

Dual Exposure Stereo for Extended Dynamic Range 3D Imaging

Juhhyung Choi^{*} Jinnyeong Kim^{*} Seokjun Choi^{*} Jinwoo Lee[†]
 Samuel Brucker[‡] Mario Bijelic[§] Felix Heide[§] Seung-Hwan Baek^{*}
^{*} POSTECH [†] KAIST [‡]Torc Robotics [§] Princeton University

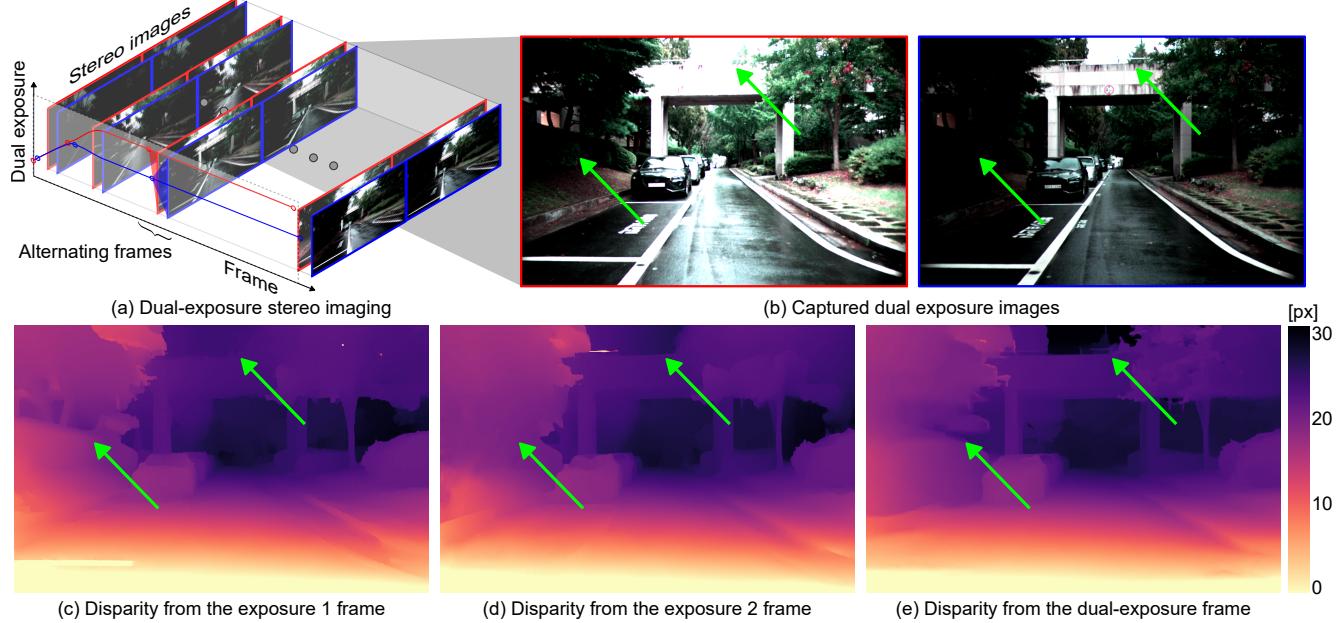


Figure 1. We introduce dual-exposure stereo, a method for extended dynamic range (DR) 3D imaging. (a) We control the dual exposures synchronously set for the stereo camera to expand the effective DR of 3D imaging. From (b) the captured dual-exposure stereo images, we estimate (e) a disparity map that preserves details in both the under- and over-exposed images of the dual-exposure pair, which cannot be faithfully reconstructed in (c)&(d) one-exposure results.

Abstract

Achieving robust stereo 3D imaging under diverse illumination conditions is an important however challenging task, due to the limited dynamic ranges (DRs) of cameras, which are significantly smaller than real world DR. As a result, the accuracy of existing stereo depth estimation methods is often compromised by under- or over-exposed images. Here, we introduce dual-exposure stereo for extended dynamic range 3D imaging. We develop automatic dual-exposure control method that adjusts the dual exposures, diverging them when the scene DR exceeds the camera DR, thereby providing information about broader DR. From the captured dual-exposure stereo images, we estimate depth using motion-aware dual-exposure stereo network. To validate our method, we develop a robot-vision system, collect stereo video datasets, and generate a synthetic dataset. Our method outperforms other exposure control methods.

1. Introduction

Robust 3D imaging is critical for autonomous systems, such as robots and self-driving vehicles, which depend on depth perception to navigate and interact with their environments. Stereo imaging is a popular 3D imaging technique that estimates depth from disparity by matching corresponding pixels in images captured by two cameras. Recent advancements in neural networks have significantly improved stereo disparity estimation, making stereo imaging a practical and cost-effective solution [27].

However, achieving robust 3D imaging with stereo cameras remains challenging, especially in real-world scenes that exhibit lighting conditions with ultra-wide dynamic ranges (DRs). Conventional cameras have limited DR capabilities [36], so in scenes with extremely wide DRs, bright regions may become overexposed while dark regions are underexposed, leading to suboptimal disparity estimation. Existing auto exposure control (AEC) methods adjust cam-

era exposure to capture the scene DR, however they do not expand the camera’s native DR, as each stereo frame is often processed individually [8]. Exposure bracketing techniques capture multiple images with different exposures to expand the effective DR through multi-exposure image processing [9, 28], however these often rely on predefined exposures that do not adapt to the scene DR, increasing capture time and computational overhead.

In this paper, we introduce dual-exposure stereo, a method for extended dynamic range 3D imaging (Figure 1). In alternating frames, we capture stereo images with dual exposures, producing two pairs of stereo images where each pair is captured at different exposure settings in successive frames. By combining AEC and exposure bracketing, we dynamically adjust the dual exposures: when the scene DR exceeds the camera’s native DR, the dual exposures diverge to capture bright and dark regions across two successive frames. When the scene DR is within the camera’s DR, the exposures converge to capture the full scene DR within the camera’s native DR. Each stereo image captured under dual exposures retains the same DR as the camera, however different exposure settings enable coverage of bright and dark regions. To leverage these images, we develop a dual-exposure depth estimation method that fuses dual-exposure features in a motion-aware manner across alternating frames. Our approach effectively extends the DR for 3D imaging, regardless of the original bit depth of the cameras.

To validate our method, we design a robot-vision system equipped with stereo cameras and a LiDAR sensor. Using this setup, we collect a dataset of stereo videos and LiDAR point clouds across indoor and outdoor environments with various lighting conditions. We also generate synthetic datasets with dense ground-truth depth maps. Our experiments demonstrate that the proposed method outperforms previous exposure control methods, enabling depth estimation in scenes with a wide range of DRs. Code and datasets will be made publicly available.

In this paper, we make the following contributions:

- We introduce dual-exposure stereo for extended dynamic range 3D imaging, developing an automatic dual-exposure control method that combines conventional AEC and exposure bracketing. Our dual-exposure disparity estimation method then utilizes dual-exposure stereo images to increase effective camera DR for robust 3D imaging.
- We develop a robot-vision system with stereo cameras and a LiDAR sensor mounted on a wheeled robot, collecting a real-world stereo video dataset and rendering a synthetic dataset with dense ground truth.
- We validate our method on both synthetic and real-world datasets, demonstrating improved performance over existing exposure control methods.

2. Related Work

HDR 3D Imaging Active imaging systems with engineered illumination have enabled 3D imaging under high-dynamic-range (HDR) environments. Examples are synchronized projector-camera systems [33, 34] and time-of-flight cameras [7, 41]. Without additional illumination modules, passive 3D imaging systems exploit unconventional sensors for HDR imaging, such as single-photon avalanche diodes [17], quanta image sensors [11, 13], event cameras [46, 54, 58], and modulo cameras [56, 57]. Using conventional cameras, capturing scenes with multiple exposures is a standard approach for HDR 3D imaging [4, 5, 23]. However, these methods rely on predetermined exposure settings, which cannot adapt to changing lighting conditions, leading to loss of detail in overexposed or underexposed regions and unnecessarily long acquisition times when scene DR is low. Our method automatically adjusts dual exposure for stereo cameras, improving performance for varying-DR scenes by expanding effective camera DR for 3D imaging.

AEC and Exposure Bracketing Single-camera AEC has been extensively studied using histogram analysis [2, 39, 44], model-predictive correction [35, 42, 43, 45], entropy analysis [25, 31, 53], and semantic analysis [32, 51, 52]. Extending AEC to stereo cameras, there are two common approaches. The first is to control exposure for each camera individually. However, this deteriorates stereo correspondence due to stereoscopic intensity inconsistency [21, 55]. The second approach is to use a synchronized AEC for stereo cameras, maintaining intensity consistency [21, 40]. However, existing methods in this category struggle when scene DR is larger than camera DR, leading to overexposed or underexposed regions. Exposure bracketing alternates multiple exposures and processes the multi-exposure images to capture a broader DR than the original camera DR [14, 15, 30, 32, 37, 47–49]. Most existing methods use a single camera [16, 50] and rely on predetermined long and short exposure settings [3, 6, 12, 18–20, 26], which limits capture efficiency when scene DR fluctuates. Our method combines the principles of AEC and exposure bracketing. We control the dual exposures of stereo cameras, thus maintaining intensity consistency between stereo images as well as expanding effective DR for 3D imaging.

Stereo Depth Estimation Estimating depth from stereo disparity has been studied for decades [38]. Recent neural-network solutions have significantly improved stereo matching by using deep feature extraction and cost volume construction [22, 24]. These models iteratively refine disparity maps using 3D convolutions or recurrent units. However, they fail when scene DR is wider than camera DR,

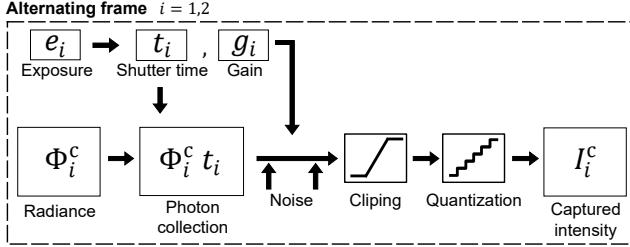


Figure 2. Image Formation. For alternating frame $i = \{1, 2\}$, scene radiance Φ_i , and exposure e_i , we simulate the captured intensity I_i^c for the camera $c \in \{\text{left, right}\}$. We consider photon collection, pre- and post-gain noise, clipping, and quantization.

because of under- and over-exposed regions. Our disparity estimation method addresses this problem by fusing the dual-exposure images while compensating the motion between consecutive frames.

3. Image Formation

We introduce the dual-exposure image formation model for stereo cameras. We denote the dual exposure as e_i , where $i \in \{1, 2\}$ is an alternating frame index. We convert the exposure e_i to shutter time $t_i = e_i/g_i$ and gain $g_i = \max(1, e_i/t_{\max})$, where t_{\max} is the maximum shutter time, allocating shutter time as long as possible capped by the maximum value to reduce image noise by a high gain. Given the shutter time t_i , gain g_i , and the incident scene radiance Φ_i , we model the intensity I_i^c captured by the stereo camera $c \in \{\text{left, right}\}$ at the frame i as:

$$I_i^c(p_i^c) = \text{quant}(\text{clip}(g_i(\Phi_i t_i + n_i^{\text{pre}}) + n_i^{\text{post}})), \quad (1)$$

where p_i^c is a camera pixel. The noise terms n_i^{pre} and n_i^{post} are the pre-gain and post-gain noise, sampled from zero-mean Gaussian distributions with standard deviations σ_{pre} and σ_{post} , respectively: $n_i^{\text{pre}} \sim \mathcal{N}(0, \sigma_{\text{pre}})$, $n_i^{\text{post}} \sim \mathcal{N}(0, \sigma_{\text{post}})$. The function $\text{clip}(\cdot)$ limits intensity values by the camera DR, and $\text{quant}(\cdot)$ quantizes the intensity to integer values. The overall procedure of the image formation is shown in Figure 2.

4. Auto Dual Exposure Control

Our ADEC method uses the left-camera images $\{I_1^{\text{left}}, I_2^{\text{left}}\}$ captured by the dual exposure $\{e_1, e_2\}$ to estimate the next dual exposure $\{\hat{e}_1, \hat{e}_2\}$. Pseudo code of our ADEC method is shown in Algorithm 1. Figure 3 shows an example scenario of applying the ADEC method.

Metric Our ADEC method uses statistical metrics to control next-frame dual exposures. Specifically, for each frame $i \in \{1, 2\}$, we compute the intensity histogram h_i of the image I_i^{left} , and calculate the histogram skewness S_i , describing whether the histogram is skewed towards low intensity

Algorithm 1 Pseudocode for ADEC.

```

Require: Dual exposure values  $e_1, e_2$  and corresponding captured images  $I_1^{\text{left}}, I_2^{\text{left}}$ 
Ensure: Next dual exposure  $\hat{e}_1, \hat{e}_2$ 
1: // Compute metrics
2: for each frame  $i \in \{1, 2\}$  do
3:    $h_i \leftarrow \text{ExtractHistogram}(I_i^{\text{left}})$ 
4:    $S_i \leftarrow \text{Skewness}(h_i)$ 
5:    $L_i, H_i \leftarrow \text{ExtremePixelRatios}(h_i)$ 
6: end for
7: // Adjust dual exposure
8: if Scene DR > camera DR:  $L_i > \tau_h$  and  $H_i > \tau_h$  for any  $i$  then
9:   if Dual exposure gap is low:  $\Delta e = |e_1 - e_2| \leq \tau_{\Delta e}$  then
10:     // Diverge dual exposure
11:      $\hat{e}_1, \hat{e}_2 \leftarrow \text{DivergeDualExposure}(e_1, e_2, L_1, L_2, H_1, H_2)$ 
12:   end if
13: else
14:   // Scene DR > camera DR or scene DR is uncertain
15:   for each frame  $i \in \{1, 2\}$  do
16:     // Adjust dual exposure towards zeroing the skewness
17:      $\hat{e}_i \leftarrow \text{MakeSkewnessZero}(e_i, S_i)$ 
18:   end for
19: end if

```

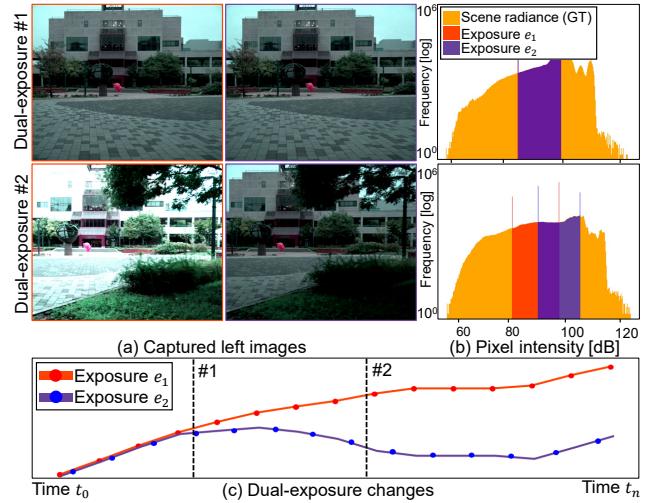


Figure 3. ADEC Method Exposure Selection. The initial dual exposures are set to be the same in this example. If the scene DR is estimated as wider than the camera DR, the dual exposures diverge to capture both dark and bright regions in alternating frames.

(negative skewness) or high intensity (positive skewness) as

$$S_i = \sum_{j=0}^K \left(\frac{j - K/2}{K/2} \right)^3 \frac{h_i(j)}{N}, \quad (2)$$

where K is the maximum detectable intensity of the camera depending on its bit depth, N is the number of pixels, and $h_i(j)$ is the frequency of intensity j .

We then compute the ratios of under- and over-exposed

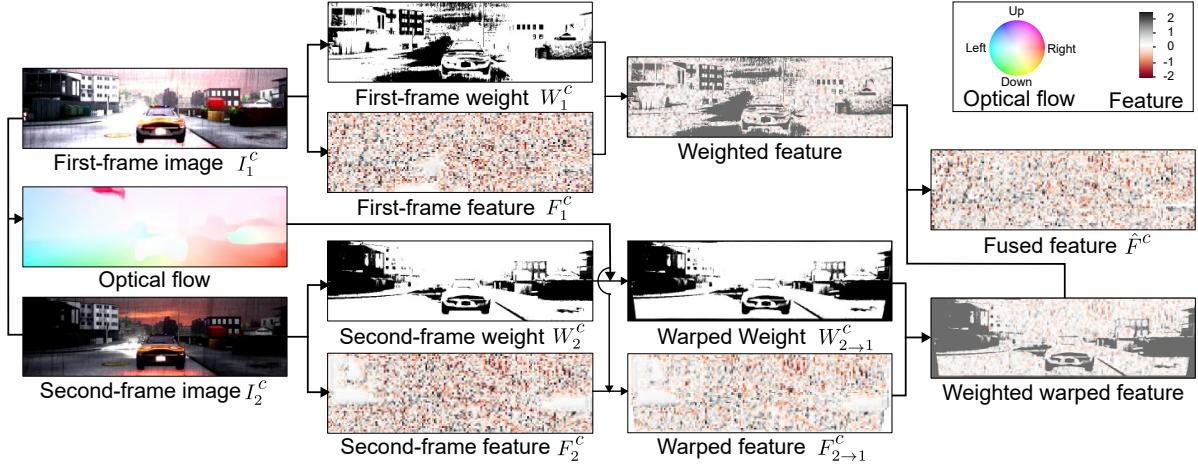


Figure 4. Dual Exposure Feature Fusion. For each camera c , we extract features F_1^c and F_2^c , optical flow f^c , and weight maps W_1^c and W_2^c of the dual-exposure images I_1^c and I_2^c . The second-frame features and weight map are warped to the first-frame view, and fused to the final feature \hat{F}^c with the weighted summation, encoding dual-exposure information.

pixels, denoted as L_i and H_i , representing the proportions of pixels near the minimum and maximum intensities

$$L_i = \sum_{j=0}^{T_{\text{low}}} \frac{h_i(j)}{N}, \quad H_i = \sum_{j=T_{\text{high}}}^K \frac{h_i(j)}{N}, \quad (3)$$

where $T_{\text{low}} = \lfloor K \times 0.05 \rfloor$ and $T_{\text{high}} = \lfloor K \times 0.95 \rfloor$ are clamping thresholds.

Below, we use the skewness S_i , extreme-valued pixel ratios L_i, H_i to determine the next-frame dual exposures.

Diverging Dual Exposure For scenes whose DR exceeds the camera native DR, we diverge the dual exposure to capture a broader DR across the alternating frames. Such cases are identified with the conditions $L_i > \tau_h$ and $H_i > \tau_h$ for at least one frame i , where $\tau_h = 0.05$. We diverge the dual exposure with magnitudes proportional to L_i and H_i as

$$\begin{aligned} \hat{e}_1 &= e_1 + \alpha L_1, & \hat{e}_2 &= e_2 - \alpha H_2, & \text{if } e_1 > e_2 \\ \hat{e}_1 &= e_1 - \alpha H_1, & \hat{e}_2 &= e_2 + \alpha L_2, & \text{otherwise,} \end{aligned} \quad (4)$$

where $\alpha = 0.5$ is a constant controlling the divergence step.

To prevent excessive divergence that could lead to unstable stereo imaging, we limit the exposure difference $\Delta e = |e_1 - e_2|$. If Δe exceeds a threshold $\tau_{\Delta e} = 2.5$, no further divergence is applied.

Towards Zeroing Skewness When the scene DR is uncertain or lower than the camera DR, indicated by $L_i < \tau_h$ and $H_i < \tau_h$ for both frames $i \in \{1, 2\}$, we adjust the exposures towards zeroing the skewness S_i . This makes the intensity histogram to be balanced, capturing the scene DR:

$$\hat{e}_i = e_i - \alpha \times S_i. \quad (5)$$

This process makes the dual exposure converge to a similar value, to fully cover the scene DR in a balanced manner, if the scene DR is lower than camera DR. For scenes with uncertain DR compared to the camera DR, this method moves the dual exposure to be with zero skewed, facilitating better identification of the scene DR.

5. Dual-exposure Stereo Disparity Estimation

We introduce our disparity estimation method for dual-exposure stereo images: $I_1^{\text{left}}, I_1^{\text{right}}$ of the first frame, and $I_2^{\text{left}}, I_2^{\text{right}}$ of the second frame.

Motion Estimation Between alternating frames $i \in \{1, 2\}$, the locations of corresponding pixels can move for dynamic movements of objects or cameras, modeled as the optical flow f^c :

$$p_1^c = p_2^c + f^c(p_2^c). \quad (6)$$

To account for such motion, we estimate the optical flow f^{left} between I_1^{left} and I_2^{left} , and f^{right} between I_1^{right} and I_2^{right} using a pretrained optical flow network robust to intensity difference between images [29]. Note that the dual-exposure images I_1^c and I_2^c have overlapped contents as we do not allow for extremely-diverged dual exposures as described in Section 4.

The estimated optical flow f^c allows us to define a warping function that transforms data X_2^c from the second frame ($i = 2$) to the first frame ($i = 1$):

$$X_{2 \rightarrow 1}^c = \text{warp}(X_2^c, f^c), \quad (7)$$

where $X_{2 \rightarrow 1}^c$ is the warped data. The warping function entails resizing the optical flow field to the corresponding resolution of the input data X_2^c .

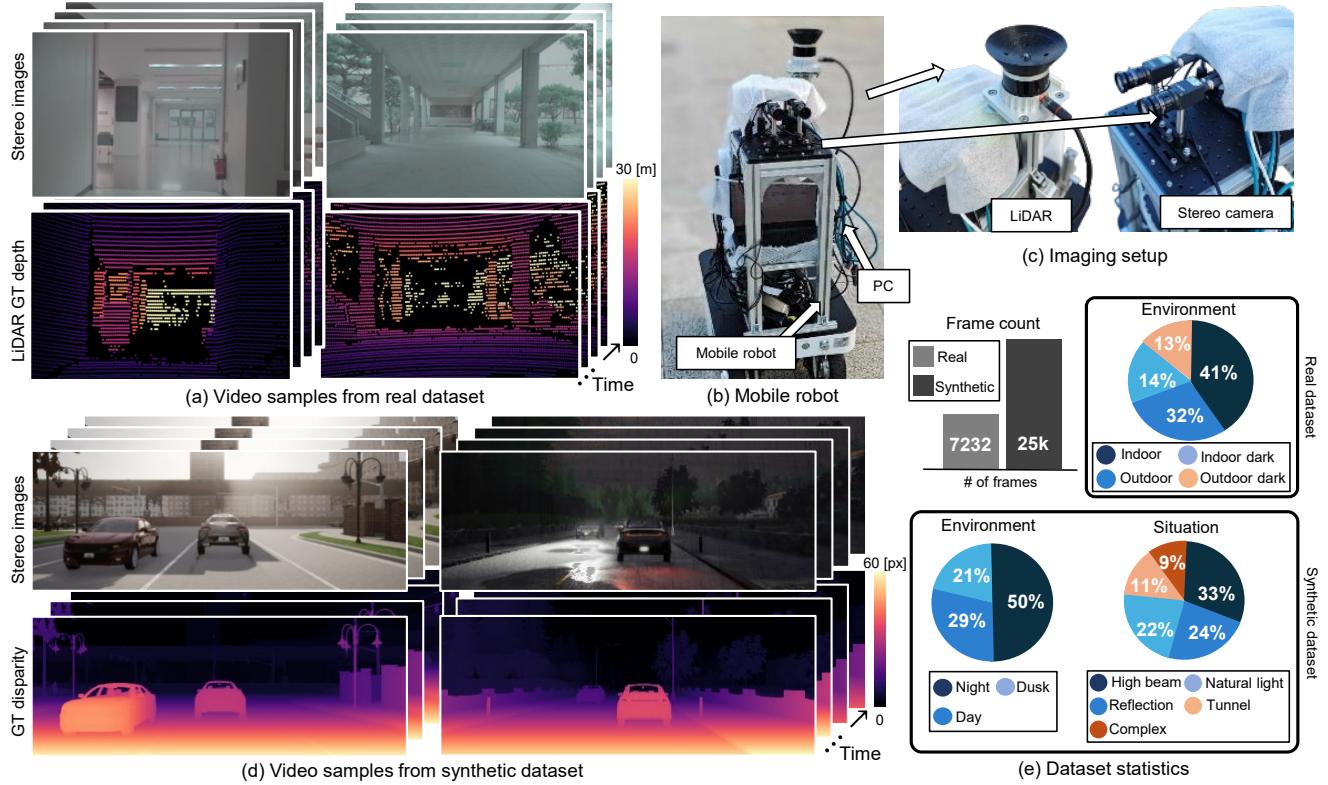


Figure 5. Stereo Video Datasets. (a) Samples of our real-world dataset: sequence of stereo images and corresponding LiDAR point cloud. (b)&(c) Our imaging setup (c) mounted on a mobile robot (b), equipped with a LiDAR sensor and stereo cameras. (d) Samples of our synthetic dataset: stereo images and ground-truth disparity maps. (e) Statistics of the synthetic dataset about time, location, and scenarios.

Dual-exposure Feature Fusion To exploit dual-exposure images $\{I_1^c, I_2^c\}$ with potentially different exposures, we develop a dual-exposure feature fusion method. Figure 4 shows the overview of our dual-exposure feature fusion. We extract features F_i^c from images I_i^c using a pretrained feature extractor $\text{FE}(\cdot)$ [24]:

$$F_i^c = \text{FE}(I_i^c). \quad (8)$$

We then warp the second-frame feature to the first frame using the warping function based on the estimated optical flow, ensuring that feature from the second frame becomes spatially aligned with the first-frame feature:

$$F_{2 \rightarrow 1}^c = \text{warp}(F_2^c, f^c). \quad (9)$$

As we now have the dual exposure features $F_{2 \rightarrow 1}^c$ and F_1^c spatially aligned, we fuse the two features using the weighted sum:

$$\hat{F}^c = \frac{W_1^c \cdot F_1^c + W_{2 \rightarrow 1}^c \cdot F_{2 \rightarrow 1}^c}{W_1^c + W_{2 \rightarrow 1}^c + \epsilon}, \quad (10)$$

where \hat{F}^c is the fused feature, ϵ is a small constant to avoid division by zero.

We define the weight map W_i^c using an intensity-based trapezoidal function [19] to exploit well-exposed pixel intensity:

$$W_i^c(p_i^c) = \begin{cases} \frac{I_i^c(p_i^c)}{\alpha} & \text{if } I_i^c(p_i^c) < \alpha, \\ 1 & \text{if } \alpha \leq I_i^c(p_i^c) \leq \beta, \\ 1 - \frac{1}{1-\beta} (I_i^c(p_i^c) - \beta) & \text{if } I_i^c(p_i^c) > \beta, \end{cases} \quad (11)$$

where $\alpha = 0.02$ and $\beta = 0.98$ are the thresholds. The weight maps W_1^c and $W_{2 \rightarrow 1}^c$ are computed for the first frame and the second frame followed by being warped to the first-frame using the estimated optical flow.

We apply the dual-exposure feature fusion of Equation (10), obtaining the fused feature maps \hat{F}^{left} and \hat{F}^{right} .

Stereo Disparity Estimation Using the fused feature maps, we construct correlation volumes C as

$$C(x, y, d) = \hat{F}^{\text{left}}(x, y) \cdot \hat{F}^{\text{right}}(x + d, y), \quad (12)$$

where x, y are the pixel coordinates and d is the disparity. Our dual-exposure feature fusion encodes both dark and bright features of dual-exposure stereo images in the correlation volume, allowing for effectively extended DR

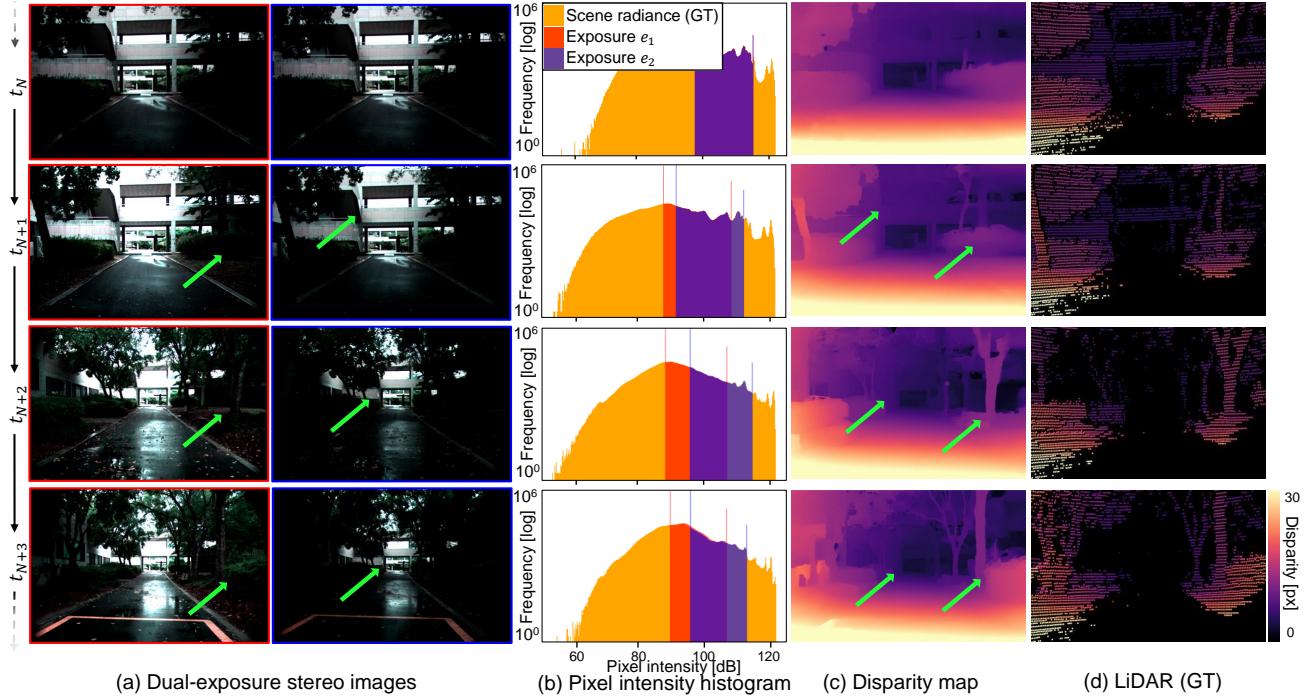


Figure 6. **Extended-DR 3D imaging.** Our ADEC controls the dual exposures for extended-DR depth estimation. The dual-exposure depth estimation module obtains depth both for bright and dark regions which cannot be captured by only one image.

for 3D imaging. We also apply a multi-scale feature fusion approach for robustness. We then estimate a disparity map from the correlation volume using a disparity estimation network [24]. We finetune the network on our synthetic dataset (Section 6.2).

6. Stereo Video Datasets

To validate our method, we introduce two stereo video datasets: one for real-world scenes and the other for synthetic scenes. Figure 5 visualizes samples from our datasets.

6.1. Real-world Dataset

Prototype System We developed an imaging system consisting of stereo RGB cameras (LUCID Triton 5.4MP) and a LiDAR sensor (Ouster OS-1). The cameras capture linear stereo images with 24-bit depth, which are then compressed to simulate 8-bit images using Equation (1) to evaluate whether our method can expand effective DR for 3D imaging. Note that our method can be applied to cameras with any bit depth. We performed intrinsic and extrinsic calibration of the stereo camera system using a chessboard-based method. Subsequently, we estimated the extrinsic transformation between the LiDAR and the left camera through ICP alignment process. LiDAR point clouds are projected onto the left-camera view, providing pseudo ground-truth sparse depth maps. The stereo images and LiDAR depth maps are time-synchronized. We mount the imaging system on a

wheeled robot (AgileX Ranger-Mini 2.0), as shown in Figure 5(c), enabling both indoor and outdoor captures. We configured a stereo camera system with a 110 mm baseline to enable effective depth estimation in both indoor and outdoor environments. The depth measurement range of our system extends up to 60 meters. For system details, we refer to the Supplementary Document.

Captured Dataset Our real-world dataset encompasses a broad range of environments and lighting conditions, thus appropriate for evaluating our method. The dataset includes 33 scenes and 7432 frames, with a resolution of 1440×928 pixels. The dataset is balanced with 41% of frames captured in indoor scenes, 32% in outdoor scenes, 14% in indoor low-light scenes, and 13% in outdoor low-light scenes. Further details are provided in the supplementary material.

6.2. Synthetic Dataset

We use the CARLA simulator [10] to generate a synthetic dataset for diverse automotive scenarios with varying lighting conditions. We render synchronized stereo images with 32 bit depth, which is used to simulate 8-bit images using Equation (1). We also render ground-truth dense disparity maps. Our synthetic dataset comprises 1,000 training videos and 200 testing videos. Training videos consist of 20 frames each, and testing videos contain 100 frames each. We simulate various lighting conditions (day, dusk, and night), as shown in Figure 5, which vary within each

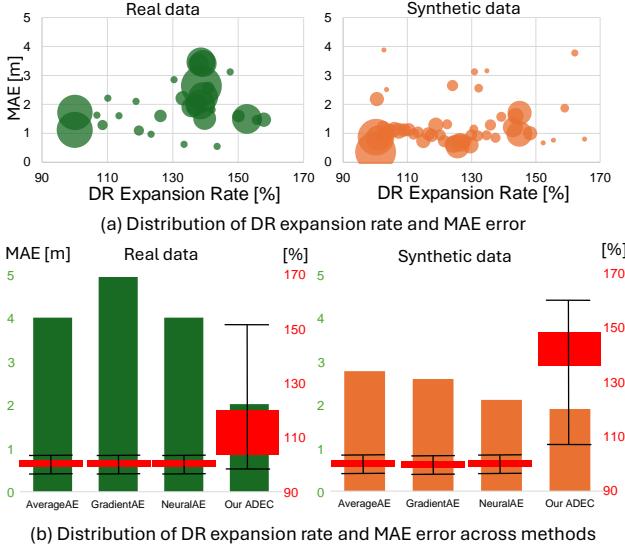


Figure 7. Disparity accuracy vs. DR expansion rate. (a) We demonstrate effective DR expansion up to 160% without significant performance drop. (b) Other AEC methods [1, 32, 40] cannot expand DR resulting in large error.

video, introducing abrupt changes in dynamic range due to environmental effects such as high beams at night, intense reflections, and sunlight emerging from tunnels. Further details are provided in the Supplementary Material.

7. Experiments

Extended-DR 3D Imaging We leverage complementary information of two-frame stereo images captured with dual exposure, which is estimated by our ADEC module. By fusing dual-exposure features, we obtain accurate disparity map that preserves details across a wider DR than the native camera DR. Figure 6 reports the estimated disparity maps for input dual-exposure stereo images. Our method enables expanding effective DR for disparity estimation with both details of dark and bright regions, which cannot be solely captured with the camera DR.

We assess the trade-off between the depth accuracy and the DR expansion rate. Compared to the maximum DRs of 42dB corresponding to 8-bit depth, we calculate the effectively-enhanced DR covered by the dual-exposure frames in our method. Figure 7 shows that our method retains high depth accuracy for the DR expansion rate of 160%, which demonstrates the effectiveness of our method. In contrast, depth estimation using other state-of-the-art AEC methods [1, 32, 40] cannot expand the DR, resulting in large error in depth reconstruction.

Comparison We compare our method with three state-of-the-art AEC methods that control and process single exposure [1, 32, 40]. We use the RAFT-Stereo model [24] for

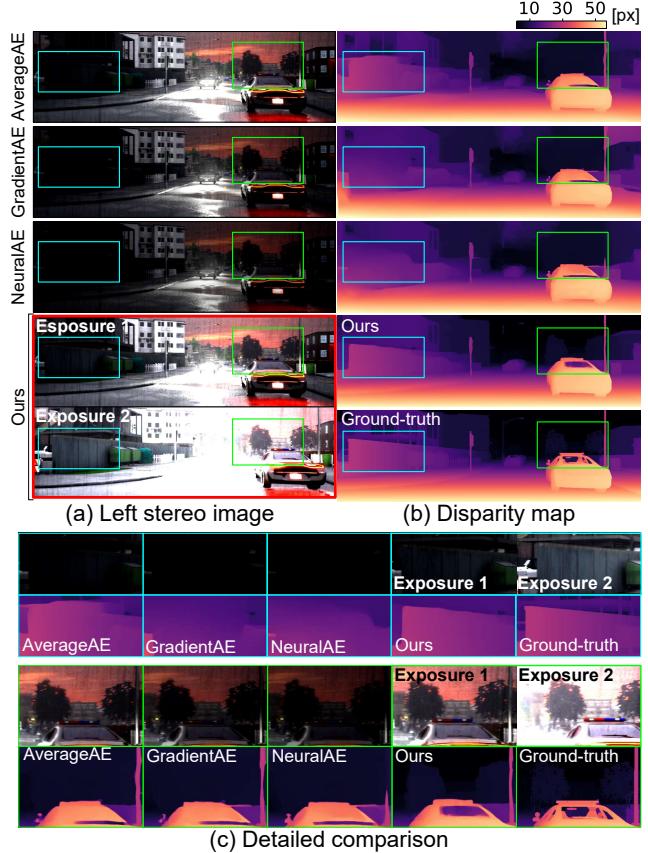


Figure 8. Disparity accuracy using our ADEC compared with other AEC methods. Our method outperforms the other AEC methods: AverageAE [1], GradientAE [40], and NeuralAE [32].

their depth estimation. Note that ours is the only method that combines the principles of AEC and exposure bracketing, which is exploited by our dual-exposure disparity estimation module. Figures 8 and 9 show the qualitative results on a synthetic scene and a real-world scene, demonstrating that our method only recovers details in both dark and bright regions by effectively expanding DR for 3D imaging. Table 1 confirms that our method also quantitatively outperforms the other AEC methods for accurate 3D imaging on both synthetic and real-world datasets. We also compare the speeds of AEC methods and our ADEC method considering its real-time use cases, which is an important factor for any exposure control methods. Our ADEC can run at more than 120 FPS, supporting real-time applications, while the neural AEC [32] fails to support real-time imaging.

Ablation Study We evaluate the importance of the three core components: ADEC module, weighted feature fusion, and motion compensation. Table 2 and Figure 10 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

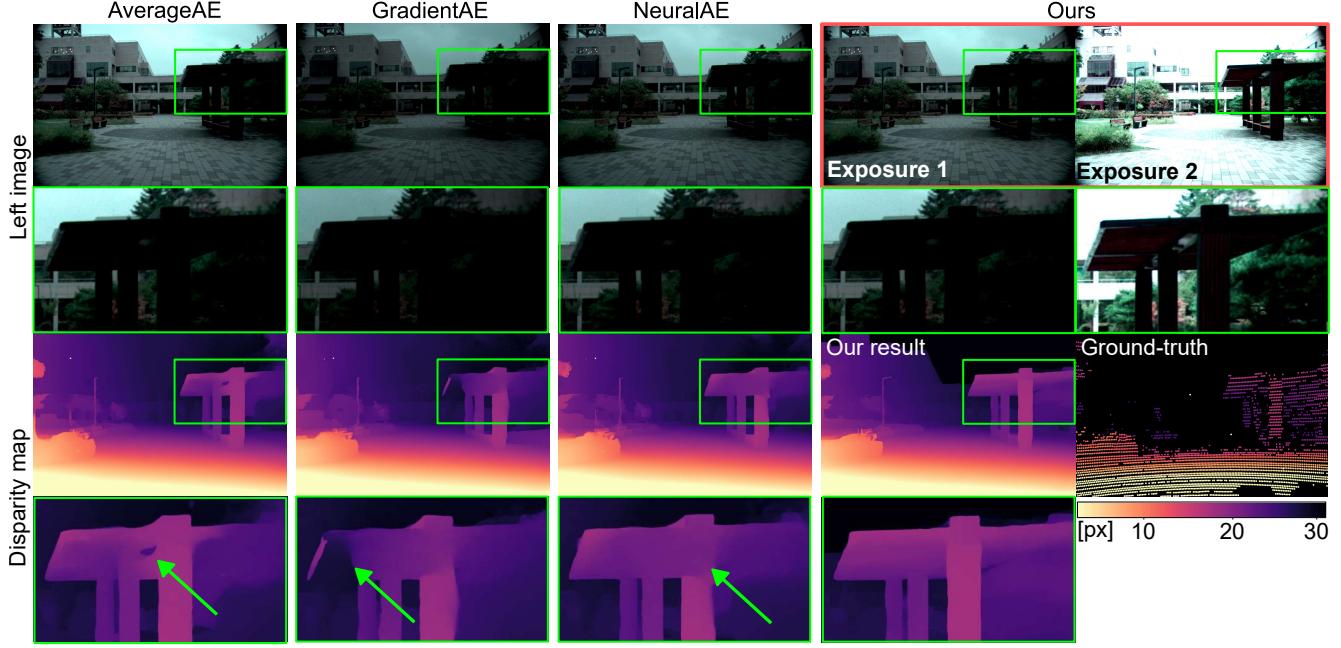


Figure 9. **Disparity-estimation results using our ADEC compared with other AEC methods.** Our ADEC method outperforms the other AEC methods for subsequent extended-DR depth estimation: AverageAE [1], GradientAE [40], and NeuralAE [32].



Figure 10. **Ablation study.** We evaluate the impact of three modules: ADEC, weighted feature fusion, and motion compensation. Our full method enable the most accurate reconstruction.

	AverageAE [1]	GradientAE [40]	NeuralAE [32]	ADEC (ours)
Synthetic Data Disp. MAE [px] ↓	2.823	2.948	<u>2.778</u>	1.355
Real Data Depth MAE [m] ↓	2.7679	2.5847	<u>1.9232</u>	1.9142
FPS↑	616.27	42.10	0.25	<u>124.58</u>

Table 1. **3D imaging accuracy and FPS comparison of our ADEC against other methods.** Our method outperforms the other AEC methods in terms of disparity and depth MAE on synthetic and real datasets, respectively. We also compare FPS. Best numbers are in **bold** and the second best are in underline.

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 features 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 2. **Ablation study.**

8. Conclusion

In this work, we introduce a dual exposure stereo for extended DR 3D imaging. We devise a ADEC module and dual-exposure depth estimation method, expanding the effective DR for robust 3D imaging. To validate this method, we report a stereo video dataset consisting of stereo videos and LiDAR pointclouds, collected by our robot vision system. We evaluate the effectiveness of our method, outperforming conventional AEC methods across all experimental settings we tested. As a method that can expand the DR of any stereo camera for 3D imaging, we hope the proposed approach can be a step to depth imaging in even more extreme lighting and environmental conditions, including photon-starved captures in fog, rain, or snow.

References

- [1] ARM. Mali-C71. <https://www.arm.com/products/silicon-ip-multimedia/image-signal-processor/mali-c71ae>, 2020. Camera product. 7, 8
- [2] Sebastiano Battiatto, Arcangelo Ranieri Bruna, Giuseppe Messina, and Giovanni Puglisi. *Image processing for embedded devices*. Bentham Science Publishers, 2010. 2
- [3] 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. 2
- [4] Yeyao Chen, Mei Yu, Ken Chen, Gangyi Jiang, Yang Song, Zongju Peng, and Fen Chen. New stereo high dynamic range imaging method using generative adversarial networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3502–3506. IEEE, 2019. 2
- [5] Yeyao Chen, Gangyi Jiang, Mei Yu, You Yang, and Yo-Sung Ho. Learning stereo high dynamic range imaging from a pair of cameras with different exposure parameters. *IEEE Transactions on Computational Imaging*, 6:1044–1058, 2020. 2
- [6] 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 (ICCV)*, pages 12760–12769, 2023. 2
- [7] Miguel Heredia Conde, Klaus Hartmann, and Otmar Loeffel. Adaptive high dynamic range for time-of-flight cameras. *IEEE Transactions on Instrumentation and Measurement*, 64(7):1885–1906, 2014. 2
- [8] Sascha Cvetkovic, Helios Jellema, and Peter HN de With. Automatic level control for video cameras towards hdr techniques. *EURASIP Journal on Image and Video Processing*, 2010:1–30, 2010. 2
- [9] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 643–652. 2023. 2
- [10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 6
- [11] Eric R Fossum, Jiaju Ma, Saleh Masoodian, Leo Anzagira, and Rachel Zizza. The quanta image sensor: Every photon counts. *Sensors*, 16(8):1260, 2016. 2
- [12] Natasha Gelfand, Andrew Adams, Sung Hee Park, and Kari Pulli. Multi-exposure imaging on mobile devices. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 823–826, 2010. 2
- [13] Abhiram Gnanasambandam and Stanley H Chan. Hdr imaging with quanta image sensors: Theoretical limits and optimal reconstruction. *IEEE transactions on computational imaging*, 6:1571–1585, 2020. 2
- [14] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 1
- [15] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7312–7321, 2017. 2
- [16] Xiangyu Hu, Liqian Shen, Mingxing Jiang, Ran Ma, and Ping An. La-hdr: Light adaptive hdr reconstruction framework for single ldr image considering varied light conditions. *IEEE Transactions on Multimedia*, 2022. 2
- [17] Atul Ingle, Trevor Seets, Mauro Buttafava, Shantanu Gupta, Alberto Tosi, Mohit Gupta, and Andreas Velten. Passive inter-photon imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8595, 2021. 2
- [18] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep hdr video from sequences with alternating exposures. In *Computer graphics forum*, pages 193–205. Wiley Online Library, 2019. 2
- [19] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen. Patch-based high dynamic range video. *ACM Trans. Graph.*, 32(6):202–1, 2013. 5
- [20] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions on Graphics (TOG)*, 22(3):319–325, 2003. 2
- [21] Hyun-Woo Kim, Soon Kwon, Jung Je-Kyo, and JaeWook Ha. Auto-exposure control method for a stereo camera robust to brightness variation. *International Journal of Control and Automation*, 7(1):321–330, 2014. 2
- [22] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022. 2
- [23] Huei-Yung Lin and Wei-Zhe Chang. High dynamic range imaging for stereoscopic scene representation. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 4305–4308. IEEE, 2009. 2
- [24] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 2, 5, 6, 7
- [25] Huimin Lu, Hui Zhang, Shaowu Yang, and Zhiqiang Zheng. A novel camera parameters auto-adjusting method based on image entropy. In *RoboCup 2009: Robot Soccer World Cup XIII 13*, pages 192–203. Springer, 2010. 2
- [26] Stephen Mangiat and Jerry Gibson. High dynamic range video with ghost removal. In *Applications of Digital Image Processing XXXIII*, pages 307–314. SPIE, 2010. 2
- [27] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 1

- [28] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer graphics forum*, pages 161–171. Wiley Online Library, 2009. [2](#)
- [29] Henrique Morimitsu, Xiaobin Zhu, Roberto M Cesar, Xiangyang Ji, and Xu-Cheng Yin. Rapidflow: Recurrent adaptable pyramids with iterative decoding for efficient optical flow estimation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2946–2952. IEEE, 2024. [4](#)
- [30] Ratnajit Mukherjee, Miguel Melo, Vitor Filipe, Alan Chalmers, and Maximino Bessa. Backward compatible object detection using hdr image content. *IEEE Access*, 8: 142736–142746, 2020. [2](#)
- [31] Jingyi Ning, Tiejun Lu, Liyan Liu, Liye Guo, and Xiaofeng Jin. The optimization and design of the auto-exposure algorithm based on image entropy. In *2015 8th International Congress on Image and Signal Processing (CISP)*, pages 1020–1025. IEEE, 2015. [2](#)
- [32] Emmanuel Onzon, Fahim Mannan, and Felix Heide. Neural auto-exposure for high-dynamic range object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7710–7720, 2021. [2, 7, 8](#)
- [33] Matthew O’Toole, John Mather, and Kiriakos N Kutulakos. 3d shape and indirect appearance by structured light transport. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3246–3253, 2014. [2](#)
- [34] Matthew O’Toole, Sreeth Achar, Srinivasa G Narasimhan, and Kiriakos N Kutulakos. Homogeneous codes for energy-efficient illumination and imaging. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015. [2](#)
- [35] SangHyun Park, GyuWon Kim, and JaeWook Jeon. The method of auto exposure control for low-end digital camera. In *2009 11th International Conference on Advanced Communication Technology*, pages 1712–1714. IEEE, 2009. [2](#)
- [36] Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010. [1](#)
- [37] Pinar Satilmis, Thomas Bashford-Rogers, Alan Chalmers, and Kurt Debattista. Per-pixel classification of clouds from whole sky hdr images. *Signal Processing: Image Communication*, 88:115950, 2020. [2](#)
- [38] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002. [2](#)
- [39] Simon Schulz, Marcus Grimm, and R Grigat. Using brightness histogram to perform optimum auto exposure. *WSEAS Transactions on Systems and Control*, 2(2):93, 2007. [2](#)
- [40] Inwook Shim, Tae-Hyun Oh, Joon-Young Lee, Jinwook Choi, Dong-Geol Choi, and In So Kweon. Gradient-based camera exposure control for outdoor mobile platforms. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(6):1569–1583, 2018. [2, 7, 8](#)
- [41] Gal Shtendel and Ayush Bhandari. Hdr-tof: Hdr time-of-flight imaging via modulo acquisition. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3808–3812. IEEE, 2022. [2](#)
- [42] Yuanhang Su and C-C Jay Kuo. Fast and robust camera’s auto exposure control using convex or concave model. In *2015 IEEE International Conference on Consumer Electronics (ICCE)*, pages 13–14. IEEE, 2015. [2](#)
- [43] Yuanhang Su, Joe Yuchieh Lin, and C-C Jay Kuo. A model-based approach to camera’s auto exposure control. *Journal of Visual Communication and Image Representation*, 36:122–129, 2016. [2](#)
- [44] Juan Torres and José Manuel Menéndez. Optimal camera exposure for video surveillance systems by predictive control of shutter speed, aperture, and gain. In *Real-Time Image and Video Processing 2015*, pages 238–251. SPIE, 2015. [2](#)
- [45] Quoc Kien Vuong, Se-Hwan Yun, and Suki Kim. A new auto exposure and auto white-balance algorithm to detect high dynamic range conditions using cmos technology. In *Proceedings of the world congress on engineering and computer science*, pages 22–24. San Francisco, USA: IEEE, 2008. [2](#)
- [46] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 155–171. Springer, 2020. [2](#)
- [47] Jian-Gang Wang and Lu-Bing Zhou. Traffic light recognition with high dynamic range imaging and deep learning. *IEEE Transactions on Intelligent Transportation Systems*, 20(4): 1341–1352, 2018. [2](#)
- [48] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6849–6857, 2019.
- [49] Xu Wang, Zhenhao Sun, Qiudan Zhang, Yuming Fang, Lin Ma, Shiqi Wang, and Sam Kwong. Multi-exposure decomposition-fusion model for high dynamic range image saliency detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4409–4420, 2020. [2](#)
- [50] Zhouxia Wang, Jiawei Zhang, Mude Lin, Jiong Wang, Ping Luo, and Jimmy Ren. Learning a reinforced agent for flexible exposure bracketing selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1820–1828, 2020. [2](#)
- [51] Huan Yang, Baoyuan Wang, Noranart Vesdapunt, Minyi Guo, and Sing Bing Kang. Personalized exposure control using adaptive metering and reinforcement learning. *IEEE transactions on visualization and computer graphics*, 25(10):2953–2968, 2018. [2](#)
- [52] Ming Yang, Ying Wu, James Crenshaw, Bruce Augustine, and Russell Mareachen. Face detection for automatic exposure control in handheld camera. In *Fourth IEEE International Conference on Computer Vision Systems (ICVS’06)*, pages 17–17. IEEE, 2006. [2](#)
- [53] Chi Zhang, Zheng You, and Shijie Yu. An automatic exposure algorithm based on information entropy. In *Sixth*

- International Symposium on Instrumentation and Control Technology: Signal Analysis, Measurement Theory, Photo-Electronic Technology, and Artificial Intelligence*, pages 152–156. SPIE, 2006. [2](#)
- [54] Xiang Zhang, Wei Liao, Lei Yu, Wen Yang, and Gui-Song Xia. Event-based synthetic aperture imaging with a hybrid network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14235–14244, 2021. [2](#)
 - [55] Xiaoying Zhang, Xiaojun Tang, Liping Yu, and Bing Pan. Automated camera exposure control for accuracy-enhanced stereo-digital image correlation measurement. *Sensors*, 22(24):9641, 2022. [2](#)
 - [56] Hang Zhao, Boxin Shi, Christy Fernandez-Cull, Sai-Kit Yeung, and Ramesh Raskar. Unbounded high dynamic range photography using a modulo camera. In *2015 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2015. [2](#)
 - [57] Chu Zhou, Hang Zhao, Jin Han, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Unmodnet: Learning to unwrap a modulo image for high dynamic range imaging. *Advances in Neural Information Processing Systems*, 33:1559–1570, 2020. [2](#)
 - [58] Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to reconstruct high speed and high dynamic range videos from events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2024–2033, 2021. [2](#)

Dual Exposure Stereo for Extended Dynamic Range 3D Imaging

Juhhyung 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
3. Datasets	6
3.1. Stereo Real Video Dataset	6
3.2. Stereo Synthetic Video Dataset	7
4. Dual-Exposure Depth Estimation	9
4.1. Network Architecture	9
4.2. Training Details	10
5. Additional Results	11
5.1. Additional Evaluation on Real Dataset	11
5.2. Additional Evaluation on Synthetic Dataset	13
5.3. Additional Ablation Experiments	14
6. Additional Discussion	19
6.1. Motion blur in dataset acquisition	19
6.2. Challenges with LiDAR points in outdoor scenarios	20
6.3. Initial Exposure Setting	21

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.

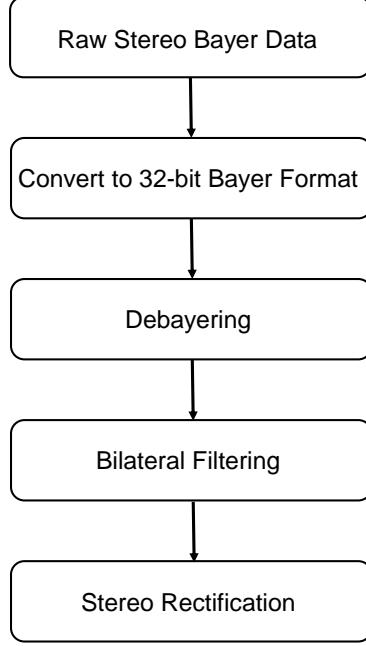


Figure 1. Image Pre-Processing Pipeline. The pipeline includes (1) loading raw Bayer data, (2) converting 24-bit raw Bayer patterns to 32-bit Bayer format, (3) performing debayering with custom bilinear interpolation and OpenCV color conversion, (4) applying bilateral filtering to reduce grid-like artifact, (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 $c \in \{\text{left, right}\}$ at pixel p_i^c is modeled as:

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

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.
- **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.

3. Datasets

3.1. Stereo Real Video Dataset

The dataset was captured using our stereo camera system described in main paper Section 3, 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.



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.

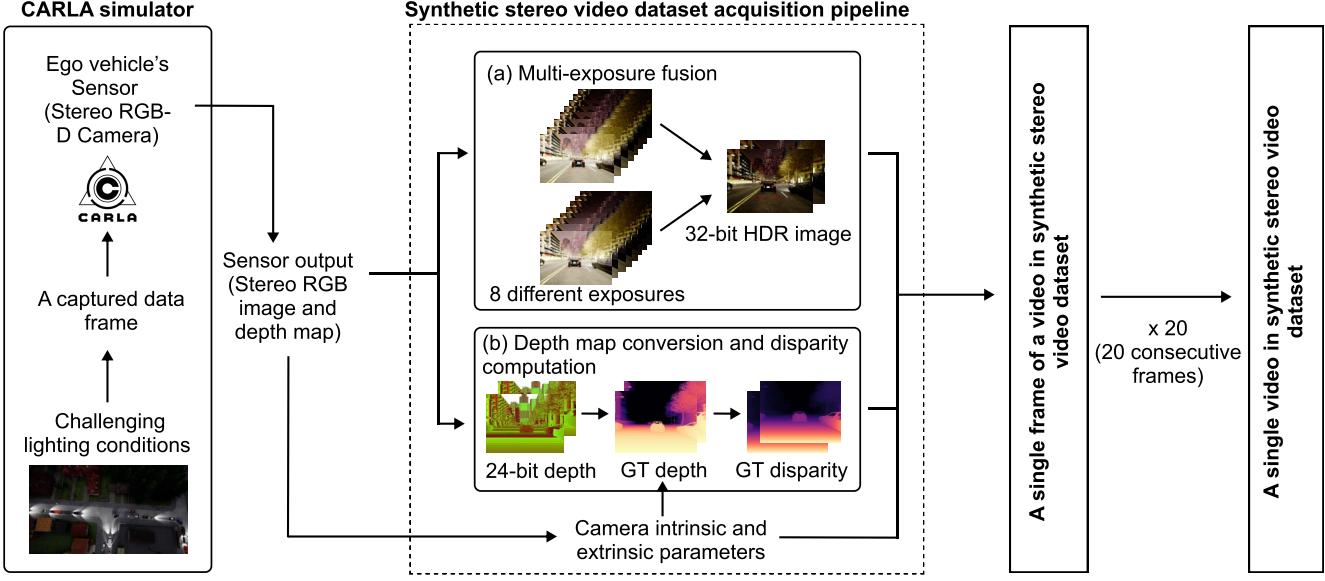


Figure 3. Overview of the synthetic stereo video dataset acquisition pipeline. With specific environmental settings on CARLA simulator to generate adverse lighting conditions, the equipped RGB cameras capture images to reconstruct HDR scenes and the depth cameras create corresponding ground truth disparity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

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.

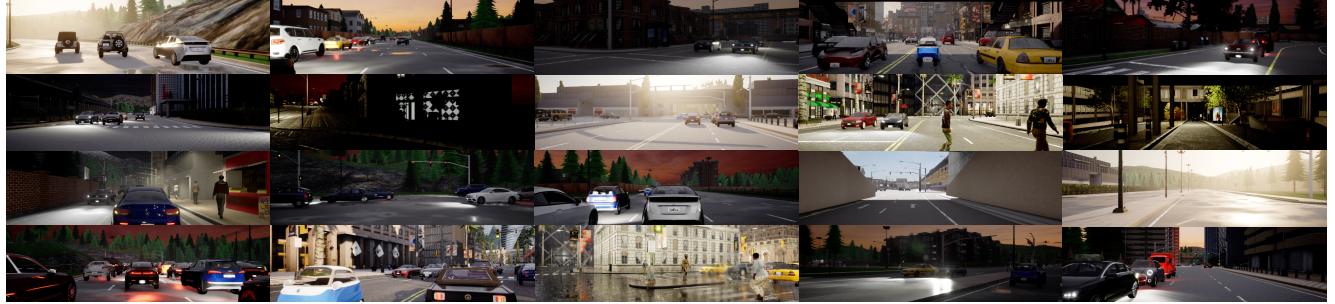
Dataset Acquisition Pipeline Figure 3 shows an overview of our stereo synthetic video dataset acquisition pipeline. In our CARLA simulator ego-vehicle’s capture setup, stereo RGB cameras and paired depth cameras—each with a resolution of 1280×384 pixels, a horizontal field of view of 75 degrees, and 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 $t \in \left\{ \frac{1}{500n} (\text{sec}) \mid n \in \{1, 2, \dots, 8\} \right\}$ 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 $t \in \left\{ \frac{1}{50n} (\text{sec}) \mid n \in \{1, 2, \dots, 8\} \right\}$ 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

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°

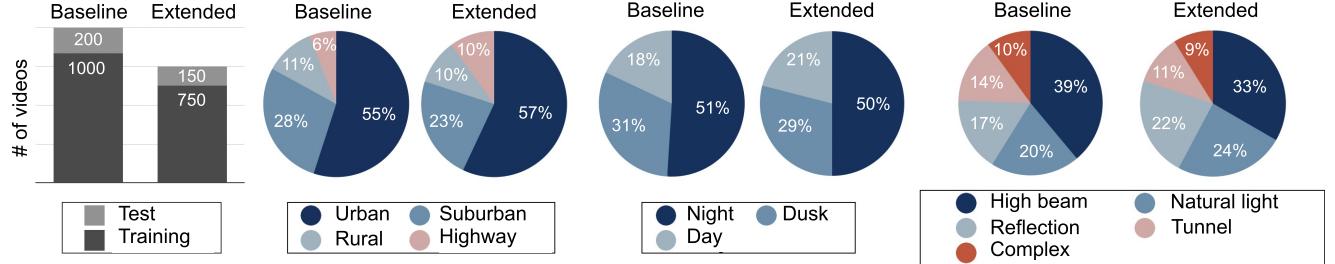
Table 2. **List of sensors used for CARLA simulator ego-vehicle’s capture setup.** Here, ISO is configured based on the temporal condition. Four categories of sensors are mounted at $x = 2.5$ m, $y = \pm 0.2$ m, $z = 1.4$ m with respect to the ego-vehicle’s centroid. Note that the given coordinates follow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



(a) Dataset thumbnails with diverse driving environments



(b) Dataset scene label statistics

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.

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

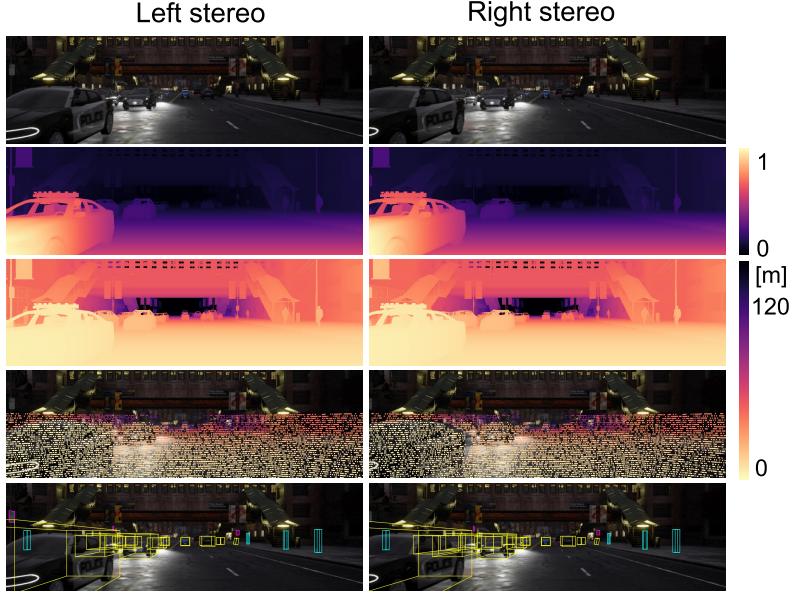


Figure 5. Extended synthetic stereo video datasets. The bottom two rows depict LiDAR point clouds and 3D bounding boxes projected onto the tone-mapped stereo HDR images. To ensure consistency across different modalities, both the depth and LiDAR maps share the same color bar, while the disparity map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

locations include urban and suburban areas, rural areas, and highways. 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 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.

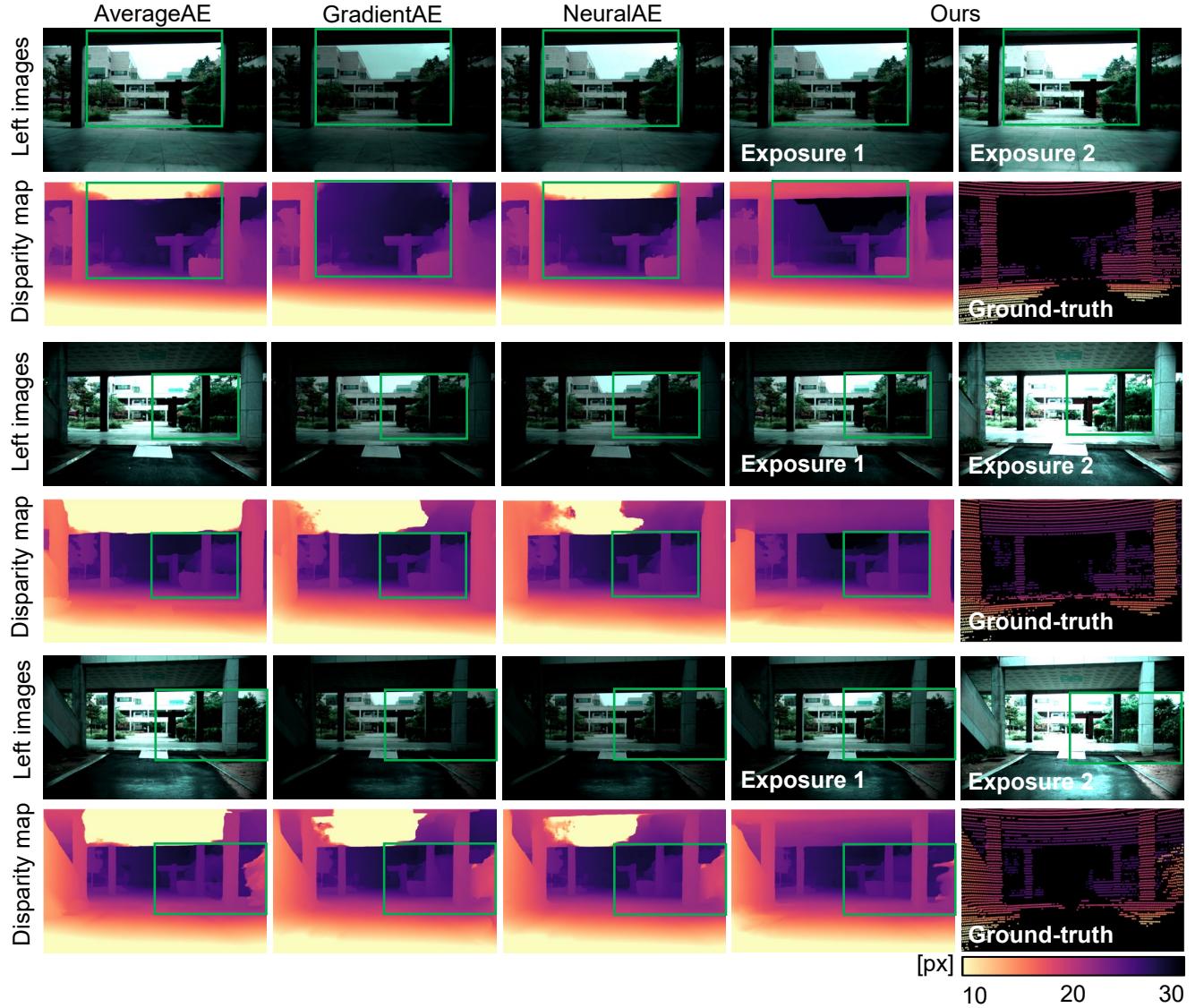


Figure 6. **Disparity-estimation results using our ADEC compared with other AEC methods in outdoor scene** Our ADEC method outperforms the other AEC methods for subsequent extended-DR depth estimation : AverageAE [1], GradientAE [11], NeuralAE [7]

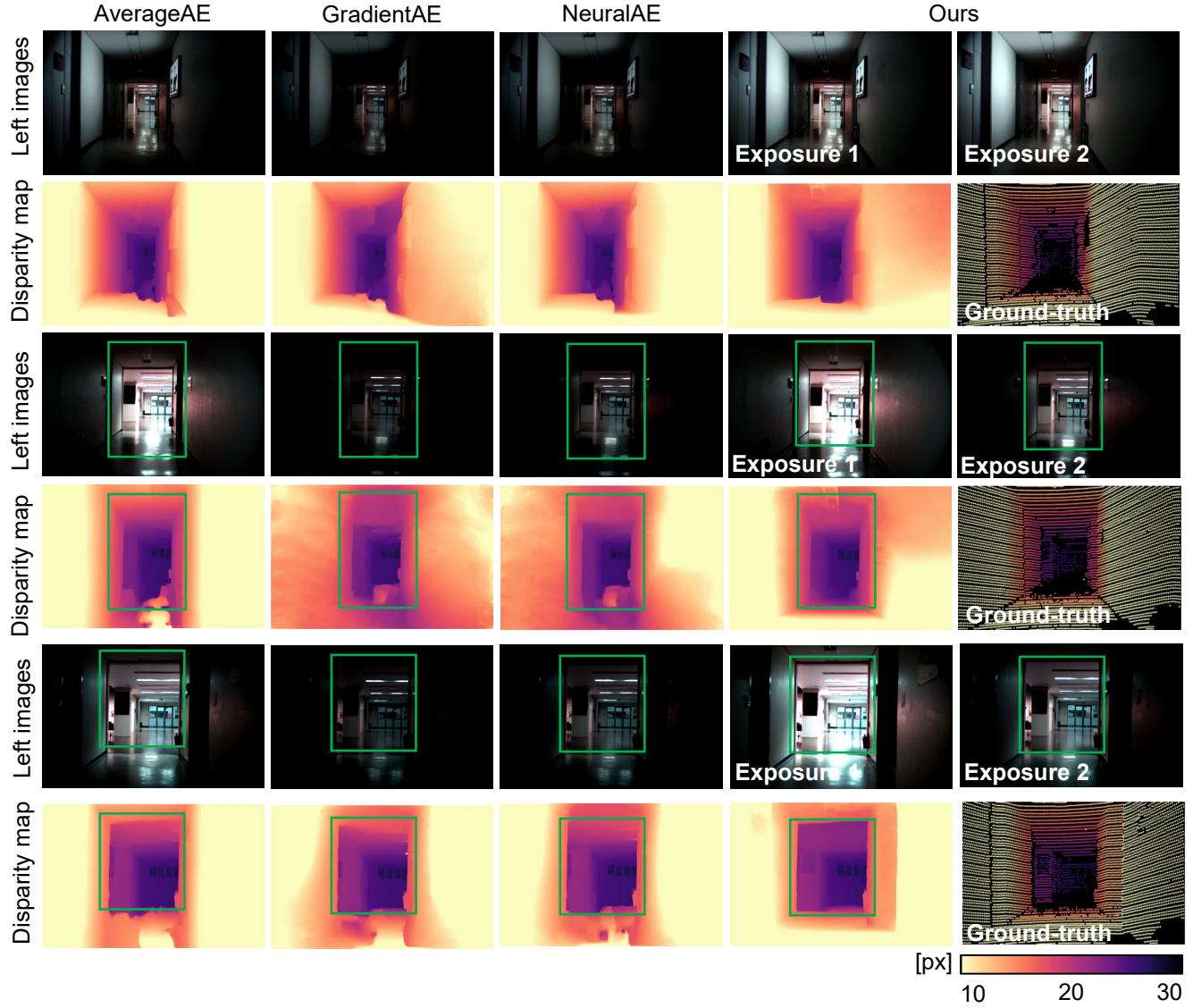


Figure 7. Disparity-estimation results using our ADEC compared with other AEC methods in indoor scene Our ADEC method outperforms the other AEC methods for subsequent extended-DR depth estimation : AverageAE [1], GradientAE [11], NeuralAE [7]

5.2. Additional Evaluation on Synthetic Dataset

We conducted a 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.

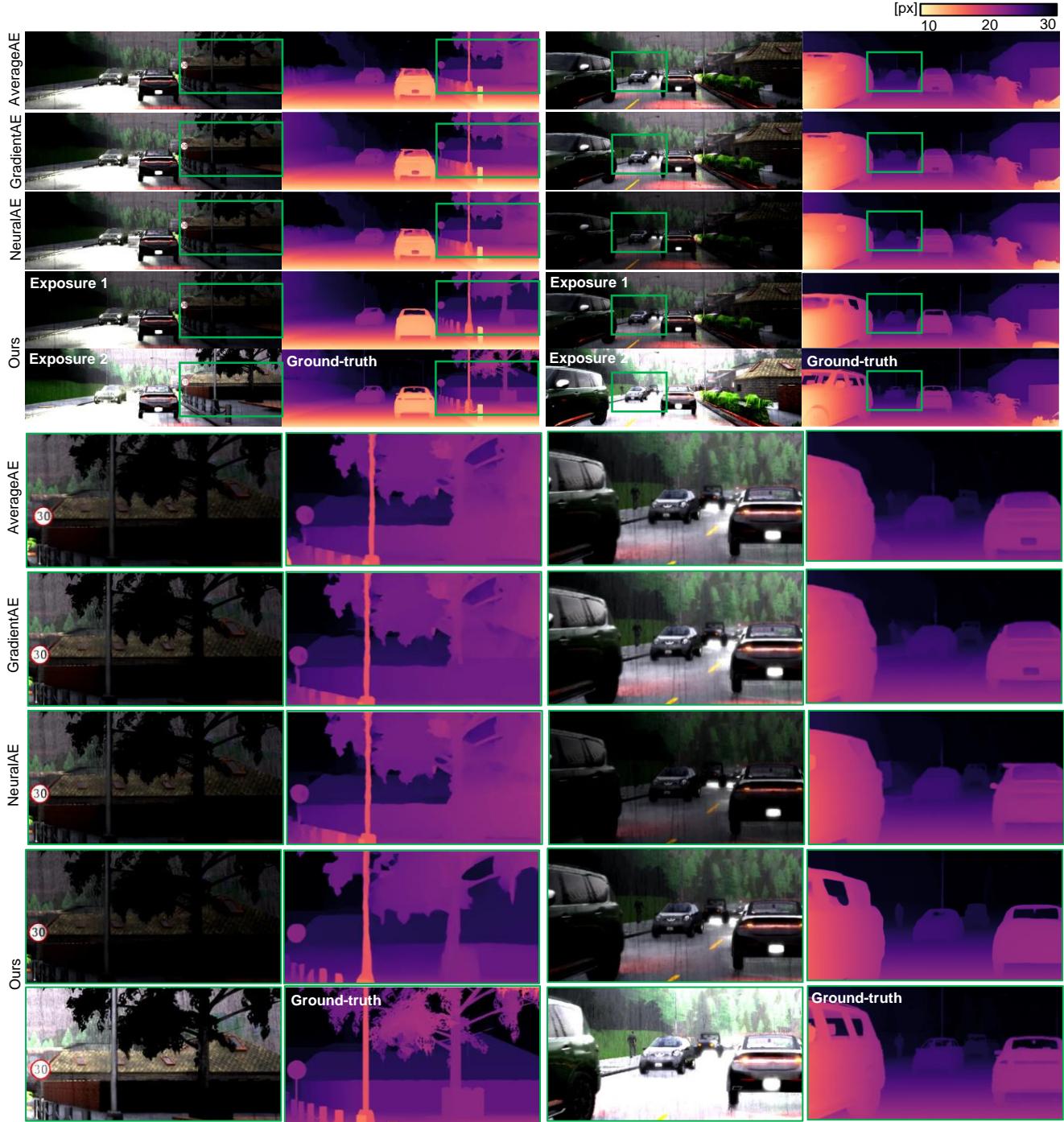


Figure 8. **Disparity-estimation results using our ADEC compared with other AEC methods** Our ADEC method outperforms the other AEC methods for subsequent extended-DR depth estimation : AverageAE [1], GradientAE [11], NeuralAE [7].

5.3. Additional Ablation Experiments

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 9, 10, and 11. 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	<u>2.8759</u>
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 In Figure 9, 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.

Exposure Increase Rate In Figure 10 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 11, 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.

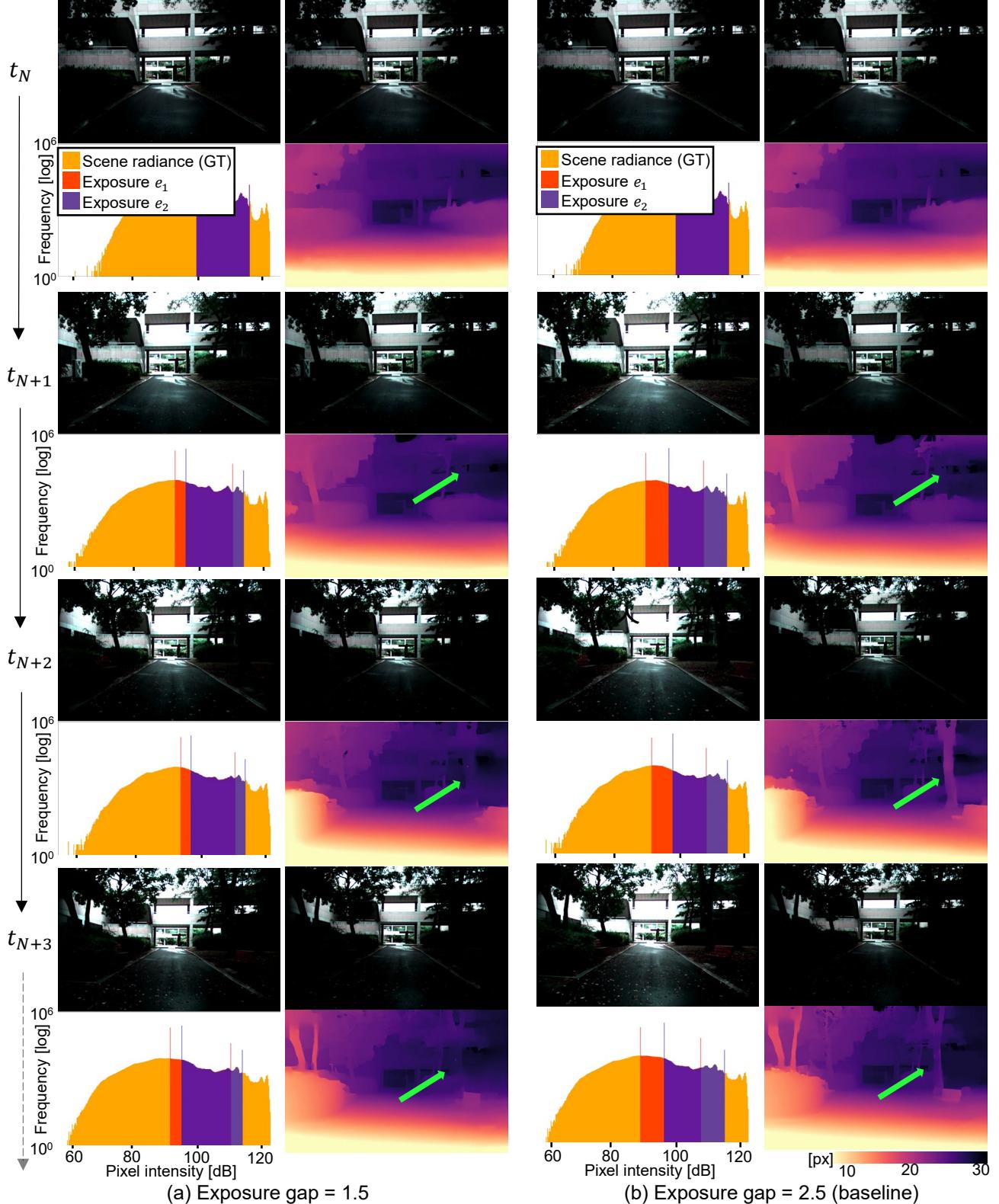


Figure 9. Impact of exposure gap settings on disparity estimation. This figure illustrates the effect of varying the exposure gap during dual-exposure control. The baseline model (exposure gap = 2.5) captures sufficient details over time, whereas the ablation model (exposure gap = 1.5) struggles to widen 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.

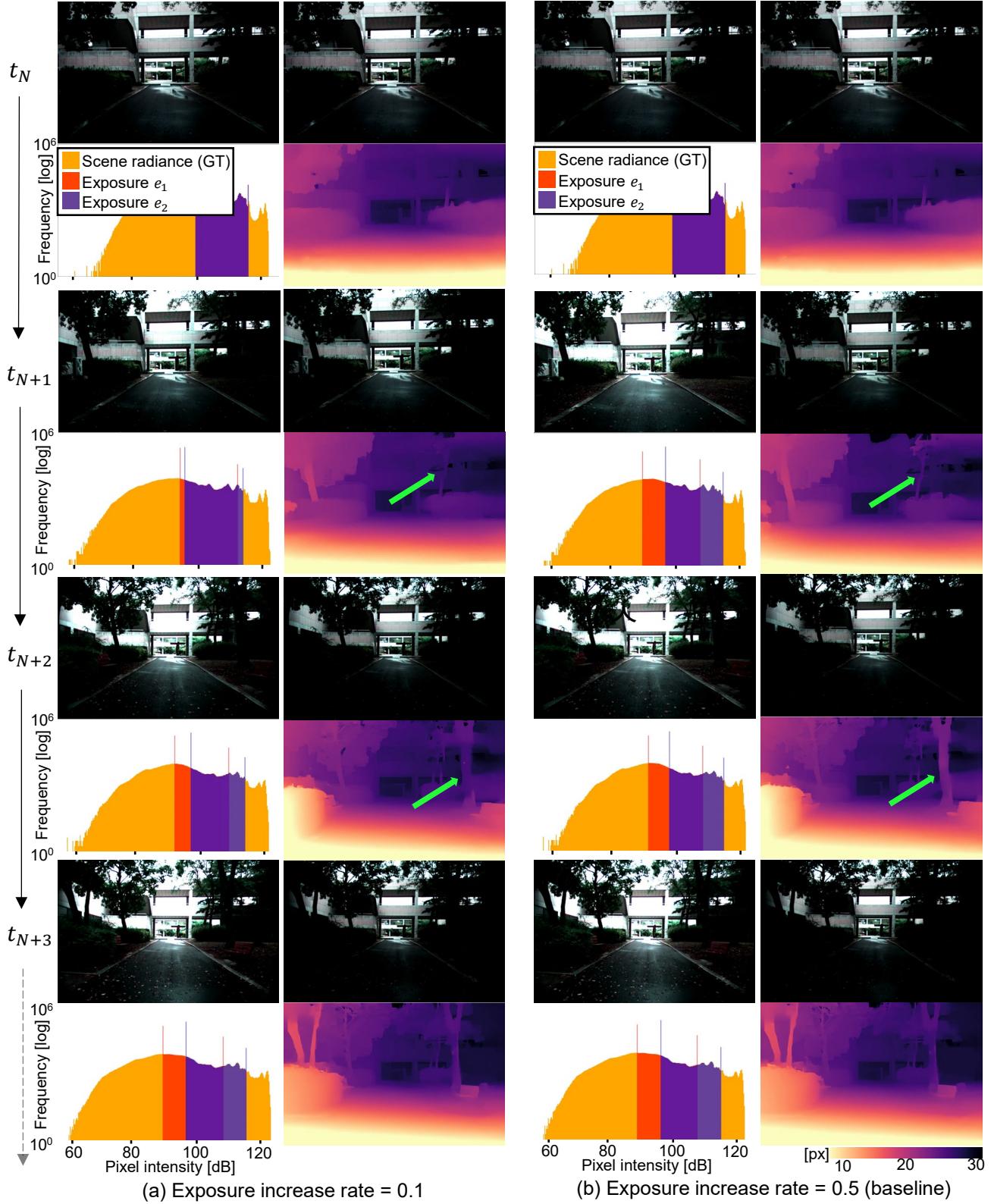


Figure 10. Impact of exposure increase rate on disparity estimation. This figure demonstrates the effect of modifying the exposure increase rate during dual-exposure control. The baseline model, with its default increase rate, quickly expands the exposure gap in the initial time steps, enabling effective 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.

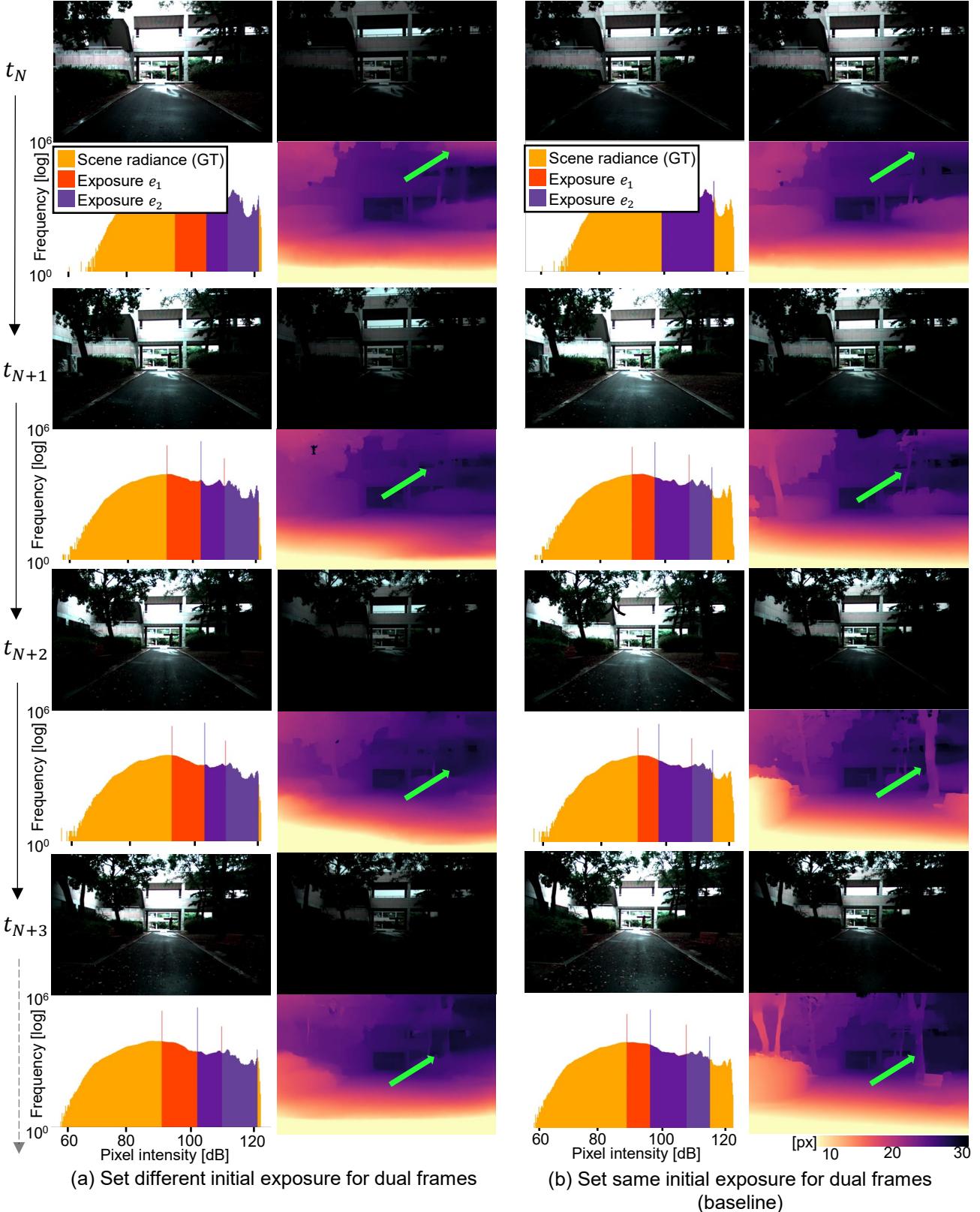


Figure 11. Impact of initial exposure settings on disparity estimation. This figure compares baseline and ablation models with different initial exposures. The baseline model uses equal exposures, ensuring consistent detail capture over time. The ablation model starts with unequal exposures, capturing more detail initially but losing balance in later time steps.

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 12 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.

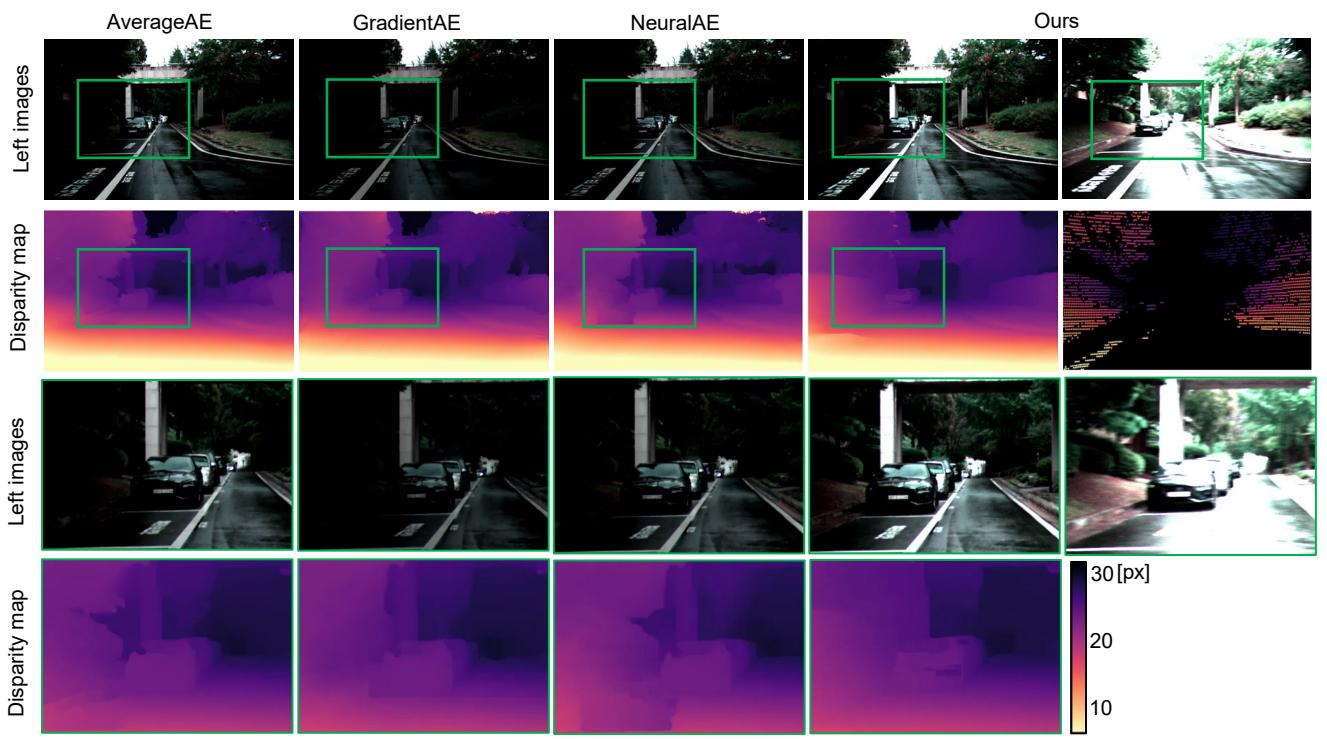


Figure 12. **Impact of Motion Blur on Disparity Estimation.** (a) An example of motion blur in one frame due to robot movement during dataset acquisition. (b) Disparity maps generated by different methods for the same scene, showing the sensitivity of our method to motion blur.

6.2. 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 13, 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 inhibits the generation of accurate ground-truth disparity maps, especially in regions where the surface is wet or reflective. Figure 13 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.

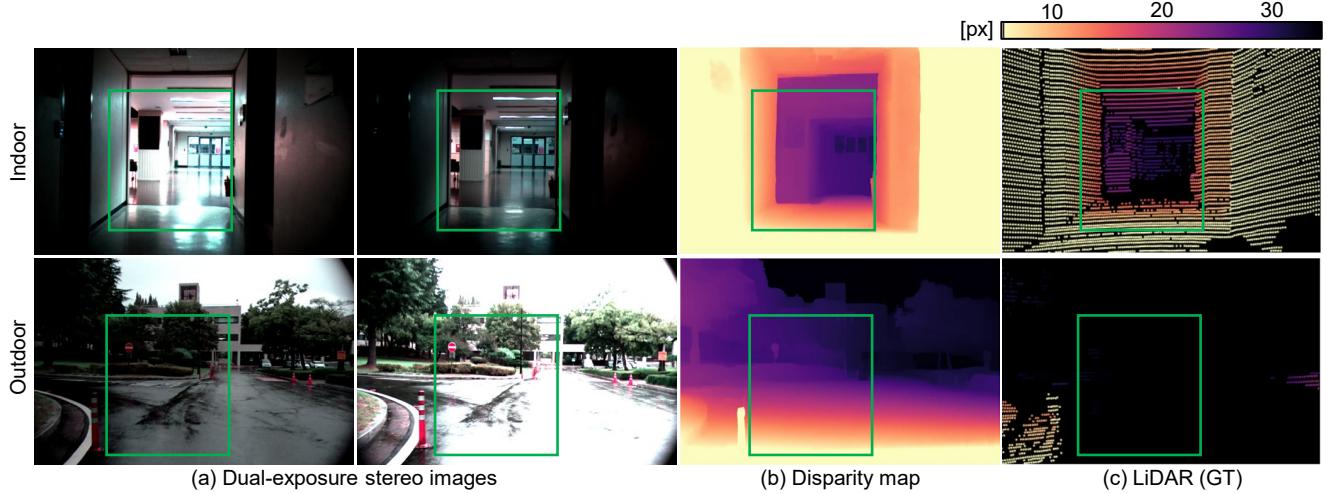


Figure 13. **Challenges with LiDAR Points in Indoor and Outdoor Scenarios.** (a) Dual-exposure stereo images for indoor and outdoor scenes. (b) Disparity maps generated by our method, showing accurate reconstruction for indoor and outdoor scenes (c) LiDAR ground-truth points, illustrating the variation in point density between indoor and outdoor scenes, particularly on wet ground in the outdoor scenario.

6.3. 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 11, 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.. 11, 12, 13, 14
- [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). 9
- [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. 9, 10
- [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. 7
- [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. 11, 12, 13, 14
- [8] Jitendra Malik Paul Debevec. 1997. Recovering High Dynamic Range Radiance Maps from Photographs. *SIGGRAPH* 29, 6, 1–10. 7
- [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. 7
- [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*. 7
- [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. 11, 12, 13, 14