

Gated2Depth: Real-Time Dense Lidar From Gated Images

Tobias Gruber^{1,3} Frank Julca-Aguilar² Mario Bijelic^{1,3} Felix Heide^{2,4}

¹Daimler AG ²Algolux ³Ulm University ⁴Princeton University

Abstract

We present an imaging framework which converts three images from a gated camera into high-resolution depth maps with depth accuracy comparable to pulsed lidar measurements. Existing scanning lidar systems achieve low spatial resolution at large ranges due to mechanically-limited angular sampling rates, restricting scene understanding tasks to close-range clusters with dense sampling. Moreover, today's pulsed lidar scanners suffer from high cost, power consumption, large form-factors, and they fail in the presence of strong backscatter. We depart from point scanning and demonstrate that it is possible to turn a low-cost CMOS gated imager into a dense depth camera with at least 80m range – by learning depth from three gated images. The proposed architecture exploits semantic context across gated slices, and is trained on a synthetic discriminator loss without the need of dense depth labels. The proposed replacement for scanning lidar systems is real-time, handles back-scatter and provides dense depth at long ranges. We validate our approach in simulation and on real-world data acquired over 4,000 km driving in northern Europe. Data and code are available at <https://github.com/gruberto/Gated2Depth>.

1. Introduction

Active depth cameras, such as scanning lidar systems, have not only become a cornerstone imaging modality for autonomous driving and robotics, but are emerging in applications across disciplines, including autonomous drones, remote sensing, human-computer interaction, and augmented or virtual reality. Depth cameras that provide dense range allow for dense scene reconstructions [26] when combined with color cameras, including correlation time-of-flight cameras (C-ToF) [19, 30, 33] such as Microsoft's Kinect One, or structured light cameras [1, 42, 43, 49]. These acquisition systems facilitate the collection of large-scale RGB-D data sets that fuel research on core computer vision problems, including scene understanding [23, 53] and action recognition [40]. However, while existing depth cameras provide high-fidelity depth for close ranges indoors [26, 39], dense depth imaging at long ranges and in dynamic outdoor scenes is an open challenge.

Active imaging at long ranges is challenging because diffuse scene points only return a small fraction of the emitted photons back to the sensor. For perfect Lambertian surfaces, this fraction decreases quadratically with distance, posing a fundamental limitation as illumination power can only be increased up to the critical eye-safety level [51, 54, 60]. To tackle this constraint, existing pulsed lidar systems employ sensitive silicon avalanche photo-diodes (APDs) with high photon detection efficiency in the NIR band [60]. The custom semiconductor process for these sensitive detectors restricts current lidar systems to a single (or few) APDs instead of monolithic sensor arrays, which requires point-by-point scanning. Although scanning lidar approaches facilitate depth imaging at large ranges, scanning reduces their spatial resolution quadratically with distance, prohibiting semantic tasks for far objects, as shown in Figure 1. Recently, single-photon avalanche diodes (SPADs) [4, 5, 41, 46] are emerging as a promising technology that may enable sensor arrays in the CMOS process [59] in the future. Although SPADs are sensitive to individual photons, existing designs are highly photon-inefficient due to very low fill factors around 1% [58] and pile-up distortions at higher pulse powers [12]. Moreover, passive depth estimation techniques do not offer a solution, including stereo cameras [20, 49] and depth from monocular imagery [13, 16, 48]. These approaches perform poorly at large ranges for small disparities, and they fail in critical outdoor scenarios, when ambient light is not sufficient, e.g. at night, and in the presence of strong back-scatter, e.g. in fog or snow, see Figure 2.

Gated imaging is an emerging sensing technology that tackles these challenges by sending out pulsed illumination and integrating a scene response between temporal gates. Coarse temporal slicing allows for the removal of back-scatter due to fog, rain and snow, and can be realized in readily available CMOS technology. In contrast to pulsed lidar, gated imaging offers high signal measurements at long distances by integrating incoming photons over a large temporal slice, instead of time-tagging the first returns of individual pulses. However, although gated cameras offer an elegant, low-cost solution to outdoor imaging challenges, the sequential acquisition of the individual slices prohibits their use as depth cameras today, restricting depth information to

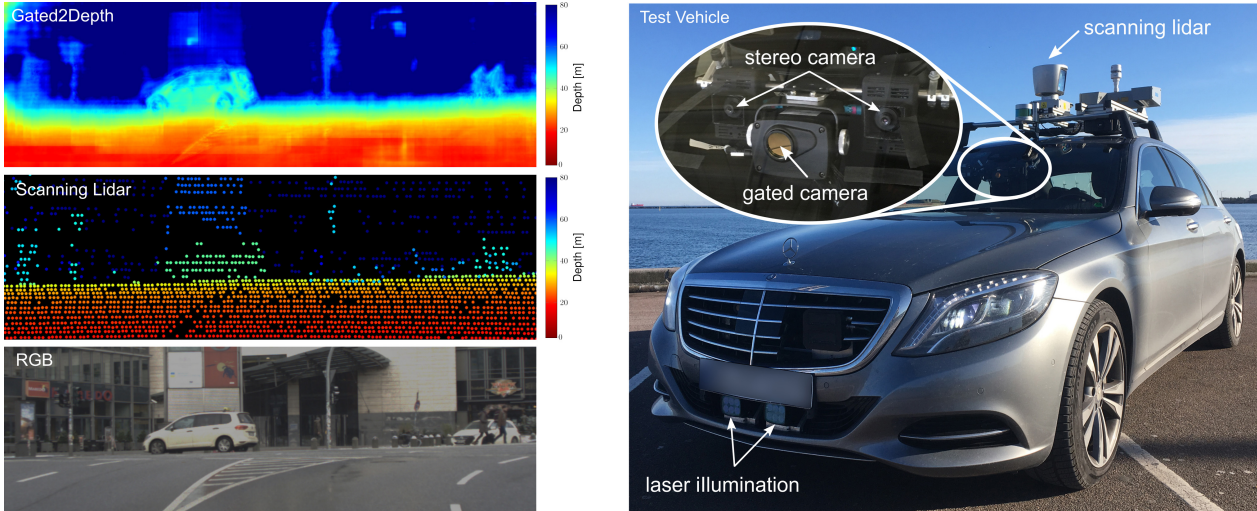


Figure 1: We propose a novel real-time framework for dense depth estimation (top-left) without scanning mechanisms. Our method maps measurements from a flood-illuminated gated camera behind the wind-shield (inset right), captured in real-time, to dense depth maps with depth accuracy comparable to lidar measurements (center-left). In contrast to the sparse lidar measurements, these depth maps are high-resolution enabling semantic understanding at long ranges. We evaluate our method on synthetic and real data, collected with a testing and a scanning lidar Velodyne HDL64-S3D as reference (right).

a sparse set of wide temporal bins spanning more than 50 m in depth. Note that using narrow slices does not offer a solution, because the slice width is inversely proportional to the number of captures, and thus frame-rate and narrow slices also means integrating less photons. With maximum frame rates of 120 Hz to 240 Hz, existing systems [18] are limited to a range of 4 to 7 slices for dynamic scenes.

In this work, we present a method that recovers high-fidelity dense depth from sparse gated images. By learning to exploit semantic context across gated slices, the proposed architecture achieves depth accuracy comparable to scanning based lidar in large-range outdoor scenarios, essentially turning a gated camera into a low-cost dense flash lidar that captures dense depth at long distances and also sees through fog, snow and rain. The method jointly solves depth estimation, denoising, inpainting of missing or unreliable measurements, shadow and multi-path removal, while being highly efficient with real-time frame rates on consumer GPUs.

Specifically, we make the following contributions:

- We introduce an image formation model and analytic depth estimation method using less than a handful of gated images.
- We propose a learning-based approach for estimating dense depth from gated images, without the need for dense depth labels for training.
- We validate the proposed method in simulation and on real-world measurements acquired with a prototype system in challenging automotive scenarios. We show

that the method recovers dense depth up to 80 m with depth accuracy comparable to scanning lidar.

- We provide the first long-range gated data set, covering over 4,000 km driving throughout northern Europe. The data set includes driving scenes in snow, rain, urban driving and sub-urban driving.

2. Related Work

Depth Estimation from Intensity Images. A large body of work explores methods for extracting depth from conventional color image sensors. A first line of research on structure from motion methods sequentially captures a stack of monocular images and extracts geometry by exploiting temporal correlation in the stack [29, 56, 57, 63]. In contrast, multi-view depth estimation methods [20] do not rely on sequential acquisition but exploit the disparity in simultaneously acquired image pairs [52]. Recent approaches to estimating stereo correspondences allow for interactive frame-rates [8, 28, 44]. Over the last years, a promising direction of research aims at estimating depth from a single monocular image [9, 13, 16, 32, 48], no longer requiring multi-view or sequential captures. Saxena et al. [48] introduce a Markov Random Field that incorporates multiscale image features for depth estimation. Eigen et al. [13] demonstrate that CNNs are well-suited for monocular depth estimation by learning priors on semantic-dependent depth [10, 16, 32]. While consumer time-of-flight cameras facilitate the acquisition of large datasets for small indoor scenes [23, 53], supervised training in large

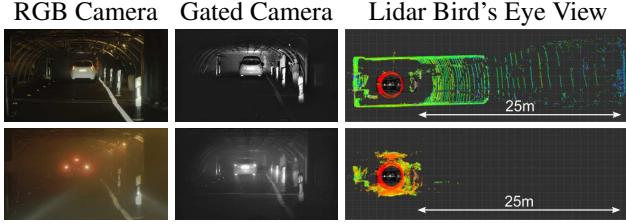


Figure 2: Sensor performance in a fog chamber with very dense fog. The first row shows recordings without fog while the second row shows the same scene in dense fog.

outdoor environments is an open challenge. Recent approaches tackle the lack of dense training data by proposing semi-supervised methods relying on relative depth [10], stereo images [15, 16, 31], sparse lidar points [31] or semantic labels [62]. Passive methods have in common that their precision is more than an order of magnitude below that of scanning lidar systems which makes them no valid alternative to ubiquitous lidar ranging in autonomous vehicles [51]. In this work, we propose a method that allows to close this precision gap using low-cost gated imagers.

Sparse Depth Completion. As an alternative approach to recover accurate dense depth, a recent work proposes depth completion from sparse lidar measurements. Similar to monocular depth estimation, learned encoder-decoder architectures have been proposed for this task [11, 27, 37]. Jaritz et al. [27] propose to incorporate color RGB data for upsampling sparse depth samples but also require sparse depth samples in down-stream scene understanding tasks. To allow for an independent design of depth estimation and scene analysis algorithms, the completion architecture has to be trained with varying sparsity patterns [27, 37] or additional validity maps [11]. While these depth completion methods offer improved depth estimates, they suffer from the same limitation as scanned lidar: low spatial resolution at long ranges due to limited angular sampling, low-resolution detectors, and costly mechanical scanning.

Time-of-Flight Depth Cameras. Amplitude-modulated C-ToF cameras [19, 30, 33], such as Microsoft’s Kinect One, have become broadly adopted for indoor sensing [23, 53]. These cameras measure depth by recording the phase shift of periodically-modulated flood light illumination, which allows to extract the time-of-flight for the reflected flood light from the source to scene and back to the camera. However, in addition to the modulated light, this sensing approach also records all ambient background light. While per-pixel lock-in amplification removes background components efficiently in indoor scenarios [33], and learned architectures can alleviate multi-path distortions [55], existing C-ToF cameras are limited to ranges of a few meters in outdoor scenarios [22] in strong sunlight.

Gated cameras send out pulses of flood-light and only

record photons from a certain distance by opening and closing the camera after a given delay. Gated imaging has first been proposed by Heckman et al. [21]. This acquisition mode allows to gate out backscatter from fog, rain, and snow [18]. Busck et al. [3, 6, 7] use gated imaging for high-resolution depth sensing by capturing large sequences of narrow gated slices. However, as the depth accuracy is inversely related to the gate width, and hence the number of required captures, sequentially capturing high-resolution gated depth is infeasible at real-time frame-rates. Recently, a line of research proposes analytic reconstruction models for known pulse and integration shapes [34, 35, 61]. These approaches require perfect knowledge of the integration and pulse profiles, which is impractical due to drift, and they provide low precision for broad gating windows in real-time capture settings. Adam et al. [2], and Schober et al. [50], present Bayesian methods for pulsed time-of-flight imaging of room-sized scenes. These methods solve probabilistic per-pixel estimation problems using priors on depth, reflectivity and ambient light, which is possible when using nanosecond exposure profiles [2, 50] for room-sized scenes. In this work, we demonstrate that exploiting spatio-temporal scene semantics allows to recover dense and lidar-accurate depth from only three slices, with exposures two orders of magnitude longer (> 100 ns), acquired in real-time. Using such wide exposure gates allows us to rely on low-cost gated CMOS imagers instead of detectors with high temporal resolution, such as SPADs.

3. Gated Imaging

In this section, we review gated imaging and propose an analytic per-pixel depth estimation method.

Gated Imaging Consider the setup shown in Figure 3, where an amplitude-modulated source flood-illuminates the scene with broad rect-shaped “pulses” of light. The synchronized camera opens after a delay ξ to receive only photons with round-trip path-length longer than $\xi \cdot c$, where c is the speed of light. Assuming a dominating lambertian reflector at distance r , the detector gain is temporally modulated with the gating function g resulting in the exposure measurement

$$I(r) = \alpha C(r) = \int_{-\infty}^{\infty} g(t - \xi) \kappa(t, r) dt, \quad (1)$$

where κ is the temporal scene response, α the albedo of the reflector, and $C(r)$ the range-intensity profile. With the reflector at distance r , the temporal scene response can be described as

$$\kappa(t, r) = \alpha p\left(t - \frac{2r}{c}\right) \beta(r). \quad (2)$$

where p is here the laser pulse profile and atmospheric effects, e.g. in a scattering medium, are modeled by the

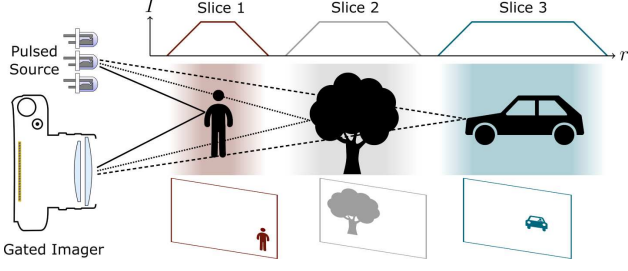


Figure 3: A gated system consists of a pulsed laser source and a gated imager that are time synchronized. By setting the delay between illumination and image acquisition, the environment can be sliced into single images that contain only a certain distance range.

distance-dependent function β . Note that we ignore ambient light in Eq. (2) which is minimized by a notch-filter in our setup and eliminated by subtraction with a separate capture without active illumination. In order to prevent the laser from overheating, the number of laser pulses in a certain time is limited and therefore a passive image can be obtained at no cost during laser recovery. The exposure profiles are designed to have the same passive component. The range-intensity profile $C(r)$ can be calibrated with measurements on targets with fixed albedo. We extract depth from three captures with different delays $\xi_i, i \in \{1, 2, 3\}$, resulting in a set of profiles $C_i(r)$ and measurements $I_i(r)$. We approximate the profiles with Chebychev polynomials of degree 6 as $\tilde{C}(r)$. Figure 4 shows the range-intensity profiles used in this work and their approximations, see supplemental material for details on the exposure profile design. The final measurement, after read-out, is affected by photon shot noise and read-out noise as

$$z = I(r) + \eta_p(I(r)) + \eta_g, \quad (3)$$

for a given pixel location, with η_p being a Poissonian signal-dependent noise component and η_g a Gaussian signal-independent component, which we adopt from [14].

Measurement Distortions A number of systematic and random measurement distortions make depth estimation from gated images challenging. Scene objects with low reflectance only return few signal photons, prohibiting an unambiguous mapping from intensities to depth and albedo in the presence of the Poissonian-Gaussian measurement fluctuations from Eq. (3). Systematic distortions include multi-path bounces of the flash illumination, see also [55]. In typical driving scenarios, severe multi-path reflection can occur due to wet roads acting as mirroring surfaces in the scene. Note that these are almost negligible in line or point-based scanning-lidar systems [1]. Automotive applications require large laser sources that cannot be placed next to the camera, inevitably resulting in shadow regions without measurements available. Severe ambient sunlight, present

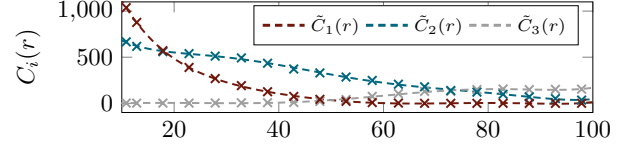


Figure 4: Discrete measurements (marked with crosses) of the three range-intensity profiles $C_i(r), i \in \{1, 2, 3\}$ used in this work, and their continuous Chebychev approximations $\tilde{C}_i(r)$ plotted with distance r [m].

as an offset in all slices, reduces the dynamic range of the gated measurements. In this work, we demonstrate a reconstruction architecture which addresses all of these issues in a data-driven approach, relying on readily available sparse lidar depth as training labels. Before describing the proposed approach, we introduce a per-pixel baseline estimation method.

Per-Pixel Least-Squares Estimate. Ignoring all of the above measurement distortions, assuming no drift in the pulse and exposure profiles and Gaussian noise only in Eq. (3), an immediate baseline approach is the following per-pixel least-squares estimation. Specifically, for a single pixel, we stack the measurements $z_{\{1,2,3\}}$ for a sequence of delays $\xi_{\{1,2,3\}}$ in a single vector $\mathbf{z} = [z_1, \dots, z_3]$. We can estimate the depth and albedo jointly as

$$\hat{r}_{LS} = \underset{r, \alpha}{\operatorname{argmin}} \left\| \mathbf{z} - \alpha \tilde{\mathbf{C}}(r) \right\|_2^2, \quad (4)$$

where $\tilde{\mathbf{C}}(r) = [\tilde{C}_1(r), \dots, \tilde{C}_3(r)]$ is a Chebychev intensity profile vector. Since the range-intensity profiles are non-linear, we solve this nonlinear least-squares estimation using the Levenberg-Marquardt optimization method, see details in the supplemental document.

4. Learning Depth from Gated Images

In this section, we introduce the *Gated2DepthNet* network. The proposed model is the result of a systematic evaluation of different input configurations, network architectures, and training schemes. We refer the readers to the supplemental document for a comprehensive study on all evaluated models.

The proposed network architecture is illustrated in Figure 5. The input to our network are three gated slices, allowing it to exploit the corresponding semantics across the slices to estimate accurate pixel-wise depth. An immediately apparent issue for this architecture is that dense ground truth depth for large-scale scenes is not available. This issue becomes crucial when designing deep models that require large training datasets to avoid overfitting. We address this problem with a training strategy that transfers dense depth semantics learned on synthetic data to a network trained on sparse lidar data.

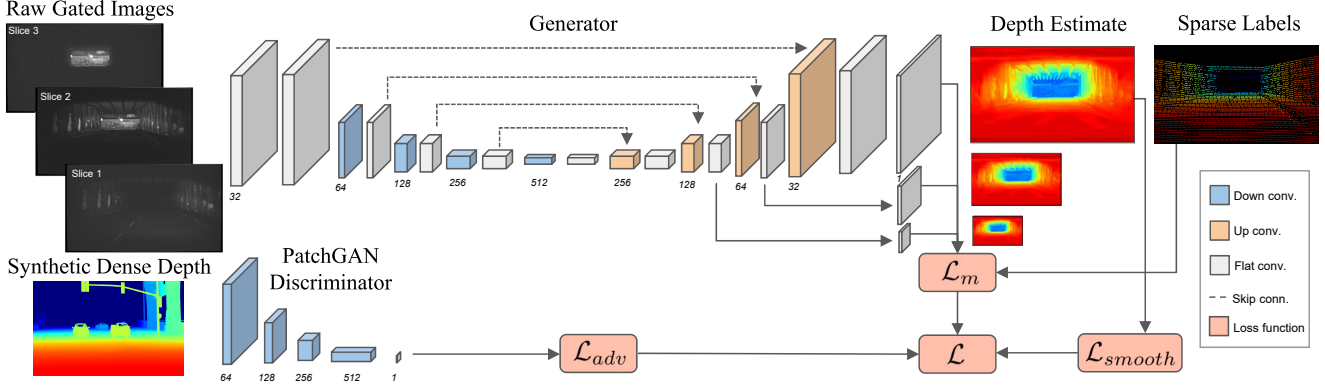


Figure 5: The proposed GATED2DEPTH architecture estimates dense depth from a set of three gated images (actual reconstruction and real captures shown). To train the proposed generator network G using sparse depth from lidar point samples, we rely on three loss-function components: a sparse multi-scale loss $\mathcal{L}_{\text{mult}}$ which penalizes sparse depth differences on three different binned scales, a smoothness loss $\mathcal{L}_{\text{smooth}}$, and an adversarial loss \mathcal{L}_{adv} . The adversarial loss incorporates a discriminator network which was trained on synthetic data, using a separate throw-away generator, and allows to transfer dense depth details from synthetic data without domain adaptation.

The proposed *Gated2DepthNet* is composed of a generator G , which we train for our dense depth estimation task. G is a multi-scale variant of the popular U-net [47] architecture. To transfer dense depth from synthetically generated depth maps to sensor data, we introduce a discriminator D , a variant of PatchGAN [25], and train the network in a two-stage process. In the first stage, we train a network (G and D) on synthetic data as generative adversarial network [17]. The generator and discriminator are trained in alternating fashion in a least-square GAN [38] approach: G is trained to generate accurate dense depth estimations, using *synthetic ground truth*, and to convince D that the estimations correspond to a real depth maps; D is trained to detect whether a dense depth map comes from G or is a real one. In the second stage, we train the network on real gated images that follow the target domain distribution. We now use sparse lidar measurements as groundtruth and keep the discriminator fixed. To use sparse lidar measurements in the final training stage, we introduce a multi-scale loss (see Section 4.1) that penalizes differences to sparse lidar points by binning these to depth maps at multiple scales.

Our generator consists of 4 pairs of convolutions with a max pooling operation after each pair. The encoder portion produces internal maps $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{16}$ of the original input size. The decoder consists of four additional convolutions, and transposed convolutions after each pair. As the depth estimate shares semantics with the input, we use symmetric skip connections, see Figure 5.

In the discriminator, we use a PatchGAN variant to best represent high-frequency image content. To this end, we define a fully convolutional network with five layers, each layer consisting of 4×4 kernels with stride 2, and leaky ReLUs with slope 0.2. The network classifies overlapping patches of a dense depth map instead of the whole map.

4.1. Loss Function

We train our proposed network to minimize a three-component loss, \mathcal{L} , with each component modeling different statistics of the target depth

$$\mathcal{L} = \mathcal{L}_{\text{mult}} + \lambda_s \mathcal{L}_{\text{smooth}} + \lambda_a \mathcal{L}_{\text{adv}} \quad (5)$$

Multi-scale loss ($\mathcal{L}_{\text{mult}}$) This loss component penalizes differences between the ground truth labels and the depth estimates. We define $\mathcal{L}_{\text{mult}}$ as a multi-scale loss over the generator’s output d and its corresponding target \tilde{d}

$$\mathcal{L}_{\text{mult}}(d, \tilde{d}) = \sum_{i=1}^M \lambda_{m_i} \mathcal{L}_{\text{L1}}(d^{(i)}, \tilde{d}^{(i)}), \quad (6)$$

where $d^{(i)}$ and $\tilde{d}^{(i)}$ are the generator’s output and target at a scale (i) , $\mathcal{L}_{\text{L1}}(d^{(i)}, \tilde{d}^{(i)})$ is the loss at scale (i) , and λ_{m_i} is the weight of the loss at the same scale. We define three scales $1/2^i$ with $i \in \{0, 1, 2\}$, binning as illustrated in Figure 5. For a scale (i) , we define $\mathcal{L}_{\text{L1}}(d^{(i)}, \tilde{d}^{(i)})$ as the mean absolute error

$$\mathcal{L}_{\text{L1}}(d^{(i)}, \tilde{d}^{(i)}) = \frac{1}{N} \sum_{j,k} |d_{jk}^{(i)} - \tilde{d}_{jk}^{(i)}|, \quad (7)$$

with the subscript jk indicates here a discretized bin corresponding to pixel position (j, k) . When training with synthetic data, we compute \mathcal{L}_{L1} over all pixels. For training with real data, we only compute this loss at bins that include at least one lidar sample point. \mathcal{L}_{L1} is formally defined as

$$\mathcal{L}_{\text{L1}}(d^{(i)}, \tilde{d}^{(i)}) = \frac{1}{N} \sum_{j,k} |d_{jk}^{(i)} - \tilde{d}_{jk}^{(i)}| m_{jk}^{(i)} \quad (8)$$

where $m_{jk} = 1$ when the bin (j, k) contains at least one lidar sample, and $m_{jk} = 0$ otherwise. For smaller scales, we average all samples per bin.

Weighted Smoothness Loss ($\mathcal{L}_{\text{smooth}}$) We rely on an additional smoothness loss $\mathcal{L}_{\text{smooth}}$ to regularize the depth estimates. Specifically we use a total variation loss weighted by the input image gradients [62], that is

$$\mathcal{L}_{\text{smooth}} = \frac{1}{N} \sum_{i,j} |\partial_x d_{i,j}| \epsilon^{-|\partial_x z_{i,j}|} + |\partial_y d_{i,j}| \epsilon^{-|\partial_y z_{i,j}|}, \quad (9)$$

where z is here the input image. As sparse lidar data is sampled on horizontal lines due to the rotating scanning setup, a generator trained on this data is biased to outputs with similar horizontal patterns. We found that increasing the weight of the vertical gradient relative to the horizontal one helps to mitigate this problem.

Adversarial loss (\mathcal{L}_{adv}) We define the adversarial loss following [38] with the PatchGAN [25] discriminator:

$$\mathcal{L}_{\text{adv}} = \frac{1}{2} \mathbb{E}_{y \sim p_{\text{depth}}(y)} [(D(y) - 1)^2] + \frac{1}{2} \mathbb{E}_{x \sim p_{\text{gated}}(x)} [(D(G(x)))^2] \quad (10)$$

Note the discriminator is fixed in the second training stage.

4.2. Training and Implementation Details

We use ADAM optimizer with the learning rate set to 0.0001. For the global loss function, we experimentally determined $\lambda_s = 0.0001$ and $\lambda_a = 0.001$. For the multi-scale loss, we define $\lambda_{m_0} = 1$, $\lambda_{m_1} = 0.8$, and $\lambda_{m_2} = 0.6$. The full system runs at real-time rates of 25 Hz, including all captures and inference (on a single TitanV).

5. Datasets

In this section, we describe the real and synthetic data sets used to train and evaluate the proposed method.

Real Dataset To the best of our knowledge, we provide the *first long-range gated dataset*, covering snow, rain, urban and sub-urban driving during 4,000 km in-the-wild acquisition. To this end, we have equipped a testing vehicle with a standard RGB stereo camera (Aptina AR0230), lidar system (Velodyne HDL64-S3) and a gated camera (BrightwayVision BrightEye) with flood-light source integrated into the front bumper, shown in Figure 1. Both cameras are mounted behind the windshield, while the lidar is mounted on the roof. The stereo camera runs at 30 Hz with a resolution of 1920x1080 pixels. The gated camera provides 10bit images with a resolution of 1280x720 at a framerate of 120 Hz, which we split up in three slices plus an additional ambient capture without active illumination. The car is equipped with two vertical-cavity surface-emitting laser (VCSEL) modules, which are diffused, with a wavelength of 808 nm and a pulsed optical output peak power of 500 W each. The peak power is limited due to

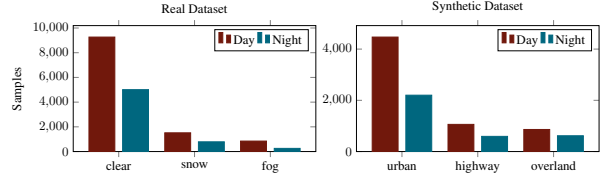


Figure 6: Dataset distribution.

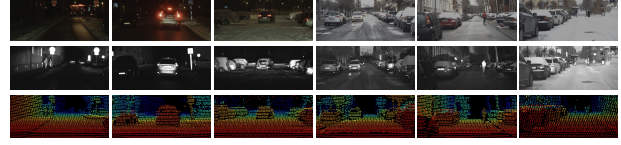


Figure 7: Examples of real dataset (rgb/gated/lidar).

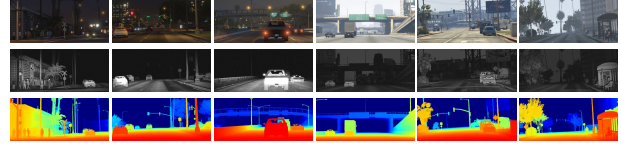


Figure 8: Examples of synthetic dataset (rgb/gated/depth).

eye-safety regulations. Our reference lidar systems is running with 10 Hz and yields 64 lines. All sensors are calibrated and time-synchronized. During a four-week acquisition time in Germany, Denmark and Sweden, we recorded 17,686 frames in different cities (Hamburg, Copenhagen, Gothenburg, Vårgårda, Karlstad, Örebro, Västerås, Stockholm, Uppsala, Gävle, Sundsvall, Kiel). Figure 6 visualizes the distribution of the full dataset, and Figure 7 shows qualitative example measurements. We captured images during night and day and in various weather conditions (clear, snow, fog). The samples in clear weather conditions (14,277) are split into a training (7,478 day/4,460 night) and test set (1,789 day/550 night). Since snow and fog disturbs the lidar data, we do not use snowy nor foggy data for training.

Synthetic Dataset While existing simulated datasets contain RGB and depth data, they do not provide enough information to synthesize realistic gated measurements that require NIR modeling and sunlight-illumination. We modify the GTA5-based simulator from [45] to address this issue. Please see the supplemental document for detailed description. We simulate 9,804 samples, and use 8,157 (5,279 day/2,878 night) for training and 1,647 (1,114 day/533 night) for testing. See Figure 6 and Figure 8 for visualizations.

6. Assessment

Evaluation Setting We compare the proposed method against state-of-the-art depth estimation methods. As per-pixel baseline methods, we compare to the least-squares baseline from Eq. (4) and against the Bayesian estimate from Adam et al. [2]. We compare against recent methods using monocular RGB images [16], stereo images [8], and

RGB images in combination with sparse lidar points [37]. For completeness, we also evaluate monocular depth estimation [16] applied on the integral of the gated slices, i.e. an actively illumination scene image without gating, which we dub full gated image. Moreover, we also demonstrate Gated2Depth trained on full gated images only, validating the benefit of the coarse gating itself. For the method of Godard et al. [16], we resized our images to the native size the model was trained on, as we noticed a substantial drop in performance when changing resolution at test time. For all other algorithms, we did not observe this behavior and we used the full resolution images. For a fair comparison, we finetuned [16] on RGB stereo pairs taken from the training set of our real dataset starting from the best available model. For the comparisons in simulation, we calibrated the sampling pattern of the experimental lidar system and use this pattern for the Sparse-to-dense [37] method. For [24] we only had a hardware implementation available running in our test vehicle which does not allow synthetic evaluations.

We evaluate the methods with the metrics from [13], namely RMSE, MAE, ARD and $\delta_i < 1.25^i$ for $i \in \{1, 2, 3\}$. On the synthetic dataset, we compute the metrics over the whole depth maps. On the real dataset, we compute the metrics only at the predicted pixels that correspond to measured sparse lidar points. We observed that our lidar reference system degrades at distances larger than 80 m and therefore we limit our evaluation to 80 m. For a fair comparison to methods that rely on laser illumination, we do not evaluate on non-illuminated pixels and introduce at the same time a completeness metric that describes on how many ground truth pixels is evaluated. Being $[z_1, z_2, z_3]$ a set of input gated slices, we define non-illuminated pixels as the ones that satisfy $\max([z_1, z_2, z_3]) - \min([z_1, z_2, z_3]) < 55$. This definition allows us to avoid evaluation over outliers at extreme distances and with very low SNR.

6.1. Results on Synthetic Dataset

Table 1 (top) shows that the proposed method outperforms all other reference methods by a large margin. The second-best method without gated images is the depth completion based on lidar and RGB [36], which yields better results than monocular or stereo methods because it uses sparse lidar ground truth samples as input. While monocular approaches struggle to recover absolute scale, stereo methods achieve low accuracy over the large distance range due to the limited baseline.

Figure 9c shows an output example of our method and compares it with others. Our method captures better fine-grained details of a scene at both close and far distances.

6.2. Results on Real Dataset

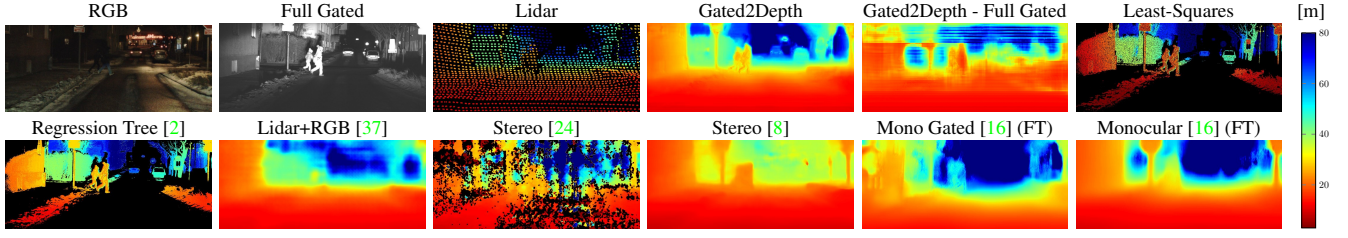
Table 1 (bottom) shows that the proposed method outperforms all compared methods, including the one that uses

METHOD	RMSE [m]	ARD	MAE [m]	δ_1 [%]	δ_2 [%]	δ_3 [%]	Compl. [%]
Simulated Data – Night (Evaluated on Dense Ground Truth Depth)							
DEPTH FROM MONO ON RGB [16]	74.40	0.62	58.47	7.76	13.67	29.17	100
DEPTH FROM MONO ON FULL GATED [16]	84.48	0.69	68.74	2.53	7.03	20.33	100
DEPTH FROM STEREO [8]	72.67	0.67	59.94	4.73	10.88	19.05	100
SPARSE-TO-DENSE ON LIDAR (GT INPUT) [37]	64.08	0.33	42.33	56.74	63.19	67.87	100
DEPTH FROM ToF, REGRESSION TREE [2]	40.33	0.45	26.03	37.33	55.96	68.47	45
LEAST SQUARES	30.45	0.29	18.66	60.82	77.41	83.61	34
GATED2DEPTH	12.99	0.07	3.96	94.24	97.28	98.34	100
Simulated Data – Day (Evaluated on Dense Ground Truth Depth)							
DEPTH FROM MONO ON RGB [16]	75.68	0.63	59.95	6.27	14.14	28.28	100
DEPTH FROM MONO ON FULL GATED [16]	81.67	0.69	66.44	2.71	8.43	20.04	100
DEPTH FROM STEREO [8]	75.04	0.70	62.06	3.76	8.86	14.97	100
SPARSE-TO-DENSE ON LIDAR (GT INPUT) [37]	60.97	0.31	39.63	58.84	65.30	69.77	100
DEPTH FROM ToF, REGRESSION TREE [2]	27.17	0.52	20.05	25.53	47.77	66.30	23
LEAST SQUARES	15.52	0.36	10.32	55.44	73.29	82.35	16
GATED2DEPTH	9.10	0.05	2.66	96.41	98.47	99.16	100
Real Data – Night (Evaluated on Lidar Ground Truth Points)							
DEPTH FROM MONO ON RGB [16]	16.87	0.38	11.64	21.74	63.15	80.96	100
DEPTH FROM MONO ON RGB [16] (FT)	11.41	0.23	6.18	76.64	89.53	94.19	100
DEPTH FROM MONO ON FULL GATED [16]	16.26	0.36	10.19	54.03	74.44	85.00	100
DEPTH FROM MONO ON FULL GATED [16] (FT)	15.41	0.52	11.33	31.72	71.23	88.74	100
DEPTH FROM STEREO [8]	14.58	0.21	8.34	68.75	82.63	89.36	100
DEPTH FROM STEREO [24]	15.51	0.36	8.75	63.94	76.19	82.31	63
SPARSE-TO-DENSE ON LIDAR (GT INPUT) [37]	8.79	0.21	4.38	87.64	93.74	95.88	100
DEPTH FROM ToF, REGRESSION TREE [2]	10.54	0.24	6.01	76.73	89.74	93.45	40
LEAST SQUARES	13.13	0.42	8.88	43.60	55.80	63.54	31
GATED2DEPTH - FULL GATED	14.86	0.29	8.84	58.79	58.79	79.84	100
GATED2DEPTH	8.39	0.15	3.79	87.52	93.00	95.21	100
Real Data – Day (Evaluated on Lidar Ground Truth Points)							
DEPTH FROM MONO ON RGB [16]	17.67	0.37	12.28	13.87	60.93	79.17	100
DEPTH FROM MONO ON RGB [16] (FT)	10.24	0.18	5.47	80.49	91.78	95.61	100
DEPTH FROM MONO ON FULL GATED [16]	13.89	0.24	8.50	60.05	79.62	89.92	100
DEPTH FROM MONO ON FULL GATED [16] (FT)	13.33	0.40	9.51	36.64	81.63	92.86	100
DEPTH FROM STEREO [8]	13.94	0.19	7.78	71.32	84.67	91.38	100
DEPTH FROM STEREO [24]	9.63	0.17	4.59	85.80	92.72	95.20	86
SPARSE-TO-DENSE ON LIDAR (GT INPUT) [37]	8.21	0.16	4.05	88.52	94.71	96.87	100
DEPTH FROM ToF, REGRESSION TREE [2]	15.83	0.49	11.40	56.30	75.54	82.45	23
LEAST SQUARES	19.52	0.75	14.05	43.42	54.63	63.76	16
GATED2DEPTH - FULL GATED	13.75	0.26	8.16	62.48	62.48	82.93	100
GATED2DEPTH	7.61	0.12	3.53	88.07	94.32	96.60	100

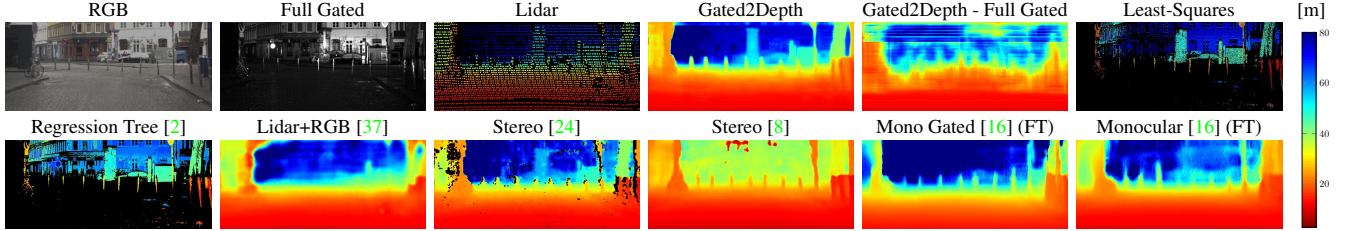
Table 1: Comparison of our proposed framework and state-of-the-art methods on unseen synthetic and real test data sets. GT INPUT: uses sparse ground truth as input. FT: model finetuned on our real data.

ground truth lidar points as input [37]. Hence, the method achieves high depth accuracy comparable to scanning lidar systems, while, in contrast, providing dense depth. Moreover, Table 1 validates the benefit of using multiple slices compared to a single continuously illuminated image.

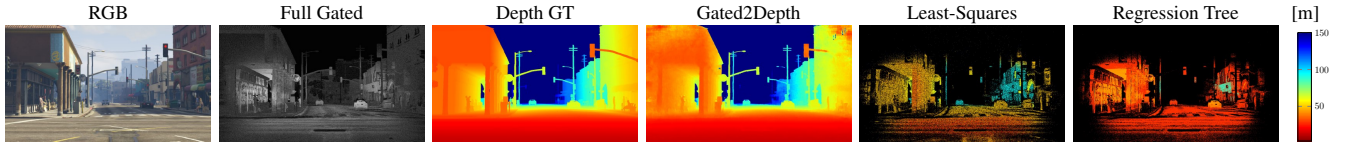
Figures 9a and 9b visualizes the dense depth estimation, and scene details captured by our method in comparison to state-of-the-art methods. Especially for fine details around pedestrians or small scene objects, the proposed method achieves higher resolution. In the example from Figure 9a our method shows all scene objects (two pedestrians, two cars), which are also recovered in both gated per-pixel estimation methods, but not at high density. While the sparse depth completion method misses major scene objects, our method preserves all of them. The same can be observed in the second example for the posts and the advertising column in Figure 9b. Figure 10 illustrates the robustness of our method in (unseen) snowing conditions. While the lidar shows strong clutter, our method provides a very clear depth



(a) Experimental night time results.



(b) Experimental day time results.



(c) Daytime simulation results.

Figure 9: Qualitative results for our method and reference methods over real and synthetic examples. For each example, we include the corresponding RGB and full gated image, along with the lidar measurements. Our method generates more accurate and detailed maps over different distance ranges of the scenes in comparison to the other methods. For the simulation results in (c) we only show models finetuned on simulated data.

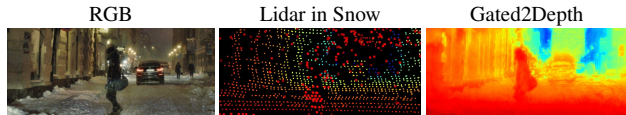


Figure 10: Results for strong backscatter in snow, with lidar clutter (larger points) around a pedestrian and in the sky.



Figure 11: Multipath Interference. In contrast to existing methods, such as the Least-Squares method, our method eliminates most multi-path interference (on the road here).

estimation, as a by-product of the gated imaging acquisition itself. Figure 11 compares per-pixel estimation with the proposed approach. The proposed method is able to fill in shadows and surfaces with low reflectance. Multi-path interference is suppressed by using the contextual information present in the whole image.

7. Conclusions and Future Work

In this work, we turn a CMOS gated camera into a cost-sensitive high-resolution dense flash lidar. We propose a

novel way of transfer learning that allows us to leverage datasets with sparse depth labels for dense depth estimation. The proposed method outperforms state-of-the-art methods, which we validate in simulation and experimentally on outdoor captures with large depth range of up to 80 m (limited by the range of the scanned reference lidar system).

An interesting direction for future research is the inclusion of RGB data, which could provide additional depth clues in areas with little variational information in the gated images. However, fusing RGB images naively as an additional input channel to the proposed architecture would lead to severe bias for distortions due to backscatter, see Figure 2, which is properly handled by the proposed system. Exciting future applications of the proposed method include large-scale semantic scene understanding and action recognition using the proposed architecture either for dataset generation or in an end-to-end-fashion.

This work has received funding from the European Union under the H2020 ECSEL Programme as part of the DENSE project, contract number 692449. Werner Ritter supervised this project at Daimler AG, and Klaus Dietmayer supervised the project portion at Ulm University. We thank Robert Böhler, Stefanie Walz and Yao Wang for help processing the large dataset. We thank Fahim Manan for fruitful discussions and comments on the manuscript.

References

- [1] Supreeth Achar, Joseph R. Bartels, William L. Whittaker, Kiriakos N. Kutulakos, and Srinivasa G. Narasimhan. Epipolar time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 36(4):37, 2017. 1, 4
- [2] 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. 3, 6, 7, 8
- [3] Pierre Andersson. Long-range three-dimensional imaging using range-gated laser radar images. *Optical Engineering*, 45(3):034301, 2006. 3
- [4] Brian F. Aull, Andrew H. Loomis, Douglas J. Young, Richard M. Heinrichs, Bradley J. Felton, Peter J. Daniels, and Deborah J. Landers. Geiger-mode avalanche photodiodes for three-dimensional imaging. *Lincoln Laboratory Journal*, 13(2):335–349, 2002. 1
- [5] Danilo Bronzi, Yu Zou, Federica Villa, Simone Tisa, Alberto Tosi, and Franco Zappa. Automotive three-dimensional vision through a single-photon counting SPAD camera. *IEEE Transactions on Intelligent Transportation Systems*, 17(3):782–795, 2016. 1
- [6] Jens Busck. Underwater 3-D optical imaging with a gated viewing laser radar. *Optical Engineering*, 2005. 3
- [7] Jens Busck and Henning Heiselberg. Gated viewing and high-accuracy three-dimensional laser radar. *Applied Optics*, 43(24):4705–10, 2004. 3
- [8] 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, 6, 7, 8
- [9] Richard Chen, Faisal Mahmood, Alan Yuille, and Nicholas J Durr. Rethinking monocular depth estimation with adversarial training. *arXiv preprint arXiv:1808.07528*, 2018. 2
- [10] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*, pages 730–738, 2016. 2, 3
- [11] Zhao Chen, Vijay Badrinarayanan, Gilad Drozdov, and Andrew Rabinovich. Estimating depth from RGB and sparse sensing. *Proceedings of the IEEE European Conf. on Computer Vision*, Sep 2018. 3
- [12] Patricia Coates. The correction for photon ‘pile-up’ in the measurement of radiative lifetimes. *Journal of Physics E: Scientific Instruments*, 1(8):878, 1968. 1
- [13] 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
- [14] Alessandro Foi, Mejdí Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. 4
- [15] 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. 3
- [16] Clément Godard, Oisín 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. 1, 2, 3, 6, 7, 8
- [17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. www.deeplearningbook.org. 5
- [18] Yoav Grauer. Active gated imaging in driver assistance system. *Advanced Optical Technologies*, 3(2):151–160, 2014. 2, 3
- [19] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012. 1, 3
- [20] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 2
- [21] Paul Heckman and Robert T. Hodgson. Underwater optical range gating. *IEEE Journal of Quantum Electronics*, 3(11):445–448, 1967. 3
- [22] Felix Heide, Wolfgang Heidrich, Matthias Hullin, and Gordon Wetzstein. Doppler time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 34(4):36, 2015. 3
- [23] Steven Hickson, Stan Birchfield, Irfan Essa, and Henrik Christensen. Efficient hierarchical graph-based segmentation of RGBD videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 344–351, 2014. 1, 2, 3
- [24] Heiko Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, Feb 2008. 7, 8
- [25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5, 6
- [26] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 1
- [27] 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. 3
- [28] 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
- [29] Jan J. Koenderink and Andrea J. Van Doorn. Affine structure from motion. *Journal of the Optical Society of America A*, 8(2):377–385, Feb 1991. 2

- [30] 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. 1, 3
- [31] Yevhen Kuznetsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2215–2223, 2017. 3
- [32] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *International Conference on 3D Vision (3DV)*, pages 239–248, 2016. 2
- [33] Robert Lange. 3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. 2000. 1, 3
- [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. 3
- [35] Martin Laurenzis, Frank Christnacher, and David Monnin. Long-range three-dimensional active imaging with super-resolution depth mapping. *Optics letters*, 32(21):3146–8, 2007. 3
- [36] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *IEEE International Conference on Robotics and Automation*, 2019. 7
- [37] 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, 7, 8
- [38] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 5, 6
- [39] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2015. 1
- [40] Bingbing Ni, Gang Wang, and Pierre Moulin. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013. 1
- [41] Cristiano Niclass, Alexis Rochas, P-A Besse, and Edoardo Charbon. Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes. *IEEE Journal of Solid-State Circuits*, 40(9):1847–1854, 2005. 1
- [42] Matthew O’Toole, Felix Heide, Lei Xiao, Matthias B Hullin, Wolfgang Heidrich, and Kiriakos N Kutulakos. Temporal frequency probing for 5D transient analysis of global light transport. *ACM Transactions on Graphics (ToG)*, 33(4):87, 2014. 1
- [43] Matthew O’Toole, Ramesh Raskar, and Kiriakos N Kutulakos. Primal-dual coding to probe light transport. *ACM Transactions on Graphics (ToG)*, 31(4):39–1, 2012. 1
- [44] Andrea Pilzer, Dan Xu, Mihai Puscas, Elisa Ricci, and Nicu Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *International Conference on 3D Vision (3DV)*, pages 587–595, 2018. 2
- [45] Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2232–2241, 2017. 6
- [46] Alexis Rochas, Michael Gösch, Alexandre Serov, Pierre-André Besse, Rade S. Popovic, Theo Lasser, and Rudolf Rigler. First fully integrated 2-D array of single-photon detectors in standard CMOS technology. *IEEE Photonics Technology Letters*, 15(7):963–965, 2003. 1
- [47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5
- [48] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems*, pages 1161–1168, 2006. 1, 2
- [49] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 2003. 1
- [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. 3
- [51] Brent Schwarz. Lidar: Mapping the world in 3D. *Nature Photonics*, 4(7):429, 2010. 1, 3
- [52] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 519–528, 2006. 2
- [53] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015. 1, 2, 3
- [54] James D. Spinhirne, Jonathan A.R. Rall, and V. Stanley Scott. Compact eye safe lidar systems. *The Review of Laser Engineering*, 23(2):112–118, 1995. 1
- [55] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. Deep end-to-end time-of-flight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6383–6392, 2018. 3, 4
- [56] Philip H.S. Torr and Andrew Zisserman. Feature based methods for structure and motion estimation. In *International workshop on vision algorithms*, pages 278–294. Springer, 1999. 2
- [57] 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

- [58] Chockalingam Veerappan, Justin Richardson, Richard Walker, Day-Uey Li, Matthew W Fishburn, Yuki Maruyama, David Stoppa, Fausto Borghetti, Marek Gersbach, Robert K Henderson, et al. A 160×128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter. In *IEEE International Solid-State Circuits Conference*, pages 312–314, 2011. [1](#)
- [59] Federica Villa, Rudi Lussana, Danilo Bronzi, Simone Tisa, Alberto Tosi, Franco Zappa, Alberto Dalla Mora, Davide Contini, Daniel Durini, Sasha Weyers, et al. CMOS imager with 1024 SPADs and TDCs for single-photon timing and 3-D time-of-flight. *IEEE journal of selected topics in quantum electronics*, 20(6):364–373, 2014. [1](#)
- [60] George M. Williams. Optimization of eyesafe avalanche photodiode lidar for automobile safety and autonomous navigation systems. *Optical Engineering*, 56(3):1 – 9 – 9, 2017. [1](#)
- [61] 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. [3](#)
- [62] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *Proceedings of the Asian Conference on Computer Vision*, 2018. [3](#), [6](#)
- [63] 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](#)