Gated2Gated: Self-Supervised Depth Estimation from Gated Images

Amanpreet Walia*1,3, Stefanie Walz*2, Mario Bijelic4, Fahim Mannan1
Frank Julca-Aguilar1, Michael Langer3, Werner Ritter2, Felix Heide1,4
1Algolux 2Mercedes-Benz AG 3McGill University 4Princeton 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 backscatter. 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 backscatter 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 multi-modal 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
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 backscatter 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
reflected light echo with a $\xi$ 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

$$Z^i_t(r) = \alpha C^i_t(r) = \alpha \int_{-\infty}^{\infty} g_i(t-\xi)p_i(t-\frac{2r}{c}) \beta(r)dt, \quad (1)$$

where $Z^i_t(r)$ is the gated exposure, indexed by $i$, at distance $r$ and time $t$; $C^i_t(r)$ is the range intensity profile, i.e., the convolution of the gated slice and its corresponding pulse profile; $\alpha$ is the surface reflectance (albedo), and $\beta$ 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 to the high spectral solar power within the NIR band that leads to a significant number of unmodulated photons captured as an ambient light $\Lambda$ component. We modify the model from [23] as

$$Z^i_t(r) = \alpha C^i_t(r) + \Lambda. \quad (2)$$

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

$$Z^i_t = \alpha C^i_t(r) + \Lambda + \eta_g + \eta_p. \quad (3)$$

In this work, at a given time $t$, we capture three sequential gated slices with spatial resolution $H \times W$ as $Z^i_t \in \mathbb{R}_+^{H \times W}$ with delays $\xi_{\{0,1,2\}}$ and an unmodulated NIR passive image $Z^0_t$ 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 $\hat{t}_t$ 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 $Z^i_t$, for $i = 0, 1, 2$, at each time $t$. The gated measurements $Z^i_t$ are concatenated in a tensor $Z_t$ 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 a 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 $Z_t$, denoted as $f_x : Z_t \rightarrow \hat{X}_t$. The second network also takes $Z_t$ as input, and predicts ambient and albedo, denoted as $f_{\Lambda\alpha} : Z_t \rightarrow (\hat{\Lambda}_t, \hat{\alpha}_t)$. The third network takes two temporally adjacent gated tensors as input $(Z_t, Z_{t-n})$, and predicts a rigid 6 DoF pose transformation $\hat{X}$ from $Z_t$ to $Z_{t-n}$, denoted as $X_n = \begin{bmatrix} \hat{R}_{t-n} & t_{t-n} \end{bmatrix}$. With $\hat{R}_{t-n} \in SO(3)$ and $t_{t-n} \in \mathbb{R}^{3 \times 1}$ generated by $f_{t-n} : (Z_t, Z_n) \rightarrow \hat{X}_{t-n}$.

The learned function $f_x$ is optimized to predict the absolute depth value and is supervised using the other two auxiliary functions, $f_{\Lambda\alpha}$ and $f_{t-n}$. 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.
4.1. Cyclic Gated Consistency

The cyclic gated consistency loss supervises the predicted depth $\hat{r}_t$, ambient $\hat{Α}_t$, and albedo $\hat{α}_t$, 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

$$\hat{Α} = η_b + η_p + Λ.$$  

The final model then can be written as

$$\hat{Z}_i = \hat{α}_t C_i(\hat{r}_t) + \hat{Α}_t,$$  

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 $T_n$

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

up to order of $N = 6$. The ambient predictions $\hat{Α}$ can be directly supervised using the ground truth captured passive images $Z_t^p$ and the photometric loss $L_p$.

As loss function $L_p$ [18], we use Structural Similarity (SSIM) [57] and $L_1$ norm, that is

$$L_p(Z_i^t, \hat{Z}_i) = 0.85 \cdot \frac{1 - SSIM(Z_i^t, \hat{Z}_i)}{2} + 0.15 \cdot ||Z_i^t - \hat{Z}_i||_1.$$  

Gated2Gated Cyclic Loss Masks. The gated slices $\hat{Z}_i$ can be reconstructed using the proposed cyclic gated consistency, which enforces a match between a predicted depth estimate and the ground truth $Z_i$ measurement it came from. Specifically, having predicted depth $r$ using $f_r$ and $\alpha$ using $f_{Λα}$, 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,

$$D_{xy} := \{ (x, y) | \max_i (p_{xy}^i) - \min_i (p_{xy}^i) > \theta \}. \quad (8)$$

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

$$M_{xy} := \{ (x, y) | \max_i (p_{xy}^i) < γ \}. \quad (9)$$

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

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

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 $\hat{r}$ 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

$$E_{xy} := \{ (x, y) | (\hat{r}K^{-1}x_t)n < h \}, \quad (11)$$
ing the camera matrix and the cyclic loss function is defined as shown below using a valid pixel mask v.

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).

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 $\hat{Z}_t$ from temporal neighbors $Z_n$ using the camera matrix $K$, the predicted depth $\hat{r}_t$ and camera pose transformation $\hat{X}_{t \rightarrow n}$. Considering $x_t$ and $x_n$ homogeneous (pixel) coordinates from $Z_t$ and $Z_n$, the mapping from source pixels $x_t$ to target pixels $x_n$ is defined as follows

$$x_n \sim K \hat{X}_{t \rightarrow n} \hat{r}_t K^{-1} x_t. \quad (14)$$

Similar to [61], we compare the reconstructed view $\hat{Z}_t$ with $Z_n$ using photometric loss and use it to train the depth prediction network $f_r$. 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 $Z^0_t$ and last slice $Z^n_t$. This allows to find valid pixels in a range from 3-80 m, where $C_0(s_0) = cC_2(s_0)$. The second set $S^2$ analyses the distance in the medium ranges 18-81 m (18 m due to the specific gate selection we made, see Supplemental Document) comparing the second $Z^1_t$ and last slice $Z^n_t$ with $C_1(s_1) = cC_2(s_1)$. In terms of measured pixel intensities the sets can be given as

$$S^1_{xy} := \{(x, y) \mid Z^0_t(x, y) \geq c \cdot Z^n_t(x, y)\}, \quad (15)$$

$$S^2_{xy} := \{(x, y) \mid Z^1_t(x, y) \geq c \cdot Z^n_t(x, y)\}. \quad (16)$$

This allows us to filter points illuminated by the last slice ranging from $\max(s_0, s_1)$-16 m (specific to our gates again) and to find close regions with a binary mask $m$

$$m_{xy} = \begin{cases} 1 & \text{if } (x, y) \in S^1_{xy} \lor (x, y) \in S^2_{xy} \\ 0 & \text{otherwise}. \end{cases} \quad (17)$$

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

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

with set $V_{xy} := \{(x, y) \mid m_{xy} = 1 \land \hat{r}_t(x, y) < 2 \cdot \bar{r}\}$. 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
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
\[
\mathcal{L}_{\text{temp}} = \min_{n=\{t-1,t+1\}} \sum_{i=0}^{2} \left( \mathcal{L}_p \left( v \odot Z_t, v \odot Z_i(n) \right) \right).
\]
The complete loss function is
\[
\mathcal{L} = \mathcal{L}_{\text{temp}} + \lambda_{\text{eye}} \mathcal{L}_{\text{eye}},
\]
where $\lambda_{\text{eye}} = 0.01$ 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.

---

**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 $\sum_{i=0}^{2} Z_i - 2Z_t$).

<table>
<thead>
<tr>
<th>Method</th>
<th>Modality</th>
<th>Train</th>
<th>RMSE</th>
<th>ARD</th>
<th>MAE</th>
<th>$d_{1}$</th>
<th>$d_{2}$</th>
<th>$d_{3}$</th>
<th>Compl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>D</td>
<td>14.93</td>
<td>0.21</td>
<td>6.34</td>
<td>40.72</td>
<td>71.37</td>
<td>89.30</td>
<td>100</td>
</tr>
<tr>
<td>PSNN [25]</td>
<td>RGB</td>
<td>D</td>
<td>15.51</td>
<td>0.26</td>
<td>6.93</td>
<td>65.94</td>
<td>76.19</td>
<td>82.31</td>
<td>100</td>
</tr>
<tr>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>D</td>
<td>15.41</td>
<td>0.52</td>
<td>8.88</td>
<td>51.33</td>
<td>70.36</td>
<td>88.74</td>
<td>100</td>
</tr>
<tr>
<td>Regression</td>
<td>LiDAR+RGB</td>
<td>D</td>
<td>8.75</td>
<td>0.57</td>
<td>7.67</td>
<td>97.64</td>
<td>95.34</td>
<td>95.68</td>
<td>100</td>
</tr>
<tr>
<td>Stereo-SGM</td>
<td>LiDAR+RGB</td>
<td>D</td>
<td>10.34</td>
<td>0.51</td>
<td>9.01</td>
<td>76.73</td>
<td>80.74</td>
<td>95.45</td>
<td>100</td>
</tr>
<tr>
<td>PackNet-Sfm</td>
<td>RGB</td>
<td>D</td>
<td>11.45</td>
<td>0.23</td>
<td>6.18</td>
<td>76.64</td>
<td>89.53</td>
<td>90.65</td>
<td>100</td>
</tr>
<tr>
<td>PackNet-Sfm</td>
<td>RGB</td>
<td>D</td>
<td>11.41</td>
<td>0.23</td>
<td>6.18</td>
<td>76.64</td>
<td>89.53</td>
<td>90.65</td>
<td>100</td>
</tr>
<tr>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>D</td>
<td>14.92</td>
<td>0.38</td>
<td>9.98</td>
<td>39.85</td>
<td>68.57</td>
<td>83.99</td>
<td>100</td>
</tr>
<tr>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>D</td>
<td>14.92</td>
<td>0.38</td>
<td>9.98</td>
<td>39.85</td>
<td>68.57</td>
<td>83.99</td>
<td>100</td>
</tr>
<tr>
<td>PackNet-Sfm</td>
<td>RGB</td>
<td>D</td>
<td>14.92</td>
<td>0.38</td>
<td>9.98</td>
<td>39.85</td>
<td>68.57</td>
<td>83.99</td>
<td>100</td>
</tr>
<tr>
<td>PackNet-Sfm</td>
<td>RGB</td>
<td>D</td>
<td>14.92</td>
<td>0.38</td>
<td>9.98</td>
<td>39.85</td>
<td>68.57</td>
<td>83.99</td>
<td>100</td>
</tr>
<tr>
<td>PackNet-Sfm</td>
<td>RGB</td>
<td>D</td>
<td>14.92</td>
<td>0.38</td>
<td>9.98</td>
<td>39.85</td>
<td>68.57</td>
<td>83.99</td>
<td>100</td>
</tr>
<tr>
<td>PackNet-Sfm</td>
<td>RGB</td>
<td>D</td>
<td>14.92</td>
<td>0.38</td>
<td>9.98</td>
<td>39.85</td>
<td>68.57</td>
<td>83.99</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Modality</th>
<th>Train</th>
<th>RMSE</th>
<th>ARD</th>
<th>MAE</th>
<th>$d_{1}$</th>
<th>$d_{2}$</th>
<th>$d_{3}$</th>
<th>Compl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Data – Day (Evaluated on Lidar Ground Truth Points)</td>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>15.84</td>
<td>0.19</td>
<td>7.84</td>
<td>71.32</td>
<td>84.47</td>
<td>95.38</td>
<td>100</td>
</tr>
<tr>
<td>Real Data – Day (Evaluated on Lidar Ground Truth Points)</td>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>15.84</td>
<td>0.19</td>
<td>7.84</td>
<td>71.32</td>
<td>84.47</td>
<td>95.38</td>
<td>100</td>
</tr>
<tr>
<td>Real Data – Day (Evaluated on Lidar Ground Truth Points)</td>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>15.84</td>
<td>0.19</td>
<td>7.84</td>
<td>71.32</td>
<td>84.47</td>
<td>95.38</td>
<td>100</td>
</tr>
<tr>
<td>Real Data – Day (Evaluated on Lidar Ground Truth Points)</td>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>15.84</td>
<td>0.19</td>
<td>7.84</td>
<td>71.32</td>
<td>84.47</td>
<td>95.38</td>
<td>100</td>
</tr>
<tr>
<td>Real Data – Day (Evaluated on Lidar Ground Truth Points)</td>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>15.84</td>
<td>0.19</td>
<td>7.84</td>
<td>71.32</td>
<td>84.47</td>
<td>95.38</td>
<td>100</td>
</tr>
<tr>
<td>Real Data – Day (Evaluated on Lidar Ground Truth Points)</td>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>15.84</td>
<td>0.19</td>
<td>7.84</td>
<td>71.32</td>
<td>84.47</td>
<td>95.38</td>
<td>100</td>
</tr>
<tr>
<td>Real Data – Day (Evaluated on Lidar Ground Truth Points)</td>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>15.84</td>
<td>0.19</td>
<td>7.84</td>
<td>71.32</td>
<td>84.47</td>
<td>95.38</td>
<td>100</td>
</tr>
<tr>
<td>Real Data – Day (Evaluated on Lidar Ground Truth Points)</td>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>15.84</td>
<td>0.19</td>
<td>7.84</td>
<td>71.32</td>
<td>84.47</td>
<td>95.38</td>
<td>100</td>
</tr>
<tr>
<td>Real Data – Day (Evaluated on Lidar Ground Truth Points)</td>
<td>Stereo-SGM</td>
<td>RGB</td>
<td>15.84</td>
<td>0.19</td>
<td>71.32</td>
<td>84.47</td>
<td>95.38</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

---

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 $f_{\theta}$ depth maps with the PackNet [24] network architecture. To estimate ambient $\alpha_{\theta}$ and albedo $\alpha_{\tau}$, we use a U-Net-based [47] architecture with a single encoder and two decoder heads. Pose transformation $f_{\theta + \tau}$ 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-
fine a $512 \times 1024$ 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 $f_d$ and pose $f_{e \rightarrow n}$ prediction networks using temporal consistency only, and the last 20 epochs to jointly train ambient+albedo $f_{\lambda \alpha}$ 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 $v$; 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 $Z_t$ with three gated slices ($Z_t^0$, $Z_t^1$, $Z_t^2$) 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.

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


[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 Zielenksi. Three-dimensional range-gated imaging at infrared wavelengths with superresolution depth mapping. In *SPIE Infrared Technology and Applications XXXV*, volume 7298, 2009. 2


