Online Temporally Consistent Indoor Depth Video Enhancement via Static Structure

Lu Sheng, *Student Member, IEEE*, King Ngi Ngan, *Fellow, IEEE*,
Chern-Loon Lim, *Member, IEEE*, and Songnan Li, *Member, IEEE*

***Abstract*— In this paper, we propose a new method to online enhance the quality of a depth video based on the intermediary of a so-called static structure of the captured scene. The static and dynamic regions of the input depth frame are robustly separated by a layer assignment procedure, in which the dynamic part stays in the front while the static part fits and helps to update this structure by a novel online variational generative model with added spatial refinement. The dynamic content is enhanced spatially while the static region is otherwise substituted by the updated static structure so as to favor the long-range spatio-temporal enhancement. The proposed method both performs long-range temporal consistency on the static region and keeps necessary depth variations in the dynamic content. Thus, it can produce flicker-free and spatially optimized depth videos with reduced motion blur and depth distortion. Our experimental results reveal that the proposed method is effective in both static and dynamic indoor scenes and is compatible with depth videos captured by Kinect and time-of-flight camera. We also demonstrate that excellent performance can be achieved by the proposed method in comparison with the existing spatio-temporal approaches. In addition, our enhanced depth videos and static structures can act as effective cues to improve various applications, including depth-aided background subtraction and novel view synthesis, showing satisfactory results with few visual artifacts.**

***Index Terms*— Static structure, temporally consistent depth video enhancement, online estimation, layer assignment.**

I. Introduction

ACQUIRING high-quality and well-defined depth data from real scenes has been a key problem in computer vision with the prevalence of various 3D applications in manufacturing and the entertainment industry, in uses that include virtual reality, 3DTV and free-viewpoint TV, game controller and robot vision. Recently a variety of systems have been proposed to obtain depth information of a real scene, from passive stereo vision system to active sensors like real-time structured-light depth sensors (*e.g.*, Kinect), Time-of-Flight (ToF) cameras or laser scanners. Unfortunately most systems suffer from low quality of the acquired depth maps, typically in terms of low resolution, noise and outliers, and missing depth regions (or holes) without depth measurements. These shortcomings obstruct the direct usage of depth information of captured scenes for different 3D applications.

Even though spatial enhancement of depth maps has been extensively studied in recent years, in area such as energy minimization methods [1]–[3] or filtering methods based on high-dimensional filtering [4]–[6], as well as other methods like patch matching [7], [8] and so on, the temporal inconsistency problem is nevertheless neglected since the necessary temporal relationship between adjacent frames has not been taken into consideration, thus severe flickering artifacts become an urgent issue to tackle. However, due to various complex and even unpredictable dynamic contents, as well as outliers in a depth video, it is not easy to exactly locate the regions where temporal consistency should be enforced. Several existing methods [9], [10] employ the temporal texture similarity to extract 2D motion information, but correct depth variation cannot always be maintained thus causing severe motion blur. In addition, typical treatments always apply temporal consistency over a short-length sequence (usually 2∼3 frames), which is otherwise insufficient to generate stable and temporally consistent results over *hundreds* of frames. Furthermore, over-smoothing around the boundaries between dynamic objects and static scenes should be eliminated to produce high quality and well-defined depth video.

In this paper, we present an alternative method to enhance a depth video both spatially and temporally by addressing two aspects of these problems: 1) efficiently and effectively enforcing the temporal consistency where it is necessary, and 2) enabling online processing. A common fact is that regions in one frame with various motion patterns (*e.g.*, static, slowly/fast moving and etc.) belong to different objects or structures and require temporal consistencies with different levels. For instance, the static region needs a long-range temporal enhancement to ensure that it is static over a long duration, while dynamic regions with slow/rapid motions expect short-term or no temporal consistency. However, it is difficult to accurately enhance arbitrary and complex dynamic contents in the temporal domain without apparent motion blurs or depth distortions. Thus we propose an intuitive compromise to cancel the temporal enhancement in the dynamic region as long as its spatial enhancement is sufficiently satisfactory,

Manuscript received July 4, 2014; revised December 9, 2014 and March 12, 2015; accepted March 12, 2015. Date of publication March 25, 2015; date of current version April 14, 2015. This work was supported by the University of Malaya, Kuala Lumpur, Malaysia, under Project UM.C/625/1/HIR/MOHE/ENG/42. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Brendt Wohlberg.

L. Sheng, K. N. Ngan, and S. Li are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: lsheng@ee.cuhk.edu.hk; knngan@ee.cuhk.edu.hk; snli@ee.cuhk.edu.hk).

C.-L. Lim is with the University of Malaya, Kuala Lumpur 50603, Malaysia (e-mail: limchernloon@gmail.com).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The video file shows four RGB-D test videos and compares their enhanced results generated by the proposed method and three references. The total size of the video is 23.7 MB. Contact lsheng@ee.cuhk.edu.hk for further questions about this work.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2416658

in which the necessary depth variation will not be distorted while the temporal artifacts are not as easy as those in the static region to be perceived. Therefore, we aim at strengthening long-range temporal consistency around the static region whilst maintaining necessary depth variation in the dynamic content. To accurately separate the static and dynamic regions, we *online* track and incrementally refine a probabilistic model called *static structure*, which acts as a medium to indicate the region that is static in the current frame. By online fusing the static region of the current frame into the static structure with an efficient variational fusion scheme, this structure has implicitly gathered all the temporal data at and before the current frame that belong to it. Substituting the static region by the updated static structure, it is thus temporally consistent and stable in a long range accordingly. Moreover, it is also suitable for online processing the streaming depth videos (3D teleconference, 3DTV and etc.) without the necessity of storing amounts of adjacent frames, thus is memory and computationally efficient.

Overall, the temporally consistent depth video enhancement is performed at two layers: 1) the static region of the input frame revealing the static structure is enhanced spatially and temporally by an online fusion technique combining it with the static structure, and 2) the dynamic content is enhanced spatially without temporal smoothness. In addition to the advantages stated aforementioned, enhancing the static and dynamic regions separately also effectively eliminates artifacts that frequently occur in conventional depth video enhancements, like the blurring artifacts or the unreliable depth propagation, across the boundaries between dynamic objects and static objects/background. Furthermore, when the depth video contains severe holes, the static structure can fill static holes convincingly and leave the rest holes filled by the dynamic content so as to avoid the inpainting artifacts. Since fully dynamic depth videos usually have weak temporal consistency thus our proposed algorithm is relegated to a spatial enhancement approach, which does not force the enhanced depth video to bear unnecessary temporal smoothness.

Based on the conference version [11] of this work, more technical details and theoretic analyses about the formulation of the static structure, effective layer assignment as well as a sound spatial enhancement of the static structure are discussed in this paper. Furthermore, a complete framework about temporally consistent depth video enhancement, a thorough experimental evaluation as well as discussions about its applications and limitations are also provided.

The rest of the paper is organized as follows. Section II reviews existing works in spatial and temporal depth video enhancement, as well as approaches on static scene reconstruction, which is indeed related to our formulation of the static structure. Section III describes our proposed framework of online estimation of the static structure and the approach regarding temporally consistent depth video enhancement. Experimental results and discussions of our method can be found in Section IV. Discussions about its limitations and applications are presented in Section V. Concluding remarks and discussions on future work are given in Section VI.

II. Related Work

A. Spatial Enhancement

On the aspect of global optimization, the pioneering work was done by Diebel and Thrun [1] utilizing the pixel-wise MRF model with the guidance of texture to denoise the depth map. Several augmented models were also proposed to handle inpainting and super-resolution [2], [3], [12]–[14], with special choices of the data and smoothness terms as well as additional regularization terms (TV-ℓ^1 norm [14], etc.), enabling a reasonable performance even without texture information [14]. But the high computational cost of these methods hinders real-time applications. Another choice is high-dimensional filtering. One variant is high-dimensional average filtering [4]–[6], [9], [15], whose weights are defined by the spatial nearness and feature proximity. The feature can be texture/depth intensities or patches [6], [16] and other user-defined ones. The main problems here are edge blurring and texture mapping. Another variant uses the median of the depth candidate histogram instead [17], [18], producing more robust results but also suffering from quantization error and slower speed. Weighted mode filtering [10], [19] otherwise looks for the histogram's global mode, and has similar artifacts. In addition, spatial enhancement, especially super-resolution and inpainting, can be performed by patch matching throughout the depth map, which achieved satisfactory visual results [7], [8] but with high computational complexity.

B. Temporal Enhancement

Existing temporal enhancement approaches usually employ the guidance of temporal texture consistency, especially by fusing the previous depth frame onto the current one according to the motion vectors estimated between the corresponding adjacent color frames [9], [10]. However, the neglect of additional motion vectors in z-axis reduces the warping accuracy. 3D motion estimation is typically adopted to solve the problem in [20]–[22]. Following them, the temporal fusion between current and warped previous frames are usually based on weighted average or weighted median filters, and energy minimization as well [9], [10], [23], [24]. Therefore the performance, on one hand, relies heavily on the accuracy of motion estimation, which is difficult to be satisfied. On the other hand, the temporal continuity is only preserved among few adjacent frames, which does not meet the demand of constraining long-range temporal consistency. To fix such an issue, Lang *et al.* [25] proposed to offline filter the paths which are the vectors of all the pixels that correspond to the motion of one scene point over time. It provides a practical and remarkable solution to enhance a depth video with long-range temporal consistency both effectively and efficiently. Our work is related to, but has essential differences from the layer denoising and completion proposed by Shen and Cheung [26], which offline trained background layer models beforehand to label the foreground and background of the input depth frame, and no temporal consistency was strengthened. Conversely, our method estimates the static structure in an online fashion and there is no need to have a series of depth frames capturing

purely static scenes. Moreover, the temporal consistency is maintained where it is required. That aside, only the spatial enhancement is taken into consideration as presented in [26].

C. Static Scene Reconstruction

The static structure estimation is related to the static scene reconstruction by fusing a series of depth maps. A majority of these works are offline methods [27]–[31] which fuse a set of depth maps to output a single geometric structure, while the rest are online approaches that receive depth measurements sequentially and incrementally estimate the current geometric structure. Offline methods always extract a batch of depth frames together so that the complexity becomes unbearable when the number of frames is large. One of the offline approaches by Zitnick *et al.* [31] employed the consistency of both the multiple view colors and disparities, which is analogous to our constraint of temporal consistency, to regularize the disparity space distribution so as to bring about the refined disparity map. Most online methods quantize the 3D space into grids [32]–[35] to reduce the memory and computational cost. Thus they are always deficient in sub-grid accuracy, but one additional approach exploits a weighted sum of truncated signed distance function (TSDF) [33], [34] over depth measurements. However it is sensitive to outliers and thus not robust to estimate a static scene containing dynamic objects and heavy outliers. To robustly estimate the static scene captured by noisy and cluttered data, some researchers have proposed a variety of measurement models with parameters describing the nature of the noise and outliers. Several methods [32], [36] need parameters learned from ground truth data or those tuned empirically. One successful model that requires fewer manually tuned parameters is the generative model, which has the ability to derive the model of the noise and clutter characteristics from the input data. Vogiatzis and Hernández [37] proposed a generative Gaussian plus uniform model that simultaneously infers the depth and outlier ratio per pixel using an efficient online variational scheme, which meets the clutter characteristics of depth maps generated by stereo. Our static structure estimation is similar as an online generative model considering both noise and outliers as well as a special treatment of the dynamic scenes.

III. Approach

The static structure can be regarded as an intrinsic depth structure (and texture structure when the registered color video is available) underneath the captured scene,[1] which always lies on or behind the surface of the input depth frame. As shown in Figure 1, any moving or foreground object stays in front of the static structure whereas the static objects or visible static background are usually on it, *i.e.*, the depth value of the static structure at one pixel is always deeper than that of a dynamic object at the same place. But it is different from the "background" of a scene, because we focus more on the "static" geometric structure rather than the distance from the camera. Since the temporal consistency around static or slowly moving regions are required to be enforced, the "static" nature is more useful than the idea of "background".

[1]Within the scope of this paper, we assume the target depth video is captured by a static depth sensor hence the captured scene is static except the dynamic objects. Although the enhancement of depth video captured by moving cameras is a more general topic, we will refer it to our future work.

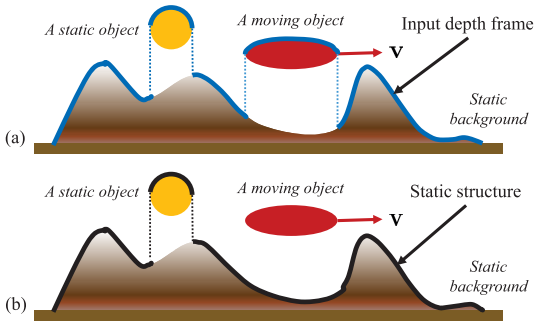


Fig. 1. The illustration of the static structure in comparison with the input depth frame. (a) shows the input depth frame (in blue curve) lies on the captured scene, (b) represents the static structure (in black curve). The depth sensor is above the captured scene. The static structure includes the static objects as well as the static background.

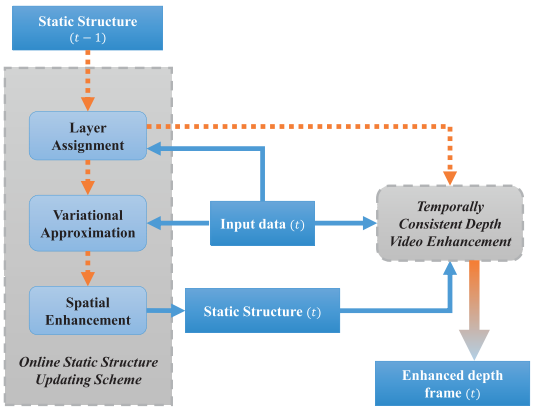


Fig. 2. Flowchart of the overall framework of the proposed method on the estimation of static structure and depth video enhancement. Please refer to the text for the detailed description.

To handle artifacts like noise, outliers and holes as well as complex dynamic contents in the input depth frame, we propose a probabilistic generative mixture model to describe the static structure as well as the characteristics of noise and outliers (Section III-A). We also define an efficient layer assignment leveraging dense conditional random fields to accurately label input depth frame into dynamic and static regions (Section III-D). For the sake of memory and calculation efficiency, as well as the ability to process streaming data, the static structure is *online* updated (Section III-E) via a variational approximation (Section III-B) governed by a first order Markov chain, which effectively fuses the labeled static region in the current depth frame with the previous estimated structure. It is further refined spatially to fill holes and regularize the structure (Section III-E). The updated static structure in turn substitutes the static region of the input depth frame, resulting in a temporally consistent depth video enhancement (Section III-F). The framework of the online static structure update scheme and temporally consistent depth video enhancement is referred to in the flowchart in Figure 2.

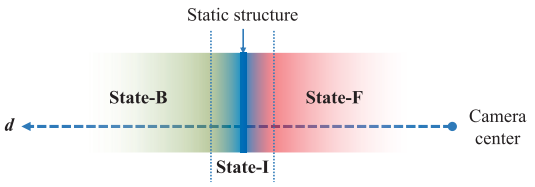


Fig. 3. Illustration of three states of input depth measurements with respect to the static structure on one line-of-sight. The current static structure refers to the blue stick in the middle. Decision boundaries are marked as blue dot lines. The depth measurement d is categorized into **state-I** when it is placed around the static structure. When d is in front of this structure, we denote it as **state-F**. While it is far behind the static structure, the state is **state-B**.

Notation: The data sequence is denoted as $\mathcal{S}$ and formed by a depth video $\mathcal{D} = \{\mathbf{D}^t | t = 1, 2, \ldots, T\}$ as $\mathcal{S} = \mathcal{D}$, or as a pair of aligned depth plus color videos as $\mathcal{S} = \{\mathcal{D}, \mathcal{I}\}$, where $\mathcal{I} = \{\mathbf{I}^t | t = 1, 2, \ldots, T\}$. The data in each frame is $\mathbf{S}^t = \mathbf{D}^t$ or $\{\mathbf{D}^t, \mathbf{I}^t\}$. The pixel location is defined as $\mathbf{x}$, and its depth value at t is $d_{\mathbf{x}}^t$ and its corresponding color is $\mathbf{I}_{\mathbf{x}}^t$. The parameter set for the probabilistic model at each frame t is denoted as $\mathcal{P}^{\mathcal{S},t}$, and $\mathcal{P}_{\mathbf{x}}^{\mathcal{S},t}$ is defined for each pixel $\mathbf{x}$, whose elements are defined in detail in the following sections.

A. A Probabilistic Generative Mixture Model

At the very beginning, we only consider the case where $\mathcal{S} = \mathcal{D}$. Denote the sequentially incoming depth samples of pixel $\mathbf{x}$ on and before time t as forming a set $\mathcal{D}_{\mathbf{x}}^t = \{d_{\mathbf{x}}^\tau | \tau = 1, 2, \ldots, t\}$. The depth value of the static structure in the pixel $\mathbf{x}$ is $Z_{\mathbf{x}}$, whose noise is conveniently governed by a Gaussian distribution. We also propose two individual outlier distributions to describe the outliers before and after the static structure respectively. Hence, they do not only describe the depth distribution but also provide evidence to indicate the state to which the current depth sample belongs.

1) State Description: The three states $\Psi = \{I, F, B\}$ are illustrated in Figure 3 and listed as follows.

a) State-I (fitting the static structure): If $d_{\mathbf{x}}^t$ belongs to the static structure, we assume that it follows a Gaussian distribution centered at $Z_{\mathbf{x}}$ as $\mathcal{N}\left(d_{\mathbf{x}}^t | Z_{\mathbf{x}}, \xi_{\mathbf{x}}^2\right)$, where $\xi_{\mathbf{x}}$ denotes the noise standard deviation, and is predefined based on the systematic error of the depth sensor. For instance, the noise variance of Kinect is actually related to the depth so it is appropriate to set $\xi_{\mathbf{x}}$ depth-dependently.

b) State-F (forward outliers): On the other hand, the depth measurements from moving objects or outliers in front, follow a clutter distribution like $\mathcal{U}_f(d_{\mathbf{x}}^t | Z_{\mathbf{x}}) = U_f \cdot 1_{[d_{\mathbf{x}}^t < Z_{\mathbf{x}}]}$, where $1_{[\cdot]}$ is an indicator function that equals to 1 when the input argument is true, and 0 otherwise. This state is activated when $d_{\mathbf{x}}^t$ is smaller than $Z_{\mathbf{x}}$, and switched off if it is larger than $Z_{\mathbf{x}}$. It can be inferred from this state that not only are the outliers in front, but also the dynamic objects are at the given location.

c) State-B (backward outliers): Furthermore, it is possible that the input depth measurements are outliers lying behind the current estimation of the static structure. Another similar indicator distribution is introduced as $\mathcal{U}_b\left(d_{\mathbf{x}}^t | Z_{\mathbf{x}}\right) = U_b \cdot 1_{[d_{\mathbf{x}}^t > Z_{\mathbf{x}}]}$. It can naturally represent outliers that have larger depth values than a given structure. Meanwhile, it provides a cue to infer the risk whether current static structure estimation is incorrect.

An additional hidden variable $\mathbf{m}_{\mathbf{x}} = \left[m_{\mathbf{x}}^I, m_{\mathbf{x}}^F, m_{\mathbf{x}}^B\right]^\top$ is introduced as the *state indicator* to represent these states, where $m_{\mathbf{x}}^k \in \{0, 1\}, k \in \Psi$. In this case, only one specific state $m_{\mathbf{x}}^k = 1$ and the rest are 0s, thus $\sum_{k\in\Psi} m_{\mathbf{x}}^k = 1$.

2) A Generative Model: The reason to introduce the generative model is that it can simulate the static structure as well as its noise and outliers, thus in case there are no observed measurements at the current frame (*e.g.*, depth holes), we can still give a reasonable static structure. Moreover, given suitable parametric forms of these distributions, the generative model can be online estimated and refined by updating the parameters with sequentially incoming depth samples.

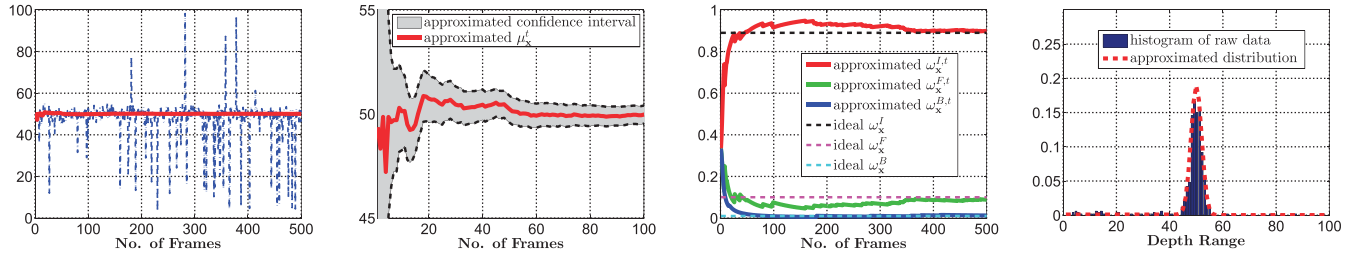
a) Likelihood: Appending the state indicator $\mathbf{m}_{\mathbf{x}}$, the likelihood of $d_{\mathbf{x}}^t$ conditioned on $\mathbf{m}_{\mathbf{x}}$ and the static structure $Z_{\mathbf{x}}$ is a product of the distributions of the three states as $p(d_{\mathbf{x}}^t | \mathbf{m}_{\mathbf{x}}, Z_{\mathbf{x}}) = \mathcal{N}(d_{\mathbf{x}}^t | Z_{\mathbf{x}}, \xi_{\mathbf{x}}^2)^{m_{\mathbf{x}}^I} \mathcal{U}_f\left(d_{\mathbf{x}}^t | Z_{\mathbf{x}}\right)^{m_{\mathbf{x}}^F} \mathcal{U}_b(d_{\mathbf{x}}^t | Z_{\mathbf{x}})^{m_{\mathbf{x}}^B}$. It equals to one required state distribution by triggering off this state indicator $m_{\mathbf{x}}^k = 1, k \in \Psi$.

b) Prior: Let the prior for $Z_{\mathbf{x}}$ also be a Gaussian distribution with the mean $\mu_{\mathbf{x}}$ and the standard deviation $\sigma_{\mathbf{x}}$, written as $p(Z_{\mathbf{x}}) = \mathcal{N}\left(Z_{\mathbf{x}} | \mu_{\mathbf{x}}, \sigma_{\mathbf{x}}^2\right)$. $\sigma_{\mathbf{x}}$ is different from $\xi_{\mathbf{x}}$ since it represents the possible range of the static structure rather than its noise level. The prior of the chance to activate one state is a categorical distribution $\mathrm{Cat}(\mathbf{m}_{\mathbf{x}} | \boldsymbol{\omega}_{\mathbf{x}})$ [38], where $\boldsymbol{\omega}_{\mathbf{x}} = \left[\omega_{\mathbf{x}}^I, \omega_{\mathbf{x}}^F, \omega_{\mathbf{x}}^B\right]^\top$ and $\sum_{k\in\Psi} \omega_{\mathbf{x}}^k = 1, \omega_{\mathbf{x}}^k \in (0, 1)$. This parameter reveals the opportunities to induce these states in advance of the input depth samples. And $\boldsymbol{\omega}_{\mathbf{x}}$ is further modeled by a Dirichlet distribution $p(\boldsymbol{\omega}_{\mathbf{x}}) = \mathrm{Dir}(\boldsymbol{\omega}_{\mathbf{x}} | \boldsymbol{\alpha}_{\mathbf{x}})$, where $\boldsymbol{\alpha}_{\mathbf{x}} = [\alpha_{\mathbf{x}}^I, \alpha_{\mathbf{x}}^F, \alpha_{\mathbf{x}}^B]^\top, \alpha_{\mathbf{x}}^k \in \mathbb{R}^+$ and corresponds to $\omega_{\mathbf{x}}^k$.

c) Posterior: Two posteriors are in fact essential for the static structure estimation. One is $p(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}} | \mathcal{D}_{\mathbf{x}}^t)$, which jointly presents the depth distribution of the static structure and the popularity densities of these three states given the current and all previous depth frames. The other is the posterior of the state indicator $p(\mathbf{m}_{\mathbf{x}} | \mathcal{D}_{\mathbf{x}}^t)$, which represents the possible states at the current frame. Based on the estimated posteriors, we can evaluate the most probable depth values of the static structure by calculating the expectation of $p(Z_{\mathbf{x}} | \mathcal{D}_{\mathbf{x}}^t)$ as $\mathbb{E}_{p(Z_{\mathbf{x}} | \mathcal{D}_{\mathbf{x}}^t)}[Z_{\mathbf{x}}]$. The reliability of current estimation refers to $\mathbb{E}_{p(\boldsymbol{\omega}_{\mathbf{x}} | \mathcal{D}_{\mathbf{x}}^t)}\left[\omega_{\mathbf{x}}^I\right]$, which means that the larger the portion of input depth samples that agree with the model, the more reliable the estimation is. The most possible state that $d_{\mathbf{x}}^t$ should occupy is calculated straightforwardly from $\arg\max_{\mathbf{m}_{\mathbf{x}}} p(\mathbf{m}_{\mathbf{x}} | \mathcal{D}_{\mathbf{x}}^t)$.

B. Variational Approximation

However it is almost unfeasible to solve these posteriors analytically because it is not independent between $Z_{\mathbf{x}}$ and $\boldsymbol{\omega}_{\mathbf{x}}$ for $p(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}} | \mathcal{D}_{\mathbf{x}}^t)$, and $p(Z_{\mathbf{x}} | \mathcal{D}_{\mathbf{x}}^t)$ and $p(\boldsymbol{\omega}_{\mathbf{x}} | \mathcal{D}_{\mathbf{x}}^t)$ do not exactly follow Gaussian and Dirichlet distributions any more. Therefore, variational approximation [38] of the posteriors is introduced to provide sufficiently accurate approximated posteriors efficiently. It minimizes the Kullback-Leibler divergence between the approximated and the original posteriors. The variationally approximated posteriors are



(a) $Z_{\mathbf{x}}^t$ *v.s.* $\mathcal{D}_{\mathbf{x}}^t$ (b) Confidence interval *w.r.t.* $Z_{\mathbf{x}}^t$ (c) Evolution of each states' portions (d) $q^T(d_{\mathbf{x}}|\mathcal{P}_{\mathbf{x}}^{\mathcal{D},T})$ *v.s.* data histogram

Fig. 4. Variational approximation of the parameter set of the static structure for a 1D depth sequence. The number of frames is $T = 500$. (a) The expected depth sequence of the static structure versus the raw depth sequence, where the ideal $Z_{\mathbf{x}} = 50$. (b) The confidence interval of $Z_{\mathbf{x}}^t$, the interval is centered $\mu_{\mathbf{x}}^t$ and between $\mu_{\mathbf{x}}^t \pm 2\sigma_{\mathbf{x}}^t$ with 95% confidence. (c) The evolution of the portions (defined by the expected value of $\boldsymbol{\omega}_{\mathbf{x}}$ at frame t, denoted by $[\omega_{\mathbf{x}}^{I,t}, \omega_{\mathbf{x}}^{F,t}, \omega_{\mathbf{x}}^{B,t}]$) of the three states. The ideal portions are $\boldsymbol{\omega}_{\mathbf{x}} = [0.89, 0.1, 0.01]$. (d) The estimated distribution $q^T(d_{\mathbf{x}}|\mathcal{P}_{\mathbf{x}}^{\mathcal{D},T})$ versus the normalized histogram estimated by $\mathcal{D}_{\mathbf{x}}^T$ when $T = 500$. The estimated depth of the static structure goes to the ideal value only with a few samples. Its confidence interval shrinks rapidly, which means the uncertainty is reduced very fast. The portion of each state is evolved with the raw depth sequence, and they match their ideal value with enough depth samples. When $T = 500$, the estimated data distribution fits the data histogram compactly.

required to own the same parametric forms as the priors thus they also produce analytical solutions to approximate $\mathbb{E}_{p(Z_{\mathbf{x}}|\mathcal{D}_{\mathbf{x}}^t)}[Z_{\mathbf{x}}]$ and $\mathbb{E}_{p(\boldsymbol{\omega}_{\mathbf{x}}|\mathcal{D}_{\mathbf{x}}^t)}[\omega_{\mathbf{x}}^I]$. The approximation starts from factorizing the posterior $p(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}}|\mathcal{D}_{\mathbf{x}}^t)$ into the product of independent Gaussian distribution $q^t(Z_{\mathbf{x}}) = \mathcal{N}\left(Z_{\mathbf{x}}|\mu_{\mathbf{x}}^t, (\sigma_{\mathbf{x}}^t)^2\right)$ and Dirichlet distribution $q^t(\boldsymbol{\omega}_{\mathbf{x}}) = \mathrm{Dir}(\boldsymbol{\omega}_{\mathbf{x}}|\boldsymbol{\alpha}_{\mathbf{x}}^t)$ as

$$q^t\left(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}}\right) = q^t(Z_{\mathbf{x}})q^t(\boldsymbol{\omega}_{\mathbf{x}}) \sim p(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}}|\mathcal{D}_{\mathbf{x}}^t). \tag{1}$$

Not only that, but the exact estimation also depends on all the previous depth samples $\mathcal{D}_{\mathbf{x}}^t$. Too many frames will bring about unbearable complexity and memory requirement. We admit a first order Markov chain into our framework so as to favor the *online* estimation. It means that we can estimate the current posterior just based on the current likelihood and the posterior of the last frame, therefore it is memory- and computationally efficient. We reformulate the posterior as a sequential parameter estimation problem

$$\begin{aligned} q^t\left(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}}\right) &\sim p\left(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}}|\mathcal{D}_{\mathbf{x}}^t\right) \\ &\sim p(d_{\mathbf{x}}^t|Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}})q^{t-1}\left(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}}\right)/q^t(d_{\mathbf{x}}^t) \\ &= Q(Z_{\mathbf{x}}, \boldsymbol{\omega}_{\mathbf{x}}|d_{\mathbf{x}}^t), \end{aligned} \tag{2}$$

where the parameters of the left-hand side are estimated by matching moments between the distributions of left- and right-hand sides [38]. This only considers the current data samples and the previous estimated parameters to approximate the current parameters. We define the parameter set estimated at $t-1$ is $\mathcal{P}_{\mathbf{x}}^{\mathcal{D},t-1} = \{\mu_{\mathbf{x}}^{t-1}, \sigma_{\mathbf{x}}^{t-1}, \boldsymbol{\alpha}_{\mathbf{x}}^{t-1}\}$, while the required parameter set is $\mathcal{P}_{\mathbf{x}}^{\mathcal{D},t}$. By matching the first and the second moments between $Q(Z_{\mathbf{x}}|d_{\mathbf{x}}^t)$ and $q^t(Z_{\mathbf{x}})$ as well as those between $Q(\boldsymbol{\omega}_{\mathbf{x}}|d_{\mathbf{x}}^t)$ and $q^t(\boldsymbol{\omega}_{\mathbf{x}}|d_{\mathbf{x}}^t)$ [39], we can obtain a closed-form solution for any parameter in $\mathcal{P}_{\mathbf{x}}^{\mathcal{D},t}$. Please refer to the supplementary materials for their detailed derivations.

Hence, recall the problem addressed in Section III-A, the approximated posterior with respect to the state indicator $\mathbf{m}_{\mathbf{x}}$ is $q^t(m_{\mathbf{x}}^k = 1|d_{\mathbf{x}}^t), k \in \Psi$, which is a suitable approximation of $p(\mathbf{m}_{\mathbf{x}}|\mathcal{D}_{\mathbf{x}}^t)$ and also has a closed-form solution.

Apart from that, the most probable depth value of the static structure at pixel $\mathbf{x}$ and time t is

$$Z_{\mathbf{x}}^t = \mathbb{E}_{p(Z_{\mathbf{x}}|\mathcal{D}_{\mathbf{x}}^t)}[Z_{\mathbf{x}}] \simeq \mu_{\mathbf{x}}^t, \tag{3}$$

and the reliability of current estimation of the static structure is the expectation of $\omega_{\mathbf{x}}^I$ as

$$r_{\mathbf{x}}^t = \mathbb{E}_{p(\boldsymbol{\omega}_{\mathbf{x}}|\mathcal{D}_{\mathbf{x}}^t)}\left[\omega_{\mathbf{x}}^I\right] \simeq \alpha_{\mathbf{x}}^{I,t} / \sum_{k\in\Psi} \alpha_{\mathbf{x}}^{k,t}. \tag{4}$$

As shown in Figure 4, an example of the variational approximation of the parameter set for a 1D depth sequence illustrates the potential of the proposed method to capture the nature of the input depth sequence.

C. Improvement With Color Video

The above discussion only considers the estimation and update of the static structure with the depth video. A more complete treatment is together with the registered color video, in which case an improved probabilistic generative model can be formulated as follows.

1) Prior: We introduce another prior over $\mathbf{C}_{\mathbf{x}}$, the color value of the static structure at $\mathbf{x}$ as $p(\mathbf{C}_{\mathbf{x}}) = \mathcal{N}\left(\mathbf{C}_{\mathbf{x}}|\mathbf{U}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}\right)$ with two parameters: the mean $\mathbf{U}_{\mathbf{x}}$ and the variance $\boldsymbol{\Sigma}_{\mathbf{x}}$.

2) Likelihood: The likelihood of input depth and color samples $d_{\mathbf{x}}^t$ and $\mathbf{I}_{\mathbf{x}}^t$ conditioned on $\mathbf{m}_{\mathbf{x}}$ given $Z_{\mathbf{x}}$ and $\mathbf{C}_{\mathbf{x}}$ is

$$\begin{aligned} p(d_{\mathbf{x}}^t, \mathbf{I}_{\mathbf{x}}^t|\mathbf{m}_{\mathbf{x}}, Z_{\mathbf{x}}, \mathbf{C}_{\mathbf{x}}) = \mathcal{U}_f\left(d_{\mathbf{x}}^t|Z_{\mathbf{x}}\right)^{m_{\mathbf{x}}^F} \mathcal{U}_b\left(d_{\mathbf{x}}^t|Z_{\mathbf{x}}\right)^{m_{\mathbf{x}}^B} \\ \cdot \left[\mathcal{N}\left(d_{\mathbf{x}}^t|Z_{\mathbf{x}}, \xi_{\mathbf{x}}^2\right)\mathcal{N}\left(\mathbf{I}_{\mathbf{x}}^t|\mathbf{C}_{\mathbf{x}}, \boldsymbol{\Xi}_{\mathbf{x}}\right)\right]^{m_{\mathbf{x}}^I}, \end{aligned} \tag{5}$$

where $\boldsymbol{\Xi}_{\mathbf{x}}$ denotes the variance matrix for the color noise. A step further we have the likelihood of $d_{\mathbf{x}}^t$ and $\mathbf{I}_{\mathbf{x}}^t$ conditioned on $Z_{\mathbf{x}}$ and $\mathbf{C}_{\mathbf{x}}$ accordingly. This formulation improves the inference since the input depth sample will belong to the static structure only when both the depth and color samples agree with the previous model. Therefore, the risk of false estimation is reduced.

3) Posterior and Variational Approximation: In a similar fashion in Section III-B, we can derive the approximated posterior when color video exists. The parameter set $\mathcal{P}_{\mathbf{x}}^{\mathcal{S},t} = \left\{\mu_{\mathbf{x}}^t, \sigma_{\mathbf{x}}^t, \mathbf{U}_{\mathbf{x}}^t, \boldsymbol{\Sigma}_{\mathbf{x}}^t, \boldsymbol{\alpha}_{\mathbf{x}}^t\right\}$, $\mathcal{S} = \{\mathcal{D}, \mathcal{I}\}$ can also be estimated online and analytically. Furthermore, the most probable depth $Z_{\mathbf{x}}^t$ and color $\mathbf{C}_{\mathbf{x}}^t$ of the static structure are achieved based on $\mu_{\mathbf{x}}^t$ and $\mathbf{U}_{\mathbf{x}}^t$. The approximate posteriors $q^t(\mathbf{m}_{\mathbf{x}}^k|d_{\mathbf{x}}^t, \mathbf{I}_{\mathbf{x}}^t), k \in \Psi$ are also derived accordingly.

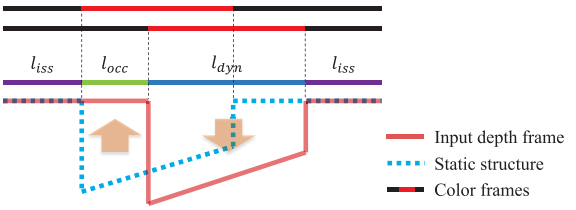


Fig. 5. One toy example illustrates the layer assignment. The cyan dot line indicates the current estimated depth structure of the static structure, and the red solid line is from the input depth frame. If color frames are available, they provide additional constraints to regularize the assignment, where the upper line corresponds to the current estimated texture structure of the static structure, and the lower one refers to the input color frame.

D. Layer Assignment

In this section, we would like to find the static region of the input depth frame so as to robustly update the model of the static structure and find the dynamic region. Specifically, we label the input depth frame in three layers $\mathcal{L} = \{l_{iss}, l_{dyn}, l_{occ}\}$:

- l_{iss}: agree with estimated static structure;
- l_{dyn}: belong to dynamic objects in its front; or
- l_{occ}: refer to the once occluded structure behind it.

The additional label l_{occ} is essential because the regions belonging to the once occluded structure do not fit the current model, but they reveal the hidden structure behind the current estimated static structure. It also points out that current estimation produces bias at these regions, in which the depth structure from the input depth frame $\mathbf{D}^t$ would be a more reasonable substitution to rectify the previous estimation.

One toy example is shown in Figure 5, where $\mathbf{D}^t$ provides a different layout from the current static structure. Intuitively, l_{occ} occurs when the input depth frame provides larger depth values and exposes the hidden static structure. l_{dyn}, on the contrary, encourages smaller depth values. Furthermore, the failure of inference due to depth holes, noise and outliers can be eliminated by the introduction of texture information, which also provides additional cues to regularize their spatial layout.

To improve the expressive power to label complex structures that is employed frequently in our case, we exploit a fully connected conditional random field (fully-connected CRF) [40] to strengthen the spatial long-range relationship. Assume a random field $\mathbf{L} = \{l_{\mathbf{x}} \in \mathcal{L} \mid \forall \mathbf{x}\}$ conditioned on the input data $\mathbf{S}^t$ and the previous model parameter set $\mathcal{M} = \mathcal{P}^{\mathcal{S},t-1}$. The Gibbs energy of a label assignment $\mathbf{L}$ is

$$E(\mathbf{L}|\mathbf{S}^t, \mathcal{M}) = \sum_{\mathbf{x}} \psi_u\left(l_{\mathbf{x}}|\mathbf{S}^t, \mathcal{M}\right) + \frac{1}{2}\sum_{\mathbf{x}\neq\mathbf{y}} \psi_p\left(l_{\mathbf{x}}, l_{\mathbf{y}}|\mathbf{S}^t, \mathcal{M}\right), \tag{6}$$

where $\mathbf{x}$ and $\mathbf{y}$ are pixel locations. $\psi_u(\cdot)$ and $\psi_p(\cdot,\cdot)$ indicate the unary and pairwise potentials. $\mathbf{S}^t = \mathbf{D}^t$ or $\{\mathbf{D}^t, \mathbf{I}^t\}$.

1) Definition of Unary and Pairwise Potentials: We define the unary potentials and pairwise potentials as follows:

a) Unary potentials: The unary potentials are negative logarithms of the approximated posteriors $q^t(\mathbf{m}_{\mathbf{x}}|\mathbf{S}^t_{\mathbf{x}})$, indicating the chance that the current depth samples should follow the previous estimation (*i.e.*, l_{iss} requires $m^I_{\mathbf{x}} = 1$), or in its front (*i.e.*, l_{dyn} needs $m^F_{\mathbf{x}} = 1$) or at its back (*i.e.*, l_{occ} refers to $m^B_{\mathbf{x}} = 1$). In detail, we have $\psi_u\left(l_{\mathbf{x}} = l_k|\mathbf{S}^t, \mathcal{M}\right) = -\ln q^t(m^k_{\mathbf{x}} = 1|\mathbf{S}^t_{\mathbf{x}})$, and l_k and $m^k_{\mathbf{x}}$ follow the correspondences listed above.

b) Pairwise potentials: The pairwise potential between pixels $\mathbf{x}$ and $\mathbf{y}$ is a weighted mixture of Gaussian kernels as

$$\begin{aligned}\psi_p(l_{\mathbf{x}}, l_{\mathbf{y}}|\mathbf{S}^t_{\mathbf{x}}, \mathcal{M}_{\mathbf{x}}) &= 1_{[l_{\mathbf{x}}\neq l_{\mathbf{y}}]} \\ &\cdot\Big\{w_s \exp\left(-\tau_\alpha\|\mathbf{x}-\mathbf{y}\|^2/2\right) \\ &\quad + w_r \exp\left(-\|\Delta_t\mathbf{f}_{\mathbf{x}} - \Delta_t\mathbf{f}_{\mathbf{y}}\|^2_{\Sigma_\beta}/2 - \tau_\gamma\|\mathbf{x}-\mathbf{y}\|^2/2\right)\Big\}.\end{aligned} \tag{7}$$

We define $\Delta_t\mathbf{f}_{\mathbf{x}} = \mathbf{f}^{I,t-1}_{\mathbf{x}} - \mathbf{f}^t_{\mathbf{x}}$ to measure the difference between the features of the static structure and those of the input data. When $\mathbf{S}^t = \mathbf{D}^t$, $\mathbf{f}^t_{\mathbf{x}}$ and $\mathbf{f}^{I,t-1}_{\mathbf{x}}$ are the normalized $d^t_{\mathbf{x}}$ and $Z^{t-1}_{\mathbf{x}}$, by a whitening process of the overall variance $(\tilde{\xi}^t_{\mathbf{x}})^2 = (\sigma^{t-1}_{\mathbf{x}})^2 + \xi^2_{\mathbf{x}}$. If $\mathbf{S}^t = \{\mathbf{D}^t, \mathbf{I}^t\}$, let $\mathbf{f}^t_{\mathbf{x}}$ and $\mathbf{f}^{I,t-1}_{\mathbf{x}}$ be the concatenations of the normalized vectors $\left[d^t_{\mathbf{x}}; \mathbf{I}^t_{\mathbf{x}}\right]$ and $\left[Z^{t-1}_{\mathbf{x}}; \mathbf{C}^{t-1}_{\mathbf{x}}\right]$. The color features are normalized with the variance $\tilde{\mathbf{\Xi}}^t_{\mathbf{x}} = \mathbf{\Xi}_{\mathbf{x}} + \mathbf{\Sigma}^{t-1}_{\mathbf{x}}$.

The indicator function $1_{[l_{\mathbf{x}}\neq l_{\mathbf{y}}]}$ lets the pairwise potentials be Potts model. It encourages a penalty for nearby pixels that are assigned different labels but they have similar features. The first kernel is a smoothness kernel that removes small isolated regions and is adjusted by τ_α. The second kernel is a range kernel trying to force nearby pixels with similar depth and/or color variation to share the same label, with a given parameter τ_γ to set the degree of nearness. $\|\Delta_t\mathbf{f}_{\mathbf{x}} - \Delta_t\mathbf{f}_{\mathbf{y}}\|^2_{\Sigma_\beta}$ is the Mahalanobis distance between $\Delta_t\mathbf{f}_{\mathbf{x}}$ and $\Delta_t\mathbf{f}_{\mathbf{y}}$, where the covariance matrix Σ_β encodes the feature proximity. The weight of the range kernel is set as w_r. If we only have the range kernel, the result tends to be noisy, while if we only have the smooth kernel, the structure cannot be well regularized.

2) Inference: We exploit an efficient mean field inference method for fully-connected CRF when the pairwise potentials are Gaussian [40]. It turns out to be an iterative estimation process convolving several runs of real-time high dimensional filtering characterized by the pairwise potentials (7).

E. Online Static Structure Update Scheme

The online static structure updating scheme is actually a sequential variational parameter estimation problem with a layer assignment to exclude the dynamic objects and include the once occluded static structure. A spatial enhancement is appended to regularize the spatial layout of the structure. The sketch of the algorithm is given in Algorithm 1.

An initialization of the parameter set $\mathcal{P}^{\mathcal{S}}$ is necessary. We set the initial $\mu^0_{\mathbf{x}} = d^0_{\mathbf{x}}$, where $d^0_{\mathbf{x}} \in \mathbf{D}^0$ from the first frame of the depth video. Similarly, let $\mathbf{U}^0_{\mathbf{x}} = \mathbf{I}^0_{\mathbf{x}}$, where $\mathbf{I}^0_{\mathbf{x}} \in \mathbf{I}^0$ from the color video. The noise parameters $\xi_{\mathbf{x}}$ and $\mathbf{\Xi}_{\mathbf{x}}$ are user-specified constants which should be large enough to enable sufficient variance of input data. $\sigma^0_{\mathbf{x}}$ and $\mathbf{\Sigma}^0_{\mathbf{x}}$ will be initialized as large values as well. The parameters of $\boldsymbol{\omega}_{\mathbf{x}}$ are also set up with given constants $\boldsymbol{\alpha}^0_{\mathbf{x}}$. A convenient setup is $\alpha^{I,0}_{\mathbf{x}} = \alpha^{F,0}_{\mathbf{x}} = \alpha^{B,0}_{\mathbf{x}}$. The user-given initialization parameter set is $\mathcal{P}^{\mathcal{S}}_{init} = \{\xi_{\mathbf{x}}, \sigma^0_{\mathbf{x}}, \boldsymbol{\alpha}^0_{\mathbf{x}} \mid \forall\mathbf{x}\}$ when $\mathcal{S} = \mathcal{D}$ and $\mathcal{P}^{\mathcal{S}}_{init} = \{\xi_{\mathbf{x}}, \sigma^0_{\mathbf{x}}, \mathbf{\Xi}_{\mathbf{x}}, \mathbf{\Sigma}^0_{\mathbf{x}}, \boldsymbol{\alpha}^0_{\mathbf{x}} \mid \forall\mathbf{x}\}$ when

Algorithm 1 Online Static Structure Update Scheme

Input : Data sequence $\mathcal{S} = \{\mathbf{S}^\tau | \tau = 0, 1, 2, \ldots\}$;
Initial parameter set $\mathcal{P}^{\mathcal{S}}_{init}$;
Output: Current parameter set $\mathcal{P}^{\mathcal{S},t}$;

// initialization
$t \leftarrow 0$, $\mathcal{P}^{\mathcal{S},0} \leftarrow$ param_init($\mathbf{S}^0$, $\mathcal{P}^{\mathcal{S}}_{init}$);
while $\mathcal{S} \neq \emptyset$ **do**
  $t \leftarrow t + 1$;
  // 1.layer assignment
  $\mathcal{M} \leftarrow \mathcal{P}^{\mathcal{S},t-1}$, $\mathbf{L} \leftarrow \arg\min_{\mathbf{L}} E(\mathbf{L}|\mathbf{S}^t, \mathcal{M})$;
  // 2.parameter update
  for $\forall \mathbf{x}$ **do**
    if $l_{\mathbf{x}} = l_{iss}$ **then** $\mathcal{P}^{\mathcal{S},t}_{\mathbf{x}} \leftarrow$ vari_approx($\mathbf{S}^t_{\mathbf{x}}$, $\mathcal{P}^{\mathcal{S},t-1}_{\mathbf{x}}$)
    else if $l_{\mathbf{x}} = l_{occ}$ **then** $\mathcal{P}^{\mathcal{S},t}_{\mathbf{x}} \leftarrow$ param_init($\mathbf{S}^t_{\mathbf{x}}, \mathcal{P}^{\mathcal{S}}_{init}$)
    else if $l_{\mathbf{x}} = l_{dyn}$ **then** $\mathcal{P}^{\mathcal{S},t}_{\mathbf{x}} \leftarrow \mathcal{P}^{\mathcal{S},t-1}_{\mathbf{x}}$
  // 3.spatial enhancement
  $Z^t_{\mathbf{x}} \leftarrow \mu^t_{\mathbf{x}}, \forall \mathbf{x}$;
  $\tilde{\mathbf{Z}}^t \leftarrow$ spatial_enhance($\mathbf{Z}^t, \mathcal{P}^{\mathcal{S},t}$), $\mu^t_{\mathbf{x}} \leftarrow \tilde{Z}^t_{\mathbf{x}}, \forall \mathbf{x}$;

$\mathcal{S} = \{\mathcal{D}, \mathcal{I}\}$. In addition, the layer assignment is not applied in the initialization step.

At the t^{th} frame, the layer assignment is applied at first based on the previous parameter set $\mathcal{P}^{\mathcal{S},t-1}$ and the input data $\mathcal{S}^t$. The region in which $l_{\mathbf{x}} = l_{iss}$ will perform the variational parameter estimation to obtain a renewed $\mathcal{P}^{\mathcal{S},t}_{\mathbf{x}}$. If $l_{\mathbf{x}} = l_{dyn}$, it belongs to a dynamic object so that $\mathcal{P}^{\mathcal{S},t}_{\mathbf{x}} = \mathcal{P}^{\mathcal{S},t-1}_{\mathbf{x}}$. But on the other hand, if $l_{\mathbf{x}} = l_{occ}$, the parameter set of this pixel is re-initialized as in the initialization step, but $\mu^t_{\mathbf{x}} = d^t_{\mathbf{x}}$, $\mathbf{U}^t_{\mathbf{x}} = \mathbf{I}^t_{\mathbf{x}}$. Furthermore, it is a common phenomenon that the input depth frames contain holes without depth measurements. In this case, $\mu^t_{\mathbf{x}}$ and $\lambda^t_{\mathbf{x}}$ will not be updated in these special regions.

The spatial enhancement, including hole filling, smoothing and regularization, is necessary to generate a spatially refined static structure. It is performed after the parameter estimation in each frame, where we have obtained the most probable depth map $\mathbf{Z}^t$ ($Z^t_{\mathbf{x}} \in \mathbf{Z}^t$). A variational inpainting method incorporating a TV-Huber norm and a data term by Mahalanobis distance with the variance $(\tilde{\xi}^t_{\mathbf{x}})^2$ is employed for spatial enhancement, which is iteratively solved by a primal-dual approach [14]. Since the solver requires hundreds of runs to converge, a trade-off between speed and accuracy is adopted by fixing the number of iterations and borrowing the spatially enhanced result in the last frame $\tilde{\mathbf{Z}}^{t-1}$ as the initialization. To reduce error propagation, unreliable pixels in the input depth map $\mathbf{Z}^t$ are deleted according to the reliability check $r^t_{\mathbf{x}} > 0.5$ (c.f., equation (4)). Given the most probable color image of the current static structure $\mathbf{C}^t$, the spatial enhancement in $\mathbf{Z}^t$ can absorb the texture information to guide the propagation of the local structures. In the end, the enhanced depth map $\tilde{Z}^t_{\mathbf{x}}$ will substitute $\mu^t_{\mathbf{x}}$ in $\mathcal{P}^{\mathcal{S},t}_{\mathbf{x}}$.

F. Temporally Consistent Depth Video Enhancement

Apart from spatial enhancement, it is preferred to employ temporal enhancement to produce a flicker-free depth video. To enable long-range temporal consistency and allow online processing, we exploit the static structure of the captured scene as a medium to find the region in the input frame exhibiting long-range temporal connection. The static region is enhanced by fusing the input depth measurements with the static structure according to the online static structure update scheme in Section III-E. Thus the static regions are well-preserved and incrementally refined over time. The idea behind this is that we restrict the temporal consistency to be enforced only around static region or slowly moving objects. This assumption is somewhat restrictive but is still suitable to process normal depth videos. One additional advantage of the proposed method is that it can prevent bleeding artifacts that propagate depth values from moving objects into the static background as long as the layer assignment is robust.

Given the resulting layer assignment of the current frame, the static region is where $l_{\mathbf{x}} \in \{l_{iss}, l_{occ}\}$, including the regions referring to the static structure and those belonging to the once occluded static structure. They both expose the current visible static structure of the captured scene, thus shall be enhanced separately from the dynamic objects. The enhanced version is obtained by substituting it with its counterpart in the static structure, which has already been updated in the temporal domain and enhanced in the spatial domain (see Section III-E). The dynamic region can be enhanced by various approaches explored in the literature, while in this paper we exploit a conventional joint bilateral filter, both to fill holes and to perform edge-preserving filtering in the dynamic region.

The proposed method is both memory- and computationally efficient. The memory requested for the proposed method only goes to storing the parameter set for each pixel, thus is efficient to process streaming videos or long sequences of high quality. Excepting the cost of the spatial enhancement, the complexity for temporal enhancement hinges on that of the online static structure update scheme, in which all the required parameters have analytical solutions whilst the layer assignment is efficient thanks to the constant-time implementations in solving the fully-connected CRF model. Provided with an efficient spatial enhancement approach, for example, the domain transform filter [41] or the proposed one with the help of multi-thread techniques or GPGPUs [42], the entire temporally consistent depth video enhancement procedure can be achieved in real-time.

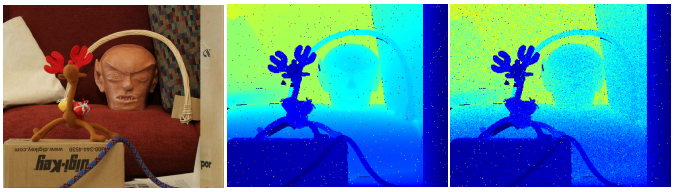
IV. Experiments and Discussions

In this section, we present our experiments on synthetic and real data to demonstrate the effectiveness and robustness of our static structure estimation and depth video enhancement.

Section IV-A numerically evaluates the performance of our method for static structure estimation using synthetic depth videos[2] generated from the Middlebury dataset [43], [44]. Our method is not sensitive to the user-given parameters, and outperforms various methods about static scene estimation with running time comparable to temporal median filtering.

In Section IV-B, we evaluate the performance on real data captured by Kinect and ToF cameras. Both static and dynamic indoor scenes are taken into consideration. Apart from

[2]The depth of one pixel in the depth frame is proportional to the reciprocal of the disparity at the same place in the corresponding disparity frame.



(a) Reindeer (b) I : $(10^{-2}, 2)$ (c) II : $(10^{-2}, 2)$

Fig. 6. Sample frames of the input depth video with two types of noise and outliers. (a) is the sample color frame, (b) and (c) are the contaminated depth frames with $\sigma_n = 2$ and $\omega_n = 10^{-2}$. (b) is type-I but (c) is type-II. Type-II error is worse than type-I error with the same parameters.

the estimation of static structure, we also evaluate the performance of the static scene reconstruction and most importantly, the temporally consistent depth video enhancement in Section IV-C.

Initial parameters are simply set as $\boldsymbol{\alpha}_{\mathbf{x}}^0 = [1, 1, 1]^\top$, $\sigma_{\mathbf{x}}^0$ is the 10% of the depth range of the input scene. And initial parameter $\boldsymbol{\Sigma}_{\mathbf{x}}^0$ is a diagonal matrix with each diagonal entity the square of 10% of the color range.

A. Numerical Evaluation of the Static Structure Estimation by Synthesized Data

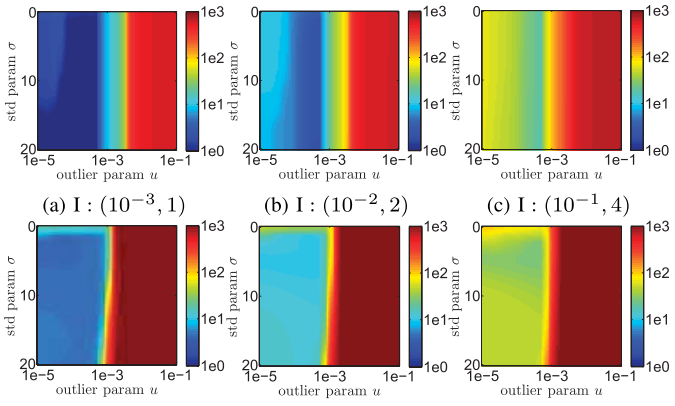
We used two types of noise and outliers, which are illustrated in Figure 6, to contaminate the depth video so that we could evaluate the performance of our method with respect to different kinds of errors from different types of depth sensors.

Type-I: We contaminated the depth map via $p(d_{\mathbf{x}}|Z_{\mathbf{x}}) = (1-\omega_n)\mathcal{N}\left(d_{\mathbf{x}}|Z_{\mathbf{x}}, \sigma_n^2\right) + \omega_n \mathcal{U}\left(d_{\mathbf{x}}\right)$, where $\mathcal{U}(d_{\mathbf{x}})$ is the reciprocal of the depth range. It is a general model of noise and outliers.

Type-II: We damaged the disparity map by $p(d_{\mathbf{x}}^{\text{disp}}|Z_{\mathbf{x}}^{\text{disp}}) = (1-\omega_n)\mathcal{N}(d_{\mathbf{x}}^{\text{disp}}|Z_{\mathbf{x}}^{\text{disp}}, \sigma_n^2) + \omega_n \mathcal{U}(d_{\mathbf{x}}^{\text{disp}})$ and rounded it. The disparity map was transformed into the depth map. $\mathcal{U}(d_{\mathbf{x}}^{\text{disp}})$ was the reciprocal of the disparity range. It mimicked the outliers in common depth videos captured by stereo or Kinect.

1) Analysis of User-Given Parameters: We first evaluated the user-given parameters for the outlier parameters U_f, U_b and the noise standard deviation $\xi_{\mathbf{x}}$. In case-I, we set $\xi_{\mathbf{x}} = \sigma$ as a constant throughout the pixel domain. For case-II, the choice of $\xi_{\mathbf{x}}$ should be suitable to dispose of the non-uniform quantization error due to disparity-depth conversion as $\xi_{\mathbf{x}} = \sigma \frac{d_{\mathbf{x}}^2}{fB}$.[3] Meanwhile, we set $U_f = U_b = u$. The experiments were evaluated by the RMSE score with varying u and σ under different levels of noise (σ_d) and outliers (ω_d). The results are shown in Figure 7, where the test video had 100 frames. We set $\sigma \in [0, 20]$ and $u \in [10^{-5}, 10^{-1}]$. Notice that the tested scene was static thus there was NO need to perform layer assignment. The spatial enhancement was also skipped.

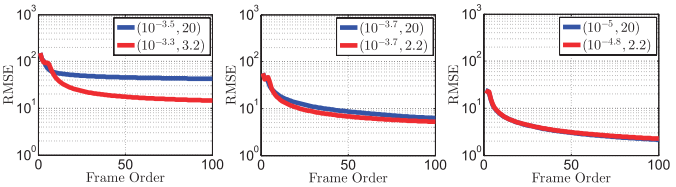
[3] f is the focal length and B is the baseline, both of which are provided in the Middlebury dataset. The conversion relationship is derived in the supplementary materials.



(a) I : $(10^{-3}, 1)$ (b) I : $(10^{-2}, 2)$ (c) I : $(10^{-1}, 4)$

(d) II : $(10^{-3}, 1)$ (e) II : $(10^{-2}, 2)$ (f) II : $(10^{-1}, 4)$

Fig. 7. RMSE maps with varying u and σ under different noise and outlier parameter pairs (ω_n, σ_n). (a)-(c) were contaminated by type-I, while (d)-(e) were contaminated by type-II. Best viewed in color.



(a) (b) (c)

Fig. 8. Performance comparisons between the constant and depth-dependent $\xi_{\mathbf{x}}$ under different type-II noise and outlier parameter pairs (ω_n, σ_n). The red curve is by depth-dependent $\xi_{\mathbf{x}}$, and the blue curve is by constant $\xi_{\mathbf{x}}$. Each curve is obtained at its own optimal parameter pair (u, σ), as shown in the legends. (a) $(10^{-1}, 4)$. (b) $(10^{-2}, 2)$. (c) $(10^{-3}, 1)$. Best viewed in color.

The proposed method achieves satisfactory performances and is insensitive to $\xi_{\mathbf{x}}$, but a slightly bigger $\xi_{\mathbf{x}}$ turns out to be more robust. On the other hand, we obtain low RMSE scores when u is around or smaller than the reciprocal of the depth range ($\leq 10^{-3}$ in the test depth videos). Although smaller u can still achieve good performance, its range tends to be narrower when noise level is increased. In practice, setting the U_f and U_b to be the reciprocal of the depth range is sufficient and convenient, since it actually means that the outliers may uniformly occur inside the depth range.

In addition, the depth-dependent noise parameter $\xi_{\mathbf{x}}$ performs superior to the constant $\xi_{\mathbf{x}}$ in dealing with type-II error. A shown in Figure 8, comparisons of the results by optimal parameter pairs (u, σ) of both cases[4] reveal that a larger constant $\xi_{\mathbf{x}}$ is required to catch severer noise presented at larger depth values due to the property of type-II error. In comparison with the depth-dependent noise, constant $\xi_{\mathbf{x}}$ might be sufficient for slightly noisy depth videos as shown in Figure 8(c), but lacks capability to catch severe noise, as shown in Figure 8(a) and (b).

2) Comparison of Synthetic Static Scenes: As some online 3D scene reconstruction methods can also successfully perform the static scene estimation in an online fashion, we numerically compared several state-of-the-art candidates, *i.e.*, the truncated signed distance function (TSDF) [33], [34]

[4] The optimal results were obtained by exhaustive search of 400 uniformly-sampled parameter pairs in the range $\sigma \in [0, 20]$ and $u \in [10^{-5}, 10^{-1}]$.

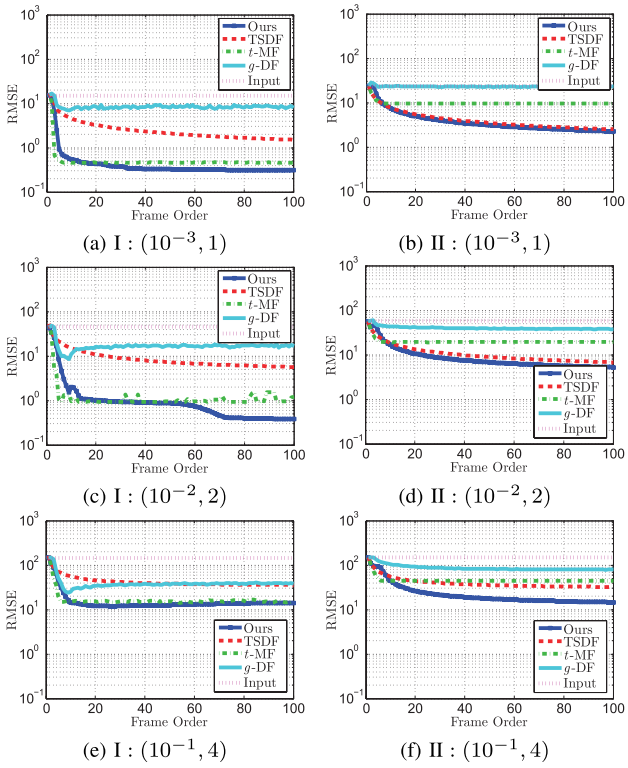


Fig. 9. Comparison with other methods on static structure estimation of the synthetic static scenes. Three levels of noise and outlier parameter pairs (ω_n, σ_n) were tested. (a), (c) and (e) were of type-I. (b), (d) and (f) were of type-II. The x-axis marks the frame order, and y-axis is the RMSE score.

in KinectFusion, the temporal median filter (t-MF) and the generative model for depth fusion (g-DF) [35], with our method. The grid number per pixel was set as 100, for both TSDF and g-DF. The temporal window size of t-MF was 5 in our experiments. As shown in Figure 9, as with all other methods, our methods tend to decrease the RMSE progressively with more frames included. However, our method is robust to the noise and outliers for both the type-I and type-II errors, and has a faster rate, *i.e.*, uses a smaller number of frames to converge and achieve a stable performance. The severer the noise is, the more superior the proposed method can be. Because TSDF is always slower to converge and g-DF suffers from quantization errors, they cannot usually achieve the same performance our method was able to achieve. In fact with a very large window size, t-MF might obtain RMSE scores lower even than those of our method, but would require more memory and will tend to be slower. Furthermore, t-MF does not provide confidence of its output as our method does. Due to the quantization artifact of g-DF, even in an optimal setting, g-DF will generally exhibit a lower performance than that of the proposed method. The occupancy grid forbids g-DF to obtain a sub-grid accuracy [35].

The per-frame running time comparison is listed in table I, where our method is comparable with t-MF. The t-MF with window size 5 has a slightly smaller computational cost, but when the window size is 10, its running time exceeds that of our method. g-DF and TSDF require much more time to process a single frame, but their performances are still not comparable to our method.

TABLE I
PER-FRAME RUNNING TIME COMPARISON (MATLAB PLATFORM)

Algorithms	t-MF (w=5/10)	g-DF	TSDF	Ours
Running time (s)	0.0188 / 0.0309	1.9186	0.6847	0.0223

B. Evaluation of the Static Structure Estimation by Real Data

To validate our algorithm with the real data, we picked several depth video sequences captured by Kinect and ToF cameras. Both static and dynamic scene were tested.

1) Static Scenes: Figure 10 shows the results of two real indoor scenes captured by Kinect. The first row shows the raw depth and color video sequences. Notice that there are severe holes presented, and fine details of the scene are susceptible to be missed or in fault depth values. Nevertheless, their corresponding color frames are always well-defined everywhere to provide enough cues to regularize the structures.

We first estimate the static structure just by raw depth frames without spatial enhancement. See the second row in Figure 10. Our method can robustly fill holes as long as sufficient depth samples in previous frames are available. In the case where only depth video is applicable, spatial enhancement is only constrained by the depth information. Even though the results are more spatially regular than those without spatial enhancement, the inpainting artifacts occur inside sufficient large holes, and edges are blurred. Furthermore, wrong measurements in the depth frames will be retained in the static structure and cannot be eliminated. As illustrated in the last row of Figure 10, spatial enhancement based on both depth and texture information produces refined static structures which are both reliable and user-acceptable. The results in green boxes show the differences between two types of spatial enhancements.

Directly employing spatial enhancement in raw depth frames cannot obtain stable results since randomly occurring holes and outliers destroy the consistency between frames and prevent the regularizing of the depth map into a temporally stable one. The static structure, in contrast, enforces the long-range temporal connection and incrementally refines the static scene. As shown in red circles in Figure 10, the missed structures cannot be inferred satisfactorily just by conventional methods, but they are refined and converged as time goes on.

The reliability of the estimated static structure (shown in Figure 11) is measured by the proportion of samples that agree with the static structure as per equation (4), which indicates that flat or smooth surfaces in the static structure are of high reliability. Simply marking unreliable pixels by $r^t_{\mathbf{x}} \leq 0.5$, many unreliable pixels are around discontinuities or occlusions. It is reasonable that measurements around such regions tend to be unreliable due to the systematic limitations of Kinect and related depth sensors. The static structure can be spatially regularized further in conjunction with the reliability map by reducing the data confidence in the unreliable region. Our reliability map is data-driven unlike those by heuristic methods [15] that need user-tuned parameters.

2) Dynamic Scenes: Our method can effectively extract the dynamic content from a static scene and further estimate and

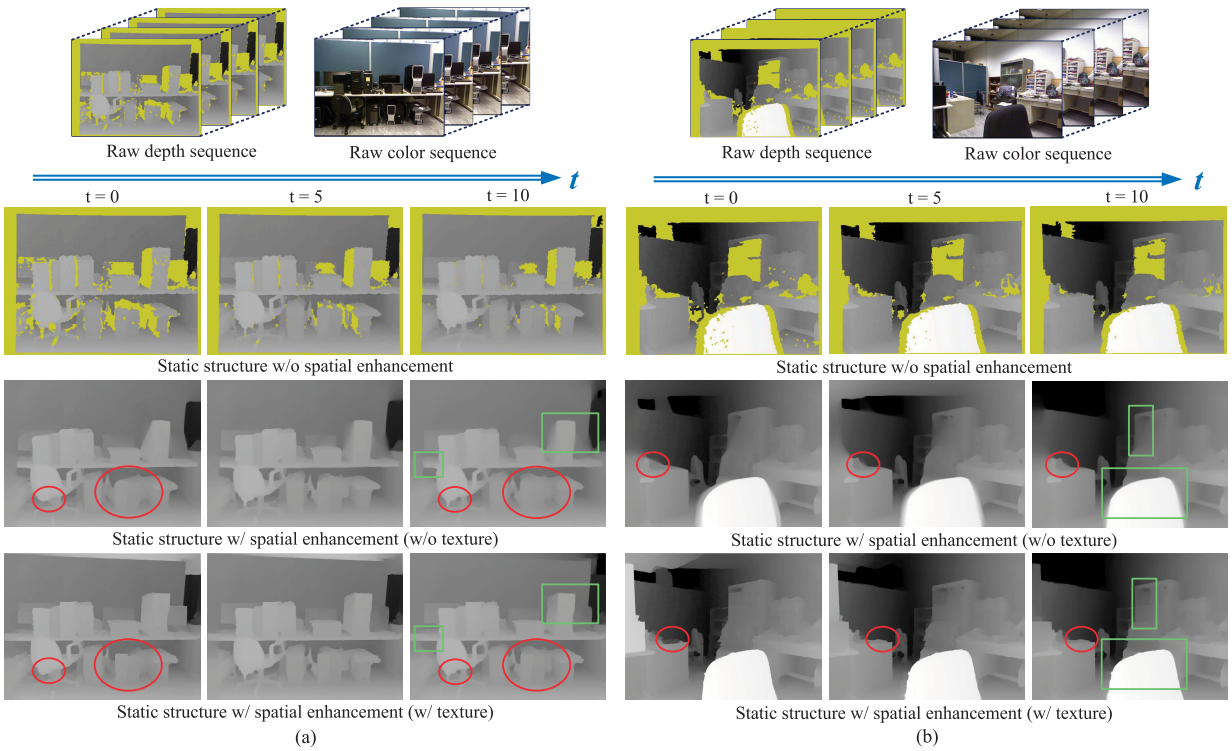


Fig. 10. Visual evaluation on real indoor static scenes. (a) and (b) are the results of two sequences `Indoor_Scene_1` and `Indoor_Scene_2`, captured from two real indoor scenes. The first row shows the raw depth sequences and color sequences. The second row is the selected results of the estimated static structures without spatial enhancement at frame $t = 0, 5, 10$ respectively. The third row shows corresponding spatially enhanced static structure without texture information, while the last row exhibits the results with the guidance of texture information. The yellow color in the second row marks missed depth values (holes). Gray represents depth value, lighter meaning a nearer distance from the camera. Best viewed in color.

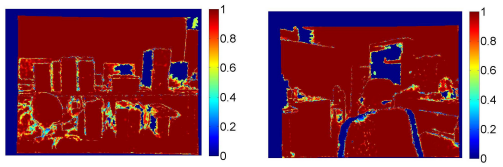


Fig. 11. Reliability maps of two test sequences of indoor static scenes. (a) `Indoor_Scene_1`. (b) `Indoor_Scene_2`. Best viewed in color.

refine the static structure in the static region. Two videos were evaluated. One was captured by Kinect, a real indoor scene with people moving around (`dyn_kinect_tl`). The second was a hand sequence by a ToF camera (`dyn_tof_tl`).

a) Kinect sequence: `dyn_kinect_tl` is a time-lapse (30×) Kinect sequence. Figure 12 shows the results of the first five frames. The parameter set for layer assignment: $w_r = 5, w_s = 10, \tau_\alpha = 16^{-2}, \tau_\gamma = 3^{-2}, \Sigma_\beta = \mathbf{I}$. Our proposed method can rapidly capture the static structure (both the depth and color) with very few frames. The artifact in Figure 12(d) is partially due to unreliable initialization, and partially because of the limited number of iterations of hole filling in the spatial enhancement. The latter one can be solved gradually after a few frames, as shown in the 3rd and 4th frames in (d). The former problem will be relieved by deleting unreliable area in the future frames according to the reliability map.

b) ToF sequence: The ToF sequence `dyn_tof_tl` [9] is time-lapse (10×) and has no color sequence embedded,

Fig. 12. Static structure estimation on `dyn_kinect_tl`. (a) and (b) are the first five frames of the input sequence. (c) shows the layer assignment results. Red, green, blue denote $l_{iss}, l_{dyn}, l_{occ}$, respectively. (d) represents the depth map of the static structure, and (e) shows the corresponding color map. The first frame is for initialization.

as shown in Figure 13. The parameter set for layer assignment: $w_r = 20, w_s = 10, \tau_\alpha = 5^{-2}, \tau_\gamma = 1^{-2}, \Sigma_\beta = \mathbf{I}$. Similar to the results from `dyn_kinect_tl`, the layer assignment can effectively exclude depth values from dynamic foregrounds ($l_\mathbf{x} = l_{dyn}$) and include those from once occluded static structures ($l_\mathbf{x} = l_{occ}$). Nevertheless, the blurs around boundaries and

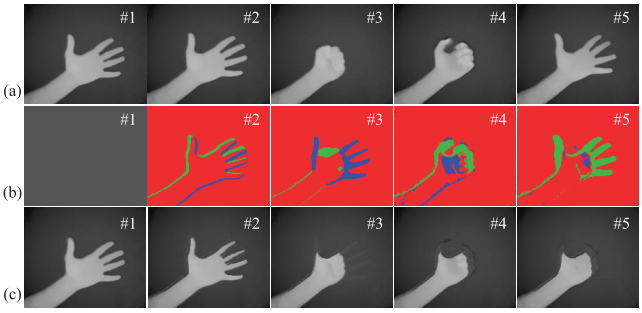


Fig. 13. Static structure estimation on `dyn_tof_t1`. (a) shows the first 5 frames of the input sequence. (b) shows the layer assignment results. Red, green, blue denote $l_{iss}, l_{dyn}, l_{occ}$, respectively. (c) represents the depth map of the static structure. The first frame is for initialization.

high noise level in the raw depth frames lead to halo artifacts in the resultant static structures at the first few frames, because in this case the layer assignment cannot definitively point out the exact boundaries between layers. Fortunately later frames provide more reliable depth samples in such regions, thus eliminating these artifacts. See the difference from the 3rd to the 5th frame in Figure 13 (c).

C. Temporally Consistent Depth Video Enhancement

Our depth video enhancement works in conjunction with the online static structure update scheme. The quality of the static structure determines the resulting performance from enhancing the tested frame spatially and temporally. Thanks to the robustness and effectiveness of our proposed method, this temporally consistent enhancement outperforms most existing representative approaches and shows comparable results with current state-of-the-art long-range temporally consistent depth video enhancement [25]. We tested several RGB-D sequences to verify our conclusion and highlight the advantages of the proposed method. These videos and their results by the proposed method and the reference approaches are available in the supplementary materials.

As shown in Figure 14, the selected frames from the sequence `dyn_kinect_1` are 113th, 133rd, 153rd, 173rd, 193rd and 213th, from left to right. Severe holes occurring in each frame are partially because of occlusion and partially due to the absorbent or reflecting materials in the captured scene. Worse still, the depth values around the boundaries of captured objects tend to be erratic. The raw depth and color frames are shown in Figure 14(a) and (b). The reference methods are the coherent spatio-temporal filtering [9] (CSTF), the weighted mode filtering [10] (WMF) and temporally consistent depth upsampling by Lang *et al.* [25]. Their parameters were set up as their default values as shown in their papers. The reference results are shown in (c), (d) and (e) of Figure 14 and the results of the proposed method are listed in Figure 14(f).

CSTF is inclined to be more blurring than the rest of the methods, especially inside the holes around the boundaries between the foreground objects and the background scene. WMF needs to quantize the depth frame into finite bins (in this experiment, 256 bins were applied), thus resulting in quantization artifacts even though it encourages sharper boundaries without blurring. Referring to any frame in Figure 14(c) and Figure 14(d), neither of these two methods can fill the depth holes with satisfactory accuracy, and the latter one performs worse in stabilizing these holes. On one hand, the reason is that they are not able to fill large holes without propagating wrong depth structure when the texture is less informative. On the other hand, the temporal consistency is enhanced only within a small temporal window, thus the structure insides the holes cannot be preserved over a long time.

A recent practical and remarkable improvement attributable to Lang *et al.* [25] is a practical long-range temporal consistency enhancement. Its results shown in Figure 14(e) present its superiority both in structure regularization as well as temporal stabilization over the previous two reference methods. Not only does the method by Lang *et al.* temporally stabilize the static objects and/or background, but also enforces the long-range temporal consistency of the dynamic objects. In comparison with it, the proposed method cannot preserve the temporal consistency inside the dynamic objects. However, the bleeding artifacts in the hole regions still cannot be eliminated immediately and are vulnerable to be propagated over the adjacent frames. Although this method is efficient in calculation thanks to the approximation solver by constant-time domain transform filtering [41], this method is globally optimized thus it often requires to store all frames into memory.

In comparison with the prior arts, the proposed method outperforms CSTF and WMF both spatially and temporally. Furthermore, it generally has a performance comparable to that of Lang *et al.*, sometimes even superior around static holes between dynamic objects and the static background, and in stabilizing the static region of each frame. Figure 14(g) compares the results of the enlarged sub-regions denoted by the red boxes in the original frames, in which our method features superior performance in regularizing these depth structures. In addition, by observing the static background behind the moving people, the proposed method offers much more stable results around regions where there were large holes, *e.g.*, the black computer cases and monitors placed on and under the white tables. It both preserves the long-range stability of the depth structure in the holes of the static region and at the same time prevents depth propagation from the dynamic objects to the static background. Meanwhile, the spatially enhanced static structure by the proposed method can incrementally refine itself by following the guidance of the corresponding color map, and gradually converges to a stable output, just as discussed in Section IV-B1.

Two additional results by the proposed method and Lang *et al.* [25] are presented in Figure 15, in which the proposed method provides comparable quality while encourages even more delicate details around the hands and heads, as well as blur-free boundaries between the human and the background, owing to the success of layer assignment in Section III-D. However, because the proposed method cannot extract a static foreground object from the static background, blurring artifacts or false depth propagation may happen around their boundaries, just as with the aforementioned

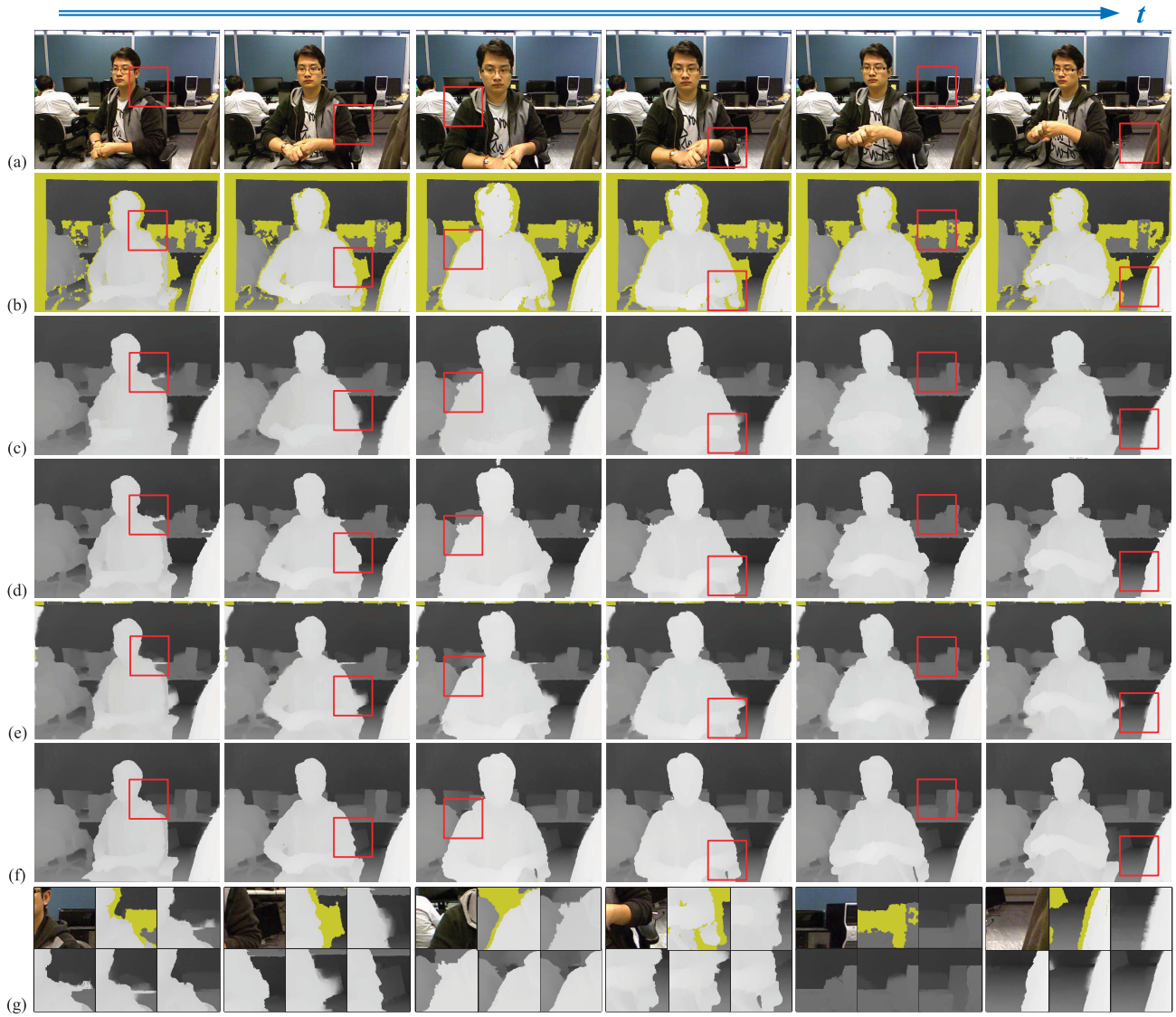


Fig. 14. Comparison on depth video enhancement. (a) and (b) are selected frames from the test RGB-D video sequences. From left to right: the 113rd, 133th, 153th, 173th, 193th and 213th frame. (c) shows the results by CSTF [9], and (d) by WMF [10]. (e) by Lang *et al.* [25] (f) is generated by the proposed method. (g) compares the performances among these methods in the enlarged sub-regions (shown in raster-scan order). Best viewed in color.

state-of-the-art method by Lang *et al.* and the filtering-based approaches like CSTF and WMF. As referring to the standing person near the background in Figure 15(b): both the proposed method and that by Lang *et al.* falsely propagated the depth values from his left arm into the computer case in the background.

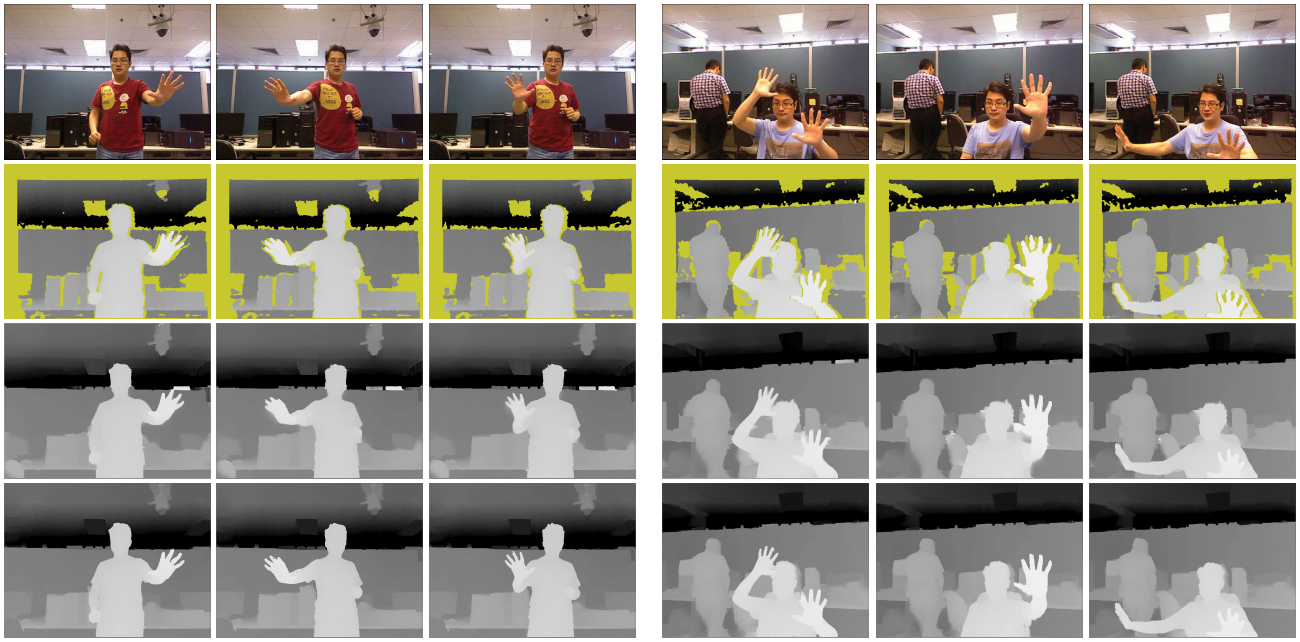
V. Limitations and Applications

A. Limitations

One limitation is that the proposed method has only been tested with indoor Kinect and ToF depth videos. To verify the reliability and generality of the proposed method, more diverse sources of depth data, *e.g.*, depth videos capturing indoor or outdoor scenes, by Kinect, ToF or laser scanners, as well as stereo vision, should be evaluated thoroughly.

For RGB-D video enhancement, the proposed method is constrained by the assumption that the static structure is "static" both in the depth and color channels. The static structure estimation may thus fail if the captured scene has varying illumination, in which case, the spatio-temporal enhancement turns into a conventional spatial enhancement approach. Another possible drawback of the proposed method is that the false estimation in the static structure cannot be eliminated if future frames cannot provide enough reliable depth samples at the same location. For example, the artifacts marked by the red dotted boxes in the enhanced depth frames (c.f. Figure 16) correspond to the holes in the input depth frames. The input depth frames cannot provide effective and reliable depth samples at these regions thus the artifacts cannot explicitly be detected by the proposed model. One possible improvement might heuristically define a threshold to delete such regions from the static structure when no reliable depth samples are received within a sufficient long time.

The proposed method only models the captured scene with dynamic and static layers, and is not capable to immediately



(a) dyn_kinect_2 (b) dyn_kinect_3

Fig. 15. Comparison of depth video enhancement. (a) and (b) are selected frames from two different RGB-D video sequences. From top to bottom: the RGB frames, the raw depth frames, results by Lang *et al.* [25] and results by the proposed method. Best viewed in color.

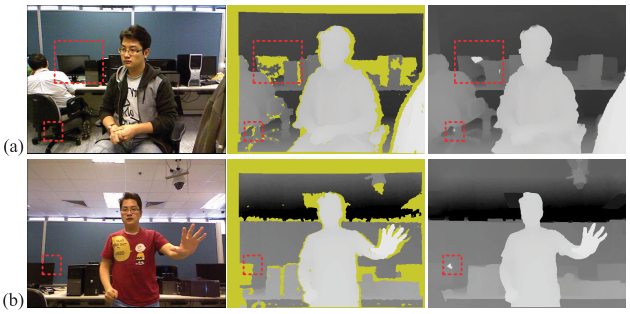


Fig. 16. Failure cases of the proposed method. (a) and (b) are two representative results. From left to right: color frame, raw depth frame and the enhanced depth frame. Artifacts are bounded by the red dot boxes.

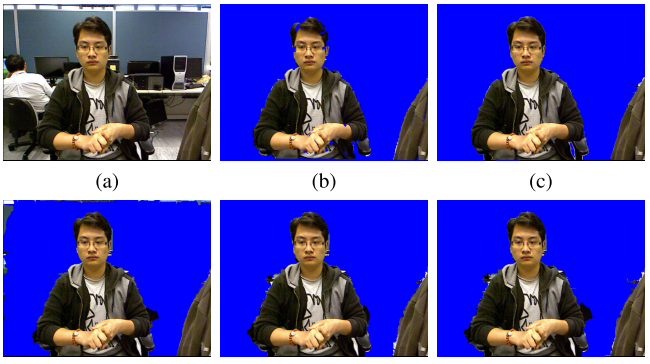


Fig. 17. Examples of the background subtraction. Best viewed in color. (a) RGB frame. (b) Raw depth frame. (c) Ours. (d) Lang *et al.* [25]. (e) CSTF [9]. (f) WMF [10].

extend to *multiple* (*e.g.*, more than 3) layers. Although it is a tough question to define and model these layers properly, we believe that more accurate results are possible by introducing such extension. For instance, the relationship between different dynamic objects can be well-defined if multiple dynamic layers compactly represent the local statistics of these objects. In this case, the spatial enhancement of each object can be handled separately and/or hierarchically, while the temporal enhancement can be adjusted to fit their distinctive motion patterns. Therefore, this meaningful extension is worthy being explored in depth as a future topic.

B. Applications

A high quality depth video improves various applications in the fields of image and graphics processing, and computer vision as well. In the following two successful applications, the enhanced depth videos by the proposed method act as an effective cue to improve their performances.

1) Background Subtraction: We can use the processed RGB-D videos to improve the segmenting of the foreground objects from the background. As shown in Figure 17, we tested one pair of RGB-D frames for background subtraction by simply extracting the region with depth values smaller than a constant threshold (in this case, we set the threshold as $1500mm$) and replacing the background by blue color. Note that there was no boundary matting applied in all the cases. The proposed method (c.f. Figure 17(c)) shows a much more refined and complete foreground segment than those by the reference methods.

2) Novel View Synthesis: A variant of novel view synthesis, named depth image-based rendering (DIBR) [45] applies the depth information to guide the warping of the texture map of one view to another synthesized view. It is a popular technique for immersive telecommunication or 3D and freeview TVs. However, the performance is hampered by the quality of the depth video. As presented in Figure 18, the novel view generated by the raw depth frame and the registered

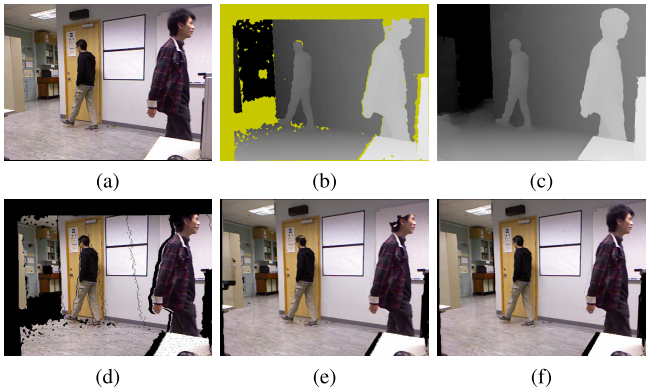


Fig. 18. Examples of the novel view synthesis. (a) and (b) are the input RGB and depth frames. (c) is the enhanced depth frame by the proposed method. (d) is the synthesized view by the raw depth frame and the RGB frame. Image holes in (d) is filled by the static structure, as shown in (e). (f) is the synthesized view based on the enhanced depth frame and the image holes are also filled by the estimated static structure. Best viewed in color.

RGB frame contains severe holes and cracks, as well as structure distortion. The static structure is appropriate to fill the image holes, but it may replace the structure of the foreground objects by mistake. The enhanced depth frame by the proposed method can preserve the depth structures well so that less structure distortion occurs in its synthesized view. Thus the synthesized view is visually plausible without apparent artifacts.

VI. Conclusion and Future Work

In this paper, we present a novel method for robust temporally consistent depth enhancement by introducing the static structure of the captured scene, which is estimated online by a probabilistic generative mixture model with efficient parameter estimation, spatial enhancement and update scheme. After segmenting the input frame with an efficient fully-connected CRF model, the dynamic region is enhanced spatially while the static region is substituted by the updated static structure so as to favor a long-range spatio-temporal enhancement. Quantitative evaluation shows the robustness of the parameters estimation on the static structure and illustrates a superior performance in comparison to various static scene estimation approaches. Qualitative evaluation demonstrates that our method operates well on various indoor scenes and two kinds of sources (Kinect and ToF camera), and proves that the proposed temporally consistent depth video enhancement works satisfactory in comparison with existing methods.

As our future work, an extension to deal with moving cameras will be a meaningful topic for study. Furthermore, we will improve the algorithm to reduce the effect of wrong estimation and design an efficient reliability check to increase the accuracy of the estimated static structure. Last but not the least, a more general probabilistic framework to handle multiple dynamic and static layers is necessary to explore for inherently increasing the performance of the proposed method.

References

[1] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Advances in Neural Information Processing Systems*, vol. 18. Cambridge, MA, USA: MIT Press, 2005, pp. 291–298.

[2] J. Yang, X. Ye, K. Li, and C. Hou, "Depth recovery using an adaptive color-guided auto-regressive model," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 158–171.

[3] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1623–1630.

[4] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. ID 96.

[5] J. Dolson, J. Baek, C. Plagemann, and S. Thrun, "Upsampling range data in dynamic environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1141–1148.

[6] B. Huhle, T. Schairer, P. Jenke, and W. Straßer, "Fusion of range and color images for denoising and resolution enhancement with a non-local filter," *Comput. Vis. Image Understand.*, vol. 114, no. 12, pp. 1336–1345, 2010.

[7] L. Sheng and K. N. Ngan, "Depth enhancement based on hybrid geometric hole filling strategy," in *Proc. 20th IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 2173–2176.

[8] J. Lu, H. Yang, D. Min, and M. N. Do, "Patch match filter: Efficient edge-aware filtering meets randomized search for fast correspondence field estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1854–1861.

[9] C. Richardt, C. Stoll, N. A. Dodgson, H.-P. Seidel, and C. Theobalt, "Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos," *Comput. Graph. Forum*, vol. 31, no. 2, pp. 247–256, May 2012.

[10] D. Min, J. Lu, and M. N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1176–1190, Mar. 2012.

[11] L. Sheng, K. N. Ngan, and S. Li, "Temporal depth video enhancement based on intrinsic static structure," in *Proc. IEEE Int. Conf. Image Process.*, Paris, France, Oct. 2014, pp. 2893–2897.

[12] H. C. Daniel, J. Kannala, L. Ladický, and J. Heikkilä, "Depth map inpainting under a second-order smoothness prior," in *Image Analysis*. Berlin, Germany: Springer-Verlag, 2013, pp. 555–566.

[13] H. C. Daniel, J. Kannala, P. Sturm, and J. Heikkilä, "A learned joint depth and intensity prior using Markov random fields," in *Proc. IEEE Int. Conf. 3D Vis. (3DV)*, Jun./Jul. 2013, pp. 17–24.

[14] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[15] F. Garcia, B. Mirbach, B. Ottersten, F. Grandidier, and A. Cuesta, "Pixel weighted average strategy for depth sensor data fusion," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 2805–2808.

[16] E. S. L. Gastal and M. M. Oliveira, "Adaptive manifolds for real-time high-dimensional filtering," *ACM Trans. Graph.*, vol. 31, no. 4, Jul. 2012, Art. ID 33.

[17] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu, "Constant time weighted median filtering for stereo matching and beyond," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 49–56.

[18] Q. Yang *et al.*, "Fusion of median and bilateral filtering for range image upsampling," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4841–4852, Dec. 2013.

[19] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[20] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," in *Proc. 17th IEEE Int. Conf. Comput. Vis.*, vol. 2. Sep. 1999, pp. 722–729.

[21] C. Vogel, K. Schindler, and S. Roth, "3D scene flow estimation with a rigid motion prior," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1291–1298.

[22] C. Vogel, K. Schindler, and S. Roth, "Piecewise rigid scene flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1377–1384.

[23] S.-Y. Kim, J.-H. Cho, A. Koschan, and M. A. Abidi, "Spatial and temporal enhancement of depth images captured by a time-of-flight depth sensor," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2358–2361.

[24] J. Zhu, L. Wang, J. Gao, and R. Yang, "Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 899–909, May 2010.

[25] M. Lang, O. Wang, T. Aydin, A. Smolic, and M. Gross, "Practical temporal consistency for image-based graphics applications," *ACM Trans. Graph.*, vol. 31, no. 4, Jul. 2012, Art. ID 34.

[26] J. Shen and S.-C. S. Cheung, "Layer depth denoising and completion for structured-light RGB-D cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1187–1194.

[27] R. Szeliski, "A multi-view approach to motion and stereo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 1999, pp. 1157–1163.
[28] P. Merrell *et al.*, "Real-time visibility-based fusion of depth maps," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[29] S. Liu and D. B. Cooper, "A complete statistical inverse ray tracing approach to multi-view stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 913–920.
[30] Y. M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun, "Multi-view image and ToF sensor fusion for dense 3D reconstruction," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops*, Sep./Oct. 2009, pp. 1542–1549.
[31] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 600–608, Aug. 2004.
[32] K. Pathak, A. Birk, J. Poppinga, and S. Schwertfeger, "3D forward sensor modeling and application to occupancy grid based sensor fusion," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct./Nov. 2007, pp. 2059–2064.
[33] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proc. 23rd ACM SIGGRAPH*, 1996, pp. 303–312.
[34] R. A. Newcombe *et al.*, "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, Oct. 2011, pp. 127–136.
[35] O. J. Woodford and G. Vogiatzis, "A generative model for online depth fusion," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 144–157.
[36] S. Thrun, "Learning occupancy grids with forward models," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, vol. 3. Oct./Nov. 2001, pp. 1676–1681.
[37] G. Vogiatzis and C. Hernández, "Video-based, real-time multi-view stereo," *Image Vis. Comput.*, vol. 29, no. 7, pp. 434–441, 2011.
[38] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.
[39] T. P. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2001.
[40] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2011, pp. 109–117.
[41] E. S. L. Gastal and M. M. Oliveira, "Domain transform for edge-aware image and video processing," *ACM Trans. Graph.*, vol. 30, no. 4, Jul. 2011, Art. ID 69.
[42] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 993–1000.
[43] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[44] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[45] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Proc. SPIE*, vol. 5921, pp. 93–104, May 2004.

Lu Sheng (S'13) received the B.E. degree in information science and electronic engineering from Zhejiang University, in 2011. He is currently pursuing the Ph.D. degree with the Image and Video Processing Laboratory, Department of Electronic Engineering, the Chinese University of Hong Kong. His current research interests include 3D image/video processing and computer vision, in particular, RGBD video enhancement, 3D reconstruction, and novel view synthesis.

King Ngi Ngan (M'79–SM'91–F'00) received the Ph.D. degree in electrical engineering from Loughborough University, Leicester, U.K. He was a Full Professor with Nanyang Technological University, Singapore, and the University of Western Australia, Australia. He has been a Chair Professor with the University of Electronic Science and Technology, Chengdu, China, under the National Thousand Talents Program, since 2012. He is currently a Chair Professor with the Department of Electronic Engineering, The Chinese University of Hong Kong. He holds honorary and visiting professorships with numerous universities in China, Australia, and South East Asia.

He has authored extensively, including three authored books, seven edited volumes, over 350 refereed technical papers, and edited seven special issues in journals. He holds 15 patents in the areas of image/video coding and communications. He was a Chair and Co-Chair of a number of prestigious international conferences on image and video processing, including the 2010 IEEE International Conference on Image Processing, and served on the advisory and technical committees of numerous professional organizations. He served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *Journal on Visual Communications and Image Representation*, the *EURASIP Journal of Signal Processing: Image Communication*, and the *Journal of Applied Signal Processing*.

Prof. Ngan is a fellow of the Institution of Engineering and Technology, U.K., and the Institution of Engineers, Australia, and the IEEE Distinguished Lecturer from 2006 to 2007.

Chern-Loon Lim (M'15) received the B.Eng., M.Eng.Sc. and Ph.D. degrees from the Department of Electrical Engineering, University of Malaya, Malaysia, in 2005, 2007, and 2013, respectively. He was a Visiting Scholar with the Chinese University of Hong Kong, Hong Kong. He is currently a Post-Doctoral Researcher with the University of Malaya. His research interests include the field of visual quality assessment, video processing, and pattern recognition.

Songnan Li (M'13) received the B.Sc. and M.Phil. degrees in computer science and technology from the Harbin Institute of Technology, China, in 2004 and 2006, respectively, and the Ph.D. degree in electronic engineering from The Chinese University of Hong Kong (CUHK), in 2012. He joined CUHK as a Research Assistant in 2007. From 2012 to 2014, he was a Post-Doctoral Fellow with the Department of Electronic Engineering, CUHK. He is currently a Research Assistant Professor with the Department of Electronic Engineering. His research interests include image and video processing, RGBD computer vision, and visual quality assessment.