BDIS: Bayesian Dense Inverse Searching Method for Real-Time Stereo Surgical Image Matching

Jingwei Song, Qiuchen Zhu, Jianyu Lin, and Maani Ghaffari, Member, IEEE

Abstract—In stereoscope-based minimally invasive surgeries (MISs), dense stereo matching plays an indispensable role in 3-D shape recovery, AR, VR, and navigation tasks. Although numerous deep neural network (DNN) approaches are proposed, the conventional prior-free approaches are still popular in the industry because of the lack of open-source annotated dataset and the limitation of the task-specific pretrained DNNs. Among the prior-free stereo matching algorithms, there is no successful real-time algorithm in none GPU environment for MIS. This article proposes the first CPU-level real-time prior-free stereo matching algorithm for general MIS tasks. We achieve an average 1.4—17 Hz on 640 × 480 images with a single-core CPU (i5-9400) for surgical images. Meanwhile, it achieves slightly better accuracy than the popular efficient large-scale stereo matching (ELAS) method. The patch-based fast disparity searching algorithm is adopted for the rectified stereo images. A coarse-to-fine Bayesian probability and a spatial Gaussian mixed model were proposed to evaluate the patch probability at different scales. An optional probability density function estimation algorithm was adopted to quantify the prediction variance. Extensive experiments demonstrated the proposed method’s capability to handle ambiguities introduced by the textureless surfaces and the photometric inconsistency from the non-Lambertian reflectance and dark illumination. The estimated probability managed to balance the confidence of the patches for stereo images at different scales. It has similar or higher accuracy and fewer outliers than the baseline ELAS in MIS, while it is 4–5 times faster.

Index Terms—Bayesian theory, posterior probability inference, stereo matching.

I. INTRODUCTION

MINIMALLY invasive surgery (MIS) technique is widely adopted in modern surgery since it mitigates postoperative infections and enables faster recovery of the patients. However, the surgeons suffer from the small field of view because the procedures are performed in a narrow space with elongated tools and without direct 3-D vision. Hence, MIS poses more difficulties to surgeons than open surgeries [1]. To overcome this challenge, stereoscopes are integrated into the operational imaging system to provide 3-D stereo images instead of single 2-D images. Moreover, the recovered 3-D shape can be further applied in applications including dense simultaneous localization and mapping (SLAM) [2], [3], AR [4], [5], and diseases diagnosis [6], [7]. Among the many off-the-shelf stereo matching methods, real-time implementation [8] is essential for tasks like surgeon-centered AR, reduction of error, decision making, or autonomous surgery’s safety boundary. We would like to emphasize that the term “real time” is task-dependent and has no clear definition of frame rate in the community. This research defines “real time” as over 10 Hz, which is enough to serve most computer-aided tasks in MIS or robotic surgery, such as visualization and surgical navigation.

The state-of-the-art stereo intraoperative tissue shape reconstruction techniques strictly follow the pin-hole camera projection model [9] and bridge the 3-D shape and 2-D stereo image. The major difference exists in the workflow for estimating the disparities from the left-right image pairs. Stereo matching algorithms can be classified as prior-free and prior-based. Prior-free approaches refer to aligning left and right images pixel-wisely using explicit handcrafted features for corner points matching (feature-metric), photometric consistency presumption for direct dense pixel searching (photo-metric) [10], or the combination of both [11], [12], [13], [14]. Differently, incorporated with the annotated disparity, prior-based methods, mostly deep neural network (DNN) based approaches, directly learn the complex “images to disparity” process with the black box model driven by the training dataset [15], [16], [17], [18], [19], [20], [21].

The prior-free and prior-based categories do not contradict. Although DNN-based methods [15], [16] are reported to be more efficient, they come with several disadvantages. First, predictions from DNN-based methods may be invalidated with changing parameters, such as focal length and baseline or a significant difference in the texture between the training and testing data [22], [23]. Moreover, the computational resources are insufficient for DNN-based methods when lacking high-end GPU or saving GPU for other tasks, such as SLAM [3],
the relationship is much more complicated than the affine scenario. Nevertheless, the affine modeling strategy cannot be applied for estimating the disparities from images recording continuous surfaces, such as the MIS scenario. However, the basic assumption of fast LK, the left-right image photometric consistency, cannot be fully satisfied in MIS. The following three factors contribute jointly to the inconsistency: textureless surface, dark region, and non-Lambertian reflectance. Unlike the indoor/outdoor scenario with abundant texture, surgical scenes contain textureless soft-tissue surfaces. Furthermore, the scope’s point source illumination leads to some dark surface regions, which exacerbates the textureless issue. The non-Lambertian reflectance brings uneven brightness to the surfaces, and it cannot be eliminated by just enforcing the patch normalization [29]. Previous research works [30], [31] demonstrated that the non-Lambertian reflectance could be modeled with an affine lighting model on the observed images. Consequently, the accuracy of the navigation system can be notably improved in the indoor scenario. Therefore, a coarse-to-fine fast LK has the potential to be adopted for estimating the disparities from images recording continuous surfaces, such as the MIS scenario. The deterministic fast LK algorithm from [10] is designed to allow the probability to cover all scales. The probabilities quantify the confidence of multiple overlapping disparities of DIS patch suffering from the textureless surface, dark region, and non-Lambertian reflectance in MIS. To overcome the three major factors mentioned above, the probability measuring module is employed for both local overlapping patch estimation fusion and outlier filtering. In general, this work has the following contributions.

1) To the best of authors’ knowledge, BDIS is the first single-core CPU stereo matching approach that achieves similar performance to the near real-time method ELAS.

2) A computationally economic probabilistic model is developed to quantify the posterior probability of the patch based on a simplified Bayesian assumption.

3) An uncertainty-aware estimation of the disparity using spatial sGMM is proposed to quantify the pixels’ confidence within the patch.

4) A coarse-to-fine probability propagation algorithm is proposed to quantify the confidence of patch parallax on all scales.

5) A maximum a posteriori (MAP) based probabilistic density function (PDF) estimation algorithm is explored, which can correctly estimate the pixel-wise variance if the assumption of Gaussian distribution is valid and fast LK converges close to global minima.

6) An open-source C++ implementation is released along with the synthesized stereo images and depths.

This article is an extension of our preliminary work [32] and contributes additionally in the following ways. Unlike the separate probability estimation in each level, this article proposes a coarse-to-fine strategy to allow the finer prediction to encode the probability of its coarser parents. Moreover, a dynamic variance selection strategy is proposed for CRFs-based probability estimation. Based on the probability within the window, a MAP-based PDF estimation is proposed to quantify the pixel-wise variance. Finally, an extensive amount of in-vivo/ex-vivo real-world experiments with reference depth1 and several ablation studies are conducted to validate the performance of the proposed method.

The rest of this article is organized as follows. Section II provides an overview of the related work. Section III covers the methodology with all the technical details. Section IV conducts experiments to validate the proposed method thoroughly. These include the qualitative and quantitative tests on the synthetic, in-vivo dataset, and ex-vivo dataset. An ablation study

---

1In the robotic community, “true value” or “reference value” obtained from hardware/software with higher precision is often termed as “ground truth.” Following [33], we name it as “reference” considering its precision inaccuracies. The remaining content, figures, and tables use “reference” instead of “ground truth.”
Stereo matching is one of the most heavily investigated topics in computer vision [34]. Trying to mimic human vision, the stereo system comprises two cameras and a computing device to triangulate the 3-D shape by estimating the parallax between the left and right images. Efficient and accurate stereo matching algorithms are essential or helpful for many tasks in MIS, such as scope navigation, AR/VR, and disease diagnosis. Regarding the theory and requirement, stereo matching algorithms can be categorized into two groups, i.e., prior-free and prior-based methods.

A. Prior-Free Methods.

The traditional prior-free methods directly estimate the parallax by addressing the similarity between the left and right image. The similarity refers to the hand-crafted features, illumination invariance, or other specially designed metrics, such as zero-mean normalized cross correlation [35]. The global methods [11], [36], [37], [38] minimize a cost function that contains a similarity data term and a smoothness term. The similarity data term enforces the illumination invariance or some similarity metrics, while the smoothness term regularizes the disparities between neighboring pixels. To save the computational resources, local approaches [39], [40], [41], [42] simplify the optimization by aggregating neighboring matching costs of the pixels. Among these approaches, the semiglobal block matching (SGBM) [11], [38] is one of the most widely applied methods in the academy and industry. SGBM constructs the matching cost volume with the range of the predefined depth. Then, it aggregates the cost with the winner-takes-all strategy and computes the disparity directly without iterative optimizations in the global methods. The recovered depth is further refined with various strategies, such as eliminating small patches and filtering low contrast pixels. The aggregation and winner-takes-all strategies ensure SGBM obtains high-quality depth without heavy computational burdens in the optimization step. ELAS [13], which deviates from these cost-volume based depth searching techniques, has also been widely used in industry [24], [25] and academy [2], [3], [26], [27]. Its procedure consists of sparse and dense steps. In the sparse matching, ELAS uses Sobel masks to conduct sparse corner matching as the supporting points set. The aligned sparse supporting points are used for Delaunay triangulation to initialize pixel-wise disparity. After the initialization, a densification step is carried out by maximizing the posterior probability defined by the photometric consistency. ELAS requires around 0.25 – 1 second with a single-core modern CPU. Thus, the real-time version needs to be implemented on the GPU end (the code on GPU is publicly available). Aiming at real-time disparity estimation, DIS [10] proposed a fast LK searching technique for stereo matching. The fast LK, combined with the coarse-to-fine strategy, estimates the parallax in real time. Nevertheless, the deterministic fast LK was not applied in the MIS community because it is vulnerable to imperfect image pairs contaminated by textureless surface, non-Lambertian reflectance, and dark illumination. Thus, this research adopts the deterministic fast LK algorithm and proposes a probabilistic formulation that is robust to these obstacles.

B. Prior-Based Methods

With the presence of high-quality training dataset, prior-based end-to-end stereo matching methods also serve the stereo matching tasks. These DNN-based methods have an advantage over traditional prior-based methods in their ability to learn more complex end-to-end searching. The delicately designed hand-crafted similarity term in traditional prior-based methods cannot fully handle the complex image to parallax process. The DNN-based methods, instead, directly learn the nonlinear relationships from the annotated training dataset. The first widely used DNN-based method GCNet [43] follows the conventional stereo matching algorithm, such as SGBM by building a 3-D cost volume based on the left and right feature map. The disparity is obtained by searching the cost volume. PSMNet [44] further improves GCNet by introducing pyramid spatial pooling and more convolution layers for cost aggregation. It is reported to have better accuracy over GCNet. Later, GwcNet [45] modifies the structure of the 3-D hourglass and introduces group-wise correlation to form a group-based 3-D cost volume. In CAS domain, Ye et al. [15] is one of the earliest research works, which adopt DNN-based stereo matching techniques in the CAS domain. PSMNet [44] and GwcNet [45] also attract the attentions from the CAS community [23]. Other stereo matching techniques [17] and [20] were also tested and recommended by the CAS community, which are capable of accomplishing the stereo matching tasks. The latest work [21] proposed a transformer-based stereoscopic depth perception algorithm.

The research of the prior-free CPU-level real-time stereo matching techniques still significantly benefits the MIS community. Despite their efficiency, the DNN-based methods require a decent amount of annotated high-quality dataset for training. First, the dataset is extremely difficult to obtain in the medical domain due to ethnic and hardware limitations. For example, the publicly available dataset with [23] adopted a structured light along with the endoscope to collect the image and shape from the porcine. The image and depth are aligned based on the kinematic information of both sensors. Moreover, to solve the synchronization problem of the two sensors, the data are collected in the porcine, which remains still during the entire procedure. Their complicated data collection procedure in porcine implies the greater difficulty in conducting the same process in the human body. Second, the limited (high-end) GPU resource also limits the real-time prediction of these approaches. Take PSMNet as an example; it costs about 4G memory and around 400 ms to predict a KITTI stereo pair, even on high-end GPUs. Besides, other tasks also involve the computation on GPU, which makes it more difficult to maintain the GPU requirement of DNN-based methods. Therefore, the prior-based and prior-free methods complement each other in different scenarios (with or without qualified training dataset). Thus, we aim at proposing...
Fig. 1. Illustrated is the framework of the proposed BDIS. It uses three scale levels as an example. The parallax of the patch is initialized with the prediction from the last level. The fast LK algorithm is applied to estimate the disparity. Then, our sGMM and CRFs algorithms are used to yield the probability of the patch in the current scale as well as the probability propagated from the last level. Finally, the predictions from the batches are fused by addressing the predicted probabilities.

the first CPU-level real-time stereo shape recovery algorithm for the MIS community.

III. METHODOLOGY

Fig. 1 demonstrates the workflow of the proposed BDIS. The left to right parallax estimation starts from the coarsest level. A fast LK algorithm is applied to estimate the initial disparity of the patch. Then, our sGMM and CRFs-based probability propagation module evaluate the appropriate pixel-wise probabilities based on the predictions from the fast LK. Next, the predictions of all overlapping patches are fused to yield the optimal probability, which is then used as an initialization for the next finer-scale processing. The fusion module strictly addresses the pixel-wise probability. The matching and fusion process is iteratively implemented until it reaches the finest level. In the probability estimation procedure, the probability in the coarser level is propagated to the finer level to account for multi-scale confidences.

A. Fast LK Algorithm

The fast LK was proposed by [10], [46]. For completeness, we briefly introduce the fast LK algorithm. In the authentic LK [28], the optimal parallax searching is realized by updating the parallax \( u_k^{(n)} \) (\( k \)th patch on scale level \( n \)) and \( n \in \mathcal{N} \) where \( \mathcal{N} \) covers all scales) by iteratively searching the optimal parallax \( \Delta u' \). This is achieved by minimizing the following objective function, which is

\[
\Delta u' = \arg \min_{\Delta u'} \sum_x \left[ I_r^{(n)}(x + u + \Delta u') - I_l^{(n)}(x) \right]^2
\]

where \( x \) is the processed center position of the patch, \( u \) is the estimated parallax at one searching stage, and it keeps being updated in each loop, \( I_l^{(n)} \) and \( I_r^{(n)} \) are the left image patch and entire right image.

Since the discrete image introduces nonlinear behavior to the system (1), the linearization of (1) is implemented on \( \Delta u' \) regarding \( I_l^{(n)}(\cdot) \) at position \( x + u \). Therefore, \( u \) is updated in each iteration, and the Jacobian and Hessian matrix need to be re-evaluated in each iteration correspondingly. Kroeger et al. [10], Baker and Matthews [46] proposed the inverse of the roles between left image patch \( I_l^{(n)}(\cdot) \) and right image \( I_r^{(n)} \).

That is switching the roles and optimizing \( \sum_x [I_r^{(n)}(x + u) - I_l^{(n)}(x - \Delta u')]^2 \). In the new formulation, the linearization is carried out on the left image patch \( I_l^{(n)} \), which can be predefined. Therefore, the linearization of the left image patch can be predetermined, and the Jacobian and Hessian only need to be calculated once in the entire optimization process.

After the patch-wise disparity searching, the predictions of the overlapping patches are fused. In DIS [10], the optimal disparity at the location \( x \) was fused by the overlapping patch predictions based on the deterministic weights. The fusion weights were determined by the normalized inverse of the left-to-right photometric residuals. The optimal disparity is

\[
\hat{u}_x^{(n)} = \sum_{k \in \Omega} \frac{1}{\max:\sum_{k \in \Omega} 1/\max(\|I_l^{(n)}(x + u_k^{(n)}) - I_r^{(n)}(x)\|^2, 1)}u_k^{(n)}
\]

where \( \Omega \) is the set of patches covering the position \( x \). The pixel-wise disparity \( \hat{u}_x^{(n)} \) is the weighted average of the estimated disparities from all patches, wherein the weight is the inverse residual of brightness.
Ψ(\(P\)) is a potential function parameterized by \(P\) \(\in \mathcal{P}\), and \(\theta\) \(\in \mathbb{R}^d\). In general, \(P\) is a small window adopted and \(\theta\) is chosen to be small, but the stereo matching on the textureless surface (the right one) is much less reliable. Furthermore, the photometric consistency presumption is also seriously violated on the surface affected heavily by the non-Lambertian reflectance. Ambiguities arise when the photo-consistency assumption is violated, e.g., at intensive reflection or very dark pixels. Even the Lambertian reflectance pollutes the photometric consistency because the reflected intensity is related to the incident angle [29]. In the large-scale environment, affine lighting formulation [30], [31] was enforced by the SLAM systems and reported to handle the photometric inconsistency well. The modeling enforces extra affine modeling of the illumination of the target image. Nevertheless, the affine modeling cannot fully tackle the complex and severe non-Lambertian reflectance in MIS. Therefore, the weights from (2) are misleading. This article seeks the CRFs to depict the confidence of the patch’s prediction.

CRFs is a sequential modeling technique that presents transitional probabilities between finite states based on a well-defined distribution over observations. CRFs formulates the joint probability of the states using a single exponential formulation instead of perstate models for simplification [47]. Given state \(s\) and target \(t\), CRFs is expressed as a Boltzmann distribution [48], which is a function to measure the probability of the state as “state’s energy.” The probability is represented in an exponential form [49] as

\[
p(t \mid s, \theta) = \frac{\exp(\Psi(t, s, h ; \theta))}{\sum_{t, h} \exp(\Psi(t, s, h ; \theta))}
\]

where \(\Psi(\cdot)\) is a potential function parameterized by \(\theta\), while \(h\) is a set of hidden variables that cannot be directly observed [50]. Specifically, the distant observations from the target employ low potential energy in our scenario. Hence, a proper potential function follows the condition that its output is inversely relevant to the distance metric.

Since there is no prior knowledge of the uncertainty distribution of the left image patch and right image, it is difficult to infer the posterior probability in terms of disparity directly. Inspired by CRFs, we choose to implicitly infer the probability with Bayesian modeling using CRFs [51]. The posterior probability of the patch-wise disparity \(u_k^{(n)}\) on level \(n\) is

\[
p(u_k^{(n)} \mid I_l^{(n)}, I_r^{(n)}) \propto \frac{p(I_l^{(n)} \mid I_r^{(n)}, u_k^{(n)})}{p(I_r^{(n)} \mid I_l^{(n)}, u_k^{(n)})} \propto \frac{p(I_r^{(n)} \mid I_l^{(n)}, u_k^{(n)})}{\sum_{u_k^{(n)} \in \mathcal{P}} p(I_r^{(n)} \mid I_l^{(n)}, u_k^{(n)})} \tag{4}
\]

where \(\mathcal{P}\) is the window of all possible choices of \(u_k^{(n)}\). In general, naturally picking \(u_k^{(n)}\) among \(\mathcal{P}\) can be regarded as an event of equal probability. Hence, the chance of each \(u_k^{(n)}\) to be chosen is a fixed constant. Denote \(p(u_k^{(n)} \mid I_l^{(n)}, I_r^{(n)})\) as the probability calculated in level \(n\). Specifically, one pixel on level \(n\) represents a \(2^n \times 2^n\) pixel patch on the original image. The computation of (4) is time-consuming, and we approximate it with a small window as

\[
p(u_k^{(n)} \mid I_l^{(n)}, I_r^{(n)}) \propto \frac{p(I_r^{(n)} \mid I_l^{(n)}, u_k^{(n)})}{\mathcal{P}'} \propto \frac{p(I_r^{(n)} \mid I_l^{(n)}, u_k^{(n)})}{\mathcal{P}' + \mathcal{P}''} \tag{5}
\]

where \(\mathcal{P}' = \sum_{u_k^{(n)} \in \mathcal{P}'} p(I_r^{(n)} \mid I_l^{(n)}, u_k^{(n)})\) and \(\mathcal{P}'' = \sum_{u_k^{(n)} \in \mathcal{P}''} p(I_r^{(n)} \mid I_l^{(n)}, u_k^{(n)})\). \(\mathcal{P}'\) is the small window adopted and \(\mathcal{P}''\) is the rest window (\(\mathcal{P} = \mathcal{P}' \cup \mathcal{P}'\)). (5) simplifies \(\mathcal{P}'\) to a small window \(\mathcal{P}'\) considering the rest candidates are numerically trivial. The compensation ratio \(r = \frac{\mathcal{P}'}{\mathcal{P} + \mathcal{P}'}\) and is close to 1. Since probability is assumed to follow Gaussian distribution, the marginal patches in \(\mathcal{P}''\) are numerically small. Our experiment shows that \(r\) ranges from 0.99 – 1 and thus can be regarded as constant.

Equation (5) reveals that the posterior probability of the disparity prediction can be retrieved by traversing the prior probability on all possible positions \(u_k^{(n)}\) in the window \(\mathcal{P}'\) with size \(s\). In addition to CRFs, we present another straightforward explanation. The direct posterior probability measurement, such as the inverse residual (2), is unavoidable to suffer from the textureless surface, non-Lambertian reflectance, and dark illumination. In contrast, the posterior probability implicitly inferred from the prior probability is robust to these factors because the impact of these issues is consistent on all patches and can be compensated on the prior probabilities within the small window. Therefore, we model the prior probability \(p(I_r^{(n)} \mid I_l^{(n)}, u_k^{(n)})\) based on the Boltzmann distribution, which is

\[
p(I_r^{(n)} \mid I_l^{(n)}, u_k^{(n)}) = \exp \left(- \frac{\|I_r^{(n)}(\mathbf{x}) - I_l^{(n)}(\mathbf{x} + u_k^{(n)})\|^2}{2\sigma_r^{(n)} + 2\sigma_i^{(n)}} \right) \tag{6}
\]
where scale $\sigma_i^{(n)}$ is the hyperparameter. Different from our preliminary work [32], which uses an arbitrary value, $\sigma_i^{(n)}$ is set as the standard deviation of $||I_i^{(n)}(x) - I_i^{(n)}(x + u_i^{(n)})||_2^2$. This dynamic parameter ensures that the scale is reasonable within the window. Results show it helps improve the accuracy and is absent of parameter tuning. Therefore, the relative posterior probability can be obtained with (5) and (6).

Fig. 2 elaborates on the benefit of the implicit inference of the posterior probability. For the well-textured and well-illuminated image pairs, contrary to the deterministic inverse residual weight (2), the residual has strong responses (large inverse residual) at the correct parallax while the small response in another parallax. However, the response curve has smaller contrast for the textureless/dark/highlighted image pairs, and the weight from (2) is always large in the residual. Therefore, (2) assigns the same confidence to both situations. Equation (5) solves this problem with the implicit inference and concludes that the former situation has much higher confidence than the latter. Moreover, CRFs also tests the local convergence to filter saddle points. The predictions identified as saddle points are discarded. It should be noticed that, theoretically, this principle does not apply to small-scale disturbances identify the small highlighted patch as well-textured. The small indistinguishable highlighted patch is mixed with texture. This issue cannot be handled unless some handcrafted threshold or prior-based method is applied. Nevertheless, we can hardly identify this issue in the experiments.

C. sGMM-Based Spatial Gaussian Probability

On top of the posterior probability quantification for the entire patch in Section III-B, we propose a probabilistic patch kernel using spatial sGMM [52] to estimate pixel-wise probability within the patch.

Theoretically, any complex distribution can be decomposed and fit with a set of simple Gaussian distributions. The target distribution can be represented as a Gaussian mixture, which is the superposition of several Gaussian kernels. Specifically, the GMM is a probability density function in the weighted sum of Gaussian processes [52] as

$$p(x | \lambda) = \sum_i w_i g(x | \mu_i, \sigma_i)$$

where $w_i$ is the weight of the kernel $g(x | \mu_i, \sigma_i)$. For a balanced observation, all the $w_i$ can be set as the same value.

GMM consists of various Gaussian kernels with diverse parameters. A unified form of each Gaussian density kernel [53] can be expressed as the Gaussian kernel

$$g(x | \mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left( -\frac{(x - \mu_i)^2}{2\sigma_i^2} \right)$$

where $\mu_i$ and $\sigma_i$ are the mean and variance in the kernel.

The authentic GMM is simplified as sGMM, which directly uses the prior knowledge of the parameters. Jiang and Leung [54] simplified GMM by assigning $\sigma_i$ with a constant and only estimated the nonexponential coefficient of Gaussian kernels. Differently, Ge and Fan [55] simplified GMM by fixing the weights and only estimate a unified $\sigma$ to all $\sigma_i$. We push them further by treating $\sigma$ as a hyperparameter while setting all weights as 1. Thus, to avoid misunderstanding, we term it sGMM.

In MIS, the images collected are natural and fit the presumption in sGMM. Hence, a multivariate Gaussian distribution is adopted to measure the confidence of the pixel-wise probability using a Gaussian mask similar to (7). Considering that the confidence is measured patch-wisely, the patch’s center accumulates a higher probability than the marginal pixels since the multivariate Gaussian distribution is enforced on the patch. The central pixels preserve more information. Assuming all pixels in the patch are independent identically distributed, we have

$$p(u_k^{(n)} | I_l^{(n)}, I_r^{(n)}, x) \propto \sum_{x_i \in \xi} \exp \left( -\frac{||x - x_i||_2^2}{2\sigma_i^2} \right)$$

where $\xi^{(k)}(x)$ is the set of all pixel positions within the patch $k$ in the image coordinate. $\sigma_i$ is the 2-D spatial variance of the probability in (9). Note that (9) is independent of the patch and can therefore be precomputed before the process. Combining (5), (6), and (9), the final pixel-wise posterior probability distribution on level $n$ can be represented as

$$p(u_k^{(n)} | I_l^{(n)}, I_r^{(n)}, x) \propto \sum_{x_i \in \xi} \exp \left( -\frac{||x - x_i||_2^2}{2\sigma_i^2} \right) \frac{\exp \left( -\frac{||I^{(n)}_l(x) - I^{(n)}_r(x + u_i^{(n)})||_2^2}{2\sigma_i^2} \right)}{\sum_{u_i^{(n)} \in \mathcal{U}} \exp \left( -\frac{||I^{(n)}_l(x) - I^{(n)}_r(x + u_i^{(n)})||_2^2}{2\sigma_i^2} \right)}. \tag{10}$$

To avoid misunderstanding, we should emphasize that (6) and (9) are not the cost functions but probability functions for each patch or pixel. The probability quantifies the weight of the fusion process. Our probabilistic Bayesian method does not require heavy computational costs in the optimization.

D. Probability Propagation From Coarse-to-Fine

In addition to the Bayesian probability in (5) at one level, the measured probabilities on the coarser levels should also be propagated to this level since the probability should describe multiscale observations. The probability on the finest level should count the probabilities of all scales. However, as the estimated disparity from the coarser level is only used as initialization for the finer level inverse searching, the exact form of probability propagation cannot be accurately obtained. We approximate it by treating the disparity estimation on each level as independent processes with different weights. This is reasonable because the inverse searching on each level is bounded, and the coarsest prediction does not deviate much from the final prediction. Therefore, all scales contribute to the searching of the finest disparity unevenly.

We approximate it with an empirical formulation. Define $\Gamma(p(u_k^{(n)} | I_l^{(n)}, I_r^{(n)}))$ as the probability contributed from level
n. The probability which considers coarser levels can be obtained as
\[ p(u_k^{(n)}|I_l^{(n)}, I_r^{(n)}, x) \propto \sum_{n \in N} \Gamma(p(u_k^{(n)}|I_l^{(n)}, I_r^{(n)}, x)). \] (11)

In this article, \( \Gamma(p(u_k^{(n)}|I_l^{(n)}, I_r^{(n)})) \) is approximated as the following because the finer scale is 2 times larger in one dimension, which is
\[ \Gamma(p(u_k^{(n)}|I_l^{(n)}, I_r^{(n)}, x)) = \frac{2^n}{\sum_{n \in N} 2^n p(u_k^{(n)}|I_l^{(n)}, I_r^{(n)}, x)}. \] (12)

It is worth noting that any successful observation at an arbitrary level can lead to an effective observation from global perspective. Hence, the formulated probability is in the form of superposition to represent such a logical “or” relationship.

The pixel-wise weighting in (2) is substituted with
\[ \tilde{u}_k^{(n)} = \sum_{k \in \Omega} \Gamma(p(u_k^{(n)}|I_l^{(n)}, I_r^{(n)}, x))u_k^{(n)}. \] (13)

Based on the fused disparity in the finest scale \( \tilde{u}_x^{(f)} \) (\( f \) is the finest scale), the depth can be obtained following the routine stereo vision process, which is
\[ d_x = \frac{fb}{\tilde{u}_x^{(f)}} \] (14)
where \( f \) and \( b \) are the focal length and baseline.

### E. PDF Estimation of the Depth

This section provides an optional inaccurate depth variance estimation method in addition to the workflow, as shown in Fig. 1. Although (5) obtains the patch’s posterior probability \( u_k \), the parameter of PDF, specifically the variance, is unknown. With the Gaussian noise assumption, confidence quantification plays an important role in 3-D geometry and sensor fusion. Assuming the errors are predominated by Gaussian distribution, we push (5) toward the estimation of the PDF. PDF estimation technique [56] is employed and modified. Denote the parameter set as \( \theta_k = \{ u_k, \sigma_k \} \), where \( \sigma_k \) is the standard deviation of the patch \( k \). Following the authentic definition of the MAP, \( \theta_k \) can be obtained as
\[ \theta_k = \arg \max_{\theta_k} p(\theta_k | P) = \arg \max_{\theta_k} \frac{p(P | \theta_k)p(\theta_k)}{p(P)} \]
\[ = \arg \max_{\theta_k} p(P | \theta_k)p(\theta_k) \]
\[ = \arg \max_{\theta_k} \sum_{i \in P} \ln p(P | \theta_k) + \ln p(\theta_k) \] (15)
where \( p(\theta_k) \) is the prior distribution of the parameter \( \theta_k \) and \( P \) is the full window of samples in (4). \( \ln p(P | \theta_k) \) is equivalent to the loss of fitting \( p(I_l^{(f)}|I_l^{(f)}, u_k^{(f)}) \) with the PDF’s optimal parameter \( \theta_k \). \( P \) is substituted with \( P' \) for faster computation. The MAP formulation (15) can be applied in BDIS after three modifications. First, the prior distribution of \( u_k \) (mean) is fixed at the fast LK’s estimation \( u_k^{(f)} \) while the variance is unknown. Second, the prior knowledge of the distribution of \( \sigma_k^{(f)} \) (standard deviation on the finest level) is unknown, and we do not enforce additional prior knowledge terms. Lastly, (5) indicates an approximate ratio \( r(f) \), an extra value \( c_k \) helps compensating the impact of inaccuracy resulted from the small processing window \( P' \). Applying the three modifications in the MAP process, we have
\[ \min_{\sigma_k^{(f)}, \sigma_k \in \Omega} \sum_{i \in P'} c_k \sigma_k^{(f)} \sqrt{2\pi} \exp \left( -\frac{(u_k^{(f)} - u_k^{(f)})^2}{2\sigma_k^{(f)^2}} \right) - p(I_r^{(f)}|I_l^{(f)}, u_k^{(f)}). \] (16)

To enable efficient optimization of (16), natural logarithm function \( \ln(\cdot) \) is applied
\[ \min_{\sigma_k^{(f)}, \sigma_k \in \Omega} \ln c_k - \frac{1}{2} \ln \left( \sigma_k^{(f)^2} \right) - \frac{(u_k^{(f)} - u_k^{(f)})^2}{2\sigma_k^{(f)^2}} \]
\[ - \ln p(I_r^{(f)}|I_l^{(f)}, u_k^{(f)}). \] (17)

Gauss–Newton (GN) algorithm is adopted to solve (17). The linearized system is solved with Cholesky decomposition for fast solving. The initial \( \sigma_k^{(f)} \) and \( c_k \) are set as \( \sqrt{0.1} \) and 1. To reveal the reliability of GN in solving (17), a massive amount of Monte Carlo experiments were conducted. The experiment indicates (17) has a local minimum if \( \sigma_k^{(f)} \) is set much larger than the reference. Thus, the initial \( \sigma_k^{(f)} \) is set as \( \sqrt{0.1} \), which is enough because the variance in practice should be much larger than \( \sqrt{0.1} \). Meanwhile, the convergence of (17) with GN is very fast. Even though five iterations are enough in our experiments, we fix the optimization iteration to 10. The extremely small size of the state vector and Jacobian matrix makes the computation small.

Additionally, (17) needs to be handled with two exceptions. First, the residual of (17) is too large. The threshold for residual in (17) is set as 0.1. Second, the unconvergence in the minimization process. For example, all probabilities are almost equal, or the probability at \( u_k^{(f)} \) is not the largest. Both situations hint the inconsistency of the Gaussian distribution assumption. Therefore, we compromise this paradox by manually setting the detected patch with a large standard deviation which alleviates the issue. Finally, it is worth clarifying that \( \sigma_x \) is assigned as the patch searching threshold in the fast LK process. This is reasonable because the patch’s probability (and weight) is small in both situations. It is worth emphasizing again that although some PDFs are inconsistent with the Gaussian distribution presumption, the proportional probability (10) is still valid.

The patch-wise variance \( \sigma_k^{(f)^2} \) is propagated to pixel-wise depth \( \sigma_x^2 \) (at position \( x \)) following weights defined in (10) and the disparity to depth process defined in (13). The propagation function is
\[ \sigma_x^2 = \frac{\left( \frac{fb}{\tilde{u}_x^{(f)}} \right)^2}{u_k^{(f)}} \left( \sum_{k \in \Omega} \Gamma(p(u_k^{(f)}|I_l^{(f)}, I_r^{(f)}, x)) \sigma_k^{(f)^2} \right). \] (18)
of the type of distribution. The PDF estimation, however, requires the consistency of the Gaussian distribution presumption of the error. There is no guarantee for consistency. In many datasets, we do notice a large amount of probability (6) within the window perfectly follows the Gaussian curve. Furthermore, multiscale fast LK suffers from local minima or bad convergence, which can be categorized as epistemic uncertainty and cannot be evaluated. Thus, our uncertainty estimation method is optional and for reference only.

F. Technical Details

This section covers important technical details in the implementation. First, in the fast LK (1) process of some patches, the algorithm stops before the convergence due to reaching the maximum iteration limit. Consequently, the un converged parallax prediction is not the minimum in the CRFs process (5). Once detected, the un converged predictions are discarded. Second, some image patch has invalid pixels caused by image padding, undistortion, or rectification. We follow ELAS [13] and use threshold γ to filter small patches without enough valid predicted pixels in the final step. Third, the exact exponential computation consumes massive computational resources. Schraudolph [57] adopted the practical approximation method, which takes advantage of the manipulation of a standard floating point representation. Our experiment shows the approximated exponential computation only takes less than 1/5 computational time of the original exponential function in the standard C++ library. Schraudolph [57] reported that the error is bounded for less than 4% which is acceptable in our case.

Finally, we would like to clarify that the Bayesian probability quantification method in Sections III-B, III-C, III-D is independent of the choice of the base method or even DNN. It is coupled with fast LK because we aim at CPU-level real-time stereo matching algorithm while preserving high accuracy. It has the potential to be applied to other stereo matching algorithms, including the DNN-based methods.

IV. RESULTS AND DISCUSSION

The efficiency of BDIS was validated from the perspectives of accuracy and time consumption. This section first introduces the datasets. Next, the accuracy comparisons on both the synthetic and in-vivo datasets are presented. The following content summarizes the time consumption of all the baseline methods. Lastly, an ablation study was conducted to discuss the contributions of different strategies and modules.

A. Baseline Algorithms, Datasets, and Metrics

Prior-free methods DIS [10], SGBM [11], and ELAS [13] were adopted as the baseline methods to be compared with BDIS in the in-vivo and synthetic datasets experiments. Following the recommendations from [23], the DNN-based methods PSM-Net [44] and GwcNet [45] were also employed for comparison. The computation platform was a commercial desktop with CPU i5-9400 and GPU GTX 1080ti. DIS and SGBM are open-sourced and provided from OpenCV (C++ version). The code of ELAS is also publicly available. We implemented BDIS in C++ based on the code from [10]. All the prior-free methods were compiled and run on the CPU end (single). The two open-sourced DNN-based methods were implemented on PyTorch [58] and run on the GPU end.

A virtual surgical dataset was synthesized to control factors, including the intrinsic parameters, reference depth, and calibration. The synthesized dataset ensures absolute accuracy in camera parameter calibration, reference depth retrieval, and image rectification and undistortion. Our synthetic dataset was generated from an off-the-shelf virtual phantom of a male’s digestive system. The stereoscope was placed inside the colon, and we implemented a module to enable it to move automatically in the colon. The 3-D game engine Unity3D³ was adopted to generate the sequential RGB stereo and depth images, which strictly follows the pin-hole camera projection model. The size of all images is 640 × 480. The number of stereo image pairs collected from the colon is 310 (300 was collected for training DNN-based methods only). For better comparison, both diffuse lighting and non-Lambertian reflectance were simulated. It should be noticed that the dark region near the edges is in low illumination, not pure black. Sample images are presented in Fig. 3. Details of the training dataset (DNN-based methods only) will be described in each experiment section.

Moreover, two in-vivo and one ex-vivo datasets were adopted for comparison. The first in-vivo dataset comes from the public Hamlyn in-vivo stereo videos [59]. We adopted 200 image pairs with size 640 × 480 and 200 image pairs with size 288 × 360. The Hamlyn dataset does not contain the reference depth. Sample images can be found in Fig. 4. The second in-vivo dataset (we name it SCARED) is from stereo correspondence and reconstruction of endoscopic data subchallenge, which was organized during the Endovis challenge at the conference MICCAI 2019 [23] (samples images are shown in Fig. 5). This dataset consists of the stereo images with the annotated depth captured using a da Vinci Xi surgical robot. The size of the images is 1280 × 720. It has 7 training datasets and 2 testing datasets (5 keyframes for each dataset). Only five training datasets and two testing datasets were used because the rest 2 are associated with incorrect calibration parameters. Keyframe 4 in dataset 1 is also incorrect. All the in-vivo stereo images (34 overall) were rectified, undistorted, calibrated, and vertically aligned with the provided intrinsic and extrinsic parameters. The ex-vivo dataset SERV-CT [33] was also obtained from da Vinci Surgical System. SERV-CT contains 16 stereo endoscopic image pairs with reference anatomical segmentation derived from CT. Two different ex-vivo porcine samples were imaged using the straight and 30° endoscopes. A CT scan provided the reference from the O-arm system. The scan contains both the anatomy and the endoscope, facilitating constrained manual alignment to provide the pose of the viewed anatomical surface relative to the endoscope. As sample calibrated left images in Fig. 6 indicate, the retrieved images in SERV-CT are less textured than SCARED dataset.

²Readers are encouraged to watch the attached video. The code also provides visualization.

³https://unity.com/
We used the following parameter setting: Coarsest scale was $2^5$, and the finest scale was $2^1$. Min/max iterations 12, early stopping parameters 0.05 0.95 0.10, patch size 10, patch overlap 0.55, and no left to right consistency check was enforced; Patch mean normalization was enforced to reduce the impact of illumination; No M-estimator was used. $\gamma$ was set to 0.75 for $640 \times 480$ and 0.25 for $288 \times 360$ data to discard the patch without enough valid pixels. The sampling within one Bayesian window was 5; the disturbance from the convergence was 0.5 and 1 pixel; The minimal ratio of the valid patch was set to 0.75; $\sigma_s$ was set to 4. Pixel-wise threshold 0.15 is enforced to filter unreliable predictions. Regarding the preliminary work [32], most parameters were the same except patch overlap set 0.55 instead of 0.6, allowing almost 2 times faster speed.

For each pixel $x$ with estimated depth $d_x$, define the absolute depth error as

$$e_x = \|d_x - \hat{d}_x\|$$
where $d(x)$ is the reference depth at pixel $x$. For each image, we count the mean and median errors in the entire image. For the dataset with multiple images, the metrics “average mean error” and “average median error” refer to the average of mean and median errors in the dataset.

The numbers of predicted pixels were counted as “validity” or “valid pixels” in the experiments. For robustness, some predicted pixels’ disparity are ignored by ELAS, SGBM, and BDIS. It is reasonable to discard relatively small regions that are believed to be not confident.

### B. Comparisons on the Synthetic Dataset

The proposed BDIS was first compared with the prior-free methods ELAS, SGBM, DIS, and our preliminary work [32] on the synthesized dataset. Moreover, DNN-based methods PSMNet [44] and GwcNet [45] were also tested following the recommendation of Allan et al. [23]. They were first trained on the 300 additionally synthesized training image pairs in the virtual colon. The 300 labeled pairs were employed to fine-tune the pretrained network, which was obtained by training on the synthesized dataset from the beginning. Unfortunately, we could not successfully train the DNN based on the generated datasets.

The performance of DNN-based methods is much inferior in the synthetic dataset test (see Table I) than in the following in-vivo dataset. The DNN’s performances in Table I deviate from the conclusions in [23]. It also contradicts to the following in-vivo test (see Section IV-C). The failure should be credited to insufficient parameter tuning. Specifically, there is a large texture gap between the synthetic training dataset and the dataset adopted in training the pretrained DNN model. The pretrained model is obtained from the city-scape dataset Kitti2015 [60], [61]. The texture of the synthetic dataset was simulated from the game engine and bereft of many imaging details and cues. That is to say, the synthetic dataset is not the nature image as the Kitti2015. Previous research works [63], [64], [65] suggest that the performance of the convolution neural network is heavily dependent on the texture of the images. This phenomenon also explains the massive effort devoted to bridging the domain gap between synthetic data and real-world data in the autonomous driving community. In our case, the pretrained weights of kernels contribute poorly to the convergence of the model. Thus, instead of fine-tuning the pretrained model, the DNN-based models should be trained on the synthetic dataset from the beginning. Unfortunately, we could not successfully train the DNN based on the generated datasets for the two DNN-based methods. The quantitative real-world

|                | Diffuse light | Non-Lambertian reflectance |
|----------------|---------------|----------------------------|
| ELAS           | 0.178         | 0.161                      |
| SGBM           | 0.512         | 0.616                      |
| DIS            | 0.251         | 0.542                      |
| GwcNet         | 0.417         | 0.616                      |
| PSMNet         | 0.161         | 0.542                      |
| BDIS           | 0.157         | 0.338                      |
| BDIS           | 1.211         | 0.338                      |
| ELAS           | 0.061         | 0.142                      |
| SGBM           | 0.528         | 0.221                      |
| DIS            | 0.248         | 0.702                      |
| GwcNet         | 0.475         | 0.850                      |
| PSMNet         | 0.702         | 0.316                      |
| BDIS           | 0.221         | 0.106                      |

Results of the prior-free methods ELAS, SGBM, DIS, and the proposed BDIS are presented. Prior-based methods GwcNet and PSMNet are also shown. “Median” refers to the median error. “Mean” refers to mean error. “Validity” is the number of valid predicted pixels. The error is presented in mm. The number of valid pixels is 1000. BDIS refers to our preliminary work [32]. “Time” refers to time consumption measured in seconds. The bold entities represent the best performances.
C. Quantitative Comparisons on the In-Vivo/Ex-Vivo Dataset

On top of the qualitative comparison, we validated the performance of BDIS on the real-world SCARED (in-vivo) and SERV-CT (ex-vivo) stereo datasets. For the SCARED dataset, two separate experiments were conducted. In the first experiment, following the recommendation of Allan et al. [23], only the last ten image pairs (images in datasets 8 and 9) were tested since the DNN-based methods require the rest images (24 stereo image pairs) for fine-tuning the network. Among the 24 stereo image pairs, 18 were adopted to fine-tune PSMNet and GwcNet, while the validation was done on the other six stereo image pairs. The fine-tuned networks were further applied to predict the rest 10 testing datasets. Fig. 8 shows sample reconstructions of the baseline methods. The second experiment is for prior-free methods only; since the prior-free approaches are free of training data, all available (34 totally) stereo image pairs were used for more comprehensive comparisons.

The frame-wise mean absolute depth errors between the state-of-the-art approaches and the proposed BDIS on the SCARED stereo dataset are demonstrated in Table II. The last ten stereo image pairs of the testing keyframes designated by Allan et al. [23] are individually presented. It also summarizes the median error, average error, and the number of valid depth pixel predictions of these approaches. The results indicate that BDIS is 6.33% more accurate than ELAS regarding the average median error. Compared with ELAS, the proposed BDIS improves the performance on the average median error by 0.44%. Nevertheless, BDIS has 5.67% more valid predictions than ELAS. Moreover, Table III indicates that SGBM can predict similar depth as BDIS and ELAS, but it is not robust. It shows that the trained GwcNet is comparable to BDIS and ELAS. The results of GwcNet validate our assumption (see Section IV-B) and conclusions from the research works [63], [64], [65] that the performance of DNN is dependent on the texture similarity between the training and testing images. The images in the SCARED stereo dataset are natural images, and their textures are much more abundant than the synthetic dataset (results in Section IV-B). However, it should be emphasized that the accuracy of the DNN-based method is heavily dependent on the quality of the training dataset. The comparisons of the prior-free and the DNN-based methods are for reference only. Generally, it confirms our claim that BDIS and ELAS have similar accuracy and predict the same amount of valid depth.

Table III shows the comparisons of the prior-free methods ELAS, SGBM, DIS, and BDIS. The frame-wise mean and median errors of all keyframes are summarized. In general, the accuracy of ELAS is 7.42% higher in average mean error over BDIS, while BDIS is 7.55% better regarding the average median error. Moreover, the valid depth pixels of BDIS is 5.93% more than ELAS. This result is consistent with other quantitative experiments in this article.
Fig. 8. Sample reconstructions of the SCARED stereo dataset. DXY means the keyframe Y in dataset X. The error map (in mm) of the last column is the error of BDIS. Readers are encouraged to refer to the attached video to appreciate the recovered shapes.

TABLE II
MEAN ABSOLUTE DEPTH ERROR COMPARISONS ON THE SCARED STEREO DATASET

|       | ELAS | SGBM | DIS | GwcNet* | PSMNet* | BDIS delta | BDIS   |
|-------|------|------|-----|---------|---------|------------|--------|
| D8K1  | 6.452| 6.522| 6.847| 7.726   | 14.759  | 6.593      | **6.386** |
| D8K2  | 3.813| 3.785| 4.036| **3.642**| 7.151   | 3.947      | 3.913  |
| D8K3  | 1.642| 3.258| 1.904| **1.513**| 3.210   | 1.782      | 1.754  |
| D8K4  | 2.008| 1.912| 2.434| **1.824**| 3.691   | 2.106      | 2.105  |
| D8K5  | 2.927| 3.187| 2.625| **1.599**| 3.220   | 1.906      | 1.880  |
| D9K1  | 3.197| **2.539**| 3.399| 2.670   | 6.577   | 2.630      | 2.607  |
| D9K2  | **0.683**| 1.452| 0.995| 0.826   | 1.833   | 0.766      | 0.704  |
| D9K3  | **0.834**| 1.360| 1.862| 0.859   | 2.064   | 0.968      | 0.940  |
| D9K4  | 0.806| 19.094| 7.441| 1.032   | 1.220   | 0.694      | **0.681** |
| D9K5  | 0.385| 30.608| 13.456| 0.509   | 0.888   | 0.358      | **0.343** |

**Average mean error** | **2.275** | **7.372** | **4.500** | **2.220** | **4.461** | **2.175** | **2.131** |
**Average median error** | 1.825 | 3.223 | 3.012 | **1.556** | 3.215 | 1.820 | 1.817 |
**Valid depth pixel** | 674.71 | 560.42 | 817.37 | 826.25 | 826.46 | 739.06 | 713.58 |

DXY means the keyframe Y in data set X. The average of the mean and the average median errors of these approaches are also listed.
Data set 8 and 9 are designated as the testing data set in [23].

* means the method requires extra training data. Errors are in mm.
The number of valid depth pixels is in 1000. BDIS+ refers to our preliminary work [32].
The bold entities represent the best performances.
ELAS, SGBM, DIS, PSMNet, GwcNet, and BDIS were also tested on the SERV-CT dataset. Since only 16 image pairs are provided in SERV-CT, the trained DNNs of PSMNet and GwcNet on SCARED datasets were adopted directly. Thus, the results of PSMNet and GwcNet are cross the textural domains. Sample qualitative comparisons are presented in Fig. 9. Quantitative comparison is shown in Table IV. Please note that only the first 15 datasets were tested because ELAS did not yield a valid prediction of the 16th image pair.

Table IV also shows that the two DNN methods are not satisfying. The major reason is that both DNN methods predicted many outliers on the edges, which suffer from the darkness. Thus, their average median errors are much better than the mean error. Another reason (minor) may be the cross-domain application because the image textures of SERV-CT are different from SCARED. Moreover, although the table shows that DNNs’ results are not satisfying, it should be emphasized that Fig. 9 shows that the central region of the two DNN methods is acceptable.

D. Qualitative Comparison With ELAS on In-Vivo Dataset

The proposed BDIS was compared with ELAS, DIS, and SGBM on the in-vivo Hamlyn datasets. Only qualitative comparisons of the prior-free methods ELAS, SGBM, DIS, GwcNet, PSMNet, and the proposed BDIS are presented. The cells are the average absolute mean error (in mm). BDIS* refers to our preliminary work [32]. The bold entities represent the best performances.

Fig. 9 and Table IV indicate that ELAS has much smaller valid predictions compared to BDIS, which contradicts the results in SCARED experiments. The reason is that the region of valid pixels is within the Delaunay polygon defined by the sparse supporting corner points. Since the textural and illuminational quality of images in SERV-CT (see Fig. 6) data are worse than images in SCARED (see Fig. 5), the sparse corner points matching module in ELAS does not provide satisfying point pairs. Thus, the accuracy of ELAS with small valid regions is for reference only. In general, BDIS achieves the best mean and median absolute error performance than the other three prior-free methods.

D. Qualitative Comparison With ELAS on In-Vivo Dataset

The proposed BDIS was compared with ELAS, DIS, and SGBM on the in-vivo Hamlyn datasets. Only qualitative comparisons of the prior-free methods can be given on this dataset because no reference labels are provided for training DNNs. Like the SCARED dataset, the Hamlyn dataset was collected...
Fig. 9. Sample reconstructions of the SERV-CT stereo dataset. The rows show dataset 1, 5, 7, 9, 12, and 15, which differ significantly in texture. The error map (in mm) in the last column is the error of BDIS. Readers are encouraged to refer to the attached video for more results.

| Reference | ELAS | SGBM | DIS | PSMNet*GwcNet* | BDIS |
|-----------|------|------|-----|----------------|------|
| 1         |      |      |     |                |      |
| 5         |      |      |     |                |      |
| 7         |      |      |     |                |      |
| 9         |      |      |     |                |      |
| 12        |      |      |     |                |      |
| 15        |      |      |     |                |      |

The Hamlyn dataset is categorized into five groups by the camera-to-surface distance since the distances in this dataset vary significantly. Fig. 10 shows the sample comparisons between baseline prior-free methods and BDIS. Table V shows the average accuracy of all frames within each category. It indicates that the proposed method achieves around 10% (median error) and 15% accuracy over the original DIS algorithm. Although
no reference is provided, this comparison shows BDIS achieves an average 0.4 – 1.66 mm (median error) and 0.65 – 2.32 mm (mean error) difference from ELAS’s results.

Fig. 10 shows BDIS can achieve similar or even slightly better robustness in handling the photometric inconsistency in the stereo matching process. The photometric inconsistency poses great difficulty to the surgical stereo matching process since the stereo images are vulnerable to non-Lambertian reflectance, dark region, or textureless surfaces. Figs. 8 and 10 indicate BDIS has fewer outliers at the image edges. BDIS overcomes this issue mainly due to the following three strategies: probabilistic inverse residual-based patch fusion, initialization strategy, and coarse-to-fine probability propagation. The probabilistic inverse residual-based patch fusion uses the predictions from multiple patches to mitigate the invalid disparity pixels (ambiguous predictions). These invalid predictions mainly result from the edges of the rectified image after the image undistortion, occlusions from objects to the left or right cameras, or unsuccessful predictions due to insufficient information. The dubious and unsuccessful coarse-level predictions pass incorrect initialization to the finer scale. Our probabilistic inverse residual-based patch fusion quantifies the posterior probability; the initialization strategy discards the patch that does not converge; coarse-to-fine probability propagation provides an important indicator to help lower the patches’ probabilities with invalid pixels. Since there are enough overlaps between the neighboring patches, the low confidence initialization may be compensated by its neighbors. In the worst case, following ELAS, the predictions with very low probability are removed.

On top of the initialization issue, the ambiguous local minima pose noticeable difficulty in the stereo matching process. The local minima issue heavily impacts the performance of fast LK since the corresponding cost function under the photometric consistency assumption is nonconvex. The toy example in Fig. 2 shows that our estimated coarse-to-fine strategy measures the probability well, and the fusion module alleviates the ambiguity of the disparity with the neighboring patch predictions. This article’s sample figures demonstrate that BDIS has fewer local minima than the baseline prior-free approaches. This is also one reason for the high accuracy of BDIS in qualitative comparisons.

In addition to testing the well-illuminated soft tissue shown in Fig. 2, we further presented the scenario of bad illumination, which corresponds to the non-Lambertian dataset (see Fig. 7). In these cases, the light intensity is roughly (not accurate) proportional to \( \cos(\alpha) \) where \( \alpha \) is the angle between the surface’s normal and the viewing direction. The texture vanishes when \( \alpha \) is close to zero (sharply lighted patch) or when \( \alpha \) is large (dark region). Both bad-textured regions are obstacles to the cost function. As demonstrated in Fig. 11, the center of the soft tissue is exposed to intense lighting while the marginal region
is dark. Note that the marginal regions are dark but not invalid. As depicted in Fig. 11, our proposed BDIS outperforms ELAS in estimating or filtering dark pixels.

### E. Comments on the Comparison With DNN-Based Methods

We provide our comments on the comparisons with DNN-based methods in Sections IV-B and IV-C. It should be emphasized that the comparison with the DNN-based methods is for the completeness of all the comparisons, and our presented conclusions regarding DNN-based methods are for reference only. First, comparing prior-free methods with DNN-based methods is unfair. DNN-based methods’ performance heavily relies on the size and quality of the training data, which varies greatly in CAS. Up to now, learning-based methods cannot fully substitute prior-free methods due to prior-free methods’ robustness to parameter tuning and being free of training data. Thus, the conclusion that DNN-based methods work in one case cannot be generalized to other cases. Moreover, DNN consumes heavy computational resources of GPU, while all the prior-free methods use only one CPU core. Additionally, BDIS and DNN methods can be complementary. Since DNN is on GPU and dependent on the training dataset, CPU-based prior-free BDIS can assist in data fusion or cross-validation.

### F. Time Consumption

Table VI presents the time consumption comparisons in predicting the disparity of the prior-based methods ELAS, DIS, SGBM, BDIS, and the prior-based methods PSMNet and GwcNet. The experiments of the prior-free methods are all carried out on a single core of the CPU (i5-9400). The experiments of prior-based methods are implemented on GPU (GTX 1080ti).

#### Table VI

 Illustrated Is the Average of 10 Times of Time Consumption of Algorithms PSMNet, GwcNet, ELAS, DIS, SGBM, Our Preliminary Work [32] and the Proposed Method

| Method      | Synthetic | SCARED | SERV-CT |
|-------------|-----------|--------|---------|
|             | 640 × 480 | 1280 × 720 | 720 × 576 |
| PSMNet*     | 0.327     | 0.566  | 0.431   |
| GwcNet*     | 0.284     | 0.430  | 0.391   |
| ELAS        | 0.246     | 0.291  | 0.264   |
| SGBM        | 0.124     | 0.346  | 0.219   |
| DIS         | 0.038     | 0.093  | 0.054   |
| BDIS        | 0.097     | 0.131  | 0.114   |
| BDIS(VAR)   | 0.072     | 0.104  | 0.087   |
| BDIS        | 0.064     | 0.072  | 0.068   |

The two prior-based methods are marked as *because the computation is conducted on GPU. BDIS(VAR) refers to our preliminary work [32]. BDIS(VAR) refers to enabling the optional variance estimation function. The time consumption is measured in seconds.

#### Table VII

 Coverage Rate of the Estimated Pixel-Wise Variance (18)

| A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|----|----|----|----|----|----|----|----|
| 93.87 | 92.35 | 87.37 | 64.01 | 90.13 | 65.76 | 82.35 | 89.02 |
| A9 | A10 | a1 | a2 | a3 | a4 | a5 | a6 |
| 88.48 | 92.28 | 93.09 | 92.11 | 86.72 | 56.9 | 90.9 | 90.75 |
| a7 | a8 | a9 | a10 | S1 | S2 | S3 | S4 |
| 85.33 | 85.66 | 73.26 | 93.52 | 46.27 | 42.16 | 34.56 | 57.30 |
| S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 |
| 67.30 | 49.25 | 64.04 | 64.32 | 93.56 | 93.68 | 94.00 | 92.71 |
| S13 | S14 | S15 | D1K1 | D1K2 | D1K3 | D1K5 | D2K1 |
| 90.25 | 93.35 | 94.46 | 93.04 | 92.3 | 92.81 | 91.29 | 87.01 |
| D2K2 | D2K3 | D2K4 | D2K5 | D3K1 | D3K2 | D3K3 | D3K4 |
| 97.06 | 84.53 | 87.43 | 90.94 | 93.82 | 89.21 | 89.79 | 90.35 |
| D3K5 | D6K1 | D6K2 | D6K3 | D6K4 | D6K5 | D7K1 | D7K2 |
| 90.49 | 66.38 | 45.02 | 57.98 | 57.8 | 50.64 | 50.87 | 63.51 |
| D7K3 | D7K4 | D7K5 | D8K1 | D8K2 | D8K3 | D8K4 | D8K5 |
| 47.97 | 38.66 | 35.01 | 50.35 | 33.12 | 52.61 | 43.92 | 61.84 |
| D9K1 | D9K2 | D9K3 | D9K4 | D9K5 |
| 80.75 | 88.36 | 85.68 | 74.13 | 83.77 |

It summarizes the ratio of pixel-wise errors that falls within the bound $e_k$ falls within the range of $[-1.96\sigma_k, 1.96\sigma_k]$. The ideal coverage rate is 95%. A and a are the results of the synthetic data sets colon diffuse lighting and non-Lambertian reflectance. SX are the results on the SERV-CT data set. The results of the in-vivo data sets are also tested. It is marked as DAKY where Y is the keyframe index and X is data set index. All values are in percentage.
TABLE VIII
AVERAGE MEAN AND THE MEDIAN ABSOLUTE DEPTH ERROR COMPARISONS ON THE SCARED STEREO DATASET AND THE SYNTHETIC DATASET COLON (IN DIFFUSE LIGHTING AND NON-LAMBERTIAN REFLECTANCE)

|                  | BDIS | BDIS (no sGMM) | BDIS (no Coarse-to-fine) | BDIS (no dynamic sigma) |
|------------------|------|----------------|--------------------------|-------------------------|
| SCARED           | Median | 1.778         | 1.792                    | 1.799                    | 1.784                   |
|                  | Mean   | 2.158          | 2.204                    | 2.197                    | 2.213                   |
|                  | Median (Normalize) | 1.000  | 1.008                    | 1.012                    | 1.036                   |
|                  | Mean (Normalize)   | 1.000  | 1.021                    | 1.018                    | 1.025                   |

| Synthetic (diffuse lighting) | Median | 0.161 | 0.160 | 0.163 | 0.163 |
|                              | Mean   | 0.202 | 0.210 | 0.207 | 0.207 |
|                              | Median (Normalize) | 1.000  | 0.994  | 1.012  | 1.012  |
|                              | Mean (Normalize)   | 1.000  | 1.040  | 1.020  | 1.020  |

| Synthetic (non-Lambertian) | Median | 0.138 | 0.140 | 0.138 | 0.139 |
|                            | Mean   | 0.224 | 0.247 | 0.240 | 0.236 |
|                            | Median (Normalize) | 1.000  | 1.020  | 1.000  | 1.007  |
|                            | Mean (Normalize)   | 1.000  | 1.102  | 1.071  | 1.053  |

The full BDIS, BDIS without GMM, BDIS without a coarse-to-fine strategy, and BDIS without dynamic sigma $\sigma_{x}^{(n)}$ ($\sigma_{r}^{(n)}$) is fixed to 8 in our preliminary research [32] are presented. For the benefits of analyzing, the normalized results are also provided.

The processing time of BDIS is double as DIS owing to the extra time spent on patch-wise window traversing. Since the size of the window $P'$ is chosen as 5, BDIS needs to spend quintuple time on the patch convolution. In general, BDIS achieves a similar/better performance over ELAS and runs 4–5 times faster. Table VI shows that BDIS achieves the second-fastest speed (except DIS) over the other algorithms with 17 Hz and 14 Hz, respectively, on both datasets. The frame rate meets the requirement for real-time SLAM [2], [3] and AR [4], [5] tasks in CAS. The time consumption comparison validates our claim that BDIS is the only real-time single-core CPU-level stereo matching in CAS. BDIS can substitute near real-time ELAS because it is much faster and similar or slightly better in terms of accuracy. Moreover, readers may notice that the proposed BDIS is almost two times faster than our preliminary work [32]. The reason is that we reduce the patch ratio from 0.6 to 0.55. A smaller patch ratio means much fewer patches to be processed. Our new strategies, coarse-to-fine and dynamic sigma $\sigma_{r}^{(n)}$, contribute to better accuracy, and thus, we can sacrifice small accuracy in exchange for faster speed. The new strategies consume very limited extra time in the implementation. Considering this, we balanced the time speed and performance by reducing 0.6 to 0.55.

G. Variance Estimation

This section validates the optional variance estimation technique described in Section III-E. Ideally, 95% of absolute depth error $e_{x}$ falls within the range of $[-1.96\hat{\sigma}_{x}, 1.96\hat{\sigma}_{x}]$.

Table VII shows the frame-wise coverage rate. The experiments indicate that the variance estimation algorithm performs unstably. It indicates that results with higher accuracy have more accurate variance estimation. As exemplified in the accuracy of in-vivo experiments (Table III), the average mean error of less than 2 mm (dataset 1, 2, 3, and 9) has a much better coverage rate. The synthetic dataset’s test trend is not as apparent as in the in-vivo case. One explanation is that the disparity largely deviates from the reference, and the local fine-scale variance fails to describe the uncertainty. The large errors in the coverage rate of some frames confirm our claim that the variance estimation is partial and inaccurate. The aetoratic uncertainty can be heavily mixed with the epistemic uncertainty in the fast LK process. Moreover, a large number of pixels do not obey the Gaussian probability presumption. Finally, Table VI shows an extra 12% to 40% percent of processing time of in variance quantification.

H. Ablation Study

Several ablation studies were conducted to investigate the modules’ contributions in the proposed BDIS. We tested the performance of the sGMM (see Section III-C), coarse-to-fine probability propagation (see Section III-D), and the dynamic choice of $\sigma_{x}^{(n)}$ (see Section III-B) on the SCARED stereo dataset and the synthetic dataset (in diffuse lighting and non-Lambertian reflectance). CRFs (see Section III-A) were not tested in the ablation study because it is the major framework.

Table VIII summarizes the accuracy (average mean and average median) comparisons. The sGMM module was first disabled by setting the pixels in the patch in equal probabilities. We only observe that BDIS is inferior in the average median error in the synthetic diffuse lighting dataset. Next, the coarse-to-fine strategy was disabled, and the obtained probability only indicates the processing level. All results suggest the accuracy is improved from 1%–3%. The coarse-to-fine strategy is the major innovation regarding the preliminary work [32]. Tables I–III all validate its performance. Lastly, the dynamic sigma $\sigma_{r}^{(n)}$ enables tuning-free parameters and slightly better accuracy. Finally, our experiments show that all the new strategies in this article have a negligible impact on time consumption.

V. CONCLUSION

This article proposes BDIS as the first CPU-level real-time prior-free stereo matching in the surgical scenario. It is both novel and practical in many real-world CAS applications, e.g.,
lacking (high-end) GPUs, saving GPUs for other tasks, lacking annotated surgical dataset for training the DNN model, etc. The proposed BDIS inherits the speed of the fast LK algorithm and overcomes major obstacles in surgical image stereo matching, i.e., textureless/dark-non-Lambertian reflectance tissue surfaces, by adopting three strategies to the deterministic fast LK: sGMM, CRFs, and a Bayesian coarse-to-fine probability propagation techniques. The proposed BDIS correctly describes the relative confidence of the pixel-wise disparity. A variance estimation algorithm is also introduced based on the estimated probability. In this way, the pixels with low confidence are filtered out.

Experiments show that the prior-free BDIS achieves an average 17 Hz on 640 × 480 image with a single core of the CPU (i5-9400) for surgical images, which satisfies most SLAM, AR, and VR requirements. Its accuracy is also slightly better than the popular near real-time ELAS. The proposed method correctly quantifies the pixel-wise relative probability, which benefits outlier filtering steps. The C++ code is open-sourced for the benefit of the community. It is supposed to assist tasks like surgeon-centered AR, reduction of error, decision making, or safety boundaries for autonomous surgery.

We plan to implement BDIS on GPU end for faster performance. Theoretically, BDIS should be faster than ELAS on GPU. Future work may also focus on the adaption of the proposed Bayesian strategy to the DNN-based methods. It is also interesting to investigate the fusion of BDIS and DNN-based methods because the prior-free CPU-based BDIS and the DNN-based GPU-based methods are complementary. When DNN is carried out on GPU and dependent on the training dataset, CPU-based prior-free BDIS can assist in data fusion or cross-validation.

ACKNOWLEDGMENT

The authors would like to thank A. Smidt from the University of British Columbia for his helpful suggestion.

REFERENCES

[1] P. Mountney and G.-Z. Yang, “Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping,” inProc. IEEE Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., 2009, pp. 1184–1187.
[2] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissnayake, “Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery,” IEEE Robot. Autom. Lett., vol. 3, no. 1, pp. 155–162, Jan. 2018.
[3] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissnayake, “MIS-SLAM: Real-time large-scale dense deformable SLAM system in minimal invasive surgery based on heterogenous computing,” IEEE Robot. Autom. Lett., vol. 3, no. 4, pp. 4068–4075, Oct. 2018.
[4] N. Haouchine, J. Dequidt, I. Peterlik, E. Kerriën, M.-O. Berger, and S. Cotin, “Image-guided simulation of heterogeneous tissue deformation for augmented reality during hepatic surgery,” inProc. IEEE Intern. Symp. Mixed Augment. Reality, 2013, pp. 199–208.
[5] A. R. Widyat et al., “3D reconstruction of whole stomach from endoscope video using structure-from-motion,” inProc. IEEE 41st Ann. Int. Conf. IEEE ENG. Med. Biol. Soc., 2019, pp. 3900–3904.
[6] F. Mahmood, Z. Yang, R. Chen, D. Borders, W. Xu, and N. J. Durr, “Polyp segmentation and classification using predicted depth from monocular endoscopy,” Proc. SPIE, vol. 10950, 2019, Art. no. 1095011.
[7] X. Jia et al., “Automatic polyp recognition in colonoscopy images using deep learning and two-stage pyramid feature prediction,” IEEE Trans. Autom. Sci. Eng., vol. 17, no. 3, pp. 1570–1584, Jul. 2020.
[8] P. Edwards et al., “The challenge of augmented reality in surgery,” in Digit. Surg. Berlin, Germany: Springer, 2021, pp. 121–135.
[9] A. M. Andrew, “Multiple view geometry in computer vision,” Kybernetes, vol. 30, pp. 1333–1341, 2001.
[10] T. Kroeger, R. Timofte, D. Dai, and L. Van Gool, “Fast optical flow using inverse search,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 471–488.
[11] H. Hirschmuller, “Accurate and efficient stereo processing by semi-global matching and mutual information,” inProc. IEEE Conf. Comput. Vis. Pattern Recognit., 2005, vol. 2, pp. 807–814.
[12] D. Stoyanov, M. V. Scarzanella, P. Pratt, and G.-Z. Yang, “Real-time stereo reconstruction in robotically assisted minimally invasive surgery,” inProc. Int. Conf. On Med. Image Comput. Comput.-Assist. Interv., 2010, pp. 275–282.
[13] A. Geiger, M. Roser, and R. Urtasun, “Efficient Large-Scale Stereo Matching,” inProc. Asian Conf. Comput. Vis., 2010, pp. 25–38.
[14] J. K. Rappel, “Surgical stereo vision systems and methods for microsurgery,” U.S. Patent 9 330 477, May 3, 2016.
[15] M. Ye, E. Johns, A. Handa, L. Zhang, P. Pratt, and G.-Z. Yang, “Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery,” inProc. Hamlyn Symp. Med. Robot., 2017, p. 27.
[16] M. Turan, Y. Almaliglah, H. Araujo, E. Konukoglu, and M. Sitti, “Deep ENDOVO: a recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots,” Neurocomputing, vol. 275, pp. 1861–1870, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092523121731665X
[17] G. Yang, J. Manela, M. Happod, and D. Ramanan, “Hierarchical deep stereo matching on high-resolution images,” inProc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5515–5524.
[18] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. D. Stefano, “Real-time self-adaptive deep stereo,” inProc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 195–204.
[19] H. Xu and J. Zhang, “Aanet: Adaptive aggregation network for efficient stereo matching,” inProc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 1959–1968.
[20] P. Brandao, D. Psychogyios, E. Mazomenos, D. Stoyanov, and M. Janatka, “HAPNet: Hierarchically aggregated pyramid network for real-time stereo matching,” Comput. Methods Biomech. Biomed. Eng.: Imag. Visual., vol. 9, pp. 219–224, 2021.
[21] Y. Long et al., “E-DSSR: Efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception,” inProc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv., 2021, pp. 415–425.
[22] P. Pratt, C. Bergeles, A. Darzi, and G.-Z. Yang, “Practical intraoperative stereo camera calibration,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv., 2014, pp. 667–675.
[23] M. Allan et al., “Stereo correspondence and reconstruction of endoscopic data challenge,” 2021, arXiv:2101.01133.
[24] G. Zampokas, K. Tsolias, G. Puleka, I. Mariolis, S. Malafosios, and T. Tzovaras, “Real-time 3D reconstruction in minimally invasive surgery with quasi-dense matching,” inProc. IEEE Int. Conf. Imaging Sys. Techn., 2018, pp. 1–6.
[25] J. Cartucho, S. Tukra, Y. Li, D. S. Elson, and S. Giannarou, “VisionBlender: A tool to efficiently generate computer vision datasets for robotic surgery,” Comput. Methods Biomech. Biomed. Eng.: Imag. Visual., vol. 9, no. 4, pp. 331–338, 2021.
[26] L. Zhang, M. Ye, P. Giataganas, M. Hughes, and G.-Z. Yang, “Autonomous scanning for endomicroscopic mosaicing and 3D fusion,” inProc. IEEE Int. Conf. Robot. Autom., 2017, pp. 3587–3593.
[27] J. Zhan, J. Cartucho, and S. Giannarou, “Autonomous tissue scanning under free-form motion for intraoperative tissue characterisation,” inProc. IEEE Int. Conf. Robot. Autom., 2020, pp. 11147–11154.
[28] B. D. Lucas et al., “An iterative image registration technique with an application to stereo vision,” in Proc. 7th Int. Joint Conf. Artif. Intell., 1981, pp. 674–679.
[29] Y. Shimasaki, Y. Iwahori, R. J. Neog, R. J. Woodham, and M. Bhuyan, “Generating Lambertian image with uniform reflectance for endoscope image,” inProc. Int. Workshop Adv. Imag. Technol., 2013, pp. 1–7.
[30] J. Engel, F. Sückerl, and D. Cremers, “Large-Scale Direct SLAM With Stereo Cameras,” inProc. IEEE/RSJ Int. Conf. Intell. Robots Syst., 2015, pp. 1935–1942.
[31] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 3, pp. 611–625, Mar. 2018.
[32] J. Song, Q. Zhu, J. Lin, and M. Ghaftari, “Bayesian dense inverse searching algorithm for real-time stereo matching in minimally invasive surgery,” inProc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv., 2022, pp. 333–344.
received the B.E. and M.S. degrees in remote sensing from Wuhan University, Wuhan, China, and the University of Chinese Academy of Sciences, Beijing, China, in 2012 and 2015, respectively, and the Ph.D. degree in robotic from the University of Technology Sydney, Ultimo, NSW, Australia, in 2020. He is currently a Research Associate with the Computational Autonomy and Robotics Laboratory, Department of Naval Architecture and Marine Engineering, University of Michigan, Ann Arbor, MI, USA. His research interests include machine vision, simultaneous localization and mapping, and robotics.

Jingwei Song received the B.E. and M.S. degrees in biomedical engineering from Xi’an Jiaotong University, Xi’an, China, in 2008 and the M.Res. and Ph.D. degrees in computing from Imperial College, London, U.K., in 2017. He is currently working toward the Ph.D. degree in image processing with the School of Electrical and Data Engineering, University of Technology Sydney, Ultimo NSW, Australia. His research interests include machine vision, image processing, probabilistic representation, and uncertainty of deep learning.

Quichen Zhu received the B.Eng. degree in control from the Huzhou University of Science and Technology, Wuhan, China, in 2017. He is currently working toward the Ph.D. degree in image processing with the School of Electrical and Data Engineering, University of Technology Sydney, Ultimo NSW, Australia. His research interests include machine vision, image processing, probabilistic representation, and uncertainty of deep learning.

Jianyu Lin received the B.Eng. degree in biomedical engineering from Xi’an Jiaotong University, Xi’an, China, in 2008 and the M.Res. and Ph.D. degrees in computing from Imperial College, London, U.K., in 2017. He is currently working toward the Ph.D. degree in image processing with the School of Electrical and Data Engineering, University of Technology Sydney, Ultimo NSW, Australia. His research interests include machine vision, image processing, probabilistic representation, and uncertainty of deep learning.