Gated2Gated: Self-Supervised Depth Estimation from Gated Images

Amanpreet Walia*[1,3] Stefanie Walz*[2] Mario Bijelic[4] Fahim Mannan[1]
Frank Julca-Aguilar[1] Michael Langer[3] Werner Ritter[2] Felix Heide[1,4]

[1]Algolux [2]Mercedes-Benz AG [3]McGill University [4]Princeton University

Abstract

Gated cameras hold promise as an alternative to scanning LiDAR sensors with high-resolution 3D depth that is robust to back-scatter in fog, snow, and rain. Instead of sequentially scanning a scene and directly recording depth via the photon time-of-flight, as in pulsed LiDAR sensors, gated imagers encode depth in the relative intensity of a handful of gated slices, captured at megapixel resolution. Although existing methods have shown that it is possible to decode high-resolution depth from such measurements, these methods require synchronized and calibrated LiDAR to supervise the gated depth decoder – prohibiting fast adoption across geographies, training on large unpaired datasets, and exploring alternative applications outside of automotive use cases. In this work, propose an entirely self-supervised depth estimation method that uses gated intensity profiles and temporal consistency as a training signal. The proposed model is trained end-to-end from gated video sequences, does not require LiDAR or RGB data, and learns to estimate absolute depth values. We take gated slices as input and disentangle the estimation of the scene albedo, depth, and ambient light, which are then used to learn to reconstruct the input slices through a cyclic loss. We rely on temporal consistency between a given frame and neighboring gated slices to estimate depth in regions with shadows and reflections. We experimentally validate that the proposed approach outperforms existing supervised and self-supervised depth estimation methods based on monocular RGB and stereo images, as well as supervised methods based on gated images. Code is available at https://github.com/princeton-computational-imaging/Gated2Gated.

1. Introduction

Depth sensing has become a cornerstone imaging modality for 3D scene understanding. Depth is used directly as input to a vision module, or indirectly in training datasets, to supervise models relying on other modalities across a wide range of applications such as perception and planning in autonomous driving, robotics, remote sensing, augmented and virtual reality [40, 56].

The most successful methods for depth estimation are either based on pulsed scanning LiDAR [43], or passive RGB sensors, i.e., monocular [14,17,48] and stereo RGB [26,49]. LiDAR depth sensors [51] measure the time-of-flight of pulses of light emitted into the scene and returned to the sensor along a coaxial path that is sequentially scanned across a scene. As a result, these sensors deliver precise depth with high spatial resolution at short distances. However, they are expensive, suffer from quadratic decreasing spatial resolution at longer distances, e.g., resulting in a few measurement points for pedestrians at 100 m distance [51], and they fail in the presence of strong back-scatter. Monocular and stereo RGB methods offer a substantially cheaper alternative, but they struggle to achieve depth precision comparable to time-of-flight imaging, struggle in low-light scenarios, and at long distances that map to small disparities.

Recently, gated imaging has been proposed as an alternative sensor modality for depth estimation and 3D detection [22, 32] which promises to overcome the spatial resolution limitation of scanning LiDAR while providing comparable depth precision. Gated cameras combine low-cost CMOS sensors with analog gated readout and active flash illumination, allowing to capture a sequence of gated image slices that each encode time-resolved illumination via their relative intensity and, as such, provide depth cues not present in RGB cameras. Thanks to this active gated flash acquisition mode, gated imaging methods have shown to be more robust in low-light scenarios and in the presence of strong back-scatter that can be suppressed during acquisition [22]. For a comprehensive review of Gated Sensors see Sec. 3. Gated images provide dense depth information at megapixel resolution of the gated camera, allowing for long-range perception where LiDAR-based methods fail [32]. However, all of these existing methods require calibrated and synchronized LiDAR data for training supervision. Although commercial gated cameras have become available [20], the requirement of such a multimodal capture system prohibits the rapid adoption of gated depth imaging not only in automotive applications but also in other robotic use cases. Moreover, large unpaired sequences of gated imagery cannot be exploited in training existing gated depth estimation methods. In this work, we propose the first self-supervised method for depth estimation from gated cameras. The proposed method takes gated

* These authors contributed equally to this work.

slices as input and predicts the scene albedo, depth, and ambient illumination. We reconstruct the input slices using calibrated gated profiles enforcing *measurement cycle consistency* and warping from the nearby slices in a temporal window enforcing *temporal consistency*. To this end, we introduce a differentiable gated image formation model that uses depth-dependent calibrated gating profiles for self-supervised measurement cycle loss. To learn in the absence of reliable gated measurements due to shadows, we utilize temporal depth consistency using differentiable structure-from-motion and the proposed image formation model.

Specifically, we make the following main contributions:

- We propose a novel self-supervised method that uses measurement cycle consistency and temporal consistency as training signals.
- The proposed model is trained end-to-end, and by exploiting calibrated gating profiles, the method is able to accurately estimate metric depth using the cycle consistency component.
- We validate that the proposed method outperforms self-supervised and supervised depth estimation using monocular and stereo RGB images and supervised gated depth estimation methods.

We have released models and code used to reproduce the results from this work.

2. Related Work

Depth from Time-of-Flight Time-of-Flight (ToF) cameras acquire depth by measuring the round-trip time of modulated flood-illuminated light returned from a scene. Existing methods can be classified into three categories: correlation [25,35,37], pulsed ToF [51] and gated imaging [20,27]. Correlation ToF cameras [25,35,37] estimate the depth from the phase difference of the sent and received laser pulse, which allows high spatial resolution, but is limited to short distances and indoor environments [28]. Pulsed ToF sensors [51] emit pulses of light and directly measure the round-trip time of the returned pulses, but require a scanning mechanism for large distances reducing the spatial resolution. Additionally, experiments have shown that they can be affected by adverse weather disturbances [4, 8, 31]. Gated imaging [5, 20, 27] records the returned light within a short integration time on the imaging sensor. This limits the capture to certain depth ranges and allows short-range back-scatter to be ignored. In [2, 6, 7] a sequence of three gated slices was used to reconstruct depth information. Further methods introduced analytical approximations [38, 39, 58] or learned the depth prediction through Bayesian methods [1, 50] and deep neural networks [23].

Supervised Depth Estimation Learning to predict depth from intensity images requires appropriate ground truth data. Methods either use the supervision from time-of-flight data [9, 14, 22, 30, 43] or ground truth depth from multi-view systems [33, 40, 44]. Previous imaging systems either process images from monocular cameras [12, 36, 40], can reason about multiple views [9, 60], or utilize the combination of monocular images with sparse LiDAR point-clouds [30, 43]. All of these existing methods can fail in low-light or low-contrast scenarios, for example, at night or in cluttered scenes, that active methods [22] tackle using illumination. Furthermore, RGB-only monocular depth estimation can only reconstruct up to an unknown scale factor.

Self-Supervised Depth Estimation Acquiring ground truth data for supervised depth estimation methods is challenging. An extensive process was applied in [16, 53] to overcome the limited range and spatial resolution by requiring a thorough LiDAR camera synchronization with the ego-motion correction and accumulating single point clouds into LiDAR maps as ground truth. Nonetheless, the application spectrum is limited, disallowing the use in areas such as scattering media, where LiDAR data is cluttered [3] or for vehicle fleet data without expensive ground truth sensors. To tackle this challenge, self-supervised training approaches exploit multiview geometry by aligning stereo image pairs [15, 18] or making use of image view synthesis between temporally consecutive frames [19, 24, 61]. Aligning stereo images pairs for depth prediction was initially proposed by Garg et al. [15]. Here, a neural network predicts the disparity from monocular camera images and supervises it by warping the stereo images. Image synthesis between temporally consecutive monocular image frames was introduced in [18, 54]. This approach utilizes two independent networks, one predicting the depth and the other estimating a rigid body transformation between two temporally adjacent frames. A reprojection error between two frames is then formulated to supervise the depth estimate. The following monocular methods investigate novel neural architectures [15, 19, 24] or extensions in the loss formulation [13, 18, 19, 24, 42, 46, 55, 59]. However, they have inherent scale ambiguity, which can be reduced by relying on vehicle velocity or LiDAR ground-truth depth measurements at test-time [24]. Departing from these approaches, the proposed depth estimation method relies on the calibrated gate profiles used in a measurement cycle consistency loss to enforce scale accuracy. Moreover, we extract further depth cues from the motion between intra-capture gated frames that are sequentially acquired for different gates.

3. Gated Imaging

Before presenting the proposed method, we briefly review the principles of gated imaging. Figure 1 illustrates a gated imaging system which consists of a synchronized camera and flash illumination source. In contrast to scanning LiDAR systems, the flash illumination source illuminates the scene through a laser pulse p, before capturing the

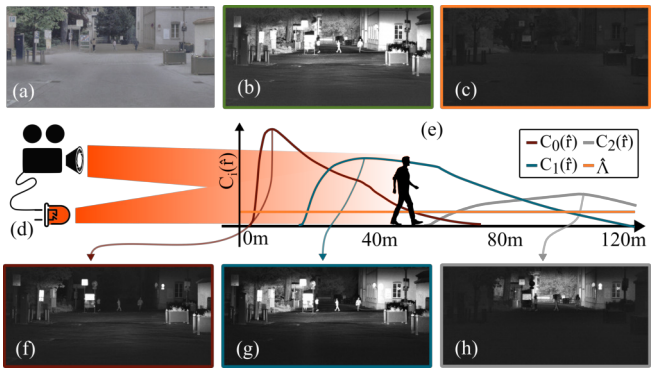

Figure 1. A gated camera consists of a synchronized gated camera and a flash pulsed illumination source (d). Using different exposure gates, the image formation can be described with three range-intensity profiles C_i, $i \in \{0, 1, 2\}$ plotted depending on distance $r[m]$, and an ambient, unmodulated $\hat{\mathbf{\Lambda}}$ scene contribution. An overlay of all exposures is shown in green (b). Individual range intensity profiles are shown for (f) short distance 3-72 m in deepred, (g) moderate distance 18-123 m in petrol, (h) far distance 57-176 m in gray and (c) ambient light ($\mathbf{Z}_t^p$) in orange. A corresponding RGB capture of the scene is shown in (a).

reflected light echo with a ξ delay. The reflected light is captured through a CMOS imaging sensor, which only captures photons arriving in a given temporal gate with profile g. Following Gruber *et al.* [23], we denote a single gated exposure as

$$\begin{aligned} Z_t^i(r) &= \alpha\, C_i(r) \\ &= \alpha \int_{-\infty}^{\infty} g_i(t-\xi) p_i\left(t - \frac{2r}{c}\right) \beta(r) dt, \end{aligned} \tag{1}$$

where $Z_t^i(r)$ is the gated exposure, indexed by i, at distance r and time t; $C_i(r)$ is the range intensity profile, i.e., the convolution of the gated slice and its corresponding pulse profile; α is the surface reflectance (albedo), and β the attenuation along a given path due to atmospheric interactions. The depth dependent path attenuation becomes unity in the absence of any participating media.

We note that during daytime, this model is incomplete due the high spectral solar power within the NIR band that leads to a significant number of unmodulated photons captured as an ambient light Λ component. We modifiy the model from [23] as

$$Z_t^i(r) = \alpha\, C_i(r) + \Lambda. \tag{2}$$

Similar to other CMOS-based sensing methods, gated imaging is also affected by noise, that can be modelled with a signal-dependent Poisson η_p and Gaussian η_g, resulting in

$$Z_t^i = \alpha\, C_i(r) + \Lambda + \eta_g + \eta_p. \tag{3}$$

In this work, at a given time t, we capture three sequential gated slices with spatial resolution $H \times W$ as $\mathbf{Z}_t^i \in \mathbb{R}_+^{H \times W}$ with delays $\xi_{\{0,1,2\}}$ and an unmodulated NIR passive image $\mathbf{Z}_t^p$ as illustrated in Figure 1. With a native frame rate of 120 Hz, the proposed gated camera provides a full set of observations at 30 Hz.

4. Self-Supervised Gated Depth Estimation

The proposed method learns to predict depth $\hat{\mathbf{r}}_t$ without ground truth supervision from LiDAR or simulation. To this end, we exploit the cyclic measurement consistency of gated images and temporal consistency in the depth predictions. Self-supervision allows us to overcome the limited depth range (80 m) of methods trained on LiDAR ground-truth and removes complex synchronization processes between LiDAR and cameras. Furthermore, we can train our models on harsh weather conditions, e.g., fog, rain, or snow, where LiDAR-based ground-truth is not available.

The proposed Gated2Gated architecture is illustrated in Figure 2. While our model is general in terms of the input gated slices, we consider three slices $\mathbf{Z}_t^i$, for $i = 0, 1, 2$, at each time t. The gated measurements $\mathbf{Z}_t^i$ are concatenated in a tensor $\mathbf{Z}_t$ that is fed to three convolutional neural networks that disentangle the input into albedo, ambient light, and depth, which are then used to reconstruct the input slices using a cyclic loss. The disentanglement is similar to [41, 52], where a network is trained to decompose an conventional RGB image into albedo and lighting components. In addition to this novel gated imaging-based training signal, we exploit temporal consistency between temporally adjacent gated frames to handle regions with shadows and multi-path reflections.

Specifically, the proposed architecture is composed of three networks. The first network predicts a dense depth map per gated tensor $\mathbf{Z}_t$, denoted as $f_{\mathbf{r}} : \mathbf{Z}_t \rightarrow \hat{\mathbf{r}}_t$. The second network also takes $\mathbf{Z}_t$ as input, and predicts ambient and albedo, denoted as $f_{\mathbf{\Lambda}\alpha} : \mathbf{Z}_t \rightarrow (\hat{\mathbf{\Lambda}}_t, \hat{\boldsymbol{\alpha}}_t)$. The third network takes two temporally adjacent gated tensors as input $(\mathbf{Z}_t, \mathbf{Z}_n)$, and predicts a rigid 6 DoF pose transformation $\hat{\boldsymbol{X}}$ from $\mathbf{Z}_n$ to $\mathbf{Z}_t$, denoted as $\hat{\boldsymbol{X}}_n = \begin{pmatrix} R_{\mathbf{t}\rightarrow\mathbf{n}} & t_{\mathbf{t}\rightarrow\mathbf{n}} \\ 0 & 1 \end{pmatrix}$, with $R_{\mathbf{t}\rightarrow\mathbf{n}} \in SO(3)$ and $t_{\mathbf{t}\rightarrow\mathbf{n}} \in \mathbb{R}^{3\times 1}$ generated by $f_{\mathbf{t}\rightarrow\mathbf{n}} : (\mathbf{Z}_t, \mathbf{Z}_n) \rightarrow \hat{\boldsymbol{X}}_{\mathbf{t}\rightarrow\mathbf{n}}$.

The learned function $f_{\mathbf{r}}$ is optimized to predict the absolute depth value and is supervised using the other two auxiliary functions, $f_{\mathbf{\Lambda}\alpha}$ and $f_{\mathbf{t}\rightarrow\mathbf{n}}$. The first auxiliary function is used to exploit cyclic measurement consistency with the measured gated slice, i.e., enforce that the predicted depth is consistent with a gated measurement. The second auxiliary function allows us to exploit temporal consistency between nearby gated frames. Using these cues, the proposed method resolves scale ambiguity inherent in monocular depth estimation. The two consistency components are discussed in the following sections.

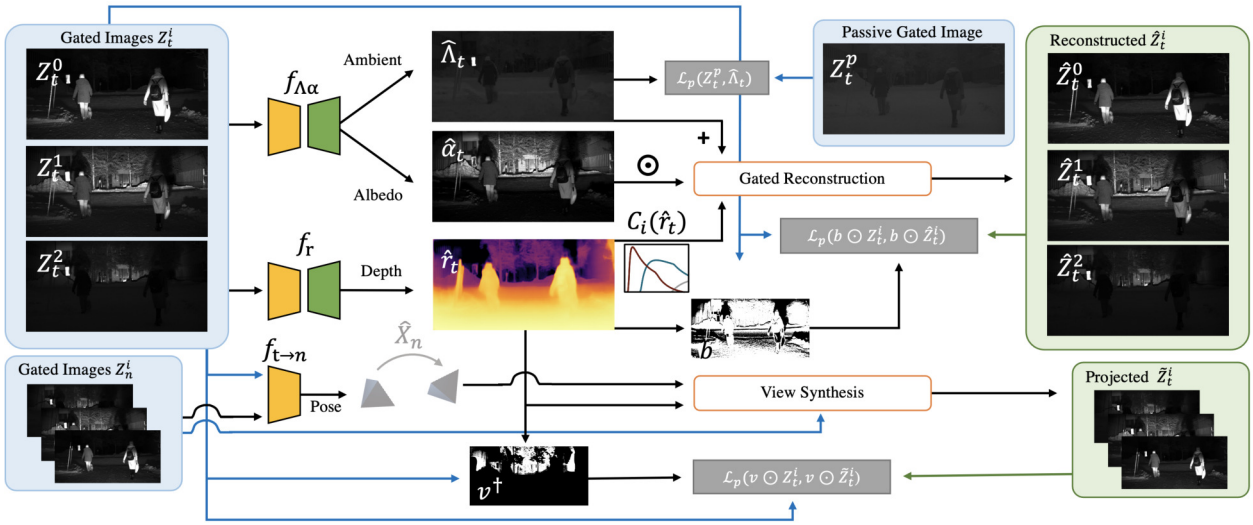

Figure 2. The proposed self-supervised Gated2Gated architecture estimates dense depth from a set of three gated images, by learning from cyclic gated and temporal consistency. Thereby, the inherent scale ambiguity is solved through the range intensity profiles introducing a scale cue during training. Additional, mask b, v resolve multi-path and shadow artifacts breaking Eq. (5).

4.1. Cyclic Gated Consistency

The cyclic gated consistency loss supervises the predicted depth $\hat{\mathbf{r}}_t$, ambient $\hat{\mathbf{\Lambda}}_t$ and albedo $\hat{\boldsymbol{\alpha}}_t$, by reconstructing the gated slices. To this end, we use a simplified version of Equation 3, in which we model ambient and noise together, that is

$$\hat{\mathbf{\Lambda}} = \eta_g + \eta_p + \Lambda. \tag{4}$$

The final model then can be written as

$$\hat{\mathbf{Z}}_t^i = \hat{\boldsymbol{\alpha}}_t C_i(\hat{\mathbf{r}}_t) + \hat{\mathbf{\Lambda}}_t, \tag{5}$$

with C_i being range-intensity profiles (as defined in Figure 1). The C_i profiles are measured experimentally with calibrated targets and approximated with Chebyshev polynomials T_n

$$T_0 = 1, \quad T_1 = x, \quad T_{n+1} = 2xT_n - T_{n-1}, \tag{6}$$

up to order of $N = 6$. The ambient predictions $\hat{\mathbf{\Lambda}}$ can be directly supervised using the ground truth captured passive images $\mathbf{Z}_t^p$ and the photometric loss $\mathcal{L}_p$.

As loss function $\mathcal{L}_p$ [18], we use Structural Similarity (SSIM) [57] and $\mathcal{L}_1$ norm, that is

$$\mathcal{L}_p(\mathbf{Z}_t^i, \hat{\mathbf{Z}}_t^i) = 0.85 \cdot \frac{1 - SSIM(\mathbf{Z}_t^i, \hat{\mathbf{Z}}_t^i)}{2} + 0.15 \cdot ||\mathbf{z}_t^i - \hat{\mathbf{z}}_t^i||_1. \tag{7}$$

Gated2Gated Cyclic Loss Masks. The gated slices $\hat{\mathbf{Z}}_t^i$ can be reconstructed using the proposed cyclic gated consistency, which enforces a match between a predicted depth estimate and the ground truth $\mathbf{Z}_t^i$ measurement it came from. Specifically, having predicted depth r using $f_{\mathbf{r}}$ and α using $f_{\mathbf{\Lambda}\alpha}$, we can predict a gated image using Eq. (5). However, adopting a photometric loss [18,61] between the measurement and the prediction fails in practice as severe multipath effects, missing illumination due to occlusion, and saturation due to retro-reflective signs can break the model in Eq. (5). We illustrate these issues in Figure 3. To this end, we introduce the following pixel masks that optimize the performance of the cyclic self-supervised model in those conditions.

Pixel Variance. Pixels p_{xy} with similar intensity across all slices are not modulated and are either beyond the range of the illumination, or highly absorptive. We then define a mask to filter out pixels with low variance as follows,

$$D_{xy} \coloneqq \{\, (x,y) \mid \left(\max_i \left(p_{xy}^i\right) - \min_i \left(p_{xy}^i\right) \right) > \theta \,\}. \tag{8}$$

Saturated Pixels. We also exclude saturated pixels, i.e., pixels with high-intensity values in all three gated slices, using the mask

$$M_{xy} \coloneqq \{\, (x,y) \mid \max_i (p_{xy}^i) < \gamma \,\}. \tag{9}$$

The pixel variance and saturated pixels masks are combined into a binary mask b'_{xy}, defined as

$$b'_{xy} = \begin{cases} 1 & \text{if } (x,y) \in D_{xy} \wedge (x,y) \in M_{xy} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Multipath Correction. The binary mask b' is further refined by modeling multipath effects, taking advantage of the view geometry, as illustrated in Figure 3. In automotive scenes, the most severe multipath effects result from reflective road surfaces. Using the camera intrinsics, we estimate a conservative constant ground plane with normal n and height h. Furthermore, we estimate an approximated depth measurement $\tilde{\mathbf{r}}$ by comparing the intensity values of the three gated slices. This allows us to filter out pixels (x, y) that get back-projected to 3D coordinates substantially lower than the ground plane

$$E_{xy} \coloneqq \{\, (x,y) \mid (\hat{\mathbf{r}} K^{-1} x_t) n < h \,\}, \tag{11}$$

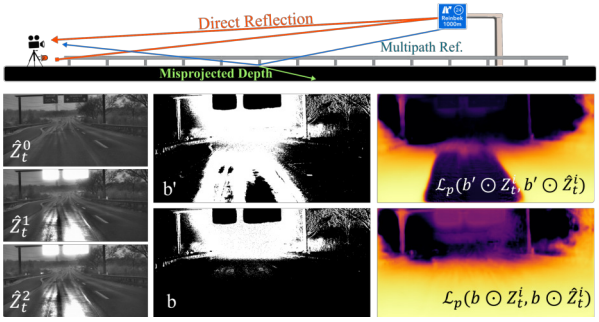

Figure 3. Illustration of the valid pixel mask b. Retro-reflective traffic signs reflect the illumination light back towards the camera (orange path), but spread out the light illuminating the ground causing multipath effects (blue path). This superposes the low intense groundplane intensity with the high intensity multi-path reflection containing further distant pulse information, leading to a wrong depth estimate (green path). Mitigation strategies are shown below using a valid pixel mask b' (middle) and next to it the depth prediction (right). The last row shows the refined mask b (middle) and the corresponding training output (right).

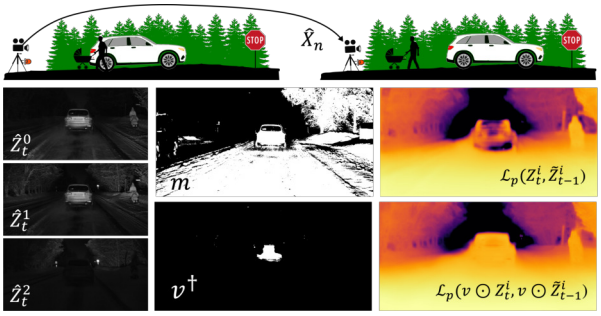

Figure 4. The top row shows two temporally adjacent frames and their rigid transformation $\hat{\mathbf{X}}_n$ as motivation for the valid mask v. Note that moving objects that remain in the same place in both frames are indistinguishable from objects at an infinite distance, causing holes in the predicted depth maps (middle right). Estimating close pixels in m (middle) infinite distances can be filtered according to the mask v (last row middle) with corresponding depth results (last row right).

where $x_t = [x, y, 1]$ denotes homogeneous pixel coordinates and K denotes the camera matrix. The final Gated2Gated cyclic loss mask b is then defined as

$$b_{xy} = \begin{cases} 1 & \text{if } (x,y) \notin E_{xy} \wedge b'_{xy} = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

and the cyclic loss function is defined as

$$\mathcal{L}_{cyc} = \sum_{i=0}^{2} \mathcal{L}_p(b \odot \mathbf{Z}_t^i, b \odot \hat{\mathbf{Z}}_t^i) + \mathcal{L}_p(\hat{\mathbf{\Lambda}}, \mathbf{Z}_t^p). \tag{13}$$

4.2. Temporal Depth Consistency

As illustrated in Figure 2, we use view synthesis to introduce temporal consistency between adjacent gated images during training. Specifically, we reconstruct the view of central gated image $\mathbf{Z}_t$ from temporal neighbors $\mathbf{Z}_n$ using the camera matrix K, the predicted depth $\hat{\mathbf{r}}_t$ and camera pose transformation $\hat{\mathbf{X}}_{t\to n}$. Considering x_t and x_n homogeneous (pixel) coordinates from $\mathbf{Z}_t^i$ and $\mathbf{Z}_n^i$, the mapping from source pixels x_t to target pixels x_n is defined as follows

$$x_n \sim K \hat{\mathbf{X}}_{t\to n} \hat{\mathbf{r}}_t K^{-1} x_t. \tag{14}$$

Similar to [61], we compare the reconstructed view $\hat{\mathbf{Z}}_t$ with $\mathbf{Z}_t$ using photometric loss and use it to train the depth prediction network $f_{\mathbf{r}}$. Unfortunately, this naive approach fails in the presence of moving occlusions due to ego-motion and the movement of non-stationary objects, which violate the rigid pose transformation. As earlier for the cyclic loss, we introduce validity masks as illustrated in Figure 4.

Infinity Correction Masks. To handle the dynamic scene objects, we use gated illumination cues to prevent projections to infinity (see Figure 4). Specifically, we define two valid sets per pixel position (x, y). The first set S^1 analyzes the pixel relation between first $\mathbf{Z}_t^0$ and last slice $\mathbf{Z}_t^2$. This allows to find valid pixels in a range from 3-s_0 m, where $\mathbf{C}_0(s_0) = c\mathbf{C}_2(s_0)$. The second set S^2 analyses the distance in the medium ranges 18-s_1 m (18 m due to the specific gate selection we made, see Supplemental Document) comparing the second $\mathbf{Z}_t^1$ and last slice $\mathbf{Z}_t^2$ with $\mathbf{C}_1(s_1) = c\mathbf{C}_2(s_1)$. In terms of measured pixel intensities the sets can be given as

$$S_{xy}^1 := \{\, (x,y) \mid \mathbf{Z}_t^0(x,y) \geq c \cdot \mathbf{Z}_t^2(x,y) \,\}, \tag{15}$$

$$S_{xy}^2 := \{\, (x,y) \mid \mathbf{Z}_t^1(x,y) \geq c \cdot \mathbf{Z}_t^2(x,y) \,\}. \tag{16}$$

This allows us to filter points illuminated by the last slice ranging from $\max(s_0, s_1)$-176 m (specific to our gates again) and to find close regions with a binary mask m

$$m_{xy} = \begin{cases} 1 & \text{if } (x,y) \in S_{xy}^1 \ \vee\ (x,y) \in S_{xy}^2 \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Then for the pixels with $P_{xy} := \{\, (x,y) \mid m_{xy} = 1 \,\}$ the average median $\bar{\mathbf{r}}_t$ is calculated. Comparing the median $\bar{\mathbf{r}}_t$ to the predicted depth $\hat{\mathbf{r}}_t$ allows us to filter mispredicted depth values stemming from moving objects which cannot be captured by the rigid transformation $\hat{\boldsymbol{X}}_n$, see Figure 4. Specifically, we omit all depth values twice the average median leading to a binary mask

$$v_{xy} = \begin{cases} 1 & \text{if } (x,y) \in V_{xy} \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

with set $V_{xy} := \{\, (x,y) \mid m_{xy} = 1 \ \wedge\ \hat{\mathbf{r}}_t(x,y) < 2 \cdot \bar{\mathbf{r}} \,\}$.

Additionally, scene points that are only visible either in the source or the target image break our image formation model. Such scenarios can be caused by foreground occluders obstructing the view of the background. In dynamic scenes, this occlusion changes in each timestep. For a triplet of adjacent frames $(t-1, t, t+1)$, we can define the minimum pixel error out of those pairwise differences, as occlusions cause a higher re-projection error. This can be

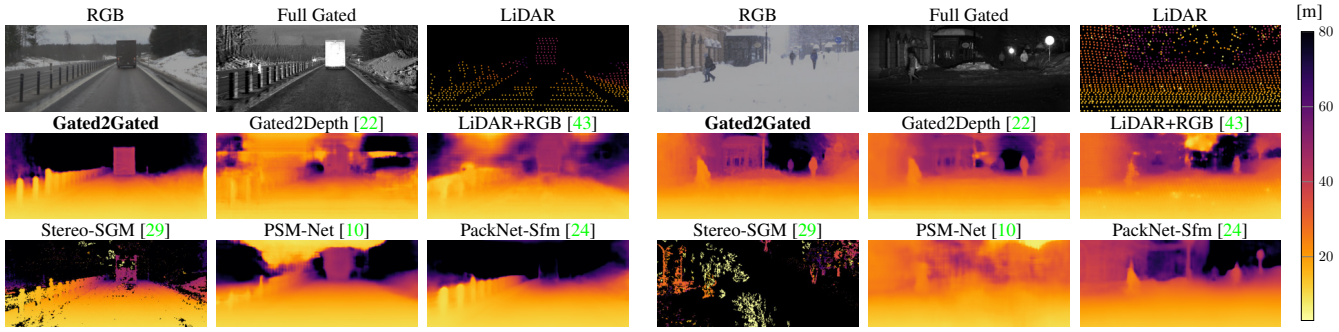

Figure 5. **Qualitative comparison of Gated2Gated and existing methods**. For two examples (a) clear day and (b) snow day, Gated2Gated predicts sharper depth maps than existing methods. (Full-Gated image refers here to an integral image $\sum_{i=0}^{2} \mathbf{Z}_t^i - 2\mathbf{Z}_t^p$).

explained as occluder and background have larger differences in texture than neighboring pixels in the background. Therefore, we calculate minimum of per-pixel loss between the re-projection from two temporal adjacent pairs similar to [19] as

$$\mathcal{L}_{temp} = \min_{n=\{t-1,t+1\}} \sum_{i=0}^{2} \left(\mathcal{L}_p \left(v \odot \mathbf{Z}_t^i, v \odot \hat{\mathbf{Z}}_t^i(n) \right) \right). \tag{19}$$

The complete loss function is

$$\mathcal{L} = \mathcal{L}_{temp} + \lambda_{cyc} \mathcal{L}_{cyc}, \tag{20}$$

where $\lambda_{cyc} = 0.01$ is determined empirically on the validation set.

5. Dataset

In order to train our proposed models, we collected 1835 video sequences, which comprised about 130,000 frames (note that previous gated imaging datasets [3, 22] do not include sequential data). Each time history is centered around one of the 13,000 middle frames and provides a temporal history of 1 second at a sampling rate of 10 Hz. The center frames are pre-selected by human annotators depending on the scene of interest from an underlying data distribution covering diverse winter road scenes collected in Northern Europe following the settings proposed in [3]. Please see the supplemental material for additional information.

We evaluate our proposed method on the open-source Gated2Depth [22] and Seeing Through Fog [3] datasets. While the first dataset contains a large variety of day and night captured images, the second contains diverse cluttered recordings in light fog, dense fog and snowfall conditions, which allow us to evaluate the performance of our model in harsh weather conditions and scenarios where obtaining ground truth data is difficult. In the last case, to filter out clutter from LiDAR ground truth, we use the DROR filtering algorithm [11], removing 8.2 % LiDAR points.

6. Experiments

In this section, we evaluate the proposed method by comparing it against the existing depth estimation approaches based on monocular gated, RGB images, and stereo RGB images; in both supervised and self-supervised training settings.

	Method	Modality	Train	RMSE [m]	ARD	MAE [m]	δ_1 [%]	δ_2 [%]	δ_3 [%]	Compl. [%]
Real Data – Night (Evaluated on Lidar Ground Truth Points)										
Supervised	PSMNet [10]	Stereo-RGB	D	14.58	0.21	8.34	68.75	82.63	89.36	100
	SGM [29]	Stereo-RGB	-	15.51	0.36	8.75	63.94	76.19	82.31	63
	Sparse-to-Dense [43]	Lidar(GT)+RGB	D	8.79	0.21	4.38	**87.64**	**93.74**	**95.88**	100
	Regression Tree [1]	Gated	D	10.54	0.24	6.01	76.73	89.74	93.45	40
	Least Squares	Gated	-	13.13	0.42	8.88	43.60	55.80	63.54	31
	Gated2Depth [22]	Full-Gated	D	14.86	0.29	8.84	58.79	58.79	79.84	100
	Gated2Depth [22]	Gated	D	**8.39**	**0.15**	**3.79**	87.52	93.00	95.21	100
Unsupervised	Monodepth [17]	RGB	S	11.41	0.23	6.18	76.64	89.53	94.19	100
	Monodepth [17]	Full-Gated	S	15.41	0.52	11.33	31.72	71.23	88.74	100
	*Packnet [24]	RGB	M	12.15	0.27	6.87	69.14	86.93	92.57	100
	*Packnet-Slim [24]	Gated	M	10.78	0.22	6.02	74.37	89.44	94.34	100
	*Monodepth2 [19]	RGB	M	14.92	0.38	9.98	39.85	68.57	83.99	100
	*Monodepth2 [19]	Gated	M	11.18	0.25	5.99	76.79	87.04	91.58	100
	Gated2Gated	Gated	MG	**9.43**	**0.21**	**4.86**	**82.17**	**91.54**	**94.48**	100
Ablation	**Gated2Gated** [v ✗, b ✗]	Gated	MG	10.05	0.27	5.36	80.06	90.44	93.75	100
	Gated2Gated [v ✓, b ✗]	Gated	MG	9.88	0.26	5.45	78.87	90.71	94.01	100
	Gated2Gated [v ✗, b ✓]	Gated	MG	9.58	0.25	5.03	80.68	91.25	94.40	100
	Gated2Gated [v ✓, b ✓]	Gated	MG	**9.43**	**0.21**	**4.86**	**82.17**	**91.54**	**94.48**	100
Real Data – Day (Evaluated on Lidar Ground Truth Points)										
Supervised	PSMNet [10]	Stereo-RGB	D	13.94	0.19	7.78	71.32	84.67	91.38	100
	SGM [29]	Stereo-RGB	-	9.63	0.17	4.59	85.80	92.72	95.20	86
	Sparse-to-Dense [43]	Lidar(GT)+RGB	D	8.21	0.16	4.05	**88.52**	**94.71**	**96.87**	100
	Regression Tree [1]	Gated	D	15.83	0.49	11.40	56.30	75.54	82.45	23
	Least Squares	Gated	-	19.52	0.75	14.05	43.42	54.63	63.76	16
	Gated2Depth [22]	Full-Gated	D	13.75	0.26	8.16	62.48	62.48	82.93	100
	Gated2Depth [22]	Gated	D	**7.61**	**0.12**	**3.53**	88.07	94.32	96.60	100
Unsupervised	Monodepth [17]	RGB	S	10.24	0.18	5.47	80.49	91.78	95.61	100
	Monodepth [17]	Full Gated	S	13.33	0.40	9.51	36.64	81.63	92.86	100
	*Packnet [24]	RGB	M	12.44	0.27	7.23	66.32	85.85	92.40	100
	*Packnet-Slim [24]	Gated	M	9.93	0.18	5.34	78.98	91.83	96.06	100
	*Monodepth2 [19]	RGB	M	13.18	0.33	8.79	44.99	75.87	90.65	100
	*Monodepth2 [19]	Gated	M	9.57	0.18	4.76	83.20	91.75	94.94	100
	Gated2Gated	Gated	MG	**8.46**	**0.17**	**4.37**	**83.56**	**93.12**	**96.09**	100
Ablation	**Gated2Gated** [v ✗, b ✗]	Gated	MG	9.29	0.22	4.99	80.74	91.88	95.40	100
	Gated2Gated [v ✓, b ✗]	Gated	MG	9.14	0.20	4.99	80.82	92.26	95.58	100
	Gated2Gated [v ✗, b ✓]	Gated	MG	8.85	0.20	4.75	81.20	92.57	95.85	100
	Gated2Gated [v ✓, b ✓]	Gated	MG	**8.46**	**0.17**	**4.37**	**83.56**	**93.12**	**96.09**	100

Table 1. Comparison of our proposed framework and state-of-the-art methods on the Gated2Depth 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.

6.1. Implementation Details

Although the proposed approach is not limited to a specific architecture, we estimate $f_{\mathbf{r}}$ depth maps with the PackNet [24] network architecture. To estimate ambient $f_{\mathbf{\Lambda}\alpha}$ and albedo $\hat{\boldsymbol{\alpha}}_t$, we use a UNet-based [47] architecture with a single encoder and two decoder heads. Pose transformation $f_{\mathbf{t}\to\mathbf{n}}$ is learned through the model introduced by Zhou *et al.* [61], without the explainability mask. The joint neural networks are implemented using PyTorch [45]. We de-

	Method	clear					light fog					dense fog					snow				
		RMSE	MAE	δ_1	δ_2	δ_3	RMSE	MAE	δ_1	δ_2	δ_3	RMSE	MAE	δ_1	δ_2	δ_3	RMSE	MAE	δ_1	δ_2	δ_3
Day	Monodepth RGB [17]	12.74	8.43	75.11	90.18	**94.81**	14.04	9.10	72.70	88.43	94.32	14.67	10.64	63.49	82.90	91.89	13.17	8.73	71.56	89.06	**94.81**
	Sparse-to-Dense [43]	13.66	9.85	54.20	82.42	91.47	14.23	10.66	49.75	79.62	90.10	18.50	15.35	37.04	64.67	78.25	13.42	9.81	53.12	82.29	92.04
	PackNet-SLIM G [24]	16.46	11.62	56.91	78.43	88.48	16.95	11.80	59.09	78.80	88.81	17.01	12.09	54.93	76.37	88.89	15.30	10.33	62.22	82.29	90.65
	Monodepth2 G [19]	13.26	7.40	78.59	88.95	92.77	18.17	10.43	72.91	83.20	89.27	15.56	8.72	76.79	85.38	90.68	12.84	7.12	**80.04**	89.34	93.13
	Gated2Depth [22]	11.48	6.60	79.17	87.38	91.58	11.28	6.63	81.20	88.66	92.56	11.86	7.85	71.72	87.10	91.70	11.28	6.61	78.87	87.93	92.50
	Gated2Gated	**11.15**	**6.31**	**80.82**	**90.48**	93.97	**10.70**	**6.01**	**84.71**	**91.52**	**94.65**	**11.09**	**6.86**	**81.09**	**91.43**	**94.47**	**10.97**	**6.28**	80.01	**91.12**	94.63
Night	Monodepth RGB [17]	13.78	8.92	72.63	88.48	93.37	13.30	8.75	72.52	88.88	**94.45**	16.31	10.99	69.15	85.92	**91.33**	15.28	9.89	68.88	86.86	**93.44**
	Sparse-to-Dense [43]	14.43	10.40	50.32	78.66	89.58	13.92	10.01	51.88	80.41	90.70	16.54	12.07	47.15	73.45	84.36	14.08	10.05	52.91	81.03	90.85
	PackNet-SLIM G [24]	15.81	11.11	59.80	79.52	88.65	16.01	11.23	58.44	80.60	89.57	17.49	12.60	57.72	77.77	87.26	16.47	11.40	60.17	79.55	88.62
	Monodepth2 G [19]	14.52	8.30	74.45	85.41	89.73	14.21	8.29	74.56	85.06	91.01	18.33	11.88	66.61	79.25	84.48	15.11	8.46	76.15	86.38	90.52
	Gated2Depth [22]	**10.06**	**5.17**	**84.81**	**90.59**	**93.39**	**9.94**	**5.37**	**81.95**	**89.80**	93.63	**12.51**	**7.72**	**76.90**	86.59	90.81	**10.70**	**5.81**	**81.81**	89.45	93.02
	Gated2Gated	11.69	6.74	80.25	89.58	92.83	11.29	6.46	79.39	89.31	93.17	13.52	8.69	76.43	**86.70**	90.61	11.91	6.80	80.76	**90.09**	93.31

Table 2. Evaluation of the proposed Gated2Gated framework and state-of-the-art-methods on adverse weather scenes. All metrics are evaluated in bins of approximately 7m to weight all distances equally. G indicates training and evaluation on gated images. Best results in each category are in **bold** and second best are underlined.

fine a 512×1024 gated image input resolution and trained the model on an NVIDIA A100 GPU with a batch size of four. As optimizer we use ADAM [34] with $\beta_1 = 0.9$ and $\beta_2 = 0.9999$ and learning rate of 10^{-4}. In total, the model is trained for 30 epochs, with first 10 epochs used to train the depth $f_{\mathbf{r}}$ and pose $f_{\mathbf{t}\to\mathbf{n}}$ prediction networks using temporal consistency only, and the last 20 epochs to jointly train ambient+albedo $f_{\mathbf{\Lambda}\alpha}$ alongside depth and pose. For valid pixel masks, we set $\gamma = 0.98$, $\theta = 0.04$ and $c = 0.995$, chosen with grid search on the validation set.

6.2. Assessment

Experimental Setup. We compare the performance of the proposed Gated2Gated method against state-of-the-art supervised and self-supervised depth estimation methods. As supervised approaches, we compare against gated depth estimation [1, 22], LiDAR depth completion [43] and stereo vision [9, 29]. For comparison with unsupervised methods, we consider stereo [18] and temporal based self-learning approaches [19, 24]. Since the self-supervised baseline methods do not provide absolute depth predictions, we follow previous works [19, 24, 61] and scale the estimated depth maps with median ground-truth LiDAR information. For completeness, we train and evaluate these approaches both on gated and on RGB input images. All supervised methods and [18] are trained on the Gated2Depth training set, where LiDAR labels are available. The self-supervised monocular approaches are trained on the proposed temporal dataset. For further training details, we refer the readers to the supplemental document.

Following [14], we evaluate using the metrics RMSE, MAE, ARD and $\delta_i < 1.25i$ for $i \in 1, 2, 3$. These metrics are computed for distances between 3m and 80m, limited by the maximum LiDAR distance. To evaluate the long range influence in adverse weather we rely on 7 m bins. We apply binned metrics in adverse weather following [21].

Evaluation on Clear Data - Gated2Depth Dataset. Table 1 reports a quantitative comparison of our proposed Gated2Gated method and other state-of-the-art methods on the test set of the Gated2Depth dataset [22]. Our model outperforms all other self-supervised methods [18, 19, 24], and even temporal approaches [19, 24] that use LiDAR ground truth for depth scaling. The proposed method also outperforms stereo [9, 29] and Bayesian-based gated depth estimation methods [1]. Among the supervised methods, only Sparse-to-Dense [43] and Gated2Depth [22] obtain better results. We note, however, that Sparse-to-Dense [43] relies on sparse ground truth depth inputs from LiDAR sensors. Figure 5 qualitatively compares our method to baseline methods for depth estimation. In Figure 5b, our method accurately shows all scene objects located at a farther distance – car, traffic signs and two pedestrians in adverse weather conditions like snowfall, whereas RGB-based and LiDAR-based methods deliver only poor depth predictions. While Gated2Depth [22] is also able to recover the scene elements, the generated depth map shows artifacts at a farther distance in contrast to our method. Similarly in Figure 5a, Gated2Gated generates accurate depth maps whereas all other methods fail here. Figure 6 shows the qualitative comparison of the supervised Gated2Depth [22] and the proposed self-supervised approach. While the performance metrics for Gated2Depth are on par with the proposed method, the qualitative comparison shows that Gated2Gated predicts much finer grain details and sharper object contours in depth maps. Furthermore, the proposed method generalizes better to far distances: Gated2Depth often estimates far distances and sky as close regions.

Ablation Study. To evaluate the individual contributions of all components of the proposed method, we perform an ablation study reported in Table 1. Without validity masks, the proposed method performs worst and improves when adding b or v the; RMSE score increases by 9.1 %. Training with both masks is mutually beneficial and provides a significant boost to all performance metrics. We provide additional ablation experiments in the Supplemental Document.

Evaluation on Adverse Weather Scenes – Seeing Through Fog Dataset. We also evaluate the proposed method in adverse weather, adopting the test splits provided in [3]. The performance is measured in binned metrics to weight all distances equally. Table 2 shows the quantita-

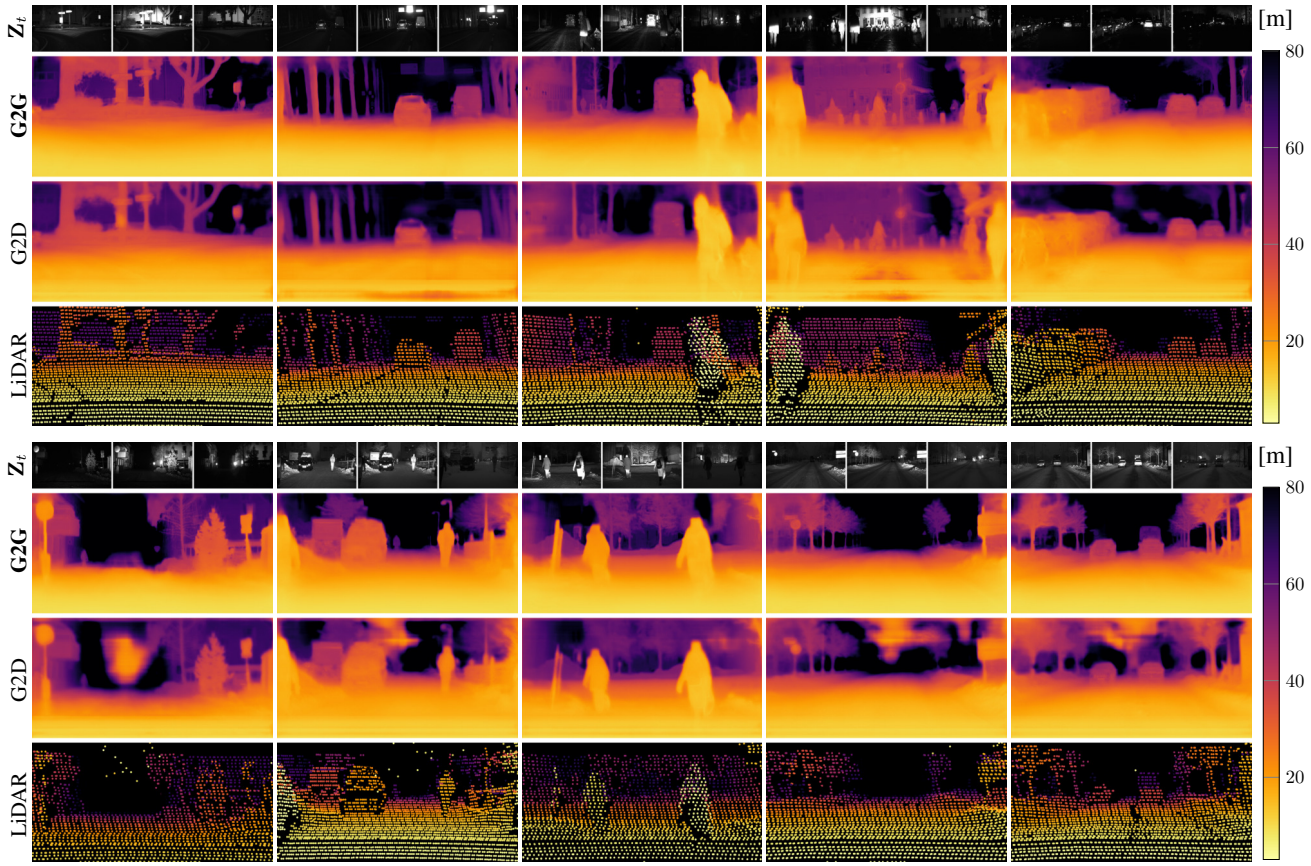

Figure 6. The top row for each example shows the gated image $\mathbf{Z}_t$ with three gated slices ($\mathbf{Z}_t^0$, $\mathbf{Z}_t^1$, $\mathbf{Z}_t^2$) for each capture, second row shows depth maps predicted by the **Gated2Gated (G2G)** – self-supervised method, third row shows depth predictions from Gated2Depth (G2D) [22] – supervised, and the bottom row shows corresponding LiDAR point clouds in gated view.

tive results of the Gated2Gated method and state-of-the-art methods. We note that absolute metrics may improve in adverse weather conditions, as the number and range of ground-truth LiDAR points decreases with worse weather conditions. We validate that Gated2Gated achieves robust performance overall weather conditions. In contrast, Monodepth2 and Sparse-to-Dense struggle to maintain performance in adverse weather. Since Sparse-to-Dense uses LiDAR points as additional inputs, wrong depth measurements from backscatter negatively impact the predicted depth maps. Furthermore, Table 2 validates that the proposed approach performs on par with Gated2Depth, and for daytime as well as in harsh weather scenarios, Gated2Gated even outperforms Gated2Depth. These results highlight the generalization capabilities of the proposed method over a wide range of distances and weather conditions.

7. Conclusion

We introduce Gated2Gated, a method that learns to estimate depth from gated images in a self-supervised fashion – only by observing gated video sequences. The proposed method exploits cyclic measurement and temporal consistency cues as training signals. This approach allows us to resolve monocular scale ambiguity by relying on gated illumination profiles and shadow/multi-path reflection via multi-view observations. To train our method, we create a novel gated video dataset containing 130,000 frames from 1835 sequences. We validate Gated2Gated in extensive real-world experimentation, where it outperforms fully supervised methods by up to 1.25 m (↓ 11.25%) in RMSE in daytime adverse weather and by at least 1.2 m (↓ 9.73%) independently of adverse weather. In the future, we plan to add wide-baseline active stereo cues to our self-supervised method by using two synchronized gated imagers.

8. Acknowledgments

This work was supported by Mitacs through the Mitacs Accelerate program. The work also received funding by the AI-SEE project with national funding from the FFG, BMBF, and NRC-IRA. We also thank the Federal Ministry for Economic Affairs and Energy for support within "VVM-Verification and Validation Methods for Automated Vehicles Level 4 and 5", a PEGASUS family project. Felix Heide was supported by an NSF CAREER Award (2047359), a Sony Young Faculty Award, and a Project X Innovation Award. We thank Karina Müller and Tom Riley for comments on the manuscript.

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, 6, 7

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

[3] 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, 7

[4] 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

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

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

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

[8] 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

[9] 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, 7

[10] 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

[11] Nicholas Charron, Stephen Phillips, and Steven L. Waslander. De-noising of lidar point clouds corrupted by snowfall. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 254–261, 2018. 6

[12] Richard Chen, Faisal Mahmood, Alan Yuille, and Nicholas J Durr. Rethinking monocular depth estimation with adversarial training. *arXiv preprint arXiv:1808.07528*, 2018. 2

[13] 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

[14] 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

[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. 2

[16] 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

[17] 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. 1, 6, 7

[18] 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, 7

[19] 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

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

[21] Tobias Gruber, Mario Bijelic, Felix Heide, Werner Ritter, and Klaus Dietmayer. Pixel-accurate depth evaluation in realistic driving scenarios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 95–105. IEEE, 2019. 7

[22] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide. Gated2depth: Real-time dense lidar from gated images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 6, 7, 8

[23] 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. 2, 3

[24] 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

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

[26] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

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

[28] Felix Heide, Wolfgang Heidrich, Matthias Hullin, and Gordon Wetzstein. Doppler time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 34(4):36, 2015. 2

[29] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, Feb 2008. 6, 7

[30] 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

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

[32] Frank Julca-Aguilar, Jason Taylor, Mario Bijelic, Fahim Mannan, Ethan Tseng, and Felix Heide. Gated3d: Monocular 3d object detection from temporal illumination cues. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[33] 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

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[35] 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

[36] 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

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

[38] 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

[39] 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

[40] 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. 1, 2

[41] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9039–9048, 2018. 3

[42] 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

[43] 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. 1, 2, 6, 7

[44] 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

[45] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[46] 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

[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. 6

[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

[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. 2

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

[52] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018. 3

[53] 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

[54] 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

[55] 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

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

[57] 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

[58] 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

[59] 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

[60] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 2

[61] 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, 4, 5, 6, 7