Cross-spectral Gated-RGB Stereo Depth Estimation

Samuel Brucker[1] Stefanie Walz[2] Mario Bijelic[1,3] Felix Heide[1,3]

[1]Torc Robotics [2]Mercedes-Benz [3]Princeton University

Abstract

Gated cameras flood-illuminate a scene and capture the time-gated impulse response of a scene. By employing nanosecond-scale gates, existing sensors are capable of capturing mega-pixel gated images, delivering dense depth improving on today's LiDAR sensors in spatial resolution and depth precision. Although gated depth estimation methods deliver a million of depth estimates per frame, their resolution is still an order below existing RGB imaging methods. In this work, we combine high-resolution stereo HDR RCCB cameras with gated imaging, allowing us to exploit depth cues from active gating, multi-view RGB and multi-view NIR sensing – multi-view and gated cues across the entire spectrum. The resulting capture system consists only of low-cost CMOS sensors and flood-illumination. We propose a novel stereo-depth estimation method that is capable of exploiting these multi-modal multi-view depth cues, including the active illumination that is measured by the RCCB camera when removing the IR-cut filter. The proposed method achieves accurate depth at long ranges, outperforming the next best existing method by 39% for ranges of 100 to 220 m in MAE on accumulated LiDAR ground-truth. Our code, models and datasets are available here[1].

1. Introduction

Depth estimation has become a cornerstone sensing modality for 3D scene understanding in a wide range of applications such as perception and planning in autonomous driving and robotics [32, 47, 82]. Today's fully-autonomous robots mainly rely on scanning LiDAR for depth estimation [68, 71]. However, at ranges greater than 100 m, the spatial resolution of existing sensors, with a few points per pedestrian, is not sufficient for semantic understanding. Furthermore, both frequency-modulated as well as time-of-flight LiDAR systems have proven to be unreliable in the presence of backscatter [6]. While innovations in LiDAR technology such as MEMS scanning mechanisms [81] and advanced photodiode systems [77] have substantially lowered costs and enabled the development of sensors with approximately 100 to 200 scanlines, they still fall short in comparison to the vertical resolution offered by modern HDR megapixel cameras, which can exceed 10k pixels. Wide-baseline RGB stereo depth estimation methods overcome this issue by providing depth maps at image resolution, but struggle in low-light scenes and texture-less regions. Recently, gated imaging [3, 7, 10, 11, 23, 29] has emerged as a potential alternative sensor modality for 3D detection and depth estimation, offering the capability to overcome low LiDAR-resolution, while providing comparable accuracy [24, 78, 80]. Operating in the near-infrared spectrum, gated imaging systems combine CMOS sensors with active flash illumination and analogue gated readout. This approach is robust to low-light and adverse weather conditions [7]. For depth prediction, Gated2Depth [24] employs three gated slices in a neural network which is trained via a combination of simulation and LiDAR supervision. Following this, Walia et al. [78] proposed a self-supervised training approach resulting in higher-quality depth maps. Walz et al. [80] recently introduced Gated Stereo, employing a wide-baseline stereo-gated configuration for depth estimation. These methods outperform scanning LiDAR systems in depth resolution, precision, and robustness to backscatter in fog, rain and snow. While these methods successfully outperform LiDAR in depth sensing, they are constrained by the gated imager's megapixel resolution and lack of color information. This results in diminished details, particularly noticeable at long distances. RGB-only depth methods yield high-resolution depth maps, but these are not metric and lack the precision of LiDAR-based depth measurements.

In this work, we close this gap by proposing a low-cost CMOS-only sensing method that combines multi-view RGB sensing with gated cameras, exploiting active and multi-view cues across the visible and NIR spectrum. Specifically, we propose a NIR gated camera in conjunction with a HDR RCCB camera without an IR-cut filter present. RCCB cameras incorporate clear channel filters where conventional RGGB Bayer color filters feature the green channel, which enhances their sensitivity in low-light conditions. This joint approach allows us to use the spectral overlap for estimating high-resolution depth maps at RCCB-camera resolution of 8 megapixels, an order of magnitude higher than the gated imager resolution. Previous works have recognized the capabilities of cross-spectral imaging due to

[1] https://light.princeton.edu/gatedrccbstereo/

the complementary information coming from different sensor modalities [8, 9, 28, 66]. For depth estimation, however, combining images from different spectra has proven to be difficult due to the differing appearance of the images [55, 75, 92]. Our approach combines two multi-view stereo views across the spectrum and an active illuminator (visible by both) by fusing the features of both modalities of the respective viewpoints. Specifically, to recover depth, we propose a stereo depth estimation method that incorporates a novel cross-spectral fusion module which leverages intermediate depth outputs for accurate registration of feature maps from both modalities, a pose refinement step and attention-based feature fusion. The merged features encompass complementary data from both spectra, enabling their use in the stereo network to generate accurate depth maps in any lighting conditions.

We validate our method on automotive driving data in urban, suburban and highway environments in varying illumination, and we find that the method compares favorably to existing active and hybrid methods. We also demonstrate that the high-resolution depth enables new applications, such as detecting small lost cargo objects in high-way scenarios that cannot be resolved by conventional methods.

Specifically, we make the following contributions:

- We propose a novel cross-spectral depth estimation approach that recovers high-resolution dense depth maps from multi-view and time-of-flight depth cues across the visible and NIR spectrum.
- We introduce a novel cross-modal stereo network that jointly estimates the depth from passive and active RCCB and gated features and a semi-supervised training scheme to train the estimator.
- We validate that the method produces accurate depth maps on accumulated LiDAR point-clouds up to 220 m, outperforming existing methods by 39% in MAE for long ranges $\geq$ 100 m. We show that these high-resolution depth estimates enable new applications such as lost cargo detection.

2. Related Work

Depth Estimation from Monocular and Stereo Intensity Images. Depth estimation from intensity images has been thoroughly investigated using various modalities, from single-image captures [21, 26, 51, 52] to stereo images [4, 13, 53, 88] and cross modal representations using intensity images augmented with sparse LiDAR data [16, 90]. Further refinement techniques were introduced, enhancing the predicted depth maps and increasing resolution [2, 60, 63, 91]. Existing work has investigated various loss formations [17, 21, 22, 26, 56, 57, 65, 76, 89], neural architectures [4, 20, 22, 26, 51, 53, 88] and introduced consistencies [20, 21]. To exploit large unlabeled datasets, self-supervised approaches [20–22, 26, 93] exploiting stereo- [20, 21] and temporal-consistencies [22, 26, 93]. Unfortunately, these methods do not resolve the need for dense depth ground-truth for high-quality depth estimation [13, 13, 19, 34, 38, 51, 52, 58, 59]. To this end, existing methods rely on sparse LiDAR measurements as ground-truth. However, using LiDAR measurements as direct inputs [16, 31, 61, 72, 73, 84, 90] for both supervised training and inference can result propagating temporal LiDAR distortions and scan pattern artifacts.

Depth from Time-of-Flight. Unlike depth estimation from intensity images, Time-of-Flight (ToF) sensors determine depth by measuring the time it takes for emitted light to return to the detector. Acquisition approaches can be classified into correlation ToF cameras [27, 40, 41], pulsed ToF sensors [68], and gated illumination with wide depth measurement bins [23, 29]. Correlation ToF cameras use flood illumination to gauge depth from the phase difference between emitted and received light pulses, offering high spatial depth resolution [27, 40, 41]. However, these sensing modalities struggle in outdoor environments due to sensitivity to ambient light. Pulsed ToF sensors measure the round-trip time of a single light pulse to a scene point, yielding high-depth accuracy [68], however, rely on scanning that compromises spatial resolution. Moreover, these sensors degrade in fog or snow because of backscatter [6, 12, 36]. Gated cameras combine high resolution CMOS imagers with microsecond exposure times, integrating pulsed flood-illumination with adjustable delays. Through this temporal gating, backscatter is effectively reduced [7], and coarse depth is reconstructed [3, 10, 11]. Extracting more refined depth initially focused on analytical methods [42, 43, 85], Bayesian methods [1, 67] and deep neural networks [24, 78] excel in low-light and outdoor scenarios. Gruber et al. [24] predict depth using a reconstruction network rivaling conventional stereo models, while Walia et al. [78] proposed a refined self-supervised method. Later, Walz et al. [80] combined two gated imagers, optimizing depth estimation through multi-view cues. All of these methods are designed for gated imagers only, compromising resolution compared to RGB imagers and in scenarios when the NIR laser power is low compared to ambient light. We lift this limitation by combining NIR gated cameras with high-resolution visible-spectrum RCCB sensors.

Cross-Spectral Matching Conventional stereo matching algorithms assume match based on the brightness constancy assumption. However, using multiple sensors, operating in distinct spectral ranges, has been investigated as an additional source of information. Progress was reported in areas such as Face Recognition [37, 46, 49], self-driving cars [33, 86], visual surveillance [44], and smartphones [74]. Existing methods have proposed methods for matching features that may be visually distinct but remain semantically

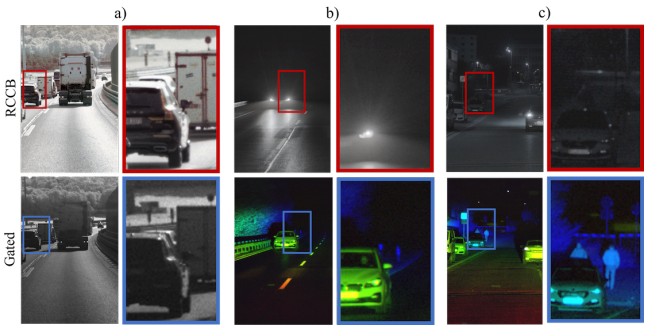

Figure 1. RCCB cameras (top row) capture 8 Mpix passive RGB images. Gated cameras (bottom row) record Time-of-Flight data of a scene by combining active flash illumination and analog gated readout. Both sensors are complementary, with distinct strengths depending on the scenario. RCCB cameras excel in daylight (a) with high dynamic range, resolution and color. At night (b, c), gated images (gated slices here RGB-color coded by mapping each slice to one RGB color) provide strong depth cues and maintain consistent scene illumination through active illumination. This work integrates both modalities to estimate depth accurately in all ambient illumination conditions.

congruent [15,30,35,39,62,70,75,92]. Early work [62] explores gradients as a robust feature for cross-modal matching, while [35] focuses on the alignment of monochrome images, which have increased light sensitivity, with RGB images to achieve precise depth in dim lighting scenarios. Recent methods [39,55,75,79,92] aim to learn cross-modal matching, where some aim to morph one modality directly into another [79,92], while others propose novel descriptors for modality matching [39,75,92].

3. Multi-view Gated and RCCB Imaging

We propose to image a scene with a gated camera stereo system and an RCCB stereo array characterized both by a baseline of $b = 0.76$ m. The gated imager is an active sensor and emits a pulse of light with a confined wavelength around 808 nm, whereas the RCCB camera is a passive sensor with a sensitivity spectrum spanning the visible band from 380 - 1050 nm. While conventional RGB cameras use color filter arrays with an RGGB pattern, often referred to as a Bayer pattern, in RCCB cameras the green channels are replaced with clear channels. The inclusion of clear channels in this pattern allows an enhanced light sensitivity, boosting its performance $\approx 30\%$ during night-time conditions. In addition, the used Onsemi AR0820AT image sensor is optimized for both low light and challenging high dynamic range scene performance, with a 2.1 µm DR Pix BSI pixel and on-sensor 140 dB HDR capture capability.

In the stereo gated camera system, a laser pulse p is emitted at $t = 0$. Following a set time delay ξ, the reflected scene is then integrated on both camera sensors. Only photons within a specific temporal gate are captured, using the gate function g, embedding depth data into 2D imagery. As detailed by Gruber *et al.* [25], these intensities, or range-intensity-profiles $C_k(z)$, are scene-independent and can be expressed as

$$\begin{aligned} I^k(z,t) &= \alpha\, C_k(z,t), \\ &= \alpha \int_{-\infty}^{\infty} g_k(t-\xi) p_k \left(t - \frac{2z}{c}\right) \beta(z) dt, \end{aligned} \tag{1}$$

where $I^k(z,t)$ is the gated exposure at distance z and time t; α represents surface reflectance, while β accounts for atmospheric attenuation. Both image sets are calibrated and rectified for aligned epipolar lines, enabling disparity d estimation. This disparity corresponds to distance $z = \frac{bf}{d}$, offering depth insights across all slices. Ambient light sources, such as sunlight or vehicle headlights influence the gated system's operation. These photons get modulated by a constant term Λ. Separately, irrespective of ambient light, there is a dark current, D_v^k, which is dependent on the gating settings. In total we model an image with

$$I_v^k(z) = \alpha\, C_k(z) + \Lambda + D_v^k. \tag{2}$$

We follow [80], capturing additional passive HDR images with fixed exposure times of 21 µs and 108 µs during the day, and extending these to 805 µs and 1745 µs at night.

When integrating both gated and RCCB stereo systems, each camera is represented by its calibration matrix K. The relative orientation and position between cameras in a stereo pair are captured by the rotation matrix, $R \in SO(3)$, and the translation vector, $t \in \mathbb{R}^{3\times 1}$.

4. Depth from RCCB and Gated Stereo

In this section, we introduce our cross-modal fusion technique for depth prediction, which relies on multi-view cues from RCCB stereo and gated stereo images. By registering and fusing cross-spectral features through an attention mechanism and prior pose refinement within the stereo network, we capitalize complementary information from different camera modalities in Sec. 4.1. We integrate this feature fusion in a stereo network described in Sec. 4.2 which we jointly train uni-modal and multi-modal, facilitating a holistic feature representation across modalities and minimizing domain differences between modalities as detailed in Sec. 4.1. The training approach is detailed in Sec. 4.3.

4.1. Cross-Spectral Matching

We align and combine cross-modal features in a two-stage approach, where we warp features first into a shared space based on a refined pose. With these aligned features in hand, we perform an attention-based fusion as input to the remainder of the stereo estimation network. An overview of cross-spectral matching (CSM) is illustrated in Fig. 2.

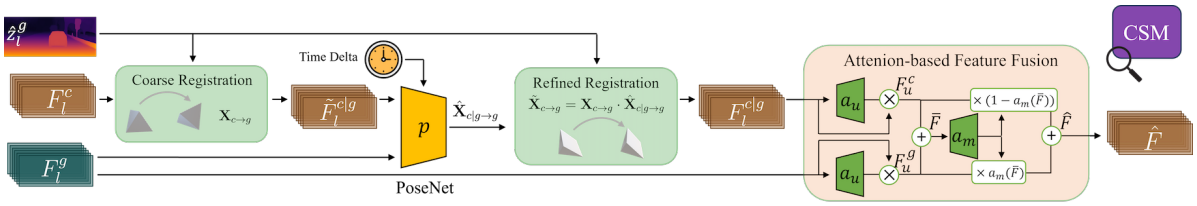

Figure 2. Cross-Spectral Matching (CSM). The layer fuses encoded features from RCCB (F_l^c) and gated (F_l^g) images. In the coarse registration step, RCCB features are aligned with gated features based on calibrated poses $X_{c\to g}$. Registration is refined based on residual pose $\hat{X}_{c|g\to g}$ estimated from coarse aligned images and measured time delta with PoseNet. Registered images are fused with attention-based fusion retaining complementary information in $\hat{F}$.

Feature Extraction and Alignment. We utilize two feature extractor backbones for color f_b^c and gated f_b^g, and share the weights for each view I_l^m, I_r^m for $m \in \{c, g\}$, that is

$$f_b^c : I_l^c, I_r^c \to F_l^c, F_r^c, \tag{3}$$
$$f_b^g : I_l^g, I_r^g \to F_l^g, F_r^g. \tag{4}$$

As a feature extractor, we use MPViT [45], a powerful vision transformer for dense prediction tasks. To align the features, we use the pose information from camera calibration $\mathbf{X}_{x\to g}$ and an intermediate depth estimation $\hat{\mathbf{z}}_l^g$ from an iterative depth estimation method, see Section 4.2, to warp corresponding views. The mapping for homogeneous coordinates x_g and x_c from I_l^g and I_l^c is defined as

$$x_g \sim K_c \mathbf{X}_{c\to g} \hat{\mathbf{z}}_l^g K_g^{-1} x_c, \tag{5}$$

where K_c and K_g are the camera matrices of the gated and RCCB camera, and $\mathbf{X}_{c\to g} = \left(\begin{smallmatrix} R_{c\to g} & t_{c\to g} \\ 0 & 1 \end{smallmatrix}\right)$ with $R_{c\to g} \in SO(3)$ and $t_{c\to g} \in \mathbb{R}^{3\times 1}$. We transform the features of the left RCCB camera, denoted F_l^c, to match the features of the left gated camera F_l^g, thus creating $\tilde{F}_l^{c|g}$.

Pose Refinement. The RCCB stereo camera and the gated stereo camera are independently synchronized to microsecond precision. However, the RCCB camera may accumulate a slight offset of up to 20 milliseconds between images because of automatic exposure and shutter timing. To address this misalignment, we utilize a lightweight Convolutional Neural Network (CNN) framework dubbed PoseNet p. This framework estimates the rotational and translational adjustments necessary for alignment, based on prealigned feature maps. The input to p is the concatenated context F_l^g, the transformed $\tilde{F}_l^{c|g}$ and the measured time offset t between modalities. The time is integrated into every downsampled layer of the pose network as additional channel, except for the final layer. This channel uniformly replicates the value of the time-offset across the spatial dimension. The computed pose update, denoted as $\hat{\mathbf{X}}_{c|g\to g}$ combines the initial pose as $\tilde{\mathbf{X}}_{c\to g} = \mathbf{X}_{c\to g} \cdot \hat{\mathbf{X}}_{c|g\to g}$. Subsequently, a second warping operation with the mapping

$$x_g \sim K_c \tilde{\mathbf{X}}_{c\to g} \hat{\mathbf{z}}_l^g K_g^{-1} x_c, \tag{6}$$

is applied, which generates the aligned features $F_l^{c|g}$.

Attention-based Feature Fusion. Following the alignment, we fuse RCCB and gated features, aiming to combine contextual information from both spectra effectively. Our approach adopts a two-step process. Firstly, we employ channel self-attention for aggregating both global and local contexts within feature maps. Secondly, we combine the individual feature maps, utilizing the predicted attention weights. The first setup is defined as

$$\bar{F} = \frac{F_u^g \oplus F_u^c}{a_u(F_l^g) + a_u(F_l^c)} \tag{7}$$
$$F_u^g = F_l^g \otimes a_u(F_l^g) \tag{8}$$
$$F_u^c = F_l^{c|g} \otimes a_u(F_l^{c|g}), \tag{9}$$

where $\oplus$ denotes element-wise addition and $\otimes$ indicates element-wise multiplication, and the attention $a_u()$ is calculated following [18]. The final fusion of features $\hat{F}$ are the result of the following weighting-operation

$$\hat{F} = \left(F_u^g \otimes a_m(\bar{F})\right) \oplus \left(F_u^c \otimes (1 - a_m(\bar{F}))\right), \tag{10}$$

where a_m follows the implementation as in [18] and denotes the multi-modal attention network, facilitating the effective combination of features from both modalities.

4.2. Stereo Matching

With the process to align features $\hat{F}$ in hand, we predict depth across all camera views. This task is executed through a stereo matching network, as depicted in Figure 3. We build on top of the CREStereo architecture [48] with major modifications to allow the development of a dynamic framework switching between modalities. This dynamic interchangeability allows us to adapt and optimize the disparity prediction in either modalities coordinate system. Such flexibility not only enhances domain generalization but also opens avenues for the application of various consistency losses, thereby improving the accuracy of our predictions.

To bridge the coordinate system we heavily rely on the CSM layers, whose predicted context feature maps are used to guide the prediction in the targeted frames. Thereby, we rely on the iterative refinement introduced in [48] and calculate the correlation volume in each step according to [48]. Here, we predict the correlation in the adaptive group correlation layers uni-modal and alternate in modality through

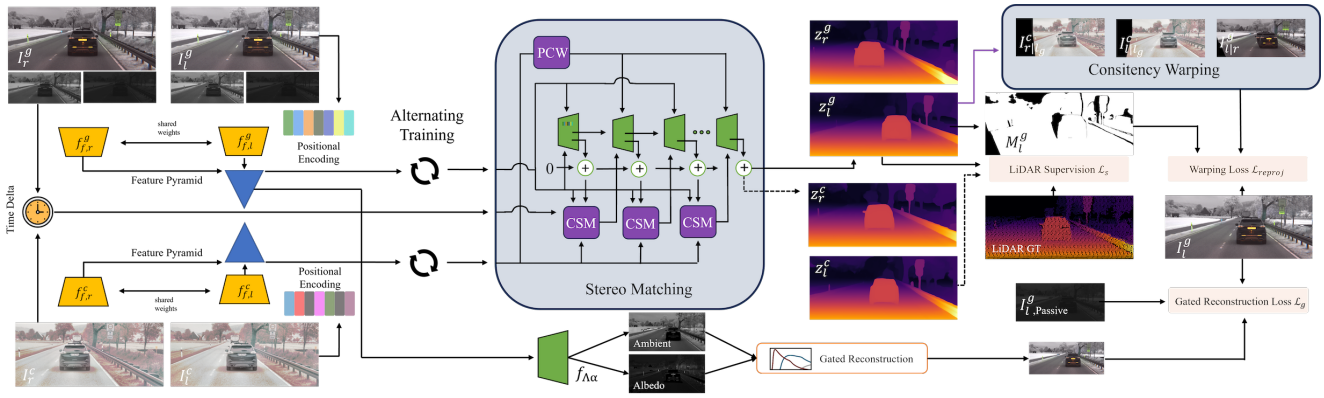

Figure 3. The proposed cross-spectral stereo architecture for depth estimation from stereo RCCB and stereo gated images incorporating our CSM layer. The network can output depth for all four input images. Intermediate depth estimates are used for iterative fusion within the CSM along the depth estimation process. The network is trained with self-supervision (Left-Right consistency for RCCB and gated images, Gated Reconstruction) and LiDAR supervision.

the iterative refinement. The secondary modality is projected into the target frame with the refined transformation $\tilde{\mathbf{X}}_{c\to g}$ in the pre-correlation warping PCW, see Fig. 3.

Then the correlation is calculated as follows,

$$\text{Corr}(x,y,k) = \frac{1}{C}\sum_{i=1}^{C} F_l^v(i,x,y)F_r^v(i,x',y'), \qquad (11)$$

where F^v is the respective feature map transformed into the modality $v \in \{c, g, c|g, g|c\}$, with camera view l, r. We follow [48] and predict $x' = x + f(k)$, $y' = y + g(k)$, with fixed offsets $f(k)$ and $g(k)$ for the k-th correlation pair, sum all channels C and apply 2D-1D alternate local search strategy for computational efficiency. Notably, the initial iteration at the coarsest scale focuses on predicting depth solely from the target modality.

4.3. Training Supervision

The network is trained to output the disparity d which is converted into the depth z for all modalities g, c and views l, r. In addition we train the stereo matching uni-modal and multi-modal, with and without cross-spectral feature enhancement to ensure optimal extracted features while sharing the stereo matching stage. This is achieved by deactivating the CSM and PCW layers. Through this alternating training, we ensure that the backbone learns relevant features for all modalities and the mix and matching between modalities forces all features to be domain independent, thereby creating robust cross-modal representations. Further this allows us to implement self-supervised and supervised loss functions for both the gated camera and the RCCB camera, as well as consistencies in between.

All self-supervised consistency losses and supervised losses are described below. Without diminishing generality, in the following all losses are defined for disparity prediction in the gated frame for better readability.

Left-Right Reprojection Consistency. The projection loss enforces the photometric consistency between the left and right camera views within each modality. Cross-modally the homogeneity between predicted depth maps is enforced. The total loss for the left gated camera g_l can be written as,

$$\mathcal{L}_w^{g_l} = \mathcal{L}_p(I_l^g, I_{r|l_g}^g) + \mathcal{L}_p(I_{l|l_g}^c, I_{r|l_g}^c) + \mathcal{L}_p(z_l^{c|g}, z_l^g), \qquad (12)$$

with $I_{r|l_g}^g$ the r right g gated image warped into the l left gated view using the predicted depth z_l^g denoted as warping operation $_{l_g}$ for the stereo pairs. For the gated warping consistency further the RCCB frames I^C are warped according to the predicted depth into the gated frame $_{l_g}$. Additionally, the predicted depth in c is transformed to the gated frame g leading to $z_l^{c|g}$. Consistencies are also applicable to the right gated frame, yielding $\mathcal{L}_w^{g_r}$, and to both left $\mathcal{L}_w^{c_l}$ and right $\mathcal{L}_w^{c_r}$ RCCB frames. The total loss can be written as $\mathcal{L}_{reproj} = \mathcal{L}_w^{c_l} + \mathcal{L}_w^{c_r} + \mathcal{L}_w^{g_l} + \mathcal{L}_w^{g_r}$. Note, $\mathcal{L}_p$ follows [21] and is a similarity loss based on the structural similarity (SSIM) metric [83] and the L_1 norm, $\mathcal{L}_p(a,b) = 0.85\frac{1-SSIM(a,b)}{2} + 0.15\|a-b\|_1$.

Gated Reconstruction Loss. To supervise the embedded time of flight information in the gated slices we adopt the cyclic gated reconstruction loss from [78], which uses measured range intensity profiles to reconstruct the input gated images from the predicted depth z, the albedo $\tilde{\alpha}$, and the ambient $\tilde{\Lambda}$. Departing from [78] who employed measured profiles, we employ an analytical gating model. We estimate the albedo $\tilde{\alpha}$ and ambient $\tilde{\Lambda}$ through an additional context encoder taking the feature pyramid as input, see Figure 3, and model a gated slice as $\tilde{I}^k(z) = \tilde{\alpha}\, C_k(z) + \tilde{\Lambda}$. The loss term incorporates both per-pixel difference and structural similarity, following

$$\mathcal{L}_{recon} = \mathcal{L}_p(M_g \odot \tilde{I}^k(z), M_g \odot I^k) + \mathcal{L}_p(\tilde{\Lambda}, \Lambda^{k_0}), \qquad (13)$$

with M_g as per-pixel SNR consistency mask [78].

LiDAR Supervision. We supervise final and interme-

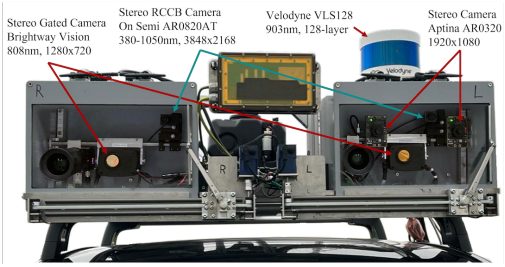

Figure 5. The sensor setup of the test vehicle used for capturing the dataset. It features a stereo gated camera, consisting of a floodlight flash source (not visible, mounted at front bumper of the car) and two gated imagers, a Velodyne VLS128 scanning lidar, a standard stereo RGB camera and the RCCB stereo camera.

diate disparity predictions. Each disparity prediction $\{\mathbf{d}_{l,i}, ..., \mathbf{d}_{l,n}\}$ is upsampled to full resolution and compared to ground-truth with a weighted combination of l_1 and l_2 defined as,

$$\mathcal{L}_{lidar} = \sum_{i=1}^{n} \gamma^{n-i} \left(\frac{2}{3} ||d_{gt} - M \odot d_l||_1 + \frac{1}{3} ||d_{gt} - M \odot d_l||_2^2 \right). \tag{14}$$

The weight γ is set to 0.9 and the mask M excludes areas without ground-truth. For ground-truth we use accumulated and sparse LiDAR measurements, more details in Section 5.

Overall Training Loss. The following loss term is obtained by combining all self-supervised and supervised loss components from above,

$$\mathcal{L}_{stereo} = c_1 \mathcal{L}_{reproj} + c_2 \mathcal{L}_{recon} + c_3 \mathcal{L}_{lidar}, \tag{15}$$

which we combine with scalar weights $c_{1,...,3}$ provided in the Supplemental Material.

Implementation Details. We refer to the Supplemental Document for implementation details, training settings, and hyperparameter settings used for the approach described.

5. Dataset

For training and testing, we use the dataset introduced by Walz et al. [80]. The dataset includes stereo gated, stereo RGB and ground-truth LiDAR data. More information is given in the Supplemental Material. In this work, we extend the dataset with RCCB stereo data, captured with an AR0820 sensor. All sensors were housed in a portable sensor cube as showcased in Figure 5. As an additional source of ground-truth, we utilize a densely constructed LiDAR map, derived from a custom adaptation of the LIO-SAM algorithm, as detailed in Shan et al. [69]. We refer to the Supplemental Document for details on the setup and dataset.

6. Assessment

In this section, we experimentally validate our proposed method. We examine the accuracy of our depth estimation under nighttime and daytime conditions and compare it to existing depth estimation techniques. Furthermore, we validate our design choices through a series of ablation studies.

	Method	Modality	Train	RMSE [m]	ARD	MAE [m]	δ_1 [%]	δ_2 [%]	δ_3 [%]
	Test Data – Night (Evaluated on LiDAR Ground-Truth Points)								
Comparison to state-of-the-art	Gated2Depth [24]	Mono-Gated	D	16.15	0.17	8.07	75.70	92.74	96.47
	Gated2Gated [78]	Mono-Gated	MG	14.08	0.19	7.95	79.84	92.95	96.59
	PackNet [26]	Mono-RGB	M	17.82	0.20	10.21	66.35	87.85	95.61
	Monodepth2 [22]	Mono-RGB	M	18.44	0.18	9.47	75.70	90.46	95.68
	SimIPU [50]	Mono-RGB	D	15.78	0.18	8.71	76.25	90.84	96.44
	AdaBins [5]	Mono-RGB	D	14.45	0.15	7.58	81.47	93.75	97.39
	DPT [64]	Mono-RGB	D	12.15	0.12	6.31	85.38	95.94	98.42
	DepthFormer [51]	Mono-RGB	D	12.15	0.11	6.20	85.18	95.76	98.47
	PSMNet [14]	Stereo-RGB	D	27.98	0.27	16.02	50.77	74.77	85.93
	STTR [54]	Stereo-RGB	D	20.99	0.19	11.14	70.84	87.70	93.46
	HSMNet [88]	Stereo-RGB	D	12.42	0.09	5.87	88.41	96.08	98.50
	ACVNet [87]	Stereo-RGB	D	11.70	0.08	5.25	89.91	96.33	98.47
	RAFT-Stereo [56]	Stereo-RGB	D	10.89	0.09	5.10	90.47	96.71	98.64
	CS-Stereo [92]	RCCB-NIR	D	21.35	0.20	11.48	72.73	89.71	95.58
	UCSSM [55]	RCCB-NIR	D	18.22	0.27	14.63	64.51	87.12	94.27
	CREStereo [48]	Stereo-RCCB	D	12.05	0.10	5.18	88.48	94.12	97.26
	Gated Stereo [80]	Stereo-Gated	DGS	6.39	0.05	2.25	96.40	98.44	99.24
	Gated RCCB Stereo	Stereo-RCCB-Gated	DGS	**6.23**	**0.04**	**2.03**	**96.69**	**98.50**	**99.26**
	Test Data – Day (Evaluated on LiDAR Ground-Truth Points)								
Comparison to state-of-the-art	Gated2Depth [24]	Mono-Gated	D	28.68	0.22	14.76	66.68	82.76	87.96
	Gated2Gated [78]	Mono-Gated	MG	16.87	0.21	9.51	73.93	92.15	96.10
	PackNet [26]	Mono-RGB	M	17.69	0.21	9.77	72.12	90.65	96.51
	Monodepth2 [22]	Mono-RGB	M	20.78	0.22	10.06	79.05	90.66	94.69
	SimIPU [50]	Mono-RGB	D	14.33	0.14	7.50	81.77	94.01	97.92
	AdaBins [5]	Mono-RGB	D	12.76	0.12	6.53	86.15	95.77	98.41
	DPT [64]	Mono-RGB	D	11.29	0.09	5.52	89.56	96.83	98.79
	DepthFormer [51]	Mono-RGB	D	10.59	0.09	5.06	90.65	97.46	99.02
	PSMNet [14]	Stereo-RGB	D	32.13	0.28	18.09	53.82	74.91	84.96
	STTR [54]	Stereo-RGB	D	16.77	0.16	8.99	78.44	93.53	98.01
	HSMNet [88]	Stereo-RGB	D	10.36	0.08	4.69	92.47	97.93	99.11
	ACVNet [87]	Stereo-RGB	D	9.40	0.07	4.08	94.61	98.36	99.12
	RAFT-Stereo [56]	Stereo-RGB	D	9.40	0.07	4.07	93.76	98.15	99.09
	CS-Stereo [92]	RCCB-NIR	D	21.51	0.22	11.87	73.70	88.77	96.06
	UCSSM [55]	RCCB-NIR	D	17.32	0.29	13.26	64.80	84.78	93.83
	CREStereo [48]	Stereo-RCCB	D	9.68	0.06	3.88	95.02	96.04	98.57
	Gated Stereo [80]	Stereo-Gated	DGS	7.11	0.05	2.25	96.87	98.46	99.11
	Gated RCCB Stereo	Stereo-RCCB-Gated	DGS	**6.89**	**0.03**	**1.95**	**97.18**	**98.55**	**99.18**

Table 1. Evaluation of the method and competing gated approaches on [80]. We compare our model to supervised and unsupervised approaches. "M" refers to methods that use temporal data for training, S for stereo supervision, "G" for gated consistency and "D" for depth supervision. Best results in each category are in **bold** and second best are underlined.

	Evaluation Range	0 - 160 m		0 - 220 m		100 - 220 m	
	Method	RMSE	MAE	RMSE	MAE	RMSE	MAE
Night	CREStereo [48]	13.58	8.60	17.64	10.05	26.39	20.24
	Gated Stereo [80]	11.45	7.36	14.03	8.93	25.55	18.36
	Gated RCCB Stereo	**10.74**	**7.02**	**12.02**	**7.94**	**15.67**	**11.15**
Day	CREStereo [48]	11.16	6.53	15.65	8.11	20.76	14.65
	Gated Stereo [80]	10.75	6.42	14.24	8.67	22.07	16.79
	Gated RCCB Stereo	**9.72**	**6.24**	**10.69**	**6.83**	**14.33**	**10.07**

Table 2. Evaluation on *Accumulated LiDAR Scans*. We compare our method to the top 3 methods from Tab. 1 using accumulated dense LiDAR as ground-truth for a range from 0 - 220 m.

Experimental Setup. The test set comprises 2463 frames, split into 1269 daytime and 1194 nighttime frames. Each frame is accompanied by high-resolution LiDAR ground-truth measurements, capturing reliable data up to 160 m. We further use the 655 frames of refined LiDAR ground-truth (303 daytime and 352 nighttime). These frames feature accumulated point clouds, allowing us to assess the methods on a dense ground-truth for accuracy up to a distance of 220 m. Our method's evaluated depth maps show

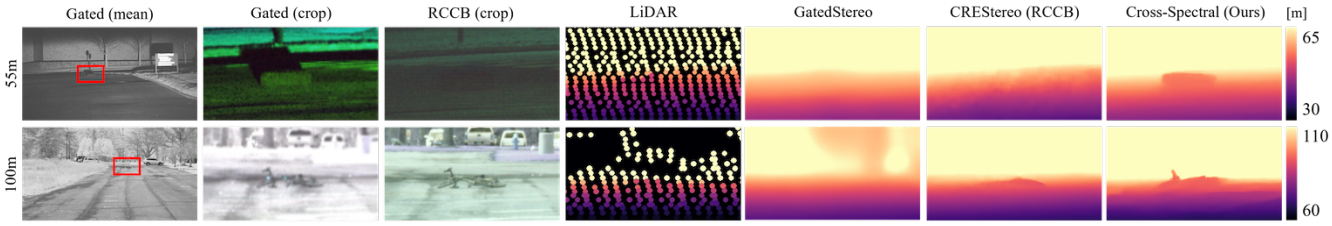

Figure 4. Depth estimation for "lost cargo", small objects at far distances on ground level that may be lost from preceding vehicles. Our method estimates accurate depth for these small objects in both daylight and nighttime conditions by integrating complementary RCCB and gated images. Single modality methods suffer from limitations: CREStereo [48] (RCCB) lacks effective illumination at night, and Gated Stereo [80] suffers from poor resolution during the day.

	Modality	Full Res.	CS Training	Pose Ref.	Att. Fusion	MPViT Backb.	RMSE [m]	MAE [m]	δ_1 [%]	δ_2 [%]	δ_3 [%]
Test Data – Night (Evaluated on LiDAR Ground-Truth Points)											
ABLATION	Stereo-RCCB-Gated	✓	✓	✓	✓	✓	**6.23**	**2.03**	**96.69**	**98.50**	**99.26**
ABLATION	Stereo-RCCB-Gated	✗	✓	✓	✓	✓	6.53	2.04	96.37	98.45	99.24
ABLATION	Stereo-RCCB-Gated	✗	✗	✓	✓	✓	6.87	2.18	96.20	98.24	99.13
ABLATION	Stereo-RCCB-Gated	✗	✗	✗	✓	✓	6.98	2.23	96.01	98.21	99.11
ABLATION	Stereo-RCCB-Gated	✗	✗	✗	✗	✓	7.23	2.42	95.89	98.20	99.10
ABLATION	Stereo-RCCB-Gated	✗	✗	✗	✗	✗	8.17	2.74	95.23	97.79	98.89
ABLATION	RCCB-Gated	✗	✗	✗	✗	✗	10.56	7.89	45.23	79.49	91.14
ABLATION	Mono-Gated	✗	✗	✗	✗	✗	10.87	4.70	89.91	95.77	97.90
Test Data – Day (Evaluated on LiDAR Ground-Truth Points)											
ABLATION	Stereo-RCCB-Gated	✓	✓	✓	✓	✓	**6.89**	1.95	**97.18**	**98.55**	**99.18**
ABLATION	Stereo-RCCB-Gated	✗	✓	✓	✓	✓	7.09	**1.93**	97.03	98.46	99.11
ABLATION	Stereo-RCCB-Gated	✗	✗	✓	✓	✓	7.57	2.12	96.62	98.35	99.04
ABLATION	Stereo-RCCB-Gated	✗	✗	✗	✓	✓	7.64	2.16	96.37	98.52	99.06
ABLATION	Stereo-RCCB-Gated	✗	✗	✗	✗	✓	7.92	2.29	96.50	98.25	98.00
ABLATION	Stereo-RCCB-Gated	✗	✗	✗	✗	✗	8.17	2.44	96.46	98.12	98.92
ABLATION	RCCB-Gated	✗	✗	✗	✗	✗	8.61	4.73	67.33	89.75	96.30
ABLATION	Mono-Gated	✗	✗	✗	✗	✗	13.71	6.05	88.99	95.56	97.71

Table 3. Ablation Experiments on the dataset from [80]. We investigate different resolution, training method, remove components of our proposed CSM, the MPViT [45] backbone and test different input modalities.

the perspective of the left gated camera and match the resolution of the RCCB images.

Our evaluation metrics are in line with those established in [19]. We use Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Absolute Relative Difference (ARD), and the threshold accuracy metric $\delta_i < 1.25^i$ for $i \in 1, 2, 3$. All methods in our evaluation have been fine-tuned on our dataset for a fair comparison.

Depth Reconstruction. Qualitative results are presented in Figure 7 and quantitative results in Table 1. Here, we compare against three recent gated [24, 78, 80], six monocular RGB [5, 22, 26, 50, 51, 64], six stereo RGB [14, 48, 54, 56, 87, 88] and two cross-spectral stereo [55, 92] methods. Compared to the next best stereo method, Gated Stereo [80], our method reduces the error by 9.7 % and 0.22 m in Mean Absolute Error (MAE) during nighttime conditions and by 13.3 % and 0.3 m during day conditions. Additionally, we compare our method to the two next-best stereo methods [48, 80] on accumulated LiDAR ground-truth maps which allow assessment up to 220 m in Table 2. Our method reduces the error of the next best method averaged over day and night, Gated Stereo [80], by 16.1 % and 1.4 m and CREStereo [48] by 17.1 % and 1.7 m. For distances between 100 and 220 m, our method achieves an improvement of 39.6 % over [80] and 39.2 % over [48], demonstrating a considerable improvement at long distances. Note that [80] is designed for distances up to 160 m only. Qualitatively, this improvement is visible in sharper edges and rendering of fine details missed by other methods. Compared to the two next-best methods [48, 80], the benefits of our cross-spectral depth estimation are highlighted for fine structures at large distances, see Fig. 6. Compared to alternative cross-spectral stereo methods like CS-Stereo [92], our method is visually and quantitatively superior by a wide margin of 83.0 % as these methods generally don't display details.

Qualitative Assessment of Lost Cargo Data Our study includes a qualitative comparison of depth estimation methods, focusing on detecting small, potentially hazardous highway objects (as shown in Figure 4, with more examples in the Supplemental Material). This is critical for autonomous driving, where early detection of such "lost cargo" is necessary for safe maneuvering. Traditional LiDAR often lacks the necessary depth detail for small objects, while passive RCCB cameras are effective in daylight but less so in low light. Gated cameras, although useful, struggle in bright conditions and have lower resolution. Our analysis highlights that high-resolution RCCB data and precise time-of-flight gated data combined with our cross-spectral gated stereo surpass single-modality sensors in detecting small objects at long distances, an essential capability for advanced autonomous driving systems.

Ablation Experiments. Next, we evaluate the effectiveness of our method by progressively removing components from the full model, see Table 3. We start with the full model, achieving the overall best metrics averaged over day and night. First, we downsample the RCCB image to a third of its original height and width, effectively setting the resolution of the depth map to be similar to the resolution of the gated image. This leads to a visible reduction of details in the depth map which cannot be measured quantitatively using sparse LiDAR. To assess the effectiveness of our cross-spectral (CS) training approach, encompassing alternating training, self-supervised losses, and dense LiDAR supervision, we then remove this component, training with sparse LiDAR supervision only, which results in an increase in MAE by 8.4%. Further simplification involves omitting

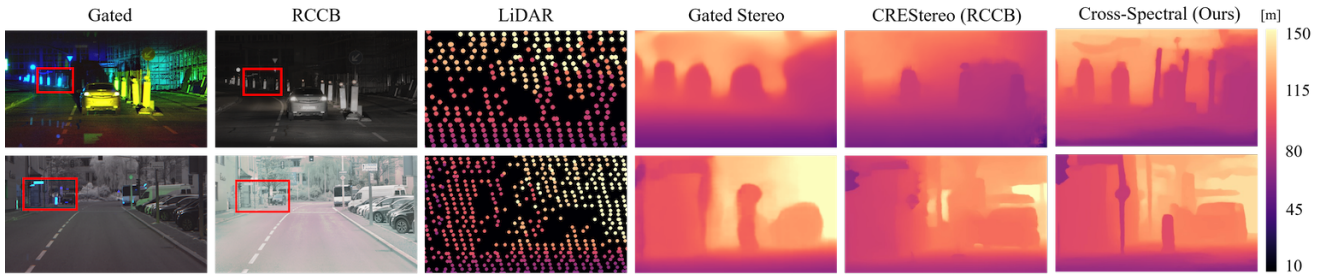

Figure 6. Comparison of our method to LiDAR and the best state-of-the-art methods that rely only on a single modality: Gated Stereo (gated images) [80] and CREStereo (RCCB images) [48]. Our method recovers fine details of distant objects irrespective of daylight and nighttime. Limitation of depth range in colorized depth maps for visualization purposes only.

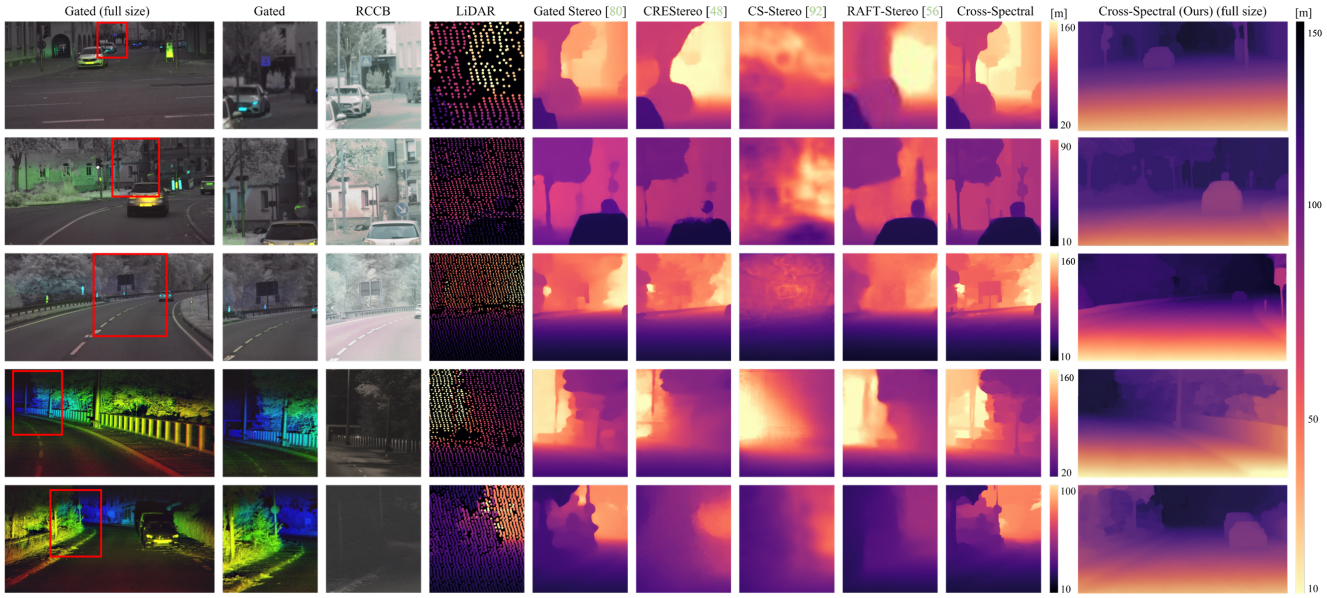

Figure 7. Qualitative comparison of our Gated-RCCB Stereo and existing methods. Our approach is unique in its ability to produce consistently accurate and high-detail depth maps regardless of the ambient illumination condition. In our depth maps, fine structures such as trees or poles are clearly visible, unlike other methods that struggle with consistent depth prediction for these elements. For enhanced visibility of distant objects, the color maps used in zoom-ins are inverted and scaled.

the pose-refinement step within our proposed CSM. This step, too, causes an increase in MAE by 2.1%, indicating the effectiveness of these components in our method. Subsequent removal of the attention-based feature fusion mechanism and replacing MPViT backbone with the backbone from [48] shows an additional decrease in MAE by 7.3% and 9.9%, respectively. Next, we analyze the impact of the dual-camera setup, comprising one RCCB and one gated camera. Discarding this setup leads to more than double the MAE, highlighting the importance of the double stereo camera configuration. Finally, we revert to a monocular depth estimation baseline, which records the highest daytime MAE, highlighting the value of stereo cues.

7. Conclusion

In this study, we devise a novel cross-spectral method for stereo depth estimation, combining active gated NIR and high-resolution HDR RCCB cameras. This approach outperforms existing LiDAR sensors in spatial resolution without compromising depth accuracy. Our method is effective in varying lighting conditions, with gated NIR excelling at night and RCCB cameras in daylight. To combine both modalities, we propose a stereo depth estimation method that hinges on a new cross-spectral fusion module trained both supervised and self-supervised losses. Economically viable, our system employs cost-effective CMOS sensors, achieving depth with superior accuracy and quality, surpassing existing methods that rely on single modalities by 39% in MAE at long ranges. This enables novel applications like long-distance detection of small ground-level objects.

Acknowledgments This work was supported by the AI-SEE project with funding from the FFG, BMBF, and NRC-IRA. Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award, and an Amazon Science Research Award.

References

[1] Amit Adam, Christoph Dann, Omer Yair, Shai Mazor, and Sebastian Nowozin. Bayesian time-of-flight for realtime shape, illumination and albedo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):851–864, 2017. 2

[2] Filippo Aleotti, Fabio Tosi, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Neural disparity refinement for arbitrary resolution stereo. In *2021 International Conference on 3D Vision (3DV)*, pages 207–217. IEEE, 2021. 2

[3] Pierre Andersson. Long-range three-dimensional imaging using range-gated laser radar images. *Optical Engineering*, 45(3):034301, 2006. 1, 2

[4] Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, and Orazio Gallo. Bi3D: Stereo depth estimation via binary classifications. In *arXiv preprint arXiv:2005.07274*, 2020. 2

[5] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 6, 7

[6] Mario Bijelic, Tobias Gruber, and Werner Ritter. A benchmark for lidar sensors in fog: Is detection breaking down? In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 760–767, 2018. 1, 2

[7] Mario Bijelic, Tobias Gruber, and Werner Ritter. Benchmarking image sensors under adverse weather conditions for autonomous driving. In *IEEE Intelligent Vehicle Symposium*, 2018. 1, 2

[8] Thirimachos Bourlai, Arun Ross, Cunjian Chen, and Lawrence Hornak. A study on using mid-wave infrared images for face recognition. In *Sensing Technologies for Global Health, Military Medicine, Disaster Response, and Environmental Monitoring II; and Biometric Technology for Human Identification IX*, volume 8371, pages 239–251. SPIE, 2012. 2

[9] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR 2011*, pages 177–184. IEEE, 2011. 2

[10] Jens Busck. Underwater 3-D optical imaging with a gated viewing laser radar. *Optical Engineering*, 2005. 1, 2

[11] Jens Busck and Henning Heiselberg. Gated viewing and high-accuracy three-dimensional laser radar. *Applied Optics*, 43(24):4705–10, 2004. 1, 2

[12] A. Carballo, J. Lambert, A. Monrroy, D. Wong, P. Narksri, Y. Kitsukawa, E. Takeuchi, S. Kato, and K. Takeda. Libre: The multiple 3d lidar dataset. In *IEEE Intelligent Vehicles Symposium (IV)*, 2020. 2

[13] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 2

[14] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 6, 7

[15] Wei-Chen Chiu, Ulf Blanke, and Mario Fritz. Improving the kinect by cross-modal stereo. 01 2011. 3

[16] Jaesung Choe, Kyungdon Joo, Tooba Imtiaz, and In So Kweon. Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation. *IEEE Robotics and Automation Letters*, 6(3):4672–4679, 2021. 2

[17] Qi Dai, Vaishakh Patil, Simon Hecker, Dengxin Dai, Luc Van Gool, and Konrad Schindler. Self-supervised object motion and depth estimation from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1004–1005, 2020. 2

[18] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3560–3569, 2021. 4

[19] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374, 2014. 2, 7

[20] Ravi Garg, B.G. Vijay Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Proceedings of the IEEE European Conf. on Computer Vision*, pages 740–756, 2016. 2

[21] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 2, 5

[22] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 2, 6, 7

[23] Yoav Grauer. Active gated imaging in driver assistance system. *Advanced Optical Technologies*, 3(2):151–160, 2014. 1, 2

[24] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide. Gated2depth: Real-time dense lidar from gated images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 6, 7

[25] Tobias Gruber, Mariia Kokhova, Werner Ritter, Norbert Haala, and Klaus Dictmayer. Learning super-resolved depth from active gated imaging. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3051–3058. IEEE, 2018. 3

[26] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020. 2, 6, 7

[27] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012. 2

[28] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for nir-vis face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 2

[29] Paul Heckman and Robert T. Hodgson. Underwater optical range gating. *IEEE Journal of Quantum Electronics*, 3(11):445–448, 1967. 1, 2

[30] Yong Seok Heo, Kyong Mu Lee, and Sang Uk Lee. Robust stereo matching using adaptive normalized cross-correlation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):807–822, 2011. 3

[31] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Towards precise and efficient image guided depth completion. 2021. 2

[32] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1

[33] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 2

[34] Maximilian Jaritz, Raoul De Charette, Emilie Wirbel, Xavier Perrotton, and Fawzi Nashashibi. Sparse and dense data with cnns: Depth completion and semantic segmentation. In *International Conference on 3D Vision (3DV)*, pages 52–60, 2018. 2

[35] Hae-Gon Jeon, Joon-Young Lee, Sunghoon Im, Hyowon Ha, and In So Kweon. Stereo matching with color and monochrome cameras in low-light conditions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4086–4094, 2016. 3

[36] Maria Jokela, Matti Kutila, and Pasi Pyykönen. Testing and validation of automotive point-cloud sensors in adverse weather conditions. *Applied Sciences*, 9, 2019. 2

[37] Felix Juefei-Xu, Dipan K Pal, and Marios Savvides. Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 141–150, 2015. 2

[38] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[39] Seungryong Kim, Dongbo Min, Stephen Lin, and Kwanghoon Sohn. Dense cross-modal correspondence estimation with the deep self-correlation descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2345–2359, 2021. 3

[40] Andreas Kolb, Erhardt Barth, Reinhard Koch, and Rasmus Larsen. Time-of-flight cameras in computer graphics. In *Computer Graphics Forum*, volume 29, pages 141–159. Wiley Online Library, 2010. 2

[41] Robert Lange. 3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. 2000. 2

[42] Martin Laurenzis, Frank Christnacher, Nicolas Metzger, Emmanuel Bacher, and Ingo Zielenski. Three-dimensional range-gated imaging at infrared wavelengths with super-resolution depth mapping. In *SPIE Infrared Technology and Applications XXXV*, volume 7298, 2009. 2

[43] Martin Laurenzis, Frank Christnacher, and David Monnin. Long-range three-dimensional active imaging with superresolution depth mapping. *Optics letters*, 32(21):3146–8, 2007. 2

[44] Ha Le, Christos Smailis, Lei Shi, and Ioannis Kakadiaris. Edge20: A cross spectral evaluation dataset for multiple surveillance problems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2685–2694, 2020. 2

[45] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7287–7296, 2022. 4, 7

[46] Jose Lezama, Qiang Qiu, and Guillermo Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[47] Jiale Li, Hang Dai, Hao Han, and Yong Ding. Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21694–21704, 2023. 1

[48] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16242–16251, New Orleans, LA, USA, Jun 2022. IEEE. 4, 5, 6, 7, 8

[49] Stan Z. Li, Dong Yi, Zhen Lei, and Shengcai Liao. The casia nir-vis 2.0 face database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2013. 2

[50] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, Junjun Jiang, Bolei Zhou, and Hang Zhao. Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1500–1508, 2022. 6, 7

[51] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022. 2, 6, 7

[52] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Mannequinchallenge: Learning the depths of moving people by watching frozen people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4229–4241, 2021. 2

[53] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6177–6186, 2021. 2

[54] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6197–6206, 2021. 6, 7

[55] Mingyang Liang, Xiaoyang Guo, Hongsheng Li, Xiaogang Wang, and You Song. Unsupervised Cross-spectral Stereo Matching by Learning to Synthesize, Mar 2019. 2, 3, 6, 7

[56] 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, 6, 7

[57] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2624–2641, 2019. 2

[58] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *IEEE International Conference on Robotics and Automation*, pages 1–8, 2018. 2

[59] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. 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*, 2016. 2

[60] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9685–9694, 2021. 2

[61] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proc. of European Conference on Computer Vision (ECCV)*, 2020. 2

[62] Peter Pinggera, T. Breckon, and Horst Bischof. On cross-spectral stereo matching using dense gradient features. In *British Machine Vision Conference*, 2012. 3

[63] Xin Qiao, Chenyang Ge, Youmin Zhang, Yanhui Zhou, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Depth Super-Resolution from Explicit and Implicit High-Frequency Features, May 2023. 2

[64] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 6, 7

[65] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12240–12249, 2019. 2

[66] Dominic Rüfenacht, Clément Fredembach, and Sabine Süsstrunk. Automatic and accurate shadow detection using near-infrared information. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1672–1678, 2013. 2

[67] Michael Schober, Amit Adam, Omer Yair, Shai Mazor, and Sebastian Nowozin. Dynamic time-of-flight. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6109–6118, 2017. 2

[68] Brent Schwarz. Lidar: Mapping the world in 3D. *Nature Photonics*, 4(7):429, 2010. 1, 2

[69] Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Daniela Rus. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 5135–5142. IEEE, 2020. 6

[70] Xiaoyong Shen, Li Xu, Qi Zhang, and Jiaya Jia. Multi-modal and multi-spectral registration for natural images. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 309–324, Cham, 2014. Springer International Publishing. 3

[71] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[72] Jiexiong Tang, John Folkesson, and Patric Jensfelt. Sparse2dense: From direct sparse odometry to dense 3-d reconstruction. *IEEE Robotics and Automation Letters*, 4(2):530–537, 2019. 2

[73] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30:1116–1129, 2020. 2

[74] Shejin Thavalengal, Petronel Bigioi, and Peter Corcoran. Evaluation of combined visible/nir camera for iris authentication on smartphones. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2015. 2

[75] Fabio Tosi, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. RGB-Multispectral Matching: Dataset, Learning Methodology, Evaluation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15937–15947, New Orleans, LA, USA, Jun 2022. IEEE. 2, 3

[76] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017. 2

[77] F Villa, B Markovic, S Bellisai, D Bronzi, A Tosi, F Zappa, S Tisa, D Durini, S Weyers, U Paschen, et al. SPAD smart pixel for time-of-flight and time-correlated single-photon counting measurements. *IEEE Photonics Journal*, 4(3):795–804, 2012. 1

[78] Amanpreet Walia, Stefanie Walz, Mario Bijelic, Fahim Mannan, Frank Julca-Aguilar, Michael Langer, Werner Ritter, and Felix Heide. Gated2gated: Self-supervised depth estimation from gated images. 2022. 1, 2, 5, 6, 7

[79] Celyn Walters, Oscar Mendez, Mark Johnson, and Richard Bowden. There and Back Again: Self-supervised Multispectral Correspondence Estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5147–5154, May 2021. 3

[80] Stefanie Walz, Mario Bijelic, Andrea Ramazzina, Amanpreet Walia, Fahim Mannan, and Felix Heide. Gated stereo: Joint depth estimation from gated and wide-baseline active stereo cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13252–13262, 2023. 1, 2, 3, 6, 7, 8

[81] Dingkang Wang, Connor Watkins, and Huikai Xie. MEMS mirrors for LiDAR: A review. *Micromachines*, 11(5), 2020. 1

[82] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8437–8445, 2019. 1

[83] Z. Wang, C Bovik, H. R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. In *IEEE Transactions on Image Processing*, 2004. 5

[84] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12747–12756, 2021. 2

[85] Wang Xinwei, Li Youfu, and Zhou Yan. Triangular-range-intensity profile spatial-correlation method for 3D super-resolution range-gated imaging. *Applied Optics*, 52(30):7399–406, 2013. 2

[86] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5363–5371, 2017. 2

[87] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12981–12990, 2022. 6, 7

[88] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 6, 7

[89] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 2

[90] Yongjian Zhang, Longguang Wang, Kunhong Li, Zhiheng Fu, and Yulan Guo. Slfnet: A stereo and lidar fusion network for depth completion. *IEEE Robotics and Automation Letters*, 7(4):10605–10612, 2022. 2

[91] Zixiang Zhao, Jiangshe Zhang, Xiang Gu, Chengli Tan, Shuang Xu, Yulun Zhang, Radu Timofte, and Luc Van Gool. Spherical Space Feature Decomposition for Guided Depth Map Super-Resolution, Aug 2023. 2

[92] Tiancheng Zhi, Bernardo R Pires, Martial Hebert, and Srinivasa G Narasimhan. Deep material-aware cross-spectral stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1916–1925, 2018. 2, 3, 6, 7

[93] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2