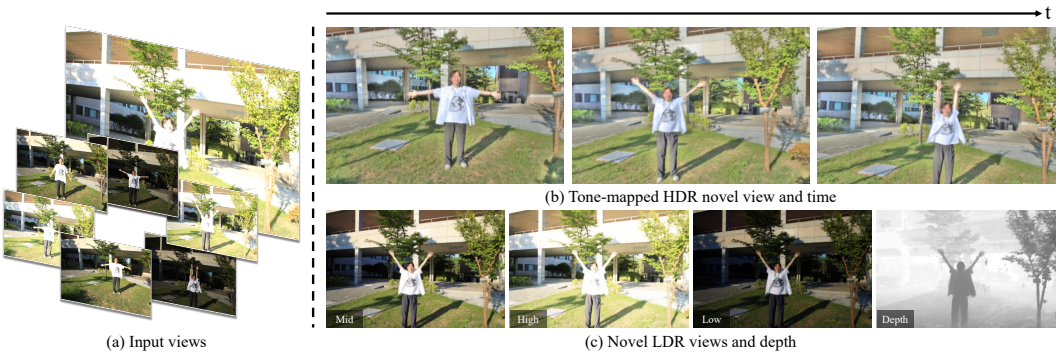
HDR-NSFF: HIGH DYNAMIC RANGE NEURAL SCENE FLOW FIELDS

Anonymous authors
Paper under double-blind review



(a) Input views (b) Tone-mapped HDR novel view and time (c) Novel LDR views and depth

Figure 1: **High Dynamic Range Neural Scene Flow Fields (HDR-NSFF)** reconstruct dynamic HDR radiance field from (a) alternatively exposed videos of dynamic scenes. 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

Radiance of real-world scenes typically spans a much wider dynamic range than what standard cameras can capture, often leading to saturated highlights or underexposed shadows. While conventional HDR methods merge alternatively exposed frames, most approaches remain constrained to the 2D image plane, failing to model geometry and motion consistently. To address these limitations, we present HDR-NSFF, a novel framework for reconstructing dynamic HDR radiance fields from alternatively exposed monocular videos. Our method explicitly models 3D scene flow, HDR radiance, and tone mapping in a unified end-to-end pipeline. We further enhance robustness by (i) extending semantic-based optical flow with DINO features to achieve exposure-invariant motion estimation, and (ii) incorporating a generative prior as a regularizer to compensate for sparse-view and saturation-induced information loss. To enable systematic evaluation, we construct a real-world GoPro dataset with synchronized multi-exposure captures. Experiments demonstrate that HDR-NSFF achieves state-of-the-art performance in novel view and time synthesis, recovering fine radiance details and coherent dynamics even under challenging exposure variations and large motions.

1 INTRODUCTION

Radiance of real-world scenes typically spans a wider dynamic range than what standard cameras can capture (see Fig. 1). As a result, captures with standard cameras often suffer from overexposed highlights or underexposed shadows, leading to severe information loss in critical regions. A widely adopted strategy to address this limitation is high dynamic range (HDR) imaging, which captures multiple low dynamic range (LDR) frames at different exposures and merges them to form a HDR image. DR has become essential for enhancing realism and preserving radiometric information

However, most existing HDR methods remain fundamentally constrained to the 2D image plane. Video-based HDR approaches (Kalantari et al., 2017; Chen et al., 2021; Chung and Cho, 2023; Xu

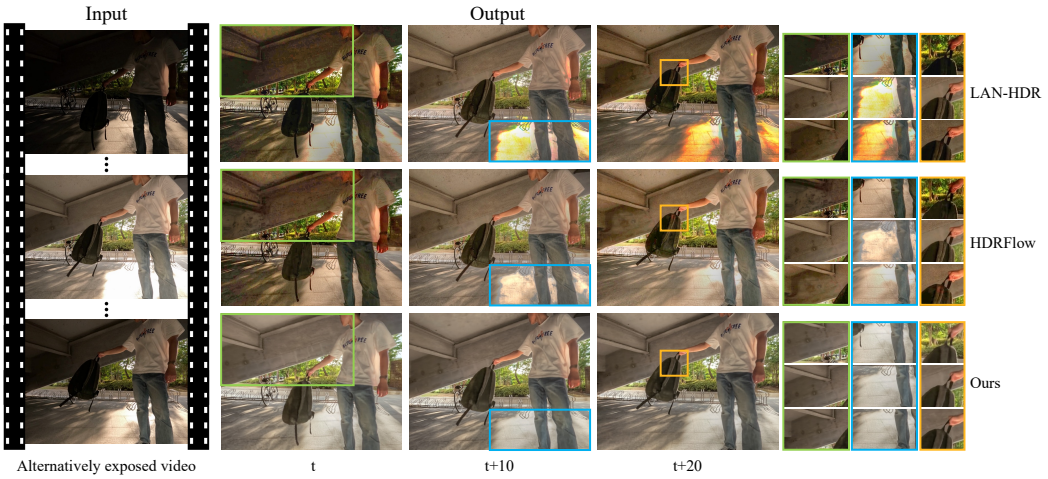


Figure 2: **Comparison of HDR video reconstruction on training views.** Given alternatively exposed video, HDR video reconstruction baselines, *i.e.*, LAN-HDR (Chung and Cho, 2023) and HDRFlow (Xu et al., 2024), fail to produce consistent results, while our model ensures temporal coherence and recovers valid information in saturated regions.

et al., 2024) typically align consecutive frames and apply refinement to suppress ghosting and motion artifacts. Operating purely in image space, these methods cannot capture 3D motion, thus failing to handle occlusions, dynamics, varying radiance, and viewpoint changes (see Fig.2). These limitations clearly indicate the need to move beyond conventional 2D fusion toward a 3D representation.

In this work, we propose a framework for reconstructing HDR dynamic radiance fields from alternatively exposed monocular videos. We represent the scene as a continuous function of both space and time, whose outputs include HDR radiance, density, and 3D motion. Due to the additional variation in exposure, the HDR dynamic reconstruction attains a higher order of ill-posedness than reconstruction only across space and time. To address, we propose a dedicated pipeline built upon a dynamic neural radiance field (Li et al., 2021), as it leverages geometric and motion priors that can counteract the ill-posedness of the problem. Then, we jointly optimize the radiance field together with a learnable tone-mapping function, enabling HDR reconstruction and tone mapping in an end-to-end manner.

In designing the pipeline, a critical challenge of alternatively exposed video lies in the severe color inconsistency across frames, which induces combinatorial degradations spanning tone mapping, geometry, and motion priors. Thus, we analyze the robustness of each component to exposure variation and investigate the optimal combination among them. In particular, for motion prior, we observe that the semantic features of DINOv2 (Oquab et al., 2023) demonstrate strong robustness to illumination changes. Inspired by this observation, we extend DINO-Tracker (Tumanyan et al., 2024) to predict dense optical flow that remains reliable under varying exposures, and integrate these predictions into scene flow learning for dynamic HDR reconstruction.

Another major challenge lies in the correlation between the sparse-view nature of monocular videos and the information loss induced by saturation under extreme exposures. In other words, the state of a moving object can only be observed at specific timesteps, and if those observations are saturated, the result is an irrecoverable loss of information. To mitigate this issue, we incorporate generative priors (Wu et al., 2025) to compensate for the loss by augmenting the single training view with multi-view information and distilling it into the radiance field.

In addition to the synthetic dataset (Wu et al., 2024a), we evaluate our method on a newly constructed real-world dataset, which spans a wide range of scenarios including indoor and outdoor environments, diverse objects, and human subjects. Across both domains, our method consistently outperforms existing baselines, including NeRF-W (Martin-Brualla et al., 2021), 4DGS (Wu et al., 2024b), MotionGS (Zhu et al., 2024) and HDR-Hexplane (Wu et al., 2024a), demonstrating superior reconstruction quality and robustness under challenging exposures.

To summarize, our key contributions are:

- **HDR-NSFF Framework:** We propose the first method that jointly models HDR scene flow fields enabling both novel view rendering and time interpolation.
- **Robust Learning Strategies:** We enhance scene flow learning by extending DINO-Tracker for exposure-robust motion estimation, and introduce generative priors as regularizers to overcome sparse-view limitations.
- **Comprehensive Evaluation:** We provide extensive experiments and a new real-world dataset with alternative exposures, demonstrating state-of-the-art performance in challenging HDR scenarios.

2 Related Work

High Dynamic Range Video Reconstruction. Creating HDR images from multi-exposure inputs is a long-studied problem in computational photography. A long line of work reconstructs HDR video by aligning and fusing alternatively exposed LDR frames (Kang et al., 2003; Kalantari et al., 2017; Chen et al., 2021; Chung and Cho, 2023; Xu et al., 2024). These approaches typically rely on optical flow or CNN-based alignment in 2D, followed by refinement to suppress ghosting. While effective for moderate motion, they remain vulnerable to occlusions, large displacements, and exposure inconsistencies. In contrast, our work reconstructs HDR video in 3D, enabling consistent rendering even under challenging dynamics.

Dynamic Scene Reconstruction. NeRF-based methods such as NSFF (Li et al., 2021), DynIBaR (Li et al., 2023), HyperNeRF (Park et al., 2021), and factorized grid models like HexPlane (Cao and Johnson, 2023) and K-Planes (Fridovich-Keil et al., 2023) have advanced free-viewpoint rendering of dynamic scenes. These methods represent a scene as a continuous function of space and time, sometimes augmented with deformation fields or canonical templates. They can synthesize novel views or even novel time steps. In parallel, 3D Gaussian Splatting has recently been extended to dynamic settings through 4DGS (Wu et al., 2024b), MotionGS (Zhu et al., 2024), Gaussian Marbles (Stearns et al., 2024), and DeformableGS (Yang et al., 2024b), achieving high efficiency and real-time rendering. Despite their success, all of these methods assume photometrically consistent LDR inputs and do not address the challenges of HDR content. Thus, they struggle to faithfully represent scenes with extreme lighting variations, whereas our approach explicitly targets HDR reconstruction of dynamic radiance fields.

High Dynamic Range Novel View Synthesis. Several recent works integrate HDR modeling into volumetric representations, mainly for static scenes. HDR-NeRF (Huang et al., 2022) and HDR-Plenoxel (Jun-Seong et al., 2022) model radiance together with tone-mapping or exposure functions, enabling HDR novel view synthesis from multi-exposure data. GaussHDR (Liu et al., 2025) extends HDR reconstruction to Gaussian Splatting with local tone mapping, while LTM-NeRF (Huang et al., 2024) embeds spatially varying tone mapping directly into NeRF. These works demonstrate the benefits of HDR-aware radiance fields but assume static content. The most relevant to our work is HDR-HexPlane (Wu et al., 2024a), which extends a factorized grid representation to dynamic HDR scenes by learning per-image exposure mappings. However, it does not explicitly model 3D motion, limiting its ability to represent complex dynamics and to perform temporal synthesis. In contrast, our method incorporates explicit motion modeling, allowing robust HDR reconstruction from real-world alternating-exposure videos and supporting both novel-view and novel-time rendering.

3 Preliminary

Neural Scene Flow Fields. Neural Scene Flow Fields (NSFF) extend NeRF (Mildenhall et al., 2020) by jointly modeling static and dynamic components of a scene. The dynamic branch, F_θ^{dy}, takes spatial location $\mathbf{x}$, view direction $\mathbf{d}$, and time t as inputs, and predicts color c_t^{dy}, density σ_t^{dy}, forward/backward scene flow F_t, and disocclusion weights W_t:

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

Scene flow is used to warp 3D points across time for enforcing temporal consistency. The static branch, F_θ^{st}, models time-invariant appearance:

$$(c^{\text{st}}, \sigma^{\text{st}}, v) = F_\theta^{\text{st}}(\mathbf{x}, \mathbf{d}), \tag{2}$$

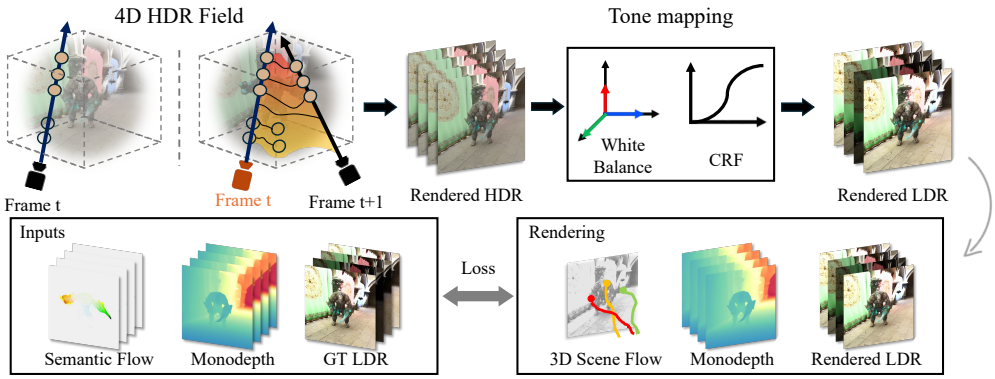


Figure 3: **Overall pipeline of our propsed method.** HDR-NSFF takes an alternatively exposed video as input and estimate 3D scene flow for the sampled points along each ray. Neighboring frames are then warped to render the HDR radiance at the target frame, which is tone-mapped to LDR via a white-balance 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 v is a blending weight. The final color is obtained by volume rendering with static–dynamic combination:

$$\hat{C}_i(r_i) = \int_{z_n}^{z_f} T_i(z) \left[v(z) c^{\text{st}}(z) \sigma^{\text{st}}(z) + (1 - v(z)) c_i^{\text{dy}}(z) \sigma_i^{\text{dy}}(z) \right] dz. \tag{3}$$

Here, $T_i(z)$ denotes transmittance along the ray. This formulation allows NSFF to capture both persistent geometry and spatio-temporal dependent motion within a unified radiance field.

4 HDR-NSFF:High Dynamig Range Neural Scene Flow Fields

Our framework builds upon Neural Scene Flow Fields (NSFF) to reconstruct dynamic HDR radiance fields from alternatively exposed monocular videos. NSFF exploits physical priors such as depth and optical flow for consistent learning in LDR videos, while recent HDR radiance field methods introduce tone-mapping modules but remain limited to static scenes. However, a direct combination of these ideas is not sufficient for HDR video.

Dynamic HDR videos present fundamental challenges: alternating exposures cause severe color inconsistency, which (i) prevents off-the-shelf models from delivering reliable performance and (ii) limits the effectiveness of tone-mapping regularization. Addressing this requires a systematic approach that disentangles and rethinks each component in light of HDR-specific aspects. Building on this perspective, HDR-NSFF is designed as an integrated framework that introduces tailored modules and empirically grounded analyses, offering a coherent solution for dynamic HDR 4D reconstruction.

HDR-NSFF integrates four core components: (i) exposure-robust semantic flow estimation for reliable motion learning (Sec. 4.1), (ii) robust depth estimation using a carefully selected model verified through empirical analysis (Sec. 4.1), (iii) NSFF-based radiance field and tone-mapping joint optimization, where we experimentally analyze tone-mapping function (Sec. 4.2, and (iv) generative prior regularization to compensate for the sparse-view limitation of monocular input (Sec. B.7). An overview of the pipeline is illustrated in Fig. 3.

4.1 Exposure Robust Learning Strategies

Semantic based Optical Flow. A key challenge in reconstructing HDR dynamic scenes from alternatively exposed video is that frame-to-frame color inconsistencies significantly degrade the reliability of conventional optical flow methods. Standard alignment techniques such as RAFT (Teed and Deng, 2020) often fail under severe exposure variations (see Fig. 4 (a)).

In this context, we focus on the abundant embedding space of the self-supervised vision foundation model DINOv2 (Oquab et al., 2023), which has demonstrated strong robustness to photometric

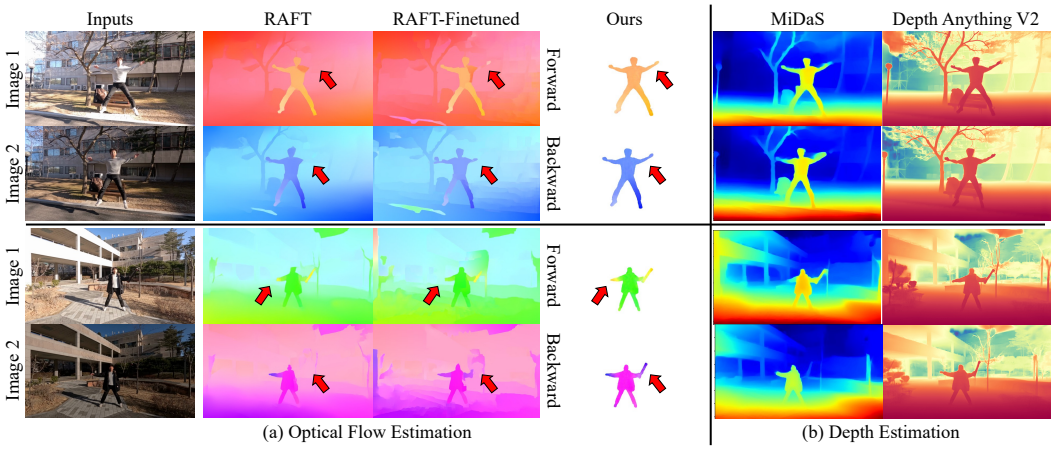


Figure 4: **Visualization of flow and depth estimation between varying exposed images.** (a) RAFT often fails under varying exposure conditions, yielding noticeable errors. Fine-tuning on synthetic varying exposed data (RAFT-Finetuned) improves performance moderately, but our semantic-based approach achieves higher accuracy. As highlighted by the red arrows, RAFT and RAFT-Finetuned miss correct motion. (b) For depth estimation, Depth Anything V2 (Yang et al., 2024a) recovers finer structural details and sharper object boundaries compared to MiDaS (Ranftl et al., 2020).

corruptions and perturbations, as shown by experiments on ImageNet-C (Hendrycks and Dietterich, 2019). We further investigate and observe the feature consistency, *i.e.*, robustness, across multi-exposure settings. These analyses are provided in the appendix. Built upon these observations, we adopt a DINOv2-based point tracking method, DINO-Tracker (Tumanyan et al., 2024), as motion estimation method, with a simple yet effective modification to ensure compatibility with our pipeline.

Since tracking errors accumulate with increasing frames under exposure variance, we redefine tracking points at each timestep and estimate only the flow between adjacent frame pairs in both forward and backward directions, as required by our pipeline. we also introduce motion masks from SAM2 (Ravi et al., 2024) to restrict DINO-Tracker to operate only within motion regions for preventing from noisy tracking performance in background. As a result, our semantic-based optical flow achieves robust motion estimation even in the presence of severe exposure variation (Fig. 4), providing consistent motion cues that are critical for HDR-NSFF.

Depth analysis under varying exposure. We investigate the robustness of off-the-shelf depth estimation methods under exposure variance. For evaluation, we synthetically simulate ±2 EV exposure changes from given RGB images. Specifically, we first convert RGB images into pseudo-RAW representations using a learning-based RAW estimation model (Xing et al., 2021), and then generate ±2 EV RGB images through the standard sRGB transformation. Following the evaluation protocol of prior work (Ke et al., 2024), we report the Absolute Mean Relative Error (AbsRel) of depth estimation methods on the NYUv2 and ScanNet datasets (Silberman et al., 2012; Dai et al., 2017), with the results summarized in Table 1. While each methods shows the suboptical accuracy on ±2 EV condition, Depth-Anything-V2 demonstrate the most robust performance, achieving the best results under these challenging conditions. Based on this analysis, we adapt the geometric prior of Depth-Anything-V2 in our pipeline.

Methods	NYUv2			ScanNet		
	Original	+2EV	-2EV	Original	+2EV	-2EV
MiDaS	9.08	13.68	9.35	8.66	13.78	10.22
DPT	9.21	12.96	8.95	8.27	13.62	9.57
Marigold	5.81	11.26	6.66	7.24	14.26	8.33
Depth-Anything-V2	4.87	7.63	5.10	4.82	10.57	6.36

Table 1: **Depth estimation results under exposure variance.** We employ AbsRel as the evaluation metric.

4.2 TONE-MAPPING

Our goal is to reconstruct HDR dynamic radiance fields, encompassing both 3D space and motion, from 2D multi-exposure LDR RGB images. A crucial component in this process is the tone-mapping module, which bridges the gap between varying 2D observations and a coherent 3D HDR

representation. Specifically, tone mapping can be expressed as:

$$C = \mathcal{T}(E, \theta) = g\big(w(E)\big), \tag{4}$$

where E denotes the rendered radiance, w the white balance correction, and θ the radiometric parameters. In the absence of a known camera response function (CRF), the choice of tone-mapping module $\mathcal{T}(\cdot, \theta)$ determines the flexibility with which HDR radiance can be effectively recovered from LDR inputs. Moreover, to build consistent HDR representations in 3D space, the tone-mapping module must also act as a regularizer, preventing fluctuations in HDR results under multi-exposure conditions. This combination of flexibility and regularization largely influences the overall quality and stability of HDR field reconstruction.

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o Tone-mapping	17.79	0.7048	0.0705	15.59	0.5577	0.1339
Fix CRF	25.55	0.8391	0.0487	20.43	0.6904	0.0911
MLP CRF	28.76	0.8861	0.0394	21.48	0.7256	0.0776
Piecewise CRF	31.01	0.9301	0.0233	22.55	0.7714	0.0697

Table 2: **Comparison of tone-mapping designs.**

From the perspectives of flexibility and regularization, we revisit the prior HDR radiance studies (Huang et al., 2022; Jun-Seong et al., 2022; Wu et al., 2024a), which explored different forms of CRF designs: **Fix CRF**, a non-learnable handcrafted mapping (HDR-HexPlane); **MLP CRF**, a fully learnable MLP converting RGB to HDR function (HDR-NeRF); and **Piecewise CRF**, a parametric form optimizing the white balance correction w and CRF (HDR-Plenoxels). The Fix CRF enforces strong regularization but lacks flexibility, whereas the MLP CRF provides high flexibility at the cost of weak regularization. We adopt the Piecewise CRF, as it is designed to balance the two by applying per-channel scaling through learnable white balance parameters. $\theta_w = [w_r, w_g, w_b]^\top$. Through experiments on our real GoPro dataset under the novel view synthesis setting, we observe that the Piecewise CRF achieves the highest quantitative scores (see Table 2). These results suggest that a fixed CRF (Wu et al., 2024a) lacks the flexibility to account for camera-specific response deviations, while fully unconstrained learning (Huang et al., 2022) often leads to convergence instability.

Reconstructing dynamic HDR scenes from monocular videos is particularly challenging due to the coupled effects of sparse temporal observations and information loss caused by saturation under extreme exposures. At each timestep, only a single viewpoint is available, and if this observation is saturated, the lost information cannot be recovered directly.

To mitigate this issue, we adopt a generative prior to compensate for information loss, extending its use from static scene reconstruction (Wu et al., 2025) to dynamic HDR scene reconstruction. During training, we periodically render unobserved views and enhance them using the pre-trained prior (see Fig. S3), which are then incorporated back into the training loop as pseudo-observations.

This iterative augmentation helps recover and enforce consistency in regions affected by saturation, leading to more reliable reconstruction of previously lost information. We also observe that it improves geometric fidelity while also enhancing perceptual quality. In practice, we apply this generative prior at controlled intervals to balance the benefit of additional supervision against potential artifacts from generative hallucination (Sim and Moon, 2025). As a result, our framework is able to recover more coherent structures and produce visually plausible HDR reconstructions, even under the severe saturated condition and view sparsity of monocular, alternatively exposed videos.

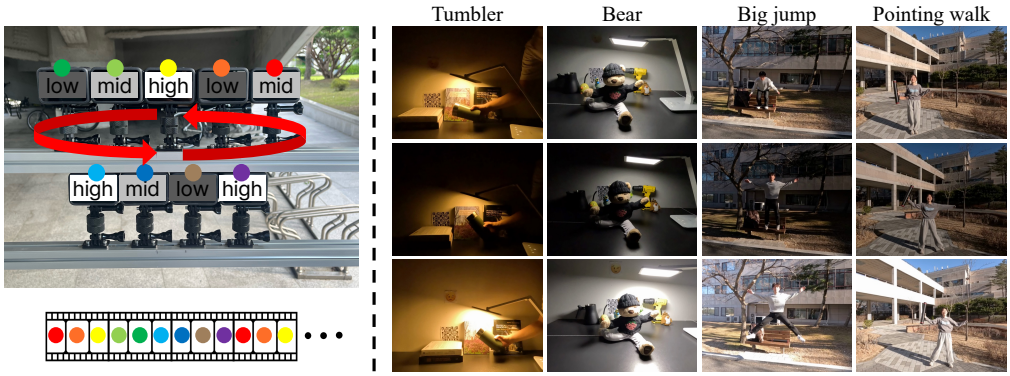
4.3 Objective Function

We train both the neural scene flow fields and the tone-mapping module by minimizing the Mean Absolute Error (MAE) between rendered LDR views and ground-truth frames. Following NSFF (Li et al., 2021), we replace the rendered color $\hat{C}$ with our tone-mapped output $\mathcal{T}(\hat{E})$, where $\hat{E}$ denotes the rendered HDR radiance. The superscript cb denotes the **combined** rendering that fuses static and dynamic components of the scene. The photometric losses are:

$$L_{cb} = \sum_{r_i} \|\mathcal{T}(\hat{E}_i^{cb}(r_i)) - C_i(r_i)\|_1, \quad \text{and} \tag{5}$$

$$L_{photo} = \sum_{r_i} \sum_{j \in \mathcal{N}(i)} \|\mathcal{T}(\hat{E}_{j\to i}(r_i)) - C_i(r_i)\|_1, \tag{6}$$

where r denotes a camera ray. Here, $\hat{E}_{j\to i}(r_i)$ denotes the HDR radiance warped from a frame j to i. We also adopt the optical flow and single-view depth prior, denoted L_{Flow} and L_{depth} to regularize



(a) Capture setup (b) Sampled sequences

Figure 5: **Evaluation setup and sampled sequences from our proposed GoPro dataset**. To evaluate novel view synthesis, we use nine GoPro Hero 13 Black cameras arranged at two height levels with fixed intervals, synchronized to record multi-view video at three exposures (mid, low, high). We construct a monocular alternatively exposed video by selecting one frame per time step across exposures, and use the remaining views for evaluation. Note that the input of our method is a monocular video and the setup described here is designed to evaluate the system.

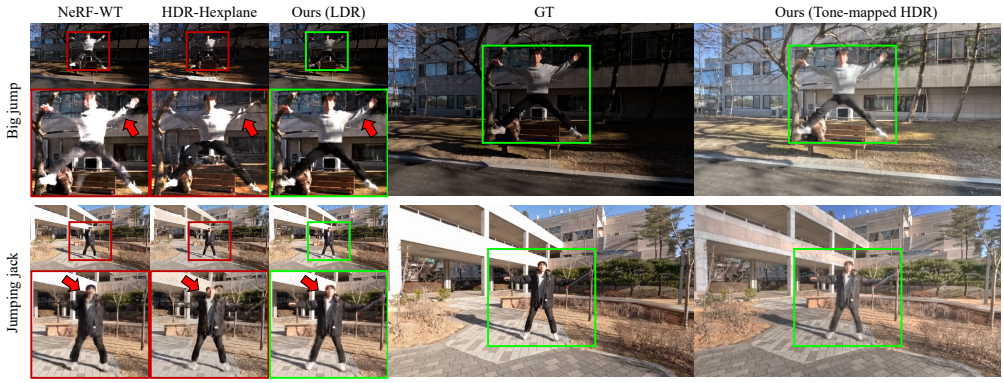


Figure 6: **Qualitative results of novel view synthesis on real multi-exposed dynamic scene dataset.** Our method maintains more consistent geometry and color across varying viewpoints, while other approaches (NeRF-WT (Quei-An, 2020), HDR-Hexplane (Wu et al., 2024a)) exhibit noticeable artifacts and geometric distortions.

monocular reconstruction followed by NSFF (Li et al., 2021). For the CRF and generative prior objective functions, we apply L_{smooth} and L_{gen}, respectively. The total objective function of our HDR-NSFF is as follows:

$$L = L_{cb} + L_{photo} + \beta_{data} L_{data} + \beta_{reg} L_{reg} + L_{smooth}, \tag{7}$$

where β are coefficients weight each term. The details can be found in the the supplementary material.

5 Experiments

HDR-NSFF takes as input an alternatively exposed monocular video of a dynamic scene and jointly reconstructs HDR radiance, 3D motion, and tone-mapping. We evaluate its performance specifically designed for each subtask, including novel view synthesis, novel time synthesis, and novel view and time synthesis. 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. All models are evaluated using standard metrics (PSNR, SSIM (Wang et al., 2003), and LPIPS (Zhang et al., 2018)), where results are reported on both real-world and synthetic datasets. Metrics are all averaged and each color stands for the best and the second best, respectively.

We compare against representative baselines across three categories: (1) dynamic scene reconstruction methods (NSFF (Li et al., 2021), 4DGS (Wu et al., 2024b), MotionGS (Zhu et al., 2024)), which

degrade under alternatively exposed inputs; (2) dynamic reconstruction with appearance embeddings (NeRF-WT (Quei-An, 2020)), which can disentangle transient appearance but cannot consistently model shared scene motion; and (3) HDR volumetric reconstruction (HDR-HexPlane (Wu et al., 2024a)), the closest prior tackling HDR in a dynamic 3D representation.

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 7 sec.

5.1 DATASETS

Proposed GoPro Dataset. While standard alternatively exposed videos are sufficient for training HDR-NSFF, a single-camera setup cannot support evaluating novel view/time synthesis under varying exposures. To address, we construct a real-world dataset captured with synchronized GoPro Hero 13 cameras at three exposures (low, mid, high). This dataset provides the first benchmark for dynamic HDR reconstruction in real-world settings with explicit exposure variation. Figure 5 illustrates the camera setup. Inspired by prior work (Yoon et al., 2020), we adopt a similar strategy but modify it to accommodate varying exposures: at each timestamp, we select one frame per viewpoint from a single camera for training, while reserving the remaining views for evaluation under exposure changes.

Synthetic Dataset. We modify the dataset proposed in HDR-HexPlane (Wu et al., 2024a). 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.

5.2 RESULTS

Novel View Synthesis. We evaluate novel view synthesis on our proposed GoPro dataset. 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 to LDR. We then compare these tone-mapped views against the GT LDR images. It directly assesses two key aspects: (1) the quality of dynamic scene modeling, and (2) the accuracy of tone-mapping functions. Table 3 shows that our approach achieves significant improvements in rendering fidelity compared to baselines, both in highly dynamic regions and across the entire scene. Figure 6 its effectiveness in reconstructing HDR scenes with fine detail across varying exposures. Methods without appearance embedding (NSFF (Li et al., 2021), 4DGS (Wu et al., 2024b), MotionGS (Zhu et al., 2024)) fail to reconstruct consistent HDR views under alternating exposures. NeRF-WT (Quei-An, 2020) and HDR-Hexplane (Wu et al., 2024a) provide limited robustness but still struggle in real-world dynamic settings.

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 (see Fig. 7). Following NSFF (Li et al., 2021), we remove every other frame from the original video sequences during training, and use the intermediate frames at held-out camera viewpoints for testing. Table 5 shows that our results outperform competing models across all evaluation metrics.

For real-world evaluation on our GoPro dataset, we extend this setting to simultaneously test novel view and time synthesis. While all camera views are retained to ensure realistic multi-view coverage, we subsample frames from each video and evaluate the model at unseen time instances and camera viewpoints. This joint evaluation directly measures the fidelity of both HDR radiance reconstruction and learned 3D motion under exposure-varying, dynamic scenes. Importantly, in this experiment as well, our model consistently surpasses all baseline methods 4.

While HDR-NSFF explicitly models 3D scene motion, enabling reliable synthesis across both space and time. In contrast, HDR-HexPlane does not incorporate explicit motion modeling, which limits its ability to handle space and time interpolation in dynamic HDR scenes.

Qualitative comparison of HDR reconstruction. To validate our HDR reconstruction, we qualitatively compare our results with ground-truth HDR images (see Fig. 8). Tone-mapped HDR views from our model closely match ground truth, preserving fine details in both under- and overexposed

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NSFF	17.04	0.6493	0.2243	15.22	0.5264	0.2644
4DGS	20.52	0.7717	0.1608	15.92	0.4992	0.2376
MotionGS	13.73	0.3282	0.3956	10.47	0.1697	0.5187
NeRF-WT	28.51	0.9215	0.0691	17.30	0.5847	0.1976
HDR-HexPlane	20.53	0.6243	0.2164	18.61	0.6370	0.1874
Ours w/ RAFT	31.00	0.9303	0.0688	23.30	0.7894	0.1168
Ours w/ Dino-Tracker	31.50	0.9363	0.0645	23.50	0.7930	0.1166
Ours w/ Generative prior	31.48	0.9359	0.0644	23.40	0.7936	0.1131

Table 3: **Averaged quantitative results of novel view synthesis on GoPro dataset.** Our HDR-NSFF achieves the best overall performance, with DINO-Tracker offering the strongest improvement in motion-consistent reconstruction and the generative prior further enhancing perceptual quality.

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NSFF	17.96	0.694	0.225	16.48	0.5548	0.294
HDR-HexPlane	20.15	0.5718	0.2182	17.37	0.5626	0.2145
Ours w/ RAFT	31.32	0.9335	0.0727	22.98	0.7687	0.1626
Ours w/ Dino-Tracker	31.75	0.937	0.0691	23.18	0.7754	0.1628
Ours w/ Generative prior	31.71	0.9365	0.0692	23.1	0.7754	0.1597

Table 4: **Averaged quantitative results of novel view and time synthesis on GoPro dataset.** Our method outperform baseline models.

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NSFF	15.98	0.6457	0.1388	16.04	0.5697	0.1527
NeRF-WT	31.10	0.9366	0.0342	21.50	0.7490	0.0895
HDR-HexPlane	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 5: **Averaged quantitative results of novel view and time synthesis on synthetic data.** Our method outperform baseline models.

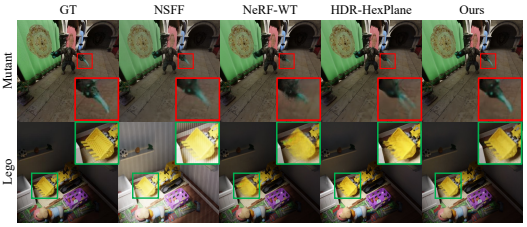


Figure 7: **Qualitative results of novel view and time synthesis on synthetic data.** Since, our approach explicitly models scene flow, it excels at time interpolation.

Figure 8: **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 and ground truth. For better visualization, we plot HDR histogram using smoothed kde method.

regions. Histograms of pixel intensities further show that our reconstructions cover the full radiance range, recovering values from very low to high intensities. In addition, novel LDR views rendered at multiple exposures confirm that our method accurately controls exposure, reproducing under- and over-saturation effects.

6 Conclusion

In this work, we introduced HDR-NSFF, the first framework that jointly reconstructs HDR radiance, 3D motion, and tone-mapping from alternatively exposed monocular videos. By explicitly modeling scene flow and integrating learnable tone-mapping, our approach addresses the fundamental limitations of prior HDR methods that operate purely in 2D image space. We further enhanced robustness through semantic-based optical flow, depth priors, and generative prior, enabling reliable reconstructions under severe exposure variations and sparse temporal observations. Extensive experiments on both real and synthetic datasets demonstrated that HDR-NSFF consistently outperforms baselines across novel view synthesis, novel time synthesis, and combined view-time synthesis. In particular, our method achieves sharper geometry, more faithful HDR radiance, and temporally coherent results compared to state-of-the-art dynamic scene and HDR reconstruction models.

ETHICS STATEMENT

Our work involves two datasets. The HDR-HexPlane synthetic dataset is distributed under the MIT license and was employed exclusively for research, including rendering experiments. As a synthetic and openly licensed resource, it entails no issues of privacy or confidentiality. We additionally collected a real-world GoPro dataset with the informed consent of participants with participants fully informed of the scope and objectives of the research. The dataset is used solely for academic research purposes.

REPRODUCIBILITY STATEMENT

Our method is built upon the open-source Neural Scene Flow Fields (NSFF) and DINO-Tracker. Implementation details are provided in Section B of the appendix, including data acquisition, objective functions, and training procedures. If accepted, we plan to release both our method and the collected dataset as open-source.

REFERENCES

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

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.

Haesoo Chung and Nam Ik Cho. Lan-hdr: Luminance-based alignment network for high dynamic range video reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12760–12769, 2023.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

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.

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.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18398–18408, 2022.

Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, and Qing Wang. Ltm-nerf: Embedding 3d local tone mapping in hdr neural radiance field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10944–10959, 2024.

Kim Jun-Seong, Kim Yu-Ji, Moon Ye-Bin, and Tae-Hyun Oh. Hdr-plenoxels: Self-calibrating high dynamic range radiance fields. In *European Conference on Computer Vision*, pages 384–401. Springer, 2022.

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

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

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024.

Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6498–6508, 2021.

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

Jinfeng Liu, Lingtong Kong, Bo Li, and Dan Xu. Gausshdr: High dynamic range gaussian splatting via learning unified 3d and 2d local tone mapping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5991–6000, 2025.

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.

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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.

Chen Quei-An. Nerf pl: a pytorch-lightning implementation of nerf. 2020.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

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.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012.

Geonhee Sim and Gyeongsik Moon. PERSONA: Personalized whole-body 3D avatar with pose-driven deformations from a single image. In *ICCV*, 2025.

Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

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

Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. In *European Conference on Computer Vision*, pages 367–385. Springer, 2024.

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.

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)*, 2024a.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024b.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *CVPR*, pages 26024–26035, 2025.

Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *CVPR*, 2021.

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.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024a.

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

Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *CVPR*, 2020.

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.

Ruijie Zhu, Yanzhe Liang, Hanzhi Chang, Jiacheng Deng, Jiahao Lu, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Motiongs: Exploring explicit motion guidance for deformable 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 37:101790–101817, 2024.

A APPENDIX

In this appendix material, we provide additional details omitted from the manuscript. Sec. B covers implementation. Sec. B.3 outlines the regularization terms used for NSFF physical prior, and Detailed description about the DINO-Tracker model. Lastly, Sec. C includes additional experimental results not shown in the manuscript due to page limits. We also provide a supplementary video that highlights novel-view rendering results.

B IMPLEMENTATION DETAILS

B.1 COUNTERPARTS

In this chapter, we briefly explain the method we compared as a counterparts in our experiments.

NeRF-WT. NeRF-W (Martin-Brualla et al., 2021) introduces per-image appearance and transient embedding, modelling to handle dynamic changes such as lighting variations and moving objects. In our experiments, we adapted NeRF-W to a dynamic HDR video (named NeRF-WT) using appearance embedding for ISP modelling and transient part for scene dynamics. We follow the hyperparameters given in the codebase. For implementation we used the codebase in `https://github.com/kwea123/nerf_pl`

HDR-Hexplane. HDR-Hexplane (Wu et al., 2024a) adopted Hexplane (Cao and Johnson, 2023) for the dynamic 3D representation and MLP with exposure embeddings accompanied with fixed gamma function to optimize ISP module. We follow the hyperparameters following manuscript. For implementation we used the codebase in `https://github.com/hustvl/HDR-HexPlane`

B.2 DATASET

Synthetic. We select four synthetic scenes for evaluation: *Lego*, *Mutant*, *Jumping Jack*, and *Stand Up*. Each image has a resolution of 800×800, with exposure values spanning from -2EV to 5EV. To maximize the influence of exposure change, we carefully adjust the camera viewpoints and lighting directions.

The sampling rate is determined based on the motion speed of each scene. Specifically, the *Lego* scene is subsampled by selecting every 10th frame, whereas the remaining scenes are sampled by skipping every two frames.

Real. For the real dataset, we preset exposure time for each cameras before acquisition. We set exposure time differently for each sequence. Sequence lengths and corresponding exposure information are detailed in the Table S1 All sequences are synchronized using the GoPro software.

name	Exp. Time [s]	# of frames
big jump	$\frac{1}{960}, \frac{1}{2880}, \frac{1}{7680}$	324
side walk	$\frac{1}{960}, \frac{1}{2880}, \frac{1}{7680}$	324
jumping jack	$\frac{1}{720}, \frac{1}{1920}, \frac{1}{7680}$	324
pointing walk	$\frac{1}{720}, \frac{1}{1920}, \frac{1}{7680}$	324
tube toss	$\frac{1}{720}, \frac{1}{1920}, \frac{1}{7680}$	324
bear	$\frac{1}{120}, \frac{1}{480}, \frac{1}{1920}$	324
dog	$\frac{1}{120}, \frac{1}{480}, \frac{1}{1920}$	324
tumbler	$\frac{1}{120}, \frac{1}{480}, \frac{1}{1920}$	324

Table S1: **Parameter setting for real dataset**

B.3 EXPERIMENTAL SETUP

To facilitate understanding of the experimental setup employed for the real dataset experiments, we provide an illustrative diagram in Fig. S1 In the novel view synthesis experiment, performance is

evaluated by measuring the differences between synthesized results and the images captured from cameras that were excluded from the training set, for all camera views i. In the novel view and time synthesis experiment, we evaluate performance by holding out certain segments of the time sequence and measuring how accurately these withheld segments are inferred.

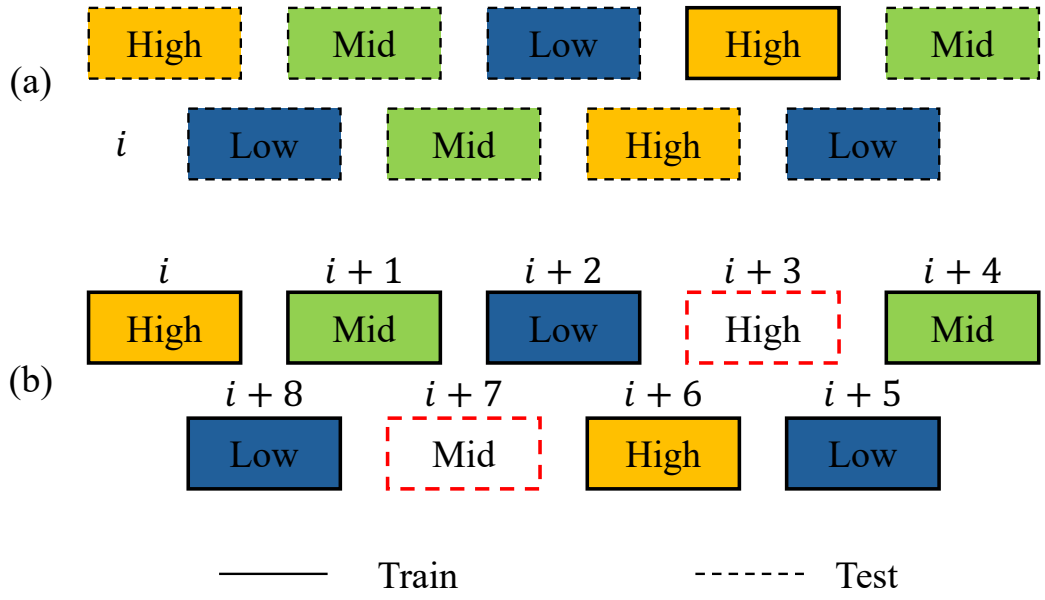


Figure S1: **Illustration of two experimental setting.** We illustrate two experimental settings described in Sec. 4.2 in manuscript: (a) Novel view synthesis (b) Novel view and time synthesis.

B.4 Details of Neural Scene Flow Fields

To model dynamic scenes, NSFF (Li et al., 2021) extend the concept of NeRF (Mildenhall et al., 2020) 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:

$$(c_t, \sigma_t, F_t, W_t) = F_\theta^{\text{dy}}(\mathbf{x}, \mathbf{d}, t). \tag{8}$$

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:

$$\hat{C}_{j\to i}(r_i) = \int_{z_n}^{z_f} T_j(z)\, \sigma_j\big(r_{i\to j}(z)\big)\, c_j\big(r_{i\to j}(z), d_i\big)\, dz, \tag{9}$$

$$\text{where} \quad r_{i\to j}(z) = r_i(z) + \mathbf{f}_{i\to j}\big(r_i(z)\big). \tag{10}$$

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

$$L_{photo} = \sum_{r_i} \sum_{j\in\mathcal{N}(i)} \|\hat{C}_{j\to i}(r_i) - C_i(r_i)\|_2^2. \tag{11}$$

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:

$$(c, \sigma, v) = F_\theta^{\text{st}}(\mathbf{x}, \mathbf{d}). \tag{12}$$

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:

$$\hat{C}_i^{cb}(r_i) = \int_{z_n}^{z_f} T_i^{cb}(z)\, \sigma_i^{cb}(z)\, e_i^{cb}(z)\, dz, \tag{13}$$

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

$$\sigma_i^{cb}(z)c_i^{cb}(z) = v(z)c(z)\sigma(z) + (1 - v(z))c_i(z)\sigma_i(z). \tag{14}$$

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:

$$L_{cb} = \sum_{r_i} \|\hat{C}_i^{cb}(r_i) - C_i(r_i)\|_2^2. \tag{15}$$

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.

Following section, we elaborate on data-driven prior loss (Flow loss and Single-view depth loss) and additional regularization terms introduced by NSFF (Li et al., 2021): Scene Flow Cycle Consistency and Low-Level regularization term. We employ additional regularization terms consistently in both our model and NSFF.

Flow Loss (L_{flow}). Flow loss operates by minimizing the discrepancy between observed 2D pixel correspondences, computed from pretrained optical flow networks and predicted 2D pixel correspondences, obtained by projecting predicted 3D scene flows. This aligns 3D scene flow with pretrained 2D motion estimation.

Given two adjacent frames at times i and $j = i \pm 1$, L_{flow} is calculated as follows. Let p_i represent a pixel location at frame i. The corresponding pixel location at frame j, denoted by $p_{i\to j}$, can be computed using pretrained 2D motion estimation $u_{i\to j}$ as $p_{i\to j} = p_i + u_{i\to j}$.

The model predicts an expected scene flow $\hat{F}_{i\to j}(r_i)$ corresponding to 3D location $\hat{X}_i(r_i)$ along the ray r_i passing through the pixel p_i via volumetric rendering. Thus, the predicted 3D displacement can be expressed as $\hat{X}_i(r_i) + \hat{F}_{i\to j}(r_i)$. Then, by applying the perspective projection operator Π_j, corresponding to the camera viewpoint at frame j , the expected 2D pixel position $\hat{p}_{i\to j}(r_i)$ at frame j is calculated as:

$$\hat{p}_{i\to j}(r_i) = \Pi_j \left(\hat{X}_i(r_i) + \hat{F}_{i\to j}(r_i) \right). \tag{16}$$

Finally, the geometric consistency loss is computed by measuring the discrepancy between these two pixel positions (observed vs. predicted) using the L1-norm:

$$L_{flow} = \sum_{r_i} \sum_{j\in\{i\pm1\}} ||\hat{p}_{i\to j}(r_i) - p_{i\to j}(r_i)||_1. \tag{17}$$

Single-view Depth Prior (L_{depth}). encourages rendered depths to match predictions from a pre-trained depth model:

$$L_{depth} = \sum_{r_i} ||\hat{Z}_i^*(r_i) - Z_i^*(r_i)||_1, \tag{18}$$

where the superscript $(*)$ denotes scale-shift invariant normalization. These priors are combined into:

$$L_{data} = L_{flow} + \beta_{depth} L_{depth}, \tag{19}$$

where $\beta_{depth} = 2$ for all experiments.

Scene Flow Cycle Consistency. To ensure plausible scene motion, the loss ensures coherence between forward and backward predicted scene flows for adjacent frames, mathematically defined as:

$$L_{cyc} = \sum_{x_i} \sum_{j \in \{i \pm 1\}} w_{i \to j} || f_{i \to j}(x_i) + f_{j \to i}(x_{i \to j}) ||_1, \tag{20}$$

where $f_{i \to j}(x_i)$ indicates the predicted displacement (scene flow) of point x_i from time i to j.

Low-Level Regularization. Spatial-temporal smoothness is enforced through $l1$ regularization on scene flow estimated between neighboring sampled 3D points along rays. This encourages 3D point trajectories to be piecewise linear. Another sparsity regularization term, calculating an $l1$ loss in flow estimation is also applied. This encourage minimal scene flow magnitudes across most spatial regions.

CRF Smoothness Loss. We impose smoothness on the estimated camera response functions (CRFs) to ensure plausible variations (Debevec et al., 2023):

$$L_{\text{smooth}} = \sum\nolimits_{i=1}^{N} \sum\nolimits_{e \in [0,1]} g_i''(e), \tag{21}$$

where $g_i''(e)$ denotes the second-order derivative of CRFs with respect to their input domain. We incorporate a smoothness loss to enforce that CRF varies smoothly in a physically plausible manner (Debevec et al., 2023). It is defined as follows:

$$\mathcal{L}_{smooth} = \sum\nolimits_{i=1}^{N} \sum\nolimits_{e \in [0,1]} g_i''(e), \tag{22}$$

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:

$$L = L_{cb} + L_{photo} + \beta_{data} L_{data} + \beta_{reg} L_{reg} + L_{smooth}, \tag{23}$$

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

Generative Prior Loss. To mitigate the sparse-view limitation of monocular input, we adopt enhanced views generated via a diffusion-based prior (Wu et al., 2025). For these views, we apply a patch-wise perceptual loss to encourage realistic and view-consistent appearance:

$$L_{\text{gen}} = \sum_{p \in \mathcal{P}} \| \phi(\hat{C}_p) - \phi(C_p^{\text{gen}}) \|_1, \tag{24}$$

where ϕ denotes a perceptual feature extractor, and p indexes sampled patches. Since generative priors may introduce hallucinations, we carefully balance their contribution by (i) delaying their use until a stable stage of training (200K iterations), and (ii) training with enhanced views at a low probability (10%) per iteration.

B.5 DINO-TRACKER

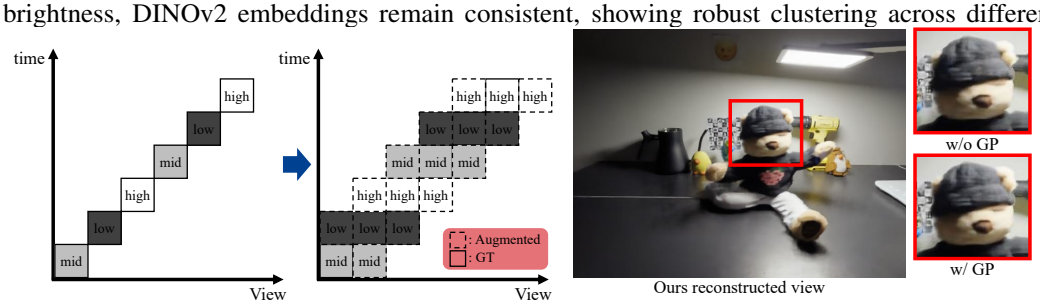
DINO-Tracker is a self-supervised framework designed to accurately track points over long sequences of video frames. Given an initial query point in an early frame of video, it estimates the trajectory of these points throughout subsequent frames. The method leverages pretrained deep features from the DINOv2-ViT (Oquab et al., 2023) model, which are refined by learning residual features via a small, trainable CNN module. DINO feature and residual feature are aggregated to find correspondence heatmap computed by cost volume. Lastly, additional CNN-refiner follows to further enhance matching.

Optimization is performed using several losses

- **Flow Loss (L_{flow})**: Ensures predicted trajectories align closely with short-term optical flow correspondences.

(a) Scene 1 (b) Scene 2

Figure S2: **DINOv2 feature visualization under varying exposures.** Despite large changes in brightness, DINOv2 embeddings remain consistent, showing robust clustering across different



(a) Multi-view augmentation using generative prior (b) After distilling generative prior (GP) to radiance fields

Figure S3: **Visualization of difix enhancement.**

- **DINO Best-Buddies Loss ($L_{\text{dino-bb}}$)**: Contrastively aligns refined features based on semantic matches from original DINO embeddings.
- **Refined Best-Buddies Loss ($L_{\text{rfn-bb}}$)**: Similar to DINO best-buddies loss but applied to newly detected reliable matches among refined features.
- **Cycle-Consistency Loss ($L_{\text{rfn-cc}}$)**: Encourages consistency in predicted trajectories, penalizing trajectories that fail a cycle-consistency criterion.
- **Prior Preservation Loss (L_{prior})**: Regularizes the refined features to remain close in norm and direction to original DINO features, ensuring semantic coherence is preserved.

In contrast to the original DINO-Tracker, our proposed approach introduces a novel utilization of this framework explicitly aimed at enhancing the robustness and accuracy of 2D dense correspondence estimation. Specifically, we propose deriving dense matching from consecutive frames using the trained DINO-Tracker model itself. Leveraging the semantic matching capability inherent to DINO features, our method provides robust optical flow estimates even in challenging conditions such as alternatively exposed video settings, where conventional texture-based methods typically degrade due to information loss. Figure S2 shows that DINOv2 features is robust to exposure variance.

B.6 Generative prior as a regularizer

B.7 Generative Prior for Recovering Saturated Information

In HDR-NSFF, we additionally employ Difix (Wu et al., 2025) as a regularizer to stabilize training under severe exposure inconsistencies. Difix provides a diffusion-based enhancement prior that guides the radiance field toward semantically consistent reconstructions when input frames suffer from brightness fluctuations or missing details. Concretely, we periodically generate pseudo-observations by enhancing intermediate renderings with the Difix prior and incorporate them into the optimization loop. This regularization not only improves geometric and radiometric stability but also enforces stronger multi-view consistency in dynamic scenes, where exposure variations and motion often break correspondences across views. As a result, HDR-NSFF achieves more coherent reconstructions that generalize better to unseen exposures and viewpoints.

Methods	real dataset-Full							
	big jump	side_walk	one arm swing	two arm swing	jumping jack	pointing walk	sit down up	tube toss
	PSNR							
NSFF	16.89	16.85	17.85	17.96	17.84	17.96	18.58	18.35
NeRF-WT	24.88	23.21	30.67	30.84	29.30	25.59	32.11	29.69
HDR-Hexplane	18.94	19.34	20.17	19.92	17.32	17.08	17.41	17.52
Ours	28.77	28.70	32.73	33.11	31.34	28.84	32.89	31.69
	SSIM							
NSFF	0.6391	0.6458	0.6580	0.6687	0.7616	0.7423	0.7644	0.7588
NeRF-WT	0.8780	0.8747	0.9345	0.9312	0.9189	0.9024	0.9410	0.9308
HDR-Hexplane	0.4844	0.5025	0.5401	0.5347	0.4639	0.4572	0.4618	0.4763
Ours	0.9066	0.9136	0.9431	0.9456	0.9328	0.9196	0.9419	0.9375
	LPIPS							
NSFF	0.0948	0.0831	0.0799	0.0746	0.0534	0.0709	0.0547	0.0525
NeRF-WT	0.0518	0.0487	0.0226	0.0250	0.0298	0.0374	0.0190	0.0236
HDR-Hexplane	0.1879	0.1656	0.1597	0.1402	0.1437	0.1452	0.1368	0.1516
Ours	0.0348	0.0301	0.0174	0.0163	0.0227	0.0276	0.0174	0.0197

Methods	real dataset-Dynamic only							
	big jump	side_walk	one arm swing	two arm swing	jumping jack	pointing walk	sit down up	tube toss
	PSNR							
NSFF	13.97	13.21	15.81	16.29	17.52	13.98	17.27	16.70
NeRF-WT	13.30	11.46	19.96	20.17	18.33	13.50	18.98	17.28
HDR-Hexplane	13.73	15.90	19.86	20.19	17.96	15.72	20.04	17.61
Ours	19.42	19.15	25.55	26.45	22.77	18.67	25.34	23.04
	SSIM							
NSFF	0.4308	0.4003	0.6535	0.6618	0.5039	0.4977	0.6547	0.6592
NeRF-WT	0.3566	0.2865	0.8208	0.8272	0.5182	0.3965	0.6891	0.6963
HDR-Hexplane	0.2853	0.5376	0.7814	0.7697	0.5067	0.4973	0.7015	0.7036
Ours	0.6329	0.6322	0.9132	0.9177	0.7392	0.6207	0.8604	0.8549
	LPIPS							
NSFF	0.1815	0.1689	0.0992	0.0859	0.1102	0.2423	0.1081	0.0754
NeRF-WT	0.2273	0.1957	0.0453	0.0521	0.1412	0.1743	0.1044	0.0741
HDR-Hexplane	0.2589	0.1403	0.0803	0.0991	0.1386	0.1301	0.0998	0.0744
Ours	0.1168	0.1198	0.0224	0.0223	0.0812	0.1040	0.0540	0.0372

Table S2: **Quantitative results of novel view synthesis on real data.** The green and yellow colors stand for the best and the second best, respectively.

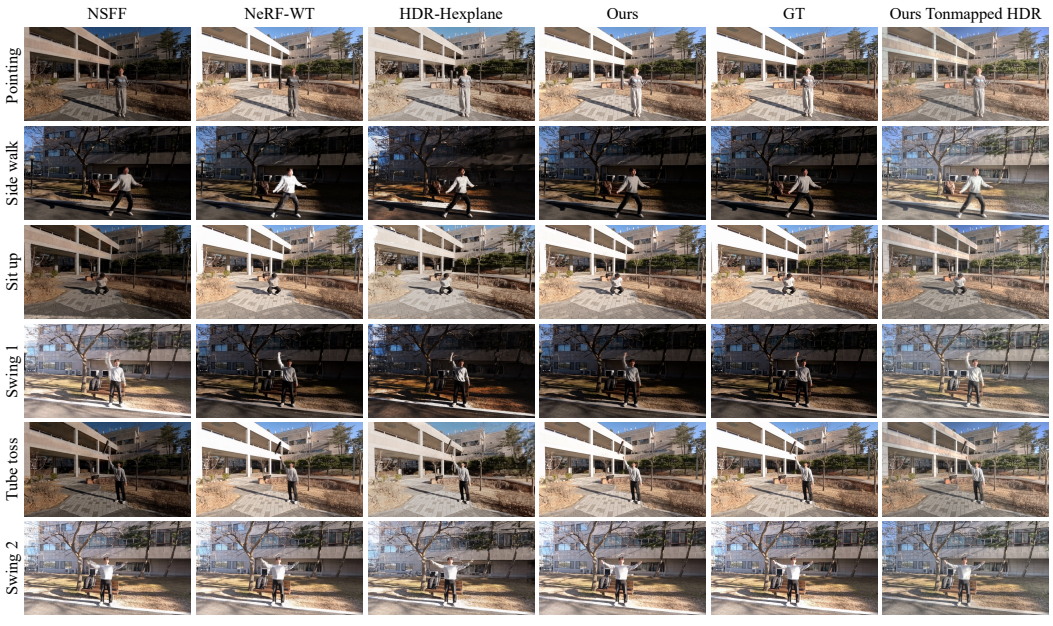


Figure S4: **Qualitative results on real dataset.**

C ADDITIONAL EXPERIMENT RESULTS

We provide additional experimental results that could not be included in the main manuscript, due to page limit. Specifically, Tables S2, S3, & S4 present per-scene quantitative results for each experiment. Figure S4 illustrates qualitative outcomes for additional real datasets not shown in the

Methods	big jump	side_walk	one arm swing	two arm swing	jumping jack	pointing walk	sit down up	tube toss
real dataset-Full								
PSNR								
NSFF	17.20	16.97	17.88	17.29	18.21	17.64	18.61	18.36
NeRF-WT	24.16	24.33	29.73	29.99	29.61	25.53	30.94	29.66
HDR-Hexplane	21.39	21.87	23.14	22.33	21.03	20.57	21.26	20.85
Ours	29.24	31.30	32.34	32.67	31.79	30.81	32.82	30.87
SSIM								
NSFF	0.6362	0.6328	0.6592	0.6590	0.7469	0.7353	0.7694	0.7839
NeRF-WT	0.8563	0.8727	0.9163	0.9192	0.9065	0.8943	0.9281	0.9187
HDR-Hexplane	0.6267	0.6575	0.6952	0.6443	0.6406	0.6347	0.6589	0.6439
Ours	0.9048	0.9259	0.9402	0.9430	0.9331	0.9292	0.9409	0.9263
LPIPS								
NSFF	0.0993	0.1060	0.0848	0.0899	0.0666	0.0816	0.0488	0.0518
NeRF-WT	0.0643	0.0545	0.0303	0.0301	0.0358	0.0444	0.0244	0.0295
HDR-Hexplane	0.1581	0.1360	0.1219	0.1598	0.1377	0.1443	0.1285	0.1390
Ours	0.0435	0.0287	0.0203	0.0190	0.0260	0.0270	0.0197	0.0256

Methods	big jump	side_walk	one arm swing	two arm swing	jumping jack	pointing walk	sit down up	tube toss
real dataset-Dynamic only								
PSNR								
NSFF	13.79	14.53	15.39	15.69	17.89	14.01	17.02	17.56
NeRF-WT	13.10	12.11	18.64	19.24	18.79	14.42	17.41	17.73
HDR-Hexplane	14.27	15.37	20.63	20.03	19.82	17.19	19.39	18.40
Ours	19.89	22.27	24.60	25.32	23.57	21.79	24.24	20.89
SSIM								
NSFF	0.4580	0.5581	0.6323	0.6304	0.5281	0.4991	0.6861	0.7189
NeRF-WT	0.2591	0.3614	0.7041	0.7617	0.5221	0.4939	0.5923	0.7160
HDR-Hexplane	0.2854	0.4683	0.8260	0.7820	0.5696	0.5752	0.6573	0.7692
Ours	0.6459	0.8001	0.9010	0.9059	0.7720	0.7852	0.8404	0.8194
LPIPS								
NSFF	0.2241	0.2104	0.1267	0.1202	0.1610	0.3262	0.1018	0.0752
NeRF-WT	0.2789	0.2231	0.0770	0.0734	0.1444	0.1829	0.1259	0.0895
HDR-Hexplane	0.2618	0.1766	0.0695	0.0985	0.1371	0.1387	0.1199	0.0750
Ours	0.1810	0.1025	0.0445	0.0385	0.1139	0.0923	0.1079	0.0789

Table S3: **Quantitative results of novel time synthesis on real data.** The green and yellow colors stand for the best and the second best, respectively.

Methods	Lego	Mutant	Standup	Jumping Jack
synthetic dataset-Full				
PSNR				
NSFF	15.45	16.97	13.47	15.53
NeRF-WT	29.55	33.06	32.55	29.25
HDR-Hexplane	28.58	30.88	30.83	29.50
Ours	34.64	36.13	35.80	33.72
SSIM				
NSFF	0.6472	0.6348	0.4958	0.6551
NeRF-WT	0.9595	0.9114	0.9556	0.9200
HDR-Hexplane	0.9443	0.8526	0.9112	0.9137
Ours	0.9670	0.9278	0.9564	0.9348
LPIPS				
NSFF	0.1556	0.1243	0.2368	0.1364
NeRF-WT	0.0171	0.0316	0.0224	0.0655
HDR-Hexplane	0.0257	0.0708	0.0603	0.0539
Ours	0.0147	0.0305	0.0249	0.1229

Methods	Lego	Mutant	Standup	Jumping Jack
synthetic dataset-Dynamic				
PSNR				
NSFF	15.94	18.43	10.25	13.74
NeRF-WT	22.32	27.58	19.77	16.33
HDR-Hexplane	24.61	29.71	21.59	19.57
Ours	28.77	31.80	24.98	23.21
SSIM				
NSFF	0.6145	0.5152	0.1601	0.5795
NeRF-WT	0.8517	0.8289	0.7741	0.5412
HDR-Hexplane	0.8626	0.8443	0.7665	0.7262
Ours	0.9062	0.9115	0.8816	0.8349
LPIPS				
NSFF	0.1528	0.1708	0.3097	0.1345
NeRF-WT	0.0592	0.0845	0.0988	0.1154
HDR-Hexplane	0.1217	0.0724	0.1547	0.0794
Ours	0.0426	0.0590	0.0749	0.0538

Table S4: **Quantitative results of novel view and time synthesis on synthetic dataset.** The green and yellow colors stand for the best and the second best, respectively.

main paper. Moreover, supplementary videos include more HDR, LDR, and novel view rendering results. Please refer supplementary video for further visualization results.

C.1 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 4, the original RAFT model was not trained on multi-exposed images, resulting in high errors when applied directly in our

Methods	Full			Dynamic only		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours w/ RAFT (Teed and Deng, 2020)	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 w/ Dino-Tracker (Tumanyan et al., 2024)	31.01	0.9301	0.0233	22.55	0.7714	0.0697

Table S5: **Ablation study of flow model.** To compare the effect of flow regularization, we compare NVS performance of our approach against the baseline optical flow model (RAFT Teed and Deng (2020)) and a stronger baseline fine-tuned RAFT on a multi-exposure adaptation of the FlyingThings3D dataset.

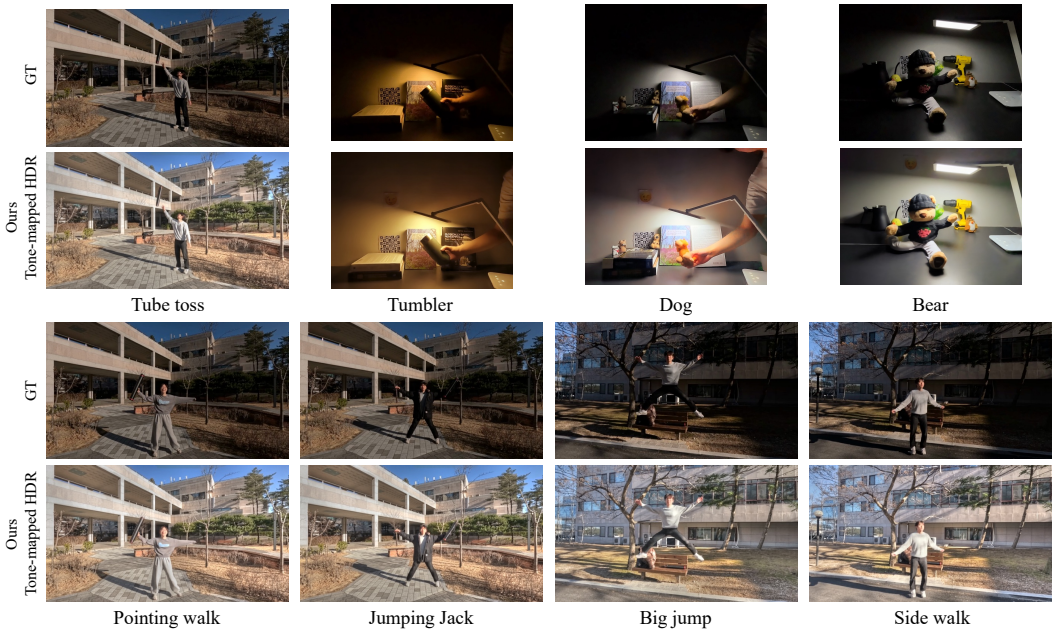


Figure S5: **Qualitative results on GoPro dataset.**Compared with the ground truth LDR views, tone-mapped HDR views reveal the details of over-exposure and under-exposure areas.

setting. By fine-tuning it on synthetic data, the performance is improved. As shown in Table S5, our proposed method achieves the best results.

Use of Large Language Models

A large language mode was used only for minor assistance in writing and improving the clarity of presentation.