ExpoMotion: A Large-Scale Benchmark and A Householder Projection Network for Multi-Exposure Fusion

Yao Liu[1,3,5⋆], Lishen Qu[1,3,5⋆], Shihao Zhou[3], Jie Liang[5], Hui Zeng[5], Yabin Peng[3], Huipeng Lin[3], Lei Zhang[4], and Jufeng Yang[1,2,3⋆⋆]

[1] Nankai International Advanced Research Institute (SHENZHEN·FUTIAN)
[2] Pengcheng Laboratory
[3] College of Computer Science, Nankai University
[4] The Hong Kong Polytechnic University
[5] OPPO Research Institute
liuyao@mail.nankai.edu.cn, yangjufeng@nankai.edu.cn

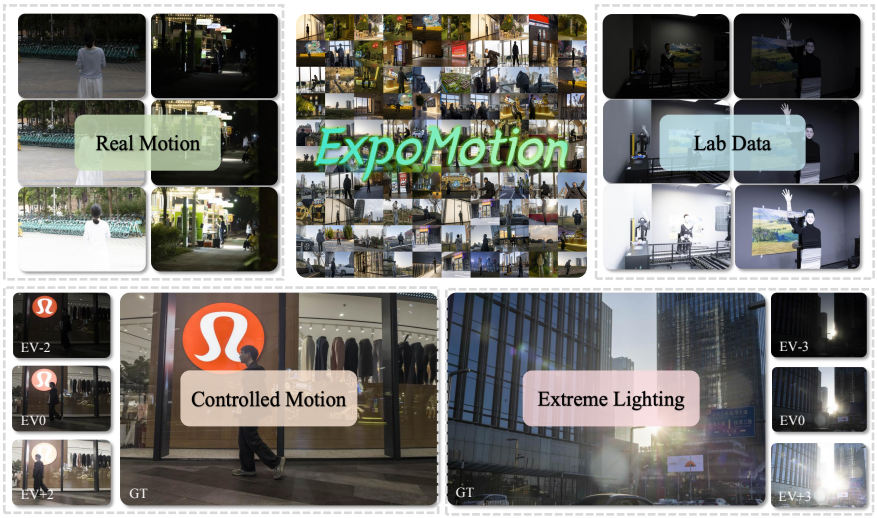

Fig. 1: Overview of the proposed large-scale MEF dataset. Deviating from the conventional focus on perceptual quality in static scenes, our work introduces a large-scale MEF dataset dedicated to simultaneous detail restoration and motion artifact suppression. We integrate data from controlled and real-world motion, extreme lighting environments, and laboratory setups.

Abstract. Multi-Exposure Fusion (MEF) effectively extends dynamic range, but practical deployment is hindered by motion-induced ghosting and the scarcity of high-quality dynamic benchmarks. Current benchmarks largely neglect dynamic scenes and lack reliable ground truth, making it difficult to handle the complexity of real-world motions. In response, we introduce ExpoMotion, a large-scale benchmark designed to

⋆ Both authors contributed equally to this research.
⋆⋆ Corresponding author.

evaluate deghosting capabilities. Comprising 1,738 sequences and 10,909 images across diverse environments, it covers a wide range of motions and provides high-fidelity GTs constructed through an expert-guided acquisition pipeline. To tackle the complex dynamics and extreme conditions captured in this benchmark, we propose the Householder Orthogonal Projection network (HOP), which revisits MEF deghosting from a mathematical perspective via Householder transformation, decoupling multi-frame alignment into exposure pre-alignment and ghost filtering. Specifically, the Global Priors Illumination Alignment (GPIA) module first rectifies drastic dynamic range discrepancies by utilizing global statistics for exposure harmonization. Regarding ghost removal, our Householder Orthogonal Attention (HOA) models artifacts as orthogonal perturbations. By employing a dynamic Householder reflector, HOA effectively projects ghosts out of the feature manifold while preserving high-frequency details. Experiments demonstrate that our ExpoMotion dataset enables superior generalization and artifact-free detail restoration, while also validating the effectiveness and efficiency of the HOP method. The dataset and code are available at https://github.com/Leo-LiuYao/ExpoMotion.

Keywords: Multi-Exposure Fusion · Dynamic Range · Deghosting

1 Introduction

Although real-world scenes exhibit vast dynamic ranges spanning over 10 orders of magnitude, ranging from starlight to direct sunlight, standard imaging sensors possess a limited full-well capacity [1, 14, 23]. This physical limitation imposes a trade-off where preserving highlights crushes shadows, and exposing for shadows blows out highlights. Consequently, computational solutions are required to bridge this perceptual gap [8, 12]. Two primary paradigms exist to recover these lost details. HDR Reconstruction recovers a linear radiance map, but requires complex Tone Mapping Operators for display, often introducing secondary artifacts like halos or contrast degradation [36]. Multi-Exposure Fusion (MEF) offers a compelling alternative by directly synthesizing a high-quality image in the display domain [7, 10, 28]. By bypassing explicit radiance estimation, MEF avoids tone-mapping distortions, aiming to seamlessly blend the most informative features into a perceptually pleasing result suitable for standard displays.

Despite its popularity, MEF struggles significantly in dynamic environments. Camera shake and moving objects introduce severe misalignments, manifesting as ghosting or tearing artifacts [6, 48, 51]. Current benchmarks for multi-exposure fusion primarily focus on static scenes and detail restoration, with few considering performance in dynamic settings. To date, a large-scale, dedicated dataset capable of comprehensively evaluating both deghosting capability and perceptual visual quality in real-world dynamic scenes remains absent [9, 49]. Consequently, a common workaround is to tone-map HDR reconstruction datasets and use the resulting LDR images as supervision for MEF. While convenient, this practice inevitably inherits the limitations of tone mapping, such as contrast distortion.

As a result, the generated supervision often struggles to faithfully reflect perceptual preferences [13, 22, 58].

Current benchmarks encounter a dilemma where realistic motion lacks reliable GT, and synthetic motion lacks realism [4, 16, 38, 42, 59]. To address these issues, we introduce ExpoMotion, a large-scale benchmark specifically designed for the joint evaluation of deghosting performance and perceptual visual quality. We integrate several established capture strategies to curate a challenging benchmark that features in-the-wild object motion alongside extreme exposure disparities [11]. Our dataset encompasses a diverse range of real-world scenarios captured across both daytime and nighttime conditions, including restaurants, beaches, staircases, streets, and other daily-life environments. The dataset includes motion types such as controlled motion, for quantitative evaluation, and real-world motion, for practical validity. For label generation, we first produce candidate ground truths via multiple algorithms, followed by selection and refinement by professional imaging experts. This annotation strategy ensures that our ground truths are not only robust to extreme motion artifacts and exposure variations but also align with human perceptual preferences.

Most existing MEF and HDR methods depend on deformable alignment, such as Optical Flow, and standard Attention to aggregate features [23, 24, 43]. However, these approaches face intrinsic limitations. Flow-based warping inevitably introduces geometric distortions in occluded regions [5, 53]. More critically, standard attention mechanisms calculate aggregation weights based on feature similarity [6, 51]. While this design principle effectively aligns features in general restoration, it faces an intrinsic paradox in MEF. The core objective here is to retrieve complementary details from auxiliary frames, such as textures in over-exposed shadows or under-exposed highlights, that are precisely missing or severely degraded in the reference view [25, 57]. To address this, we augment standard attention with the Householder transformation. By explicitly modeling feature rejection through geometric projection, our method enables the network to distinguish between useful complementary details and misalignment artifacts [32, 33]. We reconceptualize MEF alignment not as a single-step feature-matching task but as a decoupled process consisting of exposure pre-alignment and ghost filtering. This perspective addresses the twin challenges of MEF: the extreme brightness disparities that confuse standard matching metrics, and the motion artifacts that require strict geometric exclusion. To implement this, we propose the Householder Orthogonal Projection network (HOP). To address exposure discrepancies, the Global Priors Illumination Alignment (GPIA) module utilizes global statistics to rectify dynamic range variations, thereby pre-aligning auxiliary features to the reference exposure. Building on this exposure pre-alignment, we introduce Householder Orthogonal Attention (HOA) for deghosting. HOA constructs dynamic reflectors from learned misalignment vectors. Drawing an analogy to optical mirroring, this facilitates an operation that geometrically reflects auxiliary features to match the reference structure, effectively orthogonally projecting out motion-induced artifacts [45, 46]. By enforcing

this mathematically grounded constraint, HOP effectively improves deghosting fusion. Our primary contributions are threefold:

- We construct **ExpoMotion**, a large-scale dynamic MEF benchmark with 1,738 exposure stacks, curated via diverse acquisition strategies across both natural and laboratory settings for rigorous artifact evaluation.
- We propose the **Householder Orthogonal Projection (HOP)** network, which decouples alignment by utilizing global priors to harmonize illumination discrepancies and Householder transformations to project out ghosts.
- Extensive experiments demonstrate that HOP outperforms state-of-the-art methods in both quantitative metrics and perceptual quality.

2 Related Work

Multi-Exposure HDR Imaging. Existing approaches for high-dynamic-range imaging generally follow two paradigms: HDR reconstruction in the linear domain and Multi-Exposure Fusion (MEF) in the LDR domain [7, 34]. HDR reconstruction methods aim to recover linear irradiance maps by inverting the Camera Response Function (CRF) [14]. Pioneering deep learning works utilized Convolutional Neural Networks (CNNs) to learn exposure fusion weights directly from data [14, 48], establishing a strong baseline. Later advancements introduced optical flow to explicitly mitigate motion misalignment during the CNN-based merging process [16, 20, 30]. To address the limited receptive field of CNNs, recent research has shifted towards Transformer-based architectures [6, 43, 51], employing self-attention mechanisms to model long-range ghosting dependencies [24, 40]. Furthermore, generative diffusion models have been explored to hallucinate missing details in saturated regions, treating HDR reconstruction as a conditional generation task [52]. Alternatively, Multi-Exposure Fusion (MEF) bypasses CRF calibration by directly fusing the LDR sequence to produce a visually pleasing result [7, 13]. Traditional optimization-based methods focused on crafting weight maps based on hand-crafted features like local contrast and saturation [21, 27, 28]. In the deep learning era, MEF has evolved from simple pixel-wise fusion [31, 35] to sophisticated end-to-end architectures that jointly optimize for structural fidelity and perceptual aesthetics [1, 50, 60]. Across both paradigms, the central challenge lies in deghosting—mitigating artifacts caused by camera shake or dynamic objects. While explicit optical flow [14, 48] and implicit attention mechanisms [6, 51] attempt to align features, large displacements and occlusions often break correspondence assumptions, leaving residual artifacts in the fused output.

HDR Datasets and Benchmarks. Curating high-quality HDR datasets is notoriously difficult due to the requirement of capturing multiple exposures in rapid succession. In dynamic environments, unavoidable motion between frames precludes the existence of a perfect, naturally aligned ground truth. To synthesize labels, earlier works adopted controlled protocols, such as recombining static brackets to simulate stop-motion dynamics [12, 39]. A prominent baseline, Kalantari et al. [14], introduced a dataset with controllable motion, though

these synthesized movements often lack the complex motion blur and occlusions found in real-world captures. Recent efforts utilize graphics engines to generate large-scale synthetic datasets with perfect pixel-wise alignment [2, 47]. However, synthetic data suffers from a distinct domain gap—simulated noise and rendering artifacts often differ significantly from the physics of real sensor optics, limiting generalization. Furthermore, benchmarks specifically designed for MEF are largely restricted to static scenes [4, 59]. High-quality dynamic datasets that support rigorous evaluation of both HDR reconstruction and MEF deghosting remain scarce [13], underscoring the necessity of more robust evaluation protocols.

3 Proposed Dataset

Multi-Exposure Image Collection. Existing dynamic HDR datasets often rely on static bracketing to circumvent the difficulty of synthesizing ground truth (GT) from misaligned exposures. Specifically, Kong *et al.* [16] introduce camera perspective shifts between static captures (Multi-View Static-Bracketing), while Kalantari *et al.* [14] require subjects to pose in distinct stationary states (Multi-Pose Static-Bracketing). We adopt the two aforementioned protocols and introduce a third scheme: Static–Dynamic Exposure Bracketing. We capture two bracketed sets for each scene under identical settings: a static set for synthesizing reliable ground truth, and a dynamic set for input, where the reference frame is swapped with its static counterpart. This design introduces more motion patterns while maintaining high-quality supervision. Since these three protocols enforce static scenes during the exposure sequence to ensure artifact-free GT generation, they inherently fail to capture the complexity of diverse, real-world motion. We refer to the data acquired under these constrained protocols as controlled data.

In addition to the in-the-wild sequences, we captured a complementary subset of data in a professional motion laboratory. This controlled environment allows us to precisely manipulate motion patterns and illumination conditions, and to record sequences with repeatable scene dynamics. Additionally, we captured real-world dynamic scenes, which serve as a benchmark for no-reference perceptual evaluation. By integrating these strategies, we construct a large-scale MEF dataset comprising 1,738 sequences covering 3, 5, and 7-frame settings with exposures ranging from EV$-$3 to EV$+$3, amounting to a total of 10,909 images.
Data Filtering. During data collection, we captured 14,235 multi-exposure sequences under different exposure settings. Since many sequences contain distorted or low-quality frames, we perform a strict filtering process to curate a high-quality dataset. We first discard sequences with evident degradations, such as defocus blur, strong sensor noise, noticeable handshake, or other visible imaging artifacts. Moreover, to facilitate reliable ground-truth (GT) construction, we remove sequences that are not strictly static within each exposure bracket, *i.e.*, any intra-bracket motion that breaks the static-scene assumption required for synthesizing high-quality MEF supervision. For repeated sequences, we manu-

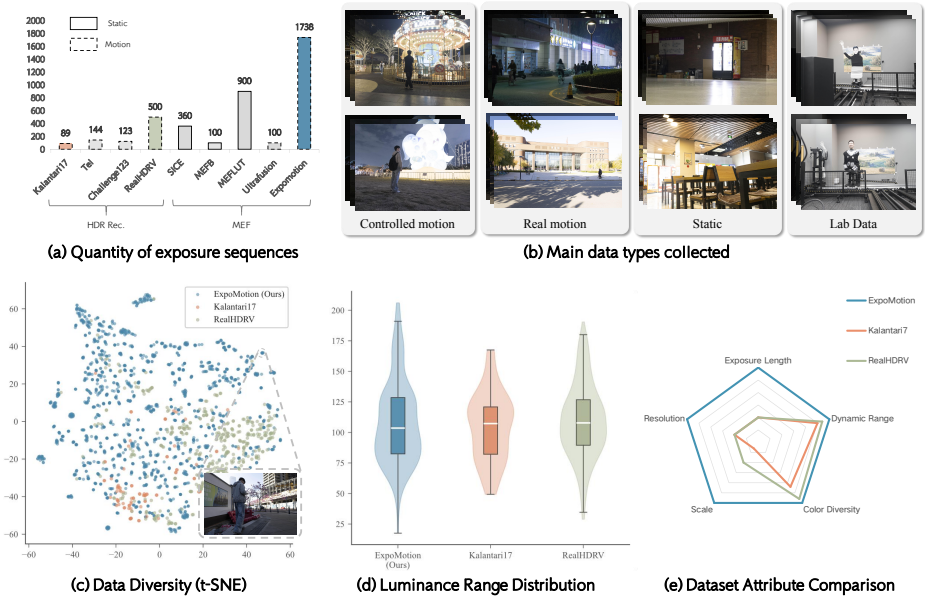

Fig. 2: Statistics and diversity of the collected ExpoMotion dataset. (a) Quantitative comparison showing our dataset is the largest among existing real-world MEF and HDR benchmarks. (b) Representative samples of the four data acquisition types. (c) T-SNE visualization showing superior feature diversity compared to related datasets. (d) Luminance range distribution. (e) A radar chart highlighting our comprehensive advantages in scale, dynamic range, and color diversity.

ally select the one with the best alignment quality and clearest content. After this stage, more than 80% of the collected sequences are filtered out, leaving about 2,850 candidate sequences. We then apply representative state-of-the-art MEF methods to these candidates to generate fused reference images, and further prune sequences according to the quality and consistency of the resulting references.

Ground Truth Generation. We construct the GT labels using a human-in-the-loop pipeline that combines automated generation with professional refinement. (1) Candidate Generation: For every sequence, we generate a diverse set of fusion results using both classic algorithmic baselines [3, 17–19, 26, 26, 29, 37] and commercial high-dynamic-range software, including Adobe Lightroom and Photomatix. This ensures a wide search space covering different de-ghosting strategies and tone-mapping operators. (2) Subjective Selection: A panel of raters screens these candidates to identify the best initialization, prioritizing results free from motion artifacts and saturation clipping. (3) Expert Refinement: The selected optimal priors are then meticulously refined by five professional photographers/imaging experts. They manually adjust the tone curves and color balance to recover shadow details and suppress highlight blowout. This manual intervention is critical for producing GTs that are not only artifact-free but also

photorealistic, avoiding the unnatural halos or color shifts often introduced by fully automated pipelines.

Dataset Analysis and Comparison. To demonstrate the superiority of our ExpoMotion dataset, we conduct a comprehensive statistical analysis comparing it with existing mainstream HDR and MEF datasets, including Kalantari17 [14], Tel [42], Challenge123 [16], RealHDRV [38], SICE [4], MEFB [59], MEFLUT [13] and UltraFusion [7]. Fig. 2 provides a holistic visualization of these comparisons.

As illustrated in Fig. 2 (a), our dataset contains a total of 1,738 multi-exposure sequences, significantly surpassing existing benchmarks in scale (e.g., 19× larger than Kalantari17 and 3× larger than RealHDRV). Unlike previous works that rely heavily on static scenes, ExpoMotion features a motion-centric composition that focuses primarily on dynamic sequences (spanning controlled, real-world, and laboratory motion), while retaining a smaller subset of static data. In addition to capturing diverse in-the-wild real motions, we established a professional laboratory setting to record precise controlled motion and scene dynamics. To evaluate the semantic richness of the collected scenes, we visualize their distribution using t-SNE. Fig. 2 (c) demonstrates that our dataset (represented by blue points) covers a much broader semantic space compared to Kalantari17 (orange) and RealHDRV (green), indicating superior diversity in scene content and texture. This wide distribution suggests that models trained on ExpoMotion are likely to achieve better generalization across a broader range of real-world scenarios. Beyond semantic content, we analyze the photometric properties essential for high-dynamic-range imaging. Fig. 2 (d) plots the distribution of luminance ranges. Our dataset exhibits a more extensive distribution with longer tails, indicating that we capture scenes with more extreme lighting conditions (both deep shadows and bright highlights). Finally, the radar chart in Fig. 2 (e) summarizes five key attributes: scale, dynamic range, resolution, exposure length, and color diversity. ExpoMotion fully encompasses the attributes of previous datasets, establishing a new benchmark for dynamic exposure fusion. Specifically, the average image resolution of ExpoMotion is 2563×1896, which offers approximately 3× the pixel count of Kalantari17 and RealHDRV dataset.

4 Proposed Method

Multi-Exposure Fusion (MEF) aims to integrate the high dynamic range information from auxiliary exposures ($\mathbf{I}_{over}, \mathbf{I}_{under}$) into a reference frame ($\mathbf{I}_{ref}$). The core challenge lies in the tension between maximizing information gain and strictly suppressing motion-induced artifacts (ghosting). In this work, we formalize the MEF task as taking a triplet of differently exposed images as input consisting of an over-exposed frame $\mathbf{I}_{over}$, a reference frame $\mathbf{I}_{ref}$, and an under-exposed frame $\mathbf{I}_{under}$ and generating a single fused image $\mathbf{I}_{pred}$ as output. The primary goal of this fusion process is to seamlessly amalgamate the distinct details preserved in each exposure: recovering shadow information from $\mathbf{I}_{over}$, highlight details from $\mathbf{I}_{under}$, and retaining the balanced mid-tones from $\mathbf{I}_{ref}$,

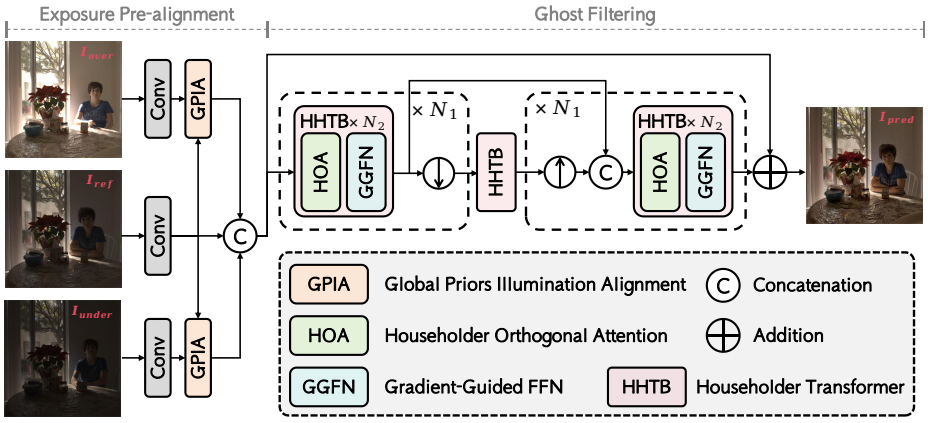

Fig. 3: Pipeline of the proposed Householder Orthogonal Projection Network (HOP). MEF is decomposed into exposure pre-alignment and ghost filtering stages.

all while producing a perceptually pleasing result free of motion-induced ghosting artifacts.

To address the issues of inconsistent brightness and motion artifacts across the three input images, we propose the Householder Orthogonal Projection Network (HOP). We decompose the MEF task into two core stages: exposure pre-alignment and ghost filtering. Specifically, we first align the brightness of the two auxiliary frames to the reference frame using global illumination priors. Then, inspired by the Householder transformation, we model the artifact removal problem as geometrically mirroring misaligned auxiliary features to match the structural pose of reference features. The detailed pipeline is illustrated in Fig. 3. We first perform exposure pre-alignment on the three input images. Subsequently, the features of these three images are concatenated and fed into a U-shaped Transformer architecture, where artifact rejection and filtering are conducted within each Transformer block.

4.1 Global Priors Illumination Alignment (GPIA).

A major challenge in HDR features is the brightness discrepancy. Inspired by the Retinex theory, where illumination is often low-frequency, we introduce the GPIA module to pre-align the features before encoding. Given the initial feature maps $\mathbf{F}_{Ref}$ and $\mathbf{F}_{Aux}$. We first extract the global illumination priors via average pooling with a 16×16 kernel:

$$\mathbf{L}_{Ref} = \phi(\text{AvgPool}(\mathbf{F}_{Ref})), \quad \mathbf{L}_{Aux} = \phi(\text{AvgPool}(\mathbf{F}_{Aux})) \tag{1}$$

where $\phi(\cdot)$ is a lightweight mapping function. The illumination ratio map $\mathcal{R}$ is computed to modulate the auxiliary stream:

$$\mathcal{R} = \frac{\text{Up}(\mathbf{L}_{Ref}) + \epsilon}{\text{Up}(\mathbf{L}_{Aux}) + \epsilon}, \quad \mathbf{F}'_{Aux} = \mathbf{F}_{Aux} \odot \mathcal{R} \tag{2}$$

By aligning the exposure levels of the auxiliary frame to the reference beforehand, this module decouples illumination correction from motion handling, allowing subsequent layers to focus exclusively on artifact removal.

4.2 Householder Orthogonal Attention (HOA)

The Householder Orthogonal Attention (HOA) module is designed to merge the reference features $\mathbf{F}$ and the pre-aligned features $\mathbf{F}_{SA}$ while discarding misaligned ghosts. Structurally, it functions as a coarse-to-fine alignment pipeline. It first employs Transposed Channel Self-Attention to compute the global covariance between frames, which corresponds to $\mathbf{F}_{SA}$ module illustrated in Fig. 4. Subsequently, the refined features are passed to our core module: the **Householder Orthogonal Projection Unit (HOPU)**. HOPU treats the deghosting process as a geometric projection problem, utilizing a dynamic Householder reflection matrix to separate the aligned signal subspace from the artifact-inducing noise subspace.

Geometric Intuition of Householder Rejection. Formally, a Householder matrix $\mathbf{H} \in \mathbb{R}^{d\times d}$ is a fundamental linear algebra operator that performs an orthogonal reflection of a vector across a hyperplane defined by a unit normal vector $\mathbf{v}$. This reflection can be expressed as:

$$\mathbf{H} = \mathbf{I} - 2\mathbf{v}\mathbf{v}^T \tag{3}$$

where $\mathbf{I}$ is the identity matrix.

Inspired by this idea, to adapt this powerful geometric tool for deep feature learning in MEF, we propose an Adaptive Householder-like Projection:

$$\mathbf{H}_{hop} = \mathbf{I} - \alpha\mathbf{v}\mathbf{v}^T \tag{4}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^C$ is a learnable scaling vector initialized to zero, allowing the network to fine-tune the rejection intensity per channel.

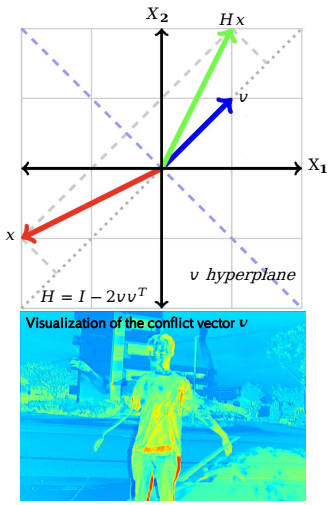

Fig. 4: Geometric intuition of the Householder transformation.

Specifically, we define the conflict vector $\mathbf{v}$ based on feature differences. The ghost artifacts in the misaligned auxiliary feature $\mathbf{F}_{SA}$ are theoretically modeled as its projection onto $\mathbf{v}$. By applying $\mathbf{H}_{hop}$, we perform an orthogonal rejection to subtract this component, recovering the clean signal $\mathbf{F}_{clean}$ on the conflict plane. This geometric mirroring operation rigorously filters out motion-induced discrepancies, offering a robust alternative to standard similarity-based aggregation.

Difference-Aware Vector Generation. Instead of concatenation, we explicitly model the conflict direction $\mathbf{v}$ from the feature difference $\mathbf{D} = \mathbf{F} - \mathbf{F}_{SA}$.

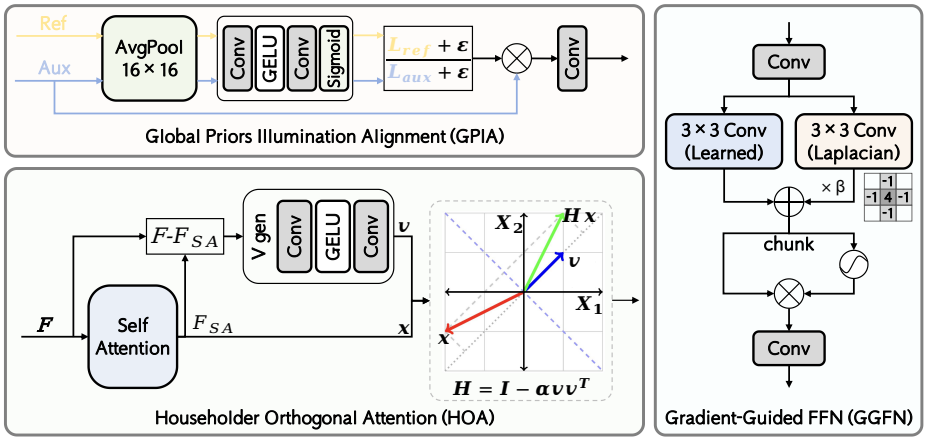

Fig. 5: Illustration of HHTB and GPIA. Specifically, HHTB is composed of two core components: Householder Orthogonal Attention (HOA) and Gradient-Guided Feed-Forward Network (GGFN).

This acts as a strong prior for ghost localization:

$$\mathbf{v}_{raw} = \mathcal{G}_{dw}(\mathbf{D}), \quad \mathbf{v} = \frac{\mathbf{v}_{raw}}{||\mathbf{v}_{raw}||_2 + \epsilon} \tag{5}$$

where $\mathcal{G}_{dw}$ denotes a parameter-efficient depth-wise convolution block. Normalization ensures $\mathbf{v}$ lies on the unit hypersphere.

Adaptive Orthogonal Filtering. To avoid suppressing useful non-ghost details and transfer the linear operator to non-linear, we introduce an intensity gate $\mathcal{M}$ derived from the magnitude of the difference map:

$$\mathcal{M} = \tanh\left(\text{AvgPool}(|\mathbf{D}|)\right) \tag{6}$$

The clean feature $\mathbf{F}_{clean}$ is obtained by projecting $\mathbf{F}_{Aux}$ onto $\mathbf{v}$ and rejecting this component:

$$\mathbf{F}_{clean} = \mathbf{F}_{Aux} - \alpha \cdot \mathcal{M}\mathbf{v} \odot (\mathbf{v}^T \mathbf{F}_{Aux}) \tag{7}$$

This formulation is computationally efficient and removes the misaligned ghost signal defined by $\mathbf{v}$.

4.3 Gradient-Guided Feed-Forward Network (GGFN)

Standard Feed-Forward Networks (FFNs) typically process features locally but often neglect explicit structural dependencies. However, in HDR deghosting, preserving high-frequency details such as edges and textures is paramount for perceptual quality. To address this, we propose the Gradient-Guided Feed-Forward Network (GGFN). Instead of relying solely on learned kernels, GGFN synergizes data-driven interpretation with a fixed geometric prior.

Specifically, the module operates through two parallel pathways:

– **Learned Branch:** A depth-wise convolution (DWConv) captures learnable semantic contexts and local patterns.
– **Laplacian Branch:** We introduce a fixed **Discrete Laplacian Operator** $\mathcal{L}$ to explicitly extract second-order spatial derivatives (i.e., curvature and boundaries). This branch serves as a hard-coded high-pass filter:

$$\mathcal{K}_{Lap} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{8}$$

The Laplacian branch acts as a hard-coded high-pass filter, forcing the network to attend to structural boundaries. The two branches are fused via a learnable mixing parameter β:

$$\mathbf{X}_{hidden} = \text{DWConv}(\mathbf{X}_{in}) + \beta(\mathbf{X}_{in} * \mathcal{K}_{Lap}) \tag{9}$$

By initializing $\beta = 0$, the network starts with standard learning and gradually incorporates the Laplacian geometric prior to sharpen textures, effectively regularizing the learning process with explicit edge information.

5 Experiments

5.1 Experimental Setting

Datasets. To evaluate motion-artifact removal and detail restoration for multi-exposure fusion (MEF), we construct two test subsets. We randomly select 132 samples with ground truth (GT) for full-reference evaluation, and 113 real-world exposure sequences without GT for no-reference evaluation. The remaining 1,493 exposure sequences are used for training. Following UltraFusion [7], we supplement our training data with HDR reconstruction datasets, Kalantari17 [14] and RealHDRV [38]. For both datasets, we apply Photomatix for tone mapping to obtain PNG format fusion targets, and we adopt their official training/testing splits.
Implementation details. Our method is implemented in PyTorch and trained end-to-end using the AdamW optimizer. The initial learning rate is set to 2×10^{-4} and decayed by cosine annealing over 150 epochs. The batch size is set to 4 per GPU. Automatic mixed precision is enabled. All experiments are conducted using distributed data-parallel training on 4 NVIDIA V100 GPUs with synchronized batch normalization. We conduct experiments on two parameter-scale versions of our method: HOP-S ($N_1 = 2$) and HOP-B ($N_1 = 3$).
Evaluation metrics. For datasets with GT, we report PSNR and SSIM as full-reference metrics. For datasets without GT, we report MUSIQ [15], DeQA-Score [55], PAQ2PIQ [54], and HyperIQA [41] as no-reference perceptual metrics. Specifically, we conduct full-reference evaluations on Kalantari17 [14], RealHDRV [38], and our controlled-motion testset, while performing no-reference evaluations on a composite test set comprising sequences from the Sen [37] and Tursen [44] datasets, as well as our real-motion testset.

Table 1: Quantitative full-reference results on Kalantari17 [14], RealHDRV [38], and our controlled-motion test set. Best in **bold**, second best underlined.

Method	Kalantari17 [14]		RealHDRV [38]		Expomotion		Time	Param
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	(ms)	(M)
SAFNet [16] (ECCV'24)	24.926	0.921	25.077	0.910	26.263	0.906	368	1.118
HDR_Trans [24] (ECCV'22)	27.032	0.930	25.666	0.924	27.877	0.915	6355	1.454
SCTNet [42] (ICCV'23)	27.100	0.931	26.420	0.923	27.806	0.912	7184	3.326
AFUNet [20] (ICCV'25)	27.582	0.934	26.963	0.926	27.672	0.917	10576	1.138
ASTv2 [61] (TPAMI'25)	28.113	0.934	27.563	0.937	28.573	0.925	1353	7.751
Restormer [56] (CVPR'22)	28.325	0.936	28.282	0.940	28.640	0.926	1860	26.10
HOP-S (Ours)	28.501	0.940	28.285	0.937	28.593	0.926	1403	3.699
HOP-B (Ours)	**28.804**	**0.943**	**28.704**	**0.942**	**28.720**	**0.927**	1576	15.69

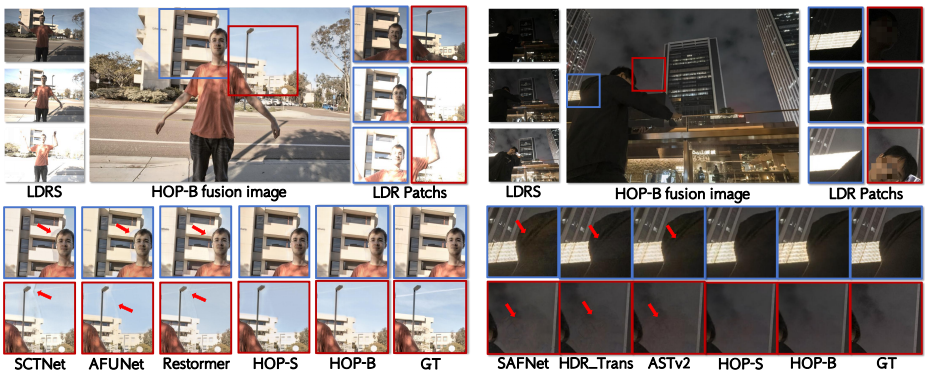

Fig. 6: Visual comparisons on Kalantari17 [14] dataset and ours. Zoom in for a better view.

5.2 Comparison with Previous Work

Compared methods. We compare our approach with representative CNN-based methods, such as SAFNet [16] and AFUNet [20], and recent Transformer-based restoration architectures, including SCTNet [42], Restormer [56], HDR-Trans [24], and ASTv2 [61]. All methods are evaluated under the same protocols when re-training is required.

Full-reference comparison. Table 1 summarizes the quantitative results on the Kalantari17 [14], RealHDRV [38], and our controlled-motion testsets. As presented, our HOP-B establishes a new state-of-the-art across all metrics. Notably, it outperforms the strong competitor Restormer [56] by approximately 0.5 dB in PSNR on Kalantari17 [14], while requiring only about 60% of its Params (15.69M vs. 26.10M). Even our lightweight variant, HOP-S, achieves remarkable efficiency-performance trade-offs: it surpasses Restormer [56] and ASTv2 [61] on the Kalantari17 [14] and RealHDRV [38] dataset with significantly reduced computational costs (only 3.699M Params). Qualitatively, as shown in Fig. 6, previous methods like AFUNet [20] and Restormer [56] suffer from noticeable ghosting artifacts in dynamic regions. Specifically, as indicated by the red arrows in the

Table 2: Cross-dataset generalization with no-reference metrics. Models are trained on different datasets and tested on (a) Sen [37] and Tursen [44], and (b) our real-motion testset.

Trainset	Method	Test on Sen [37] and Tursen [44]					Test on Ours Real-motion Testset				
		ID	MUSIQ↑	DeQA↑	PAQ2PIQ↑	HyperIQA↑	ID	MUSIQ↑	DeQA↑	PAQ2PIQ↑	HyperIQA↑
Kalantari17	Restormer [56]	(a)	57.45	3.214	67.98	0.4211	(g)	59.94	3.544	70.28	0.4314
	HOP-B	(b)	57.87	3.225	68.68	0.4328	(h)	61.08	3.55	70.83	0.4354
RealHDRV	Restormer [56]	(c)	60.50	3.366	69.09	0.4736	(i)	61.85	3.599	70.92	0.4703
	HOP-B	(d)	60.89	3.395	69.97	0.4819	(j)	62.84	3.61	71.23	0.4824
ExpoMotion	Restormer [56]	(e)	62.32	3.535	70.96	0.4839	(k)	65.40	3.867	72.19	0.5195
	HOP-B	(f)	62.74	3.546	71.34	0.4861	(l)	65.68	3.87	72.19	0.5284

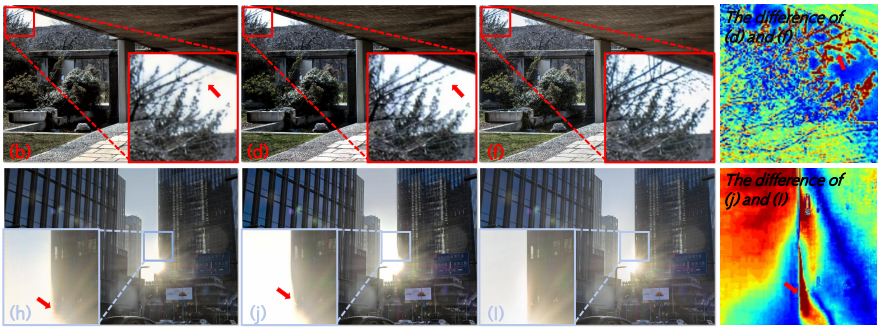

Fig. 7: Visual comparison of cross-dataset generalization. The indices of the sub-figures are consistent with those in Tab. 2.

left sub-figure, competing models struggle to distinguish foreground structures from moving backgrounds, resulting in severe ghosting artifacts and unnatural distortions around the lamp pole. Furthermore, in the challenging night scene on the right, while other models exhibit severe noise and blurring in the dark wall area, our model faithfully recovers clear details in both over-exposed light grids and under-exposed shadows. In contrast, our method significantly outperforms previous approaches in artifact removal and dark-region denoising.

No-reference comparison. Table 2 presents the perceptual evaluation using non-reference metrics on two real-world test sets (Sen [37]/Tursen [44] and ours), which lack ground truth. This experiment effectively validates the generalization capability of models trained on different datasets. Crucially, the results highlight the superiority of our ExpoMotion dataset. Specifically, on the Sen and Tursen testset, the HOP-B model trained on our dataset (f) achieves a MUSIQ score of 62.74, outperforming the same model trained on Kalantari17 (b, 57.87) and RealHDRV (d, 60.89) by significant margins of approximately 4.9 and 1.9 points, respectively. Moreover, even under identical training data, our HOP-B architecture consistently surpasses Restormer (e.g., comparing e vs. f), demonstrating the effectiveness of our design in rendering visually pleasing HDR images. Fig. 7 presents qualitative comparisons. As observed in the top row, competing datasets (Fig. 7 (b) and (d)) trained on smaller datasets struggle to handle complex motions, resulting in blurred tree branches and loss of high-frequency

Table 3: Ablation study on Kalantari17 dataset. We compare the effects of different combinations of GPIA, HOPU, and GGFN modules, with a total of seven experimental groups from (a) to (g). The GDFN module proposed in Restormer [56].

ID	GPIA	HOPU	FFN		PSNR	SSIM	Params (M)
			GDFN	GGFN			
(a)			✓		27.4526	0.9287	3.3818
(b)	✓		✓		27.8562	0.9370	3.3819
(c)		✓	✓		27.9003	0.9323	3.6332
(d)				✓	27.8332	0.9376	3.4475
(e)		✓		✓	28.2321	0.9391	3.6989
(f)	✓			✓	28.1334	0.9382	3.4476
(g)	✓	✓		✓	**28.5011**	**0.9404**	3.6990

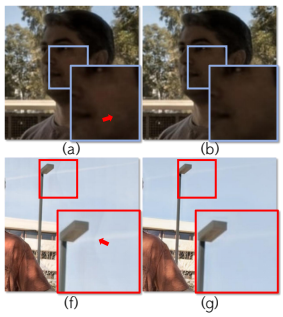

Table 4: Visual results of the ablation study

details. Similarly, in the challenging extreme-lighting scene shown in the bottom row, these methods (Fig. 7 (h) and (j)) fail to preserve the geometric integrity, resulting in noticeable structural distortions near strong light sources. This significant improvement indicates that the diverse motion patterns and realistic luminance variations captured in our dataset enable the network to learn more generalized features effectively.

5.3 Ablation Study

We utilize the Kalantari17 dataset to dissect the contribution of each component within our lightweight variant, HOP-S. As detailed in Tab. 3, the baseline (row a), devoid of our proposed modules, yields a PSNR of 27.45 dB.

Impact of Individual Modules. Integrating modules individually, like GPIA (b), HOPU (c), or GGFN (d), consistently boosts PSNR by approximately 0.4 dB compared to the baseline. Crucially, these gains incur negligible computational overhead ($\leq$0.3 M parameters), verifying the efficiency of our modular design. As shown in Tab. 4, in the top row, (b) with GPIA exhibits better noise recovery on human faces compared to (a). In the bottom row, (g) with HOPU demonstrates more significant artifact removal. These improvements validate the effectiveness of both the GPIA and HOPU modules.

Synergy and Full Model. The results further highlight the complementary nature of our components. Pairwise combinations (rows e, f) exhibit clear performance jumps over standalone configurations, confirming that the modules effectively collaborate to handle different aspects of restoration. Consequently, the full model (row g), which integrates GGFN, GPIA, and HOPU, achieves the peak performance of 28.50 dB with a substantial 1.05 dB improvement over the baseline. Remarkably, this significant gain is achieved with only a marginal 9.3% increase in Params (from 3.3818 M to 3.6990 M), demonstrating a highly favorable trade-off between reconstruction quality and computational cost.

6 Conclusion

In this work, we address the twin challenges of benchmarking and deghosting in Multi-Exposure Fusion (MEF) for dynamic scenes. To overcome the scarcity of reliable evaluation benchmarks, we construct ExpoMotion, a large-scale benchmark that assesses deghosting capabilities, detail restoration, and perceptual aesthetics via a hybrid of controlled and in-the-wild motion sequences. On the methodological front, we depart from standard correspondence learning and introduce the Householder Orthogonal Projection (HOP) network. By mathematically reframing deghosting as a geometric filtering problem, HOP effectively decouples the task: it first harmonizes exposure disparities through global statistics and subsequently projects out motion artifacts using a relaxed Householder transformation. This coherent design yields superior fusion fidelity and computational efficiency compared to state-of-the-art alternatives.

Limitation. Our model adopts a three-input fusion framework. This constraint limits the applicability of our method to datasets or devices that provide a flexible number of exposures. Future research will focus on generalizing the network design to accommodate variable input frame counts, further improving the robustness and versatility of the system.

7 Acknowledgments

This work was supported by the Shenzhen Science and Technology Program (No. JCYJ20240813114229039), PCL Major Key Project of PCL2025A17-2, 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. Bai, H., Zhang, J., Zhao, Z., Deng, L., Cui, Y., Xu, S.: Retinex-mef: Retinex-based glare effects aware unsupervised multi-exposure image fusion. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7251–7261 (2025)
2. Barua, H.B., Stefanov, K., Wong, K., Dhall, A., Krishnasamy, G.: Gta-hdr: A large-scale synthetic dataset for hdr image reconstruction. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7876–7886 (2025)
3. Bruce, N.D.: Expoblend: Information preserving exposure blending based on normalized log-domain entropy. Computers & Graphics **39**, 12–23 (2014)
4. 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)
5. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Understanding deformable alignment in video super-resolution. AAAI Conference on Artificial Intelligence (AAAI). **35**(2), 973–981 (2021)

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, Z., Wang, Y., Cai, X., You, Z., Lu, Z., Zhang, F., Guo, S., Xue, T.: Ultrafusion: Ultra high dynamic imaging using exposure fusion. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2025)
8. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. ACM SIGGRAPH Annual Conference (SIGGRAPH). pp. 369–378 (1997)
9. Fang, Y., Zhu, H., Ma, K., Wang, Z., Li, S.: Perceptual evaluation for multi-exposure image fusion of dynamic scenes. IEEE Transactions on Image Processing (TIP). **29**, 1127–1138 (2019)
10. Grosch, T., et al.: Fast and robust high dynamic range image generation with camera and object movement. Vision, Modeling and Visualization, RWTH Aachen. pp. 277–284 (2006)
11. Heo, Y.S., Lee, K.M., Lee, S.U., Moon, Y., Cha, J.: Ghost-free high dynamic range imaging. Asian Conference on Computer Vision pp. 486–500 (2010)
12. Hu, J., Gallo, O., Pulli, K., Sun, X.: Hdr deghosting: How to deal with saturation? IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1163–1170 (2013)
13. Jiang, T., Wang, C., Li, X., Li, R., Fan, H., Liu, S.: Meflut: Unsupervised 1d lookup tables for multi-exposure image fusion. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10542–10551 (2023)
14. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Transactions on Graphics (TOG). **36**(4) (2017)
15. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5148–5157 (2021)
16. Kong, L., Li, B., Xiong, Y., Zhang, H., Gu, H., Chen, J.: Safnet: Selective alignment fusion network for efficient hdr imaging. European Conference on Computer Vision (ECCV). (2024)
17. Kou, F., Li, Z., Wen, C., Chen, W.: Multi-scale exposure fusion via gradient domain guided image filtering. IEEE International Conference on Multimedia and Expo (ICME). pp. 1105–1110 (2017)
18. 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)
19. Li, S., Kang, X., Hu, J.: Image fusion with guided filtering. IEEE Transactions on Image Processing (TIP). **22**(7), 2864–2875 (2013)
20. Li, X., Ni, Z., Yang, W.: Afunet: Cross-iterative alignment-fusion synergy for hdr reconstruction via deep unfolding paradigm. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10666–10675 (2025)
21. Li, Z., Zheng, J., Zhu, Z., Wu, S.: Selectively detail-enhanced fusion of differently exposed images with moving objects. IEEE Transactions on Image Processing (TIP). **23**(10), 4372–4382 (2014)
22. Liu, R., Li, C., Cao, H., Zheng, Y., Zeng, M., Cheng, X.: Emef: Ensemble multi-exposure image fusion. AAAI Conference on Artificial Intelligence (AAAI). **37**(2), 1710–1718 (2023)
23. Liu, S., Zhang, X., Sun, L., Liang, Z., Zeng, H., Zhang, L.: Joint hdr denoising and fusion: A real-world mobile hdr image dataset. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13966–13975 (2023)

24. Liu, Z., Wang, Y., Zeng, B., Liu, S.: Ghost-free high dynamic range imaging with context-aware transformer. European Conference on Computer Vision (ECCV). pp. 344–360 (2022)
25. Luo, J., Ren, W., Gao, X., Cao, X.: Multi-exposure image fusion via deformable self-attention. IEEE Transactions on Image Processing (TIP). **32**, 1529–1540 (2023)
26. Ma, K., Duanmu, Z., Zhu, H., Fang, Y., Wang, Z.: Deep guided learning for fast multi-exposure image fusion. IEEE Transactions on Image Processing (TIP). **29**, 2808–2819 (2019)
27. 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)
28. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion. Pacific Conference on Computer Graphics and Applications (PG). pp. 382–390 (2007)
29. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion: A simple and practical alternative to high dynamic range photography. Computer graphics forum **28**(1), 161–171 (2009)
30. Prabhakar, K.R., Agrawal, S., Singh, D.K., Ashwath, B., Babu, R.V.: Towards practical and efficient high-resolution hdr deghosting with cnn. European Conference on Computer Vision (ECCV). pp. 497–513 (2020)
31. Prabhakar, K.R., Arora, R., Swaminathan, A., Singh, K.P., Babu, R.V.: A fast, scalable, and reliable deghosting method for extreme exposure fusion. IEEE International Conference on Computational Photography (ICCP). pp. 1–8 (2019)
32. Qin, R., Liu, X., Liu, X., Liu, J., Shi, J., Lin, L., Yang, J.: No pains, more gains: Recycling sub-salient patches for efficient high-resolution image recognition. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14965–14975 (2025)
33. Qin, R., Liu, X., Shi, J., Lin, L., Yang, J.: Boosting the dual-stream architecture in ultra-high resolution segmentation with resolution-biased uncertainty estimation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25960–25970 (2025)
34. Qu, L., Liu, Y., Zhou, S., Liang, J., Zeng, H., Zhang, L., Yang, J.: There and back again: A flexible-frame transformer for multi-exposure fusion (2026), https://arxiv.org/abs/2606.27905
35. Ram Prabhakar, K., Sai Srikar, V., Venkatesh Babu, R.: Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4714–4722 (2017)
36. Reinhard, E., Stark, M., Shirley, P., Ferwerda, J.: Photographic tone reproduction for digital images. Seminal Graphics Papers: Pushing the Boundaries pp. 661–670 (2023)
37. 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). pp. 203–1 (2012)
38. 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. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2879–2888 (2024)
39. 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. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2879–2888 (June 2024)

40. Song, J.W., Park, Y.I., Kong, K., Kwak, J., Kang, S.J.: Selective transhdr: Transformer-based selective hdr imaging using ghost region mask. European Conference on Computer Vision (ECCV). pp. 288–304 (2022)
41. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3667–3676 (2020)
42. Tel, S., Wu, Z., Zhang, Y., Heyrman, B., Demonceaux, C., Timofte, R., Ginhac, D.: Alignment-free hdr deghosting with semantics consistent transformer. IEEE/CVF International Conference on Computer Vision (ICCV). (2023)
43. Tel, S., Wu, Z., Zhang, Y., Heyrman, B., Demonceaux, C., Timofte, R., Ginhac, D.: Alignment-free hdr deghosting with semantics consistent transformer. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12836–12845 (2023)
44. Tursun, O.T., Akyüz, A.O., Erdem, A., Erdem, E.: An objective deghosting quality metric for hdr images. Computer Graphics Forum **35**(2), 139–152 (2016)
45. Wang, X., Zeng, H., Chen, J., Liu, S., Chen, Y., Chao, G.: Otlrm: Orthogonal learning-based low-rank metric for multi-dimensional inverse problems. AAAI Conference on Artificial Intelligence (AAAI). **39**(20), 21278–21286 (2025)
46. Wang, X., Zeng, H., Sun, B., Cao, J., Zhang, K., Shen, Q., Chen, Y.: Deep lora-unfolding networks for image restoration. IEEE Transactions on Image Processing (TIP). (2026)
47. Wang, Y., Wu, J., Bian, Y., Zhang, F., Xue, T.: S2r-hdr: A large-scale rendered dataset for hdr fusion. arXiv preprint arXiv:2504.07667 (2025)
48. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. European Conference on Computer Vision (ECCV). pp. 117–132 (2018)
49. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. European Conference on Computer Vision (ECCV). pp. 120–135 (2018)
50. 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)
51. 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. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1751–1760 (2019)
52. Yan, Q., Hu, T., Sun, Y., Tang, H., Zhu, Y., Dong, W., Van Gool, L., Zhang, Y.: Toward high-quality hdr deghosting with conditional diffusion models. IEEE Transactions on Circuits and Systems for Video Technology **34**(5), 4011–4026 (2023)
53. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep hdr imaging via a non-local network. IEEE Transactions on Image Processing (TIP). **29**, 4308–4322 (2020)
54. Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., Bovik, A.: From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3575–3585 (2020)
55. You, Z., Cai, X., Gu, J., Xue, T., Dong, C.: Teaching large language models to regress accurate image quality scores using score distribution. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14483–14494 (2025)

56. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5728–5739 (2022)
57. Zeng, H., Wang, X., Chen, Y., Su, J., Liu, J.: Vision-language gradient descent-driven all-in-one deep unfolding networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7524–7533 (2025)
58. Zhang, N., Ye, Y., Zhao, Y., Wang, R.: Revisiting the stack-based inverse tone mapping. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9162–9171 (2023)
59. Zhang, X.: Benchmarking and comparing multi-exposure image fusion algorithms. Information Fusion **74**, 111–131 (2021)
60. Zhao, Z., Deng, L., Bai, H., Cui, Y., Zhang, Z., Zhang, Y., Qin, H., Chen, D., Zhang, J., Wang, P., Van Gool, L.: Image fusion via vision-language model. International Conference on Machine Learning (ICML). (2024)
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). (2025)