarity paradox while simultaneously optimizing contrast and brightness regulation. Extensive experiments on three benchmark datasets demonstrate the effectiveness of our method, which performs favorably against state-of-the-art methods both quantitatively and qualitatively. The code is available at <https://github.com/qulishen/FreeMEF>.

Keywords: HDR Imaging · Multi-Exposure Fusion · Recurrent Fusion Mechanism

1 Introduction

Conventional imaging systems [13] are fundamentally constrained by sensor capacity [48], making it difficult to capture the full dynamic range of natural scenes.

* Corresponding Author.

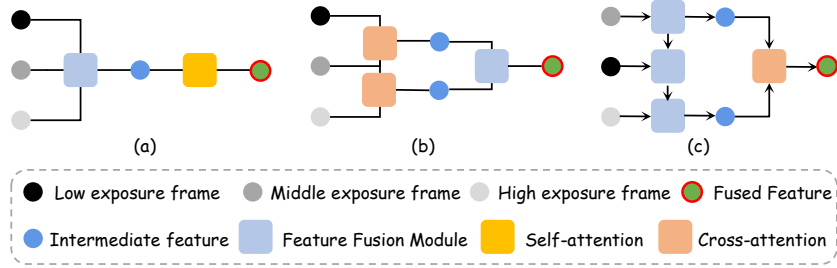


Fig. 1: Comparison of feature fusion pipelines. (a) **Fusion** $\rightarrow$ **Self-Attention**: Features from three frames are fused prior to self-attention. (b) **Cross-Attention** $\rightarrow$ **Fusion**: Reference frames interact with the base frame via cross-attention before fusion. (c) **Recurrent Fusion** $\rightarrow$ **Cross-Attention** (Ours): Features are fused sequentially, and the aggregated feature is then incorporated into the base frame via cross-attention.

As a result, high dynamic range (HDR) imaging [27, 38, 58] plays a vital role in faithfully capturing real-world scenarios with substantial illumination variations. Single-frame methods [2, 24, 30] attempt to enhance dynamic range by estimating and adjusting the brightness distribution from the individual image. Nevertheless, such approaches struggle to recover details in regions where information has already been lost due to saturation or underexposure. Therefore, multi-exposure fusion (MEF) [20] has become a widely adopted strategy, capturing multiple low dynamic range (LDR) images at different exposure levels and fusing them to reconstruct an HDR image that preserves more details. Recently, learning-based MEF approaches [29, 35, 42] have further advanced HDR imaging by leveraging deep neural networks to capture complex lighting conditions, non-linear camera responses, and dynamic scene content.

Despite their promising performance, these methods still suffer from two fundamental limitations that hinder practical application. First, existing methods universally assume that the camera uses a fixed exposure strategy. Therefore, these architectures are specifically designed for fixed exposure counts (e.g., 2, 3, or 5) and predefined exposure levels [8, 26, 29]. When the exposure strategy changes, the original model becomes inapplicable, requiring adjustments to the architecture and retraining. Second, many HDR imaging methods [35, 38, 59] employ attention mechanisms [41, 54] to capture inter-frame dependencies. Some of them [38, 39, 53] perform self-attention after feature fusion (Fig. 1 (a)), which is prone to introducing ghosting artifacts. Alternatively, others [35, 42, 59] apply cross-attention between reference and base frames before fusion (Fig. 1 (b)). While effective for fixed-length inputs, extending such pairwise interaction designs to variable-length exposure sequences is non-trivial and often requires architectural changes or repeated pairwise operations without a unified multi-reference aggregation. Furthermore, we uncover a limitation in conventional cross-attention, which we define as the *similarity paradox*. Standard cross-attention mechanisms inherently prioritize regions with high inter-frame similarity, which proves effective in tasks with consistent degradation [4, 11, 44]. How-

ever, HDR imaging demands the recovery of saturated (over-exposed) regions that exhibit extremely low pixel-wise similarity to their counterparts in normal-exposure frames [29]. This discrepancy hinders the model’s ability to effectively aggregate complementary information from neighboring frames.

To overcome these limitations, we propose FreeMEF, a flexible and robust multi-exposure fusion framework that supports arbitrary input sequence lengths during both training and inference. We reformulate the fusion paradigm by first aggregating multi-frame information into a global representation, and subsequently performing cross-attention with the base frame, as depicted in Fig. 1(c). On the one hand, we propose a recurrent state space module (RSSM), inspired by recurrent structures in video restoration [15, 28, 56] and understanding [22, 34, 55] tasks. Instead of constraining the network to a fixed number of inputs, we adopt a shared feature extraction module that processes each exposure frame independently into a latent feature space. These latent features are integrated via a recurrent mechanism that sequentially updates the aggregated representation with the current frame’s information. This design naturally accommodates arbitrary input lengths without requiring any architectural adjustments. On the other hand, we introduce a global feature guided block (GFGB) composed of an extremity-aware hybrid attention (EAHA) and an affine-injection feed-forward network (AFFN). To mitigate the similarity paradox, the EAHA utilizes a learned extremity map to adaptively balance the contributions of cross-attention and self-attention, ensuring precise feature retrieval. Subsequently, the AFFN refines the fused feature by rectifying brightness and spatial offsets via a gated affine transformation. Anchored to the base frame, these components selectively leverage global context from the RSSM to enhance restoration, thereby significantly reducing motion ghosting. Consequently, these designs allow FreeMEF to handle arbitrary input numbers with superior exposure fusion performance. In summary, our main contributions are summarized as follows:

- We propose FreeMEF, a flexible-frame transformer for multi-exposure fusion that conceptually decouples reference features from base features. This design overcomes the constraint of fixed-input architectures, allowing a single model to handle an arbitrary number of input frames during both training and inference without redesign or retraining.
- We propose RSSM, which progressively aggregates exposure features recurrently to accommodate variable input lengths, and GFGB, which exploits global context to guide the base frame, effectively resolving the similarity paradox in saturated regions and achieving precise brightness control.
- Comprehensive experiments demonstrate that our FreeMEF achieves superior performance compared to the state-of-the-art methods in both quantitative metrics and visual quality.

2 Related Work

HDR Imaging. HDR imaging methods are broadly divided into two categories: HDR reconstruction [27, 52, 53] and multi-exposure fusion (MEF) [33, 45, 58],

depending on whether fusion is performed in the linear domain or directly in the LDR domain. HDR reconstruction methods [53] typically take three bracketed exposures and aim to align all frames to a reference exposure (often the middle one) before merging. The reconstructed HDR output then requires a separate tone-mapping step [12, 49] for visualization on standard displays, which may introduce additional artifacts and errors. In contrast, MEF methods [8, 26, 37] fuse the input exposures directly in the LDR domain, avoiding camera response function (CRF) calibration and the subsequent tone-mapping stage. Considering the above, we choose to conduct the training and comparison of methods on MEF in this work.

Deep Learning-Based Methods. Early HDR methods [21, 46, 50] fused multiple LDR frames through joint optimization, but struggled to handle motion and complex textures. With the rapid development of deep learning [31, 32], learning-based methods have shown strong performance across a wide range of vision tasks, including detection [16], segmentation [43], and image restoration [10]. Kalantari *et al.* [27] constructed an HDR dataset containing motion and adopted an end-to-end pipeline that first aligns input LDR frames via optical flow [57] and then merges them into an HDR result. Kong *et al.* [29] explicitly learned masks for informative regions, concatenated the three frames, and then performed further refinement. Chen *et al.* [8] treated the MEF task as an inpainting problem, leveraging low-exposure frames to fill in over-exposed regions, which significantly reduces ghosting artifacts in dynamic scenes.

Multi-Frame Feature Fusion. Robust multi-frame feature fusion is crucial for HDR imaging and other burst image tasks [4, 20, 62]. Some methods [7, 11, 29] first align frames or features and then fuse them via feature concatenation followed by convolutional layers or attention. Others [6, 42, 53] bypass explicit alignment and directly fuse multi-exposure features, allowing the network to implicitly learn local and global motion from data. However, all of these methods assume a fixed number of input frames for the fusion process. In practical MEF applications, different devices, or even different modes of the same device, capture varying numbers of exposure brackets (commonly 2, 3, or 5) [47]. Explicit alignment methods [27, 29, 35] designed for three frames cannot be trained or evaluated on inputs with 2 or 5 frames, while alignment-free methods [18, 60, 61] still require modifications to the fusion layers (e.g., channel dimensions) and retraining when the number of input frames changes.

3 Preliminaries

Standard State Space Models. State space models (SSM) [17] map a 1D input stimulation $x(t) \in \mathbb{R}$ to an output response $y(t) \in \mathbb{R}$ through a latent state $h(t) \in \mathbb{R}^N$. This continuous system is defined by linear ordinary differential equations (ODEs) characterized by the evolution parameter $\mathbf{A}$, projection parameter $\mathbf{B}$, and output parameter $\mathbf{C}$:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad y(t) = \mathbf{C}h(t). \quad (1)$$

In modern deep learning implementations like Mamba [19, 40], this continuous system is transformed into a discrete form to operate on sampled data. Using the zero-order hold (ZOH) method with a timescale parameter Δ , the discrete parameters are derived as $\bar{\mathbf{A}} = \exp(\Delta\mathbf{A})$ and $\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}$. The discretized recurrence is formulated as:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \quad y_t = \mathbf{C}h_t. \quad (2)$$

Attentive State-Space Equation. While standard SSMs offer linear complexity, the fixed output matrix $\mathbf{C}$ restricts the model to a causal receptive field, potentially missing semantic correlations from unscanned regions. To overcome this, we adopt the attentive state-space equation (ASE) [18], which modifies the output transformation to facilitate global pixel querying. The core idea of ASE is to inject instance-specific semantics into the output matrix. This is achieved by constructing a prompt pool $\mathcal{P} \in \mathbb{R}^{N_p \times d}$, where N_p is the number of prompts and d is the hidden state dimension. To ensure parameter efficiency and interpretability, the pool is parameterized via semantic decoupling:

$$\mathcal{P} = \mathbf{M}\mathbf{N}, \quad \mathbf{M} \in \mathbb{R}^{N_p \times r}, \quad \mathbf{N} \in \mathbb{R}^{r \times d}, \quad (3)$$

where $\mathbf{N}$ is a basis matrix shared across blocks to learn common semantic features, and $\mathbf{M}$ is a block-specific coefficient matrix. r denotes the inner rank such that $r \ll \min\{N_p, d\}$. To apply these prompts spatially, a routing mechanism selects relevant prompts for each pixel. Given input features, the model predicts a routing probability distribution and employs the Gumbel-Softmax trick [25] to generate a differentiable one-hot routing matrix $\mathbf{R} \in \mathbb{R}^{L \times N_p}$, where L is the sequence length. The instance-specific prompt matrix $\mathbf{P} \in \mathbb{R}^{L \times d}$ is then obtained by $\mathbf{P} = \mathbf{R}\mathcal{P}$. Finally, the standard output equation is reformulated to include these learnable prompts, enabling the model to retrieve information from global semantic clusters dynamically:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \quad y_t = (\mathbf{C} + \mathbf{P})h_t + \mathbf{D}x_t. \quad (4)$$

By incorporating $\mathbf{P}$, the ASE mechanism effectively supplements the causal state history with non-causal global context, significantly enhancing the representation capability of the SSM.

4 Proposed Method

Given a set of LDR frames $\{\mathbf{I}_t\}_{t=0}^T$ captured at different exposure values, where $\mathbf{I}_0$ is designated as the base frame and the remaining frames $\{\mathbf{I}_t\}_{t=1}^T$ serve as auxiliary frames (i.e., T auxiliary frames and $T+1$ frames in total), our objective is to generate a fused ghost-free image $\hat{\mathbf{I}} \in \mathbb{R}^{H \times W \times 3}$ with enhanced dynamic range. Notably, unlike previous approaches [29, 35, 53], the number of auxiliary frames T need not be fixed and can vary across samples during both training and inference. The overall pipeline of our FreeMEF network is depicted in Fig. 2.

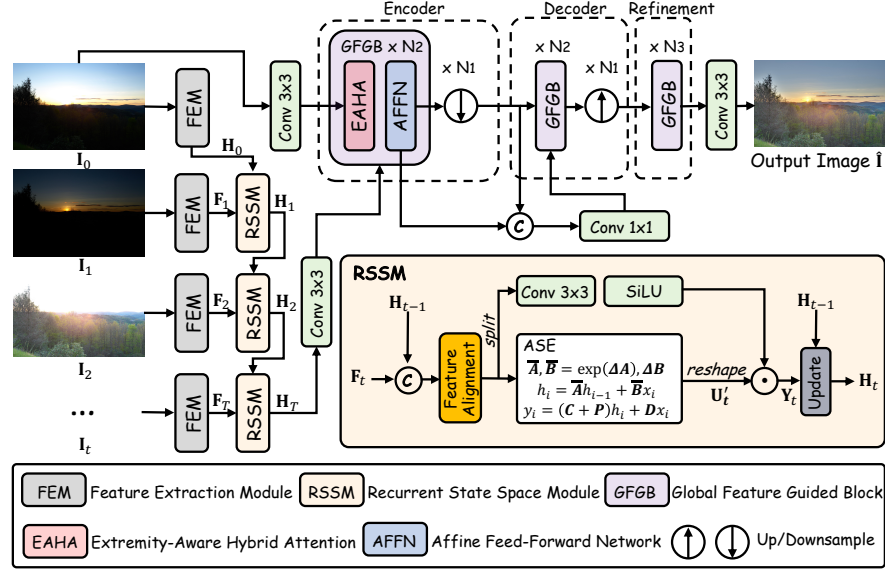


Fig. 2: Overview of our FreeMEF pipeline. Unlike previous methods [29, 35, 53] that process multiple frames in parallel, we design a recurrent fusion structure based on the ASE [18], which can adaptively exclude redundant features. Additionally, our model can be transferred to a different number of exposure levels without modification or retraining. The fused global features are then incorporated as auxiliary information to the base frame and fed into the subsequent U-shaped Transformer architecture.

First, a recurrent state space module (RSSM) aggregates the features of the base frame and all auxiliary frames into a single global fused feature $\mathbf{H}_T$. Second, a U-shaped Transformer-based architecture takes the base frame and the global fused feature as inputs and progressively refines the base feature at multiple scales using global feature guided blocks (GFGB). Specifically, the GFGB comprises two modules, namely the extremity-aware hybrid attention (EAHA) and the affine feed-forward network (AFFN).

4.1 Recurrent State Space Module

We first process the input frame $\mathbf{I}_t$ using the same lightweight feature extraction module (FEM) embedded within the recurrent cell. The shallow features $\mathbf{F}_t \in \mathbb{R}^{H \times W \times C}$ are obtained as:

$$\mathbf{F}_t = \text{FEM}(\mathbf{I}_t) = W_s^2(\text{ReLU}(W_s^1 \mathbf{I}_t)), \quad t \in \{0, 1, \dots, T\}, \quad (5)$$

where $W_s^{(\cdot)}$ denotes a standard 3×3 convolution. To effectively model temporal dependencies while handling an arbitrary number of input frames, we design a recurrent fusion scheme. Recently, state space models (SSMs) [18, 19, 40] have demonstrated significant potential in modeling long-range dependencies with



Fig. 3: The visualization of the *query* $\mathbf{Q}_{base}$ from the base feature and the *query* $\mathbf{Q}_{ref}$ from the reference feature. In the over-exposed red-boxed region, the values of $\mathbf{Q}_{base}$ are small, which leads to low attention map values when querying $\mathbf{K}$ in this region. As a result, $\mathbf{Q}_{ref}$ is required to provide supplementary information.

linear computational complexity. Inspired by this, we incorporate an attentive state-space equation (ASE) in [18] into our recurrent unit to capture global spatial-temporal contexts efficiently. Let $\mathbf{H}_t \in \mathbb{R}^{H \times W \times C}$ denote the hidden state accumulated from the t -th frame. We set $\mathbf{H}_0$ as an initial tensor which is equal to $\mathbf{F}_0$. For each step t , the module updates the hidden state by fusing the current feature $\mathbf{F}_t$ with the history state. First, we employ the deformable alignment strategy [9] to address the spatial misalignment caused by object motion between the history state and the current frame. We predict offsets based on the previous hidden state and the current feature, aligning $\mathbf{F}_t$ to $\mathbf{H}_{t-1}$:

$$\begin{aligned} \bar{\mathbf{F}}_t &= \text{DCN}(\mathbf{F}_t, \Delta \mathbf{P}_t), \quad \text{where} \\ \Delta \mathbf{P}_t &= \mathcal{F}_{\text{offset}}([\mathbf{H}_{t-1}, \mathbf{F}_t]) = W_s^4 \text{LeakyReLU}(W_s^3[\mathbf{H}_{t-1}, \mathbf{F}_t]), \end{aligned} \quad (6)$$

where $\text{DCN}(\cdot)$ is the deformable convolution. $[\cdot, \cdot]$ denotes concatenation along the channel dimension, and $\bar{\mathbf{F}}_t \in \mathbb{R}^{H \times W \times C}$ represents the aligned current feature. Then, the global features are captured using the ASE, and the output is obtained as $\mathbf{Y}_t$. Specifically, the input feature is projected into a higher-dimensional space and split into a data branch and a gate branch. This process can be represented by:

$$\begin{aligned} \mathbf{U}_t, \mathbf{Z}_t &= \text{Split}(\text{Linear}(\bar{\mathbf{F}}_t)), \quad \mathbf{U}'_t = \text{ASE}(\text{SiLU}(W_d^2(\mathbf{U}_t))), \\ \mathbf{Y}_t &= \mathbf{U}'_t \odot \text{SiLU}(W_d^1(\mathbf{Z}_t)), \end{aligned} \quad (7)$$

where $W_d^{(\cdot)}$ is the 3×3 depth-wise convolution. Then, we reshape $\mathbf{Y}'_t \in \mathbb{R}^{HW \times C}$ into $\mathbf{Y}_t \in \mathbb{R}^{H \times W \times C}$. Finally, the hidden state is updated by fusing $\mathbf{Y}_t$ with the previous state $\mathbf{H}_{t-1}$ via a sigmoid activation gate $\mathbf{G}_t = \sigma(W_s^5[\mathbf{Y}_t, \mathbf{H}_{t-1}])$, formulated as:

$$\mathbf{H}_t = \mathbf{H}_{t-1} + \mathbf{G}_t \odot (W_p^1([\mathbf{Y}_t, \mathbf{H}_{t-1}]) - \mathbf{H}_{t-1}), \quad (8)$$

where $\sigma(\cdot)$ is the sigmoid function and $W_p^{(\cdot)}$ denotes the 1×1 point-wise convolution.

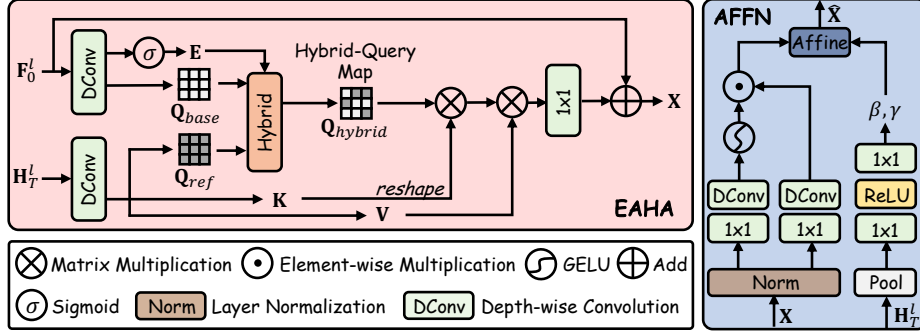


Fig. 4: Illustration of GFGB. It comprises an extremity-aware hybrid attention (EAHA) and an affine-injection feed-forward network (AFFN).

4.2 Global Feature Guided Block

The Similarity Paradox. Although the global feature $\mathbf{H}_T$ provides information about the reference frames, utilizing it to restore the base frame $\mathbf{I}_0$ introduces a fundamental conflict. The primary objective of HDR imaging is to hallucinate details in regions where the base frame is clipped (over-exposed) or noisy (under-exposed) [8, 29, 53]. However, standard attention mechanisms are based on similarity matching ($\mathbf{Q}$ vs. $\mathbf{K}$), where $\mathbf{Q}$ is extracted from the base features, and $\mathbf{K}$ is typically derived from the reference features. In over/under-exposed regions, the base features are numerically extreme and structurally degraded, leading to low correlation with the healthy global features, as shown in Fig. 3. Therefore, this leads to a similarity paradox: the attention mechanism assigns low weights to the regions that most require restoration. To resolve this, we propose the global feature guided block (GFGB), as shown in Fig. 4. It incorporates an extremity-aware hybrid attention (EAHA) that calibrates the query vectors based on the severity of saturation, ensuring robust feature query even when appearance similarity is low. Besides, we introduce affine transformations into the feed-forward network (FFN), and propose the affine-injection feed-forward network (AFFN), which leverages global features to adjust the brightness and contrast of the base frame.

Extremity-Aware Hybrid Attention (EAHA). Let $\mathbf{F}_0^l, \mathbf{H}_T^l \in \mathbb{R}^{\frac{H}{2^l} \times \frac{W}{2^l} \times 2^l C}$ denote the base frame feature and the global feature of the l -th stage, respectively. First, we generate the *key* $\mathbf{K}$ and *value* $\mathbf{V}$ from the global feature, and a primary *query* $\mathbf{Q}_{base}$ from the base frame:

$$\mathbf{Q}_{base} = W_d^Q W_p^Q \mathbf{F}_0^l, \quad \mathbf{K} = W_d^K W_p^K \mathbf{H}_T^l, \quad \mathbf{V} = W_d^V W_p^V \mathbf{H}_T^l. \quad (9)$$

Since $\mathbf{Q}_{base}$ is unreliable in saturated areas, we generate a reference *query* $\mathbf{Q}_{ref}$ derived directly from the global feature. This acts as a fallback search vector, allowing the model to perform self-correlation within the history when the base frame fails:

$$\mathbf{Q}_{ref} = W_d^{ref} W_q^{ref} \mathbf{H}_T^l. \quad (10)$$

To blend the two queries, we estimate an extremity map $\mathbf{E} \in \mathbb{R}^{\frac{H}{2^l} \times \frac{W}{2^l} \times 1}$ that highlights saturated pixels, where $\mathbf{E} = \sigma(W_d^E \mathbf{F}_0^l)$. We then synthesize a hybrid query $\mathbf{Q}_{hybrid}$ and the map $\mathbf{E}$ serves as a spatial gate:

$$\mathbf{Q}_{hybrid} = (1 - \mathbf{E}) \odot \mathbf{Q}_{base} + \mathbf{E} \odot \mathbf{Q}_{ref}. \quad (11)$$

Here, $\odot$ denotes broadcasting element-wise multiplication. In well-exposed regions ($\mathbf{E} \approx 0$), the model queries based on the base frame’s content. In saturated regions ($\mathbf{E} \approx 1$), the model switches to $\mathbf{Q}_{ref}$, effectively utilizing the history’s own structural priors to find relevant features. Finally, we compute the cross-attention along the channel dimension:

$$\text{Attention}(\hat{\mathbf{Q}}_{hybrid}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) = \hat{\mathbf{V}} \text{Softmax} \left(\frac{\hat{\mathbf{K}}^\top \hat{\mathbf{Q}}_{hybrid}}{\alpha} \right), \quad (12)$$

where $\hat{\mathbf{Q}}_{hybrid}, \hat{\mathbf{K}}, \hat{\mathbf{V}} \in \mathbb{R}^{\frac{HW}{2^l} \times 2^l C}$ are obtained by flattening the original tensors.

α is a learnable temperature scaling parameter. The resulting feature map captures global radiance information while being robust to local clipping. Essentially, $\hat{\mathbf{K}}^\top \hat{\mathbf{Q}}_{hybrid}$ is equivalent to a weighted sum of the self-attention map and the cross-attention map without additional computational overhead.

Affine-Injection Feed-Forward Network (AFFN). Recovering HDR content involves not only simple feature addition, but also addressing the domain shift between LDR and HDR by rectifying the luminance statistics (i.e., scale and shift) of the features [23, 36]. Standard FFNs lack the ability to dynamically adapt to these global exposure changes. We propose to inject global radiance information into the FFN via affine modulation. We extract a global descriptor vector from the history $\mathbf{H}_T^l$ and predict affine parameters $\gamma, \beta \in \mathbb{R}^{1 \times 1 \times C}$:

$$\gamma, \beta = \text{Split}(\text{Conv}(\text{GlobalAvgPool}(\mathbf{H}_T^l))). \quad (13)$$

Given an input tensor $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, AFFN is formulated as:

$$\begin{aligned} \hat{\mathbf{X}} &= W_p^2 \text{Gating}(\mathbf{X} \odot (1 + \gamma) + \beta), \\ \text{Gating}(\mathbf{X}) &= \text{GELU}(W_d^3 W_p^3 (\text{LN}(\mathbf{X}))) \odot W_d^4 W_p^4 (\text{LN}(\mathbf{X})). \end{aligned} \quad (14)$$

The $\text{LN}(\cdot)$ denotes the layer normalization [3]. By injecting (γ, β) , AFFN explicitly recalibrates the contrast and brightness of the local features based on the high-dynamic-range statistics preserved in the global history.

5 Experiments

5.1 Experimental settings

Datasets. We use Kalantari *et al.*’s dataset [27] and the Real-HDRV dataset [51] for training, in which each sample consists of three LDR frames and one HDR file. Following the approach in MEFLUT [26], the HDR files are tone-mapped and

Table 1: Quantitative experimental results of training and inference with 3 frames on Kalantari *et al.*'s dataset [27] and Real-HDRV dataset [51]. The best result is on **bold**, and the second best is underlined.

Method	Kalantari <i>et al.</i> [27]			Real-HDRV [51]			Flops (G)	Params (M)
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$		
HDR-Transformer [42] (ECCV'22)	27.024	<u>0.935</u>	0.093	25.131	<u>0.930</u>	<u>0.079</u>	95.261	1.454
Restormer [60] (CVPR'22)	27.149	0.930	0.097	<u>25.562</u>	0.924	0.089	141.160	26.099
MEFLUT [26] (ICCV'23)	19.362	0.747	0.270	16.392	0.655	0.405	33.060	0.356
SCTNet [53] (ICCV'23)	27.101	0.921	0.103	24.821	0.918	0.099	218.885	3.326
SAFNet [29] (ECCV'24)	24.284	0.927	0.094	21.467	0.908	0.111	42.589	1.121
ASTv2 [61] (TPAMI'25)	26.831	0.923	0.100	25.110	0.922	0.088	110.626	7.751
MambaIRv2 [18] (CVPR'25)	26.910	0.931	0.094	25.057	0.913	0.085	83.506	1.211
AFUNet [35] (ICCV'25)	<u>27.226</u>	0.925	<u>0.091</u>	25.423	0.912	0.084	75.340	1.138
FreeMEF (Ours)	28.418	0.948	0.081	26.077	0.940	0.074	41.496	8.900



Fig. 5: Visual comparisons on Kalantari *et al.*'s dataset [27] (left) and Real-HDRV dataset [51] (right). Our method achieves robust performance under both under-exposed and over-exposed conditions.

converted to PNG format for use in the MEF task. In addition, we conduct cross-dataset evaluation by training on Kalantari *et al.*'s dataset [27] and testing on the SICE dataset [5]. Specifically, since SICE includes multiple exposure levels, we evaluate on its 2-, 3-, and 5-exposure subsets to assess generalization across different frame numbers.

Baselines. We select the state-of-the-art HDR imaging methods for comparison, including MEFLUT [26], SAFNet [29], SCTNet [53], HDRTransformer [42], and AFUNet [35]. Since these methods are specifically tailored for three-frame inputs, we minimally adapt their input embedding layers for 2-frame and 5-frame settings to match the expected input dimensionality, while keeping the remaining architectures unchanged. Furthermore, we include representative models from the general image restoration field, such as Restormer [60], MambaIRv2 [18], and ASTv2 [61]. For fair comparison, these models are retrained to accommodate 2-frame and 5-frame inputs by adjusting the channel dimensions of their initial embedding layers. To train the 5-frame models, we expand Kalantari *et al.*'s dataset [27] from 3 to 5 frames through exposure simulation [1].

Table 2: Quantitative results of models trained on Kalantari *et al.*’s dataset [27] and evaluated on different exposure frames of SICE [26]. Note that previous models require modifications to their architectures and must be retrained separately on training sets corresponding to each frame number. In contrast, our FreeMEF requires no changes and can be directly tested with different frames.

Method	Test on SICE (2-frame)			Test on SICE (3-frame)			Test on SICE (5-frame)		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
HDR-Trans. [42] (ECCV’22)	15.339	0.690	0.262	18.151	0.746	0.227	20.748	0.830	0.171
Restormer [60] (CVPR’22)	14.492	0.674	0.267	17.880	0.734	0.228	21.091	0.837	0.167
MEFLUT [26] (ICCV’23)	13.415	0.535	0.301	15.747	0.621	0.298	16.214	0.645	0.280
SCTNet [53] (ICCV’23)	15.288	0.685	0.264	<u>18.724</u>	<u>0.768</u>	<u>0.210</u>	<u>21.153</u>	<u>0.842</u>	0.177
SAFNet [29] (ECCV’24)	15.161	0.627	0.277	16.947	0.738	0.223	18.798	0.820	0.170
ASTv2 [61] (TPAMI’25)	14.003	0.646	0.286	18.240	0.734	0.227	20.971	0.820	0.185
MambaIRv2 [18] (CVPR’25)	14.885	0.686	0.273	18.013	0.747	0.218	20.414	0.825	0.166
AFUNet [35] (ICCV’25)	14.591	0.647	0.290	18.669	0.758	0.218	21.121	0.838	<u>0.164</u>
FreeMEF (Ours)	17.087	0.731	0.225	19.326	0.774	0.199	22.269	0.851	0.149

Implementation Details. In our default setting, FreeMEF adopts a four-stage encoder-decoder architecture with an additional bottleneck stage. The channel dimension is set to $dim = 32$, and the numbers of blocks in the four stages are configured as $[2, 2, 2, 2]$, followed by 2 refinement blocks. During training, we crop image patches of size 256×256 and use a batch size of 2 per GPU, with geometric augmentations enabled. The model is optimized with the Adam optimizer, where the initial learning rate is set to 2×10^{-4} , the weight decay is 0, and the momentum parameters are $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is adjusted by a Cosine Annealing Restart Cyclic scheduler for a total of 300k iterations, consisting of two periods of 92k and 208k iterations, with minimum learning rates of 2.85×10^{-4} and 1×10^{-6} , respectively. Finally, the network is trained using the standard L_1 loss.

5.2 Comparison with Previous Work

We compare FreeMEF with 8 state-of-the-art methods, as shown in Tab. 1. On Kalantari *et al.*’s dataset [27], our method achieves a PSNR that is 1.192 dB higher than the second-best method, and on the Real-HDRV dataset [51], the improvement is 0.515 dB. Furthermore, as depicted in Fig. 5, our method demonstrates superior HDR performance. In the left image, our approach achieves significantly better restoration of backlit faces compared to previous methods. In the right image, previous methods introduce motion artifacts (e.g., ghosting) in the presence of moving subjects against over-exposed backgrounds, whereas our method effectively avoids such artifacts. Our method excels at maintaining low FLOPs, saving nearly 39% and 56% in computational cost relative to AFUNet [35] and HDR-Transformer [42], respectively. The number of parameters also falls within the acceptable range for deployment on mobile devices [14].

To further compare the models’ robustness, we conduct a cross-dataset evaluation. Specifically, all methods are trained on Kalantari *et al.*’s dataset [27], and



Fig. 6: Visual comparisons between models trained on Kalantari *et al.*'s dataset [27] and tested on the SICE dataset [5]. Under-exposed and over-exposed regions are marked with yellow and red boxes, respectively. Our method achieves a higher dynamic range and consistently delivers superior visual quality. Zoom in for a better view. Note that Kalantari *et al.*'s training dataset [27] contains fewer exposure levels than SICE [5], resulting in a lower dynamic range of all model outputs compared to the ground truth.

subsequently tested on the SICE dataset [5], which consists of exposure sequences with varying numbers of frames. Accordingly, we evaluate the models on three different SICE test subsets, each with a distinct number of LDR frames. The test results are shown in Tab. 2, where our FreeMEF outperforms the second-best method in terms of PSNR by 1.748, 0.657, and 1.116 dB for 2-frame, 3-frame, and 5-frame fusion, respectively. Our method demonstrates particularly significant improvements, especially in 2-frame and 5-frame fusion scenarios, which indicates better generalization to varying numbers of input frames at inference time.

We attribute this limitation to the two-frame training setting, where existing models are exposed only to long- and short-exposure images and thus fail to acquire prior knowledge of intermediate exposures. Conversely, FreeMEF benefits from training on multi-frame data, enabling it to generalize and exploit rich priors even during 2-frame inference. Regarding 5-frame fusion, fixed-input methods find it more challenging to compress all frame features into the base frame at the beginning of the network, resulting in greater information loss compared to 3-frame fusion. Our recurrent fusion architecture circumvents this bottleneck by processing features sequentially and decoupling the reference features from the base features, thereby preserving detail and significantly enhancing performance. Besides, we present visual comparison results on the highly challenging scenes of the SICE dataset [5], as shown in Fig. 6. As shown in rows 1–3 of Fig. 6, our FreeMEF demonstrates superior recovery ability and achieves a higher dynamic range, which can be attributed to the synergistic effect of RSSM and EAHA.

Table 3: Ablation studies of different components. The CNN, MDTA, and GDFN modules are drawn from Restormer [60], while the RSSM, EAHA, and AFFN modules are tailored for our FreeMEF. To ensure fairness in comparison, the parameter budgets of all modules are balanced.

	CNN	RSSM	MDTA	EAHA	GDFN	AFFN	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Flops (G)	Params (M)
(a)	✓		✓		✓		27.082	0.920	0.103	38.454	8.621
(b)		✓	✓		✓		27.565	0.937	0.090	39.315	8.633
(c)	✓			✓	✓		27.457	0.934	0.094	40.872	8.801
(d)	✓		✓			✓	27.426	0.928	0.085	38.236	8.632
(e)	✓			✓		✓	27.982	0.937	0.095	40.615	8.893
(f)		✓	✓			✓	27.837	0.938	0.089	39.117	8.639
(g)		✓		✓	✓		28.073	0.942	0.084	41.761	8.826
(h)		✓		✓		✓	28.418	0.948	0.081	41.496	8.900

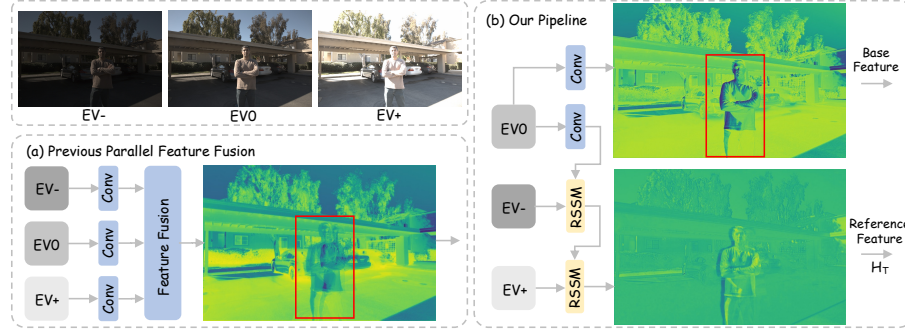


Fig. 7: Feature fusion strategy comparison. (a) The parallel fusion strategy treats all frame features equally, which can easily introduce ghosting artifacts. (b) Our method decouples the base frame from the reference features and employs a cross-attention mechanism for adaptive fusion. Unlike prior parallel fusion methods, our proposed "there and back again" mechanism effectively suppresses ghosting artifacts.

Furthermore, owing to the introduction of affine modulation in AFFN, FreeMEF achieves better contrast and color fidelity, as illustrated in row 4 of Fig. 6.

5.3 Analysis and Discussion

Effect of Individual Modules. For the ablation studies, we select three modules from Restormer [60] for comparison. Specifically, we replace the RSSM, EAHA, and AFFN modules in our FreeMEF with Restormer’s CNN embedding layer, MDTA, and GDFN, respectively, to validate the effectiveness of our proposed designs. As shown in Tab. 3 (a)-(d), RSSM, EAHA, and AFFN contribute individually to PSNR gains of 0.472 dB, 0.372 dB, and 0.347 dB, respectively. Besides, for the comparative experiments involving the removal of specific modules, i.e., (e)-(h), removing RSSM, EAHA, and AFFN results in performance drops of 0.526 dB, 0.581 dB, and 0.397 dB, respectively.

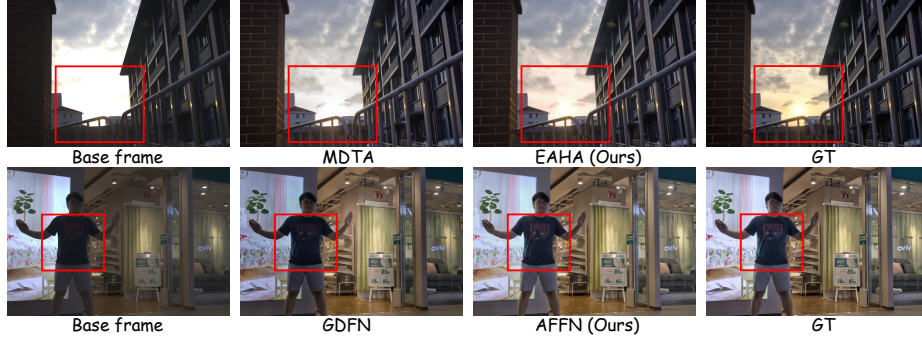


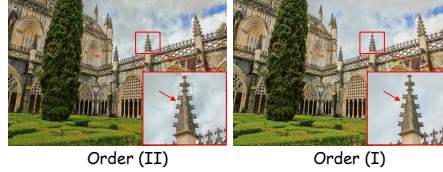
Fig. 8: Ablation study of EAHA and AFFN. Their synergy improves restoration in over-exposed regions and the performance of global brightness control.

Comparison of Fusion Mechanism. The key difference between our method and previous approaches lies in the fusion mechanism. We adopt a recurrent fusion strategy based on RSSM, whereas prior methods typically rely on parallel fusion. We analyze the differences between these two approaches at the feature level. As shown in Fig. 7 (a), the conventional parallel fusion method often leads to the introduction of severe ghosting artifacts in the fused features, which the subsequent self-attention and feed-forward networks struggle to mitigate effectively. In contrast, as shown in Fig. 7 (b), our pipeline anchors on the base frame and processes global reference information in parallel, which is subsequently fused in a selective manner within the Transformer. This paradigm reduces ghosting artifacts during exposure fusion. To intuitively illustrate the effect of individual modules, we also present qualitative comparisons with and without EAHA and AFFN, as shown in Fig. 8. Benefiting from EAHA’s extremity-aware query modulation, our model restores severely over-exposed regions more effectively. Moreover, through the explicit incorporation of affine modulation, AFFN utilizes multi-frame luminance cues to enable improved global brightness control.

Fusion Order. To assess the sensitivity of FreeMEF to fusion order, we perform a comparative study using 5-frame inference on the SICE dataset [5]. Specifically, we evaluate four different fusion strategies: (I) fusing frames in order from the darkest to the brightest exposures (EV-2, EV-1, EV+1, EV+2); (II) fusing from the brightest to the darkest exposures (EV+2, EV+1, EV-1, EV-2); (III) first fusing frames with similar exposures, followed by those with greater exposure gaps (EV-1, EV+1, EV-2, EV+2); and (IV) first fusing frames with the greatest exposure differences, then those with more similar exposures (EV-2, EV+2, EV-1, EV+1). As shown in Tab. 4, fusion orders (I), (III), and (IV) all achieve relatively robust performance, while order (II) exhibits a significant performance drop. Visualization results in Fig. 9 further reveal that positioning dark frames at the end of the fusion process tends to introduce more noise. Based on these findings, we suggest scheduling the fusion of long-exposure frames at the end of the pipeline to help optimize the final image quality.

Table 4: Quantitative results of different fusion orders on the SICE dataset [5].

Fusion Order	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
(I)	22.269	0.851	0.149
(II)	21.463	0.836	0.163
(III)	22.152	0.848	0.152
(IV)	22.181	0.849	0.151

**Fig. 9:** Visual comparison of fusion orders. (II) may introduce some noise.**Fig. 10:** Example of color shift after multi-exposure fusion.

6 Conclusion

In this paper, we propose FreeMEF, a flexible and robust transformer framework to tackle the inherent limitations of fixed-input architectures and the similarity paradox in multi-exposure fusion (MEF). Specifically, to break the constraint of fixed exposure counts, we developed a recurrent state space module (RSSM) that progressively aggregates features from an arbitrary number of input frames. This design endowed the model with the capability to gracefully handle diverse exposure configurations in real-world scenarios without retraining or architectural modification. Furthermore, to overcome the insufficient utilization of reference information in saturated regions, we decoupled the base frame from the reference frames and introduced a global feature guided block (GFGB) for selective fusion. Within this block, an extremity-aware hybrid attention (EAHA) adaptively balanced cross-attention and self-attention to ensure robust feature retrieval, while an affine-injection feed-forward network (AFFN) dynamically optimized the brightness and contrast of the fused features based on the global context. Extensive experiments demonstrate the effectiveness of our method.

Limitations. As shown in Fig. 10, the fused result of FreeMEF exhibits slight color deviations compared to the ground truth. This issue primarily stems from overly large gaps between exposure levels, where the base frame is too dark and the long-exposure frame is overly saturated, leading to colors that deviate from the true scene appearance.

Acknowledgement. This work was supported by the Shenzhen Science and Technology Program (No. JCYJ20240813114229039), the Natural Science Foundation of Tianjin, China (No. 24JCZXJC00040), the National Natural Science Foundation of China (No. 624B2072), the Doctoral Student Program of the Young S&T Talents Cultivation Project, CAST, the Fundamental Research Funds for the Central Universities (No. 63263253), the Supercomputing Center of Nankai University (NKSC) and OPPO Research Fund.

References

1. Affi, M., Derpanis, K.G., Ommer, B., Brown, M.S.: Learning multi-scale photo exposure correction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9157–9167 (2021)
2. Akyüz, A.O., Fleming, R., Riecke, B.E., Reinhard, E., Bühlhoff, H.H.: Do HDR displays support LDR content? A psychophysical evaluation. *ACM Transactions on Graphics (TOG)*. **26**(3), 38-es (2007)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. *arXiv preprint arXiv:1607.06450* (2016)
4. Bhat, G., Danelljan, M., Van Gool, L., Timofte, R.: Deep burst super-resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9209–9218 (2021)
5. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing (TIP)*. **27**(4), 2049–2062 (2018)
6. Chen, J., Yang, Z., Chan, T.N., Li, H., Hou, J., Chau, L.P.: Attention-guided progressive neural texture fusion for high dynamic range image restoration. *IEEE Transactions on Image Processing (TIP)*. **31**, 2661–2672 (2022)
7. Chen, R., Zheng, B., Zhang, H., Chen, Q., Yan, C., Slabaugh, G., Yuan, S.: Improving dynamic hdr imaging with fusion transformer. In: *AAAI Conference on Artificial Intelligence (AAAI)*. pp. 340–349 (2023)
8. Chen, Z., Wang, Y., Cai, X., You, Z., Lu, Z., Zhang, F., Guo, S., Xue, T.: Ultrafusion: Ultra high dynamic imaging using exposure fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16111–16121 (2025)
9. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 764–773 (2017)
10. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: *European Conference on Computer Vision (ECCV)*. pp. 184–199 (2014)
11. Dudhane, A., Zamir, S.W., Khan, S., Khan, F.S., Yang, M.H.: Burstformer: Burst image restoration and enhancement transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5703–5712 (2023)
12. Durand, F., Dorsey, J.: Fast bilateral filtering for the display of high-dynamic-range images. In: *ACM SIGGRAPH Annual Conference (SIGGRAPH)*. pp. 257–266 (2002)
13. Fossum, E.R.: Digital camera system on a chip. *IEEE Micro* **18**(03), 8–15 (1998)
14. Gao, T., Guo, L., Zhao, S., Xu, P., Yang, Y., Liu, X., Wang, S., Zhu, S., Zhou, D.: Quantnas: Quantization-aware neural architecture search for efficient deployment on mobile device. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1704–1713 (2024)
15. Ghasemabadi, A., Janjua, M.K., Salameh, M., Niu, D.: Learning truncated causal history model for video restoration. In: *Advances in Neural Information Processing Systems (NeurIPS)*. pp. 27584–27615 (2024)
16. Girshick, R.: Fast r-cnn. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1440–1448 (2015)
17. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: *Conference on Language Modeling (COLM)* (2024)

18. Guo, H., Guo, Y., Zha, Y., Zhang, Y., Li, W., Dai, T., Xia, S.T., Li, Y.: Mambairv2: Attentive state space restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28124–28133 (2025)
19. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.T.: Mambair: A simple baseline for image restoration with state-space model. In: European Conference on Computer Vision (ECCV). pp. 222–241 (2024)
20. Hasinoff, S.W., Sharlet, D., Geiss, R., Adams, A., Barron, J.T., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (TOG)*. **35**(6), 1–12 (2016)
21. Hu, J., Gallo, O., Pulli, K., Sun, X.: HDR deghosting: How to deal with saturation? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1163–1170 (2013)
22. Huang, D.A., Ramanathan, V., Mahajan, D., Torresani, L., Paluri, M., Fei-Fei, L., Niebles, J.C.: What makes a video a video: Analyzing temporal information in video understanding models and datasets. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7366–7375 (2018)
23. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1501–1510 (2017)
24. Huo, Y., Yang, F., Dong, L., Brost, V.: Physiological inverse tone mapping based on retina response. *The Visual Computer* **30**(5), 507–517 (2014)
25. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016)
26. Jiang, T., Wang, C., Li, X., Li, R., Fan, H., Liu, S.: Meflut: Unsupervised 1d lookup tables for multi-exposure image fusion. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10542–10551 (2023)
27. Kalantari, N.K., Ramamoorthi, R., et al.: Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (TOG)*. **36**(4), 144–155 (2017)
28. Kim, T.H., Sajjadi, M.S., Hirsch, M., Scholkopf, B.: Spatio-temporal transformer network for video restoration. In: European Conference on Computer Vision (ECCV). pp. 106–122 (2018)
29. Kong, L., Li, B., Xiong, Y., Zhang, H., Gu, H., Chen, J.: Safnet: Selective alignment fusion network for efficient hdr imaging. In: European Conference on Computer Vision (ECCV). pp. 256–273 (2024)
30. Kovaleski, R.P., Oliveira, M.M.: High-quality reverse tone mapping for a wide range of exposures. In: Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI). pp. 49–56 (2014)
31. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS)*. (2012)
32. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature*. **521**(7553), 436–444 (2015)
33. Li, H., Ma, K., Yong, H., Zhang, L.: Fast multi-scale structural patch decomposition for multi-exposure image fusion. *IEEE Transactions on Image Processing (TIP)*. **29**, 5805–5816 (2020)
34. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., Qiao, Y.: Uniformerv2: Unlocking the potential of image vits for video understanding. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1632–1643 (2023)
35. Li, X., Ni, Z., Yang, W.: Afunet: Cross-iterative alignment-fusion synergy for hdr reconstruction via deep unfolding paradigm. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10666–10675 (2025)

36. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: *Advances in Neural Information Processing Systems (NeurIPS)*. (2017)
37. Liang, P., Jiang, J., Ma, Q., Liu, X., Ma, J.: Fusion from decomposition: A self-supervised approach for image fusion and beyond. *arXiv preprint arXiv:2410.12274* (2024)
38. Liu, S., Zhang, X., Sun, L., Liang, Z., Zeng, H., Zhang, L.: Joint hdr denoising and fusion: A real-world mobile hdr image dataset. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 13966–13975 (2023)
39. Liu, X., Li, A., Wu, Z., Du, Y., Zhang, L., Zhang, Y., Timofte, R., Zhu, C.: Pasta: Towards flexible and efficient hdr imaging via progressively aggregated spatio-temporal alignment. *arXiv preprint arXiv:2403.10376* (2024)
40. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. In: *Advances in Neural Information Processing Systems (NeurIPS)*. pp. 103031–103063 (2024)
41. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 10012–10022 (2021)
42. Liu, Z., Wang, Y., Zeng, B., Liu, S.: Ghost-free high dynamic range imaging with context-aware transformer. In: *European Conference on Computer Vision (ECCV)*. pp. 344–360 (2022)
43. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3431–3440 (2015)
44. Luo, Z., Li, Y., Cheng, S., Yu, L., Wu, Q., Wen, Z., Fan, H., Sun, J., Liu, S.: Bsrt: Improving burst super-resolution with swin transformer and flow-guided deformable alignment. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 998–1008 (2022)
45. Ma, K., Li, H., Yong, H., Wang, Z., Meng, D., Zhang, L.: Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Transactions on Image Processing (TIP)*. **26**(5), 2519–2532 (2017)
46. Oh, T.H., Lee, J.Y., Tai, Y.W., Kweon, I.S.: Robust high dynamic range imaging by rank minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. **37**(6), 1219–1232 (2014)
47. Pourreza-Shahri, R., Kehtarnavaz, N.: Exposure bracketing via automatic exposure selection. In: *IEEE International Conference on Image Processing (ICIP)*. pp. 320–323 (2015)
48. Rachlin, Y., Negi, R., Khosla, P.K.: The sensing capacity of sensor networks. *IEEE Transactions on Information Theory (TIT)*. **57**(3), 1675–1691 (2011)
49. Reinhard, E., Stark, M., Shirley, P., Ferwerda, J.: Photographic tone reproduction for digital images. *ACM Transactions on Graphics (TOG)*. **21**(3), 267–276 (2002)
50. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shechtman, E.: Robust patch-based hdr reconstruction of dynamic scenes. *ACM Transactions on Graphics (TOG)*. **31**(6), 203–1 (2012)
51. Shu, Y., Shen, L., Hu, X., Li, M., Zhou, Z.: Towards real-world hdr video reconstruction: A large-scale benchmark dataset and a two-stage alignment network. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2879–2888 (2024)
52. Song, J.W., Park, Y.I., Kong, K., Kwak, J., Kang, S.J.: Selective transhdr: Transformer-based selective hdr imaging using ghost region mask. In: *European Conference on Computer Vision (ECCV)*. pp. 288–304 (2022)

53. Tel, S., Wu, Z., Zhang, Y., Heyrman, B., Demonceaux, C., Timofte, R., Ginhac, D.: Alignment-free hdr deghosting with semantics consistent transformer. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12836–12845 (2023)
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). (2017)
55. Wang, J., Chen, D., Luo, C., He, B., Yuan, L., Wu, Z., Jiang, Y.G.: Omnivid: A generative framework for universal video understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18209–18220 (2024)
56. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1954–1963 (2019)
57. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: European Conference on Computer Vision (ECCV). pp. 117–132 (2018)
58. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. **44**(1), 502–518 (2020)
59. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1751–1760 (2019)
60. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5728–5739 (2022)
61. Zhou, S., Pan, J., Yang, J.: Learning an adaptive sparse transformer for efficient image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. **47**(11), 10344–10360 (2025)
62. Zhu, P., Sun, Y., Cao, B., Hu, Q.: Task-customized mixture of adapters for general image fufsion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7099–7108 (2024)