Gated Stereo:
Joint Depth Estimation from Gated and Wide-Baseline Active Stereo Cues
(Supplementary Document)

Stefanie Walz[1] Mario Bijelic[2] Andrea Ramazzina[1] Amanpreet Walia[3] Fahim Mannan[3] Felix Heide[2,3]
[1]Mercedes-Benz [2]Princeton University [3]Algolux

This supplemental document provides additional information in support of the findings in the main manuscript. Specifically, Section 1 describes the stereo gated imaging dataset. Section 2 provides further quantitative and depth-resolved evaluations. Section 3 details the network architecture of the proposed method and Section 4 describes the training procedure. In Section 6, the gated reconstruction loss is described and Section 7 introduces the ambient consistency in more depth with additional qualitative examples. Section 8 shows the differences between illumination view consistency and left-right warping and Section 9 demonstrates qualitative advantages of fusing mono and stereo depth predictions. Finally, Section 10 provides further qualitative evaluations in support of the investigation from the main document.

Contents

Figure 1. Experimental sensor setup of the prototype vehicle for data acquisition. The car is equipped with a stereo gated imaging system with a baseline of 0.76 m, consisting of two gated imagers and a flood-light flash source, a standard RGB automotive stereo camera and a scanning LiDAR Velodyne VLS128 as reference.

1. Stereo Gated Dataset

In this section, we provide more details on the sensor setup and the statistics of the captured long-range gated stereo dataset. As described in the main document, we use this dataset (after splitting) into train, validate, and test the proposed method.

The sensor setup of our testing vehicle is shown in Figure 1 and consists of an automotive RGB stereo camera (OnSemi AR0230), a reference LiDAR system (Velodyne VLS128), and a stereo gated imaging system consisting of two gated cameras (BrightwayVision BrightEye) and a flood-light flash laser source. The experimental setup is mounted in two sensor boxes on top of the roof of an electric test vehicle. The left sensor box can be considered the master sensing box, with one gated camera in here and the second one mounted in the satellite box with a wide baseline of 0.76 m. The gated imagers have a pixel pitch of 10 μm and provide 10 bit images, captured in the NIR band at 808 nm with a resolution of 1280x720. The optics of the cameras provide field-of-view of 31.1°x17.8° (horizontal x vertical), and run at 120 Hz, which we split up into three active slices and two HDR-like passive images captured without any active illumination. This results in an overall system repetition rate of 24 Hz. The two vertical-cavity surface-emitting laser (VCSEL) modules required for active illumination of the scene are mounted on the front tow hitch of the car. Both lasers have a pulsed optical output peak power of 500 W. Number of pulses and pulse length are limited according to eye-safety regulations. We use the left gated camera as master for triggering laser pulse emission and acquisition of the right camera. The RGB Stereo system is mounted with a baseline of 0.23 m in the left master sensor cube and runs at 30Hz. This baseline corresponds to a typical multi-view and stereo systems in related literature [2, 3] and in typical ADAS systems. As such, in our comparisons, existing stereo models have also been trained with similar setups. The two On-Semi AR0230 imagers provide a resolution of 1920x 1024 and a 12 bit quantization. We use Lensagon B5M8018C optics with a focal length of 8mm to obtain a field-of-view of 39.6°x21.7°(HxV). Both stereo camera systems are synchronized and calibrated with aligned horizontal epipolar lines for efficient disparity matching along one axis of the image.

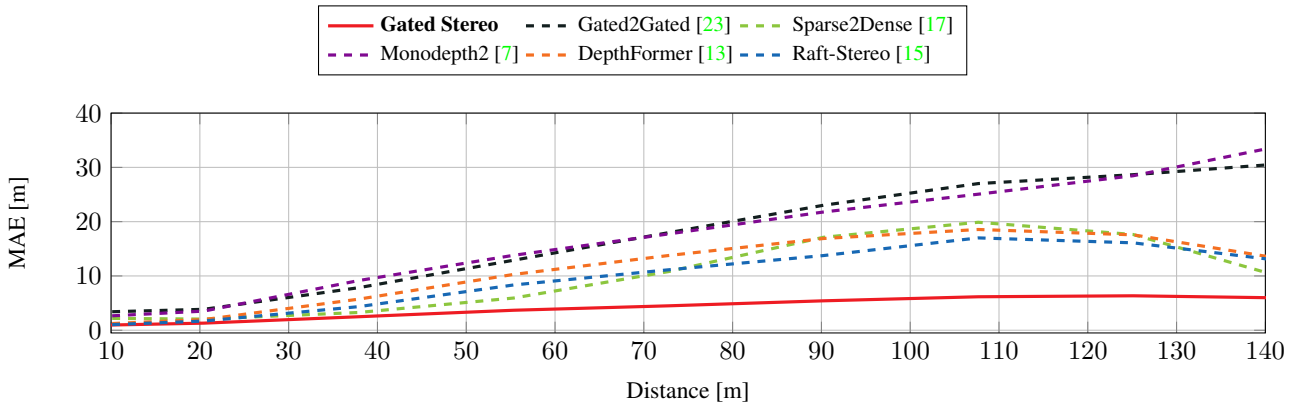
For ground-truth depth annotation, we use a Velodyne VLS128 which has a spec sheet range larger then 200 m for high reflective targets. This LiDAR operates at 905 nm and runs at 10 Hz. Furthermore, the LiDAR provides a vertical field-of-view of 40 ° with 128 non-linear distributed scanning lines with a minimum angular resolution of 0.11°. The sensor setup is implemented using the Robot Operating System (ROS) as middleware allowing to record all sensors in a common framework. For time synchronization, we extended the ROS approximate time synchronizer such that the sequential gated slices are assembled through a filter before time synchronizing with the LiDAR system and the stereo camera. This ensures that the gated slices are always kept in order with ascending illumination slices scanning from short to long distances.
In total, our dataset contains more than 100,000 samples (54,429 day/52,919 night) captured in southern Germany covering a wide variety of urban, highway, and overland scenarios. After selecting only samples in sync, we split the dataset into 54,302 (26,010 day/28,292 night) samples for training, 728 (415 day/313 night) samples for validation, and 2,463 (1,269 day/1,194 night) samples for testing. A small subset of our dataset, containing all the different sensor modalities, is shown in Figures 16, 17, and 18.

2. Additional Quantitative Evaluations

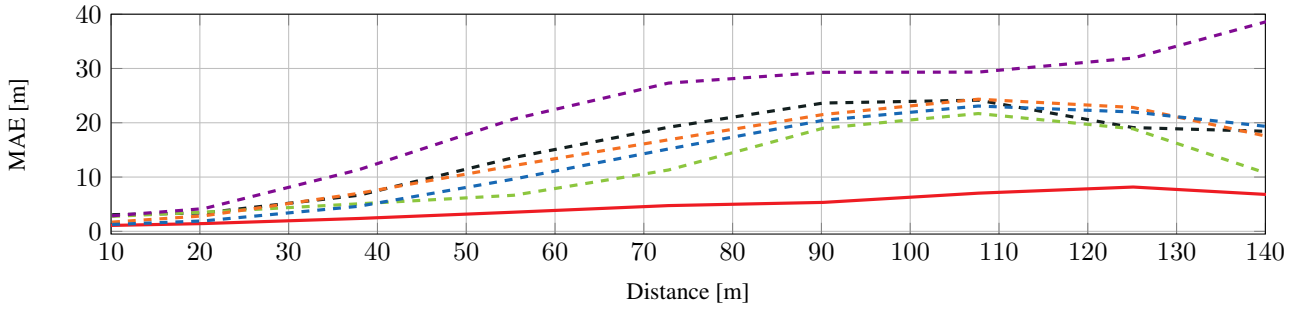
In this section, we provide additional quantitative evaluations for the proposed Gated Stereo and baseline state-of-the-art methods. See Sec. 10 for additional qualitative results.

2.1. Additional Depth-Resolved Quantitative Evaluation

Next, we present additional quantitative results for the proposed gated stereo method and alternate state-of-the-art methods. For depth ground-truth annotation, we use a Velodyne VLS-128 LiDAR. This allows us to evaluate the performance of depth estimation algorithms for distances up to 160 m which is twice as far as in previous works [9, 22, 23]. The evaluation results of the proposed Gated Stereo and other state-of-the-art methods are reported in Table 1. We here also provide depth-resolved evaluation results, proposed by [8]. Gruber et al. [8] have demonstrated that depth maps are usually not uniformly distributed, but instead close distances are more frequent than long distances due to the camera frustum. Thus, when calculating the mean error of depth maps, errors at shorter distances (which are typically smaller) contribute more to the mean than errors at large distances, resulting in an overall smaller error. In order to weigh distances more equally, we provide binned evaluations, where the different metrics are calculated in bins of approximately 16m, and the mean of the bins yields the final result. This ensures that every distance contributes equally to the evaluation metric. The quantitative results of the binned and not binned metrics are shown in Table 1. For all reference methods, the performance drops significantly for the binned metrics due to larger errors in further distances. *Only Gated Stereo is able to maintain nearly constant quality over all distances* resulting in only a slight increase in the binned metrics. This is also reflected in Figure 2, which visualizes the depth-dependent MAE for the different methods. While the MAE increases for all reference methods over long distances, Gated Stereo is able to maintain its performance at day and night.



(a) Binned MAE calculated at daytime.



(b) Binned MAE calculated at nighttime.

Figure 2. MAE calculated over depth bins of approximately 16 m in day and nighttime conditions. **Gated Stereo** outperforms all other methods especially for far distances.

	Method	Modality	Train	Test Data - Night						Test Data - Day					
				RMSE	ARD	MAE	δ_1	δ_2	δ_3	RMSE	ARD	MAE	δ_1	δ_2	δ_3
not binned	GATED2DEPTH [9]	Mono-Gated	D	16.15	0.17	8.07	75.70	92.74	96.47	28.68	0.22	14.76	66.68	82.76	87.96
	GATED2GATED [23]	Mono-Gated	MG	14.08	0.19	7.95	79.84	92.95	96.59	16.87	0.21	9.51	73.93	92.15	96.10
	SPARSE2DENSE [17]	Mono-Sparse	D	9.97	0.11	5.22	87.06	95.77	98.20	10.05	0.11	4.77	88.06	96.57	98.63
	KBNET [25]	Mono-Sparse	D	13.77	0.16	8.73	80.98	**99.33**	**99.67**	15.27	0.17	9.54	78.54	**99.31**	**99.63**
	NLSPN [18]	Mono-Sparse	D	12.19	0.09	5.42	89.63	96.84	99.03	11.78	0.08	4.99	91.41	97.70	99.24
	PENET [11]	Mono-Sparse	D	7.81	0.09	3.59	93.68	97.90	99.16	8.54	0.09	3.82	93.78	97.69	98.94
	GUIDENET [21]	Mono-Sparse	D	7.50	0.09	3.63	92.70	98.16	99.35	8.03	0.09	3.70	93.23	98.12	99.21
	PACKNET [10]	Mono-RGB	M	17.82	0.20	10.21	66.35	87.85	95.61	17.69	0.21	9.77	72.12	90.65	96.51
	MONODEPTH2 [7]	Mono-RGB	M	18.44	0.18	9.47	75.70	90.46	95.68	20.78	0.22	10.06	79.05	90.66	94.69
	SIMIPU [12]	Mono-RGB	D	15.78	0.18	8.71	76.25	90.84	96.44	14.33	0.14	7.50	81.77	94.01	97.92
	ADABINS [1]	Mono-RGB	D	14.45	0.15	7.58	81.47	93.75	97.39	12.76	0.12	6.53	86.15	95.77	98.41
	DPT [19]	Mono-RGB	D	12.15	0.12	6.31	85.38	95.94	98.42	11.28	0.09	5.52	89.56	96.83	98.80
	DEPTHFORMER [13]	Mono-RGB	D	12.15	0.11	6.20	85.18	95.76	98.47	10.59	0.09	5.06	90.65	97.46	99.02
	PSMNET [4]	Stereo-RGB	D	27.98	0.27	16.02	50.77	74.77	85.93	32.13	0.28	18.09	53.82	74.91	84.96
	HSMNET [27]	Stereo-RGB	D	12.42	0.09	5.87	88.41	96.08	98.50	10.36	0.08	4.69	92.47	97.93	99.11
	ACVNET [26]	Stereo-RGB	D	11.70	0.08	5.25	89.91	96.33	98.47	9.40	0.07	4.08	94.61	98.36	99.12
	RAFT-STEREO [15]	Stereo-RGB	D	10.89	0.09	5.10	90.47	96.71	98.64	9.40	0.07	4.07	93.76	98.15	99.09
	GATED STEREO	Stereo-Gated	DGS	**6.39**	**0.05**	**2.25**	**96.40**	98.44	99.24	**7.11**	**0.05**	**2.25**	**96.87**	98.46	99.11
binned	GATED2DEPTH [9]	Mono-Gated	D	19.98	0.17	13.45	79.09	92.67	96.92	28.60	0.23	20.34	63.76	88.48	93.62
	GATED2GATED [23]	Mono-Gated	MG	20.39	0.22	15.06	73.72	92.15	96.09	25.76	0.30	18.12	67.96	86.86	92.56
	SPARSE2DENSE [17]	Mono-Sparse	D	14.34	0.12	9.83	86.20	96.40	98.50	14.04	0.11	8.93	87.95	96.60	98.43
	KBNET [25]	Mono-Sparse	D	19.79	0.26	18.63	48.10	**99.17**	**99.69**	20.21	0.26	18.81	47.45	**99.18**	**99.66**
	NLSPN [18]	Mono-Sparse	D	21.13	0.14	16.60	68.86	95.71	98.97	19.04	0.13	14.62	70.97	97.09	99.16
	PENET [11]	Mono-Sparse	D	12.79	0.10	7.05	91.25	96.76	98.53	13.04	0.10	6.82	91.78	96.62	98.26
	GUIDENET [21]	Mono-Sparse	D	12.07	0.09	7.03	91.51	97.36	98.91	12.08	0.09	6.62	92.21	97.34	98.72
	PACKNET [10]	Mono-RGB	M	29.19	0.30	22.08	54.98	81.91	93.89	27.37	0.32	20.31	60.54	83.86	93.23
	MONODEPTH2 [7]	Mono-RGB	M	34.89	0.45	26.18	57.56	80.05	88.65	31.30	0.44	22.53	67.51	83.25	88.78
	SIMIPU [12]	Mono-RGB	D	23.37	0.25	17.63	64.90	86.81	95.54	20.48	0.20	15.04	72.87	91.75	97.32
	ADABINS [1]	Mono-RGB	D	22.61	0.24	16.64	68.03	89.03	95.93	18.94	0.17	13.37	78.44	94.17	97.93
	DPT [19]	Mono-RGB	D	20.14	0.19	14.59	74.53	92.48	97.08	16.98	0.14	11.69	82.92	95.75	98.58
	DEPTHFORMER [13]	Mono-RGB	D	19.33	0.17	13.90	75.82	93.59	98.00	15.94	0.13	10.94	84.45	96.47	98.80
	PSMNET [4]	Stereo-RGB	D	47.79	0.60	39.48	31.30	55.23	73.04	46.58	0.56	38.19	30.51	53.16	74.21
	HSMNET [27]	Stereo-RGB	D	19.93	0.15	13.84	78.48	94.21	98.34	16.25	0.12	10.88	85.99	97.03	98.95
	ACVNET [26]	Stereo-RGB	D	19.60	0.14	13.04	80.49	94.26	98.14	15.07	0.11	9.55	90.11	97.70	98.74
	RAFT-STEREO [15]	Stereo-RGB	D	18.81	0.15	12.81	80.66	94.23	98.02	15.01	0.11	9.25	89.45	97.48	98.83
	GATED STEREO	Stereo-Gated	DGS	**9.75**	**0.06**	**4.16**	**95.91**	98.34	99.20	**10.42**	**0.06**	**3.87**	**96.22**	98.09	98.86

Table 1. Comparison of our proposed method and state-of-the-art methods on the Gated Stereo test dataset. 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. All metrics are also evaluated for bins of approximately 16m to weight all distances equally.

3. Additional Network Details

In this section, we provide detailed descriptions of the network architecture for the stereo, monocular, and fusion branches of the proposed model.

3.1. Stereo Network Branch

Our stereo network builds on the RAFT-Stereo [15] architecture consisting of three main components: a feature extractor, a correlation pyramid, and a GRU-based update operator. Instead of using series of residual blocks for extracting features as in [15], we use an HRFormer [28] variant to learn dense multi-resolution representation from 5-channel active and passive inputs. In particular, we have used the HRFormer-S model to generate features at 1/4, 1/8, 1/16, and 1/32 of input image resolution. This is followed by aggregating features from each level after up-sampling to generate multi-level representation, similar to [20]. Following [15], we also use two separate feature extractors - *feature encoder* for extracting features from

both input views, and one *context encoder* for extracting features only from the target view (i.e., left) to be passed as input to GRU during each iteration. Both feature extractors generate a 256-channel final feature representation at 1/4 of input spatial resolution.

3.2. Ambient and Albedo Network Details

In addition to depth, we predict ambient and albedo information. This allows us to reconstruct a gated image using the range intensity profile information for the cyclic loss. The reconstruction process is described in more detail in Section 6. We jointly estimate ambient and albedo utilizing the intermediate feature representations from the context network and pass them through a decoder module for ambient and albedo estimation. The feature decoder network architecture is listed in Table 2.

INTERMEDIATE FEATURES (CONTEXT ENCODER)	
Layer #	**Output Shape**
1a	$64 \times H \times W$
2a	$96 \times \frac{H}{2} \times \frac{W}{2}$
3a	$128 \times \frac{H}{4} \times \frac{W}{4}$
4a	$128 \times \frac{H}{8} \times \frac{W}{8}$
5a	$128 \times \frac{H}{16} \times \frac{W}{16}$

ALBEDO (DECODER)				AMBIENT (DECODER)			
Layer #	**Layer Description**		**Output Shape**	**Layer Description**			**Output Shape**
6a	Upsampling-1	ConvTranspose2D (kernel = 2)	$128 \times \frac{H}{8} \times \frac{W}{8}$	Upsampling-1	ConvTranspose2D (kernel = 2)		$128 \times \frac{H}{8} \times \frac{W}{8}$
		BatchNorm2D			BatchNorm2D		
6b	Concat-1	Layer #6a $\oplus$ Layer #4a	$256 \times \frac{H}{8} \times \frac{W}{8}$	Concat-1	Layer #6a $\oplus$ Layer #4a		$256 \times \frac{H}{8} \times \frac{W}{8}$
6c	UpConvBlock-1	Conv (3x3)	$128 \times \frac{H}{8} \times \frac{W}{8}$	UpConvBlock-1	Conv (3x3)		$128 \times \frac{H}{8} \times \frac{W}{8}$
		LeakyReLU ($\alpha = 0.2$)			LeakyReLU ($\alpha = 0.2$)		
		BatchNorm2D			BatchNorm2D		
		Conv (3x3)			Conv (3x3)		
		LeakyReLU ($\alpha = 0.2$)			LeakyReLU ($\alpha = 0.2$)		
		BatchNorm2D			BatchNorm2D		
7a	Upsampling-2		$128 \times \frac{H}{4} \times \frac{W}{4}$	Upsampling-2			$128 \times \frac{H}{4} \times \frac{W}{4}$
7b	Concat-2	Layer #7a $\oplus$ Layer #3a	$256 \times \frac{H}{4} \times \frac{W}{4}$	Concat-2	Layer #7a $\oplus$ Layer #3a		$256 \times \frac{H}{4} \times \frac{W}{4}$
7c	UpConvBlock-2		$128 \times \frac{H}{4} \times \frac{W}{4}$	UpConvBlock-2			$128 \times \frac{H}{4} \times \frac{W}{4}$
8a	Upsampling-3		$96 \times \frac{H}{2} \times \frac{W}{2}$	Upsampling-3			$96 \times \frac{H}{2} \times \frac{W}{2}$
8b	Concat-3	Layer #8a $\oplus$ Layer #2a	$192 \times \frac{H}{2} \times \frac{W}{2}$	Concat-3	Layer #8a $\oplus$ Layer #2a		$192 \times \frac{H}{2} \times \frac{W}{2}$
8c	UpConvBlock-3		$96 \times \frac{H}{2} \times \frac{W}{2}$	UpConvBlock-3			$96 \times \frac{H}{2} \times \frac{W}{2}$
9a	Upsampling-4		$64 \times H \times W$	Upsampling-4			$64 \times H \times W$
9b	Concat-4	Layer #9a $\oplus$ Layer #1a	$128 \times H \times W$	Concat-4	Layer #9a $\oplus$ Layer #1a		$128 \times H \times W$
9c	UpConvBlock-3		$64 \times H \times W$	UpConvBlock-3			$64 \times H \times W$
10	Conv1D		$1 \times H \times W$	Conv1D			$1 \times H \times W$

Table 2. Encoder-Decoder based architecture for $f_{\mathbf{\Lambda}\alpha}$. Here, $\oplus$ defines channel concatenation across feature tensors, α defines the slope of a LeakyReLU. For the features obtained from the context network, we pass it to two decoder heads which predict albedo and ambient for the scene. Here, H and W represent the height and width of the input.

3.3. Monocular Network Branch Details

For monocular depth estimation from active and passive gated inputs, we depart from Walia et al. [23] and employ a DPT-Hybrid [19] as the choice of architecture for its ability to preserve fine-grained details and predicting globally coherent results, in contrast to PackNet used in [23]. The network architecture is adapted at the first input layer to accept 5-channel (active + passive gated channels) as input (as compared to 3-channel RGB).

3.4. Fusion Network Details

To merge the results obtained from monocular and stereo depth modality, we pass the gated image input, that consists our of 3 active gated slices + 2 passive slices, stereo depth, and monocular depth to a fusion network. The network applies an encoder-decoder architecture. We report the architecture for the encoder and decoder blocks of the proposed fusion network in Table 3 and 4. The depth outputs are obtained at multiple resolutions and scaled to the original input dimensions. The final output is obtained from the last stage, i.e., Output1.

FUSION ENCODER											
Layer #	**Layer Name**			**output channels**	**kernel**	**stride**	**Pad**	**BatchNorm**	**Non-Linearity**	**MaxPool2D**	**Output Shape**
0	Input										$7 \times H \times W$
1a	Conv1			64	7	2	3	✓	ReLU	✓	$64 \times \frac{H}{2} \times \frac{W}{2}$
2a	Layer1	Layer 1.0	Layer1.0.0	64	3	1	1	✓	ReLU	-	$64 \times \frac{H}{4} \times \frac{W}{4}$
			Layer1.0.1	64	3	1	1	✓	-	-	$64 \times \frac{H}{4} \times \frac{W}{4}$
		Layer1.1	Layer1.1.0	64	3	1	1	✓	ReLU	-	$64 \times \frac{H}{4} \times \frac{W}{4}$
			Layer1.1.1	64	3	1	1	✓	-	-	$64 \times \frac{H}{4} \times \frac{W}{4}$
3a	Layer2	Layer 2.0		128	3	2	1	✓	ReLU	-	$128 \times \frac{H}{8} \times \frac{W}{8}$
		Layer2.1		128	3	1	1	✓	ReLU	-	$128 \times \frac{H}{8} \times \frac{W}{8}$
4a	Layer3	Layer3.0		256	3	2	1	✓	ReLU	-	$256 \times \frac{H}{16} \times \frac{W}{16}$
		Layer3.1		256	3	1	1	✓	ReLU	-	$256 \times \frac{H}{16} \times \frac{W}{16}$
5a	Layer4	Layer4.0		512	3	2	1	✓	ReLU	-	$512 \times \frac{H}{32} \times \frac{W}{32}$
		Layer4.1		512	3	1	1	✓	ReLU	-	$512 \times \frac{H}{32} \times \frac{W}{32}$

Table 3. Architecture for encoder stage of the proposed fusion network block. All convolution operations refer to 2D-convolution operations. Features extracted from the encoder stage, which are used in fusion decoder stage, are listed in Table 4.

FUSION DECODER								
Layer #	**Layer Name**		**output channels**	**kernel**	**stride**	**pad**	**Non_linearity**	**Output Shape**
6	Decoder0	ConvBlock0	256	3	1	(Reflection)(1,1)	ELU($\alpha = 1.0$)	$256 \times \frac{H}{32} \times \frac{W}{32}$
		Upsample + Concat (4a)						$512 \times \frac{H}{16} \times \frac{W}{16}$
		ConvBlock1	256	3	1	(Reflection)(1,1)	ELU($\alpha = 1.0$)	$256 \times \frac{H}{16} \times \frac{W}{16}$
7	Decoder1	ConvBlock0	128	3	1	(Reflection)(1,1)	ELU($\alpha = 1.0$)	$128 \times \frac{H}{16} \times \frac{W}{16}$
		Upsample + Concat (3a)						$256 \times \frac{H}{8} \times \frac{W}{8}$
		ConvBlock1	128	3	1	(Reflection)(1,1)	ELU($\alpha = 1.0$)	$128 \times \frac{H}{8} \times \frac{W}{8}$
	Ouput4	depthconv4	1	3	1	1	ReLU	$1 \times \frac{H}{8} \times \frac{W}{8}$
		Upsample						$1 \times H \times W$
8	Decoder2 (ConvBlock0, Upsample + Concat (2a), ConvBlock1)		64	3	1	(Reflection)(1,1)	ELU($\alpha = 1.0$)	$64 \times \frac{H}{4} \times \frac{W}{4}$
	Output3	depthconv3	1	3	1	1	ReLU	$1 \times \frac{H}{4} \times \frac{W}{4}$
		Upsample						$1 \times H \times W$
9	Decoder3 (ConvBlock0, Upsample + Concat (1a), ConvBlock1)		32	3	1	(Reflection)(1,1)	ELU($\alpha = 1.0$)	$32 \times \frac{H}{2} \times \frac{W}{2}$
	Output2	depthconv2	1	3	1	1	ReLU	$1 \times \frac{H}{2} \times \frac{W}{2}$
		Upsample						$1 \times H \times W$
10	Decoder4 (ConvBlock0, Upsample, ConvBlock1)		16	3	1	(Reflection)(1,1)	ReLU	$16 \times H \times W$
	Output1	depthconv1	1	3	1	1	ReLU	$1 \times H \times W$

Table 4. Architecture for fusion network decoder. Here, the upsample operation is a bilinear interpolation resizing towards the output shape, the concat operation stacks the features obtained from Table 3 along the feature dimension. Output{1-4} refers to depth output at multiple resolutions resized to input dimensions.

4. Additional Training Details

In this section, we provide details on the loss function, pretraining and relevant hyperparameters.

For training, we use the following overall loss,

$$\mathcal{L}_{mono} = c_1\mathcal{L}_{recon} + c_2\mathcal{L}_{sup} + c_3\mathcal{L}_{smooth}, \tag{1}$$

$$\mathcal{L}_{stereo} = c_4\mathcal{L}_{reproj} + c_5\mathcal{L}_{recon} + c_6\mathcal{L}_{illum} + c_7\mathcal{L}_{sup} + c_8\mathcal{L}_{smooth}, \tag{2}$$

$$\mathcal{L}_{fusion} = c_9\mathcal{L}_{ms} + c_{10}\mathcal{L}_{sup} + c_{11}\mathcal{L}_{smooth}, \tag{3}$$

with constants $c_{1,\ldots,11}$. Those values are chosen to be $c_1 = 0.01$, $c_2 = 1.0$, $c_3 = 0.001$, $c_4 = 0.01$, $c_5 = 0.01$, $c_6 = 0.002$, $c_7 = 1.0$, $c_8 = 0.001$, $c_9 = 0.1$, $c_{10} = 1.0$, and $c_{11} = 0.005$.

As an effective training strategy, we first independently optimize the monocular and stereo networks using the losses

presented in Eq. (1), (2), and (3). The monocular network is optimized with the $\mathcal{L}_{mono}$ loss for 12 epochs using ADAMW [16] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate of 10^{-4} and of weight decay 10^{-2}. The stereo network is trained with $\mathcal{L}_{stereo}$ for 12 epochs using ADAMW [16] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, learning rate of 10^{-4} and weight decay of 10^{-2}. Finally, the fusion network is trained with frozen monocular and stereo networks with the loss $\mathcal{L}_{fusion}$ for 5 epochs using ADAMW and the losses described in Eq. (3) with a learning rate of $3 \cdot 10^{-4}$.

We use $\eta = 0.05$ for generating occlusion masks referred in Equation (4) of the main paper. The gating mask M_g and signal-to-noise ratio mask M_{SNR} is defined in Section 6. For the M_{SNR} mask we set the parameter θ to 0.04 and and γ to 0.98. All methods are trained on NVIDIA A100 GPUs with 80GB memory.

4.1. Training for Baseline Methods

To provide a fair comparison of the proposed model against state-of-the-art methods, all the baseline methods that we compare to use the same training, validation and testing dataset as used for Gated Stereo.
The four sparse depth-completion algorithms, namely Sparse2Dense [17], Calibrated Backprojection Network (KBNet) [25], NLSPN [18], PENet [11], and GuideNet [21] have been finetuned using the RGB left frame and a subset of the LiDAR pointcloud, created by randomly sampling five hundred points from the pointcloud. The monocular RGB methods PackNet [10] and Monodepth2 [7] have been finetuned through the use of temporal data of the left camera, following their respective proposed training methodologies. The other monocular algorithms SimIPU [12], AdaBins [1], DPT [19] and DepthFormer [13] have been finetuned using the sparse depth supervision provided by the LiDAR, in an analogous way to how Gated Stereo has been trained. A similar approach has been employed also for the stereo depth-supervised models PSMNet [4], STereo TRansformer (STTR) [14], HSMNet [27], ACVNet [26] and RAFT-Stereo [15], using the highest possible resolution. For all the methods mentioned above, instead of training from scratch, we start from the publicly available pretrained models best fit for our use case, which typically are the ones finetuned on the KITTI dataset, and finetune it on our data.

5. Runtime

Our optimized full network is bottlenecked by the stereo branch and optimized in our prototype system with two parallel A100 GPUs, one for the monocular branch and one for the stereo branch, and operates at 14.5Hz for FP32 and 23Hz for FP16, matching the recording rate of the sensor. The addition of the HRFormer block comes at a reduction of 27.4% in runtime. The addition of a 5-channel input tensor, including the two passive slices, to the Mono and Stereo branches comes at a runtime cost of 10% and for 1.7%, respectively. Gated2Gated [55] runs on the same hardware at 90Hz. Our non-optimized components are the following (runtime included): Fusion Stage (175Hz), MonoGated (DPTHybrid with 5 Gated slices as input) (55.18Hz), Stereo (RaftStereo+HRFormer) (6.74Hz), Stereo (RaftStereo) (9.28Hz).

6. Gated Reconstruction Consistency

In the following, before describing gated reconstruction consistency, we first demonstrate the accuracy of the proposed ambient and albedo estimation network – both components are required for the proposed gated reconstruction loss. Figure 3 shows examples of the input gated slices, the estimated depths, the estimated ambient, and albedo maps for representative scenes. The first two examples show results in nighttime scenes and the last three examples in daytime conditions. In the first two examples, the estimated mean ambient light is correctly estimated as very close to zero. Only single light sources are visible in the ambient image, such as car headlights. In daytime conditions, the ambient component is strongly present and takes a good portion of the total intensity of the gated slices. The estimated albedo maps are consistent in both day and night time conditions. Here, it is also important to note that the shadow visible behind close objects such as pedestrian and vehicles are due to the non-negligible distance between gated camera and illuminator, mounted at different positions of the vehicle as shown in Figure 4 of the main paper.

With adequately estimated albedo $\tilde{\alpha}$ and ambient Λ terms in hand, the gated reconstruction can be derived from the gated image formation, that is,

$$\tilde{I}_v^k(z) \;=\; \tilde{\alpha}\, C_k(z) + \tilde{\Lambda} + D_v^k, \tag{4}$$

with I being the outputted image, k the slice number, v the image view, $\tilde{\alpha}$ being the albedo, C_k the range intensity profile, z the scene distance, $\tilde{\Lambda}$ being the passive and D_v^k the dark current.

Here, the estimated ambient $\hat{\Lambda}$ captures the measurement noise,

$$\hat{\Lambda} = \eta_g + \eta_p + \Lambda_{noiseless}, \tag{5}$$

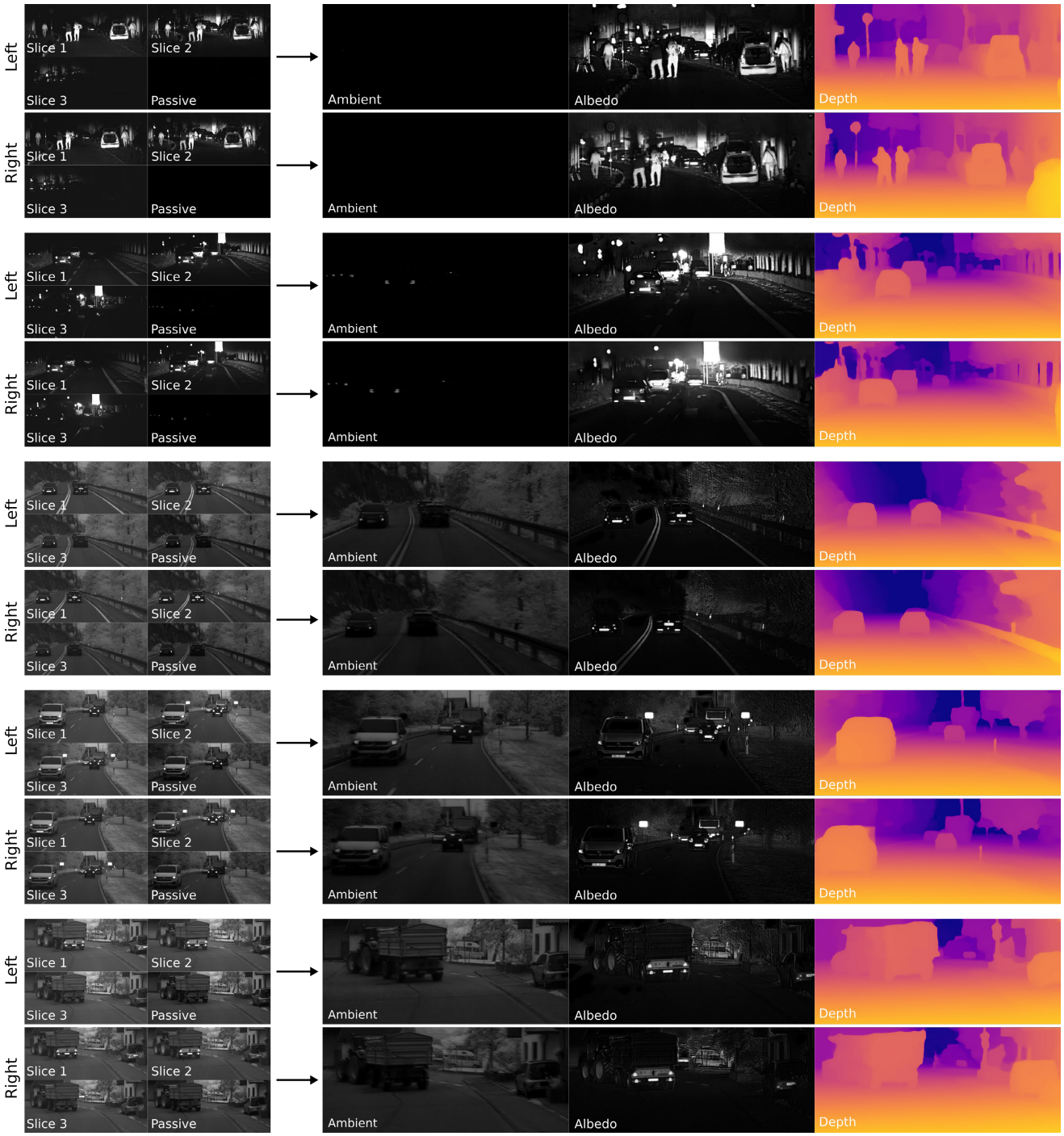


Figure 3. Qualitative examples of the three input gated slices and their passive component, and the corresponding predicted ambient, albedo, and depth for left and right images. Since we use a gated reconstruction loss [23] for additional depth supervision, as a by-product, our method reconstructs albedo and ambient illumination of the scene in addition to the depth information. While albedo represents the NIR reflectivity of objects and the laser illumination, the ambient image captures sunlight and active light sources.

where η_p models the signal-dependent Poisson photon shot noise and η_g Gaussian read-out noise [5].

The estimation of the ambient $\tilde{\Lambda}$ can be directly supervised through the passive captured gated images I_4, I_5 and the estimated Λ^{k_0}. Specifically, we use the formulation derived in the main manuscript for the ambient consistency,

$$\Lambda^{k_0} = \mu_k(I^4 + I^5 - D_v^4 - D_v^5)/(\mu_4 + \mu_5), \tag{6}$$

	Night				Day			
	Laser duration	Gate duration	Delay ξ	Pulses	Laser duration	Gate duration	Delay ξ	Pulses
I^1	240 ns	220 ns	260 ns	202	240 ns	220 ns	260 ns	101
I^2	280 ns	420 ns	400 ns	591	280 ns	420 ns	400 ns	296
I^3	370 ns	420 ns	750 ns	770	370 ns	420 ns	750 ns	385
I^4	-	1000 ns	-	805	-	120 ns	-	175
I^5	-	1000 ns	-	1745	-	120 ns	-	900

Table 5. Definitions of the gating parameters that we use for the experimental acquisition in this work.

with exposure times $\mu_{4,5,k}$ for the passive images and k for the active gated slices. With this in mind, the ambient term can be directly supervised using the photometric loss $\mathcal{L}_p$ [6], with Structural Similarity (SSIM) [24] and $\mathcal{L}_1$ norm,

$$\mathcal{L}_p(\tilde{\Lambda}, \Lambda^{k_0}) = 0.85 \cdot \frac{1-SSIM(\tilde{\Lambda}, \Lambda^{k_0})}{2} + 0.15 \cdot ||_{\tilde{\Lambda}-\Lambda^{k_0}||_1}. \tag{7}$$

For each slice, we use the parameters camera gate, laser duration, laser delay and number of pulses as in [9, 23]. The parameters are reported in Table 5. The profiles $C_k(z)$ are measured experimentally with calibrated targets and approximated with Chebyshev polynomials T_n,

$$T_0 = 1, \quad T_1 = x, \quad T_{n+1} = 2xT_n - T_{n-1}, \tag{8}$$

up to order of $N = 6$.

To guide the loss better when reconstructing the gated images we employ the mask M_g from [23] which filters for low signal-to-noise areas, saturated areas, and multi-path effects, not modeled by Eq. 4. We derive the specific mask employed in our loss in the following.

Signal-to-noise Ratio. Areas with little to no illumination, e.g., due to occlusion, the pixel variation between the three active slices is minimal. For each pixel p_{uv} at coordinate position (u, v) we can therefore define the binary mask M_{SNR},

$$M_{SNR}(u,v)\{\,(u,v) \mid \left(\max_i \left(p^i_{uv}\right) - \min_i \left(p^i_{uv}\right)\right) > \theta\,\}. \tag{9}$$

Saturated Pixels. Saturated pixels are caused by retro reflectors and introduce a non-linearity clipping all intensity values to the image maximum value. This causes that the underlying true albedo cannot be estimated. Such pixels are suppressed by the mask $S(u, v)$,

$$S(u,v)\{\,(u,v) \mid \max_i(p^i_{uv}) < \gamma\,\}. \tag{10}$$

Multipath Correction. We also remove multipath effects from the cyclic supervision. In most cases, such multi-path effects are the second-order reflections from strong reflectors as traffic signs on the road surface. Such points are projected underneath the road surface. Using the camera intrinsics K, we are able to estimate a constant ground plane with normal n and height h. Using the predicted depth z, we can filter which points are projected underneath the road and should be omitted for supervision. We define the set E of incorrectly predicted pixels (u, v) as,

$$E(u,v)\{\,(u,v) \mid (\mathbf{z}K^{-1}x_t)n < h\,\}, \tag{11}$$

where $x_t = [u, v, 1]$ represent the homogeneous pixel coordinate.

The final gated reconstruction mask M_g then is defined as,

$$M_g(u,v) = \begin{cases} 1 & \text{if } (u,v) \notin E(u,v) \wedge (u,v) \in S(u,) \wedge (u,v) \in M_{SNR}(u,v) \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

The gated reconstruction loss is defined as,

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

7. Ambient Estimation and Suppression

We briefly describe the method used to modulate the ambient light component, illustrated in Figure 4. Starting from the HDR passive captures I^4 and I^5, the ambient illumination term Λ^{k_0} is computed. Then this estimate is used to increase or decrease the ambient light in all frames, depending on the factor μ_s uniformly sampled from the interval $[0.5\mu_k, 1.5\mu_k]$, where μ_k is the exposure time of the active slices. This is equivalent to capturing between 50% and 150% of the original ambient component. To increase the accuracy of the ambient light modulation, we consider additionally the dark level (also referred as dark current) D_v^k. The D_v^k is assumed to be solely dependent on the camera and gating settings, and was calibrated offline.

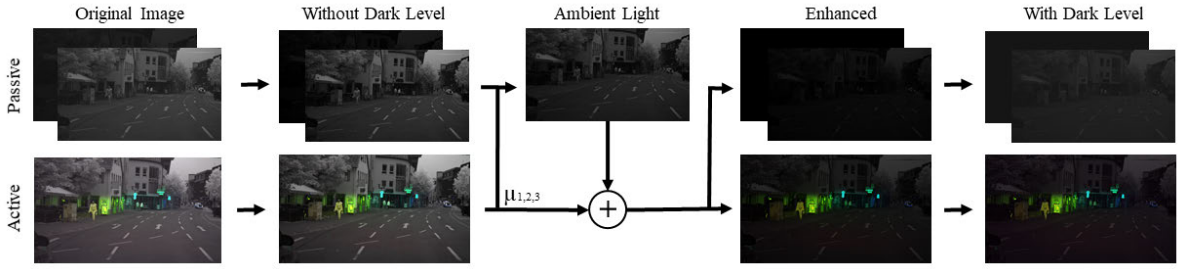


Figure 4. Illustration of the suppression of ambient illumination by estimation and subtracting the ambient term, under the consideration of the dark level, from the gated slice measurements.

Figure 5 reports examples of scene captures with an increased or decreased ambient light component. The second and third

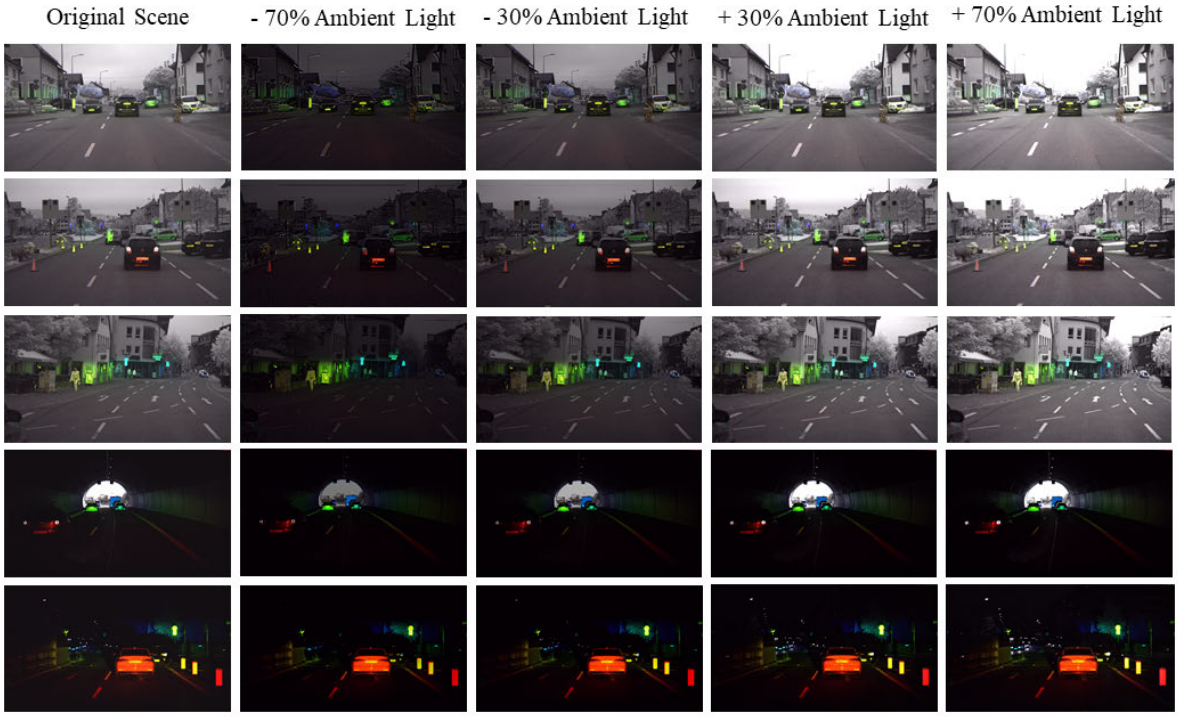
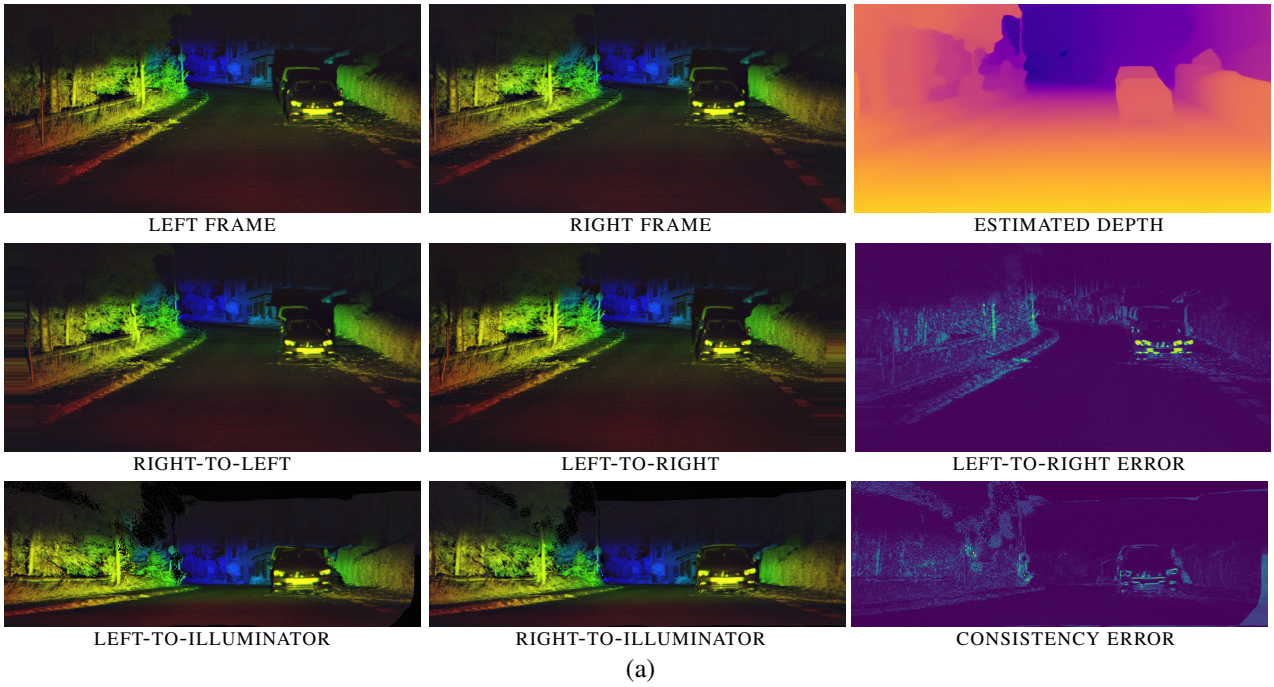


Figure 5. Example measurements with modulated ambient light components using the estimated ambient illumination term.

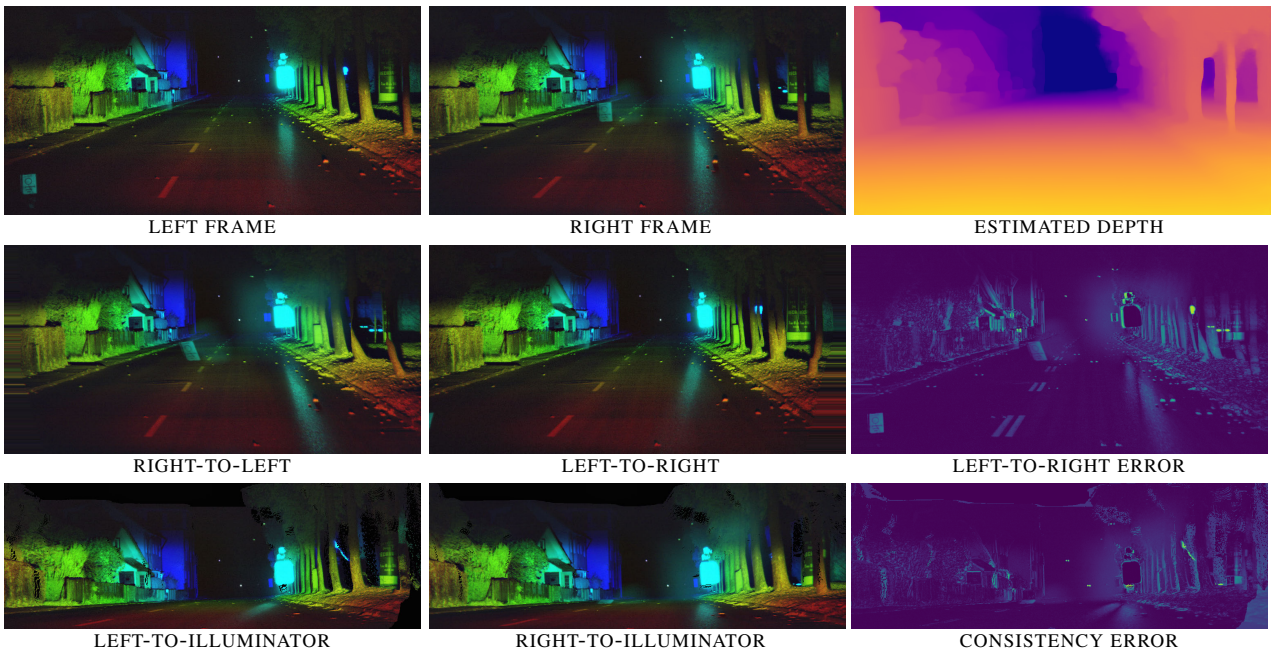
columns of the first three rows demonstrate that this process allows to modify images taken in normal daylight conditions to get captures closer to (less frequent) twilight or dawn conditions. On the other hand, as can be seen in the last column, a strong increase in ambient light might lead to a saturation of the capture and hence loss of details.

8. Illuminator View Consistency vs. Left-Right Warping

In this section, we illustrate the advantage of the illuminator view consistency. In Figure 6 qualitative warping results are depicted, showing left-right, camera-to-illuminator, and error maps. Here in sample (a) and (b), the shadow cast by the vehicle leads to inaccuracies in the left-right warping (second row), while it is completely absent in the illuminator view (third row). This behavior is because the shadows cast by the emitted light from the illuminator field of view are not visible from its view. Hence, the matching is more stable. Please also note, especially in the last example, the presence of multi-path effects on the roads, which are not the same between left and right frames. Due to this inconsistency, such effects are clearly visible in the left-right warping error. On the other hand, these effects are minor in the case of illuminator projections.



(a)



(b)

Figure 6. Examples of left-right and illuminator projections computed using the estimated depth map with corresponding error maps. For shadow areas and multi-path regions, the illuminator projections have less errors than the left-right warps.

9. Fusion Network for Mono and Stereo

In this section, we describe our approach toward fusing monocular and stereo depth. We also experimentally validate the effectiveness of our fusion module.

9.1. Depth Fusion

Stereo depth estimation computes depth by finding dense correspondence of pixels in left and right views. Given a pair of rectified stereo images with baseline B, we compute the disparity d for each pixel in the reference image, then the stereo depth z^s is calculated using $\frac{fB}{d}$. Although current state-of-the-art methods provide accurate depth estimates, stereo methods usually struggle to find correct correspondences in occluded and texture-less regions of the scene. Furthermore, the depth accuracy also depends on the baseline between the stereo camera setup. Monocular depth estimation relies on image-level cues like texture variations, gradients, occlusions, object sizes, etc. to estimate scale-ambiguous depth estimates of the scene.

To combine the strengths of two depth estimation techniques, we propose a fusion network that acts as an additional refinement stage to improve the final predicted depth. The architecture of the fusion network is discussed in Section 3.4. The central idea behind fusing the two depth maps is that monocular depth estimates are often more reliable in occluded regions, whereas non-occluded regions, especially far-away objects which are visible in both left and right views have more accurate depth estimate in the stereo depth modality. Figure 7 presents qualitative examples showing how monocular and stereo depth can be fused using occlusion masks to construct a pseudo ground-truth for the fusion stage training. In addition to quantitative improvements, we have presented a set of qualitative examples in Section 9.2.

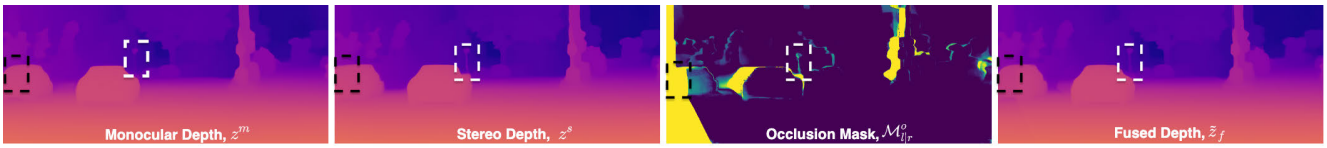
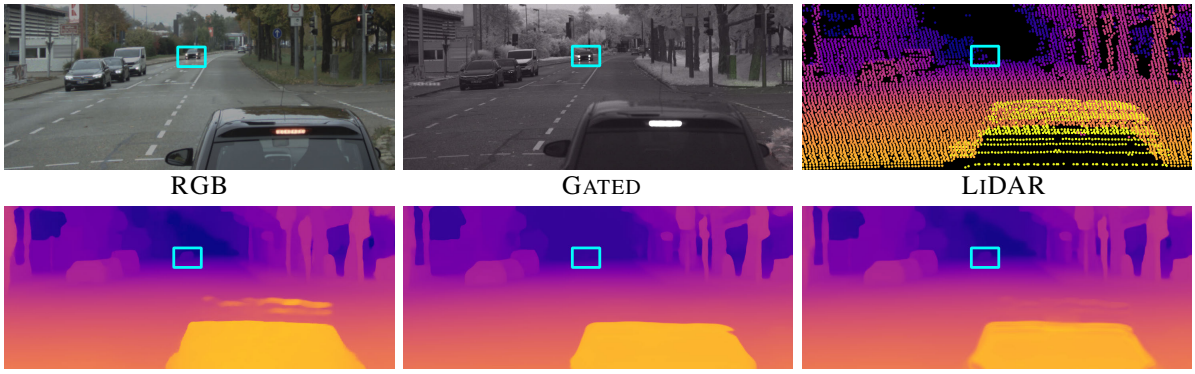


Figure 7. Qualitative example showing how occlusion masks can be used to fuse monocular depth z^m and stereo depth z^s,i.e. $\tilde{z}_f = \mathcal{M}^o_{l|r} z^m + (1 - \mathcal{M}^o_{l|r}) z^s$. Stereo depth estimation fails to correctly estimate depth in occluded and untextured regions (incomplete structure annotated with black box), whereas monocular depth performs poorly for structures located at far distance (indicated with a white box).

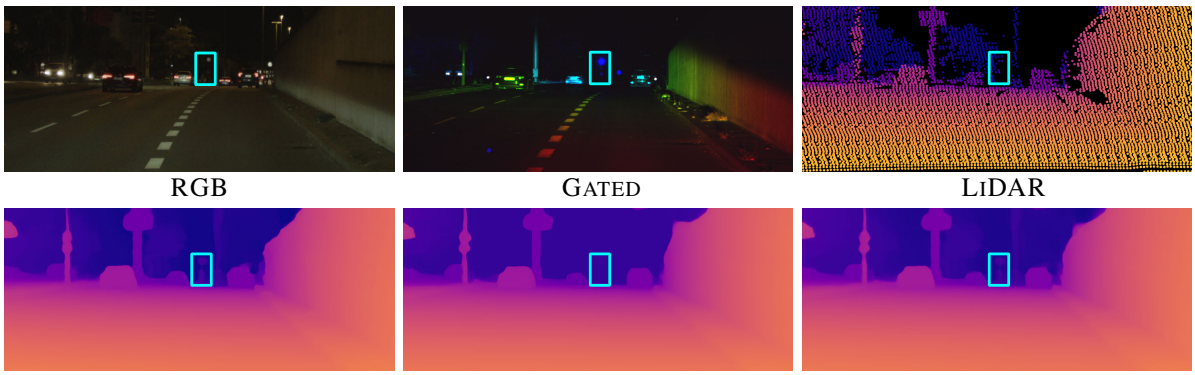
9.2. Qualitative Evaluation of Stereo-Mono Fusion

In this section, we provide qualitative examples to show how our fusion approach leads to significant qualitative improvements in depth estimation. The following Figs. 8 and 9 validate the effectiveness of the proposed fusion module at long and close distances.

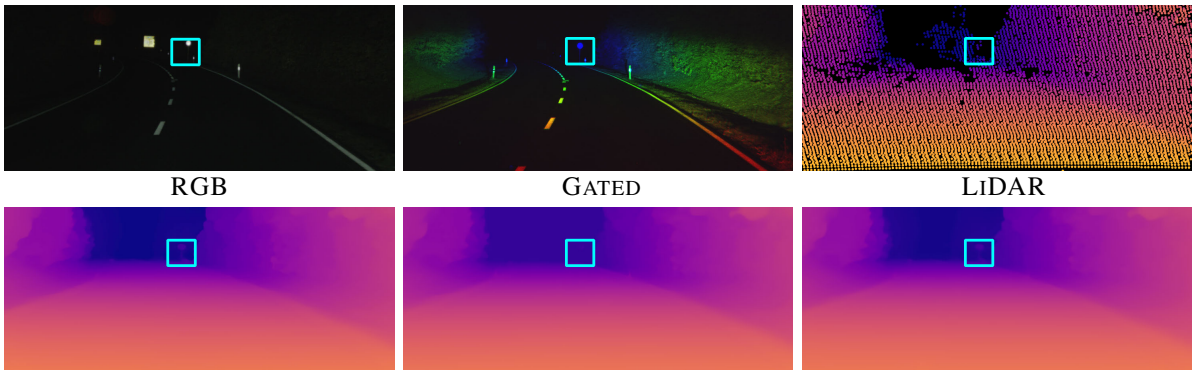
Fusion at Long Range



(a) Even in daylight conditions, a car located at far-away distance is only visible in the stereo and fused depth



(b) Far-away signboard is merged with background for monocular depth but visible in stereo and fused depth.



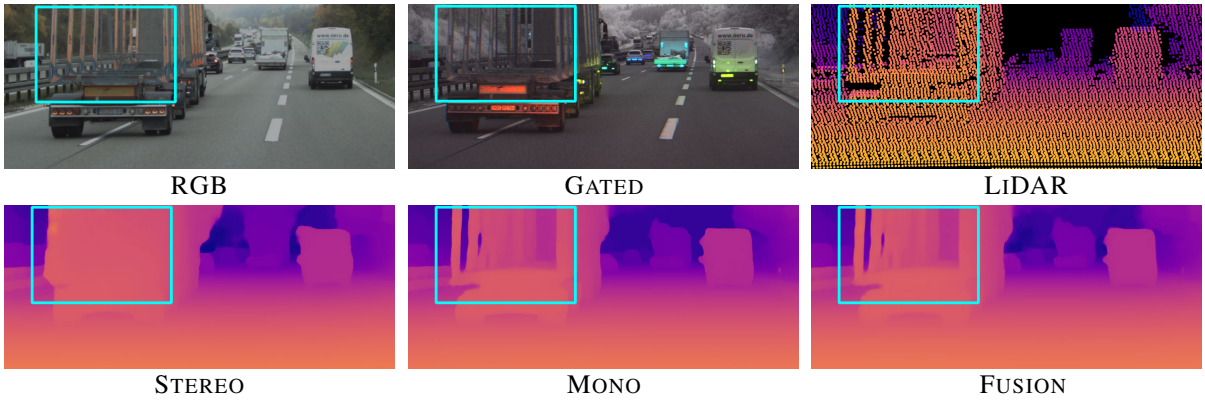
(c) Stereo and Fused depth maps capture thin structure located at long range.

Figure 8. Qualitative examples (a-c) showing significant improvements in depth maps for objects located at far distances. Since stereo-based depth is better suited for objects located at long distances, our fusion module is able to implicitly emphasize more on stereo-depth information for such objects, as shown by the marked regions.

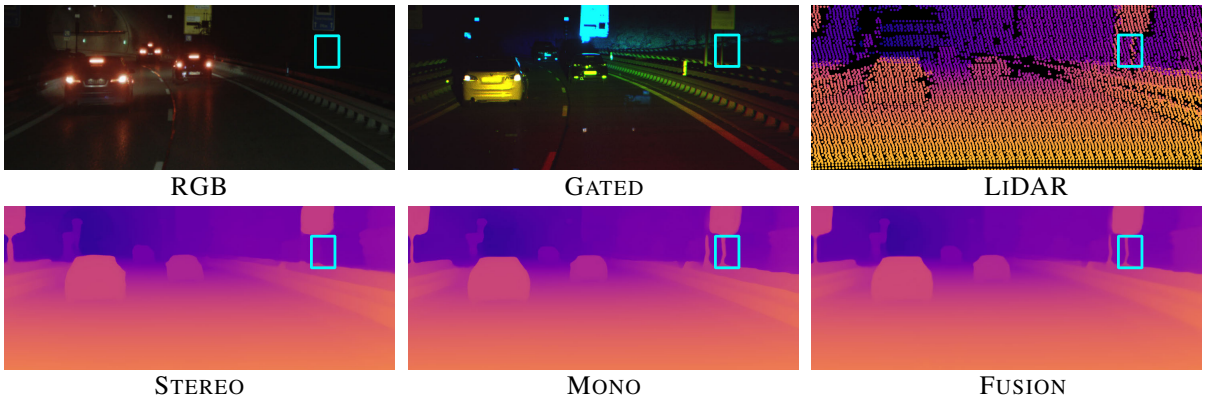
10. Additional Qualitative Evaluation

In this section, we present additional qualitative results of the proposed Gated Stereo and other state-of-the-art methods, including monocular gated [9, 23], monocular RGB [7, 10, 13, 19], sparse-depth-completion [17, 25], and stereo RGB [15, 26, 27] approaches. Figure 10, 11, 12, 13, 14, and 15 show the depth map predictions of the different approaches with the

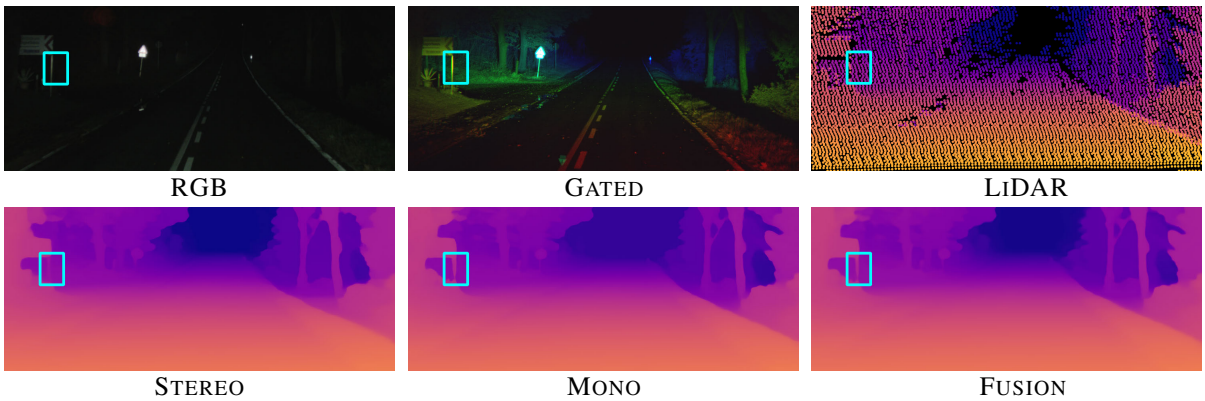
Close-Range Fusion for Thin Structures



(a) Close-range thin structures are better captured by monocular and fused depth in daylight conditions.



(b) For night, signboard pole is more visible in monocular and fused depth.



(c) For close range illuminated structures, stereo depth performs poorly as compared to monocular and fused depth.

Figure 9. Qualitative examples (a-c) showing close-range depth-map improvements. Using monocular depth, our fusion network is able to improve the depth estimates for close-range thin objects which are not captured by the stereo-based branch, as shown by marked regions.

corresponding RGB and gated image as well as the LiDAR measurements. The results demonstrate that our method is able to predict much sharper edges and less washed-out depth estimates than the other methods, especially for far distances. For example in Figure 10 and 12, Gated Stereo is the only method that is able to provide distinct object contours for the far away cars in the scenes. Furthermore, due to the additional HDR-like passive input, Gated Stereo is able to handle sunlight and maintain its performance even in the presence of strong ambient light. This is shown in Figure 13, 14, and 15, where Gated Stereo is able to provide depth maps with fine details including thin poles of traffic lights and traffic signs. We notice that the

self-supervised methods Packnet [10] and Monodepth2 [7] struggle to estimate correct depth values for moving objects (see Figure 11 and 15). Moreover, LiDAR depth completion methods [17, 25] are not able to interpolate plausible depth maps, revealed by truncated object edges, for example in Figure 14 and 15. In general, monocular and stereo RGB approaches are able to capture fine structures for close distances but have difficulty in determining exact depth for far distances. In contrast, the proposed method is able to provide accurate and detailed depth maps for near and far distances during day and night.

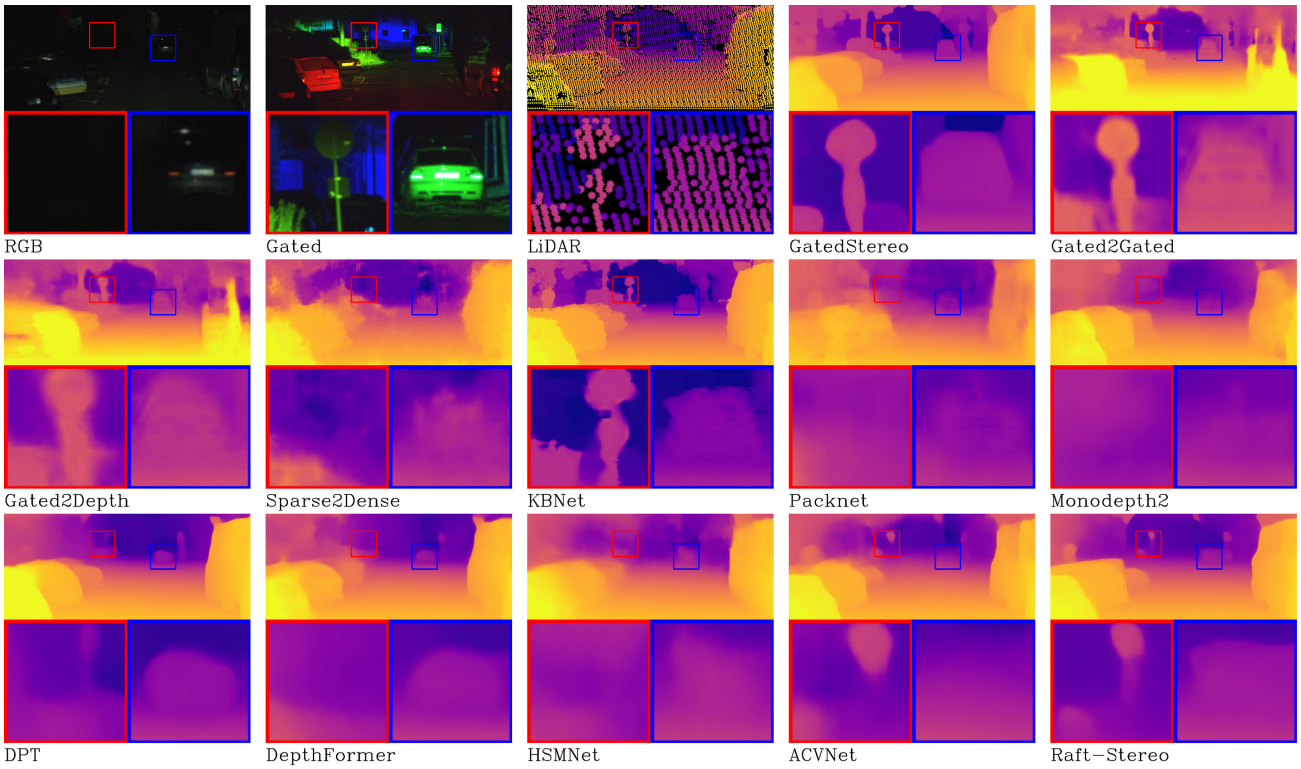


Figure 10. Qualitative comparison of the proposed Gated Stereo approach and state-of-the-art methods. For each example, we show the corresponding RGB image, the colored gated image, and the LiDAR measurements. Gated Stereo predicts fine details and sharper object contours than the other approaches.

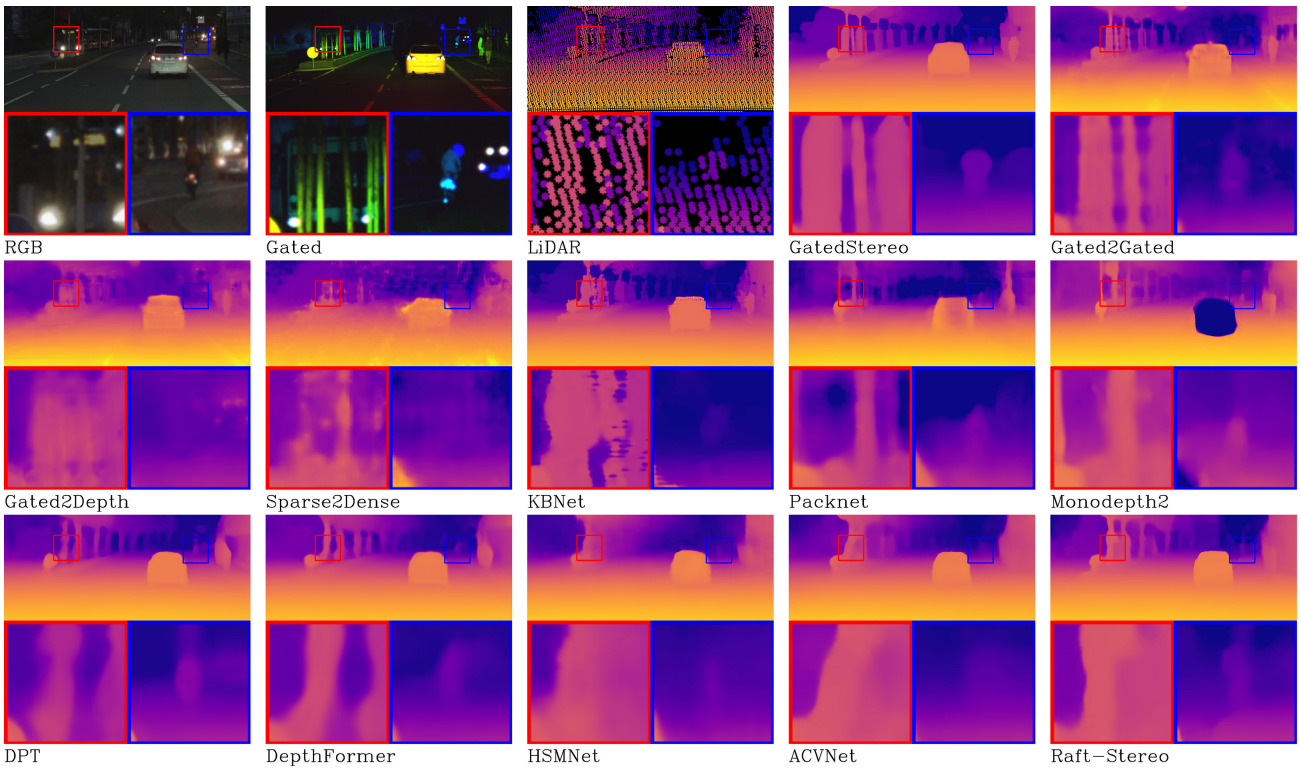


Figure 11. Additional qualitative comparison of our Gated Stereo approach and state-of-the-art methods for thin structure (red) and distant small object (blue).

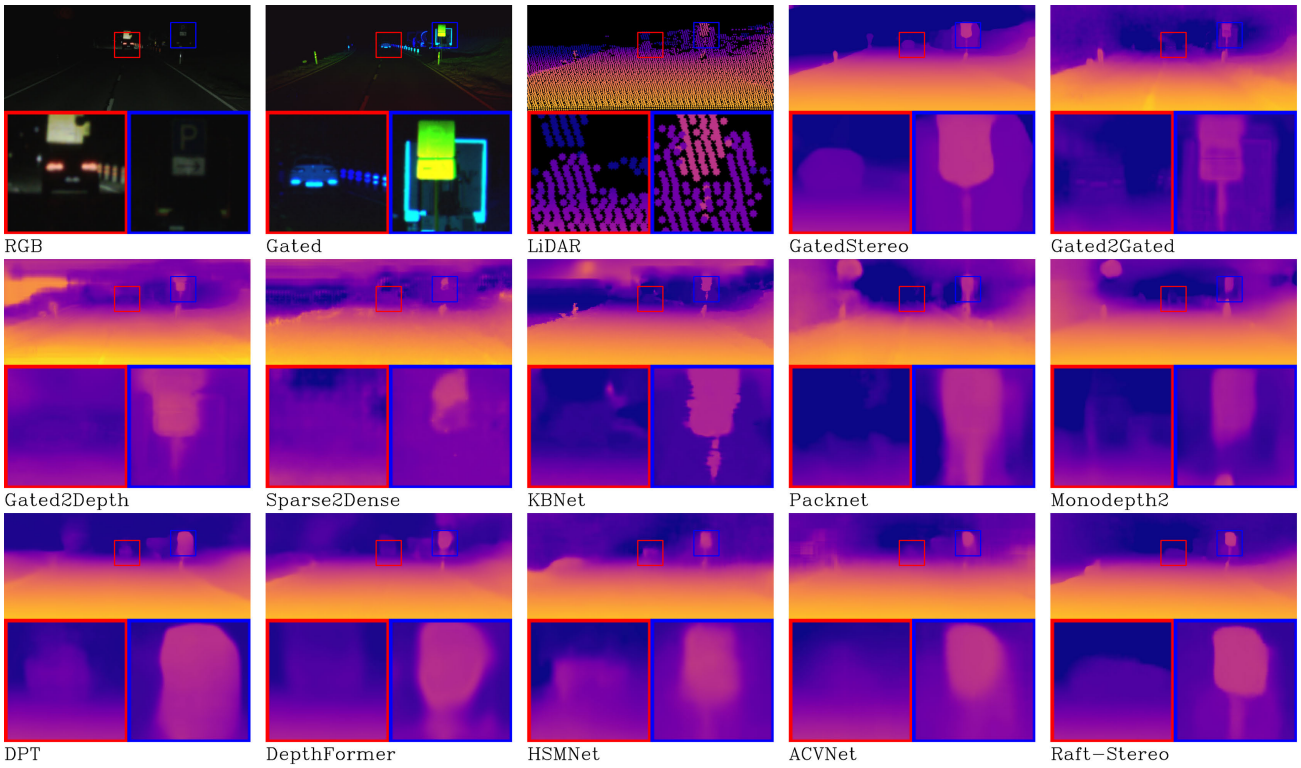


Figure 12. Additional qualitative comparison of our Gated Stereo approach and state-of-the-art methods.

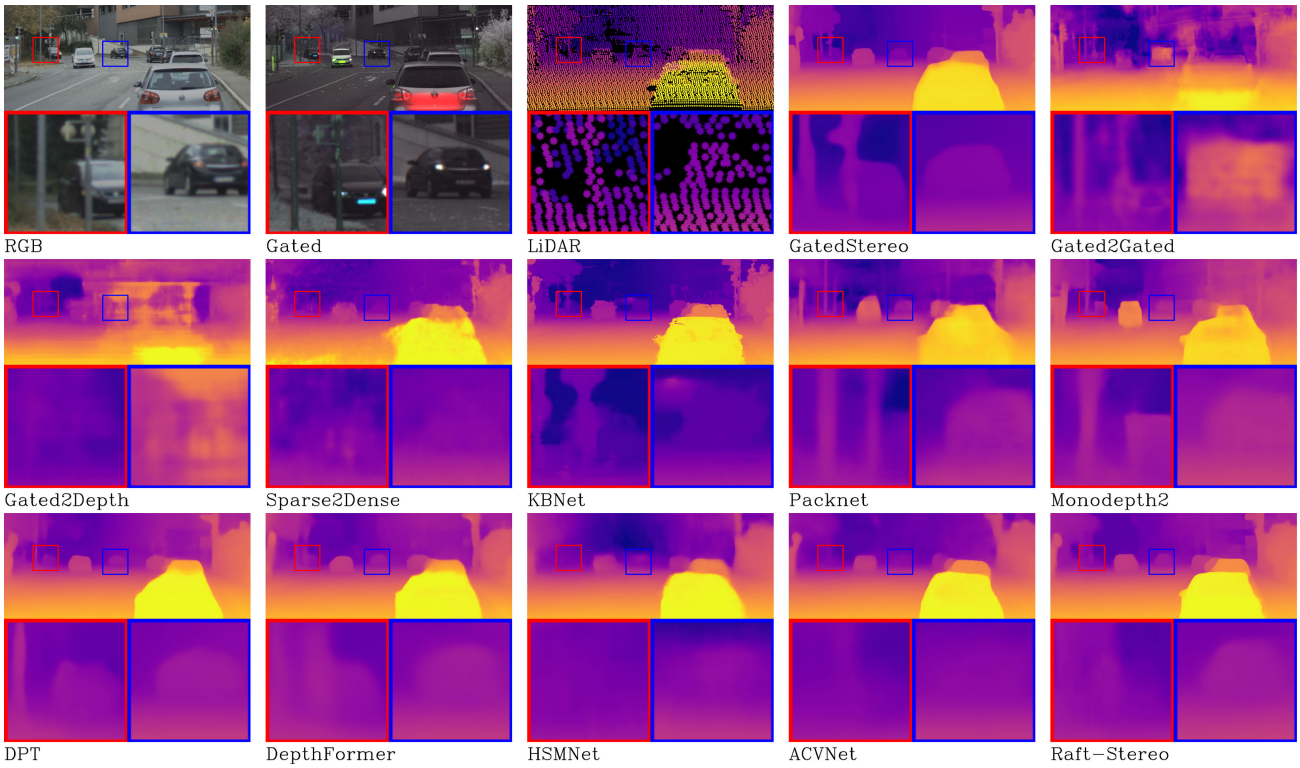


Figure 13. Additional qualitative comparison of our Gated Stereo approach and state-of-the-art methods.

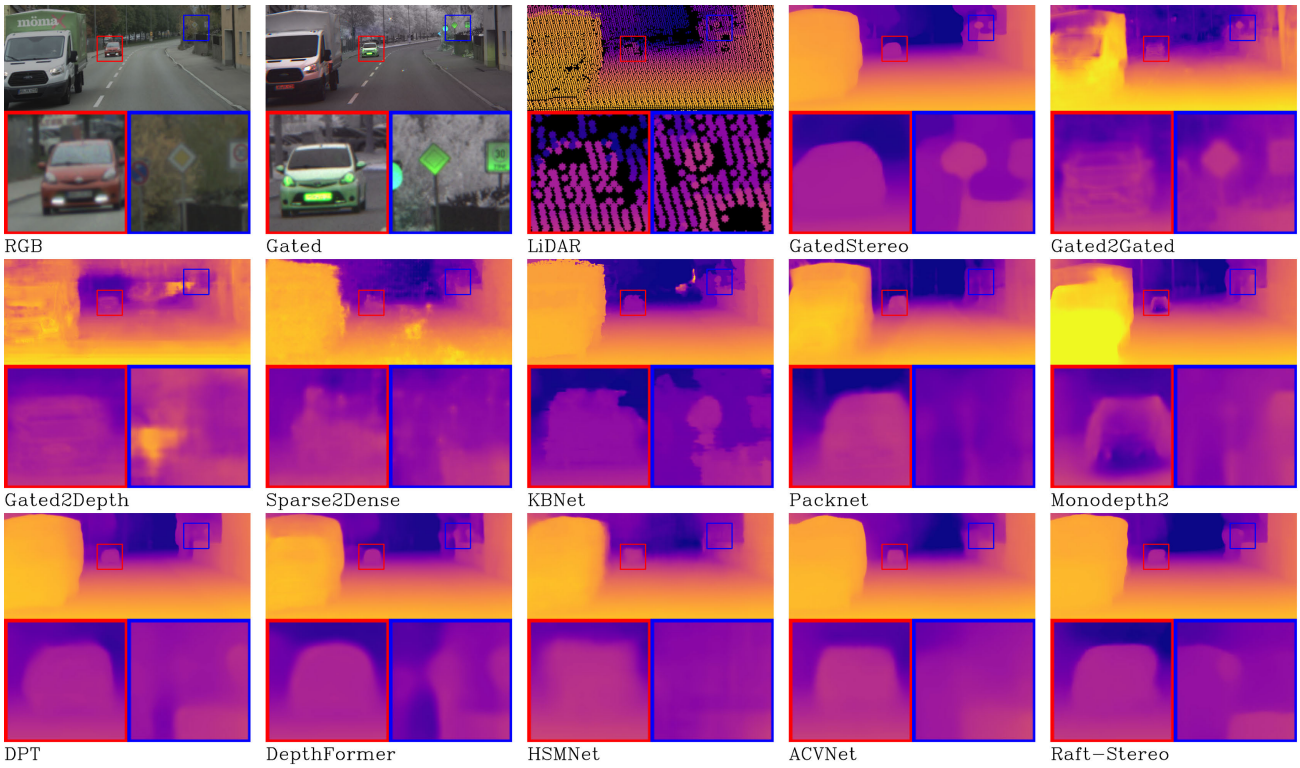


Figure 14. Additional qualitative comparison of our Gated Stereo approach and state-of-the-art methods.

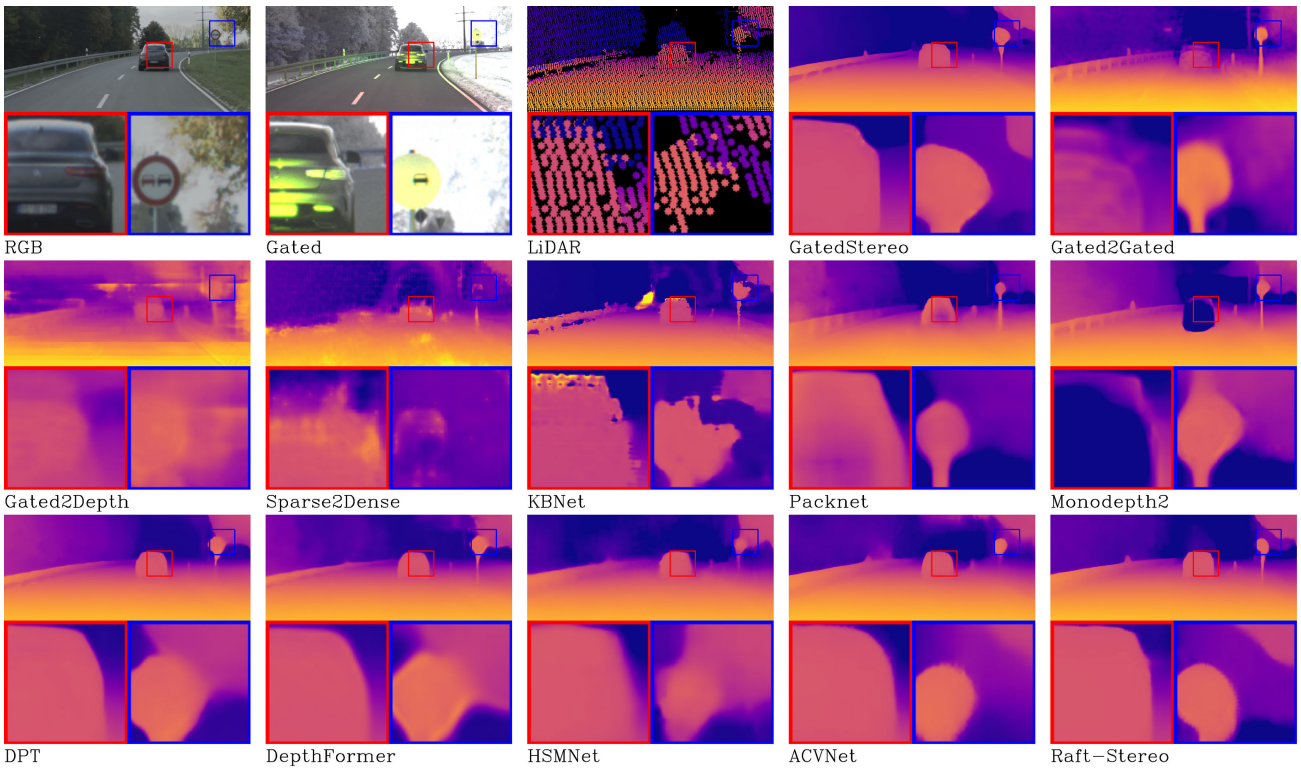


Figure 15. Additional qualitative comparison of our Gated Stereo approach and state-of-the-art methods.

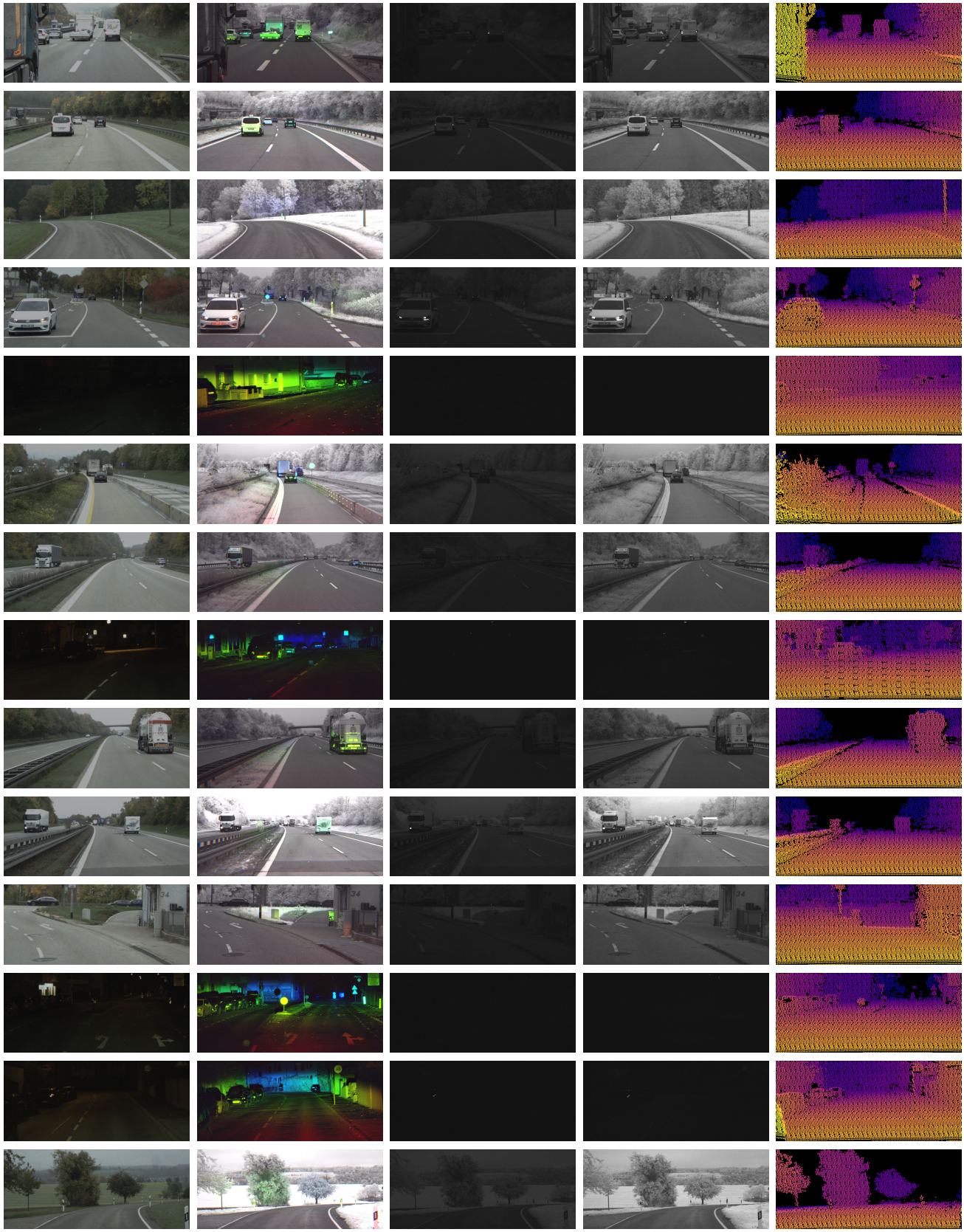


Figure 16. Random samples of our proposed Gated Stereo dataset to illustrate the diversity of scenes, illumination, and sensor modalities. From left to right: RGB; gated I^k with red for I^1, green for I^2, and blue for I^3; gated passive with low exposure time I^4; gated passive with high exposure time I^5; LiDAR.

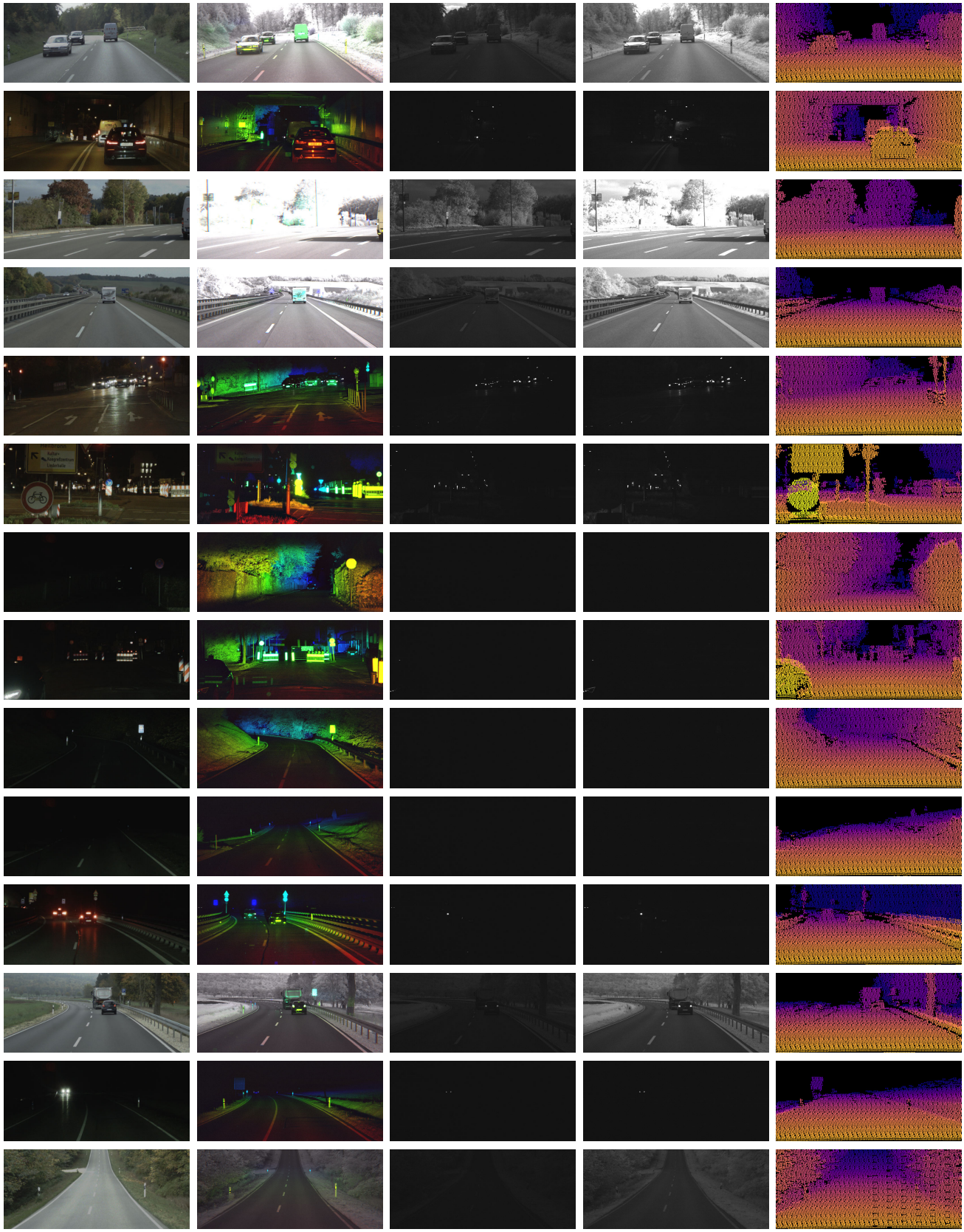


Figure 17. Random samples of our proposed Gated Stereo dataset to illustrate the diversity of scenes, illumination, and sensor modalities. From left to right: RGB; gated I^k with red for I^1, green for I^2, and blue for I^3; gated passive with low exposure time I^4; gated passive with high exposure time I^5; LiDAR.

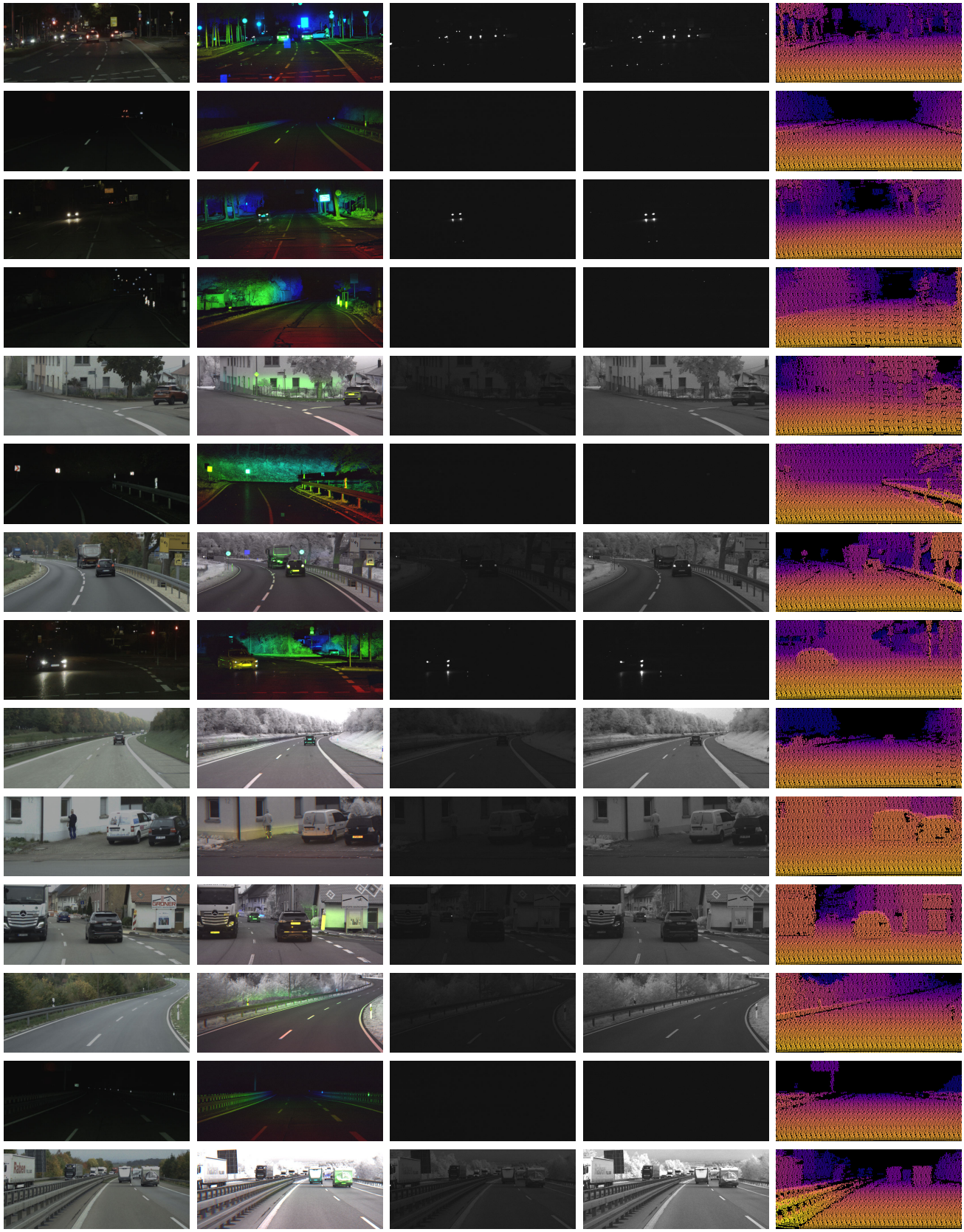


Figure 18. Random samples of our proposed Gated Stereo dataset to illustrate the diversity of scenes, illumination, and sensor modalities. From left to right: RGB; gated I^k with red for I^1, green for I^2, and blue for I^3; gated passive with low exposure time I^4; gated passive with high exposure time I^5; LiDAR.

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. 4, 7

[2] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 2

[4] 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. 4, 7

[5] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE TIP*, 17(10):1737–1754, 2008. 8

[6] 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. 9

[7] 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. 3, 4, 7, 14, 16

[8] Tobias Gruber, Mario Bijelic, Felix Heide, Werner Ritter, and Klaus Dietmayer. Pixel-accurate depth evaluation in realistic driving scenarios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 95–105. IEEE, 2019. 3

[9] 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. 3, 4, 9, 14

[10] 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. 4, 7, 14, 16

[11] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Towards precise and efficient image guided depth completion. 2021. 4, 7

[12] 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. 4, 7

[13] 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. 3, 4, 7, 14

[14] 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. 7

[15] 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. 3, 4, 7, 14

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

[17] 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. 3, 4, 7, 14, 16

[18] 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. 4, 7

[19] 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. 4, 5, 7, 14

[20] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 4

[21] 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. 4, 7

[22] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 3

[23] 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. 3, 4, 5, 8, 9, 14

[24] 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. 9

[25] 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. 4, 7, 14, 16

[26] 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. 4, 7, 14

[27] 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. 4, 7, 14

[28] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. 2021. 4