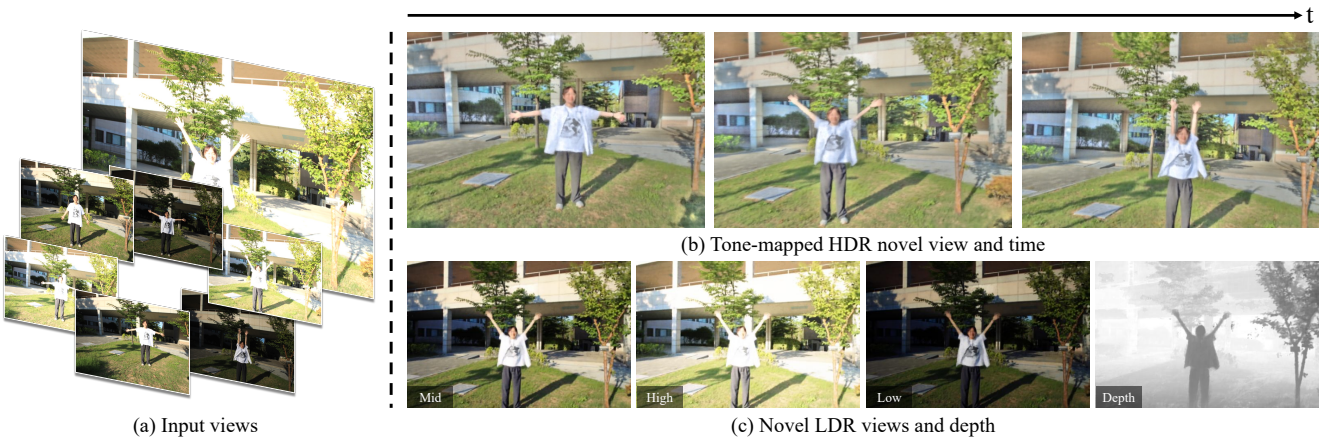
Dynamic HDR Radiance Fields via Neural Scene Flow

Shin Dong-Yeon[1] Kim Jun-Seong[1] Kwon Byung-Ki[1] Tae-Hyun Oh[2]

[1]POSTECH [2]KAIST



(a) Input views

(b) Tone-mapped HDR novel view and time

(c) Novel LDR views and depth

Figure 1. We reconstruct high dynamic range (HDR) neural scene flow fields from (a) multiple view sequences of dynamic scenes captured at different exposures (low dynamic range; LDR). Our method enables the rendering of (b) HDR novel views across both spatial and temporal domains. Additionally, we can generate (c) novel LDR views along with their corresponding depth maps.

Abstract

Reliving transient moments captured by a single camera requires reconstructing accurate radiance, geometry, and 3D motion. While significant progress has been made in dynamic 3D scene reconstruction, high-dynamic-range (HDR) radiance fields of dynamic scenes remain difficult to reconstruct. This work introduces HDR-NSFF, a novel approach to reconstructing dynamic HDR radiance fields from a monocular camera with varying exposures. HDR imaging requires multiple LDR images captured at different exposures, but capturing dynamic scenes with alternating exposures introduces challenges such as the correspondence problem, motion inconsistency, color discrepancies, and low frame rates. Here, Neural Scene Flow Fields (NSFF) is used to jointly model scene flow with neural radiance fields, enabling both novel view synthesis and temporal interpolation. NSFF is extended to HDR radiance field reconstruction by modeling learnable explicit camera response functions so that the NSFF and camera response functions can be jointly estimated in challenging dynamic scenes. Since multi-exposure images disrupt applying standard optical flow estimation due to color inconsistency, we mitigate this issue by incorporating DINOv2 semantic features, which provide exposure-invariant object-level priors for motion estimation. By integrating these components, HDR-NSFF effectively reconstructs dynamic HDR radiance fields from single-camera footage, overcoming the limitations of the previous methods and enabling novel view synthesis and high-quality time interpolation in challenging HDR scenarios.

1. Introduction

High Dynamic Range (HDR) imaging plays a pivotal role in enhancing visual realism and acquiring physical measurements for various computer vision and graphics applications, including augmented reality (AR), virtual reality (VR), and film production [1, 41, 44]. By capturing a wider range of light intensities, HDR allows for the faithful reproduction of both the bright and dark regions of a scene. Prior work [4, 12, 13, 50] has shown notable progress in 2D HDR video reconstruction, not only in static scenes but also in dynamic ones. Building on this, the focus of HDR reconstruction has shifted to 3D space. Recent works [2, 9, 48] have demonstrated robust HDR reconstruction quality using

multi-view and multi-exposure LDR images. However, this achievement is limited to static scenes, and reconstructing an HDR radiance field in dynamic 3D scenes remains an open problem, despite being a key component in the aforementioned real-world applications.

In the field of dynamic scene reconstruction, building upon the success of static scene reconstruction with Neural Radiance Fields (NeRF) [25] and 3D Gaussian Splatting (3DGS) [14], several studies [3, 7, 19, 20, 28, 29] have extended these methods to dynamic scene reconstruction. However, their approaches assume multi-view images with consistent exposure, making them struggle in multi-exposure scenarios with varying exposure times. Figure 1 illustrates this challenge. In high- and low-contrast regions caused by the sensor limitations of conventional cameras, radiance and geometry reconstruction are likely to be hindered due to color inconsistency and information loss from saturation, as they assume single-exposure inputs.

In this work, we propose HDR-NSFF, which enables the robust reconstruction of a 3D HDR dynamic radiance field (DRF), geometry, and motion from multi-exposure and multi-view images. To achieve this, we analyze the key factors that hinder 3D HDR-DRF under a multi-exposure setting and find that the motion estimator severely struggles in this scenario, providing inaccurate guidance for learning the 3D scene flow. Since the reconstruction of 3D DRF significantly rely on the 3D scene flow, degraded scene flows lead to overall degraded dynamic scene reconstruction. To robustify this, we analyze the behavior of motion estimators [39, 42] under the multi-exposure setting and observe that DINO-Tracker [42] is robust to exposure variations and captures fine-detail motions. Motivated by its robustness, we modify DINO-Tracker to seamlessly integrate with NSFF and the learnable camera tone-mapping module. As a result, we propose a complete 3D HDR-DRF reconstruction pipeline.

We evaluate our method on two tasks: novel view synthesis and novel view-and-time synthesis. Since no suitable real-world dataset exists for this purpose, we construct a multi-exposure HDR dataset capturing dynamic scenes under realistic conditions. We also evaluate on the synthetic dataset [48]. In both real and synthetic scenarios, our method consistently outperforms competing models—including NeRF-W [30] and HDR-Hexplane [48]—demonstrating our superior reconstruction quality and robustness across a wide range of challenging exposures and motions. Last but not least, our model can even reconstruct a full 3D dynamic scene in HDR from a low-FPS, multi-exposed image sequence, underscoring its robustness to limited temporal resolution and varying exposure settings. To summarize, our key contributions are:

- **HDR-NSFF:** We introduce the first method that jointly models HDR radiance field reconstruction and scene flow modeling for dynamic scenes, enabling both novel view rendering and time interpolation.
- **Exposure-Robust Optical Flow:** We introduce a novel usage of DINOv2 semantic features for dynamic HDR reconstruction by identifying its favorable property of exposure robustness. It notably improves flow estimation under multi-exposure inputs.
- **Real Evaluation Dataset:** We collect and share a real dataset of dynamic scenes captured with alternating exposures, facilitating research on dynamic HDR novel view and time interpolation.

2. Related Work

High Dynamic Range Imaging. HDR reconstruction is pioneered by Debevec and Malik [5], who propose aggregating multiple low-dynamic-range (LDR) frames to recover an HDR signal. Most subsequent HDR methods have been based upon this foundational approach. To extend HDR reconstruction to dynamic scenes, HDR video reconstruction methods [12] are introduced, typically following a two-stage process: aligning multi-exposure LDR frames and synthesizing the HDR. Optical flow or feature matching is commonly employed to establish frame correspondences, while deghosting techniques mitigate artifacts introduced during warping.

With the advent of deep learning, CNNs [12, 22, 49, 51] and Transformers [23, 37, 40, 52] have been leveraged to improve deghosting and frame alignment using large-scale datasets. However, most existing HDR video reconstruction methods rely on 2D frame alignment and struggle to synthesize novel views due to their limited representation of 3D motion. Methods constrained to 2D information approximate motion in a projected space, leading to inconsistencies when handling disocclusion and fast motion [45].

To address these challenges, we explore a 4D HDR reconstruction approach that explicitly models motion in 3D space. By leveraging multi-view 3D information, our method enhances geometric consistency and improves temporal coherence in HDR video reconstruction.

Dynamic Neural Radiance Fields. Existing dynamic view synthesis methods primarily reconstruct scenes from RGB-D images [27, 58] or monocular videos using explicit depth [15, 18, 56] representations. However, explicit representations inherently struggle to model complex structures or optical effects such as non-Lambertian reflections.

Recently, volumetric rendering methods, including NeRF [25] and 3DGS [14], have significantly advanced dynamic novel view synthesis [3, 8, 20, 28, 29, 45]. Shifting focus from static scene reconstruction, numerous methods leveraging multi-view videos have successfully reconstructed dynamic scenes [16, 17, 38, 43], enabling a wide range of practical applications. Consequently, dynamic reconstruction methods using monocular video inputs have attracted significant research interest, broadly categorized into

two approaches. The first approach learns deformations from a fixed canonical template [28, 29, 35, 54], yet is susceptible to occlusions and exhibits reduced spatial consistency when handling long sequences or large, rapid motions [8]. The second approach directly models temporal changes at individual spatial locations [8, 19, 20, 45]. Representative methods typically incorporate additional priors such as depth or optical flow to enhance temporal modeling. Although this approach involves more complex training procedures, it has demonstrated robust performance for capturing fast and complex 3D scene motions from in-the-wild videos.

In our monocular bracketing scenario, the need to acquire videos across various exposures inherently leads to reduced temporal resolution. Thus, we extend Neural Scene Flow Fields (NSFF) [19] to effectively represent and reconstruct dynamic scenes with significant motion under HDR conditions. Specifically, we propose additional HDR-robust geometric prior and a corresponding reconstruction framework. Note that our framework can easily integrate and extend beyond NSFF, leveraging various data-driven priors used by existing dynamic scene reconstruction methods.

HDR Volumetric reconstruction. Recent research actively extends conventional LDR-based volumetric reconstruction methods to the HDR domain. Several studies employed raw sensor data [10, 21, 26, 36], yet required specialized sensors and extensive calibration, limiting their practicality. In another direction, some explicitly disentangle varying scene appearance via learned embeddings [24, 32, 57], but this approach is limited to isolating appearance changes rather than reconstructing HDR radiance. Thus, approaches using widely available multi-exposure images have become prevalent, often jointly modeling HDR 3D representations with tone-mapping for LDR conversion [2, 9, 11, 33, 46].

However, existing HDR volumetric methods typically focus on static scenes, struggling with motion, common in real-world. HDR-Hexplane [48], a recent dynamic HDR method based on Hexplane representations, has two critical limitations. First, it demands densely sampled, high-FPS data that are difficult to obtain practically. Second, it employs a fixed sigmoid function instead of learnable camera-specific camera response functions (CRFs), insufficiently capturing diverse image signal processor (ISP) characteristics. Moreover, the strong reliance on inter-plane interpolation leads to artificial motions in scenes with rapid dynamics.

To overcome these issues, we propose: (1) a scene-flow-based temporal interpolation method that achieves natural temporal continuity from sparse, low-FPS multi-exposure data; (2) a robust optical flow estimator insensitive to exposure variations; and (3) explicit modeling of camera-specific learnable CRFs to enhance accuracy and adaptability.

3. Method

This section provides details of our method. We first provide a brief overview of the baseline NSFF method (Sec. 3.1) before introducing the overall pipeline (Sec. 3.2). Next, we describe our tone-mapping module (Sec. 3.3) and novel semantic-based optical flow estimation (Sec. 3.4). Finally, we detail the optimization procedure (Sec. 3.5).

3.1. Preliminary

To model dynamic scenes, NSFF [19] extend the concept of NeRF [25] by representing 3D motion as scene flow fields. NSFF learns a combination of static and dynamic NeRF representations. The dynamic model, denoted as F_θ^{dy}, explicitly models view and time dependent variations by incorporating time t as an additional input. Beyond color and density, it also predicts forward and backward 3D scene flow $F_t{=}(\mathbf{f}_{t\to t+1}, \mathbf{f}_{t\to t-1})$ and occlusion weights $W_t{=}(w_{t\to t+1}, w_{t\to t-1})$ to handle 3D motion disocclusion:

$$(c_t, \sigma_t, F_t, W_t) = F_\theta^{\text{dy}}(\mathbf{x}, \mathbf{d}, t). \tag{1}$$

To supervise scene flow estimation, NSFF uses temporal photometric consistency. Specifically, for each time i, scene flow is predicted for the 3D points sampled along rays, and this predicted flow is used to warp corresponding points from neighboring times $j \in \mathcal{N}(i)$ to time i. The color and opacity information of the warped points is then used to render the image at time i:

$$\hat{C}_{j\to i}(r_i) = \int_{z_n}^{z_f} T_j(z)\, \sigma_j\big(r_{i\to j}(z)\big)\, c_j\big(r_{i\to j}(z), d_i\big)\, dz, \tag{2}$$

$$\text{where} \quad r_{i\to j}(z) = r_i(z) + \mathbf{f}_{i\to j}\big(r_i(z)\big). \tag{3}$$

Temporal photometric consistency is enforced by minimizing the mean squared error (MSE) between the warped rendered view and the ground-truth image:

$$L_{photo} = \sum_{r_i} \sum_{j\in\mathcal{N}(i)} \|\hat{C}_{j\to i}(r_i) - C_i(r_i)\|_2^2. \tag{4}$$

The static NeRF, F_θ^{st}, represents a time-invariant scene using a multilayer perceptron (MLP). Given an input position $\mathbf{x}$ and view direction $\mathbf{d}$, it outputs the RGB color c, volume density σ, and an unsupervised 3D mixing weight v that determines the blending between static and dynamic components:

$$(c, \sigma, v) = F_\theta^{\text{st}}(\mathbf{x}, \mathbf{d}). \tag{5}$$

Here, c_t and σ_t denote the color and volume density at position $\mathbf{x}$ at time t. The final color is computed by blending the static and dynamic components using the following rendering equation:

$$\hat{C}_i^{cb}(r_i) = \int_{z_n}^{z_f} T_i^{cb}(z)\, \sigma_i^{cb}(z)\, e_i^{cb}(z)\, dz, \tag{6}$$

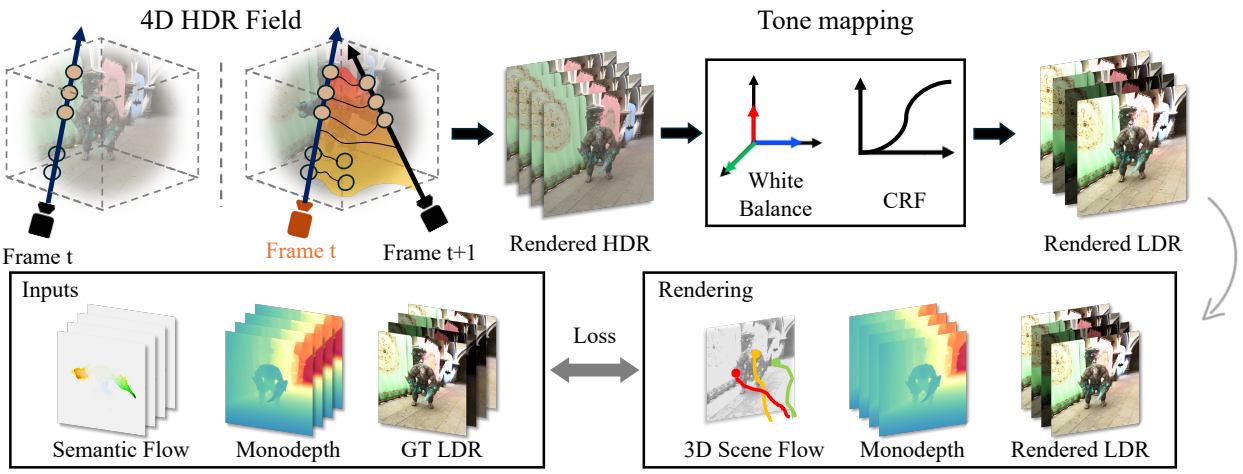


Figure 2. **Overall pipeline of our propsed method.** HDR-NSFF takes a multi-exposed image sequence of a dynamic scene as input and estimate 3D scene flow for the sampled points along each ray. Neighboring frames are then warped to render the HDR radiance at the target frame, which is tone-mapped to LDR via a white-balance and camera-response function module. Photometric loss with the ground-truth LDR images, along with optical flow and depth constraints from off-the-shelf models, jointly optimize both the scene flow fields and tone-mapping module in an end-to-end manner.

where $\sigma_i^{cb}(z)c_i^{cb}(z)$ is a linear combination of static and dynamic scene components, weighted by $v(z)$:

$$\sigma_i^{cb}(z)c_i^{cb}(z) = v(z)c(z)\sigma(z) + (1 - v(z))c_i(z)\sigma_i(z). \quad (7)$$

T_i represents the transmittance at time i, while z_n and z_f denote the near and far depths along the ray. The final rendered output $\hat{C}_i^{cb}(r_i)$ is optimized against the ground-truth pixel color $C_i(r_i)$ using a photometric loss:

$$L_{cb} = \sum_{r_i} \|\hat{C}_i^{cb}(r_i) - C_i(r_i)\|_2^2. \quad (8)$$

Reconstructing dynamic scenes from monocular input is inherently ill-posed, and relying solely on photometric consistency often leads to convergence at poor local minima. Therefore, NSFF incorporates three additional guided losses: a term enforcing monocular depth and optical flow consistency, a motion trajectory term promoting cycle-consistency and spatiotemporal smoothness, and a compactness prior encouraging binary scene decomposition and reducing floaters via entropy and distortion losses. For more details, please refer to the supplementary materials.

3.2. HDR-NSFF

We now extend NSFF to handle HDR reconstruction, introducing the HDR-NSFF pipeline. Figure 2 shows overall pipeline. Given that the radiance of a scene remains invariant to changes in camera exposure settings, our method explicitly decouples the intrinsic scene radiance and geometry from camera-dependent imaging characteristics. Specifically, HDR-NSFF first leverages the NSFF framework to model the intrinsic spatio-temporal radiance (E) and geometric structure represented by density (σ), independently from camera parameters. Subsequently, a dedicated tone-mapping module simulates the camera's physical imaging processes—including white balance and the camera response function —to convert these intrinsic radiance values into observed LDR images. Such modular decomposition allows HDR-NSFF to effectively learn exposure-invariant scene representations. In the following, we explain a detailed design of the tone-mapping module.

3.3. Tone Mapping

The tone mapping process converts HDR images into LDR images by sequentially applying exposure adjustment, white balance correction, and CRF. In HDR-NSFF, volume rendering produces an HDR radiance value E, which is processed by an explicit tone mapping module $\mathcal{T}$ with radiometric parameters θ :

$$C = \mathcal{T}(E, \theta). \quad (9)$$

The tone mapping function $\mathcal{T}$ consists of two stages: a white balance function w and a CRF g. The white balance scaling parameters are learned jointly with exposure values. The tone mapping process follows a typical digital camera's acquisition pipeline and is expressed as:

$$C = \mathcal{T}(E) = g \circ w(E). \quad (10)$$

The white balance function w applies per-channel scaling using the white balance parameter $\theta_w = [w_r, w_g, w_b]^\top \in \mathbb{R}^3$, producing a white balance-corrected image E_w. The CRF g is then applied to E_w, mapping it to the final LDR image C. The CRF is parameterized as a piecewise linear function, defined using 256 points uniformly sampled in the $[0, 1]$ range. Values exceeding the dynamic range are thresholded accordingly.

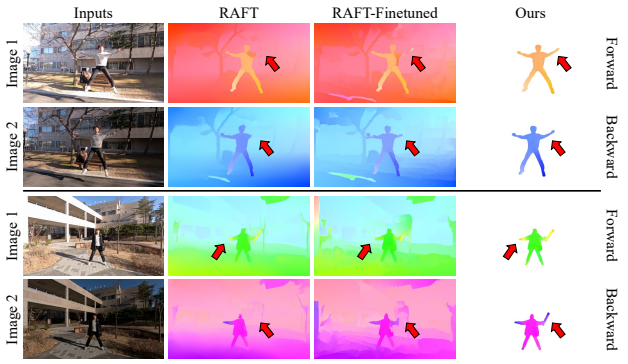


Figure 3. **Visualization of flow between multi exposed images.** RAFT struggles under multi-exposure conditions, resulting in noticeable errors. As a stronger reference, fine-tuning RAFT on synthetic multi-exposed data (called RAFT-Finetuned) provides moderate improvement, but our semantic-based approach remains more accurate overall. The red arrows highlight regions where RAFT and RAFT-Finetuned fail to capture correct motion, whereas our method recovers a more reliable flow.

During training, we adopt leaky-thresholding, which is proposed by HDR-Plenoxels [11], to reduce saturation loss in rendered images:

$$g_{\text{leaky}}(x) = \begin{cases} \alpha x, & x < 0 \\ g(x), & 0 \leq x \leq 1 \\ -\frac{\alpha}{\sqrt{x}} + \alpha + 1, & x > 1, \end{cases} \tag{11}$$

where α is the thresholding coefficient. This approach ensures effective color correction and dynamic range handling during HDR-NSFF training.

3.4. Semantic based Optical Flow

DINO-Tracker [42] leverages a transformer-based architecture to aggregate global semantic features from DINOv2, which allows it to maintain robust correspondences even under dramatic exposure variations. By computing matching scores across frames, it can reliably track objects over long sequences despite occlusions or changes in appearance. This global context enables the tracker to overcome challenges that conventional optical flow methods face in multi-exposure scenarios. However, because DINOv2 is primarily trained on object-centric data, its performance on background regions can be noisy. To address this limitation, we further refine the tracking by restricting it to adjacent frame pairs and using Segment Anything Model 2 (SAM2) [31] at inference time to designate query points corresponding to moving objects. This targeted approach significantly improves the quality of the estimated optical flow, as demonstrated in Figure 3, and significantly contributes to more accurate scene flow estimation in our HDR-NSFF framework.

3.5. Losses

Photometric Loss. To end-to-end train both the neural scene flow fields and the tone-mapping module using only LDR images, we minimize the Mean Squared Error between the LDR views rendered by our HDR-NSFF and the ground-truth (GT) LDR images.

Building on the photometric supervision in NSFF [19], we instead replace the rendered color C with our tone-mapped output $\mathcal{T}(E)$, where E denotes the rendered HDR radiance. Formally, the photometric losses are defined as:

$$L_{cb} = \sum_{r_i} \|\mathcal{T}(\hat{E}_i^{cb}(r_i)) - C_i(r_i)\|_2^2, \quad \text{and} \tag{12}$$

$$L_{photo} = \sum_{r_i} \sum_{j \in \mathcal{N}(i)} \|\mathcal{T}(\hat{E}_{j \to i}(r_i)) - C_i(r_i)\|_2^2, \tag{13}$$

where r denotes a camera ray. This loss directly supervises our network by ensuring that the tone-mapped renderings closely match the GT LDR images.

Data-Driven Priors. Monocular reconstruction of complex dynamic scenes is highly ill-posed, and multi-exposure conditions further exacerbate this challenge. Similar to prior works that integrate external cues for improved geometry and motion estimation [19, 20], we leverage data-driven priors by integrating our semantic-based optical flow and off-the-shelf single-view depth. Specifically, our optical flow loss, L_{flow}, initializes scene flow estimation by minimizing the MSE between the 2D displacement of projected 3D points and the semantic-based optical flow. In parallel, the single-view depth loss, L_{depth}, enforces temporal geometric consistency by comparing the warped, opacity-derived depth against the depth predicted by Depth-Anything-V2 [53]. The overall data-driven prior loss is formulated as:

$$L_{\text{data}} = L_{\text{flow}} + \beta_{depth} L_{\text{depth}}, \tag{14}$$

which effectively enforces motion and depth consistency. For further details, please refer to the supplementary material.

CRF Smoothness Loss. We incorporate a smoothness loss to enforce that CRF varies smoothly in a physically plausible manner [6]. Following HDR-Plenoxels [11], we penalize the second-order derivative of the CRFs: It is defined as follows:

$$\mathcal{L}_{smooth} = \sum_{i=1}^{N} \sum_{e \in [0,1]} g_i''(e), \tag{15}$$

where $g''(e)$ denotes the second order derivative of CRFs *w.r.t.* its input domain. Finally, our HDR-NSFF is end-to-end optmized using the following loss:

$$L = L_{cb} + L_{photo} + \beta_{data} L_{data} + \beta_{reg} L_{reg} + L_{smooth}, \tag{16}$$

where the β coefficients weight each term. Additional regularization terms, L_{reg} leveraging scene flow priors are detailed in the supplementary material.

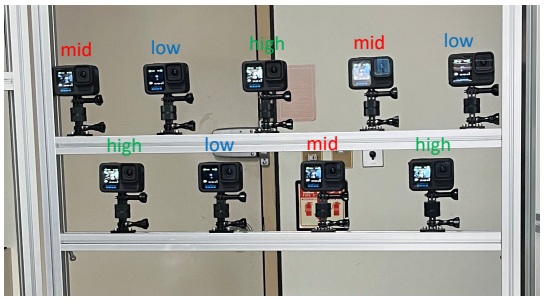


Figure 4. **A camera rig for multi-exposure multi-view dataset**. We use nine GoPro Hero 13 Black cameras arranged at two height levels with fixed spatial intervals. Each camera is synchronized to capture multi-view video simultaneously under three distinct exposure settings.

4. Experiments

4.1. Experimental Settings

We evaluate our method on both synthetic and real datasets. The synthetic dataset is sourced from HDR-Hexplane [48], while the real dataset is collected using our multi-exposure camera system.

Synthetic Dataset. We modify the dataset proposed in HDR-HexPlane [48]. The original dataset is rendered with a high frame rate and a multi-camera configuration. To better reflect real-world exposure bracketing scenarios, we select four scenes and modified. We re-render scenes in a monocular setup and uniformly sample 30 images to simulate sparse acquisition conditions. Further implementation details are provided in the supplementary materials.

Real Dataset. For further evaluation, we additionally captured real-world datasets, *Real dataset*. Go-pro dataset consists of a custom camera rig equipped with 9 GoPro Hero 13 Black. Inspired by [17] we arrange cameras at two different height levels with a fixed spatial interval. All cameras are synchronized to capture a dense multi-view video sequence simultaneously and are preconfigured with three distinct multi-exposure settings to ensure a diverse range of exposure levels in the recorded data. Figure 4 shows the camera setup. To construct a video sequence, we select one frame per viewpoint from a single camera at each timestamp simulating monocular dynamic input setup.

Implementation Details. We estimate the intrinsic and extrinsic camera parameters using COLMAP [34]. Since COLMAP assumes a static scene, we utilize the SAM2 [31] to mask out features corresponding to dynamic objects.

During training and testing, we sample 128 points along each ray and normalize the video sequence to a temporal range of $i \in [0, 1]$. Training a full model takes about 15 hours per scene using a single NVIDIA RTX 3090 GPU. Rendering at a resolution of 720×480 takes around 5 sec.

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NSFF [19]	17.79	0.7048	0.0705	15.59	0.5577	0.1339
NeRF-WT [30]	28.29	0.9139	0.0322	16.62	0.5739	0.1268
HDR-HexPlane [48]	20.75	0.6248	0.1387	17.65	0.5659	0.1316
Ours	31.29	0.9277	0.0235	22.81	0.7673	0.0718

Table 1. **Quantitative results of novel view synthesis on real data.** Metrics are averaged across all scenes. The green and yellow colors stand for the best and the second best, respectively.

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NSFF [19]	17.77	0.7028	0.0786	15.74	0.5889	0.1682
NeRF-WT [30]	27.99	0.9015	0.0392	16.43	0.5513	0.1494
HDR-HexPlane [48]	21.56	0.6502	0.1407	18.14	0.6166	0.1346
Ours	31.48	0.9304	0.0262	22.82	0.8087	0.0949

Table 2. **Quantitative results of novel view and time synthesis on real data.** Metrics are averaged across all scenes. Each color stands for the best and the second best, respectively.

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NSFF [19]	15.98	0.6457	0.1388	16.04	0.5697	0.1527
NeRF-WT [30]	31.10	0.9366	0.0342	21.50	0.7490	0.0895
HDR-HexPlane [48]	29.95	0.9055	0.0527	23.87	0.7999	0.1071
Ours	35.07	0.9465	0.0483	27.19	0.8836	0.0576

Table 3. **Quantitative results of novel view and time synthesis on synthetic data.** Metrics are averaged across all scenes. Each color stands for the best and the second best, respectively.

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours w/ RAFT [39]	30.42	0.9269	0.0246	21.38	0.7369	0.0675
Ours w/ Finetuned	30.68	0.9234	0.0253	21.51	0.7377	0.0689
Ours	31.01	0.9301	0.0233	22.55	0.7714	0.0697

Table 4. **Ablation study on real data.** To compare the effect of flow regularization, we compare our approach against the baseline optical flow model (RAFT [39]) and a stronger baseline finetuned RAFT on a multi-exposure adaptation of the FlyingThings3D dataset. The green and yellow colors stand for the best and the second best, respectively.

Evaluation. We compare our method against NSFF [19], HDR-HexPlane [48], and NeRF-WT [30]. Since the exposure values of individual images cannot be directly estimated, we assume that the scene is captured using three identical cameras. To convert HDR images into LDR images, we borrow the tone-mapping function learned from neighboring cameras. To evaluate the quality of synthesized images from novel views, we use PSNR, SSIM [47], and LPIPS [55] as metrics. To evaluate HDR images, we employ μ-law, which applies a logarithmic transformation to compress HDR pixel values, converting them into a tone-mapped image:

$$M(E) = \frac{\log(1+\mu E)}{\log(1+\mu)}, \tag{17}$$

where μ is the compression level set to 50 in this work.

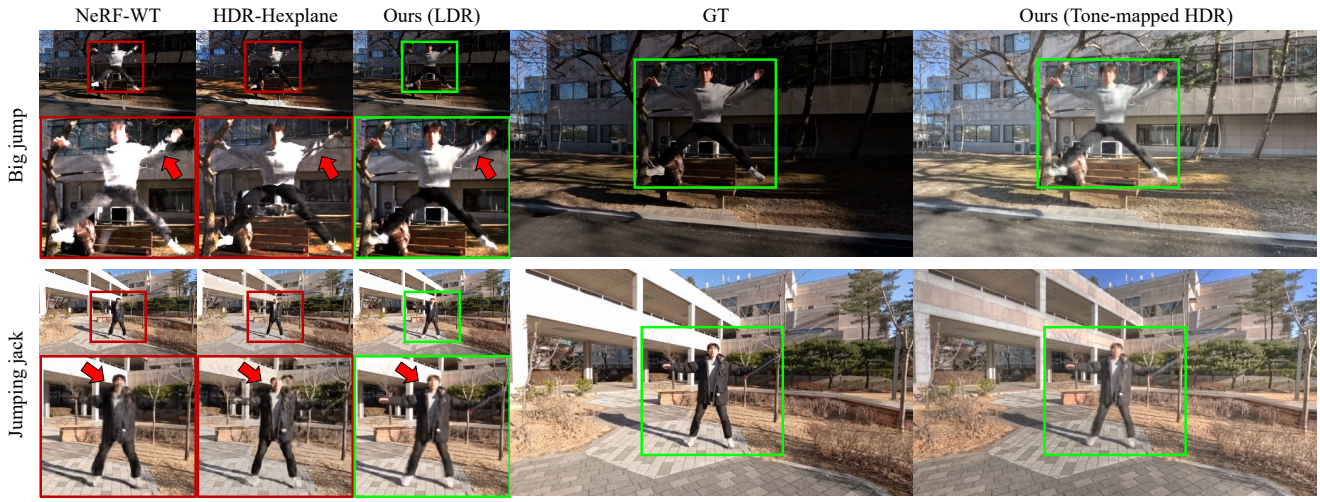


Figure 5. **Qualitative results of novel view synthesis on real multi-exposed dynamic scene dataset.** Our method maintains more consistent geometry and color across varying viewpoints, while other approaches (NeRF-WT [30], HDR-Hexplane [48]) exhibit noticeable artifacts and geometric distortions. The green and red boxes highlight regions where our method yields sharper details and fewer artifacts.

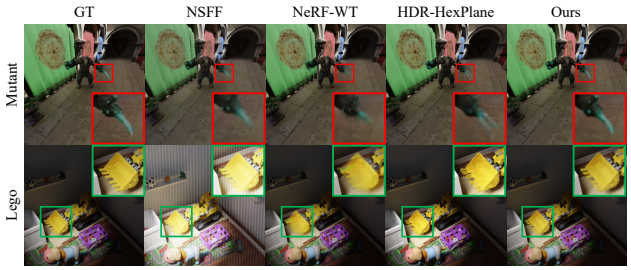


Figure 6. **Qualitative results of novel view and time synthesis on synthetic data.** Since, our approach explicitly models scene flow, it excels at time interpolation and consistently produces high-quality results. In contrast, other methods [19, 30, 48] struggle to handle the temporal dimension, leading to visible artifacts.

4.2. Quantitative Results

Novel View Synthesis. To quantitatively evaluate the performance of our method on dynamic scene reconstruction, we adopt a novel view synthesis (NVS) evaluation strategy. Specifically, for each time instance, we render the scene from all camera poses not used during training and apply the corresponding learned tone-mapping functions to convert the HDR renders back to LDR. We then compare these synthesized views against the actual LDR images captured by those cameras. By measuring how closely the rendered results match the ground-truth images, this evaluation directly assesses two key aspects: (1) the quality of our dynamic scene modeling, and (2) the accuracy of our camera-specific tone-mapping functions. Table 1 shows that our approach achieves significant improvements in rendering fidelity compared to previous single-view and multi-view baselines, both in highly dynamic regions and across the entire scene. This demonstrates its effectiveness in reconstructing HDR scenes with fine detail across varying exposures.

Novel View and Time Synthesis. We also evaluate novel view and time synthesis to demonstrate our method's ability to handle dynamic scenes with sparse temporal sampling. Following NSFF [19], we remove every other frame from the original video sequences during training, and use the intermediate frames at held-out camera viewpoints for testing. To render a new frame in fractional time, we adopt the scene motion–based splatting approach: we predict the volumetric density and color from the two nearest training frames, warp them via our scene flow, and then blend them linearly according to the target time index. After reconstructing the scene in HDR, we apply the learned tone-mapping function to convert it to LDR for direct comparison against the ground-truth intermediate frames. Tables 2 and 3 show that our results outperform competing models across all evaluation metrics.

Ablation study. We analyze the impact of our proposed semantic-based optical flow on the novel view synthesis task using 8 real dataset samples. We compare two variants of our method: (1) Ours (w/ RAFT), in which the RAFT optical flow is used without modification, and (2) Ours (w/ RAFT Finetuned), where RAFT is fine-tuned on synthetic multi-exposure data. Note that, as shown in Figure 3, the original RAFT model was not trained on multi-exposed images, resulting in high errors when applied directly in our setting. By fine-tuning it on synthetic data, the performance is improved. As shown in Table 4, our proposed method achieves the best results.

Figure 7. **Qualitative results on novel LDR/HDR view synthesis.** We visualize LDR rendering results at varying exposure levels (low, mid, and high), tone-mapped HDR rendering by ours and corresponding ground-truth HDR references. We also visualize histograms of our HDR images (the upper) and ground truth (the lower). For better visualization, we plot HDR histogram using smoothed kde method.

4.3. Qualitative Results

LDR reconstruction. We qualitatively assess our method's ability to render novel views on the real-world multi-exposure dataset. This indicates that our HDR-NSFF consistently preserves fine geometric details and color accuracy compared to other baselines. Because our model simultaneously learns geometry and a camera-specific tone-mapping module, it not only reconstructs dynamic scenes more faithfully but also generalizes to new exposure levels. that our method captures both the scene structure and the camera's response function more accurately, leading to sharper images and fewer artifacts even from unseen viewpoints.

We further demonstrate the capability of our approach to handle time synthesis under multi-exposure conditions. In contrast to methods that do not explicitly model 3D motion, our HDR-NSFF leverages scene flow to warp content from adjacent frames, enabling smooth transitions at fractional time steps. As shown in Figure 6, our model renders dynamic objects with noticeably fewer motion artifacts, even when the temporal gap is large due to their low frame-rate. This is particularly beneficial in multi-exposure scenarios, where high-FPS data acquisition is challenging. By accurately modeling motion and exposure simultaneously, our approach generates more coherent intermediate frames, outperforming baselines that struggle with large inter-frame displacements.

HDR reconstruction. To validate the effectiveness of our approach in HDR reconstruction, we conducted qualitative comparisons between HDR rendering of our model and ground-truth HDR images, as shown in Fig. 7. Tone-mapped HDR views generated by our model (Tone-mapped HDR) exhibit strong visual consistency with ground-truth HDR images, particularly in preserving fine details in challenging lighting scenarios, including both underexposed and overexposed regions.

We also visualize histograms of pixel intensity of tone-mapped HDR images, demonstrating that our reconstructed HDR images effectively cover the entire radiance range, accurately recovering radiance values from very low to high intensities, closely matching ground-truth distribution. Also, we present novel LDR views rendered at multiple exposure levels, demonstrating that our method successfully controls exposure using specified exposure times, by accurately reconstructing under and over saturation of the images.

5. Conclusion

In this work, we presented HDR-NSFF, a framework for reconstructing dynamic HDR scenes from multi-exposure video captured using a single camera. Our method tackles the inherent challenges of dynamic HDR reconstruction by simultaneously modeling intrinsic scene properties and camera-specific imaging variations. Without relying on uniform exposure assumptions or pre-calibrated camera responses, our approach jointly estimates HDR radiance, geometry, and 3D motion while learning tone-mapping parameters directly from multi-view LDR images. We also introduce a robust optical flow estimation strategy that leverages semantic cues to handle exposure variations, thereby refining scene flow predictions even under challenging conditions. Extensive evaluations on both real-world and synthetic datasets demonstrate that HDR-NSFF significantly improves rendering fidelity, geometric consistency, and temporal interpolation compared to other methods.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00397663, Real-time XR Interface Technology Development for Environmental Adaptation, 25%), Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [25ZD1160, Development of ICT Convergence Technology for Daegu-Gyeongbuk Regional Industry], Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2025-25443318, Physically-grounded Intelligence: A Dual Competency Approach to Embodied AGI through Constructing and Reasoning in the Real World), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH))

References

[1] Francesco Banterle, Alessandro Artusi, Kurt Debattista, and Alan Chalmers. *Advanced high dynamic range imaging*. AK Peters/CRC Press, 2017. 1

[2] Yuanhao Cai, Zihao Xiao, Yixun Liang, Minghan Qin, Yulun Zhang, Xiaokang Yang, Yaoyao Liu, and Alan Yuille. Hdr-gs: Efficient high dynamic range novel view synthesis at 1000x speed via gaussian splatting. In *arXiv preprint arXiv:2405.15125*, 2024. 1, 3

[3] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *CVPR*, 2023. 2

[4] Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K Wong, and Lei Zhang. Hdr video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2502–2511, 2021. 1

[5] Paul E Debevec and Jitendra Malik. Recovering High Dynamic Range Radiance Maps from Photographs. In *ACM TOG*, page 10, 1997. 2

[6] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 465–474. 2023. 5

[7] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2

[8] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *NeurIPS*, 2022. 2, 3

[9] Xin Huang, Qi Zhang, Feng Ying, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *CVPR*, 2021. 1, 3

[10] Xin Jin, Pengyi Jiao, Zheng-Peng Duan, Xingchao Yang, Chong-Yi Li, Chun-Le Guo, and Bo Ren. Lighting every darkness with 3dgs: Fast training and real-time rendering for hdr view synthesis. In *NeurIPS*, 2024. 3

[11] Kim Jun-Seong, Kim Yu-Ji, Moon Ye-Bin, and Tae-Hyun Oh. Hdr-plenoxels: Self-calibrating high dynamic range radiance fields. In *ECCV*, 2022. 3, 5

[12] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. In *ACM TOG*, 2017. 1, 2

[13] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions On Graphics (TOG)*, 22(3):319–325, 2003. 1

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM TOG*, 2023. 2

[15] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1611–1621, 2021. 2

[16] Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Ping Tan. Streaming radiance fields for 3d video synthesis. In *NeurIPS*, 2022. 2

[17] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. In *CVPR*, 2022. 2, 6

[18] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 2

[19] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 2, 3, 5, 6, 7

[20] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *CVPR*, 2023. 2, 3, 5

[21] Zhihao Li, Yufei Wang, Alex Kot, and Bihan Wen. From chaos to clarity: 3dgs in the dark. In *NeurIPS*, 2024. 3

[22] Zhen Liu, Wenjie Lin, Xinpeng Li, Qing Rao, Ting Jiang, Mingyan Han, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In *CVPR*, 2021. 2

[23] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *ECCV*, 2022. 2

[24] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 3

[25] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European conference on computer vision*, 2020. 2, 3

[26] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. NeRF in the dark: High dynamic range view synthesis from noisy raw images. 2022. 3

[27] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[28] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. In *ACM TOG*, 2021. 2, 3

[29] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *CVPR*, 2020. 2, 3

[30] Chen Quei-An. Nerf pl: a pytorch-lightning implementation of nerf. 2020. 2, 6, 7

[31] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5, 6

[32] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022. 3

[33] Darius Rückert, Linus Franke, and Marc Stamminger. ADOP: Approximate Differentiable One-Pixel Point Rendering. In *arXiv:2110.06635 [cs]*, 2021. 3

[34] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 6

[35] Richard Shaw, Michal Nazarczuk, Jifei Song, Arthur Moreau, Sibi Catley-Chandar, Helisa Dhamo, and Eduardo Pérez-Pellitero. Swings: Sliding windows for dynamic 3d gaussian splatting. In *ECCV*, 2024. 3

[36] Shreyas Singh, Aryan Garg, and Kaushik Mitra. Hdrsplat: Gaussian splatting for high dynamic range 3d scene reconstruction from raw images. *BMVC*, 2024. 3

[37] Jou Won Song, Ye-In Park, Kyeongbo Kong, Jaeho Kwak1, and Suk-Ju Kang. Selective transhdr:transformer-based selective hdr imaging using ghost region mask. In *ECCV*, 2022. 2

[38] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *ACM TOG*, 2023. 2

[39] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. *ECCV*, 2020. 2, 6

[40] Steven Tel, Zongwei Wu, Yulun Zhang, Barthélémy Heyrman, Cédric Demonceaux, Radu Timofte, and Dominique Ginhac. Alignment-free hdr deghosting with semantics consistent transformer. In *ICCV*, 2023. 2

[41] Gaurav Tiwari and Pushpi Rani. A review on high-dynamic-range imaging with its technique. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(9):93–100, 2015. 1

[42] Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. In *ECCV*, 2024. 2, 5

[43] Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, and Huaping Liu. Mixed neural voxels for fast multi-view video synthesis. In *ICCV*, 2023. 2

[44] Lin Wang and Kuk-Jin Yoon. Coaug-mr: An mr-based interactive office workstation design system via augmented multi-person collaboration. *arXiv preprint arXiv:1907.03107*, 2019. 1

[45] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. In *arXiv preprint arXiv:2407.13764*, 2024. 2, 3

[46] Yuehao Wang, Chaoyi Wang, Bingchen Gong, and Tianfan Xue. Bilateral guided radiance field processing. *ACM TOG*, 2024. 3

[47] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2003. 6

[48] Guanjun Wu, Taoran Yi, Jiemin Fang, Wenyu Liu, and Xinggang Wang. Fast high dynamic range radiance fields for dynamic scenes. In *2024 International Conference on 3D Vision (3DV)*, 2024. 1, 2, 3, 6, 7

[49] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *ECCV*, 2018. 2

[50] Gangwei Xu, Yujin Wang, Jinwei Gu, Tianfan Xue, and Xin Yang. Hdrflow: Real-time hdr video reconstruction with large motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24851–24860, 2024. 1

[51] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *CVPR*, 2019. 2

[52] Qingsen Yan, Weiye Chen, Song Zhang, Yu Zhu, Jinqiu Sun, and Yanning Zhang. A unified hdr imaging method with pixel and patch level. In *CVPR*, 2023. 2

[53] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911, 2024. 5

[54] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *CVPR*, 2024. 3

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[56] Zhoutong Zhang, Forrester Cole, Richard Tucker, William T. Freeman, and Tali Dekel. Consistent depth of moving objects in video. *ACM TOG*, 2021. 2

[57] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In *CVPR*, 2024. 3

[58] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, and Marc Stamminger. Real-time non-rigid reconstruction using an rgb-d camera. *ACM TOG*, 2014. 2