Coarse-to-fine Framework for Generative MEF via Implicit Neural Representation

Sangmin Han[1,2], Jinho Kim[1], Jinwoo Kim[1], Dongyoung Kim[3], and Seon Joo Kim[1]

[1]Yonsei University [2]AI Lab, CTO Division, LG Electronics
[3]AI Center - Toronto, Samsung Electronics

Fig. 1: Our coarse-to-fine generative MEF framework preserves the natural synthesis of generative MEF (color tone and reduced ghosting) while achieving detailed structural preservation comparable to conventional MEF. It is about 4× faster than the previous generative MEF method (Ultrafusion [8]) and runs close to conventional MEF (HSDS-MEF [46]); speed is measured on a single NVIDIA RTX A5000 with a 1228 × 1638-resolution image. The first-row insets are center-cropped.

Abstract. Multi-exposure fusion (MEF) expands the luminance range beyond what a single exposure can capture. Combining images taken at different exposure levels requires handling geometric differences while naturally merging their complementary brightness information. It often demands generative completion where details are missing. Diffusion-based generative methods address these challenges, however, they are computationally expensive and struggle to preserve fine structures in saturated regions. We propose LIIFusion, a coarse-to-fine framework that balances fusion quality and efficiency in generative MEF. The coarse stage performs low resolution generative fusion, enhanced by an adaptive exposure correction that recovers structure lost in saturated over-exposed areas. The fine stage applies an implicit neural representation, and adapts the local implicit image function for fusion, enabling resolution-agnostic, pixel-wise refinement capable of restoring fine detail. LIIFusion achieves up to 4× speed-up over existing generative methods while maintaining or improving structural fidelity and perceptual quality. We believe this framework provides an effective pathway toward making generative MEF more practical in real-world applications.

1 Introduction

A fundamental challenge in computational photography is capturing scenes with a high dynamic range (HDR). Due to the limited dynamic range of standard sensors, a single exposure is often insufficient to capture the full range of scene illumination, resulting in over-exposed backgrounds or under-exposed foregrounds. A common solution to this problem is multi-exposure fusion (MEF). This technique involves capturing a sequence of images at varying exposure levels and synthesizing them into a single high-quality image that preserves details across all regions.

The core objective of MEF is to blend salient, well-exposed regions from the input sequence while maintaining natural tonal transitions. Traditional approaches [31] tackled this using hand-crafted, multi-scale pyramid fusion, which relies on pixel-level measures such as contrast, saturation, and well-exposedness. However, these low-level heuristics often produce halo artifacts and, crucially, lack the semantic awareness to distinguish important foreground details from background texture. To address these limitations, modern deep learning methods [14,25,46,50] have improved robustness by learning end-to-end fusion strategies. These models often operate on rich, feature-level representations and may adopt multi-scale designs for computational efficiency.

However, both conventional and early learning-based approaches face significant challenges in real-world scenarios. They frequently exhibit ghosting artifacts when subject motion occurs between exposures and fail in scenes with extreme dynamic ranges, where large regions become saturated and require generative synthesis rather than simple blending. Recent works [8, 63] have introduced diffusion-based generative priors that reinterpret MEF as a guided inpainting task. While these models effectively suppress ghosting and synthesize realistic structures in severely over- or under-exposed regions, they introduce a new set of trade-offs as summarized in Tab. 1. Specifically, high-resolution fusion is computationally expensive because diffusion models operate at fixed resolutions (e.g., 512×512) and require patch-wise sampling, and purely generative fusion can still yield structural uncertainty or unstable tones when guidance is weak in saturated areas.

Table 1: Comparison between conventional MEF and generative MEF.

	Conventional MEF	**Generative MEF**	
		Patch-wise	**Coarse-to-fine**
Models	CNN, Transformer	Diffusion	Diffusion, LIIF
Dynamic Range	Limited (3-4 stops)	Extended (9 stops)	
Motion Handling	Static / Mild	Dynamic	
Fusion Behavior	Regressive	Probabilistic	Both
Speed	Fast (∼minutes)	Slow (∼hours)	Fast (∼minutes)

To overcome these limitations, we propose **LIIFusion**, a coarse-to-fine generative framework that harmonizes a diffusion prior at low-resolution with a high-resolution implicit refinement module. In the coarse stage, LIIFusion performs generative fusion on downsampled exposure pairs, operating near the diffusion model's native pre-training scale to maximize efficiency and fully exploit its inpainting priors. This design lets the diffusion model produce a single low-resolution fused image with a minimal number of forward passes, avoiding expensive patch-wise sampling. Within this stage, we introduce an adaptive exposure

correction mechanism, which conditions the input to yield structurally reliable coarse estimates even in saturated regions. In the fine stage, we employ an implicit neural representation (INR) to perform resolution-agnostic detail fusion. This module extracts LIIF-style features [7] from the low-resolution fusion and uses lightweight encoders for the high-resolution exposures, enabling efficient multi-resolution fusion. By integrating these complementary signals at the coordinate level, the INR reconstructs a high-quality, continuous fused image. To the best of our knowledge, this constitutes the first application of INR to the task of multi-exposure fusion.

We conduct extensive quantitative and qualitative evaluations on both static and dynamic multi-exposure benchmarks. Results demonstrate that LIIFusion achieves a 4× speed-up over existing generative MEF methods while maintaining superior quantitative performance and perceptual quality. We believe this framework bridges the critical gap between generative fidelity and high-resolution efficiency, paving the way for the practical deployment of generative MEF in real-world imaging pipelines.

2 Related works

High dynamic range imaging. HDR imaging aims to capture the full luminance range of real-world scenes, overcoming the sensor limitations of standard cameras such as pixel saturation and narrow dynamic range. Existing approaches generally fall into two categories: HDR reconstruction, which restores physical radiance, and Multi-Exposure Fusion (MEF), which targets perceptual quality. HDR reconstruction methods [3, 6, 15, 19, 22, 27, 34, 41, 44, 45, 47, 48, 52, 53, 61, 62] recover absolute scene radiance maps from Low Dynamic Range (LDR) inputs. While precise, these methods typically require camera response function (CRF) inversion and subsequent tone-mapping for display, which can introduce color distortion or dynamic range compression artifacts.

Multi-exposure fusion (MEF). In contrast, MEF [28, 29, 31] directly fuses the LDR sequence into a perceptually high-quality image, bypassing explicit radiance recovery. Traditional MEF methods [21, 23, 35, 60] relied on hand-crafted weights or multi-scale blending, often struggling with ghosting artifacts in dynamic scenes. Deep learning-based approaches [14, 37, 46, 49, 50, 59] have significantly improved robustness to misalignment and detail preservation. Most recently, diffusion-based methods [8, 63] have leveraged generative priors to hallucinate plausible structures in saturated regions, treating MEF as a guided inpainting task. However, to handle high-resolution inputs, these diffusion models rely on patch-wise inference. This not only incurs prohibitive computational costs due to repeated sampling but can also lead to spatial inconsistencies at patch boundaries or unstable structure when guidance is weak in saturated regions. In this work, we address these bottlenecks by decoupling the generative prior from the output resolution and pairing it with deterministic refinement.

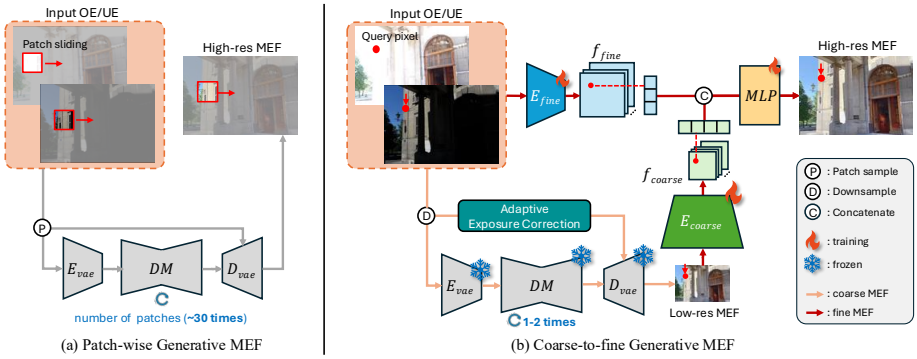

Fig. 2: Overview of LIIFusion pipeline. (a) Previous generative MEF models operate in a patch-wise manner, requiring many sampling passes. (b) Our proposed coarse-to-fine framework, LIIFusion, performs low-resolution generative fusion with adaptive exposure correction to obtain a structurally reliable result, minimizing diffusion usage to up to 2 passes depending on the image resolution ratio, followed by a LIIF-based fine stage that conducts resolution-agnostic, pixel-wise detail fusion.

Implicit neural representations (INR). INR [2, 5, 12, 13, 17, 20, 33, 36, 38, 39, 56] models visual signals as continuous functions mapping spatial coordinates to signal values. Originally popularized for 3D view synthesis in NeRF [32], INRs have been successfully adapted to 2D tasks such as super resolution (LIIF [7]), where they reconstruct fine details from discrete features via local implicit functions. Recently, this paradigm has extended to other low-level vision tasks. For instance, CoLIE [9] utilizes INRs for context-aware low-light enhancement, while INR-st [18] enables controllable style transfer via test-time training. Despite these advances, INR has not yet been explored in the context of multi-exposure fusion. We hypothesize that the continuous nature of INRs is uniquely suited for MEF, as it allows for the seamless integration of global generative priors (from the coarse stage) with high-frequency texture details (from the fine stage) at arbitrary coordinates.

3 Methodology

In this section, we describe our proposed coarse-to-fine framework for generative multi-exposure fusion (MEF). We first define the notations used throughout this paper. Let $I_{oe} \in \mathbb{R}^{H\times W\times 3}$ and $I_{ue} \in \mathbb{R}^{H\times W\times 3}$ denote the high-resolution (HR) over-exposed (OE) and under-exposed (UE) input images, respectively. Similarly, let $I_{oe}^{LR} \in \mathbb{R}^{h\times w\times 3}$ and $I_{ue}^{LR} \in \mathbb{R}^{h\times w\times 3}$ denote their low-resolution (LR) counterparts, where $H > h$ and $W > w$.

The generative MEF model g operates on these low-resolution inputs to produce a coarse fused image I_{mef}^{LR}. Subsequently, our fine stage implicit neural representation (INR) model employs two encoders: a coarse feature encoder E_{coarse} processing I_{mef}^{LR}, and a fine feature encoder E_{fine} processing the HR exposure

inputs $\{I_{oe}, I_{ue}\}$. The final output of our framework is the high-resolution fused image $I_{\text{mef}}^{HR} \in \mathbb{R}^{H\times W\times 3}$.

3.1 Preliminary

UltraFusion UltraFusion [8], which serves as the backbone for our coarse stage, is a diffusion-based generative MEF framework. Unlike conventional deep learning approaches that directly learn pixel-level fusion mappings, UltraFusion formulates MEF as a guided inpainting problem in the latent space. It leverages the generative prior of a pre-trained diffusion model to synthesize structurally consistent and artifact-free results.

Given an exposure pair $\{I_{oe}, I_{ue}\}$, UltraFusion first estimates the optical flow between the inputs using RAFT [43] to obtain dense motion fields. The UE image is then warped according to the flow to produce an aligned version $\tilde{I}_{ue}$ that spatially matches the OE image. This alignment mitigates ghosting and establishes reliable correspondence for fusion. As illustrated in Fig. 2(a), the aligned exposures $\{I_{oe}, \tilde{I}_{ue}\}$ are typically divided into patches matching the model's pre-training resolution (e.g., 512×512) and processed independently. Within each patch, the OE image acts as the reference, while information from $\tilde{I}_{ue}$ guides the reconstruction of saturated regions. Finally, a VAE decoder reconstructs the fused RGB image from the latent representations with patch guidance. Through this design, UltraFusion combines the semantic reconstruction power of diffusion models with structural guidance, providing a strong foundation for generative fusion.

Local implicit image function (LIIF) The Local implicit image function (LIIF) [7] represents an image as a continuous function mapping spatial coordinates to RGB values. Given an input image I, an encoder E extracts a latent feature map $M = E(I) \in \mathbb{R}^{H\times W\times D}$, where H and W denote the spatial dimensions and D is the feature dimension. The image coordinates are normalized within $[-1, 1]$ for both axes so that any spatial location can be represented in a scale-independent manner. A local feature vector $z \in \mathbb{R}^D$ is obtained by bilinear interpolation near a query coordinate $x = [x_h, x_w] \in [-1, 1]^2$ over the feature map M.

A lightweight MLP decoder f_θ, parameterized by θ, predicts the RGB value $s \in \mathbb{R}^3$ at the coordinate x as:

$$s = f_\theta(z, [x, c]), \tag{1}$$

where $c = [c_h, c_w]$ represents the relative size of the query pixel, referred to as the *cell*. The cell encodes the spatial extent of each query and allows the function to adapt to varying output resolutions.

This implicit representation enables the model to process and integrate image information across different resolutions in a continuous manner. In our framework, we leverage this property to jointly utilize the coarse generative fusion result I_{mef}^{LR} and the high-resolution exposure images $\{I_{oe}, I_{ue}\}$, allowing the model

to predict the RGB value of each coordinate in the target high-resolution fused image I_{mef}^{HR} through unified multi-resolution image fusion.

3.2 LIIFusion

The proposed model, **LIIFusion**, performs generative MEF in a coarse-to-fine manner. The framework consists of two complementary stages: a coarse fusion stage based on UltraFusion [8] for exposure harmonization and ghost removal, and a fine fusion stage that reconstructs high-resolution details using an implicit representation network.

Coarse fusion. In the coarse stage, LIIFusion adopts UltraFusion as the generative prior to perform low-resolution fusion efficiently. Given the high-resolution exposure pair $\{I_{oe}, I_{ue}\}$, both images are first downsampled to obtain $\{I_{oe}^{LR}, I_{ue}^{LR}\}$. Optical flow is estimated using RAFT [43], yielding dense motion fields $F_{oe\to ue}$ and $F_{ue\to oe}$. These fields are used to warp both the high- and low-resolution under-exposed images, producing $\{\tilde{I}_{ue}^{HR}, \tilde{I}_{ue}^{LR}\}$. The aligned low-resolution pair $\{I_{oe}^{LR}, \tilde{I}_{ue}^{LR}\}$ is then processed by the diffusion model g to produce a coarse fused image:

$$I_{\text{mef}}^{LR} = g(I_{oe}^{LR}, \tilde{I}_{ue}^{LR}). \tag{2}$$

Fine fusion. The fine stage reconstructs the target high-resolution fused image I_{mef}^{HR} by integrating features across resolutions. The coarse feature encoder E_{coarse} extracts representations $f_{\text{coarse}} = E_{\text{coarse}}(I_{\text{mef}}^{LR})$ from the generative output, while the fine feature encoder E_{fine} encodes the concatenated high-resolution exposures as $f_{\text{fine}} = E_{\text{fine}}([I_{oe}, \tilde{I}_{ue}^{HR}])$.

To predict the RGB value at an arbitrary query coordinate x in the high-resolution domain, we first obtain local feature vectors z_{coarse} and z_{fine} by performing bilinear interpolation on f_{coarse} and f_{fine} at location x. These vectors are concatenated and decoded by a lightweight MLP f_θ:

$$s = f_\theta([z_{\text{coarse}}, z_{\text{fine}}], [x, c]). \tag{3}$$

The final image I_{mef}^{HR} is formed by querying all coordinates in the target resolution. The network is trained using an ℓ_1 reconstruction loss against the ground truth I_{GT}:

$$\mathcal{L}_{\text{fusion}} = \|I_{\text{mef}}^{HR} - I_{GT}\|_1. \tag{4}$$

This design allows the diffusion-based coarse stage to handle exposure inconsistencies and motion artifacts globally, while the implicit representation refines geometric and texture details locally at full resolution.

3.3 Adaptive exposure correction

Diffusion-based generative fusion models typically perform exposure blending via guided inpainting in the latent space [63]. UltraFusion [8] employs a fidelity

guidance stage within the VAE decoder, utilizing input exposures to ensure structural consistency. While effective, this design struggles when the over-exposed (OE) input is severely saturated, rendering the guidance unreliable. This issue is exacerbated at low resolutions, where the model is required to distinguish fine details within a limited receptive field. To address this, we propose an adaptive exposure correction module that refines the OE image prior to its use in fidelity guidance, thereby providing a stable structural cue for the generative decoder.

Exposure difference estimation. Given normalized exposure pairs $I_{oe}^{LR}, I_{ue}^{LR} \in [0,1]^{h\times w\times 3}$, we first convert them to the LAB color space and extract their luminance channels L_{oe} and L_{ue}. An exposure difference map is computed as:

$$L_{\text{diff}} = \min(\max(L_{oe} - L_{ue}, 0), 1), \tag{5}$$

where L_{diff} highlights saturated regions exhibiting large luminance disparities. We explicitly clip values to $[0,1]$ for numerical stability.

Adaptive brightness adjustment. A per-pixel weight map is generated to modulate excessively bright regions in I_{oe}^{LR}:

$$W = (1 - \alpha \cdot L_{\text{diff}})^{1/2.2}, \tag{6}$$

where α is a scalar controlling the strength of compensation and the exponent $1/2.2$ applies perceptual gamma correction to smooth the adjustment curve [3]. The adjusted OE image is then obtained via:

$$I_{oe}'^{LR} = I_{oe}^{LR} \odot W, \tag{7}$$

where $\odot$ denotes element-wise multiplication. Finally, $I_{oe}'^{LR}$ replaces the original I_{oe}^{LR} during the fidelity guidance stage of the generative fusion pipeline.

4 Experiments

4.1 Experimental setting

Datasets. We construct our training dataset from the SICE dataset [4] by randomly selecting under- and over-exposed image pairs from each exposure bracket. For evaluation, we assess the performance of our model on both static and dynamic datasets, including MEFB [57] with 100 static exposure pairs, the UltraFusion Benchmark [8] containing 100 real-captured under/over-exposed pairs with larger exposure differences, and the RealHDRV dataset [40] that captures 50 dynamic scenes with diverse motions.

Implementation details. Since the SICE dataset contains static scenes, we follow UltraFusion's synthetic setup by multiplying an occlusion mask extracted from Vimeo-90K [51] with the UE images to simulate dynamic misalignment. SICE label data is used as ground truth, and we downsample them to random

Table 2: Quantitative comparison on the dynamic RealHDRV dataset [40] and UltraFusion Benchmark [8]. **Bold** denote the best performance and underlined indicate the second best. Total inference time for each dataset is measured on a single NVIDIA RTX A5000.

Model	RealHDRV (50 scenes)					UltraFusion Benchmark (100 scenes)				
	MUSIQ↑	DeQA-Score↑	PAQ2PIQ↑	HyperIQA↑	Time↓	MUSIQ↑	DeQA-Score↑	PAQ2PIQ↑	HyperIQA↑	Time↓
Defusion [25]	56.38	3.2867	68.31	0.4838	2 min	60.11	3.3529	71.83	0.5440	6 min
MEF-LUT [14]	62.42	3.2864	70.04	0.5020	4 sec	64.06	3.2859	71.80	0.5103	8 sec
HSDS-MEF [46]	61.82	3.6045	71.14	0.5055	18 min	65.23	3.6662	73.77	0.5786	46 min
UltraFusion [8]	<u>67.54</u>	**3.8998**	<u>73.39</u>	<u>0.5834</u>	101 min	<u>68.40</u>	**4.0123**	<u>75.18</u>	<u>0.6214</u>	203 min
Ours	**69.52**	<u>3.8908</u>	**74.06**	**0.6175**	27 min	**70.19**	<u>3.9807</u>	**75.59**	**0.6467**	59 min

Table 3: Quantitative comparison on the static MEFB dataset [57].

Model	MEFB (100 scenes)			
	MUSIQ↑	PAQ2PIQ↑	HyperIQA↑	MEF-SSIM↑
MEF-GAN [50]	51.28	70.58	0.3974	0.7197
U2Fusion [49]	64.09	72.34	0.5436	0.9212
Defusion [25]	62.70	70.82	0.5454	0.9102
MEF-LUT [14]	65.71	71.21	0.5267	0.9005
HSDS-MEF [46]	66.81	72.60	0.6026	**0.9367**
UltraFusion [8]	68.14	73.45	0.6310	0.9266
Ours	**68.78**	**73.69**	**0.6418**	0.9201

Table 4: FLOPs and parameter comparison. FLOPs are measured on a single 1988×1326 image from UltraFusion Benchmark.

Model	Parameters	TFLOPs
MEF-GAN	0.488 M	2.483
U2Fusion	0.659 M	3.477
Defusion	7.874 M	0.640
MEFLUT	0.061 M	0.003
HSDS-MEF	1.166 M	2.307
UltraFusion	1.860 B	944.4
Ours	1.872 B	282.1

scale to obtain corresponding triplet $\{I_{oe}, I_{ue}, I_{\text{mef}}^{LR}\}$ for training. The fine encoder E_{fine} receives 19×19 local patches around each target pixel from I_{oe} and I_{ue} to encode local exposure features, implemented as a 4-layer CNN. We adopt a SwinIR [24] backbone for the coarse encoder E_{coarse} such as convention of LIIF [7]. We use the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 16. The model is trained for 500 epochs on a single NVIDIA H100 GPU, requiring approximately 22 hours for full convergence.

Evaluation metrics. To evaluate both structural preservation and perceptual naturalness, we employ MEF-SSIM [30] as a structure aware fusion metric, along with four non reference image quality assessment (NRIQA) metrics: MUSIQ [16], PAQ2PIQ [54], DeQA-Score [55], and HyperIQA [42], which measure perceptual quality. MEF aims to produce a content-complete, well tone-mapped image, yet there is no unique ground truth for what constitutes a "good" tone mapping; therefore, we use NRIQA to assess perceptual naturalness while relying on MEF-SSIM to verify structural preservation. Using these metrics, we quantitatively compare LIIFusion with existing conventional and generative MEF models across the aforementioned datasets. We also report total inference time per dataset, as well as the number of parameters and FLOPs for a single image, to assess the computational efficiency of our approach compared with patch-wise generative MEF methods.

4.2 Quantitative results

Tab. 2 reports results on RealHDRV and UltraFusion Benchmark, which contain large exposure differences and diverse object motions. Our model consistently outperforms conventional approaches such as MEF-LUT [14] and HSDS-MEF [59] across all quality metrics, while achieving comparable or even superior performance to the generative baseline UltraFusion [8]. These results indicate that LIIFusion effectively handles occluded or saturated regions in dynamic scenes by adaptively integrating pixel-wise features through its implicit representation. Moreover, our approach requires only one fourth of UltraFusion's inference time, demonstrating that the proposed coarse-to-fine framework performs high-quality fusion far more efficiently than patch-wise generative models, even though lightweight methods such as MEF-LUT remain faster overall but fall short in quality. As summarized in Table 4, adding the fine fusion module slightly increases parameters over UltraFusion, but our coarse-to-fine design reduces TFLOPs by more than 3× for a single image, preserving diffusion-driven quality while alleviating runtime bottlenecks. Tab. 3 further presents results on the static MEFB dataset, where LIIFusion consistently outperforms baselines across all non-reference quality metrics. These results confirm that LIIFusion provides a strong balance between fidelity and efficiency, preserving generative-quality fusion while enabling faster inference for high resolution multi exposure imaging.

4.3 Qualitative results

We further conduct qualitative comparisons on ultra high dynamic range scenes from the UltraFusion Benchmark and dynamic motion scenes from the RealHDRV dataset. For fair comparison, we visualize results from MEF-LUT [14], HSDS-MEF [59], UltraFusion [8], and our LIIFusion.

Ultra HDR scenes. As shown in Fig. 3, MEF-LUT produces inconsistent tone transitions due to its LUT-based fusion, resulting in sky–ground discontinuities and unnatural illumination in building regions. HSDS-MEF, though content-adaptive, fails to fully cover the full dynamic range and often produces dark or low-contrast results. Both UltraFusion and our method generate visually natural outputs while our approach achieves these comparable results at one fourth of UltraFusion's inference time.

Dynamic motion scenes. Fig. 4 presents results on large-motion scenarios. MEF-LUT and HSDS-MEF suffer from severe ghosting artifacts when aligning moving objects, while UltraFusion produces smoother results via guided inpainting but still struggles in extreme motion regions. Because UltraFusion operates on patch-wise latent inpainting, its contextual range is limited, sometimes leading to uncertain synthesis in occluded areas. In contrast, LIIFusion performs coarse fusion over the entire downsampled image, enabling globally consistent generation and effective utilization of diffusion priors. As shown in the

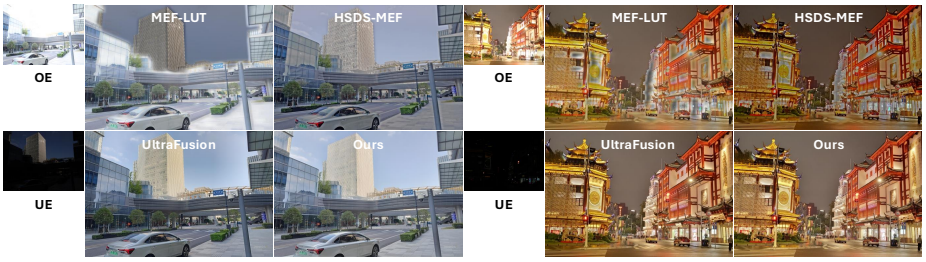

Fig. 3: Qualitative comparison on ultra HDR scenes in the UltraFusion Benchmark. Compared to prior MEF models, our approach achieves more stable exposure fusion, retaining structural detail and natural color appearance in both day and night scenes.

Fig. 4: Qualitative results on the RealHDRV dataset, showing that our method provides cleaner fusion with improved guided inpainting, robust occlusion handling, and reduced ghosting.

highlighted regions, our model successfully removes ghost artifacts and restores missing structural details, confirming its efficiency and robustness in complex dynamic scenes.

4.4 Ablation study

All ablation studies except Sec. 4.4 are conducted on the UltraFusion Benchmark dataset to ensure a consistent evaluation environment across different configurations.

Adaptive exposure correction ablation We first evaluate the contribution of the proposed adaptive exposure correction (AEC). As shown in Tab. 5, our method achieves higher overall image quality than UltraFusion regardless of whether AEC is applied. While the numerical gain from AEC is relatively modest, its qualitative impact is substantially more pronounced.

AEC is designed to address a core difficulty of low resolution generative fusion: saturated regions in over-exposed inputs often lack recoverable structure, causing the diffusion model to hallucinate blurred or distorted content during coarse fusion. To examine this effect, Fig. 5 visualizes the low resolution MEF

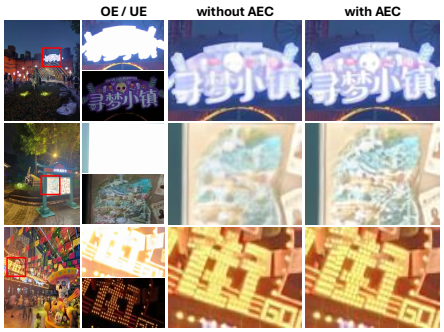

Fig. 5: Visual ablation of the adaptive exposure correction module on the UltraFusion Benchmark at low resolution, showing improved reconstruction of details lost in saturated OE regions.

Table 5: Ablation study of the adaptive exposure correction.

Method	MUSIQ	PAQ2PIQ	HyperIQA	DeQA-Score
UltraFusion	68.40	75.18	0.6214	**4.0123**
Ours (w/o AEC)	70.07	**75.64**	0.6451	3.9870
Ours (w/ AEC)	**70.19**	75.59	**0.6467**	3.9807

Table 6: Ablation study of LIIF within the multi-exposure fusion framework. LIIF (pre-trained) is trained on DIV2K [1] and LIIF (from scratch) is trained on SICE dataset [4].

Method	MEF-SSIM	HyperIQA
LIIF (pre-trained)	0.8942	0.6389
LIIF (from scratch)	0.9172	0.6230
Ours	**0.9201**	**0.6418**

outputs before fine refinement. Without AEC, the model struggles to synthesize coherent structures in heavily saturated areas, where neon signs become blurred, map details on display boards collapse, and bright decorative lights lose their shape. These artifacts originate in the coarse stage and cannot be fully restored by the fine fusion module, which refines but does not reinterpret the underlying structure.

With AEC, however, the saturated regions are attenuated before entering the diffusion model, allowing the coarse generative stage to preserve the essential structure of challenging areas. As a result, the LR fusion already contains clearer letter shapes, map boundaries, and high-frequency lighting patterns. When passed to the fine-stage implicit refinement, these structurally reliable coarse predictions lead to high resolution fusion results where both the global scene layout and fine details remain well aligned and visually coherent.

Overall, AEC strengthens the structural reliability of the coarse generative fusion, enabling the fine stage to more effectively restore details and produce consistent HDR-like outputs.

Comparison between LIIFusion and LIIF We further examine the effectiveness of adapting the LIIF architecture for MEF rather than conventional super resolution. Tab. 6 compares three configurations: (1) the pre-trained LIIF trained on DIV2K [1], (2) a LIIF model trained from scratch on SICE for super resolution, and (3) our proposed LIIFusion trained for multi exposure fusion. This ablation is performed on MEFB to measure structure consistency.

As shown in Tab. 6, the first two models achieve comparable results, indicating that conventional super resolution training, even with exposure data, does not provide sufficient cues for accurate fusion. In contrast, our LIIFusion improves all metrics including MEF-SSIM for structural fidelity and HyperIQA for

Fig. 6: Visual comparison between the pre-trained LIIF and our LIIFusion on the UltraFusion Benchmark. Our LIIFusion restores finer textual and structural details, such as characters and patterns, which are blurred or lost in the pre-trained LIIF outputs.

perceptual quality, demonstrating the benefit of learning to fuse multi exposure inputs rather than simply upsampling the coarse MEF result.

The qualitative comparison in Fig. 6 further supports this observation. The pre trained LIIF models produce blurry textures and fail to recover fine details such as text and small structural elements. UltraFusion, though structurally strong due to its generative prior, still lacks pixel-level consistency in high-frequency areas. Our model, on the other hand, synthesizes sharper and more coherent details, accurately reconstructing neon letters and map textures visible in zoomed-in regions. This confirms that framing LIIF as a fine stage fusion module, rather than as a simple super resolution network, enables precise pixel wise integration of exposure information, effectively complementing the coarse generative fusion stage.

Training dataset We additionally examine whether LIIFusion can be trained using pseudo labels instead of human-crafted ground truth. MEF datasets are difficult to scale because generating high-quality fused labels typically requires manual tone mapping or expert-designed fusion strategies. To explore whether synthetic annotations are sufficient for training, we replace the original SICE labels with the fused outputs generated by UltraFusion.

As shown in Tab. 7, training with UltraFusion pseudo labels yields performance that is highly competitive with the model trained on the original SICE labels, with only marginal differences across all evaluation metrics. A qualitative comparison in Fig. 7 further illustrates that the pseudo labels maintain comparable visual fidelity to the provided SICE labels, capturing natural tones and detailed structures without noticeable degradation.

These results indicate that LIIFusion can be effectively trained using synthetic supervision alone, reducing the dependency on manually curated MEF annotations. This also suggests that stronger generative models can serve as scalable label generators, enabling future expansion of Multi Exposure Fusion datasets without additional human labeling effort.

Fig. 7: Qualitative comparison between the original SICE labels and our generated pseudo labels.

Table 7: Quantitative comparison between using the original SICE label and our generated pseudo labels.

Dataset	MUSIQ	DeQA-Score	PAQ2PIQ	HyperIQA
SICE label	70.19	3.9807	75.59	0.6467
SICE pseudo label	69.34	3.9688	75.19	0.6323

Table 8: Ablation study on exposure inputs for fine fusion. **Bold** denote the best performance and underlined indicate the second best.

	MUSIQ	DeQA-Score	PAQ2PIQ	HyperIQA
w/o OE, UE	54.65	3.7961	74.09	0.4277
w/ UE only	62.73	3.8412	74.59	0.5773
w/ OE only	68.38	3.8438	75.58	0.6141
w/ OE, UE	**70.19**	**3.9807**	**75.59**	**0.6467**

Fig. 8: Comparison of different exposure-conditioning strategies for LIIF. Using both OE and UE exposures preserves fine details and reduces residual artifacts compared with configurations using only one exposure or none.

Ablation on exposure inputs for fine fusion To further examine whether LIIFusion performs true multi-exposure fusion rather than simple super-resolution, we analyze the contribution of each exposure input in the fine fusion stage. Specifically, we evaluate four input configurations for the LIIF module: (1) using neither OE nor UE, (2) using UE only, (3) using OE only, and (4) using both OE and UE. To isolate the effect of exposure cues, we replace the missing OE input with an all-white image and the missing UE input with an all-black image.

As summarized in Tab. 8, performance consistently improves as more exposure inputs are provided. Using only the coarse LR MEF without any exposure cues results in a significant drop across all metrics, confirming that LIIFusion depends on exposure-conditioned refinement rather than purely upsampling the coarse output. Among the two exposures, OE contributes the most, likely because over-exposed images retain more structure in bright regions where LR MEF tends to blur details. UE also provides complementary information but has a smaller effect individually.

As shown in Fig. 8, when the fine fusion module receives only the coarse result, the input lacks sufficient detail and the output remains blurry. Using UE alone can help recover details in saturated regions (e.g., the second scene),

but in the first and third scenes it tends to transfer UE noise. Using OE alone preserves HR details from the OE input, yet it cannot fully restore saturated regions. In contrast, using UE, OE, and the coarse result together yields the most coherent fusion. These results indicate that the fine fusion module adapts to the characteristics of each input and effectively integrates their complementary information to produce the final output.

5 Discussion and future work

Although LIIFusion offers clear benefits in both efficiency and fusion quality, several limitations remain. As shown in Fig. 5, the final output is still influenced by the structural reliability of the coarse generative stage. In addition, despite the reduced computational cost, our approach is still slower than lightweight methods such as MEF-LUT, limiting its applicability in real-time scenarios. Finally, in scenes with extremely large occlusions or rapid motion, our model may still produce subtle ghost artifacts, suggesting the need for more motion-aware generative mechanisms.

For future work, our findings suggest that Local Implicit Image Functions provide a flexible representation for a broad class of image synthesis and restoration tasks. By allowing the implicit model to perform fine-detail fusion with guided inputs while the diffusion module manages coarse structure, our coarse-to-fine formulation reduces the computational and structural burdens commonly seen in diffusion-based low-level vision.

6 Conclusion

In this work, we introduced LIIFusion, a novel coarse-to-fine framework for generative Multi-Exposure Fusion that combines diffusion-based global harmonization with pixel-level implicit refinement. Our approach addresses the fundamental trade-off between generative power and structural fidelity by performing coarse fusion at a low resolution using a diffusion prior, then restoring high-resolution details through a resolution-agnostic implicit representation. This architecture effectively bridges the gap between global exposure alignment and fine-grained structural reconstruction. Extensive experimental results and ablation studies demonstrate that LIIFusion delivers fusion quality comparable to, and in many cases exceeding, state-of-the-art diffusion baseline while requiring only one-quarter of their inference time. By providing a more efficient yet powerful synthesis paradigm, we hope that this framework provides a practical step toward making generative Multi-Exposure Fusion more feasible for real-world, high-resolution imaging applications and sets a new precedent for the integration of INRs in generative workflows.

Coarse-to-fine Framework for Generative MEF via Implicit Neural Representation

Supplementary Material

A Additional ablation study

We provide additional ablation studies in this section. All experiments are conducted on the UltraFusion Benchmark [8] except for Sec. A.3, and evaluation is performed using the non-reference image quality assessment metrics MUSIQ [16], PAQ2PIQ [54], and HyperIQA [55]. These studies complement the main paper by examining architectural design choices, encoder configurations, and the behavior of our multi-exposure fusion (MEF) modules under various settings.

A.1 Synthetic data augmentation

Building on our observation in Tab. 6 of the main paper that LIIFusion can be trained reliably using synthetic MEF results as pseudo labels, we further investigate whether this property can be exploited to expand the training data beyond SICE. Since the ICCV23 multi-exposure dataset [44] provides only exposure brackets without fused ground truth, we generate synthetic MEF outputs using UltraFusion and treat them as pseudo labels. We then retrain LIIFusion on the combination of the original SICE dataset and the ICCV23 pseudo-labeled set.

As shown in Tab. A, augmenting the training set with ICCV23 pseudo labels yields slight improvements in MUSIQ and HyperIQA, indicating that synthetic supervision remains effective even when the additional images are collected from a different dataset. This experiment suggests that LIIFusion can benefit from datasets without MEF ground truth by leveraging synthetic MEF generation, and highlights the potential of our framework as a practical tool for scaling training data in scenarios where ground truth fused images are not available.

A.2 Backbone of coarse feature encoder

We evaluate different backbone choices for the coarse encoder used to extract features from the low resolution fused image I_{mef}^{LR}. Following common practice in LIIF-based architectures, we test three representative encoders: EDSR-baseline [26], RDN [58], and SwinIR [24]. Tab. B compares their performance and inference time.

While all backbones exhibit comparable performance, SwinIR achieves the best overall trade-off. It achieves the highest scores on MUSIQ and PAQ2PIQ and remains competitive on HyperIQA, while also providing the fastest inference time. Based on this favorable balance of efficiency and quality, we adopt SwinIR as the coarse encoder in LIIFusion.

Table A: Effect of synthetic data augmentation.

Dataset	MUSIQ↑	PAQ2PIQ↑	HyperIQA↑
SICE label	70.19	**75.59**	0.6467
+ ICCV23 pseudo label	**70.22**	75.46	**0.6506**

Table B: Ablation study on backbone of coarse feature encoder.

Model	MUSIQ↑	PAQ2PIQ↑	HyperIQA↑	Time↓
EDSR-baseline	69.77	<u>75.52</u>	0.6351	69 min
RDN	<u>70.08</u>	75.51	**0.6473**	71 min
SwinIR (Ours)	**70.19**	**75.59**	<u>0.6467</u>	59 min

Table C: Ablation study on α of adaptive exposure correction on the MEFB dataset. Across different α values, there is a trade-off between structural preservation and fusion quality, and we choose the optimal setting $\alpha = 0.2$.

Value of α	MUSIQ	PAQ2PIQ	HyperIQA	MEF-SSIM
1.0	65.41	72.23	0.6017	0.8633
0.8	66.71	72.73	0.6100	0.9196
0.6	67.27	72.94	0.6181	0.9322
0.4	67.60	73.11	0.6227	**0.9335**
0.2 (Ours)	67.91	73.28	0.6265	0.9313
0.0 (UltraFusion)	**68.14**	**73.45**	**0.6309**	0.9266

Fig. A: Comparison of high resolution MEF results. With AEC and our fine feature encoder, LIIFusion preserves far more structure in saturated regions, recovering text and map details that become blurred or lost in UltraFusion.

A.3 Ablation on adaptive exposure correction

The adaptive exposure correction (AEC) is designed to alleviate the structural degradation that occurs when performing multi-exposure fusion at low resolution. As shown in Fig. 5 of the main paper, saturated regions in the over-exposed (OE) image often lose fine structures, yielding blurry or collapsed details that the generative MEF model cannot reliably recover. By attenuating overly bright pixels prior to coarse fusion, AEC restores structure that would otherwise be lost in these regions.

To determine the optimal correction strength, we vary the hyperparameter α, which controls Eq. (6) in the main paper. Larger values of α impose stronger suppression on saturated areas. We conduct this ablation on the MEFB dataset [57] using the low-resolution MEF outputs, as these directly reflect the structural fidelity produced by the coarse stage.

Tab. C shows that the non-reference image quality assessment metrics generally favor the setting without correction ($\alpha = 0.0$), consistent with the behavior of the original UltraFusion model. In contrast, the MEF-SSIM [30] score, which captures structural consistency, improves whenever AEC is applied and reaches its best value around $\alpha = 0.4$. Considering both perceptual quality and structure preservation, we select $\alpha = 0.2$ as the default setting, which provides a balanced performance across all metrics.

As further illustrated in Fig. A, even mild correction enhances the reconstruction of saturated details such as neon signage and high-intensity textures. These results confirm that AEC strengthens the robustness of the coarse fusion stage and supplies more structurally reliable inputs for the subsequent fine fusion process.

B Implementation details for inference

Our inference pipeline follows a coarse to fine procedure that integrates low resolution generative fusion with high-resolution implicit function based pixel level fusion. A central requirement of this pipeline is maintaining perfect pixel alignment between the low resolution fusion stage and the high-resolution fine fusion stage. The full inference process is described below.

We begin by applying zero padding to the high-resolution OE and UE images so that both become square. This step is critical for preserving consistent spatial scaling across axes. Without padding, resizing the shorter edge to a fixed size such as 512 leads to a non integer scale factor for the longer edge (for example, 773.3 becomes 773 after rounding). Even this fraction of a pixel introduces misalignment between the fused low resolution output and the original high-resolution exposures. Such sub pixel drift propagates into the fine fusion stage and results in ghosting artifacts as shown in Fig. B. Padding the images into a square ensures that both dimensions share the same integer scaling factor, preventing this issue.

The padded images are then downsampled to the target resolution (e.g., 512×512 or 768×768), producing low resolution OE and UE. Optical flow is computed between these two low resolution exposures, and the low resolution UE is backward warped to align with the low resolution OE. The same flow field is then upsampled and applied to the high-resolution UE image, ensuring alignment consistency between the low resolution and high-resolution branches.

Coarse fusion is performed using the aligned low resolution exposures through the generative MEF module. This stage resolves large exposure differences, suppresses occlusions, and generates globally coherent structures at a low computational cost due to the reduced spatial resolution. The coarse fusion output, along with the high-resolution OE and the aligned high-resolution UE, is then processed by our LIIF based fine stage. This implicit function based pixel level fusion refines the coarse output by synthesizing high-frequency textures and restoring fine structural details that are not recoverable through generative fusion alone.

Fig. B: Effect of square padding during inference. Without padding, subpixel misalignment leads to blurred fine details. With padding, alignment is preserved and details remain sharp.

Finally, we remove the original zero padding and crop the fused output back to its original aspect ratio, producing the final high-resolution MEF result. This inference strategy maintains strict pixel alignment across all stages, avoids sub pixel inconsistencies caused by uneven resizing, and effectively combines global generative reasoning with precise local pixel level fusion to produce high-quality multi exposure results.

C Justification of using LIIF in the fine stage

We choose LIIF [7] as the fine fusion module. This model choice is for efficient fusion guided by a long-standing super resolution (SR) trend : (i) avoid dense HR-grid computation and (ii) keep the refinement module resolution-agnostic. FSRCNN [11] motivates "process-then-upsample" by showing that "upsample-then-process" becomes expensive on HR grids, reporting $> 40\times$ speedup over SRCNN [10]'s upsample-then-process pipeline. LIIF takes the next step: instead of committing to discrete scales (e.g., $\times 2/\times 3/\times 4$, requiring scale-specific parameters), it represents the output as a continuous function queried at arbitrary coordinates, so one set of parameters supports any scale without retraining or storing multiple scale checkpoints. This directly matches MEF deployment, where target resolution varies by sensor and post-crop. Importantly, LIIF here is not a drop-in SR head: our ablations show that a pretrained (SR-trained) LIIF is insufficient for MEF, while LIIFusion training improves MEF metrics (MEF-SSIM / HyperIQA).

D Additional results

This section provides additional qualitative examples highlighting robustness, fine detail and ultra high dynamic range, presented in Figs. C-G.

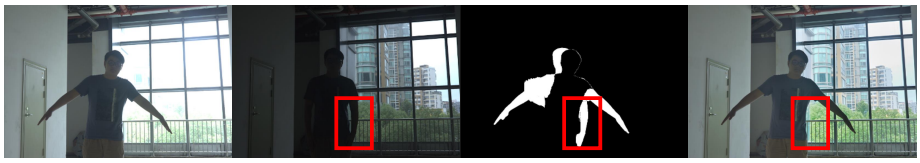

Fig. C: Robust fusion with inaccurate flow-based masks. Our LR-domain flow estimation improves reliability (main paper Fig. 4), yet correspondences may still be imperfect under extreme exposure gaps. Even in such cases, our method maintains a stable fusion result.

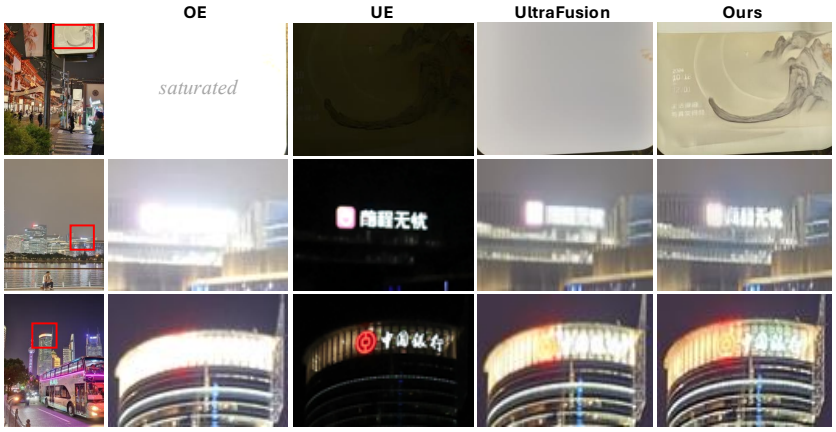

Fig. D: Comparison of additional qualitative results. The red-boxed regions highlight differences in detail preservation and appearance between UltraFusion and our method.

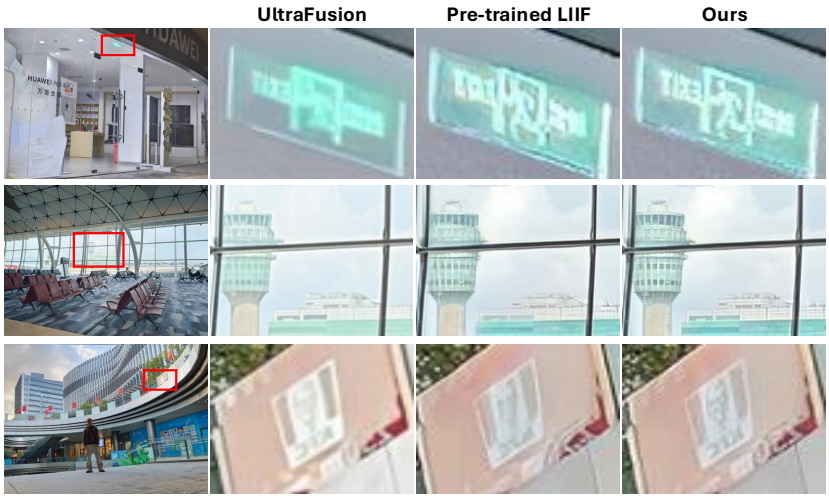

Fig. E: Comparison of additional qualitative results for LIIF. Our approach retains finer textures and more stable structural details across scenes.

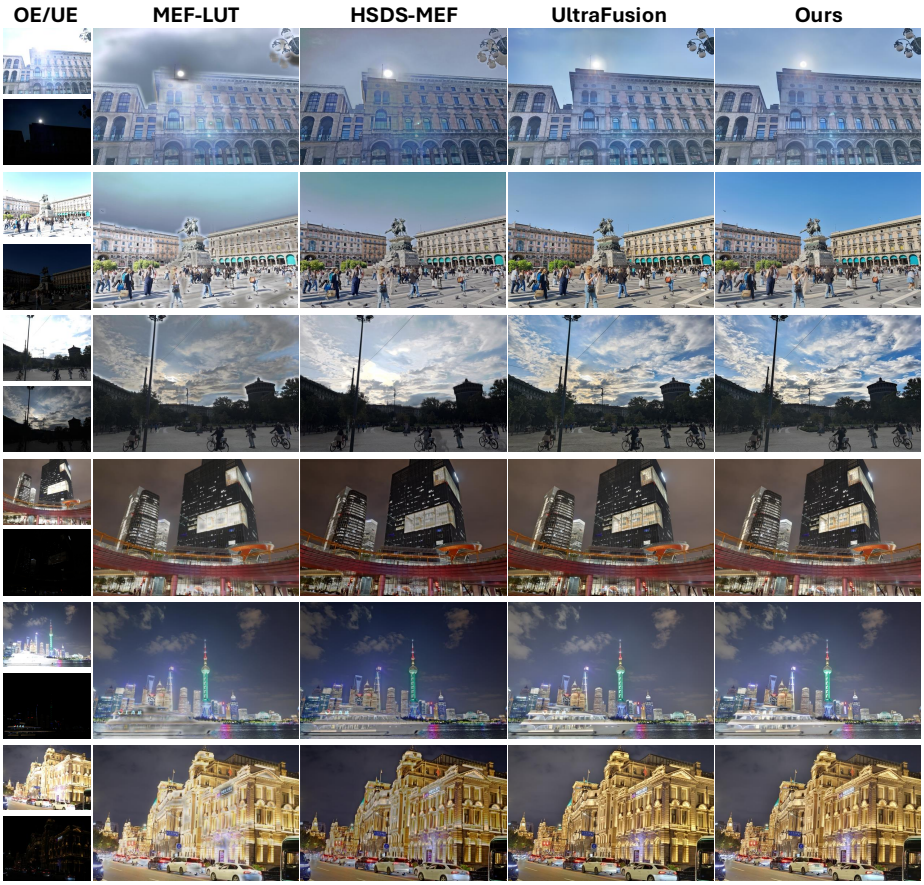

Fig. F: Comparison of additional ultra high dynamic range results, showing improved preservation of bright areas and enhanced contrast.

Fig. G: Comparison on additional multi exposure fusion results in dynamic scenes, showing that our method effectively suppresses motion-induced ghost artifacts while preserving scene details and contrast.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshops. pp. 126–135 (2017)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
3. Bemana, M., Leimkuehler, T., Myszkowski, K., Seidel, H.P., Ritschel, T.: Bracket diffusion: Hdr image generation by consistent ldr denoising. Computer Graphics Forum (Proceedings of Eurographics 2025) **44**(2) (2025), arXiv:2405.14304
4. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE transactions on image processing **27**(4), 2049–2062 (2018)
5. Cao, J., Wang, Q., Xian, Y., Li, Y., Ni, B., Pi, Z., Zhang, K., Zhang, Y., Timofte, R., Van Gool, L.: Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1796–1807 (2023)
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 **31**, 2661–2672 (2022)
7. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8628–8638 (2021)
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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16111–16121 (2025)
9. Chobola, T., Liu, Y., Zhang, H., Schnabel, J.A., Peng, T.: Fast context-based low-light image enhancement via neural implicit representations. In: European Conference on Computer Vision. p. 413–430 (2024), arXiv:2407.12511
10. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
11. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: European Conference on Computer Vision. pp. 391–407. Springer (2016)
12. Fang, W., Tang, Y., Guo, H., Yuan, M., Mok, T.C.W., Yan, K., Yao, J., Chen, X., Liu, Z., Lu, L., Zhang, L., Xu, M.: Cycleinr: Cycle implicit neural representation for arbitrary-scale volumetric super-resolution of medical data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11631–11641 (2024)
13. Feng, Y., Shang, Y., Li, X., Shao, T., Jiang, C., Yang, Y.: Pie-nerf: Physics-based interactive elastodynamics with nerf. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4450–4461 (2024)
14. Jiang, T., Wang, C., Li, X., Li, R., Fan, H., Liu, S.: Meflut: Unsupervised 1d lookup tables for multi-exposure image fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10542–10551 (2023)

15. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Transactions on Graphics (TOG) **36**(4), 144:1–144:12 (2017)
16. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
17. Kim, J., Kim, T.K.: Arbitrary-scale image generation and upsampling using latent diffusion model and implicit neural decoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9202–9211 (2024)
18. Kim, S., Min, Y., Jung, Y., Kim, S.: Controllable style transfer via test-time training of implicit neural representation. arXiv preprint arXiv:2210.07762 (2022)
19. 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. pp. 256–273. Springer (2024)
20. Lee, J., Jin, K.H.: Local texture estimator for implicit representation function. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1929–1938 (2022)
21. Li, H., Ma, K., Yong, H., Zhang, L.: Fast multiscale structural patch decomposition for multi-exposure image fusion. IEEE Transactions on Image Processing **29**, 5805–5816 (2020)
22. Li, X., Ni, Z., Yang, W.: Afunet: Cross-iterative alignment–fusion synergy for hdr reconstruction via deep unfolding paradigm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025), arXiv:2506.23537
23. Li, Z., Wei, Z., Wen, C., Zheng, J.: Detail-enhanced multi-scale exposure fusion. IEEE Transactions on Image Processing **26**(3), 1243–1252 (2017)
24. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1833–1844 (2021)
25. Liang, P., Jiang, J., Liu, X., Ma, J.: Fusion from decomposition: A self-supervised decomposition approach for image fusion. In: European Conference on Computer Vision. pp. 719–735. Springer (2022)
26. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition workshops. pp. 136–144 (2017)
27. Liu, Z., Wang, Y., Zeng, B., Liu, S.: Ghost-free high dynamic range imaging with context-aware transformer. In: European Conference on Computer Vision. pp. 344–360. Springer (2022)
28. Ma, K., Duanmu, Z., Zhu, H., Fang, Y., Wang, Z.: Deep guided learning for fast multi-exposure image fusion. IEEE Transactions on Image Processing **29**, 2808–2819 (2019)
29. 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 **26**(5), 2519–2532 (2017)
30. Ma, K., Zeng, K., Wang, Z.: Perceptual quality assessment for multi-exposure image fusion. IEEE Transactions on Image Processing **24**(11), 3345–3356 (2015)
31. Mertens, T., Kautz, J., Reeth, F.V.: Exposure fusion. In: Pacific Conference on Computer Graphics and Applications (PG'07). pp. 382–390 (2007)
32. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421 (2020)

33. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14314–14323 (2021)
34. Prabhakar, K.R., Agrawal, S., Singh, D.K., Ashwath, B., Babu, R.V.: Towards practical and efficient high-resolution hdr deghosting with cnn. In: European Conference on Computer Vision. pp. 497–513 (2020)
35. Prabhakar, K.R., Arora, R., Swaminathan, A., Singh, K.P., Babu, R.V.: A fast, scalable, and reliable deghosting method for extreme exposure fusion. In: IEEE International Conference on Computational Photography. pp. 1–8 (2019)
36. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
37. Ram Prabhakar, K., Sai Srikar, V., Venkatesh Babu, R.: Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4714–4722 (2017)
38. Sabour, S., Vora, S., Duckworth, D., Krasin, I., Fleet, D.J., Tagliasacchi, A.: Robustnerf: Ignoring distractors with robust losses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20626–20636 (2023)
39. Shi, K., Zhou, X., Gu, S.: Improved implicit neural representation with fourier reparameterized training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25985–25994 (2024)
40. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2879–2888. IEEE (2024), realHDRV dataset
41. 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. pp. 288–304 (2022)
42. 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3667–3676 (2020)
43. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision. pp. 402–419. Springer (2020)
44. Tel, S., Wu, Z., Zhang, Y., Heyrman, B., Demonceaux, C., Timofte, R., Ginhac, D.: Alignment-free hdr deghosting with semantics consistent transformer. arXiv preprint arXiv:2305.18135 (2023)
45. Wang, C., Xia, Z., Leimkuehler, T., Myszkowski, K., Zhang, X.: Lediff: Latent exposure diffusion for hdr generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 453–464 (2025), arXiv:2412.14456
46. Wu, G., Fu, H., Liu, J., Ma, L., Fan, X., Liu, R.: Hybrid-supervised dual-search: Leveraging automatic learning for loss-free multi-exposure image fusion. In: Proceedings of the AAAI conference on artificial intelligence. pp. 5985–5993 (2024)
47. 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. pp. 117–132 (2018)
48. Xu, G., Wang, Y., Gu, J., Xue, T., Yang, X.: Hdrflow: Real-time hdr video reconstruction with large motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24851–24860 (2024)

49. 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 **44**(1), 502–518 (2020)
50. Xu, H., Ma, J., Zhang, X.P.: Mef-gan: Multi-exposure image fusion via generative adversarial networks. IEEE Transactions on Image Processing **29**, 7203–7216 (2020)
51. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**(8), 1106–1125 (2019)
52. Yan, Q., Gong, D., Shi, Q., van den Hengel, A., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1751–1760 (2019)
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 **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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14483–14494 (2025)
56. Yuan, W., Zhu, Q., Liu, X., Ding, Y., Zhang, H., Zhang, C.: Sobolev training for implicit neural representations with approximated image derivatives. In: European Conference on Computer Vision, pp. 72–88. Springer (2022)
57. Zhang, X.: Benchmarking and comparing multi-exposure image fusion algorithms. Information Fusion **74**, 111–131 (2021), mEFB dataset
58. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2472–2481 (2018)
59. Zhao, Z., Deng, L., Bai, H., Cui, Y., Zhang, Z., Zhang, Y., Qin, H., Chen, D., Zhang, J., Wang, P., et al.: Image fusion via vision-language model. arXiv preprint arXiv:2402.02235 (2024)
60. Zheng, J., Li, Z.: Superpixel based patch match for differently exposed images with moving objects and camera movements. In: IEEE International Conference on Image Processing. pp. 4516–4520 (2015)
61. Zheng, J., Li, Z., Zhu, Z., Wu, S., Rahardja, S.: Hybrid patching for a sequence of differently exposed images with moving objects. IEEE Transactions on Image Processing **22**(12), 5190–5201 (2013)
62. Zhu, R., Xu, S., Liu, P., Li, S., Lu, Y., Niu, D., Liu, Z., Meng, Z., Li, Z., Chen, X., Fan, Y.: Zero-shot structure-preserving diffusion model for high dynamic range tone mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26130–26139 (2024)
63. Zhu, R., Xu, S., Liu, P., Liu, J., Lu, Y., Niu, D., Zheng, H., Chen, Y.K., Jing, M., Fan, Y.: A flexible zero-shot approach to tone mapping via structure-preserving diffusion models. IEEE Transactions on Circuits and Systems for Video Technology (2025)