Cross-spectral Gated-RGB Stereo Depth Estimation (Supplementary Information)

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

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

This supplemental document provides additional information to support the findings in the main manuscript. Specifically, we discuss details on the evaluation dataset, provide further results, ablation experiments, training details, network architecture, and runtime evaluations.

Contents

1. Dataset

In this section, we provide more details on the long-range depth dataset used for training and testing.

1.1. RCCB + Gated Dataset

We adopt the dataset from Walz et al. [23] and refine the provided RCCB stereo camera data. The RCCB camera utilizes an AR0820 sensor. All sensors were housed in a portable sensor cube as showcased in Figure 1. The RCCB stereo system captures 16-bit HDR imagery at 15 Hz with a 3848x2168 resolution. The gated camera records 10-bit images at 120 Hz and 1280x720 resolution, which is further divided into three slices alongside two additional HDR-like captures without active illumination, following [23]. The RCCBs maximum exposure time is limited to 7 ms together with a readout time of 30 ms; this allows for 24Hz repetition times matching that of the 5-slice configuration of the gating setting. Further we utilize the

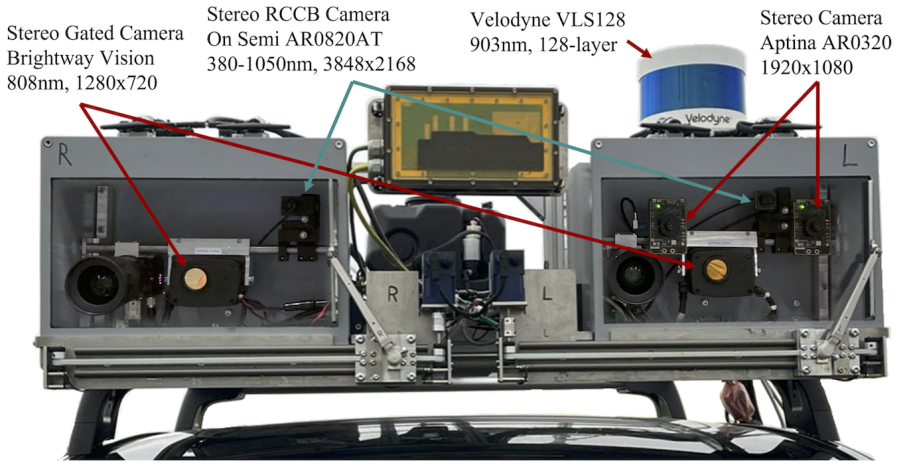

Figure 1. The prototype vehicle for data acquisition is equipped with a stereo gated imaging system, a flood-light flash source (not visible in the image, mounted at the front tow hitch), a standard RGB automotive stereo camera, a Velodyne VLS128 scanning LiDAR, and an RCCB stereo camera setup.

provided LiDAR data from [23], which runs at 10 Hz and a vertical resolutions of 128 lines and range of up to 200 m for high reflective targets. Calibration and time synchronization ensure data consistency across sensors.

The dataset comprises varying conditions: day, night and inclement weather—the dataset boasts 107,348 samples. For comparability with Gated Stereo [23], we use the same training-/validation and test-split, dividing the dataset in 54320 samples for training, 728 samples for validation, and 2463 samples for testing.

1.2. Accumulated Pointclouds

To allow for even greater evaluation distances than in [23] beyond 160 m and to achieve high resolution ground-truth, we use a densely constructed LiDAR map derived from a custom adaptation of the LIO-SAM algorithm, as detailed in Shan et al. [20]. LIO-SAM takes as input the measurements from the car IMU and Global Navigation Satellite System (GNSS) sensor in addition to the LiDAR point cloud. The mapping module consists of three components, LiDAR point cloud ego motion correction, factor graph based IMU motion prediction, and factor graph based global map optimization. Point cloud ego-motion correction is performed using the IMU measurements and odometry. The algorithm's output includes a LiDAR pointcloud along with the calculated position and orientation of the LiDAR, Gated, and RCCB camera sensors, each determined at their respective measurement timestamps. The LiDAR pointcloud is then projected into the camera view using the computed position and orientation. The LiDAR map, being an accumulation of data over an entire scene, contains depth values that extend beyond the maximum range measurable by current depth-estimation and depth-sensing techniques used for in-vehicle, on-the-fly processing. Therefore we limit maximum depth values to a range of 220m. When LiDAR points are accumulated over a scene, this process naturally includes points representing occluded objects, as well as motion artifacts from dynamic objects, when projecting these points into the camera frame. As a countermeasure, we apply a decimation to the projected LiDAR points, following the method by Uhrig et al. [21]. Different from their approach, that uses SGM [8] depth maps to clean the pointcloud projection by removing all LiDAR points exhibiting large relative errors, we deploy the CREStereo stereo [10] algorithm as we found that it generalizes well for different image modalities and is able to reliably filter outliers. Specifically, we exclude points of the projected accumulated pointcloud z_{acc} if the predicted stereo depth z_{stereo} disagrees as follows

$$M_1 = \begin{cases} 1, & \text{if } \frac{|z_{stereo} - z_{acc}|}{z_{acc}} < 0.3 \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Assigning a value of zero to the excluded points matches the treatment of missing values in sparse LiDAR data, ensuring these points are likewise disregarded in the loss calculation process. Here, M_1 is a boolean mask, which is multiplied with the accumulated LiDAR depth z_{acc} to create the filtered pointcloud

$$\hat{z}_{filtered} = M_1 \odot z_{acc}. \tag{2}$$

Subsequent to the filtering-step, the sparse LiDAR map z_{sparse} is added to the filtered pointcloud

$$z_{filtered} = M_2 \odot \hat{z}_{filtered} + z_{sparse}, \tag{3}$$

where M_2 is a boolean mask, ensuring the priorization of the sparse LiDAR measurements, when available, that is

$$M_2 = \begin{cases} 1, & \text{if } z_{sparse} = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In summary, we generate 5834 samples for the training dataset. For the test dataset, dense LiDAR maps are created for 655 frames, with alignment and accuracy ensured through manual filtering. Qualitative examples of the raw projected pointcloud and the corresponding filtered pointcloud compared to the sparse LiDAR are shown in Figure 2. A visualization of reconstructed scenes with pointclouds, is shown in Figure 3.

1.3. Lost Cargo Dataset

For additional qualitative evaluation, we captured data that includes staged scenes with small objects on the ground showcasing both the advantage of the high resolution of a color array and the depth accuracy of the gated modality. We emulate lost cargo scenarios that infrequently occur on highways and roads. Those scenarios are critical for a safe operation as they tackle sensor resolution limits, requiring an early detection at long ranges with small object dimensions. Often, these objects are caused by various cargo or vehicle parts found on the road, presenting a significant hazard. We stage these scenarios as their natural occurrence is very rare. This data was gathered during day and night in Northern America. The dataset features 'lost cargo'-like objects like tires, lost bumpers, and humans on the ground, simulating post-crash scenarios, and other items. Additional qualitative examples of lost-cargo depth estimation are provided in Section 6.1.

2. Additional Network Details

In this section, we provide detailed descriptions of the network architecture for the pose network, attention-based fusion, and stereo network of the proposed model.

2.1. Additional Details on Pose Estimation Network

Both gated and RCCB stereo setups are time-synchronized to microsecond precision. Due to automatic exposure and shutter timing, the RCCB camera can accumulate small time offsets, effectively resulting in a time-synchronization offset of up to 20ms between gated cameras and RCCB cameras. To mitigate the influence of resulting misaligned images, we employ a pose estimation network which predicts a rotation and translation component from coarsely aligned features, warped into a common view using the predicted depth and calibrated poses. This network operates at all refinement stages except for the coarsest level, utilizing feature maps of varying resolutions. To achieve consistent results from a single PoseNet, all feature maps are resized prior and have $\frac{1}{16}$th the width and height of the RCCB image. The input feature maps to the PoseNet is the gated feature map generated by the stereo networks backbone and the coarsly aligned RCCB-feature map from the preceding warping operation. In addition to the feature maps, an additional feature map containing the time offset in ms is created by repeating this value over the complete feature map input size, see Table 1. The PoseNet predicts a feature vector of length, see Table 2. The first three values in the sequence represent an Euler rotation vector, while the last three values define a translation vector. To confine the output range, we multiply the first 5 values by 0.05 and the last value by 0.8, effectively limiting the output values the PoseNet can generate. The values fall within a range consistent with what can be expected from vehicle motion. In this context, the final value indicates the vehicle longitudinal translation movement. This accounts for a larger offset, especially at higher speeds, of up to 0.8 meters.

2.2. Attention-based Fusion

In our approach, gated and RCCB features are fused after the preceding alignment step. The network has information from both the RCCB and the gated modalities and perform well in all cases, even when one modality is failing. The focus of the fused features should lie on the informative modality. To achieve this, we introduce an attention-based fusing mechanism that is able to highlight certain feature maps and features that are beneficial to the networks prediction. The network architecture is defined in Table 3. The inputs are the registered RCCB and gated images, generated with the refined pose from PoseNet. During the fusion process, we employ a combination of global and local attention mechanisms. These are capable of predicting weighted features that effectively highlight the most beneficial features.

2.3. Stereo Network

Our stereo network performs iterative refinement based on GRU layers. We adopt adaptive group correlation layers from [10]. In our feature extraction process, we inherit MPViT-S [9] to learn the dense representations $F_l^c, F_r^c, F_l^g, F_r^g$ for

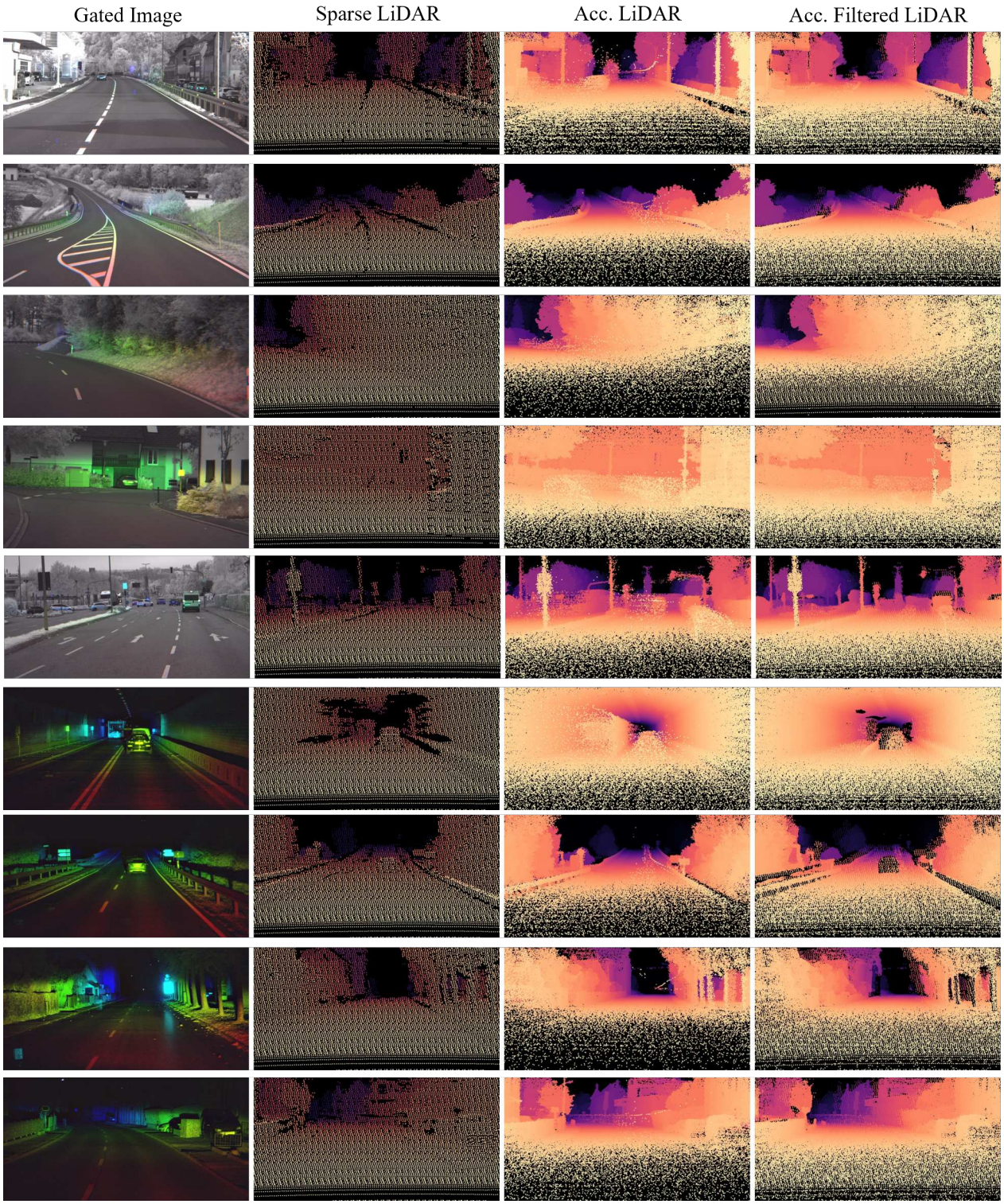

Figure 2. Sparse LiDAR and accumulated and filtered LiDAR, see text. LiDAR points are enhanced for visibility. By filtering, occluded points and artifacts from dynamic objects are removed.

the stereo matching. MPViT is a transformer-based backbone f_b^c, f_b^g that is designed to extract features for dense prediction tasks. For the shared-weight encoders for gated images, we enable a 5-channel input for the utilization of three active and two

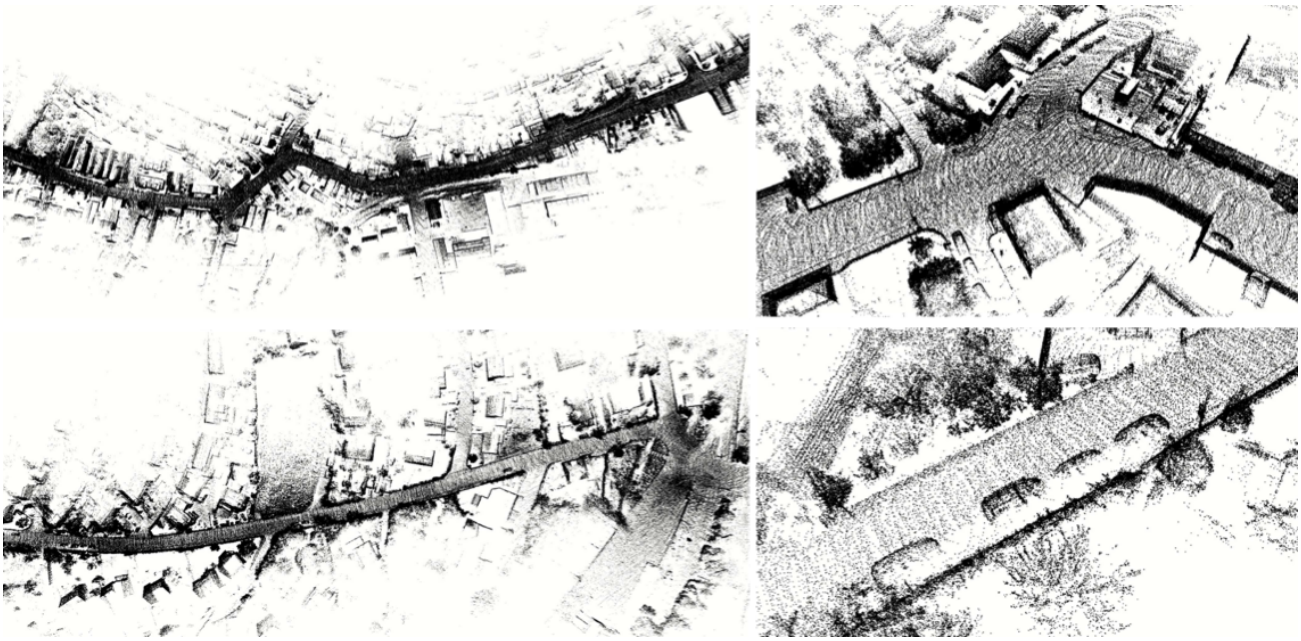

Figure 3. Visualization of an aggregated scene for evaluation, reconstructed with LiDAR points: The left side is a zoomed-out view of a larger scene, while the right side shows a zoomed-in perspective, highlighting the accurate reconstruction of cars, trees, houses, and smaller structures.

passive inputs, which are concatenated channel-wise to form five-channel inputs I_l^g, I_r^g. The shared-weight RCCB encoder encodes three-channel RGB input images I_l^c, I_r^c. The feature maps are downsampled to create four refinement stages at different resolutions. The refinement stages are defined $\frac{1}{48}, \frac{1}{24}, \frac{1}{12}, \frac{1}{4}$ for the RCCB image resolution and $\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{3}{4}$ for the gated resolution respectively. In our network, we execute our CSM in every refinement step to register and fuse cross-spectral features. Within the CSM, PoseNet (see Section 2.1) and attention-based fusion (see Section 2.2) are carried out.

3. Training and Implementation Details

Next, we provide additional details on the training process and implementation of the proposed method. We recall the overall training loss of the proposed method is defined as

$$\mathcal{L}_{stereo} = c_1\mathcal{L}_{reproj} + c_2\mathcal{L}_{recon} + c_3\mathcal{L}_{lidar}, \tag{5}$$

with constants $c_1 = 0.05$, $c_2 = 0.01$, $c_3 = 1$.

In the early stages of training, we warp with pre-computed depth map generated from a pre-trained CREStereo [10] network for RCCB view depth map generation and Gated Stereo [23] for gated view depth map generation. We found that this accelerates training significantly, as feature warping and, hence, alignment and fusion is independent from the generated

INTERMDIATE FEATURES (POSENET)		
Layer #		**Output Shape**
1a	Gated Feature Map	$128 \times \frac{H}{16} \times \frac{W}{16}$
1b	RCCB-Warp Feature Map	$128 \times \frac{H}{16} \times \frac{W}{16}$
2	Time-Offset Feature Map	$1 \times \frac{H}{16} \times \frac{W}{16}$
3	Time-Offset Feature Map	$1 \times \frac{H}{32} \times \frac{W}{32}$
4	Time-Offset Feature Map	$1 \times \frac{H}{64} \times \frac{W}{64}$

Table 1. Representation of intermediate features of PoseNet decoder. Gated feature map is product of stereo-backbone. RCCB-warp feature map is generated through preceding coarse alignment step. Time-offset feature map is measured time-offset between RCCB and gated images, repeated over respective feature map size.

POSENET (DECODER)			
Layer #	**Layer Description**		**Output Shape**
5	Concat	#1a $\oplus$ #1b	$256 \times \frac{H}{16} \times \frac{W}{16}$
6	ConvSqueeze	Conv (1x1)	$256 \times \frac{H}{16} \times \frac{W}{16}$
		ReLU	
7	Concat	#6 $\oplus$ #2	$257 \times \frac{H}{16} \times \frac{W}{16}$
8a	ConvBlock-1	Conv (3x3)	$257 \times \frac{H}{32} \times \frac{W}{32}$
8b	ScatterND	Update with #3	$257 \times \frac{H}{32} \times \frac{W}{32}$
		ReLU	
9a	ConvBlock-2	Conv (3x3)	$257 \times \frac{H}{64} \times \frac{W}{64}$
9b	ScatterND	Update with #4	$257 \times \frac{H}{64} \times \frac{W}{64}$
10	ConvBlock-3	Conv (1x1)	$6 \times \frac{H}{64} \times \frac{W}{64}$
11	Reshape		$1 \times 1 \times 6$
	Tanh		

Table 2. Architecture for PoseNet decoder. The ScatterND operation replaces the last channel of the current feature map with corresponding time-offset feature map.

FUSION			
Layer #	**Layer Description**		**Output Shape**
1a	LocalAttentionBlock-1 (Gated)		$128 \times \frac{H}{X} \times \frac{W}{X}$
1b	LocalAttentionBlock-1 (RCCB)		$128 \times \frac{H}{X} \times \frac{W}{X}$
1c	GlobalAttentionBlock-1 (Gated)		$1 \times \frac{H}{X} \times \frac{W}{X}$
1d	GlobalAttentionBlock-1 (RCCB)		$1 \times \frac{H}{X} \times \frac{W}{X}$
2a	Addition	#1a + #1c	$128 \times \frac{H}{X} \times \frac{W}{X}$
		Sigmoid	
2b	Addition	#1b + #1d	$128 \times \frac{H}{X} \times \frac{W}{X}$
		Sigmoid	
3a	Addition	#2a + #2b	$128 \times \frac{H}{X} \times \frac{W}{X}$
3b	Division	#2a / #3a	$128 \times \frac{H}{X} \times \frac{W}{X}$
3c	Division	#2b / #3a	$128 \times \frac{H}{X} \times \frac{W}{X}$
4a	Multiplication	2 $\cdot$ Gated $\odot$ #3b	$128 \times \frac{H}{X} \times \frac{W}{X}$
4b	Multiplication	2 $\cdot$ RCCB $\odot$ #3c	$128 \times \frac{H}{X} \times \frac{W}{X}$
4c	Addition	#4a + #4b	$128 \times \frac{H}{X} \times \frac{W}{X}$
5a	LocalAttentionBlock-2 (#4b)		$128 \times \frac{H}{X} \times \frac{W}{X}$
5b	GlobalAttentionBlock-2 (#4c)		$1 \times \frac{H}{X} \times \frac{W}{X}$
5c	Addition	#5a + #5b	$128 \times \frac{H}{X} \times \frac{W}{X}$
		Sigmoid	
6a	Multiplication	#5c $\odot$ #4a	$128 \times \frac{H}{X} \times \frac{W}{X}$
6b	Multiplication	$(1-$#5c$)$ $\odot$ #4b	$128 \times \frac{H}{X} \times \frac{W}{X}$
6c	Addition	#6a + #6b	$128 \times \frac{H}{X} \times \frac{W}{X}$

Table 3. Architecture for Fusion Layer. Local and Global Attention blocks are adapted from [4]. Fusion is applied in every refinement stage of the stereo network, therefore the output shape of each layer is determined by $X \in \{48, 24, 12, 4\}$ with respect to the height H and width W of the RCCB image.

depth outputs. During later stages and for testing, we fully rely on the iterative depth predictions from our cross-spectral stereo network.

3.1. Implementation Details

The backbones f_b^c, f_b^g of our stereo-network follow the MPViT-S architecture from [9]. For the attention a, we implement the MS-CAM module, introduced by [4]. To reduce computational complexity, pose refinement with PoseNet p is executed once per refinement stage, reusing the same pose for subsequent CSM passes. We crop both the RCCB and gated images to ensure similar fields of view, resulting in resolutions of 512x1024 pixels for gated and 1536x3072 pixels for RCCB images. The depth map resolution matches that of the RCCB images at 1536x3072 pixels. Data analysis shows a minimum LiDAR acquisition distance of 4.3 metres, which is recovered by our algorithm. We train our network for 30 epochs total, with a batch size of 4. For the first 20 epochs, the network is trained at 1/3 resolution, skipping the last upsample layer, with a learning rate of $5 \cdot 10^{-3}$ and a weight decay of 10^{-2} using ADAMW [15] with $\beta_1 = 0.9, \beta_2 = 0.999$. In the final 10 epochs, the network undergoes training at a constant learning rate of $3 \cdot 10^{-6}$ and incorporates full-resolution images. Here, the newly added layer is initialized with the weights from the prior refinement stage. In total we employ 28 refinement steps split into the four scales in the sequence $6, 6, 12, 4$. The network architecture is optimized on 4 NVIDIA RTX A6000 GPUs with 48GB memory each. For the reported quantitative results, we employ a total of 28 refinement steps for training and testing, distributed as $6, 6, 12, 4$ across the refinement stages at $\frac{1}{48}, \frac{1}{24}, \frac{1}{12}, \frac{1}{4}$ with respect to the RCCB image resolution, respectively.

3.2. Training of Baseline Methods

In order to fairly compare our proposed model with current state-of-the-art methods, we use the same training, validation, and testing datasets for all baseline methods as were used for our method. For [1,2,5–7,11,12,14,16,18,22,25,26], we adopt the results generated by [23]. Instead of initiating training from scratch, they opted to use pre-existing, publicly available models that were well-suited for their needs, fine-tuning them on the Gated Stereo [23] dataset. We adopt a similar approach in our methodology. We utilize CREStereo [10] as baseline for high-resolution RCCB images. We employ a model that has been fine-tuned on the ETH3D [19] dataset, and further fine-tune it on our data using sparse LiDAR depth supervision. To process the high-resolution RCCB data, we implement a two-stage inference strategy, similar to that proposed by [10]. In this method, disparity is first estimated using RCCB images downsampled to one third it's original resolution, and the model output is then used to initialize a second prediction step using the full-resolution images. CS-Stereo [27] was trained in a self-supervised manner enforcing left-right consistency. Due to the significant disparities in appearance between NIR and RGB images, it is not directly possible to enforce photometric consistency across these different modalities. Consequently, they utilize auxiliary material information for spectral translation, which aids in creating a pseudo-NIR image. Our dataset presents two major challenges that preclude the direct application of the referenced training approach. Firstly, the utilization of active gated images, containing time-of-flight data from illumination, varies greatly from passive NIR images, particularly at night, which breaks the process of spectral translation from RGB. Secondly, the training method depends on additional material information, which our dataset does not include. Therefore we finetune this method using LiDAR supervision. We train using rectified RCCB-gated stereo pairs. To achieve this, we specifically use the three active gated slices and downsample the RCCB images to one-third of their original size. This approach ensures that both images in the pair have equal resolution. We apply the same training strategy and image sets to UCSSM [13], choosing to fine-tune it using LiDAR supervision additionally. This approach was necessitated as the model did not converge when trained on our data using their originally proposed self-supervision only. The limitation arises because their method, which relies on spectral translation for left-right supervision, is not capable of generating active gated images.

4. Runtime Evaluation

To maximize efficiency, we exploit the temporal redundancy found in the scene geometry by incorporating information from preceding frames. Specifically, we inherit the approach of [3], cold-starting disparity prediction in the first frame and predicting depth as detailed in Section 3.1. For subsequent scenes, we warp previous disparity estimations into the current frame using an estimated transformation matrix utilizing Softmax Splatting to handle occlusions [17]. The warped disparity is used to warm-start disparity prediction, resulting in the need of only 2 GRU layers. Our optimized setup is optimized in our prototype system with four parallel A6000 GPUs, distributing the feature extractors, and operates with equal resolution as Gated Stereo [23] at 11.78 Hz for FP16, matching the recording rate of the LiDAR reference sensor. Doubling the resolution of the RCCB camera comes at a runtime cost of 39%.

5. Additional Ablation Experiments

In the following, we present additional ablation experiments to validate the choices we make for our method.

5.1. Evaluation of Pose Network

To validate the effectiveness of the pose aspect of our method, we disable the pose prediction step in our ablation. This leads to an increase of the MAE averaged over day and night by 2.1%. Qualitatively, the updated poses can be validated by comparing RCCB images warped into the gated frame using the calibrated and the updated poses. An example validating the effectiveness the approach is given in Figure 4. A quantitative evaluation of this kind is not possible due to the differing appearances of gated and RCCB images.

Figure 4. In certain cases, the temporal offset between the RCCB and gated cameras causes misalignment of the images. This is evident when the RCCB image is warped into the gated frame using the predicted depth, as shown in the top row. However, by using the refined pose from the pose network and subsequent adjustment, these misalignments can be significantly reduced. This is demonstrated in the images in the bottom row, where an overlay of gated and warped RCCB image is shown and a better alignment of the bus-stop sign is evident.

5.2. Evaluation of Attention-based Fusion

In this section, we further validate our attention-based fusion approach, introduced in Section 2.2. We exchange attention-based fusion with simple addition of the left gated feature map F_l^g and the left RCCB feature map warped into the gated frame $F_l^{c|g}$ for the case of depth prediction for the left gated image frame. For addition-based fusion, the fused features $\hat{F}$ are obtained through

$$\hat{F} = \frac{1}{2} F_l^g + \frac{1}{2} F_l^{c|g}. \tag{6}$$

This operation weights both feature maps the same in all cases and is not able to weight certain features more or less. In our ablation, we find that using addition-based fusion over attention-based fusion increases the MAE averaged over day and night by 7.3%. Qualitatively, the method makes less use of the RCCB features, leading to a lack of certain details, see Figure 5.

5.3. Sensor Ablation

To evaluate our dual stereo setup approach, we ablate it with the best possible methods that only employ a subset of the available sensors. Therefore, we conduct four experiments. Firstly, we evaluate the use of only one monocular gated sensor. Secondly, we compare it to RCCB stereo images. Subsequently, we compare one single RCCB imager and one single gated imager. Finally, we compare it to the Gated Stereo [23] imaging approach. Additionally, we implement our method switching the RCCB stereo camera with an RGB stereo camera.

To benchmark our method against methods that use monocular setups, we implement ManyDepth [24] utilizing a single gated camera. We extend the existing training approach with LiDAR supervision. This strong gated baseline exceeds the performance of Gated2Gated [22] through more complex neural architectures and multi-frame analysis. Secondly, for the comparison to the RCCB stereo camera setup, we build on top of CREStereo [10], achieving the best possible passive color camera setup in daytime conditions as detailed in Table 1 of the main manuscript. Subsequently, we compare it to one cross-modal stereo setup built from a single gated camera and one RCCB camera. As with the RCCB setup, it builds upon [10]. We deviate from the original implementation by learning one feature extractor per modality and allowing 5-channel inputs

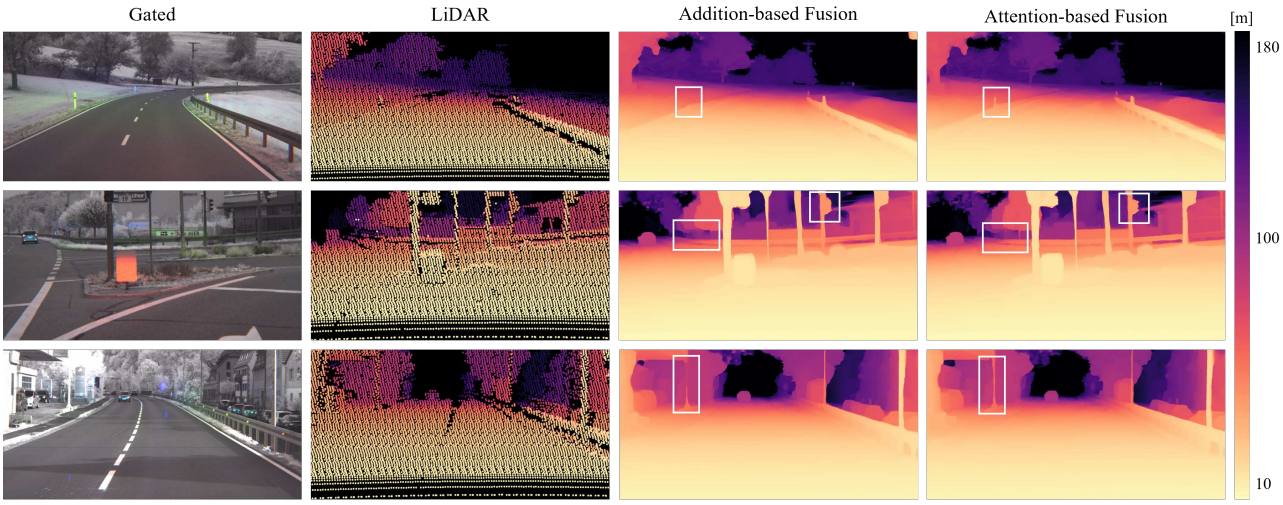

Figure 5. Qualitatively, the use of attention-based fusion leads to a more consistent depiction of fine structures, which are sometimes missed with addition-based fusion. Failure cases of addition-based fusion are highlighted. In the top row, the guide post is missed by addition-based fusion. In the second row, tree trunks and traffic light details are missed. In the third row, a pole is lost.

to accommodate the three gated slices and two passive NIR HDR images. We train this method using LiDAR supervision, achieving competitive results during the day compared to state-of-the-art mono-gated [6, 22] and cross-spectral RGB-NIR methods [13, 27]. However, this approach falls short at night, as it cannot bridge the larger domain difference between the active gated cues and passive RCCB images. Hence, also the gated cues fall short, and the approach drops below the monocular gated approach detailed in Table 3 of the main manuscript.

Finally, all single modalities don't meet the same level of the performance achieved by Gated Stereo [23], highlighting the value of a stereo-gated imaging setup, especially in scenarios with intense ambient lighting. Consequently, only the combination of cross-spectral data with our proposed training scheme to bridge the modalities can surpass the gated approach, achieving state-of-the-art results by combining the unique advantages of each modality and offering high resolution from color arrays and superior depth estimation from gated images.

For validating our decision to use RCCB images instead of RGB images which are included in the dataset as well, we train our method tailored to the usage of the RGB stereo images. The depth prediction performance is inferior to the Gated RCCB Stereo in all metrics, particularly noticeable at night with a 9.8% lower RMSE, as shown in Tab 4. This shows the benefit of the RCCB color array achieving 30% higher sensitivity compared to the RGGB array, maximizing signal-to-noise ratio in night conditions. Moreover, while the RGB camera boasts a higher resolution of 2MP compared to the gated camera 1MP, it significantly lags behind the RCCB camera 8MP resolution. Consequently, the resolution of fine details in the depth maps is lower for RGB in our setup, as illustrated in Figure 6. Most notably, Gated RGB Stereo, although inferior to Gated RCCB Stereo, shows better performance both qualitatively and quantitatively compared to Gated Stereo [23], demonstrating the effectiveness of our approach in combining complementary sensors for improved depth estimation.

6. Additional Qualitative Results

In this section, we present additional qualitative results of the proposed method and competing methods, including monocular gated [22], monocular RGB [12], stereo RGB [14], stereo RCCB [10], and cross-spectral rgb-nir [13, 27] approaches. Figures 7, 8, 9, and 10 present the depth map predictions from existing methods alongside the corresponding RCCB, gated, RGB, and LiDAR measurements. Our method produces depth estimates with enhanced sharpness and clarity both day and

Evaluated on accumulated LiDAR Ground-Truth Points										
	DAY					NIGHT				
Model	**RMSE**	**MAE**	δ_1	δ_2	δ_3	**RMSE**	**MAE**	δ_1	δ_2	δ_3
GATED STEREO [23]	14.24	8.67	86.76	98.76	99.41	14.03	8.93	84.21	98.83	99.46
GATED RGB STEREO	11.42	7.22	89.66	99.29	99.70	13.20	8.43	85.52	99.19	99.74
GATED RCCB STEREO	**10.69**	**6.83**	**90.25**	**99.57**	**99.81**	**12.02**	**7.94**	**86.05**	**99.71**	**99.90**

Table 4. Additional comparison to Gated RGB Stereo approach.

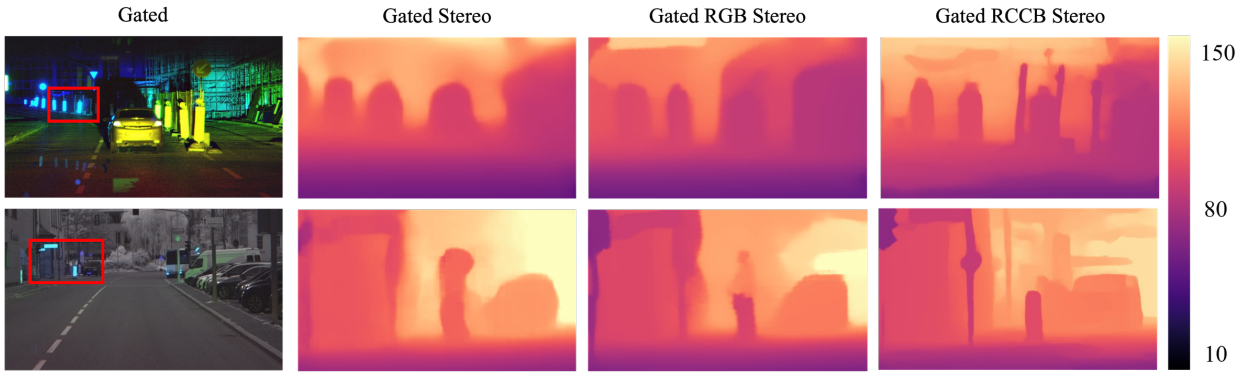

Figure 6. Qualitative comparison highlighting the advantages of Gated RCCB Stereo method, in contrast to Gated RGB Stereo and Gated Stereo [23].

night. Compared to Gated Stereo [23], which estimates depth accurately during the day through the use of passive gated slices, our method performs favorably both qualitatively and quantitatively through the utilization of high-resolution RCCB images. In Figure 7, we accurately estimate the depth of the third car and the street light, which Gated Stereo [23] fails to identify. Our cross-spectral approach, particularly effective at night, utilizes RCCB images when they are available. In scenarios where RCCB data might not be accessible due to nighttime conditions, the method effectively compensates by harnessing the additional illumination from the gated imaging system. Figure 9 and Figure 10 highlight the benefit of our method compared to CREStereo [10] and Gated Stereo [23] which each utilize only one of the modalities. Compared to state-of-the-art cross-spectral methods UCSSM [13] and CS-Stereo [27], our method shows a clear improvement, as both methods fail to discern any details both day and night. Our method performs better than both RAFT-Stereo [14] and Depthformer [12] in terms of detail and overall performance. Our advantage is especially noticeable at night, when RGB-based methods struggle with insufficient illumination. While Gated2Gated [22] outperforms RGB methods in terms of detail at night, it lacks during daytime with sufficient ambient light present and is consistently outperformed by our method in all scenarios.

6.1. Lost Cargo

We qualitatively compare our method to the best method per modality, namely Gated Stereo [23] and CREStereo [10], focusing on detecting small, potentially hazardous 'lost cargo' objects. For a more detailed description of the lost cargo dataset, see Section 1.3. Our cross-spectral method consistently estimates depth for lost cargo objects both day and night, generating depth maps with sharp edges and clear details. CREStereo [10] performs well, when passive illumination is sufficient. However, it is less consistent than our method, missing objects even during well-lit daytime conditions, as seen for the package and the tire. During the night, CREStereo [10] depth is compromised due to missing illumination, where it fails to accurately represent the shapes of the objects or misses them completely. Gated Stereo [23] performs better than CREStereo [10] at night, better representing the shape of the objects. However it is limited due to a lower resolution and performs generally not as well during the day. Our Gated RCCB Stereo approach consistently outperforms both methods by integrating the complementary features of the two camera systems. This includes the active illumination provided by the gated imaging system and the higher resolution offered by the RCCB stereo camera.

References

[1] 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. 7

[2] 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. 7

[3] Ziang Cheng, Jiayu Yang, and Hongdong Li. Stereo Matching in Time: 100+ FPS Video Stereo Matching for Extended Reality, Sep 2023. 7

[4] 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. 6, 7

[5] 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. 7

[6] 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. 7, 9

[7] 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. 7

[8] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, Feb 2008. 2

[9] 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. 3, 7

[10] 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. 2, 3, 5, 7, 8, 9, 10, 12, 16

[11] 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. 7

[12] 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. 7, 9, 10

[13] Mingyang Liang, Xiaoyang Guo, Hongsheng Li, Xiaogang Wang, and You Song. Unsupervised Cross-spectral Stereo Matching by Learning to Synthesize, Mar 2019. 7, 9, 10

[14] 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. 7, 9, 10

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2018. 7

[16] 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. 7

[17] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2020. 7

[18] 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. 7

[19] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 7

[20] 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. 2

[21] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 2

[22] 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. 7, 8, 9, 10

[23] 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, 5, 7, 8, 9, 10, 12, 15, 16

[24] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth, Jul 2021. 8

[25] 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. 7

[26] 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. 7

[27] 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. 7, 9, 10

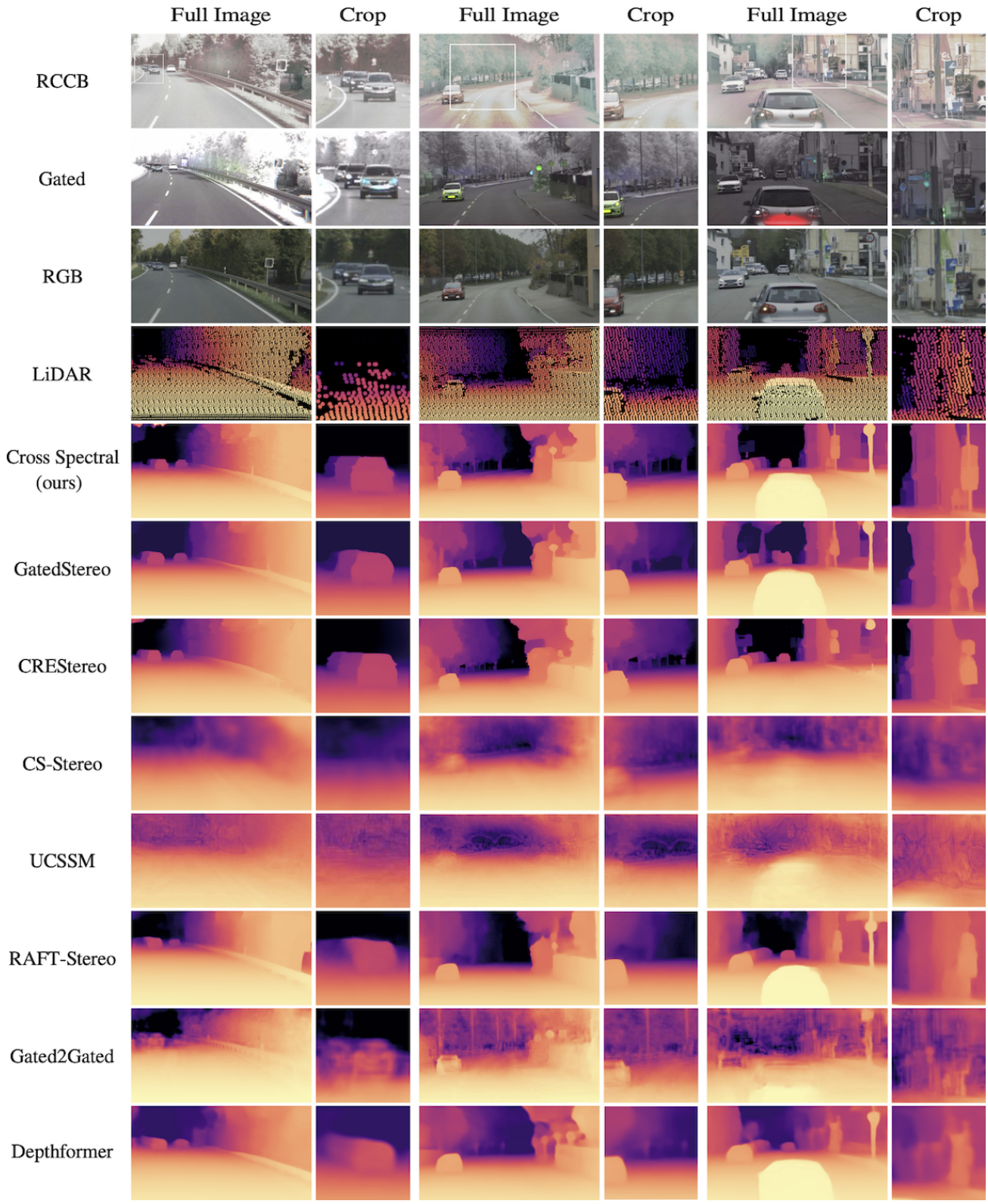

Figure 7. Additional qualitative comparison of our method to existing depth estimation methods during daytime. Our approach, in comparison to the next-best methods, Gated Stereo [23] and CREStereo [10], excels in estimating depth with finer details and sharper edges. For instance, in the first image, Gated Stereo misses the third car, and in the second image, both Gated Stereo and CREStereo are unable to accurately estimate the depth of the street light.

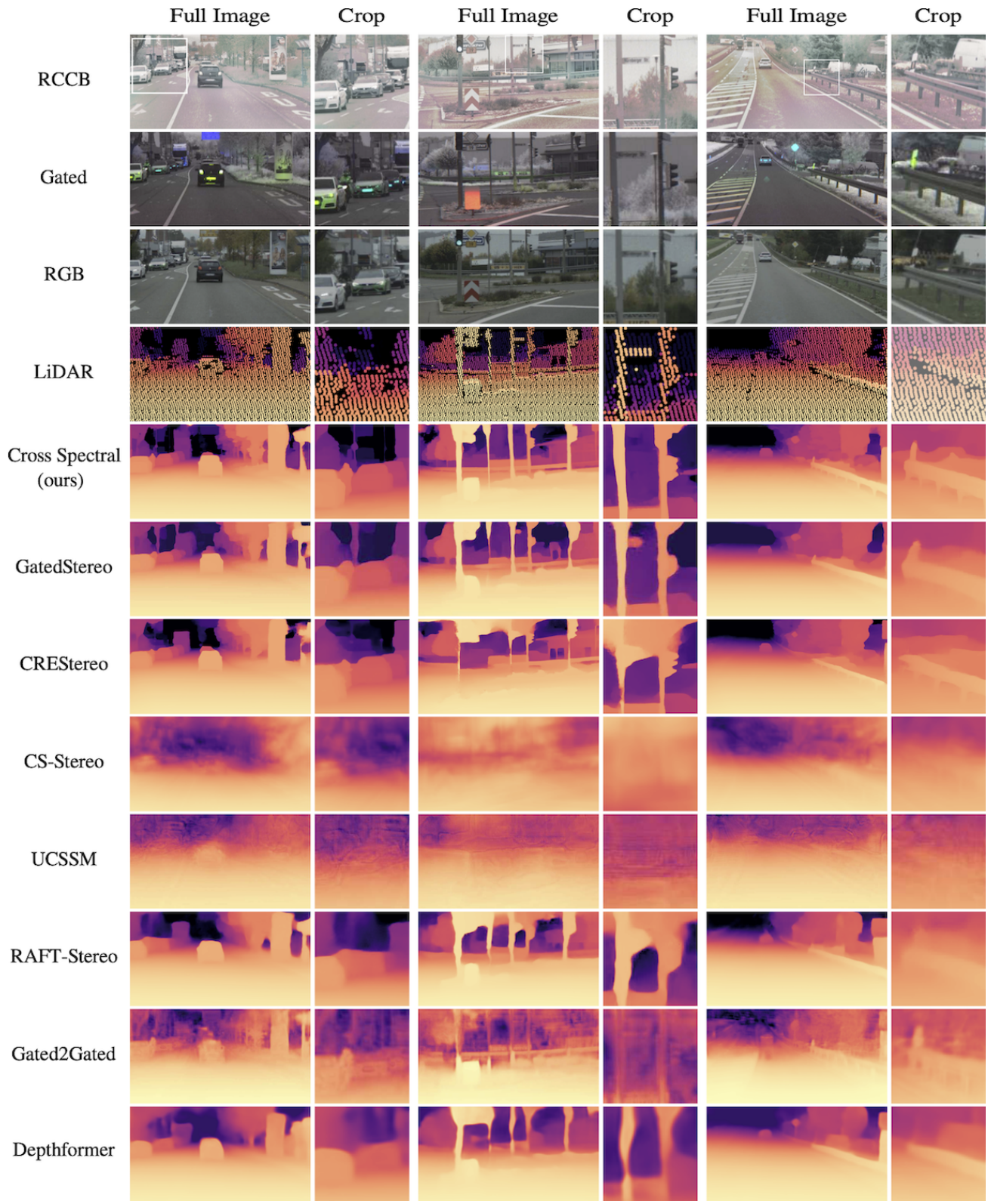

Figure 8. Additional qualitative comparison of our method to existing depth estimation methods. Our method stands out as the only one capable of accurately estimating depth for the motorcyclist, the traffic light, and the guide post.

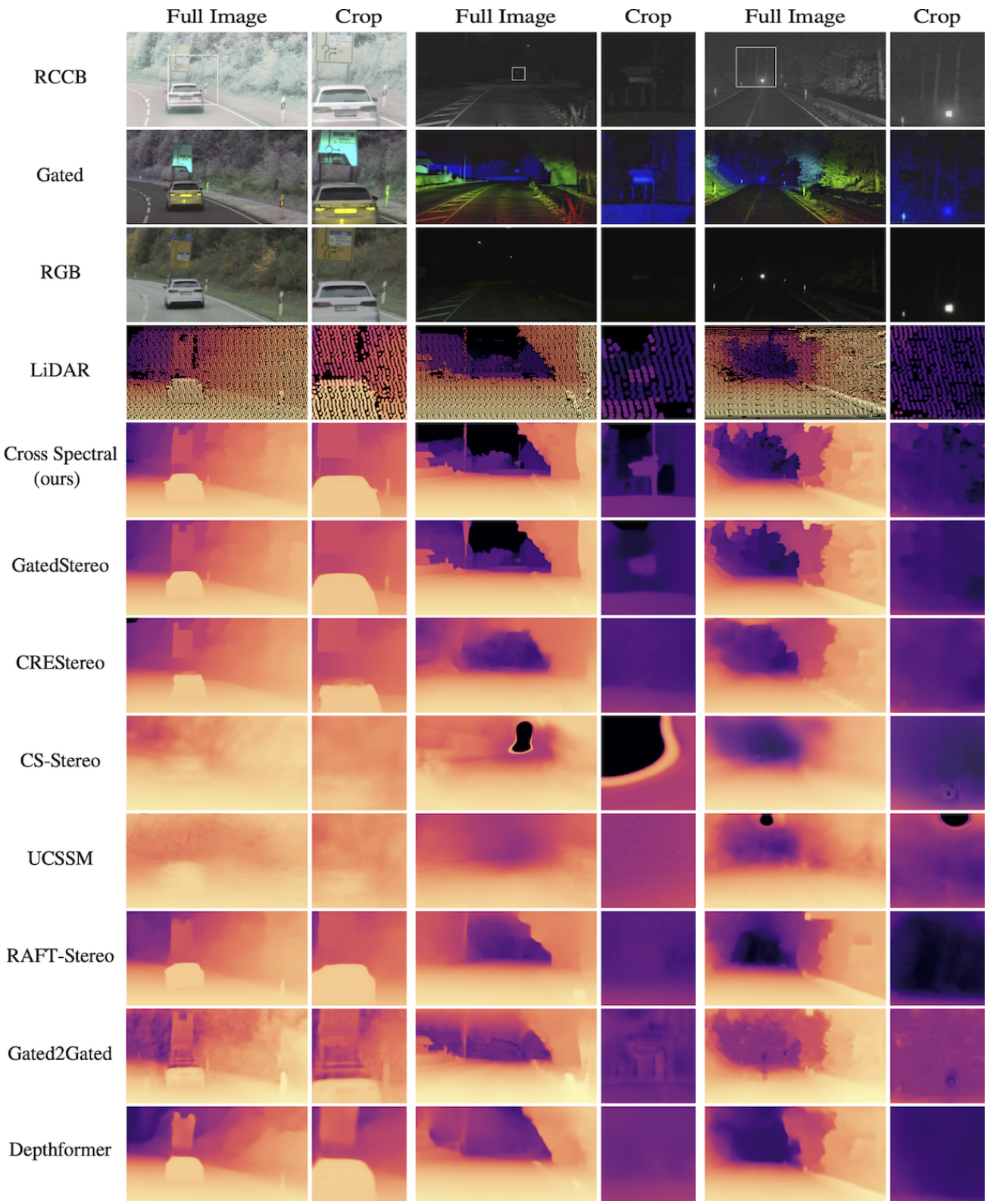

Figure 9. Additional qualitative comparison of our method to existing depth estimation methods. Particularly at night, our cross-spectral approach surpasses state-of-the-art methods in capturing details, effectively and accurately depicting distant, small structures like street signs and distant trees.

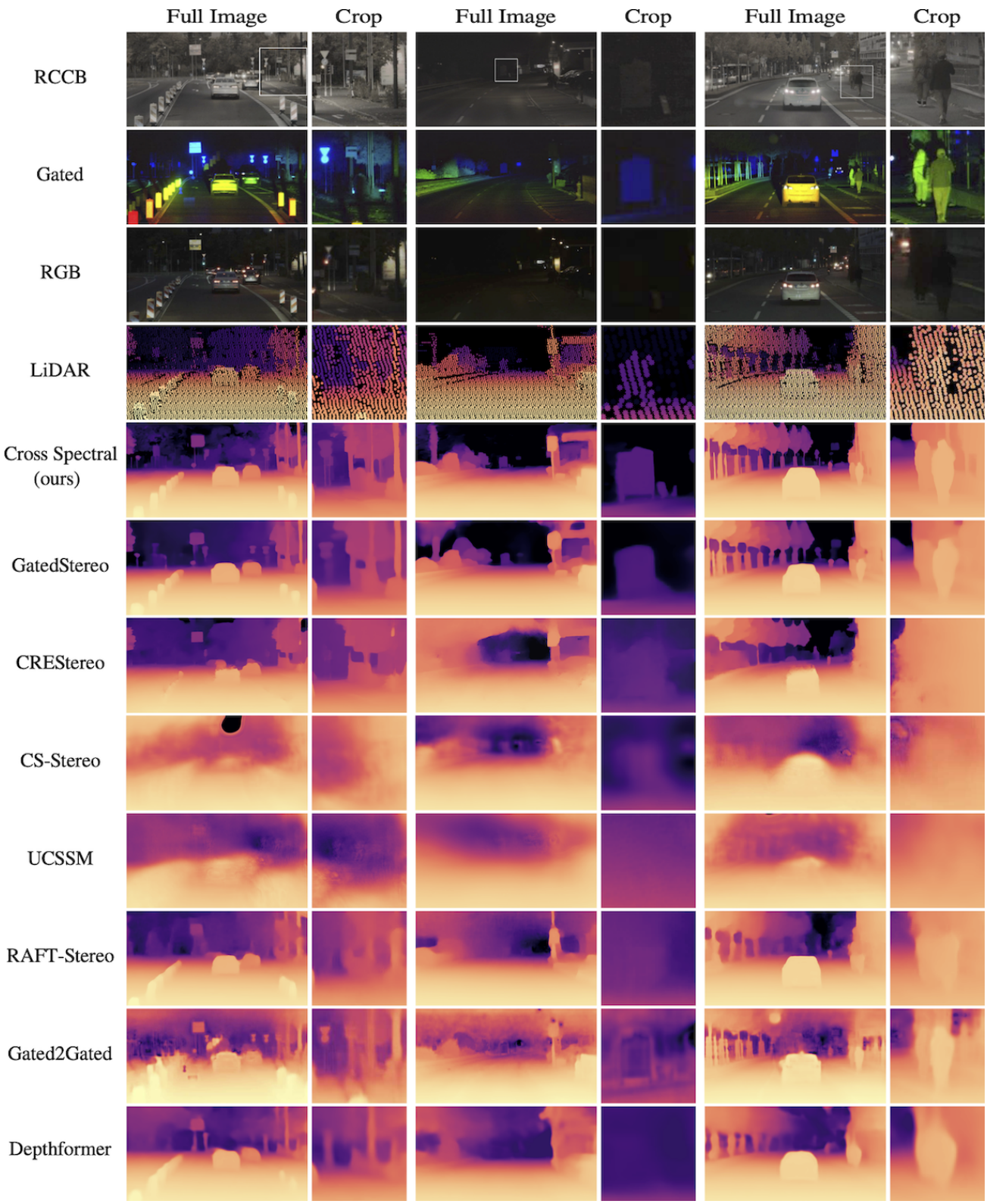

Figure 10. Qualitative comparison of our method to other state-of-the-art methods. Our Gated RCCB Stereo generates highly-detailed accurate depth maps and all other methods fail to do. In the third column, our method distinctly excels by clearly estimating the shapes of both pedestrians and objects in the background, thus significantly surpassing the quality of Gated Stereo [23].

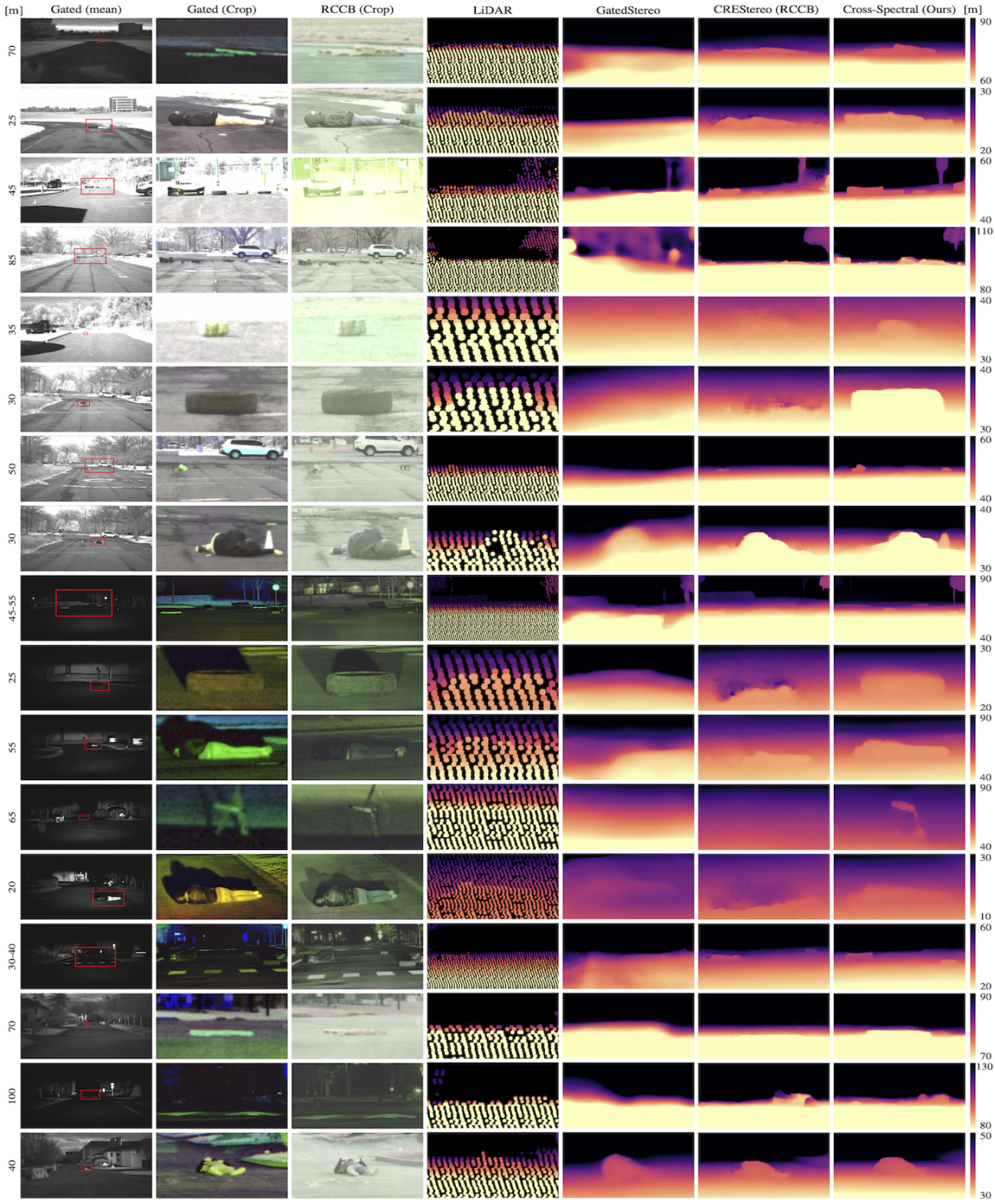

Figure 11. Lost-cargo Evaluation for Depth of Small Objects at Long Distances. We compare here our method to Gated Stereo [23] and CREStereo [10] (RCCB) as next-best methods using the gated and RGB modality. Our method reliably produces accurate depth measurements for lost-cargo objects, shown here for different ranges and varying time of day. Gated Stereo struggles with lost-cargo due to the low-resolution sensor. RGB depth performance is limited to well-lit scenarios. This highlights the benefit of our cross-spectral setup, outperforming also LiDAR where lost-cargo detection is difficult due to sparsity.