Gated Stereo:
Joint Depth Estimation from Gated and Wide-Baseline Active Stereo Cues

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

Abstract

We propose Gated Stereo, a high-resolution and long-range depth estimation technique that operates on active gated stereo images. Using active and high dynamic range passive captures, Gated Stereo exploits multi-view cues alongside time-of-flight intensity cues from active gating. To this end, we propose a depth estimation method with a monocular and stereo depth prediction branch which are combined in a final fusion stage. Each block is supervised through a combination of supervised and gated self-supervision losses. To facilitate training and validation, we acquire a long-range synchronized gated stereo dataset for automotive scenarios. We find that the method achieves an improvement of more than 50 % MAE compared to the next best RGB stereo method, and 74 % MAE to existing monocular gated methods for distances up to 160 m. Our code, models and datasets are available here[1].

1. Introduction

Long-range high-resolution depth estimation is critical for autonomous drones, robotics, and driver assistance systems. Most existing fully autonomous vehicles strongly rely on scanning LiDAR for depth estimation [51, 52]. While these sensors are effective for obstacle avoidance the measurements are often not as semantically rich as RGB images. LiDAR sensing also has to make trade-offs due to physical limitations, especially beyond 100 meters range, including range range versus eye-safety and spatial resolution. Although recent advances in LiDAR sensors such as, MEMS scanning [60] and photodiode technology [58] have drastically reduced the cost and led to a number of sensor designs with $\approx$ 100 - 200 scanlines, these are still significantly lower resolutions than modern HDR megapixel camera sensors with a vertical resolution more than $\approx$ 5000 pixels. However, extracting depth from RGB images with monocular methods is challenging as existing estimation methods suffer from a fundamental scale ambiguity [16]. Stereo-based depth estimation methods resolve this issue but need to be well calibrated and often fail on texture-less regions and in low-light scenarios when no reliable features, and hence triangulation candidate, can be found.

To overcome the limitations of existing scanning LiDAR and RGB stereo depth estimation methods, a body of work has explored gated imaging [2, 7–9, 22, 27]. Gated imagers integrate the transient response from flash-illuminated scenes in broad temporal bins, see Section 3 for more details. This imaging technique is robust to low-light, and adverse weather conditions [7] and the embedded time-of-flight information can be decoded as depth. Specifically, Gated2Depth [23] estimates depth from three gated slices and learns the prediction through a combination of simulation and LiDAR supervision. Building on these findings, recently, Walia et al. [59] proposed a self-supervised training approach predicting higher-quality depth maps. However, both methods have in common that they often fail in conditions where the signal-to-noise ratio is low, e.g., in the case of strong ambient light.

We propose a depth estimation method from gated stereo observations that exploits both multi-view and time-of-flight cues to estimate high-resolution depth maps. We propose a depth reconstruction network that consists of a monocular depth network per gated camera and a stereo network that utilizes both active and passive slices from the gated stereo pair. The monocular network exploits depth-dependent gated intensity cues to estimate depth in monocular and low-light regions while the stereo network relies on active stereo cues. Both network branches are fused in a learned fusion block. Using passive slices allows us to perform robustly under bright daylight where active cues have a low signal-to-noise ratio due to ambient illumination. To train our network, we rely on supervised and self-supervised losses tailored to the stereo-gated setup, including ambient-aware and illuminator-aware consistency along with multi-camera consistency. To capture training data and assess the method, we built a custom prototype vehicle and captured a stereo-gated dataset under different lighting conditions and automotive driving scenarios in urban, suburban and highway environments across 1000 km of driving.

Specifically, we make the following contributions:

- We propose a novel depth estimation approach using gated stereo images that generates high-resolution

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

dense depth maps from multi-view and time-of-flight depth cues.

- We introduce a depth estimation network with two different branches for depth estimation, a monocular branch and a stereo branch, that use active and passive measurement, and a semi-supervised training scheme to train the estimator.
- We built a prototype vehicle to capture test and training data, allowing us to assess the method in long-range automotive scenes, where we reduce the MAE error by 50 % to the next best RGB stereo method and by 74 % on existing monocular gated methods for distances up to 160 m.

2. Related Work

Depth from Time-of-Flight. Time-of-Flight (ToF) sensors acquire depth by estimating the round travel time of light emitted into a scene and returned to the detector. Broadly adopted approaches to time-of-flight sensing include correlation time of flight cameras [26, 32, 33], pulsed ToF sensors [51] and gated illumination with wide depth measuring bins [22, 27]. Correlation time-of-flight sensors [26, 32, 33] flood-illuminate a scene and estimate the depth from the phase difference of the emitted and received light. This allows for precise depth estimation with high spatial resolution but due to its sensitivity to ambient light existing correlation time-of-flight detectors have been limited to indoor applications. In contrast, pulsed light ToF systems [51] measure the round trip time directly from a single light pulse emitted to a single point in the scene. Although a single point measurement offers high depth precision and signal-to-noise ratio, this acquisition process mandates scanning to allow long outdoor distances and, as such, drastically reduces spatial resolution in dynamic scenes. In addition, pulsed LiDAR measurements can drastically degrade in adverse weather [6, 10, 30] due to backscattered light from fog or snow. Gated cameras [7, 22, 27] accumulate flood-illuminated light over short temporal bins limiting the visible scene to certain depth ranges. As a result, gated cameras gate-out backscatter and at short-range [7] and reconstruct coarse depth [2, 8, 9].

Depth Estimation from Monocular and Stereo Intensity Images. Depth estimation from single [19,25,37,38], single images with sparse LiDAR points [28,47,53,54,62], stereo image pairs [3, 11, 39, 65] or stereo with sparse LiDAR [14, 68] is explored in a large body of work. Monocular depth imaging approaches [38] offer low cost when a single CMOS camera is used, reduced footprint, especially compared to LiDAR systems, and, hence, also can be applied across application domains. However, monocular depth estimation methods inherit a fundamental scale ambiguity problem that can only be resolved by vehicle speed or LiDAR ground-truth depth measurements at test-time [25]. Stereo approaches, on the other hand, allow triangulating between two different views resolving the scale ambiguity [11]. As a result, these methods allow for accurate long-range depth prediction when active sensors are not present. To learn the depth prediction from stereo intensity images, existing methods employ supervised [11, 11, 16, 29, 31, 37, 38, 44, 45] and unsupervised learning techniques [17, 19, 21, 25, 70]. Supervised stereo techniques often rely on time-of-flight data [11, 16, 29, 44] or multi-view data [31, 38, 45] for supervision. As a result, the collection of suitable dense ground-truth data can be challenging. Specifically, existing work [18, 55] aims to compensate for the sparsity of LiDAR ground-truth measurements through ego-motion correction and acquisition of multiple point clouds. Moreover, such aggregated LiDAR "ground-truth" depth is incorrect in scattering media [5]. To tackle this challenge and exploit large datasets of video data without ground-truth LiDAR depth present, self-supervised stereo approaches exploit multiview geometry by aligning stereo image pairs [17, 19] or they make use of image view synthesis between temporally consecutive frames [21, 25, 70]. Garg et al. [17] train a network to predict disparities from monocular camera images by encouraging consistency when warped to stereo images. Followup work [19, 56] extends this idea to warp temporally consecutive stereo captures. To perform the warping correctly for these methods, two networks are necessary one predicting the depth and a second one predicting a rigid body transformation between two temporally adjacent frames. Existing work on depth prediction has investigated diverse neural architectures for depth estimation networks [3, 17, 21, 25, 37, 39, 65] and extensions in the loss formulation [15,19,21,25,43,49,57,66]. Recently, RAFT-Stereo [41] relies on iterative refinement over the cost volume at high resolution, thanks to the construction of a lighter cost volume and the employment of 2D convolution instead of 3D convolutions, which are memory and computationally intensive. All depth estimation methods discussed above, which are based on passive imaging, can fail in low-light or low-contrast scenarios that active gated methods [23] tackle using illumination. Alternate approaches employ sparse LiDAR measurements [14, 28, 47, 53, 54, 62, 68] not only for supervised training but also during inference time to overcome the scale ambiguity from monocular approaches, but they come with the drawback that temporal LiDAR distortions and scan pattern artefacts are passed through.

Depth Estimation from Gated Images. Gated depth estimation methods with analytical solutions guiding the depth estimation [34,35,63] had been first proposed over a decade ago. Recently, learned Bayesian approaches [1, 50] and approaches employing deep neural networks [23, 59] have achieved dense depth estimation at long-range outdoor sce-

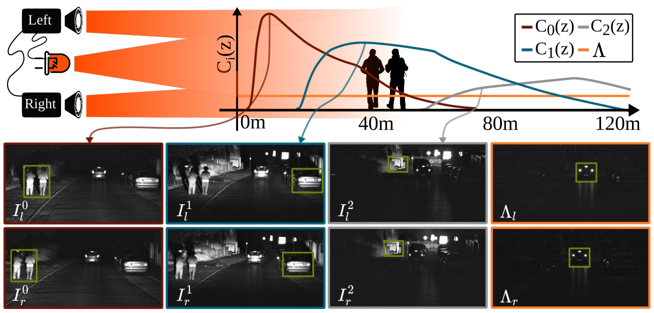


Figure 1. The proposed stereo gated camera consists of two gated cameras and a single flood-lit pulsed illumination source. Varying the delay between illumination and the synchronized cameras results in different range-intensity profiles C_k describing the pixel-intensity for distance z for each camera in addition to disparity d. For image formation in bright airlight, an additional passive component Λ is required. The resulting images for left and right camera positions illustrating gating and parallax in an example scene are illustrated at the bottom.

narios and in low-light environments. All of these existing methods rely on monocular gated imaging systems, which are able to deliver similar performances to passive color stereo approaches [23, 59]. Gruber et al. [23] introduce a fully supervised depth prediction network leveraging pre-training on fully synthetic data performing on par with traditional stereo approaches. Recently Walia et al. [59] proposed a self-supervised gated depth estimation method. Although their approach resolves the scale ambiguity, it still suffers in bright daylight in the absence of depth cue, and at long ranges due to depth quantization and lack of relative motion during training. In this work, we tackle these issues with a wide-baseline stereo-gated camera to estimate accurate depth in all illumination conditions and at long ranges.

3. Gated Stereo Imaging

This section introduces the proposed gated stereo camera. We propose a synchronized gated camera setup with a wide baseline of b =0.76 m. After flood-illuminating the scene with a single illuminator, we capture three synchronized gated and passive slices with two gated cameras. Synchronizing two gated cameras requires not only the trigger of individual single exposures as for traditional stereo cameras, but the transfer of gate information for each slice with nano-second accuracy. This level of synchronization allows us to extract slices with gated multi-view cues.

Specifically, after the emission of a laser pulse p at time $t = 0$, the reflection of the scene gets integrated on both camera sensors after a predefined time delay ξ identical on both cameras. Only photons arriving in a given temporal gate are captured with the gate function g allowing to integrate implicit depth information into 2D images. Following Gruber *et al.* [24], the distance-dependent pixel intensities are described by so-called range-intensity-profiles $C_k(z)$ which are independent of the scene and given by,

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

where $I^k(z,t)$ is the gated exposure, indexed by k for the slice index at distance z and time t; α is the surface reflectance (albedo), and β the attenuation along a given path due to atmospheric effects. Both image stacks are rectified and calibrated such that epipolar lines in both cameras are aligned along the image width and disparities d can be estimated. Epipolar disparity is consistent with the distance $z = \frac{bf}{d}$, where f is the focal length, providing a depth cue across all modulated and unmodulated slices.

In the presence of ambient light or other light sources as sunlight or vehicle headlamps, unmodulated photons are acquired as a constant Λ and added to the Eq. 1,

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

Independently from ambient light, a dark current D_v^k depending on the gate settings is added to the intensity count,

$$I_v^k(z) = \alpha C_k(z) + \Lambda + D_v^k, \tag{3}$$

which we calibrate for each gate k and camera v. We adopt the Poisson-Gaussian noise model from [59]. In contrast to prior work [23, 59], we also capture two unmodulated passive exposures in an HDR acquisition scheme. So specifically, we use three gated exposures C_1, C_2, C_3 with the same profile as in [23] and two additional passive images without illumination, that is, $C_4 = C_5 = 0$, and HDR-like fixed exposure times of 21 µs and 108 µs at daytime and 805 µs and 1745 µs at night time. This allows us to recover depth simultaneously from stereo-gated slices and passive stereo intensity cues with the same camera setup. The proposed system captures these images at 120 Hz, natively, allowing for a per-frame update of 24 Hz, which is about 2× the update rate of recent commercial scanning LiDAR systems, e.g., Luminar Hydra or Velodyne Alpha Puck.

4. Depth from Gated Stereo

In this section, we propose a depth estimation method that exploits active and passive multi-view cues from gated images. Specifically, we introduce a joint stereo and monocular network that we semi-supervise this network using several consistency losses tailored to gated stereo data. In the following, we first describe the proposed network architecture before describing the semi-supervision scheme.

4.1. Joint Stereo-Mono Depth Network

The proposed depth estimation network is illustrated in Fig. 2, which has a stereo and monocular branches, and a

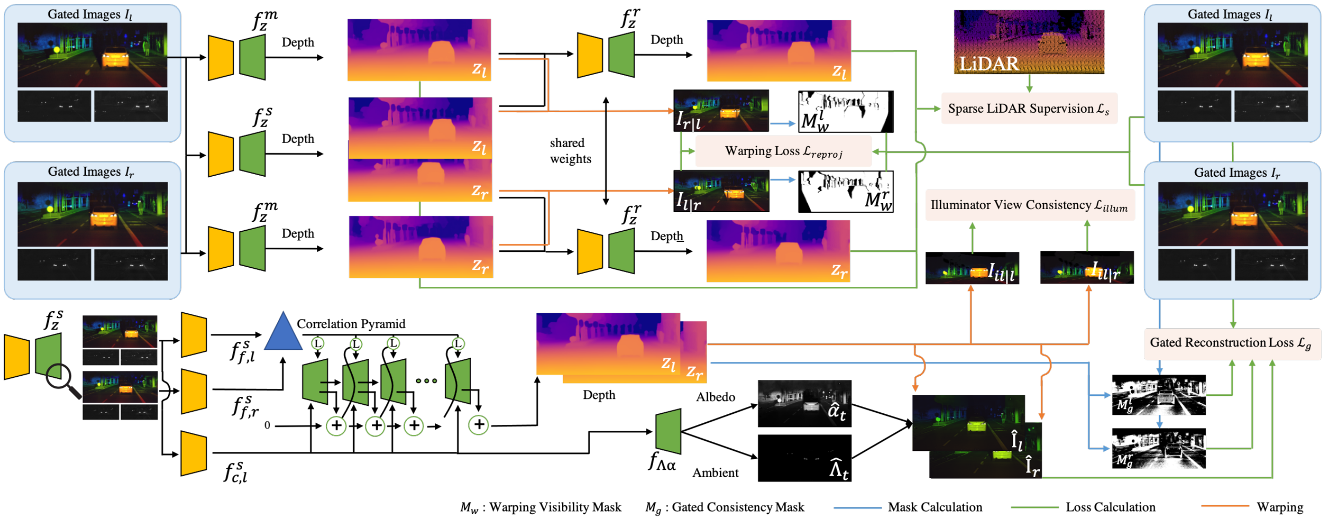


Figure 2. The proposed model architecture is composed of a stereo (f_z^s), two monocular (f_z^m), and two fusion (f_z^r) networks with shared weight. The fusion network combines the output of the monocular and stereo networks to obtain the final depth image for each view. Both stereo and monocular networks use active and passive slices as input, with the stereo network using the passive slices as context and includes a decoder ($f_{\Lambda\alpha}$) for albedo and ambient estimation which are used for gated reconstruction. The loss terms are applied to the appropriate pixels using masks that are estimated from the inputs and outputs of the networks.

final fusion network that combines the outputs from these branches to produce the final depth map.

Monocular Branch. The monocular network, $f_z^m : I \to z^m$, estimates absolute depth for a single gated image I from either of the two imagers. Unlike monocular RGB images, monocular gated images encode depth-dependent intensities which can be used by monocular depth networks to estimate *scale-accurate* depth maps [23, 59]. The proposed monocular gated network uses a DPT [48]-type architecture and outputs inverse depth bounded in $[0, 1]$ which results in absolute depth between $[1, \infty]$. For network details, we refer to the Supplemental Material.

Stereo Branch. The stereo branch, $f_z^s : (I_l, I_r) \to (z_l^s, z_r^s)$, estimates disparity from a pair of stereo images and outputs the depth for the left and right images z_l and z_r respectively. The network architecture is based on RAFT-Stereo [41] with all three active gated slices and two passive captures concatenated to a 5-dimensional input. The feature extractor is replaced with HRFormer [67], which is able to extract robust high-resolution features for downstream stereo matching. The left and right slice features $f_{f,l}^s$ and $f_{f,r}^s$ are given as input to the correlation pyramid module and the context feature $f_{c,l}^s$ are used as input for the GRU layers (see Fig. 2 bottom-left). Furthermore, the context features are fed to a decoder, $f_{\Lambda\alpha}$, to estimate the albedo and ambient components for gated slice reconstruction.

Stereo-Mono Fusion. Monocular gated depth estimates suffer from depth quantization due to the depth binning of gated slices, failure in the presence of strong ambient illumination, and illuminator occlusion. Stereo methods, in isolation, suffer from inherent ambiguity in partially occluded regions and can fail when one of the views is completely obstructed, e.g., by lens occlusions and bright illumination. Previous work [13] proposed distilling the monocular network with the stereo output, and distilling the stereo network with fused pseudo-labels. Departing from that approach, we use a light-weight 4-layer ResUNet [69] network, $f_z^r : (z^m, z^s, I) \to z^f$, that takes in monocular and stereo depth with the corresponding active and passive slices as input and produce a single fused depth map as output. The active and passive slices provide additional cues for the fusion network.

With the proposed depth estimation network in hand, we propose a set of stereo and monocular semi-supervised training signals for actively illuminated gated stereo pairs along with high dynamic passive captures.

4.2. Depth and Photometric Consistency

We rely on self-supervised consistency losses and sparse supervised losses as following.

Left-Right Reprojection Consistency. This loss enforces the photometric consistency between the left and right gated images given the per-pixel disparity,

$$\mathcal{L}_{reproj} = \mathcal{L}_p(\mathcal{M}_{l|r}^o \odot I_l, \mathcal{M}_{l|r}^o \odot I_{l|r}), \quad (4)$$

with $I_{l|r}$ the left image warped into the right view using the predicted disparity d_l. Here, $\mathcal{L}_p$ [19] is a similarity loss based on the structural similarity (SSIM) metric [61] and the L_1 norm, $\mathcal{L}_p(a, b) = 0.85\frac{1-SSIM(a,b)}{2} + 0.15\|a - b\|_1$. The occlusion mask $\mathcal{M}_{l|r}^o$ indicates pixels in the left image that are occluded in the right image and is defined as a soft mask for better gradient flow, $\mathcal{M}_{l|r}^o = 1 - \exp\left(-\eta\,|d_l + d_{l|r}|\right)$, where d_l is the left disparity and $d_{l|r}$ is the disparity of the right image projected to the left view.

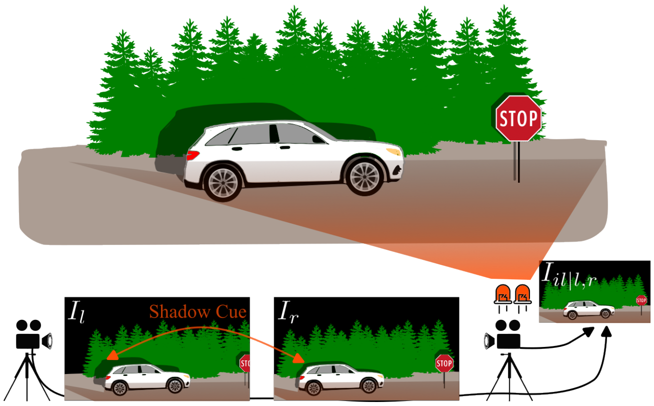


Figure 3. Scene regions occluded in the illuminator view will be in shadow in the two views (left, middle), and shadowless after projecting to the illuminator viewpoint (right).

Stereo-Mono Fusion Loss. The mono-stereo fusion loss $\mathcal{L}_{ms}$ guides the fusion network at depth discontinuities with the occlusion mask to obtain a fused depth map, $\tilde{z}_f = \mathcal{M}^o_{l|r} z^m + (1 - \mathcal{M}^o_{l|r}) z^s$, using the following loss,

$$\mathcal{L}_{ms} = \|z_f - \tilde{z}_f\|_1. \tag{5}$$

Ambient Image Consistency. The ambient luminance in a scene can vary by 14 orders of magnitude, inside a dark tunnel with bright sun at a tunnel exit, all in the same scene [46]. To tackle this extreme dynamic range, we reconstruct the ambient Λ^{k_0} in the scene from the short exposure slice μ_k, and sample Λ^{HDR} from the HDR passive captures I^4, I^5. Then, novel scene images $\hat{I}^k_v$ can be expressed as,

$$\Lambda^{HDR}_v = \mu_s (I^4_v + I^5_v - D^4_v - D^5_v)/(\mu_4 + \mu_5), \tag{6}$$

$$\Lambda^{k_0}_v = \mu_k (I^4_v + I^5_v - D^4_v - D^5_v)/(\mu_4 + \mu_5), \tag{7}$$

$$\hat{I}^k_v = \text{clip}\left(I^k_v - \Lambda^{k_0}_v + \Lambda^{HDR}_v, 0, 2^{10}\right), \tag{8}$$

with μ_s uniformly sampled in the interval from $[0.5\mu_k, 1.5\mu_k]$. We supervise the network by enforcing the depth to be consistent across different ambient illumination levels.

Gated Reconstruction Loss. We adopt the cyclic gated reconstruction loss from [59], which uses measured range intensity profiles $C_k(z)$ to reconstruct the input gated images from the predicted depth z, the albedo $\tilde{\alpha}$ and the ambient $\tilde{\Lambda}$. We estimate the $\tilde{\alpha}$ and $\tilde{\Lambda}$ from the context encoder through an additional U-Net like decoder, see Figure 2 and Supplemental Material. Specifically, the consistency loss models a gated slice as,

$$\tilde{I}^k(z) \;=\; \tilde{\alpha}\, C_k(z) + \tilde{\Lambda}. \tag{9}$$

The loss term is based on the per-pixel difference and structural similarity as follows,

$$\mathcal{L}_{recon} = \mathcal{L}_p(M_g \odot \tilde{I}^k(z), M_g \odot I^k) + \mathcal{L}_p(\tilde{\Lambda}, \Lambda^{k_0}). \tag{10}$$

Similar to [59] we utilize per-pixel SNR to obtain the gated consistency mask M_g. See the Supplemental Material for a detailed derivation. This loss enforces that the predicted depth is consistent with the simulated gated measurements.

Illuminator View Consistency. In the proposed gated stereo setup, we can enforce an additional depth consistency from the illuminator field of view. In this virtual camera view no shadows are visible as illustrated in Figure 3. This effectively makes the regions that are visible to the two cameras and the illuminator consistent. We use the gated consistency mask M_g to supervise only regions that are illuminated by the laser and project the gated views $I_{l,r}$ into the laser field of view $I_{il|r,l}$, resulting in the loss,

$$\mathcal{L}_{illum} = \mathcal{L}_p(M_g \odot I_{il|l}, M_g \odot I_{il|r}). \tag{11}$$

Image Guided Depth Regularization. Following binocular and multi-view stereo methods [20, 70], we add an edge-aware smoothness loss $\mathcal{L}_{smooth}$ as regularization to the mean normalized inverse depth estimates d,

$$\mathcal{L}_{smooth} = |\nabla_x d| e^{-|\nabla_x I|} + |\nabla_y d| e^{-|\nabla_y I|}. \tag{12}$$

Sparse LiDAR Supervision. The proposed gated stereo system has a higher update rate (24 Hz) than typical scanning LiDAR (10 Hz). Therefore, sparse LiDAR supervision can only be applied to samples fully in sync while all the previously presented self-supervised losses are applied to all samples. The LiDAR returns are first compensated for ego-motion, and then projected onto the image space. The supervision loss $\mathcal{L}_{sup}$ for view v is,

$$\mathcal{L}_{sup} = \mathcal{M}_{v|s} \odot \|z_v - z^*_{v|s}\|_1, \tag{13}$$

where $\mathcal{M}_{v|s}$ is a binary mask indicating the projection of LiDAR points on the image, and $z^*_{v|s}$ is the ground-truth depth from a single LiDAR scan projected into the image v.

Overall Training Loss. Combining all self-supervised and supervised loss components from above, we arrive at the following loss terms,

$$\mathcal{L}_{mono} = c_1 \mathcal{L}_{recon} + c_2 \mathcal{L}_{sup} + c_3 \mathcal{L}_{smooth}, \tag{14}$$

$$\mathcal{L}_{stereo} = c_4 \mathcal{L}_{reproj} + c_5 \mathcal{L}_{recon} + c_6 \mathcal{L}_{illum} + c_7 \mathcal{L}_{sup} + c_8 \mathcal{L}_{smooth}, \tag{15}$$

$$\mathcal{L}_{fusion} = c_9 \mathcal{L}_{ms} + c_{10} \mathcal{L}_{sup} + c_{11} \mathcal{L}_{smooth}, \tag{16}$$

which we combine with scalar weights $c_{1,...,11}$ provided in the Supplemental Material.

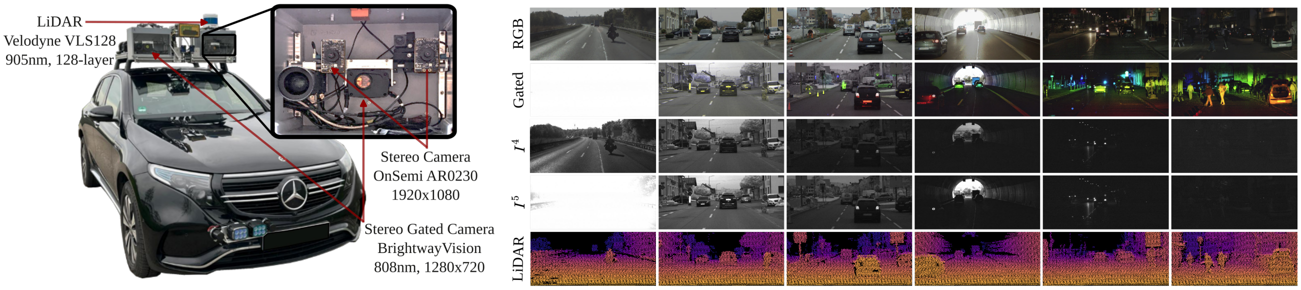


Figure 4. Illustration of the used sensor setup (left) and example captures from the wide-base gated stereo dataset (right). From top to bottom: RGB, Gated with red for slice 1, green for slice 2 and blue for slice 3, Gated Passive with low exposure time I^4, Gated Passive with high exposure time I^5, LiDAR. Note, the availability of a large number of frames with $\alpha C_k < I^k$.

4.3. Implementation Details

We first independently optimize the monocular and stereo networks using the losses presented in Sec. 4.2. Both the stereo and monocular networks are trained using the same protocol using ADAMW [42] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate of 10^{-4} and of weight decay 10^{-2}. Finally, the fusion network is trained for 5 epochs using ADAMW and the losses described in Eq. 16 with a learning rate of $3 \cdot 10^{-4}$. We used $\eta = 0.05$ for generating occlusion masks referred in Equation 4. For gated consistency masks, we set $\gamma = 0.98$, $\theta = 0.04$. All models are trained with input/output resolution of 1024×512.

5. Dataset

In this section, we describe the long-range depth dataset that we captured for training and testing. The dataset was acquired during a data collection campaign covering more than one thousand kilometers of driving in Southern Germany. We have equipped a testing vehicle with a long-range LiDAR system (Velodyne VLS128) with a range of up to 200 m, an automotive RGB stereo camera (OnSemi AR0230 sensor) and a NIR gated stereo camera setup (BrightWayVision) with synchronization. The sensor setup is shown in Figure 4 with all sensors mounted in a portable sensor cube, except for the LiDAR sensor. The RGB stereo camera has a resolution of 1920x1080 pixels and runs at 30 Hz capturing 12bit HDR images. The gated camera provides 10 bit images with a resolution of 1280x720 at a framerate of 120 Hz, which we split up into three slices plus two HDR-like additional ambient captures without active illumination. We use two vertical-cavity surface-emitting laser (VCSEL) modules as active illumination mounted on the front tow hitch. The lasers flood illuminate the scene at a peak power of 500 W each, a wavelength of 808 nm and laser pulse durations of 240-370 ns. The maximum peak power is thereby limited due to eye-safety regulations. The mounted reference LiDAR system is running with 10 Hz and yields 128 lines. All sensors are calibrated and time-synchronized and Fig. 4 provides visual examples. The dataset contains 107348 samples in day, nighttime, and varying weather conditions. After sub-selection for scenario diversity, we split the dataset into 54320 samples for training, 728 samples for validation and, 2463 samples for testing, see Supplemental Material for details.

	METHOD	Modality	Train	RMSE	ARD	MAE	δ_1	δ_2	δ_3
				$[m]$		$[m]$	[%]	[%]	[%]
	Test Data – Night (Evaluated on LiDAR Ground-Truth Points)								
COMPARISON TO STATE-OF-THE-ART	GATED2DEPTH [23]	Mono-Gated	D	16.15	0.17	8.07	75.70	92.74	96.47
	GATED2GATED [59]	Mono-Gated	MG	14.08	0.19	7.95	79.84	92.95	96.59
	SPARSE2DENSE [44]	Mono-Sparse	D	9.97	0.11	5.22	87.06	95.77	98.20
	KBNET [62]	Mono-Sparse	D	13.52	0.16	8.56	81.41	**99.33**	**99.66**
	NLSPN [47]	Mono-Sparse	D	12.19	0.09	5.42	89.63	96.84	99.03
	PENET [28]	Mono-Sparse	D	7.81	0.09	3.59	93.68	97.90	99.16
	GUIDENET [54]	Mono-Sparse	D	7.50	0.09	3.63	92.70	98.16	99.35
	PACKNET [25]	Mono-RGB	M	17.82	0.20	10.21	66.35	87.85	95.61
	MONODEPTH2 [21]	Mono-RGB	M	18.44	0.18	9.47	75.70	90.46	95.68
	SIMIPU [36]	Mono-RGB	D	15.78	0.18	8.71	76.25	90.84	96.44
	ADABINS [4]	Mono-RGB	D	14.45	0.15	7.58	81.47	93.75	97.39
	DPT [48]	Mono-RGB	D	12.15	0.12	6.31	85.38	95.94	98.42
	DEPTHFORMER [37]	Mono-RGB	D	12.15	0.11	6.20	85.18	95.76	98.47
	PSMNET [12]	Stereo-RGB	D	27.98	0.27	16.02	50.77	74.77	85.93
	STTR [40]	Stereo-RGB	D	20.99	0.19	11.14	70.84	87.70	93.46
	HSMNET [65]	Stereo-RGB	D	12.42	0.09	5.87	88.41	96.08	98.50
	ACVNET [64]	Stereo-RGB	D	11.70	0.08	5.25	89.91	96.33	98.47
	RAFT-STEREO [41]	Stereo-RGB	D	10.89	0.09	5.10	90.47	96.71	98.64
	GATED STEREO	Stereo-Gated	DGS	**6.39**	**0.05**	**2.25**	**96.40**	98.44	99.24
	Test Data – Day (Evaluated on LiDAR Ground-Truth Points)								
COMPARISON TO STATE-OF-THE-ART	GATED2DEPTH [23]	Mono-Gated	D	28.68	0.22	14.76	66.68	82.76	87.96
	GATED2GATED [59]	Mono-Gated	MG	16.87	0.21	9.51	73.93	92.15	96.10
	SPARSE2DENSE [44]	Mono-Sparse	D	10.05	0.11	4.77	88.06	96.57	98.63
	KBNET [62]	Mono-Sparse	D	15.27	0.17	9.54	78.54	**99.31**	**99.63**
	NLSPN [47]	Mono-Sparse	D	11.78	0.08	4.99	91.41	97.70	99.24
	PENET [28]	Mono-Sparse	D	8.54	0.09	3.82	93.78	97.69	98.94
	GUIDENET [54]	Mono-Sparse	D	8.03	0.09	3.70	93.23	98.12	99.21
	PACKNET [25]	Mono-RGB	M	17.69	0.21	9.77	72.12	90.65	96.51
	MONODEPTH2 [21]	Mono-RGB	M	20.78	0.22	10.06	79.05	90.66	94.69
	SIMIPU [36]	Mono-RGB	D	14.33	0.14	7.50	81.77	94.01	97.92
	ADABINS [4]	Mono-RGB	D	12.76	0.12	6.53	86.15	95.77	98.41
	DPT [48]	Mono-RGB	D	11.29	0.09	5.52	89.56	96.83	98.79
	DEPTHFORMER [37]	Mono-RGB	D	10.59	0.09	5.06	90.65	97.46	99.02
	PSMNET [12]	Stereo-RGB	D	32.13	0.28	18.09	53.82	74.91	84.96
	STTR [40]	Stereo-RGB	D	16.77	0.16	8.99	78.44	93.53	98.01
	HSMNET [65]	Stereo-RGB	D	10.36	0.08	4.69	92.47	97.93	99.11
	ACVNET [64]	Stereo-RGB	D	9.40	0.07	4.08	94.61	98.36	99.12
	RAFT-STEREO [41]	Stereo-RGB	D	9.40	0.07	4.07	93.76	98.15	99.09
	GATED STEREO	Stereo-Gated	DGS	**7.11**	**0.05**	**2.25**	**96.87**	98.46	99.11

Table 1. Comparison of our proposed framework 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. * marked method are scaled with LiDAR ground-truth. Best results in each category are in **bold** and second best are underlined.

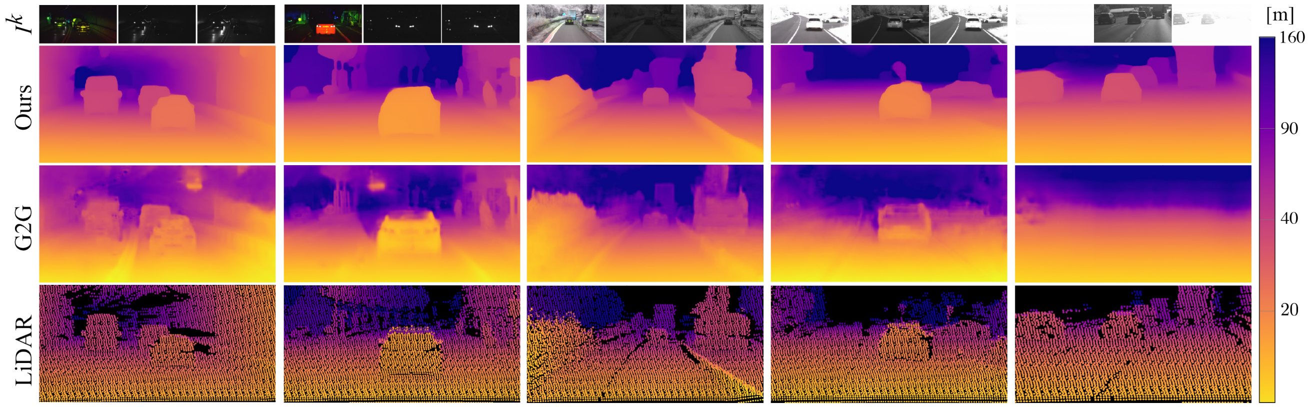


Figure 5. The top row for each example shows the concatenated gated image $I^{1,2,3}$ and the corresponding passive images I^4 and I^5. The second row shows the depth map of our proposed method, the third row illustrates the results of Gated2Gated (G2G) [59], and the bottom row depicts the projected LiDAR point cloud into the gated view. Our method handles shadow areas and high-reflectivity targets much better than G2G. Furthermore, the HDR input allows to predict accurate depth even in bright conditions.

	Modality	HDR	Ambient Con.	Cycle Con.	Warp Con.	RMSE $[m]$	MAE $[m]$	δ_1 [%]	δ_2 [%]	δ_3 [%]
	Test Data – Night (Evaluated on LiDAR Ground-Truth Points)									
ABLATION	Mono-Gated	✗	✗	✗	✗	8.03	3.36	93.45	97.56	98.91
	Mono-Gated	✓	✓	✓	✗	7.07	2.60	95.91	98.14	99.09
	Stereo-Gated [41]	✗	✗	✗	✗	7.92	3.06	95.23	97.98	99.00
	Stereo-Gated	✓	✓	✗	✗	7.38	2.41	95.63	98.01	99.02
	Stereo-Gated	✓	✓	✓	✗	7.72	2.55	95.31	97.86	98.89
	Stereo-Gated	✓	✓	✓	✓	7.33	2.39	95.84	98.09	99.02
	Mono+Stereo-Gated	✓	✓	✓	✓	**6.39**	**2.25**	**96.40**	**98.44**	**99.24**
	Test Data – Day (Evaluated on LiDAR Ground-Truth Points)									
ABLATION	Mono-Gated	✗	✗	✗	✗	11.93	5.31	90.15	95.62	97.73
	Mono-Gated	✓	✓	✓	✗	9.26	3.66	94.69	97.84	98.88
	Stereo-Gated [41]	✗	✗	✗	✗	9.77	4.03	92.15	96.69	98.28
	Stereo-Gated	✓	✓	✗	✗	7.63	2.31	96.42	98.18	98.98
	Stereo-Gated	✓	✓	✓	✗	7.87	2.27	96.46	98.13	98.92
	Stereo-Gated	✓	✓	✓	✓	7.47	**2.15**	96.72	98.29	99.00
	Mono+Stereo-Gated	✓	✓	✓	✓	**7.11**	2.25	**96.87**	**98.46**	**99.11**

Table 2. Ablation studies evaluated on the proposed **Gated Stereo** test dataset. We investigate different input modalities, feature encoders, and loss combinations for the monocular and stereo network. Our final fusion model outperforms all other methods by a significant margin.

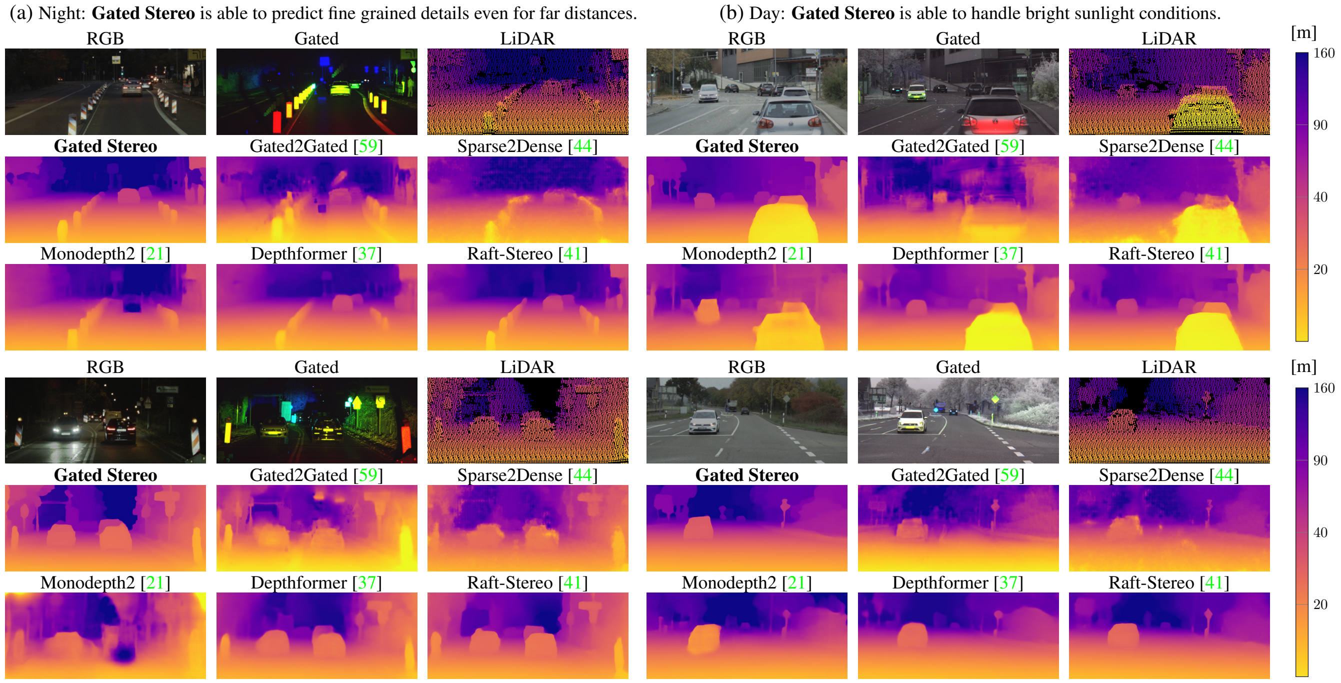
6. Assessment

In this section, we validate the proposed method experimentally. We investigate depth estimation at night, day and compared to existing depth estimation methods. Moreover, we validate design choices with ablation experiments.

Experimental Setup. We evaluate on the proposed test set consisting of 2463 (1269 day/1194 night) frames with high-resolution 128-layer LiDAR ground-truth measurements up to 200 m. Unlike existing work [23, 55, 59] which was limited to 80 m, we are therefore able to report results up to a distance of 160 m to asses long-range depth prediction. Following [16], we evaluate depth using the metrics RMSE, MAE, ARD, and $\delta_i < 1.25i$ for $i \in 1, 2, 3$ and split results for day and night. For fair comparison, *all methods we compare to have been fine-tuned* on our dataset. Details on the fine-tuning of reference methods are given in Section 4.1. of the Supplemental Material.

Depth Reconstruction. Qualitative results are presented in Figure 6 and quantitative results in Table 1. Here, we compare against two recent gated [23, 59], six monocular RGB [4, 21, 25, 36, 37, 48], five stereo RGB [12, 40, 41, 64, 65] and five monocular+LiDAR [28, 44, 47, 54, 62] methods. Comparing Gated Stereo to the next best stereo method RAFT-Stereo [41], our method reduces error by 45 % and 1.8 m in MAE in day conditions. In night conditions, the error is reduced by 56 % and 2.9 m MAE. Qualitatively this improvement is visible in sharper edges and less washed-out depth estimates. Fine details, including thin poles, are better visible due to the structure-aware refinement achieved through the monocular depth outputs. The next best gated method, Gated2Gated [59] achieves a 9.51 m MAE in day conditions and 7.95 m MAE in night conditions. Here, the performance drops significantly in day conditions due to strong ambient illumination, while Gated Stereo is capable of making use of the passive captures. This is also visible in the shown qualitative Figure 5, where Gated Stereo maintains high-quality depth outputs, while Gated2Gated fails. Overall, we report a reduction of 74 % in MAE error compared to existing gated methods. Comparing to the best monocular RGB method, Depthformer [37], textures are often wrongly interpreted as rough surfaces missing smoothness. Lastly, we compare to monocular + LiDAR methods. Note, that the methods are fed with ground-truth points and therefore achieve competitive quantitative results on par with the best stereo methods. Qualitatively, the methods are not capable of interpolating plausible depth maps, which are instead washed out, and we find that problematic texture interpretation is carried over from monocular depth estimation methods.

Ablation Experiments. To validate the contributions of each component of the proposed method, we report ablation experiments in Table 2, see Supplemental Material for qualitative results. In the following, we compare the MAE of the different models averaged over day and night. The starting point for our analysis is the monocular gated esti-


(a) Night: **Gated Stereo** is able to predict fine grained details even for far distances.

(b) Day: **Gated Stereo** is able to handle bright sunlight conditions.

Figure 6. **Qualitative comparison of Gated Stereo and existing methods**. For (a) night and (b) day conditions, Gated Stereo predicts sharper depth maps than existing methods. (In the gated image red refers to I^1, green to I^2, and blue to I^3).

mation using the proposed monocular branch with LiDAR supervision only. This method outperforms the best monocular RGB approach [37] by 23 % lower MAE error. Next, the concatenated passive images and the active slices result in an added reduction of 28 % MAE error. We analyze RAFT-Stereo with stereo gated images and HDR-like passive frames as input. With additional Ambient Aware Consistency and the proposed backbone, we reduce the MAE error by 25 % compared to the next monocular gated approach and by 36 % to a native RAFT Stereo network with gated input. The HR-Former backbone alone contributes about 10 % of the 33 % reduction in MAE. By adding the Gated Consistency loss and the warping losses for left-right consistency across views and illuminator the error further decreased by 4 %. Finally, the fusion stage combining the monocular and stereo outputs preserves the fine structures from the monocular model and the long-range accuracy of the stereo model, results in an reduction of 48 % in MAE error when compared to monocular gating.

7. Conclusion

We present Gated Stereo, a long-range active multi-view depth estimation method. The proposed method predicts dense depth from synchronized gated stereo pairs acquired in a wide-baseline setup. The architecture comprises a stereo network and per-view monocular and stereo-mono fusion networks. All of these sub-networks utilize both active and passive images to extract depth cues. Stereo cues can be ambiguous, e.g., due to occlusion and repeated structure. Similarly, monocular gated cues can be insufficient in bright ambient illumination and at long range. To this end, our proposed approach predicts stereo *and* per-camera monocular depth and finally fuses the two to obtain a single high-quality depth map. The different parts of the network are semi-supervised with sparse LiDAR supervision and a set of self-supervised losses that ensures consistency between different predicted outputs. We train and validate the proposed method on a new long-range automotive dataset with a maximum depth range twice as long as prior work. The proposed method achieves 50 % better mean absolute depth error than the next best method on stereo RGB images and 74 % better than the next best existing gated method. In the future, we hope that the proposed method may allow us to solve novel 3D vision tasks that today's LiDAR systems cannot solve due to their angular resolution, such as detecting unseen small objects as lost debris at long distances and high-quality road edge and lane detection.

Acknowledgments This work was supported in the AI-SEE project with funding from the FFG, BMBF, and NRC-IRA. We thank the Federal Ministry for Economic Affairs and Energy for support via the PEGASUS-family project "VVM-Verification and Validation Methods for Automated Vehicles Level 4 and 5". Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award, and an Amazon Science Research Award.

References

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

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

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

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

[5] 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

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

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

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

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

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

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

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

[13] Zhi Chen, Xiaoqing Ye, Wei Yang, Zhenbo Xu, Xiao Tan, Zhikang Zou, Errui Ding, Xinming Zhang, and Liusheng Huang. Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15529–15538, 2021. 4

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

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

[16] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374, 2014. 1, 2, 7

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

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 2

[19] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 2, 4

[20] 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*, 2017. 5

[21] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 2, 6, 7, 8

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

[23] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide. Gated2depth: Real-time dense lidar from gated images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 4, 6, 7

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

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

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

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

[28] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Towards precise and efficient image guided depth completion. 2021. 2, 6, 7

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

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

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

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

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

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

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

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

[37] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022. 2, 6, 7, 8

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

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

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

[41] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 2, 4, 6, 7, 8

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2018. 6

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

[44] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *IEEE International Conference on Robotics and Automation*, pages 1–8, 2018. 2, 6, 7, 8

[45] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[46] Karol Myszkowski, Rafal Mantiuk, and Grzegorz Krawczyk. *High Dynamic Range Video*. Association for Computing Machinery, 2016. 5

[47] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proc. of European Conference on Computer Vision (ECCV)*, 2020. 2, 6, 7

[48] 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, 6, 7

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

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

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

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

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

[54] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30:1116–1129, 2020. 2, 6, 7

[55] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 2, 7

[56] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. DeMoN: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

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

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

[59] Amanpreet Walia, Stefanie Walz, Mario Bijelic, Fahim Mannan, Frank Julca-Aguilar, Michael Langer, Werner Ritter, and Felix Heide. Gated2gated: Self-supervised depth estimation from gated images, 2022. 1, 2, 3, 4, 5, 6, 7, 8

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

[61] 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. 4

[62] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12747–12756, 2021. 2, 6, 7

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

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

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

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

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

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

[69] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 4

[70] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 5