Robust 3D Reconstruction of Dynamic Scenes From Single-Photon Lidar Using Beta-Divergences

Quentin Legros, Julián Tachella, *Member, IEEE*, Rachael Tobin, Aongus McCarthy, *Member, IEEE*,
Sylvain Meignen, Gerald S. Buller, Yoann Altmann, *Member, IEEE*, Stephen McLaughlin, *Fellow, IEEE*,
and Michael E. Davies, *Fellow, IEEE*

Abstract—In this article, we present a new algorithm for fast, online 3D reconstruction of dynamic scenes using times of arrival of photons recorded by single-photon detector arrays. One of the main challenges in 3D imaging using single-photon lidar in practical applications is the presence of strong ambient illumination which corrupts the data and can jeopardize the detection of peaks/surface in the signals. This background noise not only complicates the observation model classically used for 3D reconstruction but also the estimation procedure which requires iterative methods. In this work, we consider a new similarity measure for robust depth estimation, which allows us to use a simple observation model and a non-iterative estimation procedure while being robust to mis-specification of the background illumination model. This choice leads to a computationally attractive depth estimation procedure without significant degradation of the reconstruction performance. This new depth estimation procedure is coupled with a spatio-temporal model to capture the natural correlation between neighboring pixels and successive frames for dynamic scene analysis. The resulting online inference process is scalable and well suited for parallel implementation. The benefits of the proposed method are demonstrated through a series of experiments conducted with simulated and real single-photon lidar videos, allowing the analysis of dynamic scenes at 325 m observed under extreme ambient illumination conditions.

Index Terms—3D reconstruction, single-photon lidar, robust estimation, Bayesian filtering, variational methods.

I. Introduction

FAST and reliable reconstruction of 3D scenes using single-photon light detection and ranging (lidar) is extremely important for a variety of applications, including environmental monitoring [1], [2], autonomous driving [3] and defence [4], [5]. While 3D profiles can be obtained from a range of modalities, single-photon lidar (SPL) offers appealing advantages, including low-power imaging, a capability for long-range imaging [6]–[8] or imaging in complex media such as fog/smoke [5] and turbid underwater environments [9] with excellent range resolution (of the order of millimetres [10]).

Over the last few years, a wide range of reconstruction algorithms has been proposed to reconstruct individual depth images from SPL data, e.g., [11]–[18]. Several algorithms have also been proposed to analyze distributed objects [16], [19]–[23], i.e., when multiple surfaces are visible within each pixel. Irrespective of the number of surfaces visible in each pixel, one of the main goals of these algorithms is to reconstruct high quality depth profiles from as small a photon budget as possible (see also [10], [24]–[26]) and it was shown that reconstruction from as few as one photon per pixel is possible under favorable observation conditions. Since single-photon lidar technology consists of illuminating the scene with a pulsed laser and analyzing the time of arrival (ToA) of reflected photons, successful reconstruction from a few return photons enables the consideration of shorter integration/acquisition times and thus the analysis of highly dynamic scenes. Note that most existing methods are to be used offline since the computational time required to reconstruct a point cloud is usually longer than the acquisition time allocated for a single frame. However, a recent study has presented results of the reconstruction of complex scenes at video frame rates [23].

A shared property of all the algorithms mentioned above is that they concentrate on the reconstruction of one point cloud per time frame, processing a video as a sequence of independent frames. While a method was recently proposed in [27] to jointly process batches of SPL frames, it remains computationally intractable for long video sequences due to memory requirements. Thus, there is a clear need for scalable and reliable methods able to adaptively process the increasing amount of single-photon data recorded by new single-photon avalanche diode (SPAD) detector arrays [28], [29], offering

a growing number of pixels. In [30], we proposed an online reconstruction method, relying on individual photon-detection events, e.g., binary frames, which was used to reconstruct sequentially a series of depth images using at most one photon per pixel and frame. Although this sequential approach leverages correlations between successive time frames, the method does not allow the analysis of histograms (i.e., reconstruction after several illumination periods), as the estimation is performed after each illumination period. Moreover, this method is not adapted to situations where the ambient illumination levels lead to low signal-to-background ratio (in particular, smaller than one).

In this article, we consider lidar data acquired using SPAD arrays and investigate a new 3D reconstruction algorithm that accounts for the temporal correlation between successive point clouds to be reconstructed. More precisely, we address the problem of the reconstruction of a temporal series of point clouds, where each point cloud is associated with a different integration period, over which the pulsed laser emits an arbitrary number of pulses. This integration period is user-defined and we assume that the scene is quasi-static during that period. In contrast to [30], the sequential reconstruction by the proposed algorithm is performed after an arbitrary number of periods and thus does not make any assumptions regarding the number of detection events to be used in the reconstruction of each point cloud. The most basic and fastest method to estimate the distance of an object from a SPL histogram is via log-matched filtering but this method fails if the background level due to ambient illumination and light scattering is too high. In such cases where the background cannot be neglected, it is traditionally included in the observation model and is estimated to improve the depth estimation. However, the resulting model makes the estimation process more complicated and slower iterative schemes (optimization-based [19] or simulation-based [24]) are classically used. Strong ambient levels are encountered in many practical applications, for instance in long-range imaging applications in free-space [6]–[8] and challenging imaging applications through scattering media such as turbid underwater environments [9], [31], [32] and through fog/smoke [5]. It is thus extremely important to develop methods adapted to such challenging observation conditions.

In contrast to existing reconstruction methods, we propose a depth estimation method that does not require the background level to be modeled while allowing robust depth estimation when the background cannot be neglected. Instead of defining a likelihood function based on an observation model assumed fully specified and accurate, we define a pseudo-likelihood which only depends on the target depth. A robust estimation procedure is then developed to account for the mismatch between the simplified observation model and the actual data distribution. More precisely, the proposed data fidelity term is based on a β-divergence instead of the classical Kullback-Leibler divergence, allowing efficient reconstruction performance in the presence of an unknown (and high) ambient illumination level. Adopting a Bayesian approach, this pseudo-likelihood is coupled, for each pixel and each frame, with a depth prior model to derive the so-called

pseudo-posterior distribution of the position of each surface. This distribution is then used to perform surface detection (i.e., deciding if a surface is actually visible or not) and to define the prior distribution of the point cloud in the next frame, in an adaptive fashion.

The main contributions of this work are:

- A new pseudo-Bayesian model for robust 3D reconstruction using streams of photon detection events in the presence of high ambient illumination levels
- An efficient depth estimation strategy which presents a fixed and predictable computational cost (in contrast to most existing methods which are iterative and which require a convergence criterion to be reached).
- A new online/sequential estimation strategy, proposed to the best of our knowledge for the first time, for reconstruction of dynamic 3D scenes from streams of photon detection events. This method based on assumed density filtering is highly scalable and computationally attractive. It also includes an automatic and principled surface detection method originally proposed in [33], included for the first time in a sequential reconstruction process.
- A re-interpretation of the classical matched-filtering approach adopted for depth estimation using SPL as a robust estimator.

The remainder of the paper is organized as follows. Section II recalls the classical observation models for 3D reconstruction using SPL measurements in the photon-starved regime, introduces the similarity measure based on β-divergences for robust estimation, and demonstrates its benefits for the analysis of a single frame. The spatio-temporal model and new online reconstruction method are detailed in Section III. Results of simulations conducted with real and synthetic sequences of frames/histograms are presented and discussed in Section IV. Conclusions are finally reported in Section V.

II. ROBUST DEPTH ESTIMATION

A. Observation Models

In this work, we consider a sequence of N temporal frames which consist of P pixels. More precisely, for each frame, the data associated with each pixel consists of a set of photon ToAs. This article addresses the reconstruction of dynamic 3D scenes where the N frames are processed sequentially to reduce data storage requirements and account for the temporal correlation between successive frames. In this section, we discuss observation models and estimation strategies for the analysis of a pixel of a single frame. For now, we assume that one surface is visible in each pixel and we do not introduce pixel/frame indices in this section to keep notation clear.

1) Ideal Model: Assuming that the ambient illumination and detector dark counts can be neglected, the recorded ToAs are only associated with photons originally emitted by the laser source. For a given pixel, the probability density function of a photon ToA $y \in (0, T)$, where T is the repetition period of

the laser source, is given by

$$f_0(y|d) = s_0\left(y - \frac{2d}{c}\right), \tag{1}$$

where d is the distance between the imaging system and the surface of interest and c is the light speed in the medium. In (1), we assume that the scene is approximately static (within each frame) and $s_0(\cdot)$ is the normalized instrumental response function (IRF) of the lidar system which can be measured during calibration of the imaging system. To simplify notation, we assume that $s_0(\cdot)$ is the same for all the pixels but the method proposed also applies if the shape of the IRF is pixel-dependent. Note that we also assume that the shape of $s_0(\cdot)$ remains the same for all the admissible values of d.

When K photons are detected, if the dead-time of the detector can be neglected, the photon ToAs are mutually independent (given d) and the joint likelihood is given by $f(y|d) = \prod_k f_0(y_k|d)$ with $y = \{y_k\}_k$. Due to the finite timing resolution of SPAD detectors, the recorded ToAs are not continuous variables but instead live on a grid, whose resolution depends on the system used. Thus, the photon-starved regime it is possible to consider a model equivalent to (1), based on Poisson noise and which can be expressed as

$$z_t|r, d \sim \mathcal{P}\left(r\tilde{s}_0\left(t - \frac{2d}{c}\right)\right), \quad t = 1, \ldots, N_T \tag{2}$$

where N_T is the number of non-overlapping time bins spanning $(0, T)$, $\tilde{s}_0\left(t - \frac{2d}{c}\right) = \int_{y=(t-1)T/N_T}^{tT/N_T} s_0\left(y - \frac{2d}{c}\right) dy$, and z_t is the number of photons detected in the tth time bin. In (2), $\mathcal{P}(\lambda)$ denotes the Poisson distribution with mean λ. Moreover, $r \geq 0$ is an amplitude parameter which mainly depends on the number of laser pulses sent during the frame, the efficiency of the detector and the reflectivity of the object. In this case, we obtain the joint likelihood $f(z|r, d) = \prod_t f(z_t|r, d)$, where $z = [z_1, \ldots, z_{N_T}]^T$ is the ToA histogram constructed from y and the maximum likelihood estimator of d can be computed by maximizing the cross-correlation between the logarithm of $\tilde{s}_0 = [\tilde{s}_0\left(1 - \frac{2d}{c}\right), \ldots, \tilde{s}_0\left(T - \frac{2d}{c}\right)]^T$ and z [34].

2) Accounting for Background Detections: In many practical applications however, the models in (1)-(2) are not well adapted as background illumination cannot be neglected. Detection events arising from dark counts and additional sources (e.g., solar background) often present a uniform distribution and a more accurate observation model is the following mixture of distributions [13].

$$f(y|d, w) = ws_0\left(y - \frac{2d}{c}\right) + (1 - w)\mathcal{U}_{(0;T)}(y), \tag{3}$$

where w is the pixel and frame dependent probability of a detected photon to be a "signal" photon originally emitted by the laser source. Note however that other distributions could be used to account for the nature of the background photons. This probability relates to the signal-to-background ratio (SBR) defined by SBR $= w/(1 - w)$. In that case, estimating d from the joint likelihood $f(y|d, w) = \prod_k f(y_k|d, w)$ becomes more challenging since w is usually unknown and needs to be estimated jointly with d. Even if w is known, the estimation

of d is challenging as the evaluation of $f(y|d, w)$ consists of a product of K mixtures. Similarly, the model in (2) becomes

$$z_t|r, d \sim \mathcal{P}\left(r\tilde{s}_0\left(t - \frac{2d}{c}\right) + b\right), \quad t = 1, \ldots, N_T, \tag{4}$$

where $b \geq 0$ represents the average background level which can be correlated with the target reflectivity r. The relationship between r, b and w is obtained through the SBR, i.e., SBR $= w/(1 - w) = r/(bN_T)$.

Although the models described in (3)-(4) are more appropriate than those in (1)-(2), they can still fail to describe the data accurately, in particular in scattering media where the distribution of the background photons is not uniform [32]. Moreover, the joint estimation of (w, d) or (r, b, d) requires iterative algorithms (e.g., [11], [13], [24], [28]) which can result in a significant computational bottleneck. For these reasons, we investigate robust estimation of d based on (1), whereby robustness relates to the mismatch between the postulated model and the actual distribution of the ToAs.

B. Robust Estimation Using β-Divergences

Modern methods for depth estimation from single-photon data are statistical methods which use the data likelihood either in a maximum penalized likelihood fashion or within a Bayesian framework [11], [13], [24], [28]. Maximum likelihood estimation (MLE) in this context is equivalent to minimizing the Kullback-Leibler (KL) divergence

$$D_{\mathrm{KL}}(\hat{f}(y)||f(y|\boldsymbol{\theta})) = \int \hat{f}(y)\log\left(\frac{\hat{f}(y)}{f(y|\boldsymbol{\theta})}\right)dy,$$

$$= \mathrm{Const.} - \frac{1}{K}\sum_k \log\left(f(y_k|\boldsymbol{\theta})\right) \tag{5}$$

between the empirical distribution of the ToAs, denoted as $\hat{f}(y) = \frac{1}{K}\sum_k \delta(y - y_k)$, with $\delta(\cdot)$ the Dirac delta function, and the distribution $f(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the set of parameters of the postulated model (i.e., $\boldsymbol{\theta} = d$ if (1) is used, and $\boldsymbol{\theta} = (d, w)$ if (3) is used). Moreover, maximum penalized likelihood approaches can be seen as methods aiming at minimizing a penalized KL divergence (see discussion in Section II-C).

In this work, we propose to investigate a robust divergence instead of the classical KL divergence when estimating d to account for the mismatch between $\hat{f}(y)$ and the postulated observation model. Moreover, since (1) is simple, i.e., it involves a single parameter, and yields satisfactory results in the low to moderate background regime, it seems reasonable to use this model $f_0(y|d)$ in our 3D imaging strategy instead of (3).

Among the different families of divergences, we concentrate on β-divergences defined by

$$D_\beta(g||h) = \frac{1}{\beta}\int g(y)^{1+\beta}dy$$
$$- \frac{\beta + 1}{\beta}\int g(y)h(y)^\beta dy + \int h(y)^{1+\beta}dy, \tag{6}$$

with $\beta > 0$, which generalize the KL divergence [35]. In addition to being robust to model mismatch (as will be

shown in Section II-D), this family of divergences allows a computationally attractive estimation of d, since the evaluation and minimization of the β-divergence is simple (e.g., simpler than γ-divergences [36]). Note that the KL divergence is asymptotically recovered when $\beta \to 0$. For our problem, we obtain

$$D_\beta(\hat{f}(y)||f_0(y|d)) = \text{Const.} - \frac{\beta + 1}{\beta K} \sum_k f_0(y_k|d)^\beta \quad (7)$$

where the constant (which depends on β) does not depend on d since we assume that the shape and the integral of $s_0(\cdot)$ does not depend on d, i.e., $\int f_0(y|d)^{1+\beta} dy$ does not depend on d, for any d in its admissible set. The performance of any depth estimator based on this β-divergence depends on the value of the divergence parameter β. Its impact on the proposed method will be discussed in Section II-D.

An interesting result is that using the histogram z (defined in the paragraph below (2)) instead of y in (7), i.e., using discretized ToAs, minimizing the β-divergence reduces to maximizing the cross-correlation between z and $\tilde{s}_0^\beta$, where the exponential function is applied element-wise. Indeed, the sum on the left-hand side of (7) becomes $z^T s_d^\beta$. In particular, when $\beta = 1$, the resulting estimator reduces to the depth estimator obtained via matched filtering. Matched filtering is a classical method for peak localization, and is optimal in the presence of white Gaussian noise. Nonetheless, it has been widely used in single-photon lidar analysis [37] and it has been shown empirically to provide similar or even better results than log-matched filtering. In this article, we focus on Bayesian estimation of the depth. Thus, detailed analysis of the minimum divergence estimator, i.e., using only (7) to estimate d (without additional prior information), is out of scope of this work and is left for future work.

C. Pseudo-Bayesian Estimation

While robust depth estimation using only point estimates is interesting, we are interested in computing measures of uncertainty about d as such information can also be propagated to estimate the object depth in future frames. Thus, we adopt a Bayesian viewpoint and use the β-divergence to construct a pseudo-Bayesian method. Let us assume that d is assigned a prior distribution $f(d)$. Note that this assumption is consistent with the method discussed in Section III, where the depth parameters of a given frame will be assigned a product of P independent distributions.

When considering the KL divergence as a similarity measure, the classical posterior distribution of d, can be obtained by solving

$$\min_{p(d) \in \mathbb{P}} L(p(d)), \quad (8)$$

where $\mathbb{P}$ is the set of all probability distributions, $-L(p(d))$ is the evidence lower-bound (ELBO),

$$L(p(d)) = K \mathbb{E}_{p(d)}[\text{CE}_{\text{KL}}(d)] + D_{\text{KL}}(p(d)||f(d)), \quad (9)$$

with $\mathbb{E}_{p(d)}[\cdot]$ the expectation with respect to $p(d)$ and where $\text{CE}_{\text{KL}}(d) = -\frac{1}{K} \sum_k \log(f(y_k|d))$ is the cross-entropy

between $\hat{f}(y)$ and $f_0(y|d)$ [36]. Note that in (9), the term $D_{\text{KL}}(p(d)||f(d))$ acts as a penalty enforcing the solution $p(d)$ to be similar to the prior distribution $f(d)$. Indeed the solution of (8) yields

$$p(d) = f(d|y) \propto f(d) \exp^{-K\text{CE}_{\text{KL}}(d)}. \quad (10)$$

As expected in (10), $\exp^{-K\text{CE}_{\text{KL}}(d)}$ is indeed proportional to the likelihood $f(y|d)$.

In a similar fashion, we build a pseudo-posterior distribution, which maximize the β-ELBO $-L_\beta(p(d))$, i.e.,

$$\min_{p(d) \in \mathbb{P}} L_\beta(p(d)), \quad (11)$$

where

$$L_\beta(p(d)) = K \mathbb{E}_{p(d)}[\text{CE}_\beta(d)] + D_{\text{KL}}(p(d)||f(d)), \quad (12)$$

and where

$$\text{CE}_\beta(d) = -\frac{\beta + 1}{\beta K} \sum_k f(y_k|d)^\beta \quad (13)$$

is the β-cross-entropy between $\hat{f}(y)$ and $f_0(y|d)$ [36]. The solution of (11) yields

$$p(d) = \tilde{f}(d|y) \propto f(d) \exp^{-K\text{CE}_\beta(d)}. \quad (14)$$

The solution $\tilde{f}(d|y)$ of (11) and the traditional posterior distribution $f(d|y)$ in (10) present very similar expressions, the main difference being the likelihood term in (10) which is replaced by $\exp^{-K\text{CE}_\beta(d)}$ in (14). Thus, $\tilde{f}(d|y)$ is referred to as pseudo-posterior distribution, as it relies on the pseudo-likelihood $\exp^{-K\text{CE}_\beta(d)}$.

While $\tilde{f}(d|y)$ is generally non-standard, its mean and variance can be efficiently computed via numerical integration, e.g., by discretizing the admissible domain of definition of d, especially since the expected support of d is bounded in practice. Thus, we use as depth point estimate the mean of the pseudo-posterior $\tilde{f}(d|y)$ and as measure of the uncertainty the variance of $\tilde{f}(d|y)$. In addition to providing summary statistics about the current depth, the mean and variance of the posterior distribution in (14) can also be incorporated in the prior model of the next frame, as will be discussed in Section III.

Although this article focuses on robust depth estimation from single-wavelength SPL, the approach proposed here can also be used when multispectral lidar data are available. The resulting β-divergence and pseudo-posterior distribution are detailed in Appendix.

D. Preliminary Comparative Study

Prior to applying the proposed robust depth estimation strategy to online 3D imaging, we assess the depth estimation quality for a single pixel. For this study, we consider two normalized IRFs with unitary integrals, depicted in Fig. 1 (top). We generated synthetic data using $T = 1500$ and $N_T = 1500$ time bins. The first IRF (red curve) in Fig. 1 is a real IRF measured in [38], [39] (473 nm) and presents a full width at half maximum (FWHM) of 28 bins. Each time bin corresponds to a 2 ps time interval. The second IRF (blue curve) presents a Gaussian shape with the same FWHM. This second IRF allows
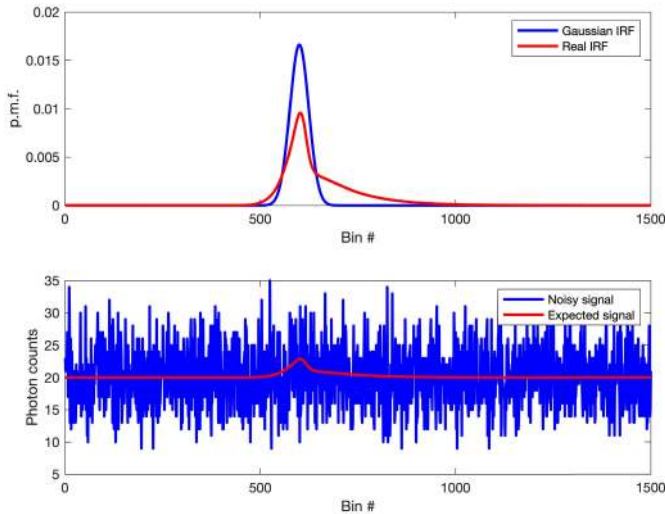
Fig. 1.   Top: Real (red) and Gaussian (blue) IRFs used to simulate synthetic data. The position of the peak is bin 600 for each IRF. Bottom: Example of expected (red) and noisy (blue) histogram of photon counts, using the real IRF, for 300 signal photons and SBR $= 10^{-2}$ (the peak is located at bin 600).
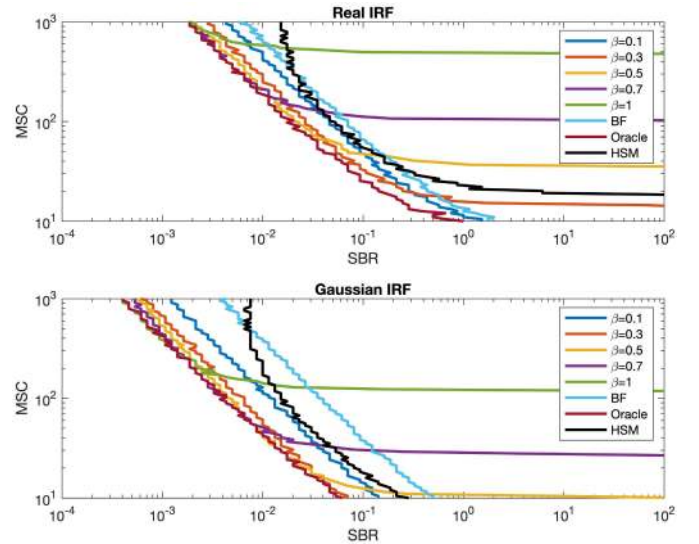

Fig. 2.   Curves of satisfactory detection ($p_d = 85\%$, with $\eta = 28$) as a function of the SBR and mean signal count (MSC) using the real IRF (top) and Gaussian IRF (bottom). This figure compares HSM and (pseudo-)MMSE estimators using the same Gaussian prior distribution with mean 600 and variance 2500.

us to investigate the impact of the asymmetry of the IRF on the depth estimation. For simplicity, we assume the position of the surface is associated with the time instant where the IRF has the highest amplitude. Given the high resolution of the discretization grid (compared to the shape of the IRF), using (3) or (4) yields similar results thus we do not distinguish the two models. For each IRF, we generated data from (4) and investigated various illumination scenarios with SBR in the interval $[10^{-4}, 10^2]$ and mean signal counts (MSC) (in each pixel) ranging from 10 to 1000. To illustrate the difficulty of the problem, an example of histogram with 300 signal photons and SBR of 0.01 is depicted in Fig. 1 (bottom). For each couple of SBR/signal photon values, we generated $N_{\text{MC}} = 2000$ histograms, drawing each depth parameter from a Gaussian distribution with mean 600 and variance 2500.

Here, the depth estimation performance for a single pixel is assessed through a comparison of (pseudo-)posterior means based on a fixed depth prior distribution. For such estimators, the prior distribution used is the Gaussian distribution with mean 600 and variance 2500 used to generate the data, which is a relatively weakly informative prior distribution. The reference estimator, referred to as "Oracle", is the minimum mean squared error (MMSE) estimator of d associated with (4), assuming that (r, b) is perfectly known. Similarly the MMSE estimator associated with (2) is referred to as "BF" for *background-free*. We also include a non-parametric estimator of d, namely the half-sample mode estimator [40], denoted by "HSM". Additional robust estimators could be considered, such as the Huber estimator [41], [42]. However, the latter requires sensitive parameter tuning (depending on the SBR) and does not provide satisfactory results in the low SBR regime of interest here, where the photon detections considered as outliers can represent more than 99% of the detected photons. Thus, we only report the results of HSM. Finally, we consider the pseudo-posterior mean (pseudo-MMSE estimator) obtained from (14). It is referred to as "PB" for *pseudo-Bayesian*.

To assess if the competing methods can accurately estimate the depth parameters, we define the empirical probability of "satisfactory" detection [33] as

$$p_d = \frac{1}{N_{\text{iter}}} \sum_{n=1}^{N_{\text{iter}}} \mathbb{I}\left(|\hat{d}_n - d_n| < \eta\right) \tag{15}$$

where d_n (resp. $\hat{d}_n$) is the actual (resp. estimated) depth estimate and $\mathbb{I}(\cdot)$ is the indicator function, which is equal to 1 if $|\hat{d}_n - d_n| < \eta$ and 0 otherwise. Moreover, η is a parameter reflecting which error is deemed acceptable. Here we set $\eta = 28$ (the FWHM of the IRFs).

Fig. 2 compares the depth estimation performance of HSM, Oracle, BF and PB for five values of β. This figure displays, for each method, the curve of satisfactory detection ($p_d = 85\%$) as a function of the SBR and signal photon counts. On the right-hand side of each curve, each method yields $p_d > 85\%$. As mentioned above, the prior distribution has been set to be weakly informative to better highlight the behavior of the different methods. If the prior distribution was more informative and correct (e.g., properly centered around the actual depth value), it would dominate the (pseudo-)likelihood factors in the (pseudo-)posterior distributions and all the methods would present similar behaviors and improved performance. Fig. 2 thus illustrates how the different methods perform when limited information is available about the unknown depth. As expected, the Oracle provides the best results for both IRFs. The BF estimator is significantly less accurate for low SBRs. Using the PB estimators, the curves approach those of BF for small values of β, while they tend to converge towards the Oracle curves when increasing β, provided that the MSC is large enough. This figure also shows that performance of all the estimators depends on the skewness of the IRFs.
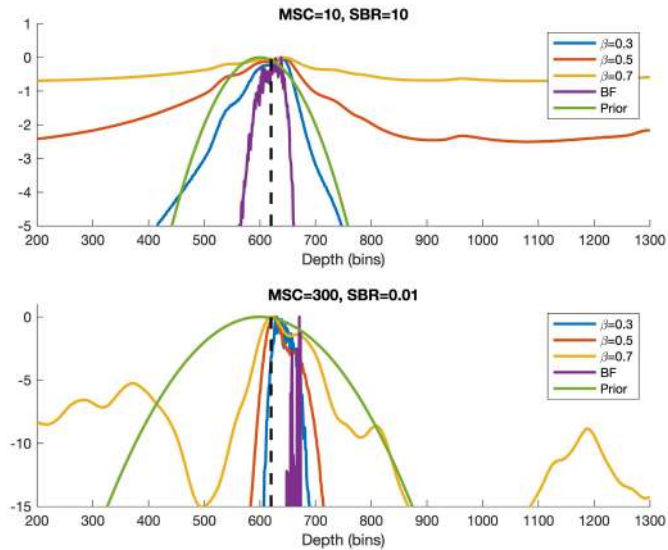
Fig. 3. Comparison of the log-likelihood (BF) and log-pseudo-likelihood terms (PB) for (MSC, SBR) = (10, 10) (top) and (MSC, SBR) = (300, 0.01) (bottom), as a function of $d \in [200, 500]$. The dashed black lines represent the actual position of the object ($d = 620$). Note that the scale of the y-axis is different in the two plots.

Overall, HSM leads to less accurate results than the other methods, partly because it does not leverage the shape of the underlying IRF, and because it does not use additional prior information. Note also that HSM is not accurate when the MSC is low since the mode of the distribution then becomes difficult to estimate (see Fig. 2 (top)).

Irrespective of the IRF shape, the parameter β of the divergence plays a key role in the working region of the resulting algorithm. For a given MSC level, increasing β allows a reduction of the limiting SBR below which the algorithm starts to fail. This limiting SBR however remains bounded by the limiting SBR of the Oracle. On the other hand, increasing β also increases the minimum MSC required for the algorithm to perform satisfactorily. For instance, in Fig. 2 (top), with $\beta = 0.5$ it is still possible to estimate the depth accurately for SBR > 1 and 35 signal photons while setting $\beta = 0.7$ in such scenarios leads to poor results, potentially worse than BF. Setting β depends on shape of the IRF but also on the expected SBR/MSC. Note that if the SBR is sufficiently large, the background effect is not significant and the depth reconstruction does not require a robust method. These first results show near-optimal results can be obtained with PB for large values of β (provided that MSC is large enough), without estimating additional model parameters and with a fixed computational budget.

To better illustrate the impact of the β-divergence in the various MSC/SBR scenarios, we show in Fig. 3 how the parameter β affects the pseudo likelihood term and in turn the pseudo-posterior distribution. In these two examples, we generated data with (MSC, SBR) = (10, 10) and (MSC, SBR) = (300, 0.01). The actual position of the object is at bin 620 and the Gaussian prior distribution has a mean of 600 and variance 2500. The curves in Fig. 3 represent the logarithm of the (pseudo)-likelihood terms for different values of β, the likelihood term assuming no background (BF) and the log-

prior. All the curves are normalized such that the maximum of each curve is 0 since vertical offsets do not affect the posterior distribution.

When (MSC, SBR) = (10, 10) (Fig. 3, top plot), the likelihood (1) is similar to the actual distribution of the data and the maximum of the log-likelihood is located around the actual depth value. When β increases, the pseudo-likelihood term becomes flatter, which gives more weight to the prior distribution in the pseudo-posterior distribution in (14). Thus the mean of the pseudo-posterior distribution tends to the mean of the prior distribution. In this regime, small values of β should be preferred. When (MSC, SBR) = (300, 0.01) (Fig. 3, bottom plot), the likelihood (1) becomes very concentrated, potentially around a value that is far from the actual depth due to the mismatch between the background-free observation model and the actual distribution of the data. This likelihood term is more concentrated than the prior and thus dominates the posterior distribution, leading to a poor MMSE depth estimate. In a similar fashion to the previous scenario, when β increases, the pseudo-likelihood term becomes flatter, which gives more weight to the prior distribution in the pseudo-posterior distribution in (14). Note that the mode of the pseudo-likelihood can also change when β changes. For instance, the mode for $\beta = 0.3$ is much closer to the actual depth than when using the likelihood (1). As discussed when analyzing Fig. 2, in this regime, larger values of β are preferred.

For practical applications, it is thus important to select β sensibly based on the expected observation conditions, balancing performance at low MSCs and low SBRs. Fortunately, it is possible to pre-compute performance bounds as in Fig. 2, for any real instrumental response. Moreover, from the preliminary experiments we conducted, it seems that using $\beta \in [0.3, 0.6]$ leads to a good trade-off for both the low SBR and low MSC regimes.

E. Target Detection and Additional Parameter Estimation

In Section II-C, we proposed to estimate the depth of a surface in a given pixel based on (1). This model assumes that an object is actually present in the pixel considered, which is not always true, especially for long range imaging applications. Thus, it is important to be able to decide whether a surface is actually present and this cannot be achieved directly using (14). To address this problem, we use the Bayesian object detection algorithm proposed in [43]. This method uses the Poisson likelihood model in (4) and assigns prior distributions $f(r)$, $f(b)$ and $f(d)$, to the reflectivity r of the target, the background level b and the target depth, respectively. The detection is finally seen as a binary hypothesis test where the background, and the target reflectivity and depth are marginalized. More precisely, the algorithm decides, a posteriori, whether $r = 0$ (no surface) or $r > 0$ (surface present). The output of the algorithm is π, the posterior probability of target presence, which can then be thresholded to derive a detection map (the interested reader is invited to consult [33], [43] for additional details about the detection method). Note that the algorithm also incorporates the prior probability of target presence π^0 (through a spike-and-slab

prior model $f(r)$). This point will be further discussed in Section III-B, together with the choice of the other prior distributions used for object detection.

After the pixel-wise detection procedure, it is possible (if needed) to estimate the average background level and the target reflectivity (if a target is detected), for instance using (4). To keep the computational overhead low, we only report a fast method here, but more complex schemes, as in [6], [11], [24] could be used. If no target is detected, the background level is estimated by dividing the photon count K by T. If an object is detected, we use as depth point estimate the mean of the pseudo-posterior distribution in (14), and the target reflectivity is estimated together with the background by MLE using (4).

III. APPLICATION TO ONLINE RECONSTRUCTION

In this section, we consider a set of N sequential temporal periods during which a ToA histogram is recorded for each of the P pixels. We denote by $\mathbf{y}_{n,p}$ the set of photon ToAs and $d_{p,n}$ the depth of the object in the pixel p and frame n.

A. Approximation Using Assumed Density Filtering

As mentioned in the introduction, our online estimation procedure consists of leveraging the temporal correlation between successive frames by incorporating the posterior distribution of the depth profile at time $(n-1)$ in the inference problem at time n. As described in [30], estimating the posterior mean and variance of $d_{p,n}$ presents a significant advantage beyond simply providing summary statistics about the current range profile. It allows the derivation of tractable adaptive estimation procedures. A classical choice for modeling relatively slowly evolving parameters is the Gaussian random walk (RW), i.e.,

$$f_t(d_{p,n}|d_{p,n-1}) \propto \exp\left\{-\frac{(d_{p,n}-d_{p,n-1})^2}{2\sigma_{\text{RW}}^2}\right\}, \quad (16)$$

controlled by the variance σ_{RW}^2. This RW mostly allows displacements smaller than $3\sigma_{\text{RW}}$ along the direction of the observation (using the 3-sigma rule of thumb). Whilst this approach is simple, it does not allow for rapid changes as might occur when the imaging system or the scene moves orthogonally to the direction of observation. To alleviate issues associated with such changes while keeping the estimation strategy tractable, we define as in [30], for each pixel, a local neighborhood $\mathcal{V}_p$ of M neighbors (including the current pixel) and define the following prior model

$$f(d_{p,n})$$
$$\propto \sum_{p'\in\mathcal{V}_p} v_{p'} \int f_t(d_{p,n}|d_{p',n-1}) q_{p',n-1}(d_{p',n-1}) \mathrm{d}d_{p',n-1},$$
$$(17)$$

where $\{q_{p,n-1}(\cdot)\}_p$ are Gaussian distributions. Basically, the prior model of $d_{p,n}$ is constructed via a Gaussian mixture model using the depth information in neighboring pixels at the previous frame, convolved by a Gaussian RW. In [30], $\{q_{p,n-1}(\cdot)\}_p$ was the set of Gaussian approximations of the

depth posterior distributions obtained at frame $(n-1)$. A similar approach is adopted here since the Gaussian approximations of the posterior distributions can be obtained as for assumed density filtering (ADF) [44], [45] and expectation-propagation [46], i.e., by minimizing the KL divergence

$$D_{\text{KL}}\left[\tilde{f}(d_{p,n-1}|\mathbf{y}_{p,n-1})||q_{p,n-1}(d_{p,n-1})\right] \quad (18)$$

w.r.t. $q_{p,n-1}(d_{p,n-1})$ which belongs to the family of Gaussian distributions and where $\tilde{f}(d_{p,n-1}|\mathbf{y}_{p,n-1})$ is the pseudo-posterior distribution described in Section II-C. This minimization reduces to matching the mean and variance of $\tilde{f}(d_{p,n-1}|\mathbf{y}_{p,n-1})$ and $q_{p,n-1}(d_{p,n-1})$, hence the discussion about the estimation of the moments of $\tilde{f}(d_{p,n-1}|\mathbf{y}_{p,n-1})$ in Section II-C.

However, one of the main limitations of the spatio-temporal model used in [30] is that it does not explicitly take into account whether objects were actually present in the neighboring pixels in frame $(n-1)$ when building the prior model for frame n. To address this problem, we incorporate the results of the detection procedure detailed in Section II-E. If an object is detected in pixel p' and frame $(n-1)$, $q_{p',n-1}(\cdot)$ is set to the Gaussian approximation of the pseudo-posterior (14) in that pixel, as in [30]. If no object is detected in pixel p' and frame $(n-1)$, $q_{p',n-1}(\cdot)$ is replaced by a Gaussian distribution with mean $(d_{\min}+d_{\max})/2$ and variance $(d_{\max}-d_{\min})^2/12$, where $(d_{\min},d_{\max})$ are the expected lower and upper bound of the scene depth. This choice of mean and variance leads to a flat prior distribution mimicking the uniform distribution defined on $(d_{\min}; d_{\max})$. Note that in practice, T is chosen large enough so that for any depth in $(d_{\min}; d_{\max})$, the shape of $s_0(y-2d/c)$ remains the same. Increasing the variance of the Gaussian distributions $q_{p',n-1}(\cdot)$ of empty pixels allows us to better detect new objects appearing in the scene and random depths. A similar strategy is adopted at the edges of the image where pixels keep M neighbors, some of them being outside the field of view and contributing to the mixture with weakly informative Gaussian distributions.

The M weights of the mixture in (17) are set to

$$v_{p'} = \begin{cases} v_0 \in [0,1], & \text{if } p'=p \\ \dfrac{(1-v_0)}{M-1} & \text{otherwise,} \end{cases} \quad (19)$$

where v_0 is a user-defined weight which controls the weight assigned to the central pixel of each neighborhood.

B. Online Target Detection

As mentioned in Section II-E, the target detection algorithm proposed in [43] requires, for each pixel of the frame n, prior distributions for the background level, the target depth, its reflectivity and a prior probability of target presence to be refined. An exponential prior model is chosen for the background, whose mean is given by the background estimate obtained at that pixel in the previous frame, assuming that the background varies slowly over time. Similarly, let $\{\pi_{p,n-1}\}_p$ be the probabilities of target presence estimated at frame $(n-1)$. The prior probabilities of target presence $\{\pi_{p,n}^0\}_p$ of

Algorithm 1 R3DSP algorithm

1: <u>Fixed input parameters</u>: Variance of RW for dynamic model: s^2, Neighborhood size M, parameter of GMM ν_0, indices of faulty pixels.
2: <u>Initialization</u> ($n = 0$)
3: Set $(q_{p,0}(\cdot), \pi_{p,0}^0), \forall p$.
4: **for** $n = 1, \ldots N$ **do**
5:    **for** $p = 1, \ldots P$ **do**
6:       Compute the prior model $f(d_{p,n}), \forall p$ from (17) via ADF.
7:       Compute the pseudo-posterior $f(d_{p,n}|\boldsymbol{y}_{p,n})$ in (14).
8:       Compute $q_{p,n}(d_{p,n})$ using (18).
9:       Estimate $\pi_{p,n}$ using [43].
10:       **if** target present, i.e., $\pi_{p,n} > 0.5$ **then**
11:          Set the estimated depth $\hat{d}_{p,n}$ as the mean of $q_{p,n}(d_{p,n})$.
12:          Estimate current background level and target reflectivity.
13:       **else**
14:          Set $\hat{d}_{p,n} = \emptyset$.
15:          Estimate current background level.
16:       **end if**
17:    **end for**
18:    Compute $\{\pi_{p,n+1}^0\}, \forall p$ using (20).
19: **end for**

---

the frame n are computed by nonlinear local averaging

$$\pi_{p,n}^0 = \sigma \left(\sum_{p' \in \mathcal{V}_p} \nu_{p'} \sigma^{-1} \left(\pi_{p,n-1} \right) \right), \qquad (20)$$

where $\sigma(x) = (1 + \exp(-x))^{-1}$ is the logistic function. Nonlinear averaging is preferred here over simple averaging as it promotes values closer to 0.5, and is hence less informative. As in [43], the prior model for the reflectivity (assuming a target is present) is a gamma distribution whose parameters are set using the calibration measurements (which provide insight about the expected signal photon counts when an object is present). In contrast to the background prior, the depth prior model used for the detection step is not set using a predictive model such as (17). Instead, it is set in an empirical Bayes fashion using the pseudo-posterior computed in (14).

The pseudo-code of the proposed method, referred to as R3DSP (for Robust 3D reconstruction using Single-Photon data) is presented in Algo. 1. While the pseudo-code includes a loop of the P pixels for each frame, it is important to remark that all the pixels of a frame can be processed independently and in parallel fashion. Thus the resulting algorithm is scalable and well adapted for GPU-based implementation. Note that the proposed algorithm can also be applied in the presence of faulty pixels for which no observations are available. In that case, we simply set $\tilde{f}(d_{n,p}|y_{np}) = f(d_{p,n})$ and $\pi_{p,n} = 0.5, \forall n$.

IV. RESULTS

To demonstrate the benefits of the proposed algorithm, we used a bistatic transceiver system incorporating a Princeton Lightwave Kestrel camera that was capable of providing picosecond resolution, time-tagged, single-photon data from its 32×32 SPAD detector array, which captures 150,400 binary frames per second (see [23] for more details). The

transceiver system used a sub-nanosecond pulsed laser source operating at a wavelength of 1550 nm to flood-illuminate the scene of interest. We acquired a series of 3D videos using $T = 153$ histogram bins (binning resolution of 3.75 cm) and here, we report results obtained from two videos, both measured in daylight conditions with significant ambient light background.

In the first experiment, we integrated the binary acquisitions into 500 lidar frames per second. At this frame rate, each lidar frame is composed of about 300 binary frames, i.e., contains at most 300 photons (see [23] for additional details about the experimental setup). We considered a dynamic scene which consists of two people, standing approximately 1.5 metres apart, exchanging a ≈ 220 mm diameter ball at a distance of 320 metres from the lidar system. Note that in this configuration, the depth estimation performance is not expected to be altered by the distance of the observed objects within the range associated with the detector gate (4-5 meters here). Approximately half of the pixels do not contain any surface and a single peak is usually observed in the remaining pixels (either one of the two pedestrians or the ball). In each pixel and frame, we observe approximately 35 photons related to dark counts or ambient illumination from solar background, and the visible surfaces lead to 55 additional photons per pixel, on average. For those pixels, the SBR is thus ≈ 1.6. The IRF of each pixel was recorded during the system calibration. Although they could have been approximated by Gaussian IRFs, we used the actual IRFs during our analysis as they are slightly skewed. The proposed algorithm has been applied to a series of 3000 successive frames, representing a total acquisition of 6 s. For this scenario, we set, $M = 5$ neighbors, $\sigma_{\text{RW}} = \sqrt{3}$ bins, $\nu_0 = 0.5$ and $\beta = 0.5$. However, we did not notice significant changes when using $\beta \in [0.4, 0.7]$. The reconstructed point clouds, together with a standard video of the scene recorded by a camera located next to the two people are presented in the Video 1. In all the videos presented in this work, the colormap of the point cloud represents the amplitude (number of signal photons) of the returns in the lidar data.

We compared the performance of the proposed method to that of two methods able to handle rapidly thousand of frames. First, we consider the classical depth MLE estimator BF assuming (2) applied independently to each pixel, and followed by an intensity-based thresholding step (5% of the IRF intensity) using the intensity estimated via MLE and (3). This thresholding step allows us to remove estimated surfaces which present too low an intensity. The second method is the RT3D algorithm [23] recently proposed for fast reconstruction of complex (multi-surface) scenes. While RT3D aims at solving a more complex problem, that is, the estimation of an unknown number of peaks per pixel, it possesses a surface detection capability which is of interest in our study. The parameters of RT3D have been tuned via cross-validation by optimizing the visual quality of the reconstruction. In particular, although the data consists only of 32×32 pixels, RT3D is set so that it reconstructs point clouds with 96×96 pixels in the transverse direction. Note that we also applied the pre-trained implementation (provided by the authors) of the deep-learning method proposed in [14].
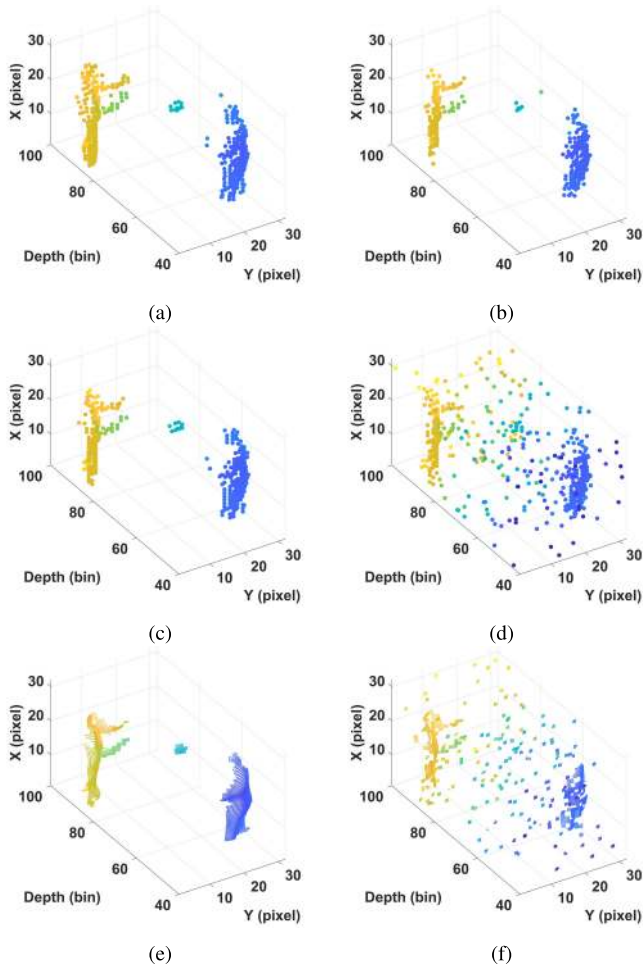
Fig. 4. Examples of 3D reconstruction using the proposed method ((a),(b)), BF ((c),(d)) and RT3D (e),(f). The left-hand plots correspond to actual measurements while the right-hand plots have been obtained from the pseudo-synthetic data generated with additional background ($b = 50$). The colormap represents the surface depth (axial direction).

The results obtained with this first dataset are similar to those obtained by other competing methods. However, this method performed worse than the other methods when considering noisier measurements (discussed next). For this reason, we did not include this method for comparison in this study. However, these first results could potentially be improved by retraining the original network in future work.

Fig. 4 (a), (c) and (e) depict an example of reconstruction (Frame #1700) using R3DSP, BF and RT3D, respectively. In this frame, the two pedestrians are in the field on view and the ball is roughly at the midpoint between them. In this scenario, the three methods provide similar results, as the SBR and MSC are high enough to allow a clear identification of the empty pixels and a satisfactory estimation of the surface depth in the other pixels. Note that the surfaces appear smoother using RT3D, partly because of the spatial smoothing involved but also because the corresponding point cloud has about 9 times more points than those of R3DSP and BF. For completeness, the point clouds and background levels estimated by the three methods for the whole sequence are presented in Video 2. For visualization purposes, the video

is played at actual speed, with 50 frames per second (the intermediate frames processed by R3DSP are not displayed). This video also compares the estimated background levels, which are consistent across the three methods and it presents the estimated surface presence maps. The proposed method is able to more efficiently detect the head of the pedestrian on the left-hand side, which presents a low reflectivity due to the wavelength used (1550 nm).

These results are used as reference to investigate more challenging scenarios, with lower SBRs. More precisely, we generated additional pseudo-synthetic datasets by artificially adding constant background levels to all the pixels of each sequence. We created two sequences using background levels of $b = 20$ and $b = 50$, leading to approximately 3060 and 7650 additional background photons per pixel and per frame. The resulting SBRs are 1.8×10^{-2} and 7.2×10^{-3}, respectively. Examples of reconstructed point clouds in the lowest SBR regime ($b = 50$) are depicted in Fig. 4 (b), (d) and (f). When the SBR decreases, BF and RT3D, which are based on intensity thresholding, generally present higher false alarm rates than R3DSP which incorporates a Bayesian test for object detection. If the thresholds of BF and RT3D are increased, the corresponding probabilities of detection decrease. Although R3DSP misses some surfaces (see Fig. 4 (a) and (b)), it is able to reconstruct the ball, using both the tailored detection strategy and its ability to promote correlation between successive frames via the spatio-temporal model. For completeness, the sequences reconstructed by the three methods for $b = 20$ and $b = 50$ are presented in Video 3 and Video 4, respectively. These videos shows that although the three methods provide similar background estimates, R3DSP consistently yields better reconstructions (visually lower false alarm rates and higher probabilities of detection). These results confirm that in the presence of significant background levels (low SBR) the proposed method is able to detect and track dynamic surfaces more efficiently.

The second experiment was conducted under the same observation conditions and we recorded the movements of a pedestrian running back and forth at about 320 m from the detector. For this scene, we integrated the binary acquisitions into 1000 lidar frames per second. In each pixel and frame, we observe approximately 18 photons related to dark counts or ambient illumination from solar background, and the visible surfaces lead to 27 additional photons per pixel, on average. The proposed algorithm has been applied to a series of 6000 successive frames, representing a total acquisition of 6 s. As before, we set $\beta = 0.5$, $M = 5$ neighbors, $\sigma_{\text{RW}} = \sqrt{3}$ bins and $\nu_0 = 0.5$. The reconstructed point clouds, together with a standard video of the scene are presented in the Video 5. Fig. 5 depicts the temporal profile of the depth estimated at a central pixel where the pedestrian is always visible. This figure shows the mean and credible interval (± 6 standard deviation interval for visualization purposes) of the depth posterior distribution. As expected, the depth uncertainty is larger at the beginning of the sequence due to the limit amount of information available about the object range. After about 300 frames (300 ms), the uncertainty becomes more stable and the algorithm is able to successfully track the position of
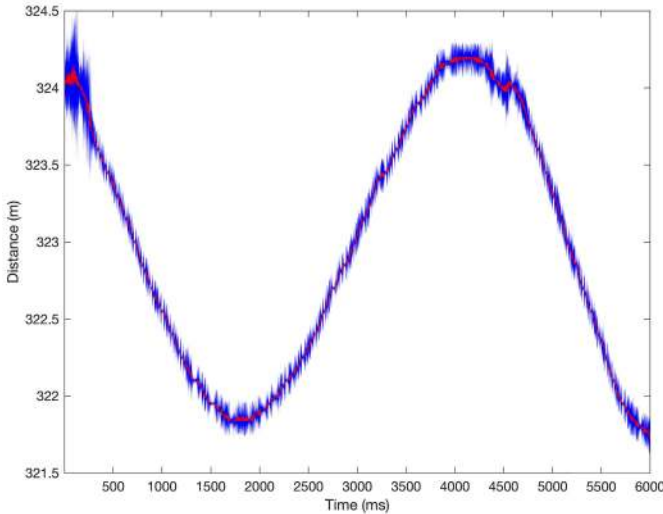
Fig. 5. Temporal evolution of the estimated depth in pixel (16,16) of the second experiment (running pedestrian). The red curve represents the mean of the depth posterior distribution obtained for each frame and the blue region represents the ± 6 standard deviation credible interval.
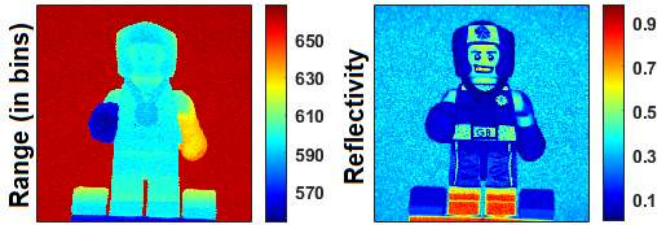

Fig. 6. Depth (left) and reflectivity (right) profiles used to simulate the dynamic scene.

the surface. Note that such depth uncertainty measures can be used for instance to quantify uncertainties associated with the instantaneous velocity or moving objects.

Although the resolution of the 32×32 Princeton Lightwave Kestrel camera used in this study limits the resolution of the reconstructed point cloud, the proposed method can be applied to larger detector arrays without significant computational degradation since most computational steps can be performed at the pixel level, using only local information. Nonetheless, coarse depth maps, as obtained here, can be used in diverse applications involving object detection or recognition [47], [48]. Higher-resolution depth images can also be obtained by resorting to deep learning approaches such as in [14], [49], [50], where associated high resolution color maps are used as guidance to improve the details of the resulting fine depth map.

A last experiment was performed using synthetic data to demonstrate that the proposed method can be applied with larger SPAD arrays. The ground truth profiles used to simulate the synthetic data consists of 200×200 pixels depth and reflectivity profiles obtained from real single-photon lidar data acquired at 532nm in [51], with a target (a figurine) located at a distance of about 1.8 m from the imaging system. A dynamic scene of 312 temporal frames was created from these profiles, such that the figurine enters in the field of view from the left with a horizontal shift of 1 pixel per frame (for the first 171 frames). Once the figurine reaches the center of the image, a 45

pixel-radius disk appears on the right hand-side of the image and is shifted to the left until it reaches the center of the image. Each pixel of the disk has a depth corresponding to 570 bins and a reflectivity of 0.8. Histograms of 1500 temporal bins (bin width of 2ps) were generated from (4) using a Gaussian IRF with a FWHM of 31 bins (same as that measured in [51]), including a spatially constant background level, such that $MSC = 35$ signal photons were recorded on average per pixel within each frame, with $SBR \approx 1.5$. The proposed method has been applied using $M = 5$, $\sigma_{\mathrm{RW}} = \sqrt{3}$ bins, $v_0 = 0.5$ and $\beta = 0.4$, even though using $\beta \in [0.4, 0.6]$ did not change significantly the results. For visual comparison, a log-matched filtering depth estimation was also performed. Estimation of the background level was then conduced by averaging over the temporal bins outside the estimated peak (using the 3-sigma rule of thumb). Those estimates were finally plugged in (4) to estimate the reflectivity profile by MLE. The ground truth depth and reflectivity profiles, together with the estimated depth and reflectivity using both log-matched filtering and the proposed online approach are presented in the Video 6. This video illustrates how the proposed spatio-temporal model regularizes the depth estimation process, while allowing new objects to enter the field of view.

V. CONCLUSION

In this work, we presented, to the best of our knowledge, a first algorithm for sequential reconstruction of dynamic 3D scenes from SPLs data, using temporal correlation. This method primarily focuses on the estimation of the surface depth using a model that does not involve the target reflectivity nor the background level. The resulting depth estimation process is particularly efficient, as is reduces to computing, pixel-wise, the cross-correlation (either discrete or continuous) between the measured photon ToAs and a modified IRF, which represents a fixed and predictable computational cost. Moreover, if the system IRF is Gaussian, the update rules are greatly simplified as the mean and variances of the pseudo-posterior distributions can be obtained in closed-form. Thanks to the proposed spatio-temporal model, most of the steps of the algorithm can be performed independently for each pixel, which is attractive for parallel/distributed implementation. In this work, we focused of the theoretical development of the method but did not fully optimize its implementation, e.g., on GPUs, mainly because the current implementation of the detection method [33] is not yet optimized. This is left for future work. We also derived the β-divergence for MSL data and the method proposed here could also be extended for online reconstruction of colored point clouds, provided that an efficient detection strategy (adapted to MSL data) is used.

APPENDIX

In this appendix, we derive the expression of the β-divergence for the estimation of the depth d in a single pixel and frame, assuming that multispectral single-photon lidar data (with L wavelengths) are recorded simultaneously. We denote by y_ℓ the time of arrival of a photon at the ℓth wavelength. Using the classical MSL systems, separate and
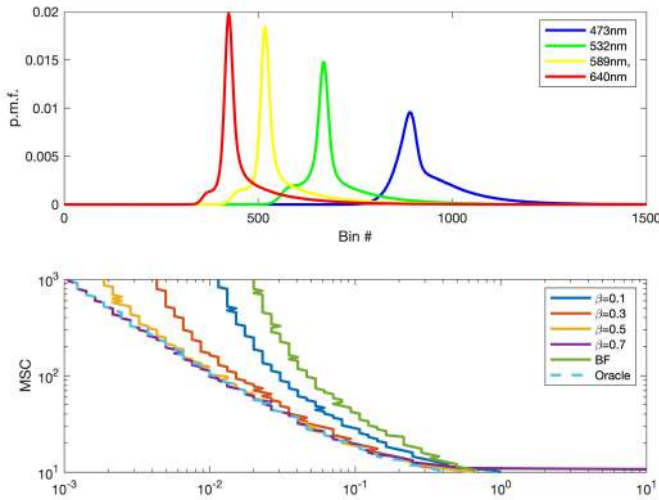
Fig. 7. Top: IRFs from [38] used to simulate MSL synthetic data. Bottom: curves of satisfactory detection ($p_d = 85\%$, with $\eta = 28$) as a function of (SBR,MSC) using the real IRF. This figure compares only (pseudo-)MLEs adapted to MSL data.

independent detectors are used such that the detection events in the L channels are mutually independent, conditioned on the configuration of the scene. Thus, for any set of L random variables $(y_1, \ldots, y_L)$ associated with each of the L bands, the ideal model in (1) can be extended as

$$f_0(y_1, \ldots, y_L|d) = \prod_{\ell=1}^{L} f_\ell(y_\ell|d) = \prod_{\ell=1}^{L} s_\ell \left(y_\ell - \frac{2d}{c} \right), \quad (21)$$

where $s_\ell(\cdot)$ is the impulse response of the ℓth band [52]. For multivariate continuous distributions the β-divergence has the same expression as in (6) and it can be easily shown in a similar fashion to (7) that, under the same mild conditions as in Section II,

$$D_\beta(\hat{f}(y_1, \ldots, y_L) || f_0(y_1, \ldots, y_L|d))$$

$$= \text{Const.} - \frac{\beta+1}{\beta} \prod_{\ell=1}^{L} \frac{1}{K_\ell} \sum_k f_\ell(y_{\ell,k}|d)^\beta, \quad (22)$$

where $\hat{f}(y_1, \ldots, y_L)$ is the product of the empirical marginal distributions of the ToAs in the L bands, i.e., $\hat{f}(y_1, \ldots, y_L) = \prod_{\ell=1}^{L} \hat{f}(y_\ell)$ and $\hat{f}(y_\ell) = \frac{1}{K_\ell} \sum_{k=1}^{K_\ell} \delta(y_\ell - y_{\ell,k})$, with $\{y_{\ell,k}\}_k$ the ToAs of the K_ℓ photons detected at the ℓth wavelength. Note that in contrast to classical maximum likelihood estimation which would introduce a sum (of the log-likelihood terms) over the L bands, in (22) we obtain a product over the L bands and sums over the detection events.

To demonstrate the benefits of the proposed robust method for MSL-based depth imaging, we generated MSL data using the $L = 4$ IRFs depicted in Fig. 7 (top), using the same SBR and same MSC for all the bands. In contrast to the results presented in Fig. 2, for simplicity we only consider estimators using only the observations, without regularization or additional prior information, i.e., pseudo-MLEs. More precisely, we compared the minimum divergence estimator derived from (22) to the BF MLE (assuming no background) for MSL data and the Oracle estimator assuming the reflectivity and

background in each band is known. The limit of the regions of successful depth estimation ($p_d > 85\%$ with $\eta = 28$) are depicted in Fig. 7 (bottom). As in the single-wavelength case, increasing β allows satisfactory depth estimation at lower SBRs than using BF and using $\beta = 0.7$ here leads to results similar to those of the Oracle, without requiring knowledge or estimation of the reflectivity and background parameters. Note also that, as in the single-band case, too large values of β (see $\beta = 0.7$) lead to poor results in the low MSC and high SBR regime.

REFERENCES

[1] C. Mallet and F. Bretar, "Full-waveform topographic Lidar: State-of-the-art," *ISPRS J. Photogramm. Remote Sens.*, vol. 64, no. 1, pp. 1–16, Jan. 2009.

[2] M. A. Canuto *et al.*, "Ancient lowland maya complexity as revealed by airborne laser scanning of northern guatemala," *Science*, vol. 361, no. 6409, Sep. 2018, Art. no. eaau0137.

[3] J. Hecht, "Lidar for self-driving cars," *Opt. Photon. News*, vol. 29, no. 1, pp. 26–33, 2018.

[4] J. Gao, J. Sun, J. Wei, and Q. Wang, "Research of underwater target detection using a slit streak tube imaging Lidar," in *Proc. Academic Int. Symp. Optoelectron. Microelectron. Technol.*, Harbin, China, Oct. 2011, pp. 240–243.

[5] R. Tobin, A. Halimi, A. McCarthy, M. Laurenzis, F. Christnacher, and G. S. Buller, "Three-dimensional single-photon imaging through obscurants," *Opt. Exp.*, vol. 27, no. 4, pp. 4590–4611, Feb. 2019.

[6] A. M. Pawlikowska, A. Halimi, R. A. Lamb, and G. S. Buller, "Single-photon three-dimensional imaging at up to 10 kilometers range," *Opt. Exp.*, vol. 25, no. 10, pp. 11919–11931, May 2017.

[7] Z.-P. Li *et al.*, "Single-photon computational 3D imaging at 45 km," 2019, *arXiv:1904.10341*. [Online]. Available: http://arxiv.org/abs/1904.10341

[8] Z.-P. Li *et al.*, "Super-resolution single-photon imaging at 8.2 kilometers," *Opt. Exp.*, vol. 28, no. 3, pp. 4076–4087, 2020.

[9] A. Maccarone *et al.*, "Underwater depth imaging using time-correlated single-photon counting," *Opt. Exp.*, vol. 23, no. 26, pp. 33911–33926, Dec. 2015.

[10] A. McCarthy, R. J. Collins, N. J. Krichel, V. Fernández, A. M. Wallace, and G. S. Buller, "Long-range time-of-flight scanning sensor based on high-speed time-correlated single-photon counting," *Appl. Opt.*, vol. 48, no. 32, pp. 6241–6251, Nov. 2009.

[11] D. Shin, A. Kirmani, V. K. Goyal, and J. H. Shapiro, "Photon-efficient computational 3-D and reflectivity imaging with single-photon detectors," *IEEE Trans. Comput. Imag.*, vol. 1, no. 2, pp. 112–125, Jun. 2015.

[12] D. Shin *et al.*, "Photon-efficient imaging with a single-photon camera," *Nature Commun.*, vol. 7, no. 1, pp. 1–8, Nov. 2016.

[13] Y. Altmann and S. McLaughlin, "Range estimation from single-photon lidar data using a stochastic em approach," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 1112–1116.

[14] D. B. Lindell, M. O'Toole, and G. Wetzstein, "Single-photon 3D imaging with deep sensor fusion," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–12, Aug. 2018.

[15] J. Rapp and V. K. Goyal, "A few photons among many: Unmixing signal and noise for photon-efficient active imaging," *IEEE Trans. Comput. Imag.*, vol. 3, no. 3, pp. 445–459, Sep. 2017.

[16] J. Tachella *et al.*, "Bayesian 3D reconstruction of complex scenes from single-photon Lidar data," *SIAM J. Imag. Sci.*, vol. 12, no. 1, pp. 521–550, Jan. 2019.

[17] J. Rapp, Y. Ma, R. M. A. Dawson, and V. K. Goyal, "Dead time compensation for high-flux ranging," 2018, *arXiv:1810.11145*. [Online]. Available: https://arxiv.org/abs/1810.11145

[18] S. Chan *et al.*, "Long-range depth imaging using a single-photon detector array and non-local data fusion," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.

[19] D. Shin, F. Xu, F. N. Wong, J. H. Shapiro, and V. K. Goyal, "Computational multi-depth single-photon imaging," *Opt. Exp.*, vol. 24, no. 3, pp. 1873–1888, 2016.

[20] R. Tobin *et al.*, "Long-range depth profiling of camouflaged targets using single-photon detection," *Opt. Eng.*, vol. 57, no. 3, 2017, Art. no. 031303.

[21] A. Halimi, R. Tobin, A. McCarthy, S. McLaughlin, and G. S. Buller, "Restoration of multilayered single-photon 3D Lidar images," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 708–712.

[22] J. Tachella, Y. Altmann, S. McLaughlin, and J.-Y. Tourneret, "3D reconstruction using single-photon Lidar data exploiting the widths of the returns," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7815–7819.

[23] J. Tachella *et al.*, "Real-time 3D reconstruction from single-photon Lidar data using plug-and-play point cloud denoisers," *Nature Commun.*, vol. 10, no. 1, p. 4984, Dec. 2019.

[24] Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, and S. McLaughlin, "Lidar waveform-based analysis of depth images constructed using sparse single-photon data," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 1935–1946, May 2016.

[25] Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, and S. McLaughlin, "Target detection for depth imaging using sparse single-photon data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 3256–3260.

[26] A. Halimi *et al.*, "Restoration of intensity and depth images constructed using sparse single-photon data," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Budapest, Hungary, Aug. 2016, pp. 86–90.

[27] A. Halimi, R. Tobin, A. McCarthy, J. Bioucas-Dias, S. McLaughlin, and G. S. Buller, "Robust restoration of sparse multidimensional single-photon LiDAR images," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 138–152, 2020.

[28] X. Ren *et al.*, "High-resolution depth profiling using a range-gated CMOS SPAD quanta image sensor," *Opt. Exp.*, vol. 26, no. 5, pp. 5541–5557, Mar. 2018.

[29] R. K. Henderson *et al.*, "A 192×128 time correlated SPAD image sensor in 40-nm CMOS technology," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 1907–1916, Jul. 2019.

[30] Y. Altmann, S. McLaughlin, and M. E. Davies, "Fast online 3D reconstruction of dynamic scenes from individual single-photon detection events," *IEEE Trans. Image Process.*, vol. 29, pp. 2666–2675, 2020.

[31] A. Halimi, A. Maccarone, A. McCarthy, S. McLaughlin, and G. S. Buller, "Object depth profile and reflectivity restoration from sparse single-photon data acquired in underwater environments," *IEEE Trans. Comput. Imag.*, vol. 3, no. 3, pp. 472–484, Sep. 2017.

[32] A. Maccarone, F. M. D. Rocca, A. McCarthy, R. Henderson, and G. S. Buller, "Three-dimensional imaging of stationary and moving targets in turbid underwater environments using a single-photon detector array," *Opt. Exp.*, vol. 27, no. 20, pp. 28437–28456, Sep. 2019.

[33] J. Tachella, Y. Altmann, S. McLaughlin, and J.-Y. Tourneret, "Fast surface detection in single-photon lidar waveforms," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–5.

[34] Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, and S. McLaughlin, "Robust Bayesian target detection algorithm for depth imaging from sparse single-photon data," *IEEE Trans. Comput. Imag.*, vol. 2, no. 4, pp. 456–467, Dec. 2016.

[35] A. Basu, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, Sep. 1998.

[36] F. Futami, I. Sato, and M. Sugiyama, "Variational inference based on robust divergences," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 84, A. J. Storkey and F. Pérez-Cruz, Eds. Canary Islands, Spain: PMLR, Apr. 2018, pp. 813–822.

[37] A. McCarthy *et al.*, "Kilometer-range depth imaging at 1550 nm wavelength using an InGaAs/InP single-photon avalanche diode detector," *Opt. Exp.*, vol. 21, no. 19, pp. 22098–22113, Sep. 2013.

[38] R. Tobin *et al.*, "Comparative study of sampling strategies for sparse photon multispectral lidar imaging: Towards mosaic filter arrays," *J. Opt.*, vol. 19, no. 9, Aug. 2017, Art. no. 094006.

[39] X. Ren, Y. Altmann, R. Tobin, A. Mccarthy, S. Mclaughlin, and G. S. Buller, "Wavelength-time coding for multispectral 3D imaging using single-photon LiDAR," *Opt. Exp.*, vol. 26, no. 23, pp. 30146–30161, 2018.

[40] D. R. Bickel and R. Frühwirth, "On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications," *Comput. Statist. Data Anal.*, vol. 50, no. 12, pp. 3500–3530, Aug. 2006.

[41] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964.

[42] R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing* (Statistical Modeling and Decision Science). Amsterdam, The Netherlands: Elsevier, 2012.

[43] J. Tachella, Y. Altmann, S. McLaughlin, and J.-Y. Tourneret, "On fast object detection using single-photon Lidar data," in *Wavelets Sparsity XVIII*, vol. 11138, D. V. D. Ville, M. Papadakis, and Y. M. Lu, Eds. Bellingham, WA, USA: SPIE, 2019, pp. 252–261.

[44] S. L. Lauritzen, "Propagation of probabilities, means, and variances in mixed graphical association models," *J. Amer. Stat. Assoc.*, vol. 87, no. 420, pp. 1098–1108, Dec. 1992.

[45] X. Boyen and D. Koller, "Tractable inference for complex stochastic processes," in *Proc. 14th Conf. Uncertainty Artif. Intell. (UAI)*. San Francisco, CA, USA: Morgan Kaufmann, 1998, pp. 33–42.

[46] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. 17th Conf. Uncertainty Artif. Intell.* San Mateo, CA, USA: Morgan Kaufmann, 2001, pp. 362–369.

[47] S. Song and J. Xiao, "Deep sliding shapes for amodal 3D object detection in RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 634–651.

[48] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 821–826.

[49] M. Ni, J. Lei, R. Cong, K. Zheng, B. Peng, and X. Fan, "Color-guided depth map super resolution using convolutional neural network," *IEEE Access*, vol. 5, pp. 26666–26672, 2017.

[50] Y. Wen, B. Sheng, P. Li, W. Lin, and D. D. Feng, "Deep color guided Coarse-to-Fine convolutional network cascade for depth image super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 994–1006, Feb. 2019.

[51] R. Tobin *et al.*, "Comparative study of sampling strategies for sparse photon multispectral lidar imaging: Towards mosaic filter arrays," *J. Opt.*, vol. 19, no. 9, Aug. 2017, Art. no. 094006.

[52] Y. Altmann *et al.*, "Robust spectral unmixing of sparse multispectral Lidar waveforms using gamma Markov random fields," *IEEE Trans. Comput. Imag.*, vol. 3, no. 4, pp. 658–670, Dec. 2017.