

Dual Exposure Stereo for Extended Dynamic Range 3D Imaging

Juhyung Choi
POSTECH

Jinnyeong Kim
POSTECH

Jinwoo Lee
KAIST

Samuel Brucker
Torc Robotics

Mario Bijelic
Princeton University

Felix Heide
Princeton University

Seung-Hwan Baek
POSTECH

In this supplemental document, we provide additional results and details in support of our findings in the main manuscript.

Contents

1. Details on Image Formation	2
1.1. Image Preprocessing	2
1.2. Image Formation	3
2. Experimental Prototype	4
2.1. Device Part List	4
2.2. Image Acquisition Pipeline	5
2.3. Calibration Details	5
2.4. Runtime	5
3. Datasets	6
3.1. Stereo Real Video Dataset	6
3.2. Stereo Synthetic Video Dataset	7
4. Dual-Exposure Depth Estimation	10
4.1. Network Architecture	10
4.2. Training Details	11
5. Additional Results	12
5.1. Additional Evaluation on Real Dataset	12
5.2. Additional Evaluation on Synthetic Dataset	14
5.3. Ablation Study on Exposure Setting	15
5.4. Ablation Study on Stereo Backbone	20
5.5. Ablation Study on Feature Fusion	20
5.6. Ablation Study on Motion Compensation and Fusion Methods	20
5.7. Ablation Study on Single-Frame vs. Dual-Frame Stereo	21
5.8. Ablation Study on Individual Modules	22
6. Additional Discussion	22
6.1. Motion Blur in Dataset Acquisition	22
6.2. Impact of Fast Motion on Depth Accuracy	23
6.3. Challenges with LiDAR Points in Outdoor Scenarios	24
6.4. Initial Exposure Setting	25

1. Details on Image Formation

1.1. Image Preprocessing

We develop a comprehensive image pre-processing pipeline. This section provides a detailed description of the pre-processing steps, including data handling, Bayer to RGB conversion, bilateral filtering, and stereo rectification.

Conversion from Bayer to RGB The raw Bayer images are first converted into 32-bit Bayer patterns, packing three 8-bit channels into a 32-bit representation. This representation is crucial for preserving the full dynamic range of the raw image data. Since the camera stores RAW image data in a custom 24-bit format, standard OpenCV functions cannot be directly applied for Debayering. To address this, we implemented a bilinear interpolation-based Debayering method. This approach reconstructs the red, green, and blue channels by interpolating the Bayer pattern, ensuring minimal color distortion. After interpolation, OpenCV's `cvtColor` function is used to convert the interpolated Bayer image into a standard RGB format.

Bilateral Filtering To reduce grid-like artifacts introduced during Bayer to RGB conversion, bilateral filtering is applied using the OpenCV's `bilateralFilter` function. We used a spatial parameter `sigmaSpace = 20` and color parameter `sigmaColor = 20` to maintain a balance between smoothing and edge retention.

Stereo Image Rectification Accurate stereo rectification is essential for consistent disparity calculation. Using calibration data, we rectified the left and right images to align their epipolar lines. The calibration data includes intrinsic matrices, distortion coefficients, rotation, and translation parameters. Stereo rectification was performed using OpenCV's `stereoRectify` and `initUndistortRectifyMap` functions. During rectification, the `alpha` parameter was set to 0, ensuring no blank regions were left in the rectified images by cropping areas outside the valid region. This approach produces rectified images suitable for disparity estimation with minimized distortions and artifacts.

Pipeline Overview The pre-processing pipeline combines raw data loading, Bayer to RGB conversion, bilateral filtering, stereo rectification, and tensor conversion. These steps collectively enhance image quality and geometric consistency, enabling accurate and robust disparity estimation in subsequent stages of the pipeline. A diagram summarizing the pipeline is presented in Figure 1.

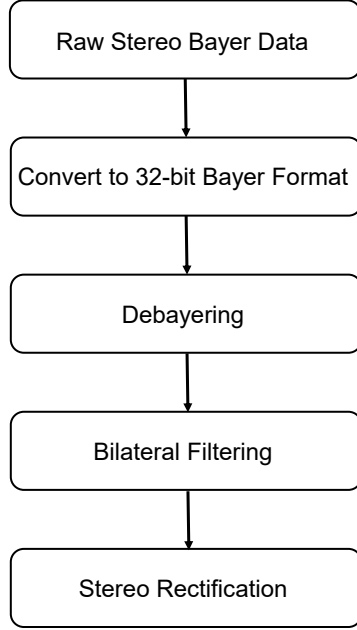


Figure 1. **Image Pre-Processing Pipeline.** The pipeline includes (1) loading raw Bayer data, (2) converting 24-bit raw Bayer patterns to 32-bit Bayer format, (3) performing debayering with custom bilinear interpolation and OpenCV color conversion, (4) applying bilateral filtering to reduce grid-like artifact, (5) rectifying stereo images using calibration parameters. This pipeline ensures high-quality and geometrically consistent inputs for disparity estimation.

1.2. Image Formation

To simulate dual-exposure stereo image captures, we model the image formation process using exposure settings and the incident scene radiance. This process is critical for accurately simulating the captured intensity values generated under various exposures. The procedure is formalized in Equation (2) and implemented in our pipeline.

Exposure Modeling We denote the exposure for each frame as e_i , where $i \in \{1, 2\}$ alternates for consecutive frames. The exposure is converted to shutter time t_i and gain g_i as follows:

$$t_i = \frac{e_i}{g_i}, \quad g_i = \max(1, \frac{e_i}{t_{\max}}), \quad (1)$$

where t_{\max} is the maximum allowable shutter time. This formulation ensures the longest possible shutter time is used to minimize noise, while higher gains compensate for cases where $e_i > t_{\max}$.

Noise Modeling and Clipping Given the incident scene radiance Φ_i , the intensity captured by camera $c \in \{\text{left}, \text{right}\}$ at pixel p_i^c is modeled as:

$$I_i^c(p_i^c) = \text{quant} \left(\text{clip} \left(g_i (\Phi_i t_i + n_i^{\text{pre}}) + n_i^{\text{post}} \right) \right), \quad (2)$$

where n_i^{pre} and n_i^{post} are pre-gain and post-gain noise terms, respectively, sampled from zero-mean Gaussian distributions:

$$n_i^{\text{pre}} \sim \mathcal{N}(0, \sigma_{\text{pre}}), \quad n_i^{\text{post}} \sim \mathcal{N}(0, \sigma_{\text{post}}).$$

The $\text{clip}(\cdot)$ function limits the intensity values within the dynamic range of the camera, defined by the bounds $[\Phi_{\text{lower}}, \Phi_{\text{upper}}]$, and $\text{quant}(\cdot)$ quantizes the intensity values to discrete levels.

Dynamic Range Clipping The simulation pipeline begins by applying the exposure settings to the reference scene radiance Φ_i , followed by noise modeling and dynamic range clipping. This process ensures that the simulated captured intensity values are consistent with the physical limitations of a camera’s dynamic range.

- **Dynamic Range Initialization.** Given the scene radiance Φ_i , the dynamic range bounds are computed based on its distribution. The midpoint of the radiance, Φ_{middle} , is defined as:

$$\Phi_{\text{middle}} = \frac{\max(\Phi_i)}{2}.$$

To determine the span of the dynamic range, we calculate an interval:

$$\text{interval} = \Phi_{\text{middle}} \cdot \frac{\text{range} - 1}{\text{range} + 1},$$

where $\text{range} = 8$ is a predefined parameter. The lower and upper bounds for the dynamic range are expressed as:

$$\Phi_{\text{lower}} = \Phi_{\text{middle}} - \text{interval}, \quad \Phi_{\text{upper}} = \Phi_{\text{middle}} + \text{interval}.$$

- **Dynamic Range Clipping.** The radiance values after exposure modeling and noise addition are clipped within the defined bounds:

$$\Phi_{\text{lower}} \leq g_i(\Phi_i t_i + n_i^{\text{pre}}) + n_i^{\text{post}} \leq \Phi_{\text{upper}}.$$

Captured Intensity Simulation To normalize the captured intensity values to the camera’s range $[0, 1]$, the clipped intensity is processed as:

$$I_{i,\text{captured}}^c(p_i^c) = \frac{I_i^c(p_i^c) - \min(I_i^c(p_i^c))}{\max(I_i^c(p_i^c)) - \min(I_i^c(p_i^c))}.$$

Quantization Quantization is a critical step in the image formation model, simulating the limited bit depth of real-world cameras by mapping scene radiance values to discrete intensity levels. This process involves scaling, clamping, and rounding intensity values to match the resolution of the target camera system, typically 8 bits. To achieve this, we use a quantization function defined as:

$$\text{quant}(x) = \frac{\text{round}(\text{clip}(x \cdot (2^8 - 1)))}{2^8 - 1}. \quad (3)$$

Here, the $\text{clip}(\cdot)$ operation restricts the intensity values to the valid dynamic range, and the $\text{round}(\cdot)$ operation maps the scaled values to the nearest discrete level. This approach ensures that the simulated intensity values align with the physical constraints of stereo cameras while maintaining compatibility with captured image formats.

Straight-Through Estimator (STE) for Backpropagation To preserve gradient flow during training, the Straight-Through Estimator (STE) framework is employed for the quantization step. STE approximates the quantization operation as an identity function during the backward pass, effectively bypassing its non-differentiable nature. The gradient of the quantized intensity I_i^{quant} with respect to the scaled intensity I_i^{scaled} is expressed as:

$$\frac{\partial I_i^{\text{quant}}}{\partial I_i^{\text{scaled}}} \approx 1. \quad (4)$$

This approximation ensures that the quantization operation does not hinder the optimization process, allowing seamless end-to-end training of the model. By combining quantization with STE, the image formation pipeline effectively replicates the behavior of real-world cameras while remaining fully differentiable.

2. Experimental Prototype

2.1. Device Part List

Our imaging system consists of a stereo camera, a mobile robot platform, a PC and a 3D LiDAR sensor. The components are selected and configured to ensure synchronized data capture and geometric consistency across diverse environments:

Item #	Part description	Quantity	Model name
1	RGB Camera	2	LUCID Triton TRI054S-CC
2	Objective lens	2	Edmund Optics #33-307
3	Mobile Platform	1	AgileX Ranger-Mini 2.0
4	LiDAR	1	Ouster OS-1 128
5	PC	1	ASUS Rog Zephyrus G14

Table 1. **Part list of out imaging system.**

- **Stereo Cameras:** Two LUCID Triton 5.4MP cameras (TRI054S-CC) capture 24-bit linear RAW Bayer color images. The cameras are connected via Ethernet and synchronized using the Precise Time Protocol (PTP), achieving sub-millisecond shutter synchronization. For exposure setting at 10 ms with a gain of 1.0, this configuration achieves up to 120 dB of dynamic range in daytime scenes.
- **Mobile Platform:** To capture images in diverse real-world environments, we employed the AgileX Ranger-Mini 2.0, a robust four-wheel robot capable of traversing challenging terrains, including urban streets, pedestrian walkways, and indoor environments. Our mobile platform operates at approximately 6 km/h.
- **LiDAR Sensor:** The Ouster OS-1 3D LiDAR sensor provides geometric data with 128 vertical beams, a maximum detection range of 200 meters, and up to 2048 samples per rotation at 20 Hz. The LiDAR’s output resolution reaches 2048×128 , offering precise depth data. The LiDAR is aligned with the left stereo camera to generate sparse depth maps for the left camera view.

2.2. Image Acquisition Pipeline

Our system is designed to capture synchronized stereo and LiDAR data in real-time. The acquisition process is split into two parallel loops:

1. **Stereo Image Capture:** The stereo cameras operate at a fixed frame rate of 5 FPS, capturing synchronized frames as 24-bit HDR images saved in .npy format. The cameras are triggered simultaneously at the start of each sequence, ensuring precise temporal alignment.
2. **LiDAR Data Capture:** The LiDAR sensor scans the environment continuously, sending acknowledgments (ACKs) for each frame. If a corresponding stereo frame is captured within 50 milliseconds of the LiDAR frame, the system associates the two, creating a single synchronized data frame.

This pipeline ensures that stereo intensity data and LiDAR measurements are aligned, enabling robust integration for depth estimation and scene analysis.

2.3. Calibration Details

Geometric Calibration Geometric calibration is performed to align the stereo camera and LiDAR sensor. The calibration parameters include:

- **Stereo Cameras:** Intrinsic matrices (focal length, principal point), distortion coefficients, and extrinsic parameters (rotation and translation) are computed using a checkerboard pattern with OpenCV.
- **LiDAR-Camera Alignment:** The extrinsic transformation matrix between the LiDAR and the left camera is calculated to project LiDAR points onto the left camera’s image plane, using the camera’s intrinsic matrix.

Radiometric Calibration To ensure consistent intensity measurements across stereo images, the camera settings (exposure and gain) are fixed, and intensity normalization is applied to compensate for sensor sensitivity differences. This step is critical for maintaining accurate depth alignment between the stereo cameras and LiDAR.

Calibration Dataset The calibration process uses 50 checkerboard images captured across various distances and angles to optimize the stereo rectification and LiDAR alignment. Reprojection error analysis confirms the geometric accuracy of the calibration parameters.

2.4. Runtime

The FPS values reported in Table 1 of the main paper correspond to the runtime of the exposure control module within each method’s pipeline. These values do not include the stereo matching process, which runs separately. Our stereo matching

pipeline operates at 1 FPS on an Nvidia RTX 3090 when executed offline. However, real-time performance can be further improved through on-device optimizations, such as quantization and other network acceleration techniques.

3. Datasets

3.1. Stereo Real Video Dataset

The dataset was captured using our stereo camera system described in main paper Section3, equipped with two LUCID Triton 5.4MP cameras for synchronized stereo imaging. Each stereo frame is accompanied by corresponding LiDAR ground truth data captured using an Ouster OS-1 3D LiDAR sensor. Figure 2 illustrates sample stereo image pairs from the dataset, along with their corresponding ground truth LiDAR points projected onto the left camera view. The stereo images showcase the variety of environments and lighting conditions present in the dataset. The LiDAR ground truth highlights the sparse yet accurate depth information used for evaluation. Some ground areas in the LiDAR data lack points due to wet surfaces after rainfall, which interferes with LiDAR capture.

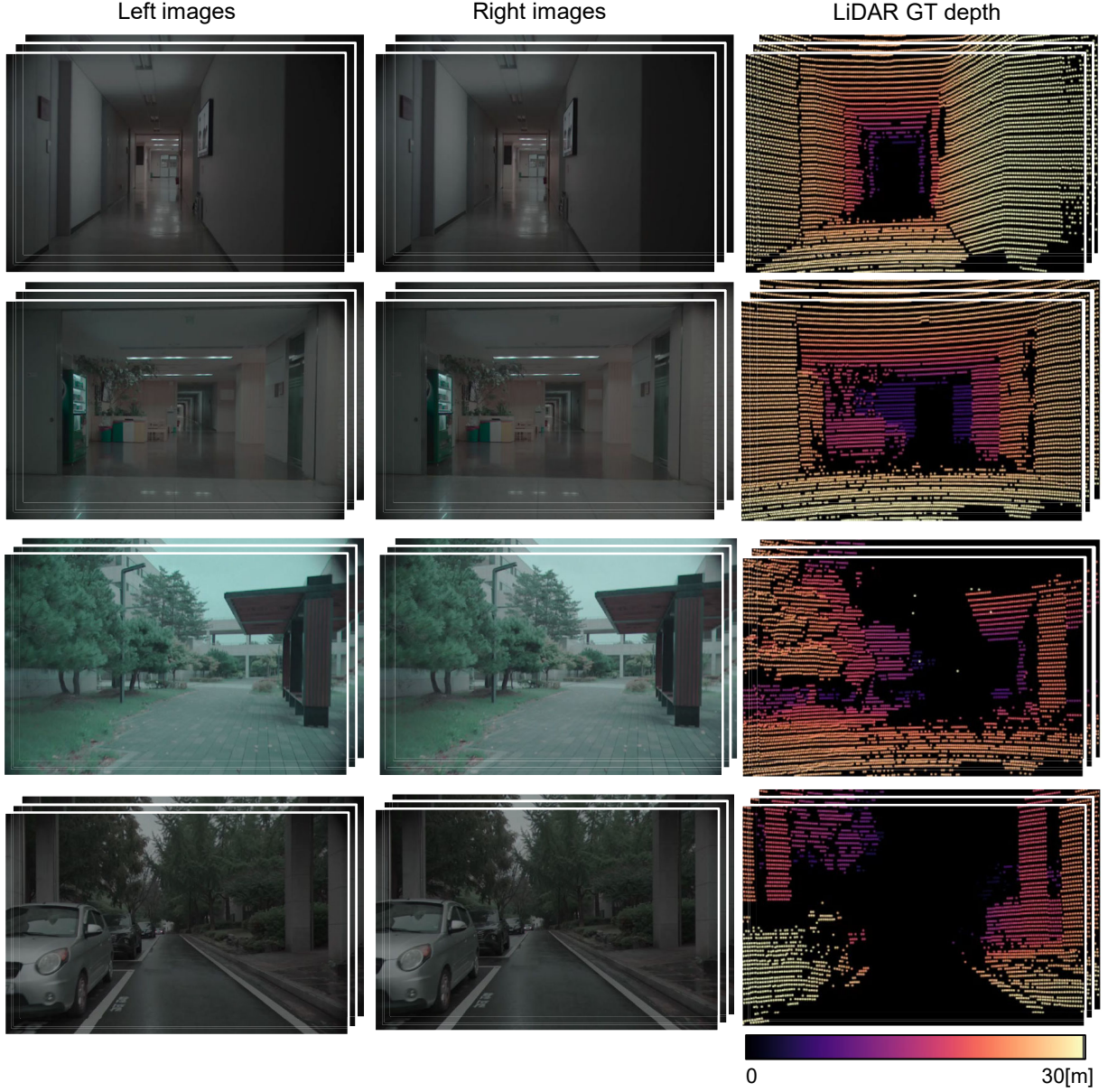


Figure 2. **Visualization of Real Dataset Samples.** Examples of dual-exposure stereo images and their corresponding LiDAR ground-truth depth maps from the captured real-world dataset. The top two rows represent indoor scenes, while the bottom two rows represent outdoor scenes. The LiDAR GT depth maps demonstrate the variability in point density and accuracy across different environments.

3.2. Stereo Synthetic Video Dataset

We use the CARLA driving simulator [2] to generate a synthetic video dataset that supports training and testing of dual-exposure stereo depth estimation in diverse automotive scenarios. Our synthetic dataset is specifically configured to capture extreme lighting scenes to simulate real-world dynamic range challenges. To simulate stereo imaging, we configured the CARLA environment with virtual side-by-side mounted RGB-D cameras to capture synchronized stereo image pairs at 1280×384 resolution. Each virtual RGB camera captures full 32-bit stereo images using multi-exposure imaging [8], while the depth camera generates a dense ground truth depth map for each frame. Hereby, the setup generates ground-truth depth maps, ground-truth disparity maps, and stereo calibration data alongside stereo images, enabling the creation of a comprehensive dataset with precise geometry and calibration details consistently across diverse driving scenarios.

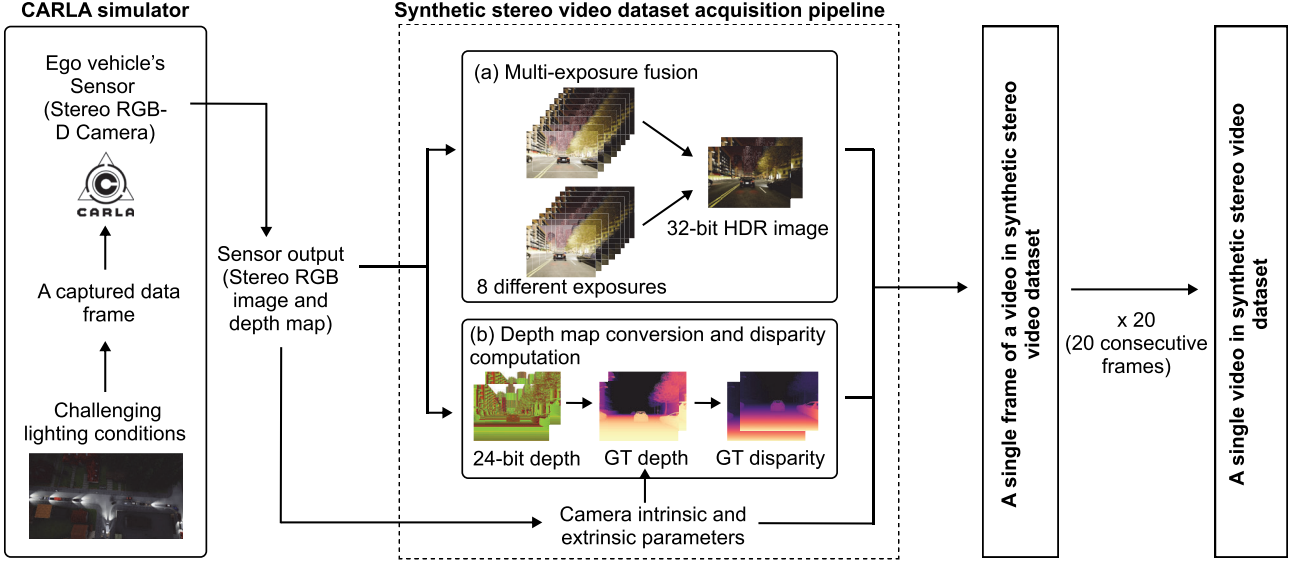


Figure 3. **Overview of the synthetic stereo video dataset acquisition pipeline.** With specific environmental settings on CARLA simulator to generate adverse lighting conditions, the equipped RGB cameras capture images to reconstruct HDR scenes and the depth cameras create corresponding ground truth disparity maps.

Dataset Acquisition Pipeline Figure 3 shows an overview of our stereo synthetic video dataset acquisition pipeline. In our CARLA simulator ego-vehicle’s capture setup, stereo RGB cameras and paired depth cameras—each with a resolution of 1280×384 pixels, a horizontal field of view of 75 degrees, and a fixed frame rate of 10 FPS—are mounted side by side on the bonnet of the test vehicle, with a baseline of 0.4 m. Refer to Table 2 for the sensor configuration details. Since the CARLA itself does not offer real-time HDR rendering, a primary process is required to reconstruct HDR images from the rendered RGB images. In each time frame, stereo RGB cameras capture images with eight different exposure times $t \in \{\frac{1}{500n}(\text{sec}) \mid n \in \{1, 2, \dots, 8\}\}$ while fixed ISO = 200 and aperture size f/1.4 in day time and dusk time. For night time, on the other hand, exposure times are given as $t \in \{\frac{1}{50n}(\text{sec}) \mid n \in \{1, 2, \dots, 8\}\}$ with fixed ISO = 1,600 and aperture size f/1.4. Here, daytime is defined as the period when the solar altitude satisfies $\alpha \geq 3^\circ$, and dusk is defined as the period when the solar altitude satisfies $-3^\circ \leq \alpha < 3^\circ$. Night time is the complement of these periods, corresponding to the range where $\alpha < -3^\circ$. Then, with multi-exposure HDR reconstruction [8], we obtain 32-bit stereo images for each frame of the scenario. Note that there are no motion artifacts between multi-exposure frames within a single time step of a dynamic automotive scene, as all RGB images are synchronously captured by virtual RGB cameras in the CARLA simulator. This eliminates the risk of failure in exposure bracketing-based HDR reconstruction for dynamic scenes, which would otherwise require addressing using various de-ghosting approaches [5, 9, 10]. Meanwhile, the depth camera captures the ground truth depth map up to 1,000m. As the CARLA provides depth information with 24-bit floating-point precision encoded across the three channels of the RGB color space, it is decoded to reconstruct the plain depth map in meters. We also compute disparity map from ground-truth depth map using stereo calibration parameters, here by acquiring the ground-truth value for disparity. In specific, given a pair of rectified stereo depth maps with depth z_s , focal length f and baseline B , the disparity d in the corresponding pixel is calculated using $f \frac{B}{z_s}$. As a result, pairs of 32-bit stereo RGB images, depth maps, disparity maps, and stereo calibration parameters (both intrinsic and extrinsic) compose a single frame of a video in the stereo synthetic video dataset, see Table 3. Additionally, by leveraging CARLA’s support for simulating diverse driving environments, both training and testing videos are selectively retrieved from the simulation, introducing abrupt changes in dynamic range and thereby reflecting real-world dynamic range challenges.

Dataset Details and Statistics Our synthetic dataset comprises 1,000 training videos and 200 testing videos, each designed to introduce dynamic range challenges across various driving conditions. Training scenarios comprise 20 consecutive stereo frames, and test scenarios contain 100 consecutive stereo frames. Scenarios represent a wide range of lighting conditions (day, dusk, and night), with distributions of approximately 50% at night, 30% during the day, and 20% at dusk. The driving locations include urban and suburban areas, rural areas, and highways. While the dataset includes extreme weather conditions

Sensor Type	Sensor Count	Output Shape	Configuration
RGB camera	8	$\mathbb{R}^{3 \times 384 \times 1280}$	Left, ISO = 200(Day, Dusk) / 1600(Night), f/1.4, FOV = 75°
RGB camera	8	$\mathbb{R}^{3 \times 384 \times 1280}$	Right, ISO = 200(Day, Dusk) / 1600(Night), f/1.4, FOV = 75°
Depth sensor	1	$\mathbb{R}^{1 \times 384 \times 1280}$	Left, FOV = 75°
Depth sensor	1	$\mathbb{R}^{1 \times 384 \times 1280}$	Right, FOV = 75°

Table 2. **List of sensors used for CARLA simulator ego-vehicle’s capture setup.** Here, ISO is configured based on the temporal condition. Four categories of sensors are mounted at $x = 2.5$ m, $y = \pm 0.2$ m, $z = 1.4$ m with respect to the ego-vehicle’s centroid. Note that the given coordinates follow the left-handed coordinate system in Unreal Engine 4.

Modality	Shape	Description
HDR image	$\mathbb{R}^{3 \times 384 \times 1280}$	A pair of left and right, 32-bit float
Depth map	$\mathbb{R}^{1 \times 384 \times 1280}$	A pair of left and right, up to 1,000m
Disparity map	$\mathbb{R}^{1 \times 384 \times 1280}$	A pair of left and right, computed from depth map
Intrinsic camera parameters	$\mathbb{R}^{3 \times 3}$	Shared between the left and right views
Extrinsic camera parameters	$\mathbb{R}^{4 \times 4}$	A pair of left and right

Table 3. **Dataset composition for a single frame in a stereo synthetic video dataset.** Each modality, its dimensions, and additional details are outlined.

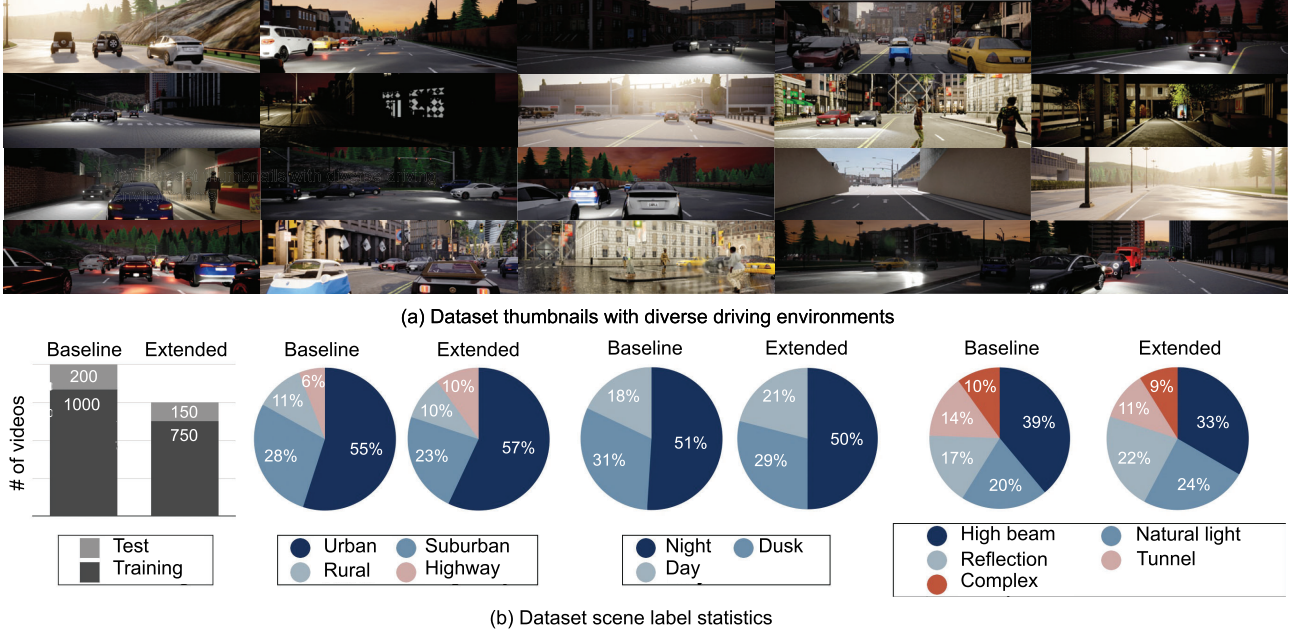


Figure 4. **Synthetic stereo video datasets** (a) Sample tone-mapped thumbnails with diverse driving environments. (b) Dataset scene label statistics for two versions of datasets.

such as rainy scenarios, it does not contain snow or foggy conditions. Each video presents challenging lighting conditions induced by various environmental factors, categorized into four major types: the vehicle’s headlights at night, intense reflections from highly reflective surfaces (such as ponds), intense natural lighting, and light passing through tunnels. Figure 4 shows the dataset thumbnails and scene statistics of the synthetic dataset, which includes diverse driving environments.

Extended Dataset for 3D Object Detection To extend the baseline dataset for the vision task of object detection, we introduce an additional dataset that facilitates both 2D and 3D object detection. Extended dataset consists of 750 driving videos for training and 150 videos for testing, both adhering to the same specifications and scene diversities as the baseline dataset, with the addition of two new modalities: (1) LiDAR point clouds, (2) per-frame object detection data annotations. The virtual LiDAR system is configured to replicate the characteristics of a Velodyne HDL-64E (64 channels, 10Hz revolution

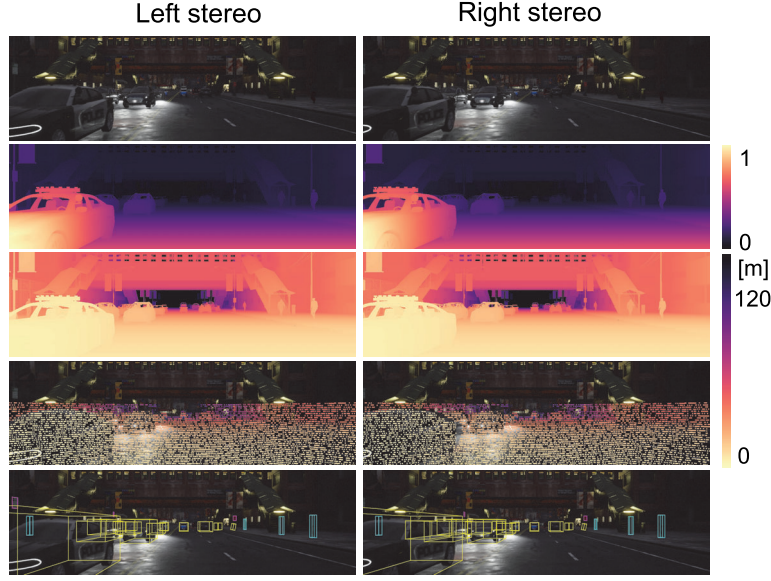


Figure 5. **Extended synthetic stereo video datasets.** The bottom two rows depict LiDAR point clouds and 3D bounding boxes projected onto the tone-mapped stereo HDR images. To ensure consistency across different modalities, both the depth and LiDAR maps share the same color bar, while the disparity map is shown after normalization.

frequency, from -24.8° to $+2^\circ$ vertical field of view and 120m maximum range) and mounted along the optical axis of the left RGB-D camera. The cameras' exposures are triggered only when the LiDAR has completed its rotation and is aligned with the optical axis of the left camera, ensuring precise cross-modality alignment between the LiDAR and the stereo RGB-D cameras. For each object within the left camera's field of view, we automatically annotate it using 3D bounding boxes in a format simplified from the KITTI [3] object detection labels. The annotations include fields for class names, truncation, occlusion state, and bounding box coordinates, all represented in the reference camera's coordinate system. Specifically, we provide annotations for three object classes: 'Vehicle', 'Pedestrian', 'Traffic Signal'. Figure 5 presents the composition of the extended dataset, highlighting the time-varying lighting conditions that contribute to challenging lighting conditions for both depth estimation and object detection.

4. Dual-Exposure Depth Estimation

4.1. Network Architecture

Our dual-exposure depth estimation model extends the RAFT-Stereo framework [4] by incorporating modules for dual-exposure feature fusion and inter-frame motion compensation. These additions enable the network to effectively utilize exposure-specific features from dual-exposure stereo inputs, enhancing disparity estimation under high dynamic range conditions.

Network Overview The architecture consists of three primary stages: 1) Optical Flow Estimation, 2) Dual-Exposure Feature Fusion, and 3) Stereo Depth Estimation. These components are seamlessly integrated into the RAFT-Stereo backbone. While the backbone's original disparity estimation modules remain unchanged, modifications were made to handle dual-exposure inputs and inter-frame alignment:

- **Optical Flow Estimation:** We introduce a pretrained optical flow network [6] to estimate motion between consecutive frames for both left and right stereo views. The optical flow enables spatial alignment of the second-frame features to the first-frame features, addressing temporal motion.
- **Dual-Exposure Feature Fusion:** A feature fusion module combines aligned features from dual-exposure stereo frames. This module uses intensity-based weight maps to ensure that well-exposed details from both bright and dark regions are effectively preserved. The fusion is applied at multiple scales to enhance robustness.
- **Stereo Disparity Estimation:** The fused features are passed through the RAFT-Stereo backbone to construct correlation

volumes and refine disparity predictions. While the correlation computation and update block follow the original RAFT-Stereo design, they now operate on fused feature maps containing dual-exposure information.

Summary of Modified Layers Table 4 summarizes the layers and modules where significant modifications were made. Components such as the correlation volume and update block are inherited directly from the RAFT-Stereo framework and are not described in detail here.

Module	Input Size	Output Size	Description
Optical Flow Network	$[B, 3, H, W]$	$[B, 2, H, W]$	Estimates motion for temporal alignment.
Warping Function	$[B, 256, H/4, W/4]$	$[B, 256, H/4, W/4]$	Aligns second-frame features using optical flow.
Dual-Exposure Fusion	$[B, 256, H/4, W/4]$	$[B, 256, H/4, W/4]$	Combines features from dual-exposure frames using intensity-based weights.

Table 4. **Modified Modules in the Proposed Network.** The table summarizes the key components added to the RAFT-Stereo backbone for dual-exposure depth estimation. Input and output sizes are for a batch size of B and image resolution $H \times W$.

Integration with RAFT-Stereo Backbone The proposed modifications are integrated into the RAFT-Stereo backbone while retaining its core functionality. Optical flow is computed between consecutive stereo frames and used to warp second-frame features to the first frame. These warped features are then fused with first-frame features using the dual-exposure feature fusion module. The fused features are passed through the correlation volume computation and update block to estimate the disparity map. This integration ensures that dual-exposure information is effectively utilized while maintaining the robustness of the original RAFT-Stereo design.

4.2. Training Details

Data Augmentation and Exposure Simulation To simulate dual-exposure stereo inputs, we generated random exposure pairs for each training batch using controlled randomization. The exposure values e_1 and e_2 for the two frames were generated as:

$$e_2 = e_1 \cdot \text{rand}(\text{min_gap}, \text{max_gap}),$$

where $e_1, e_2 \in [2^{-2}, 2^2]$ and the gap $\text{rand}(\text{min_gap}, \text{max_gap})$ was sampled uniformly between 0.5 and 3.0. This exposure simulation ensures the model is trained across diverse lighting conditions, reflecting real-world variations in dynamic range.

Loss Function For training, we employed the sequence loss function adopted from the original RAFT-Stereo framework [4]. This loss progressively refines disparity predictions over $N = 32$ iterations, with a decay factor $\gamma = 0.9$. The sequence loss is defined as:

$$\mathcal{L}_{\text{seq}} = \sum_{i=1}^N \gamma^{\frac{15}{N-1}(N-i-1)} \cdot \|\hat{d}_i - d_{\text{gt}}\|_1,$$

where \hat{d}_i represents the predicted disparity at iteration i , and d_{gt} is the ground truth disparity. A validity mask filters out invalid regions and restricts the loss to valid pixels with a maximum disparity threshold of 700. This ensures effective training of disparity refinement while avoiding the impact of large outliers.

Training Configuration The training was conducted on four NVIDIA RTX 3090 GPUs with a batch size of 4. The CARLA synthetic dataset was used to simulate extreme lighting scenarios, providing diverse and challenging conditions for training. To preserve the robustness of the pretrained RAFT-Stereo model on real-world datasets, only the GRU update block was fine-tuned during training. All other layers were frozen to retain their existing weights. This targeted fine-tuning strategy ensured that the network specialized in feature fusion and disparity refinement for dual-exposure inputs, without degrading its performance on real datasets.

5. Additional Results

5.1. Additional Evaluation on Real Dataset

To further validate our method, we conducted evaluations on real-world scenarios featuring dynamic lighting changes. These scenarios include both indoor and outdoor environments, emphasizing the robustness of our approach under challenging illumination conditions. The evaluation comprises four distinct scenarios, consisting of approximately 1000 frames in total. Figure 6 visualizes the results of outdoor scenes, with each row representing a consecutive frame in temporal order, showcasing the effectiveness of our method in handling dynamic lighting changes across time. Similarly, Figure 7 demonstrates the results for indoor scenes, where the images also follow a temporal sequence.

Method	AverageAE [1]	GradientAE [11]	NeuralAE [7]	ADEC (ours)
MAE [m]↓	2.6142	4.1859	<u>2.2869</u>	2.0251

Table 5. **Comparison of MAE across methods.** The table highlights the performance of different methods, showing that our approach (ADEC) achieves the lowest MAE compared to other baselines.

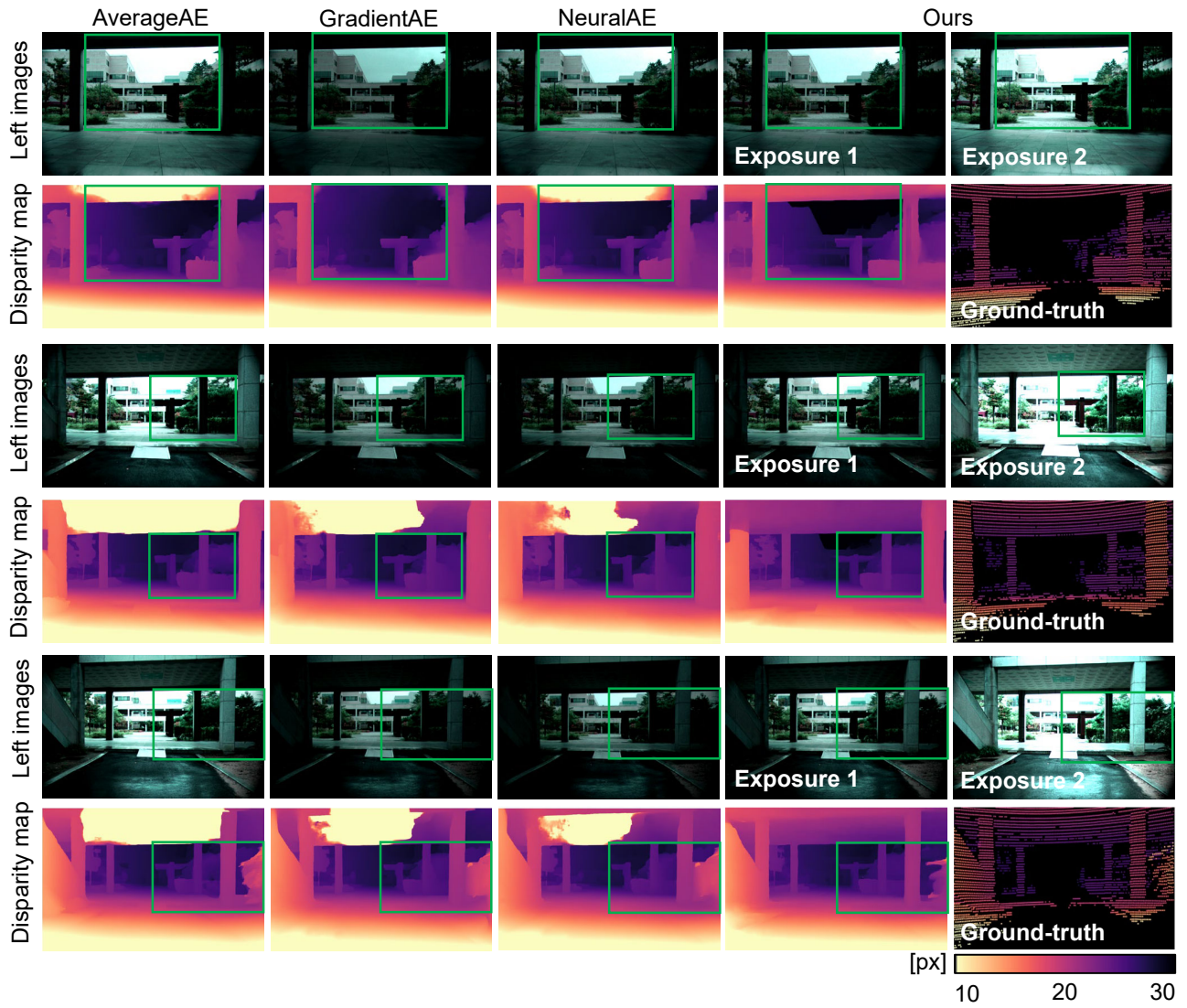


Figure 6. **Disparity-estimation results using our ADEC compared with other AEC methods in outdoor scene** Our ADEC method outperforms the other AEC methods for subsequent extended-DR depth estimation : AverageAE [1], GradientAE [11], NeuralAE [7]

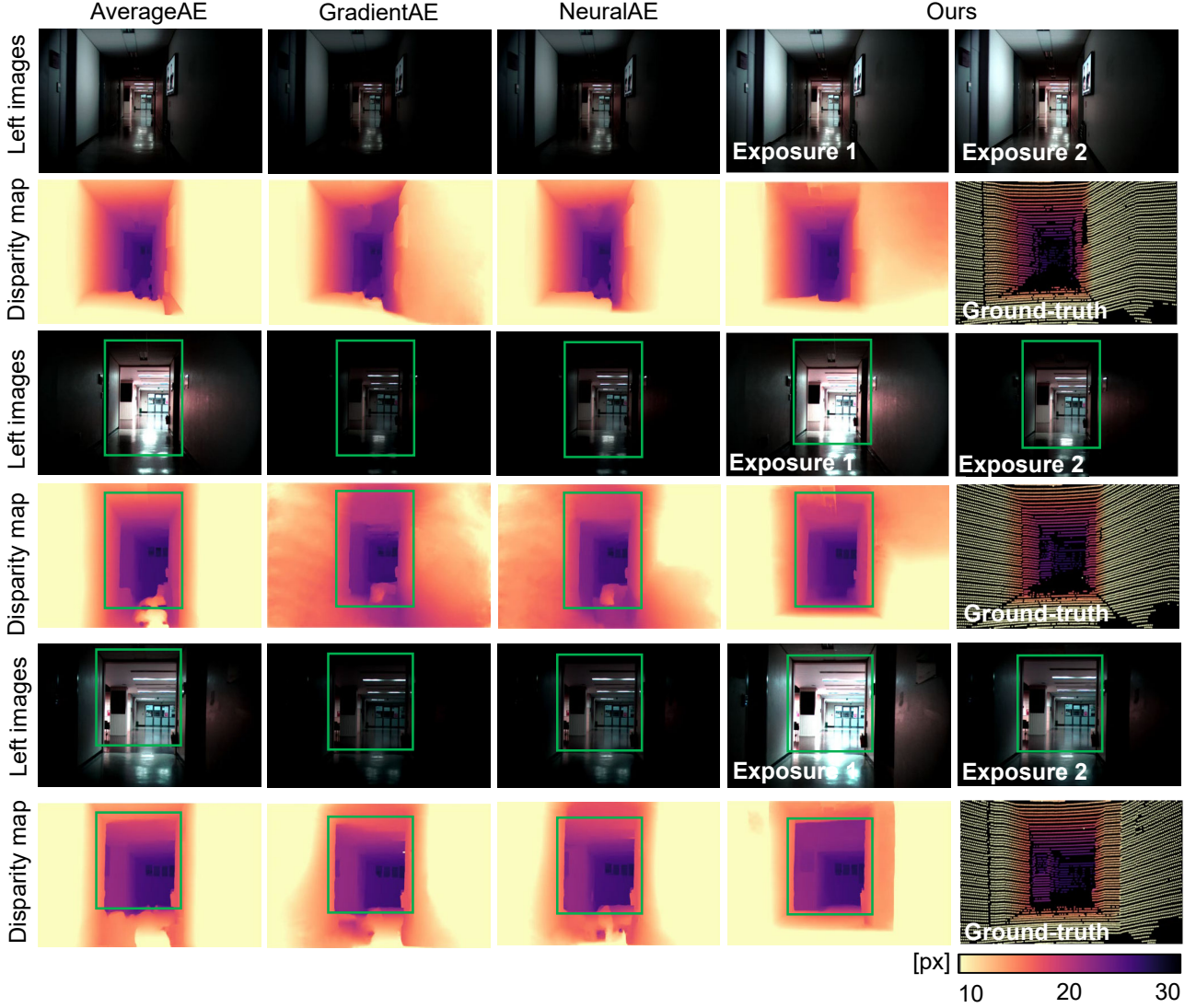


Figure 7. **Disparity-estimation results using our ADEC compared with other AEC methods in indoor scene** Our ADEC method outperforms the other AEC methods for subsequent extended-DR depth estimation : AverageAE [1], GradientAE [11], NeuralAE [7]

5.2. Additional Evaluation on Synthetic Dataset

We conducted an evaluation of our method on the CARLA synthetic dataset to further demonstrate its robustness under various exposure and lighting conditions. The comparison includes other single exposure control methods : AverageAE [1], GradientAE [11], and NeuralAE [7] finetuned using the original RAFT-Stereo framework on the our CARLA synthetic dataset. This ensures a fair comparison between our dual-exposure control approach and existing single-exposure control methods. Figure 8 illustrates qualitative results comparing disparity maps generated by each method. The dataset includes diverse scenarios, such as high-contrast outdoor environments and challenging low-light conditions. For each method, we show the left image input, the predicted disparity map, and the corresponding ground-truth disparity.

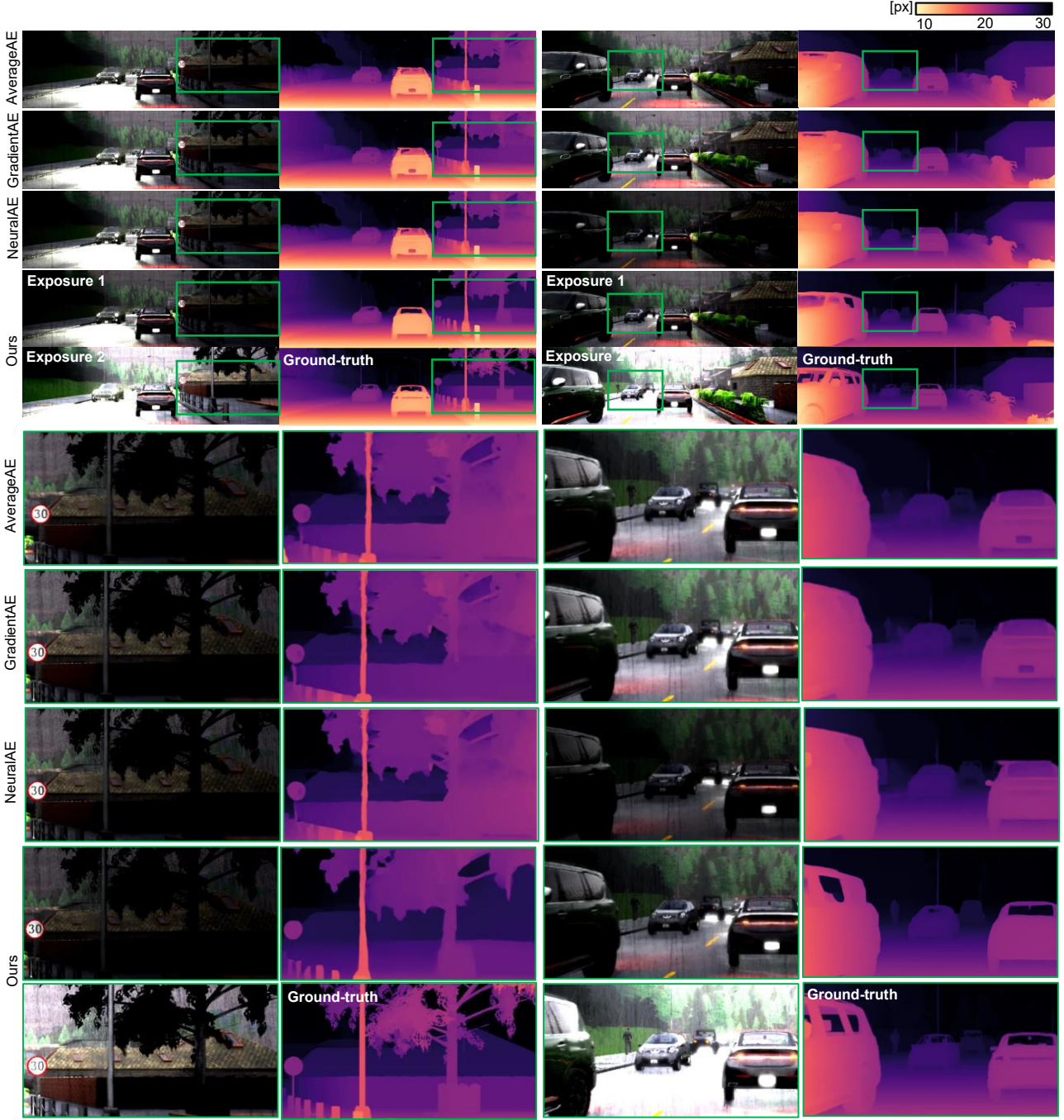


Figure 8. **Disparity-estimation results using our ADEC compared with other AEC methods** Our ADEC method outperforms the other AEC methods for subsequent extended-DR depth estimation : AverageAE [1], GradientAE [11], NeuralAE [7].

5.3. Ablation Study on Exposure Setting

We conducted additional ablation studies focusing on the exposure control module to evaluate its impact on performance. The results are presented both quantitatively and qualitatively through Table 6 and Figures 10, 11, and 12. Each figure visualizes the ablation results by comparing the baseline and ablation models across time steps. For each time step, the visualizations include dual-exposure stereo images, pixel intensity histograms, and disparity maps.

Experiment	Exposure Gap	Exposure Increase Rate	Initial Exposure Values	Disparity MAE [px]↓
Baseline	2.5	Baseline	Equal	2.7452
Ablation 1	1.5	Baseline	Equal	2.8759
Ablation 2	2.5	Reduced	Equal	2.7634
Ablation 3	2.5	Baseline	Unequal	3.2522

Table 6. **Ablation study on exposure control parameters.** The table presents the disparity MAE for different ablation settings, focusing on exposure gap, exposure increase rate, and initial exposure values. The baseline uses an exposure gap of 2.5, baseline increase rate, and equal initial exposure values, achieving the lowest MAE.

Exposure Gap Configuration Figure 9 shows that with an extremely high exposure gap $|e_1 - e_2|$, our optical flow and depth estimation gracefully degrades. This trade-off between exposure gap and depth error is quantitatively analyzed in detail in Figure 7 of the main paper. In Figure 10, we evaluate the effect of different exposure gap configurations. While the baseline model gradually increases the exposure gap, the ablation model fails to widen the gap significantly after a certain point. This results in difficulty capturing sufficient details, particularly in high-contrast regions, compared to the baseline model.

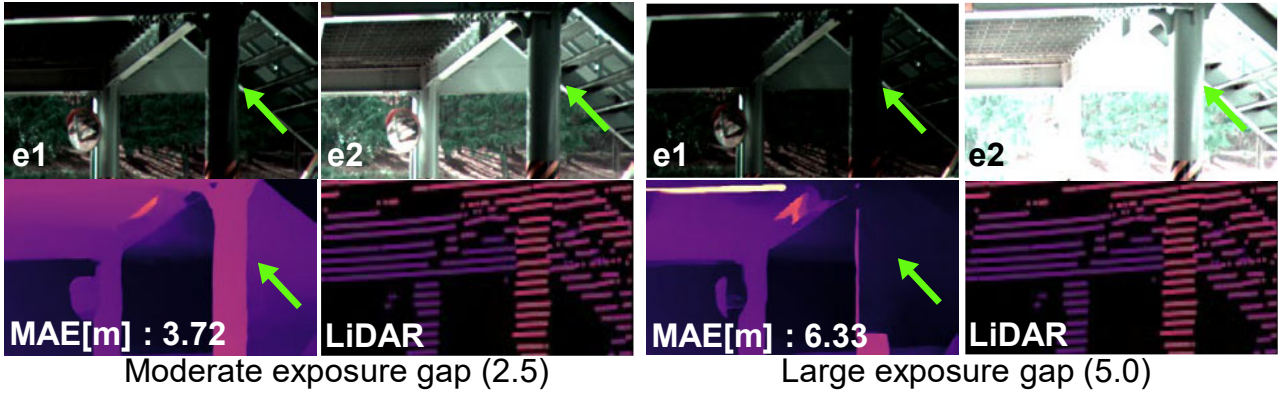


Figure 9. **Impact of exposure difference** $|e_1 - e_2|$. This figure illustrates the exposure gap $|e_1 - e_2|$ and its effect on optical flow and depth estimation. While a properly adjusted exposure gap enhances detail preservation, an excessively large gap results in a gradual decline in performance, especially in high-contrast regions.

Exposure Increase Rate In Figure 11 demonstrates the impact of modifying the scaling factor for determining the next exposure value, referred to as the exposure increase rate. Compared to the baseline model, the ablation model does not achieve a sufficiently large exposure gap in the initial time steps. As a result, the baseline model captures more details in critical regions at earlier time steps, while the ablation model struggles to do so.

Initial Exposure Values In Figure 12, we analyze the effect of setting different initial exposure values for dual-exposure frames. In the baseline model, both exposures start at the same value, while in the ablation model, one frame starts with a higher exposure and the other with a lower one. Although the ablation model benefits from a pre-established exposure gap in the first time step, the baseline model eventually outperforms it by securing more consistent details as the time steps progress.

These results illustrate how variations in exposure gap configuration, exposure increase rate, and initial exposure settings influence the ability of the model to capture and preserve sufficient detail across dynamic scenes. The figures highlight the importance of a well-balanced exposure control strategy for robust disparity estimation.

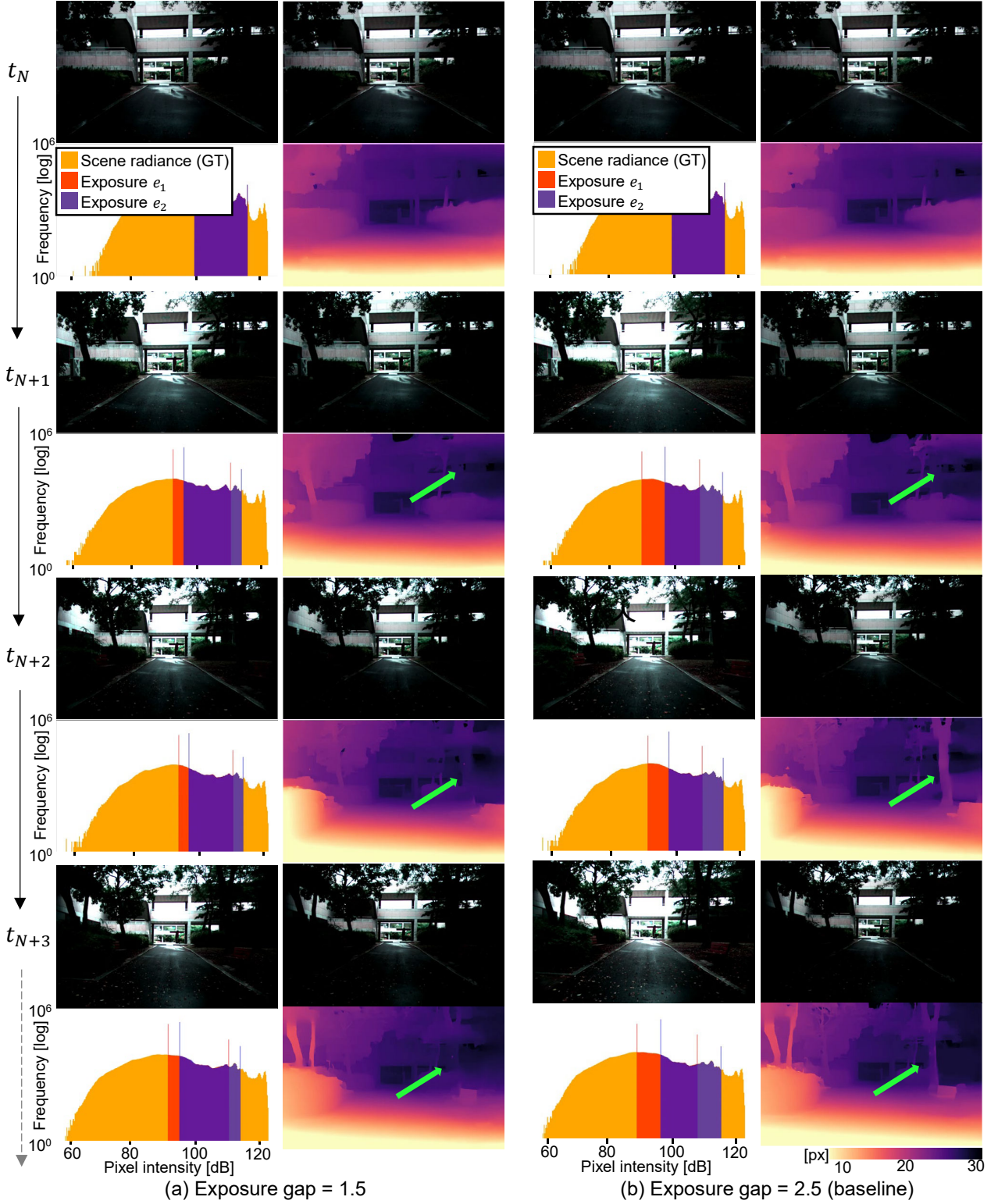


Figure 10. **Impact of exposure gap settings on disparity estimation.** This figure illustrates the effect of varying the exposure gap during dual-exposure control. The baseline model (exposure gap = 2.5) captures sufficient details over time, whereas the ablation model (exposure gap = 1.5) struggles to widen the exposure gap further, resulting in insufficient detail capture in challenging lighting conditions. Each time step showcases the dual-exposure stereo images, pixel intensity histograms, and disparity maps.

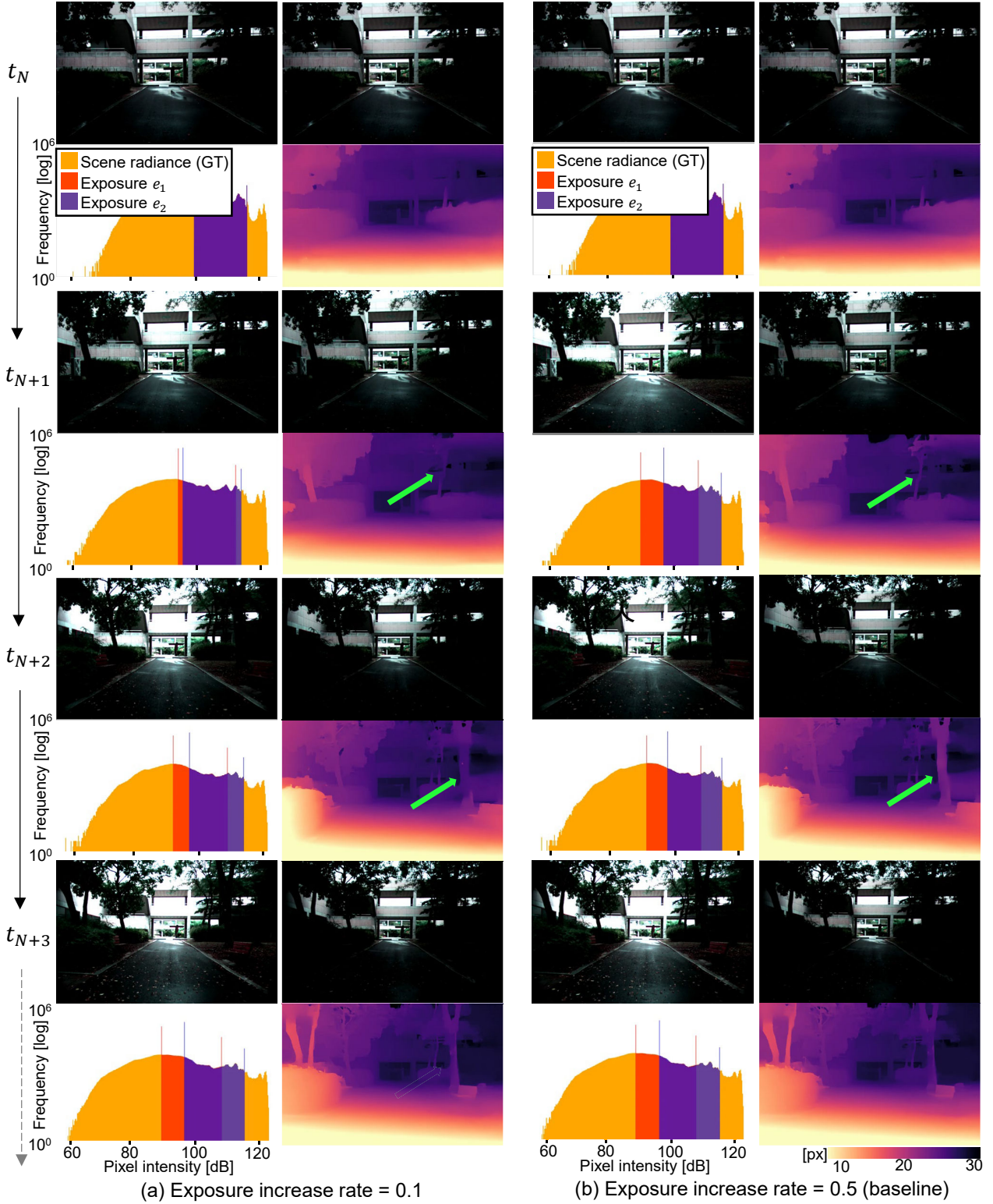


Figure 11. **Impact of exposure increase rate on disparity estimation.** This figure demonstrates the effect of modifying the exposure increase rate during dual-exposure control. The baseline model, with its default increase rate, quickly expands the exposure gap in the initial time steps, enabling effective detail capture. In contrast, the ablation model, with a reduced increase rate, shows slower gap expansion, leading to less effective detail capture in the early time steps. Each time step visualizes the dual-exposure stereo images, pixel intensity histograms, and disparity maps.

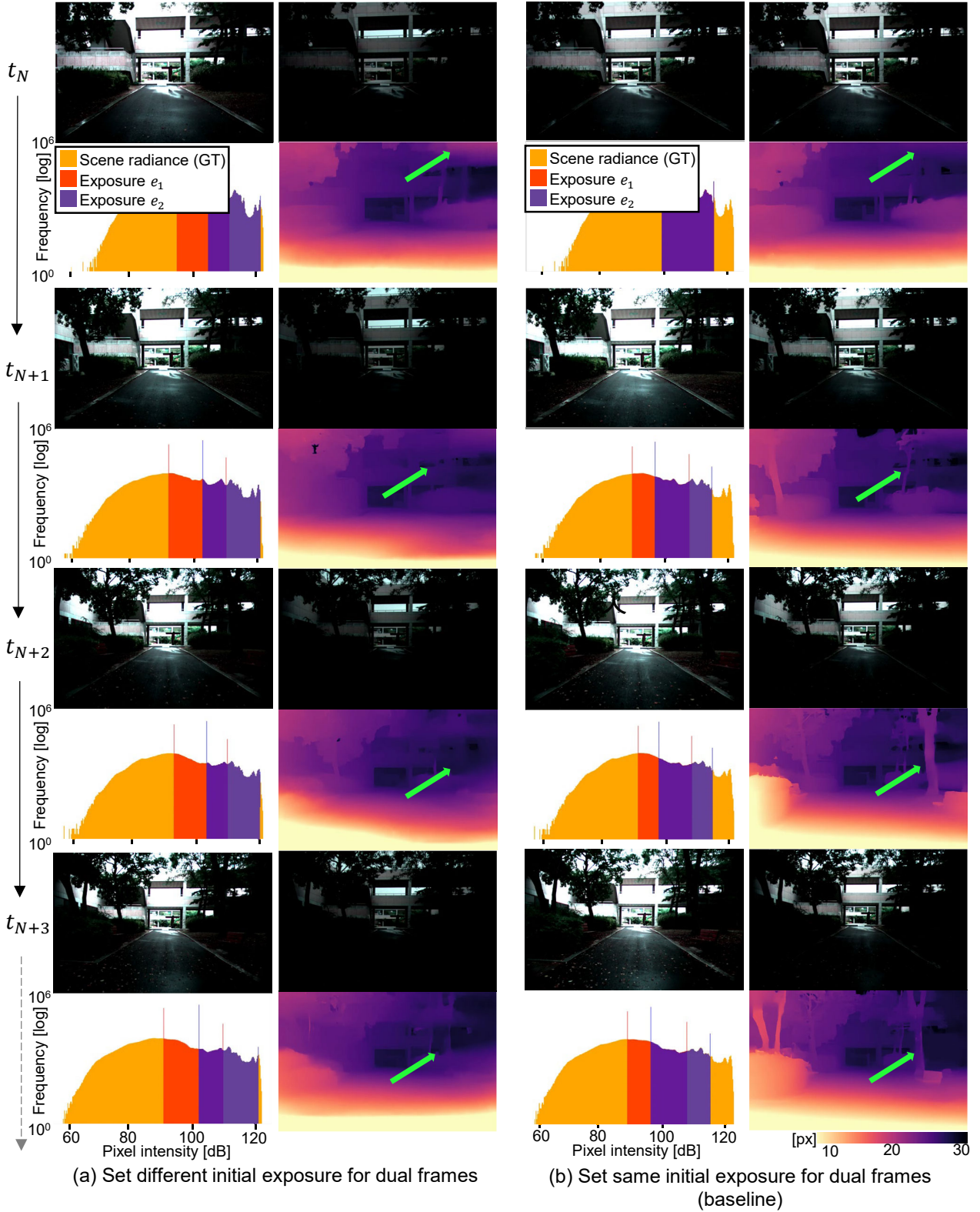


Figure 12. **Impact of initial exposure settings on disparity estimation.** This figure compares baseline and ablation models with different initial exposures. The baseline model uses equal exposures, ensuring consistent detail capture over time. The ablation model starts with unequal exposures, capturing more detail initially but losing balance in later time steps.

Backbone	Average AEC	Gradient AEC	Neural AEC	Our ADEC
RAFT-Stereo	2.5751	2.5236	1.9197	1.7557
ACVNet	2.1754	1.7883	1.6673	<u>1.4621</u>

Table 7. **Comparison of disparity MAE [px] using different stereo backbone networks** The results indicate that replacing RAFT-Stereo with a more recent stereo backbone (ACVNet) leads to improved performance, suggesting potential further gains with future state-of-the-art stereo models.

5.4. Ablation Study on Stereo Backbone

Our method uses RAFT-Stereo [4] as the stereo matching backbone. To validate the generalization capability of our approach, we conducted additional experiments using a more recent stereo matching model, ACVNet [12], as the backbone. We applied exposure module to both RAFT-Stereo and ACVNet to assess its performance across different stereo architectures. For a fair comparison, we maintained the same experimental setup, where the exposure control module and the disparity estimation module were jointly fine-tuned in an end-to-end manner. Table 7 reports the disparity MAE (in pixels) for various AEC methods when using RAFT-Stereo and ACVNet as backbones. As shown in Table 7, switching to a more advanced stereo backbone resulted in improved performance across all AEC methods, with our ADEC achieving the lowest disparity MAE. These results suggest that incorporating more recent stereo backbone networks and further optimizing the pipeline could lead to even greater improvements in performance.

5.5. Ablation Study on Feature Fusion

We conduct an ablation study comparing three different dual-exposure fusion approaches to evaluate their impact on disparity estimation accuracy. (1) Exposure Image Fusion \rightarrow Stereo Matching : The two exposure images are first fused into a single image before applying stereo matching. (2) Separate Disparity Estimation \rightarrow Disparity Fusion : Disparity maps are computed separately for each exposure and then merged. (3) Our Intermediate Feature Fusion : The two exposure images are processed separately up to an intermediate feature level, after which the features are fused before stereo matching. Table 9 presents the results, showing that our intermediate feature fusion approach achieves the lowest disparity error. This highlights the effectiveness of leveraging exposure differences at the feature level rather than at the image or disparity level, allowing the model to extract complementary information more effectively.

Index	Fusion method	Disparity MAE [px] \downarrow
#1	Exposure image fusion	2.6532
#2	Disparity late fusion	3.7485
#3	Our intermediate feature fusion	1.8179

Table 8. **Comparison of different fusion methods.** Disparity MAE [px] comparison of different dual-exposure fusion methods. Our intermediate feature fusion achieves the lowest error, demonstrating the advantage of fusing exposure information at the feature level rather than at the image or disparity level.

5.6. Ablation Study on Motion Compensation and Fusion Methods

To evaluate the impact of motion compensation and fusion strategies, we conduct an ablation study comparing different methods for disparity estimation. In this study, we analyze the effects of using single-exposure disparity estimation, image-based exposure fusion, and motion-compensated late fusion, as shown in Figure 13.

- **Single Exposure (e_1, e_2):** Disparity maps computed using a single stereo pair with a fixed exposure (e_1 or e_2).
- **Exposure Fusion (w/ MC, w/o MC):** Disparity maps obtained from images fused at the image level before stereo matching, with and without motion compensation (MC).
- **Late Fusion (w/ MC, w/o MC):** Disparity maps computed separately for each exposure and fused at the disparity level, with and without motion compensation.
- **Ours:** Our proposed dual-exposure method that integrates exposure-aware motion compensation and feature-level fusion.
- **GT:** Ground-truth disparity map.

Table 9 confirms that removing motion compensation increases disparity error, demonstrating the necessity of motion handling in dual-exposure stereo. Additionally, the disparity errors from single-exposure estimations ($e1$, $e2$) further validate the effectiveness of our dual-exposure approach, which better preserves scene details across varying lighting conditions.

Method	Motion compensation	Disparity MAE [px] ↓
Single exposure $e1$	X	2.6482
Single exposure $e2$	X	3.2221
Exposure fusion (w/ MC)	O	2.6532
Exposure fusion (w/o MC)	X	3.2173
Late Fusion (w/ MC)	O	3.7485
Late Fusion (w/o MC)	X	5.0340
Our intermediate fusion	O	1.82

Table 9. **Comparison of different fusion methods.** Disparity MAE [px] comparison of different dual-exposure fusion methods. Our intermediate feature fusion achieves the lowest error, demonstrating the advantage of fusing exposure information at the feature level rather than at the image or disparity level.

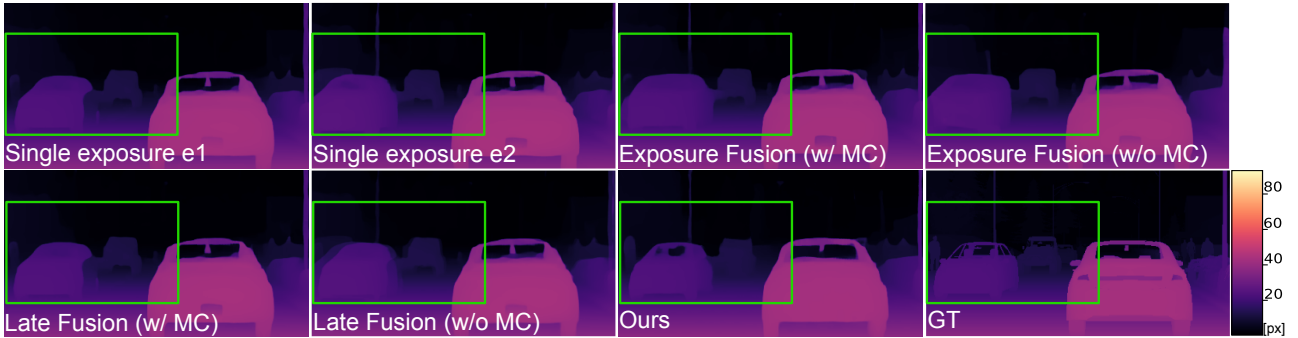


Figure 13. **Comparison of different fusion methods and motion compensation strategies.** Disparity maps generated using single-exposure images, exposure fusion, late fusion, and our method. Motion compensation (MC) significantly reduces errors in fusion-based disparity estimation.

5.7. Ablation Study on Single-Frame vs. Dual-Frame Stereo

To analyze the impact of using dual-stereo frames, we conduct an ablation study comparing single-frame stereo and dual-frame stereo without applying the exposure control module (i.e., without exposure variations). As shown in Table 10, the disparity MAE is 2.49 px for single-frame stereo and 2.21 px for dual-frame stereo. This result demonstrates that dual-frame stereo disparity estimation achieves better accuracy than single-frame stereo, even without adaptive exposure adjustments. The performance improvement indicates that utilizing two stereo pairs helps capture richer scene details, leading to more robust depth estimation.

Stereo Method	Disparity MAE [px] ↓
Single-Frame Stereo	2.49
Dual-Frame Stereo	2.21

Table 10. **Comparison of disparity MAE between single-frame and dual-frame stereo.** We compare the disparity estimation accuracy of single-frame stereo and dual-frame stereo without exposure control. The results show that dual-frame stereo achieves a lower mean absolute error (MAE) of 2.21 pixels compared to 2.49 pixels for single-frame stereo. This suggests that using two consecutive stereo frames helps capture richer scene details, improving disparity estimation even without adaptive exposure adjustments.

5.8. Ablation Study on Individual Modules

We evaluate the importance of the three core components: ADEC module, weighted feature fusion, and motion compensation. Table 11 and Figure 14 show results. First, instead of using our ADEC method, we fix the dual exposure to low and high values respectively using the average scene statistics. This results in the failure of adaptation to varying scene DR, leading to high disparity error. Second, we exclude the weighted fusion in our dual-exposure disparity estimation: we set the weight maps to be one for all pixels: $W_i^c = 1$. The resulting fused features are affected by unstable features from under- or over-exposed features, leading to disparity error. Third, we omit the motion compensation: the optical flow is set to be zero in our depth estimation process. This results in significant misalignment errors in the fused feature, making the disparity accuracy low. Our complete method enables highest accuracy.

ADEC	Weighted fusion	Motion compensation	Disparity MAE [px]↓
×	✓	✓	6.2775
✓	×	✓	3.3968
✓	✓	×	8.3657
✓	✓	✓	2.9010

Table 11. **Quantitative results of the ablation study.** We analyze the impact of removing each module on disparity accuracy. The table presents the mean absolute error (MAE) of the disparity map for different configurations: (1) without ADEC (fixed dual exposure), (2) without weighted fusion, and (3) without motion compensation. Removing ADEC results in a significantly high disparity error due to poor adaptation to scene DR. Excluding weighted fusion leads to feature instability, increasing errors. The absence of motion compensation causes severe misalignment, further degrading performance. Our complete method achieves the lowest MAE, demonstrating its effectiveness.

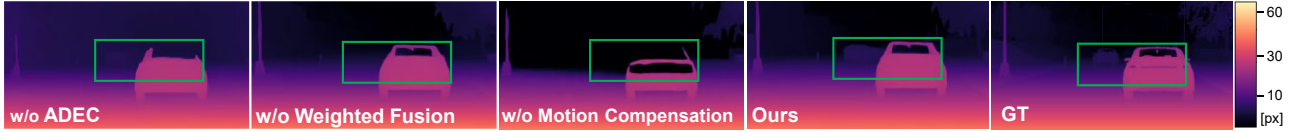


Figure 14. **Ablation study on individual modules.** We evaluate the impact of three core components: ADEC, weighted feature fusion, and motion compensation. The figure visualizes how each module affects the final disparity estimation. The absence of ADEC leads to poor adaptation to scene dynamic range (DR), increasing disparity errors. Without weighted feature fusion, unstable features from over-/under-exposed regions negatively impact the disparity map. Omitting motion compensation results in severe misalignment errors due to uncorrected optical flow. Our full method achieves the most accurate disparity reconstruction.

6. Additional Discussion

6.1. Motion Blur in Dataset Acquisition

While our method demonstrates significant improvements in disparity estimation under challenging lighting conditions, it is not without limitations. One key challenge arises during dataset acquisition, particularly when the stereo cameras are mounted on a mobile robot. Despite careful synchronization of the stereo cameras, as described in Section 2, motion blur can occur in consecutive frames if the mobile robot experiences sharp rotations or vibrations during movement. This motion blur, even in a single frame, can adversely affect our dual-exposure disparity estimation pipeline.

Figure 15 illustrates an example of this limitation. (a) shows a sample captured from our dataset, where one of the frames exhibits motion blur due to the robot’s movement. (b) compares disparity maps generated by different methods for this scene. The results indicate that our method is particularly sensitive to motion blur, as it relies on the effective fusion of details from consecutive frames. The blurred frame reduces the accuracy of feature alignment and fusion, ultimately impacting the disparity estimation.

To address the limitations posed by motion blur, several strategies can be explored. First, robust feature extraction techniques could be employed to reduce the sensitivity to motion blur. This could include pre-processing steps such as deblurring algorithms or using motion-compensated encoders to improve the quality of extracted features. Second, frames affected by severe motion blur can be automatically detected and excluded from training or evaluation using motion blur detection algorithms that analyze temporal or spatial gradients. Lastly, employing higher frame rate cameras during dataset acquisition could significantly reduce motion blur by capturing images at shorter time intervals, thereby improving the alignment and

fusion of stereo features in our pipeline. These solutions offer promising directions to enhance the robustness of our method against motion blur while maintaining its effectiveness in challenging scenarios. Future work will focus on implementing these strategies to further enhance the robustness of our method in dynamic real-world scenarios.

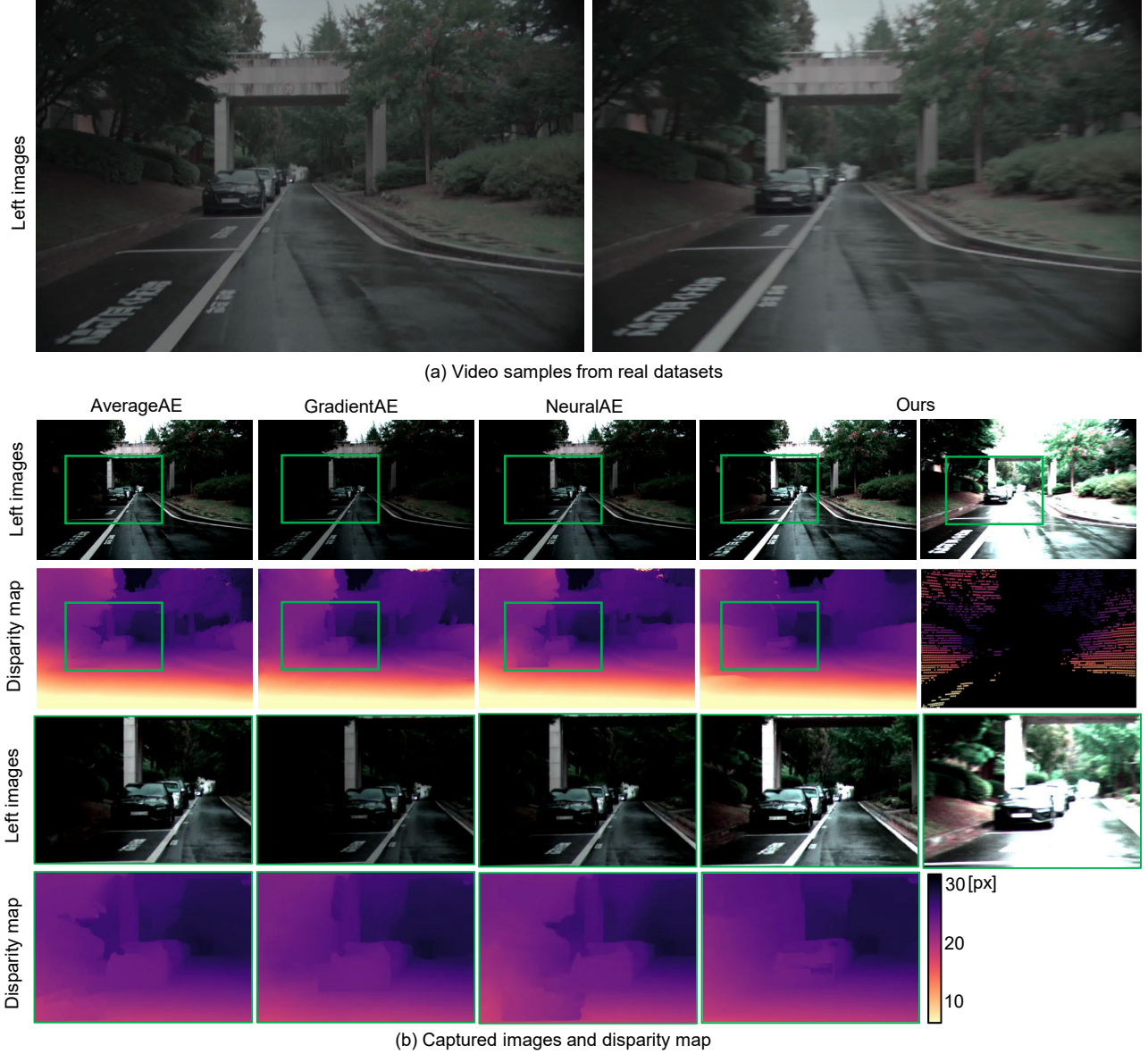


Figure 15. **Impact of Motion Blur on Disparity Estimation.** (a) An example of motion blur in one frame due to robot movement during dataset acquisition. (b) Disparity maps generated by different methods for the same scene, showing the sensitivity of our method to motion blur.

6.2. Impact of Fast Motion on Depth Accuracy

Our current prototype demonstrates robust performance for moderate human motion, as shown in Figure 16. However, for extremely fast-moving objects, such as a fast-moving car, the performance gracefully degrades due to severe motion blur. The exposure differences between stereo frames introduce additional challenges when handling high-speed motion, as rapid scene changes can exacerbate misalignment issues and depth estimation errors. To address this limitation, future work may incorporate motion blur modeling into both the image simulation process and the training pipeline. Explicitly modeling

motion blur within the stereo framework could help the network learn more robust representations for dynamic scenes. Additionally, integrating deblurring techniques or exposure-aware motion compensation strategies could further enhance accuracy in scenarios with fast-moving objects.

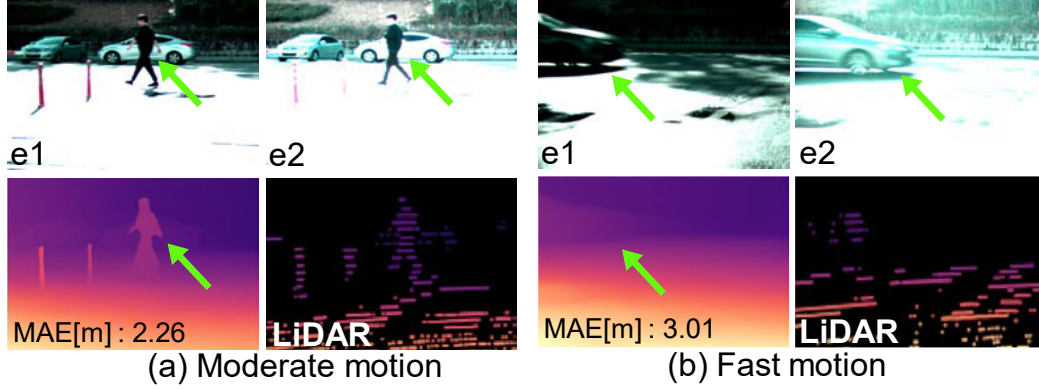


Figure 16. **Impact of fast motion on depth accuracy.** Effect of motion speed on captured images, disparity estimation, and LiDAR ground truth. (a) Under moderate motion, depth estimation remains robust with minimal errors. (b) Under fast motion, severe motion blur leads to misalignment between stereo frames, increasing depth estimation errors.

6.3. Challenges with LiDAR Points in Outdoor Scenarios

Despite the benefits of using LiDAR data as ground-truth for disparity estimation, challenges arise when capturing outdoor scenes, particularly under adverse weather conditions. Unlike indoor scenes where LiDAR points are densely distributed, outdoor environments often result in sparser point measurements due to various factors. For instance, as shown in Figure 17, outdoor scenes with wet ground caused by rain introduce significant inaccuracies in the LiDAR data. The reflective nature of the wet surface can disrupt the LiDAR signal, leading to incomplete or noisy point measurements. This limitation inhibit the generation of accurate ground-truth disparity maps, especially in regions where the surface is wet or reflective. Figure 17 illustrates this issue, where (a) depicts the dual-exposure stereo images of indoor and outdoor scenes, (b) visualizes the disparity map generated by our method, and (c) shows the corresponding LiDAR points. The difference in point density between indoor and outdoor scenes is particularly evident, highlighting the limitations of LiDAR under specific conditions.

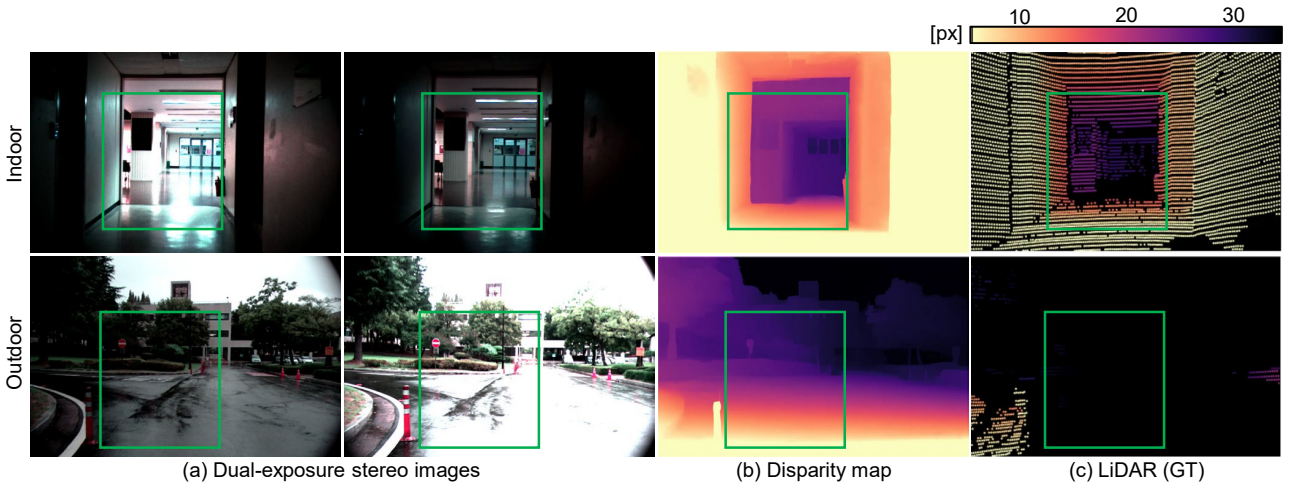


Figure 17. **Challenges with LiDAR Points in Indoor and Outdoor Scenarios.** (a) Dual-exposure stereo images for indoor and outdoor scenes. (b) Disparity maps generated by our method, showing accurate reconstruction for indoor and outdoor scenes (c) LiDAR ground-truth points, illustrating the variation in point density between indoor and outdoor scenes, particularly on wet ground in the outdoor scenario.

6.4. Initial Exposure Setting

The initial exposure setting plays a critical role in the performance of dual-exposure disparity estimation. Our exposure control mechanism increases the exposure gap when the scene is determined to have a wide dynamic range, up to a predefined exposure gap. Once this gap is reached, the control mechanism maintains the exposure gap as long as the scene continues to exhibit a wide dynamic range. However, the specific exposure values at which this gap is maintained can vary depending on the initial exposure setting and scene characteristics.

As shown in Figure 12, when the initial exposures are set to unequal values, the ablation model captures more detail in the first time step due to the larger exposure gap. However, as the exposure gap stabilizes, the model struggles to maintain optimal detail capture, resulting in suboptimal performance compared to the baseline model, which starts with equal exposures. This is particularly evident in scenarios where maintaining the exposure gap is insufficient to fully capture the details of both bright and dark regions.

This observation highlights the importance of carefully selecting the initial exposure setting to balance detail capture across the entire dynamic range of the scene. Future work could focus on adaptive initialization strategies tailored to the scene's characteristics to improve robustness and consistency.

References

- [1] ARM. 2020. Mali-C71. <https://www.arm.com/products/silicon-ip-multimedia/image-signal-processor/mali-c71ae>. Camera product.. 12, 13, 14, 15
- [2] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*. PMLR, 1–16. 7
- [3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)* (2013). 10
- [4] Lahav Lipson, Zachary Teed, and Jia Deng. 2021. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 218–227. 10, 11, 20
- [5] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. 2022. Ghost-free high dynamic range imaging with context-aware transformer. In *European Conference on Computer Vision*. Springer, 344–360. 8
- [6] Henrique Morimitsu, Xiaobin Zhu, Roberto M Cesar, Xiangyang Ji, and Xu-Cheng Yin. 2024. Rapidflow: Recurrent adaptable pyramids with iterative decoding for efficient optical flow estimation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2946–2952. 10
- [7] Emmanuel Onzon, Fahim Mannan, and Felix Heide. 2021. Neural auto-exposure for high-dynamic range object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7710–7720. 12, 13, 14, 15
- [8] Jitendra Malik Paul Debevec. 1997. Recovering High Dynamic Range Radiance Maps from Photographs. *SIGGRAPH* 29, 6, 1–10. 7, 8
- [9] K Ram Prabhakar, Susmit Agrawal, Durgesh Kumar Singh, Balraj Ashwath, and R Venkatesh Babu. 2020. Towards practical and efficient high-resolution HDR deghosting with CNN. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI* 16. Springer, 497–513. 8
- [10] Zhiyuan Pu, Peiyao Guo, M Salman Asif, and Zhan Ma. 2020. Robust high dynamic range (hdr) imaging with complex motion and parallax. In *Proceedings of the Asian Conference on Computer Vision*. 8
- [11] Inwook Shim, Tae-Hyun Oh, Joon-Young Lee, Jinwook Choi, Dong-Geol Choi, and In So Kweon. 2018. Gradient-based camera exposure control for outdoor mobile platforms. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 6 (2018), 1569–1583. 12, 13, 14, 15
- [12] Gangwei Xu, Yun Wang, Junda Cheng, Jinhui Tang, and Xin Yang. 2023. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (2023), 2461–2474. 20