Towards Dense 3D Reconstruction for Mixed Reality in Healthcare: Classical Multi-View Stereo vs Deep Learning

Kristina Prokopetc and Romain Dupont
CEA-LIST, LVML - Laboratory for Vision, Modeling and Localization
Point Courrier 173, F-91191 Gif-sur-Yvette, France
{kristina.prokopetc, romain.dupont} @cea.fr

Abstract

Faithfully reproducing surroundings in 3D is a key-component in Mixed Reality for medical training in neonatology, where a user sees a hospital room in a Virtual Reality helmet while retaining tangible interaction with a baby mannequin and various medical tools. Deep learning solutions have high claims against classical methods but their performance in real-life application remains unclear. To fill this blank, we present a comparative study of depth map based Multi-View Stereo methods for dense 3D reconstruction. We compare classical state-of-the-art methods to their learned counterparts and assess their robustness to weakly-textured and reflective surfaces as well as accuracy on thin structures both globally and locally. We also analyze the effect of depth filtering along with computational effort. Our experiments reveal various factors which contribute to the performance gap between the methods that we discuss in detail. This study is the first to evaluate traditional dense geometry reconstruction methods against brand-new deep learning models. It helps to better understand what suits best the challenges of hospital environments. Furthermore, it builds a solid analytic ground to underscore the strengths and weaknesses of the learned methods.

1. Introduction

For various computer vision applications, dense 3D reconstruction plays an important role. Supported by latest innovative tools and technology, professional training in healthcare became one of them [3, 4]. In this article we focus on neonatal resuscitation training - an essential skill for health care providers who are involved in the child delivery (*i.e.* intervention, which sometimes required shortly after birth to help a baby breathe and to help its heart beat).

Simulation scenario in neonatology includes various modalities such as human, synthetic and digital (Figure 1a-c). Recent digital simulation solutions employ Virtual Reality (VR) to enrich the experience. They however often suffer from two major shortcomings: *1)* the non-tangible side of VR often bothers learners and requires a pre-learning phase and *2)* the risk of motion sickness. As part of the collaborative effort with one of the specialized centers for medical training in France, we aim to go one step further. We develop a Mixed Reality (MR) system that solves VR-related problems and leverages all simulation types whereas allowing medical students to profit from a fully immersive experience. Figure 1 illustrates the main idea and the components of the proposed solution. For successful MR dense and faithful 3D reconstruction of the environment plays an important role. It allows to attain a consistent and accurate display between virtual and real objects. Moreover it is essential for minimizing the user's mistakes while interacting with the environment which also reassures the learner and helps to stay focus on the task.

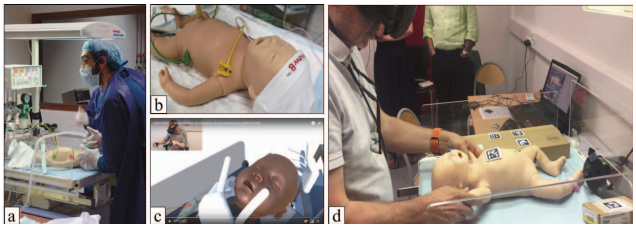


Figure 1: Learning by simulation for neonatal resuscitation. (a) - training via role-playing; (b) - high-fidelity mannequin; (c) - training in VR game; (d) - demo of the MR set-up (*i.e.* a user is interacting with the mannequin while seeing its 3D model with augmented virtual features in the VR helmet).

In this paper we focus on the Multi-View Stereo (MVS) approach for dense 3D reconstruction that have been the most successful in terms of robustness and the number of applications [10, 12, 19]. MVS methods evolved as a result of careful engineering using geometric priors and various depth cues. They vary in complexity but generally share

similar ideas and use multiple depth maps due to the flexibility and scalability of such representation [8, 6, 17, 20]. Recently, new line of methods emerged [11, 25, 26] that builds on the pivotal ideas from classical MVS within Deep Learning (DL) framework.

While classsical MVS methods and their performance has been extensively studied and tested on public benchmarks [13, 15, 18, 19, 21], the evaluation of learned MVS methods remains spurious and mainly exists within the experimental sections in the corresponding publications. This creates a significant gap in understanding how well the learned MVS stands against classical MVS. Specifically, *1)* How well learned MVS models can generalize to a unseen data? *2)* Does the inclusion of semantic features helps to better reconstruct challenging areas such as thin structures, weakly-textured and reflective surfaces? *3)* What are the crucial factors that have to be taken into account for development of a learned MVS method?

To this end, we propose a first comparative study of classical depth map based MVS methods and their learned counterparts on real-life dataset composed of challenging scenes from medical environment. We start with a review of related methods in §2. We define the evaluation protocol and provide the description of our dataset as well as evaluation criteria in §3. Our experiments reveal various factors which contribute to the performance gap between classical and learned MVS methods that we discuss in detail in §4. Finally, we synthesize our findings into conclusions in §5.

2. Multi-View Stereo in the Wild

The goal of MVS is to reconstruct a dense 3D point cloud from a collection of images taken from different views with the known camera poses. In our study we focus on methods that operate directly on the image, where the core problem is to infer depth (and normal) information for every view. Full MVS pipeline consist of three main stages: *1) sparse reconstruction*, where given the input image sequence and the corresponding camera calibration a Structure from Motion (SfM) is solved to obtain a sparse point cloud and camera poses; *2) densification*, where per-view depth (and normal) information is inferred by solving dense pixel-wise stereo-correspondence problem and *3) fusion*, where individual depth maps are merged into a single dense point cloud. In this work we aim to evaluate methods proposed for *densification* task. Therefore, we provide an overview of existing depth map based MVS methods, further 'MVS methods'.

2.1. Classical Methods

Traditional MVS methods are composed of five building blocks, namely stereo-pair selection, matching cost computation (*e.g.* photo-similarity metrics), cost aggregation, depth (and normal) computation (*e.g.* local or global), and depth refinement (*e.g.* sub-pixel accuracy). They generally build on few seminal ideas such as PatchMatch algorithm [1], originally proposed for image editing, its re-purposed version PatchMatchStereo [5] as well as PlaneSweepStereo [7].

[1] is a randomized algorithm for quickly finding approximate nearest neighbor matches between image patches. This idea was adopted to MVS under assumption of fronto-parallel scene structure until the authors of [7] proposed a way to handle slanted surfaces, where they back-project the image set onto successive virtual planes in different directions in the 3D space. Later, the authors of [5] combined both aforementioned concepts and proposed a method that randomly initializes depth values and further refines the hypothesis based on the local propagation and the random search strategies on slanted support windows.

Further extention of [5] was presented in [20], where authors' effort is focused on improving stereo-pair selection and depth refinement. They select a subset of camera pairs depending on the number of shared points computed by SfM and their mutual parallax angle followed by depth map estimation and refinement enforcing consistency among many views. [6], in turn, explores more effective depth hypothesis propagation scheme in such a way that computation can better exploit the parallelization of GPUs. Unlike [20], the authors aggregate a set of matching costs computed from different source images for each reference view. This approach suffers from decoupled depth estimation and camera pairs selection. [23] proposed an attempt to overcome this issue where the authors extended [6] with yet more efficient propagation pattern and their optimization procedure jointly considers all the views and all the depth hypotheses. For the depth estimation of weakly-textured areas, they onwards propose a multi-scale geometric consistency guidance [24].

Opposing the idea of using all images to compute the matching costs, [28] proposed an effective method to deal with stereo-pair selection. The authors designed a robust variational approximation framework with joint depth estimation and pixel-wise view selection, where they alternate depth update with a propagation as in [1]. They incorporate fixed view selection and pixel-wise view inference with a forward-backward checks for fixed depth levels. [17] extended this method with a focus on view selection, where they employ view-dependent priors and jointly estimate per-pixel depths and normals, such that the knowledge of the normals enables slanted support windows.

Although these methods are the top performing approaches in several MVS public benchmarks [13, 15, 18, 19, 21], there exist open issues. Despite good performance on well textured areas under Lambertian surface assumption, weakly-textured and reflective regions are often poorly reconstructed. Moreover, depth discontinuities pose additional challenge as well as the presence of thin structures.

2.2. Deep Learning Methods

Motivated by the success in classification and recognition tasks due to the robustness of deep features learned using convolution neural networks (CNNs) and the assumption that they can introduce global semantic information such as structural, specular and reflective priors for more robust matching, a number of efforts have been made to apply DL to dense 3D reconstruction. Deep features have been extensively used for stereo matching and similarity metric learning [9, 14, 27] to name a few. However, directly extending the learned two-view stereo to multi-view settings is less trivial. In that front there are considerably fewer pioneers. They mainly combine key contributions of learned two-view methods and borrow insights from seminal ideas in classical MVS approaches [11, 22, 25, 26].

[11] poses the depth estimation as a multi-class classification problem. First, the authors produce a set of plane-sweep volumes [7] for a reference view that contains the warped neighbor colors at every disparity, and feed these into a CNN to extract features from each patch pair (reference patch vs. patch in plane-sweep volume). Second, they use an encoder-decoder architecture with skip connections to aggregate the features across large spatial regions. Lastly, they use a max-pooling layer to aggregate the information extracted by each neighbor image and produce the final depth estimate which is then refined using conditional random fields (CRFs). This method is trained on a combination of real and synthetic datasets to circumvent the issue of incomplete ground truth, and, theoretically, can handle an arbitrary resolution.

[22] proposed a partially learned MVS method, where multiple images are first encoded into the cost volume by calculating pixel-wise absolute intensity difference and it is then passed to the encoder-decoder network along with a reference frame that estimates inverse depth maps at four resolutions. Even though, the method has a multi-view nature it is primarily targeted for continuous generation of depth maps given image-pose sequences from a localized moving camera rather than for dense 3D reconstruction.

[25] encodes camera geometries in the network as differentiable homography and infers the depth map for the reference image. In this solution, deep image features are first extracted from input images through a 2D CNN. These features are then warped into the reference camera frustum by differentiable homographies to build the feature volumes in 3D space. To handle N-view image input, a variance based cost metric is proposed to map N feature volumes to one cost volume. Similar to the learned two-view stereo, this solution regularizes the cost volume using the multi-scale 3D CNNs, and regresses the reference depth map through the soft argmin operation. A refinement network is further applied to enhance the depth map quality. The method does not include synthetic data in its training.

The major drawback of [25] is its inability to handle high-resolution images because the memory requirement of learned cost volume regularizations grows with model's resolution. To this end [26] proposed a modification, where the authors exploit recurrent neural networks (RNNs) and regularize the cost volume in a sequential manner using the convolutional gated recurrent unit (GRU). This model does not employ any specific data augmentation and produce depth maps 4x times smaller than the original input.

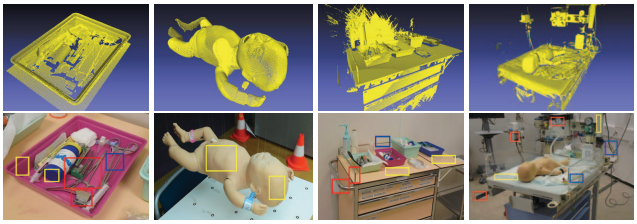
Whilst learned MVS methods specifically designed for 3D reconstruction are not many, they already exhibit an interesting research direction. Taking into account the fast development of the DL tools and solid understanding of analytical geometry this may inspire more contributions to learned MVS in near future. Consequently, the comparative evaluation with classical approaches becomes essential.

2.3. Selected Methods

Whereas we strive to identify the most suitable solution for dense 3D reconstruction of medical environment, it is necessary to choose the methods for evaluation taking into account various factors. First, the method should either represent the state-of-the-art or to be close to it, especially in case of classical MVS with its long history. Concerning learned MVS, we consider the latest contributions that target 3D reconstruction. Second, the method should have some record of testing on public benchmarks such that we can relate our findings to some reference point. Availability of an open-source implementation is another important factor that facilitates the evaluation considerably. Taking into account incremental nature of the contributions, we limit the selection to four methods with two representatives per category. These requirements lead us to the following choice of methods which we will further refer to in the remaining sections of this paper: classical MVS methods (C-MVS) [17, 20], learned MVS methods (L-MVS) [11, 26].

3. Comparative Evaluation

To evaluate the methods' performance we adopt the protocol and quantitative measures from public benchmarks [15, 18]. Given a densely reconstructed point cloud R obtained from our dataset by each selected method and a ground truth dense point cloud G acquired with high-precision laser scanner, the principle is to align R to G with the maximal possible precision and evaluate the quality and fidelity of R. The main objective is to analyze how the method behaves in the presence of thin structures, weakly-textured and reflective surfaces and examine the effect of depth filtering. Thus, our evaluation framework consist of four main steps: *1)* dataset and ground truth acquisition; *2)* execution of the selected methods; *3)* R model to G model alignment *4)* comparative evaluation through a number of experiments. The details are given below.



(a) ToolBox (b) NewBorn (c) ToolStand (d) ReanimTable

Figure 2: Images from the dataset with specular (red), weakly-textured (yellow) and thin (blue) structures and respective ground truth models. (Best viewed electronically)

3.1. Dataset and Ground Truth

Our dataset is composed of four medical scenes. It was acquired in the partnered center for medical training. A consumer grade DSLR camera was used to manually obtain the image sequences of the scene so it is always the central point of the acquisition. This is a natural setting for the user of MR system. All the scenes were obtained under same fixed and uniform illumination indoors. We used NIKON D750 with fixed intrinsic at 6016×4016 pixels image size. Figure 2 shows the sample views from each scene.

ToolBox with 20 views, being the smallest object and the shortest image sequence, shows a plastic box with some medical instruments in it (Figure 2a). It is a close-up example of a reflective and weakly-textured scene. This object is also present in the *ToolStand* scene but it has been specifically extracted as a separate entity to evaluate on.

ToolStand with 71 views shows the medical toolstand with various objects on it located near the wall in a child delivery room (Figure 2c). It also showcases the reflective and weakly-textured surface but at a bigger scale.

NewBorn with 43 views, depicts the high-fidelity baby mannequin used in the training sessions to replace a real patient (Figure 2b). This is one of the main objects that the user interacts with and on which various augmentations occurs (*e.g.* change of skin color). It is an example of smooth depth discontinuities with a homogeneous texture.

ReanimTable with 152 views, being the largest both in size and scale, shows a reanimation table in the child delivery room where the neonatal reanimation is performed (Figure 2d). This scene contains a lot of fine metallic and plastic structures while being the largest in out dataset.

To obtain the ground truth G, measurements with the high-precision laser scanner were made for each scene and around and above the objects in it, where possible. We used a portable hand-held close range (0.1 - 10 meters) laser scanner Creaform HandySCAN 3D unlike the static wide range (0.6 - 330 meters) Faro Focus X 330 used in [15, 18]. It allows continuous acquisition such that a complete 3D model can be obtained in one session with alignment and fusion of the 3D scans performed in real-time. Its maximal accuracy is 0.030 millimeters contrary to 2 millimeters in [15, 18]. Alas it requires the use of tracking targets to get the highest performance level (small circular chape stickers visible on the example of the *NewBorn* scene in Figure 2b).

Our G models are not always complete due to the vast presence of specular and semi-transparent objects which are difficult to digitize. They are not rendered as the colored point clouds using images from the camera. First, there exists a difference between the image and laser point cloud resolution where the latter is sparse at the full DSLR resolution. Second, upsampling the G model's density or downsampling the images may lead to misalignment of color and the surface geometry. This is also why we have chosen not to generate ground truth depth maps from the laser scanner.

3.2. Method Execution

All methods have been tested on a computer with AMD 16 cores (x32) CPU, 32GB RAM, GeForce GTX 1080 ti 11Gb GPU. Our ground truth does not contain the depth maps to directly evaluate the methods. Thus, we integrate them in the full MVS pipeline as described below.

In *sparse reconstruction* step we rely on SfM method in [16] due to its robustness and efficiency. Unlike [18], we do not perform realignment of the acquired images with G models at this step in order to overcome any remaining SFM drift effects. This requires to rely on the color information from the laser scanner which we do not have. Even though [26] was specifically designed to deal with high resolution images, it cannot handle the images from our dataset in their original resolution. Therefore, all images were standardized to 1152×864 size by downsampling and centered cropping to achieve a fair comparison. This is maximum size that fits to [26] while preserving the aspect ratio.

In *densification* stage we run all the selected MVS methods one by one using default parameters without downscaling the input to make the methods more comparable and preserve their original formulation.

In *fusion* we opt for customized execution. Note that open-source implementation of C-MVS methods [17, 20] is provided with a corresponding sophisticated fusion algorithms. We, however, prefer to carefully exploit the multi-view information at the level of photo-consistency, and use a rather basic fusion scheme. To this end we simplified the fusion steps in clasical methods by turning off specific parameters triggered for 'best' outcome. L-MVS methods [11, 26] are complemented with the basic fusion method from [6].

3.3. Reconstruction to Ground Truth Alignment

Because the images capture wide view of the scenes while the G models are limited to a certain part, we extract the Region of Interest (RoI) from the R prior to the

alignment. This is done by manually pre-registering G with R and removing all points of R which do not fall within the bounding box of the corresponding G model yielding R_{RoI}. Thus, the alignment is performed in two steps:

1. We pre-align R_{RoI} to G by picking 20 points correspondences. This simplifies the automatic point cloud registration and helps to achieve more accurate result.

2. We run iterative closest point algorithm (ICP) [2] with scale adjustment for fine registration. It uses the root-mean-square (RMS) of the ratio between the deviations of the centroids of each point cloud to estimate the scale. We run ICP for 200 iterations reaching final RMS as low as 1 millimeter (*mm*) .

3.4. Success Criteria and Evaluation Metrics

A suitable method should have high score in its global performance within an optimal allowed deviation τ. We choose this to be 2*mm* for all scenes according to the needs of our application and because they do not vary greatly in scale and sampling density (*i.e*. all scenes were scanned at close range). Ideally, R should have no missing parts and minimal noise in the areas of thin structures. To quantify the global performance we employ different quality metrics used in the existing benchmarks [15, 18]:

- **Accuracy** - ξ_A is a fraction (%) of the points in R that are close to the nearest points in G up to certain τ.

- **Completeness** - ξ_C is a fraction (%) of the points in G that are close to the nearest points in R up to certain τ.

- **F-1 Score** - ξ_F measures the method's performance for a certain τ. It combines accuracy and completeness in the form of harmonic mean $2 \cdot (\xi_A \cdot \xi_C)/(\xi_A + \xi_C)$.

- **Rank** denoted S is inferred from F-1 Score and defined as an average ξ_F of a method across the dataset.

ξ_A can be maximized by producing a very sparse set of precisely localized points that yields a low ξ_C and vice versa. Meanwhile only R that is both accurate and complete can guarantee a high ξ_F score for a conservative τ. We also focus on examining ξ_A and ξ_C across a range of distance thresholds which we set to $\tau = 1, 2, 5, 10, 20, 50$ *mm*.

High global ξ_A and ξ_C scores do not necessarily imply an accurate and complete reconstruction in challenging areas. Thus, it is important to perform a fine-grained examination. We propose to visually inspect G distance-color coded with respect to R in the the regions of interest to analyze the completeness in the weakly-textured and reflective surfaces and the amount of noise localized around thin structures.

Finally, we are interested in relatively fast execution as the estimation of individual depth maps is just one step in the full MVS pipeline. Thus, we measure computation time for each method in minutes across all images in a scene as well as average result per image across the dataset.

4. Experimental Results

4.1. Global Reconstruction Quality

The results in Table 1 show the global performance of all evaluated methods for each of the medical scenes. It reports the ξ_A, ξ_C, ξ_F scores for the reconstruction produced by the methods using distance threshold $\tau = 2mm$. For each method the table also provides its average rank S which, we believe, is a robust measure of the relative performance.

We found that C-MVS [17] achieves the highest accuracy on all scenes in our dataset regardless the inclusion or exclusion of the depth filtering in the fusion step. It attains the top ξ_F scores on three scenes out of four and yields the 'best' rank $S = 1.25$. On *ToolStand* it is superseded by its categorical neighbor C-MVS [20], which ξ_F score is only a small fraction higher.

Despite the competition between C-MVS [20] and L-MVS [26], the former ranks the second 'best' on our dataset with $S = 2.25$ leaving behind the latter. In particular, C-MVS [20] achieves $\xi_F = 73.38$ on the *ToolBox* scene, where L-MVS [26] follows right after with a significant gap. This situation changes on the *NewBorn* and *ReanimTable* scenes, where L-MVS [26] beats C-MVS [20]. This is expected result for the *NewBorn* scene as it depicts an object of homogeneous texture and very smooth surface where handcrafted photo-consistency metrics are unreliable. A remarkable performance gap on *Reanimtable* scene can be also attributed to the learned nature of L-MVS [26].

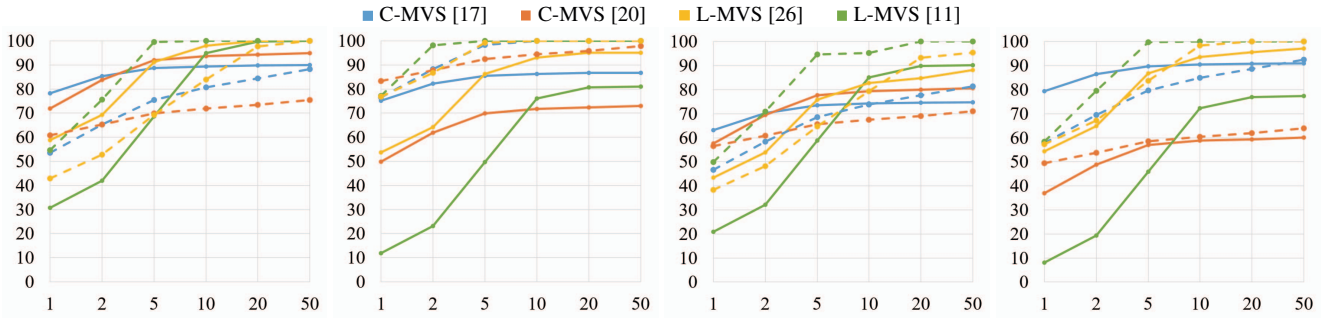
L-MVS [11] is ranked the last despite constantly providing the highest ξ_C scores. In fact, it demonstrates the situation mentioned in §3.4. In particular, ξ_C is maximized by densely covering the space with points, where only small fraction of points are actually accurate.

The exclusion of depth filtering step notably increases ξ_C while reducing ξ_A. This is a common trend across all the scenes and for all the methods. Interestingly, the inclusion of the filtering step lowers the ξ_F score. Overall, *ToolBox* and *NewBorn* appears to be the easiest datasets for the evaluated methods. *ToolStand* and *ReanimTable* are the hardest that can be ascribed to the larger scale as well as amount of specular materials and thin structures.

In addition to the global statistics on the specific distance threshold Figure 3 shows sensitivity of the methods to the different thresholds across the dataset. Concretely, each graph provides a plot of two curves representing accuracy ξ_A and completeness ξ_C for the corresponding method for each threshold τ. We found that the accuracy and completeness rankings among C-MVS methods are relatively stable with respect to τ whilst both L-MVS methods show more sensitivity. This implies that classical methods are more ro-

		ToolBox			NewBorn			ToolStand			ReanimTable		
	S	$\xi_A^{+/-}$	$\xi_C^{+/-}$	$\xi_F^{+/-}$	$\xi_A^{+/-}$	$\xi_C^{+/-}$	$\xi_F^{+/-}$	$\xi_A^{+/-}$	$\xi_C^{+/-}$	$\xi_F^{+/-}$	$\xi_A^{+/-}$	$\xi_C^{+/-}$	$\xi_F^{+/-}$
C-MVS [17]	**1.25**	**85.41 / 81.13**	65.23 / 73.05	**73.97 / 76.88**	**82.29 / 78.17**	88.22 / 98.80	**82.15 / 87.28**	**70.27 / 66.75**	58.42 / 65.43	63.80 / 66.08	**86.49 / 82.16**	69.51 / 82.16	**77.08 / 79.95**
C-MVS [20]	2.25	83.78 / 78.75	65.27 / 72.44	73.38 / 75.47	61.81 / 58.10	87.77 / 97.42	72.54 / 72.79	69.50 / 65.33	60.91 / 67.61	**64.92 / 66.45**	48.89 / 45.95	*53.81 / 59.72*	51.23 / 51.94
L-MVS [26]	2.5	69.26 / 67.87	*52.73 / 56.94*	59.87 / 61.93	64.25 / 62.96	*86.68 / 93.61*	73.80 / 75.28	53.98 / 52.90	*48.21 / 52.06*	50.93 / 52.48	64.82 / 63.52	67.13 / 72.50	65.96 / 67.71
L-MVS [11]	*4*	*42.01 / 41.58*	**75.68 / 79.46**	*54.03 / 54.60*	*23.10 / 22.86*	**98.22 / 100.0**	*37.40 / 37.43*	*32.21 / 31.88*	**70.91 / 74.45**	*44.30 / 44.65*	*19.34 / 19.14*	**79.5 / 83.48**	*31.11 / 31.14*

Table 1: Performance statistics for all methods across all scenes in the dataset for ξ_A, ξ_C, ξ_F and S at $\tau = 2mm$ distance threshold obtained with and without depth filtering (+/-) with **'best'** scores and *'worst'* scores.



(a) ToolBox (b) NewBorn (c) ToolStand (d) ReanimTable

Figure 3: Sensitivity of ξ_A and ξ_C (%, vertical axis) to the distance thresholds τ (*mm*, horizontal axis) for all methods across the dataset. Solid line denotes ξ_A and dashed line denotes ξ_C. (Best viewed electronically)

bust to the depth range related uncertainty. All the methods reach their maximal performance (both ξ_A and ξ_C) at 10*mm* on *NewBorn*, *ToolStand* and on *ReanimTable* while on *ToolBox* it is at 5*mm*. Considering ξ_A and ξ_C being over 80% as the minimal desired performance, only after 10*mm* some of the methods can reach the plank on our dataset, except the *NewBorn* scene. Figures 3c and 3d emphasize the average performance drop for the hardest scenes in the dataset.

4.2. Robustness to Challenging Structures

C-MVS methods are known to suffer in the areas of homogeneous texture and non-Lambertian materials. Depth discontinuities on object boundaries pose additional challenges. Thus, it is no surprise that reconstructed models may have missing parts and certain amount of noise localized near object edges and thin structures. L-MVS methods, in contrast, have high claims in this matter relying on semantic features which assumed to better represent the scene's context. In this work we are interested in the best performance for each category of methods. This aids the understanding of how much room remains for the progress, specifically when there exist a significant performance gap between classical and learned MVS. Figure 4 shows the full view of the models obtained by the best-performing methods. For each method and each scene, the figure provides the final reconstruction, the accuracy and completeness.

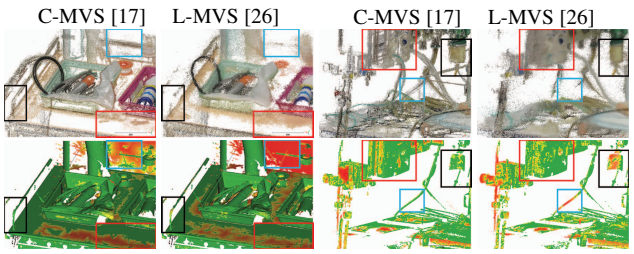
One can see that L-MVS [26] provides a better coverage of the table surface in the *ToolBox* scene even though the ξ_C does not capture this. This is also true for the remaining scenes. It, however, tends to oversmooth the reconstruction (*e.g.* the face of the baby-mannequin in the *NewBorn* scene). C-MVS [17], in turn, preserves more details. This can be seen on *ReanimTable* scene. Generally, both methods have difficulties in the same areas.

Figure 5 shows the close-up views of some of the most challenging areas in the two hardest scenes of our dataset, namely *ToolStand* and *ReanimTable* for C-MVS [17] and L-MVS [26]. Colored rectangles emphasize the differences in perfomance between the methods. Thus, L-MVS [26] struggle to reconstruct thin reflective handle of the *ToolStand* shown in black rectangle in Figure 5a. Blue rectangle highlights the area with the thin reflective structure on the homogeneous background. C-MVS [17] performs remarkably better in this area. It, however, cannot complete the flat surface of the table as well as L-MVS [26] as shown by red rectangle. We can see in Figure 5b how sophisticated depth filtering prevents C-MVS [17] to corectly reconstruct the metrology pannel as highlited by red rectangle. The effect of the depth map resolution can be observed as well. Specifically, L-MVS [26], which output depth maps are 4x smaller than the original input, fails to recover thin objects and details as highlighted with blue and black rectangles.

4.3. Computation Time

Table 2 provides the measured time in minutes. Even though, processing time strongly depends on how a method is implemented, we believe that some global useful conclusions can be drawn out of these results. C-MVS [20]

Figure 4: Reconstruction results from the best perfoming classical and leaned MVS methods [17, 26] across the dataset for $\tau = 2mm$ precision threshold. C-MVS results are shown in top three rows and L-MVS results are shown in the bottom three rows respectively. Rows 1,4 - colored R model; Rows 2,5 - the R model color coded with distance to G visualizes *accuracy*; Rows 3,6 - the G model color coded with distance to R visualizes *completeness*. (Best viewed electronically)



(a) ToolStand (b) ReanimTable

Figure 5: Performance of the best C-MVS and L-MVS on the challenging structures of two hardest scenes in the dataset. Top row - colored reconstructions and bottom row - their completeness maps. (Best viewed electronically)

	C-MVS [17]	**C-MVS [20]	L-MVS [26]	*L-MVS [11]
ToolBox (20)	9.77	1.53	**0.6**	*241.52 (>4h)*
NewBorn (43)	23.58	2.97	**1.25**	*519.268 (>8h)*
ToolStand (71)	35.72	5.63	**2.02**	*857.396 (>14h)*
ReanimTable (152)	71.47	11.32	**4.25**	*1835.552 (>30h)*
Average per image	0.48	0.06	**0.02**	*11.12*

Table 2: Runtime for each method per scene in minutes with **fastest** and *slowest* times. (*) - partial execution on CPU and GPU, (**) - multithreaded CPU only, (20) - # images.

is the fastest in its category and it runs on CPU via efficient multithreaded implementation. C-MVS [17] takes a long time. We attribute this to simultaneous estimation of depth and normal maps and photometric/geometric consistency enforcement, which creates a very large and computationally expensive optimization problem. L-MVS methods demonstrate two extremities of the computational effort with [26] - the fastest and [11] - the slowest. We attribute this to the use of RNN in the regularization phase of [26] and the unoptimized generation of plane-sweep volumes on CPU in [11]. The main drawback of the winner [26], however, is the 4x smaller size of the resulting depth maps.

4.4. Discussion

Traditional approaches are still troublesome to outperform. Learned MVS methods have certain difficulties to generalize to our dataset. This potentially can be improved via fine-tuning given that the available ground truth is 100% complete. This is hard to achieve in the settings of real medical environments due to vast amount of specular materials and weakly-textured surfaces, which are difficult to digitize.

The assumption that learned MVS can introduce global semantic information for more robust matching appears rather valid. In some cases learned MVS methods were able to recover the missing geometry where the photo-consistency metrics of classical MVS were unreliable.

Evaluated learned MVS methods mainly consider the overall accuracy by reporting global statistics of depth residuals during training. The shortage of the awareness of salient and important regions like geometric discontinuities results in apparent defects on object boundaries and thin structures. Evidently, this dictates the need to incorporate the depth discrepancy localization into the process. Also, image super-resolution becomes essential if the model's architecture restricts the input resolution to some maximum.

	C-MVS [17]	C-MVS [20]	L-MVS [26]	L-MVS [11]
global quality	****	***	**	*
textureless objects	***	***	**	**
reflective surfaces	***	***	***	**
thin structures	***	***	*	*
amount of noise	*	*	***	****
degree of detail	***	****	**	**
computation time	**	***	****	*

Table 3: Evaluation summary. The methods ranked with respect to their efficiency where * (one star) is the lowest rank and **** (four stars) is the highest rank.

Finally, one may wonder if there is a win-win situation, *i.e.* if there is a method that is not computationally demanding while being accurate and provides maximal coverage of the scene preserving its fidelity? The answer to this question is no as some methods perform well on some aspect and bad on others. Table 3 provides an evaluation summary, where methods are ranked w.r.t their efficiency. MVS method [20] seems to offer the best compromise between speed, simplicity and efficiency.

5. Conclusion

We presented a comparative evaluation of four depth map based MVS methods designed for dense 3D reconstruction. Two of these methods follow traditional practices of geometric modeling and rely on carefully engineered solutions [17, 20] while others use deep learning [11, 26]. These methods were compared based on their computational power requirements as well as their capability to correctly estimate depth information on the challenging dataset from medical environment while maximizing robustness to challenging structures. Our results prove the superiority of classical MVS in accuracy while learned approaches tend to enforce completness. We believe that further exploring the potential of learned MVS [11, 26] is a promising research direction. We plan to continue to improve and enlarge our dataset that later will be available to the community.

Acknowledgement

This work is a part of LABFORSIMS2 project funded by the National Research Agency of France. We thank the team of the LabForSIMS Simulation Center at the Faculty of Medicine of the Paris Sud University for providing all necessary facilities and participating in the discussions.

References

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, volume 28, page 24. ACM, 2009.

[2] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992.

[3] A. Blanié and M. L. Guen. Debriefing values in high-fidelity simulation. *Anaesthesia, Critical Care and Pain Medicine*, 36(4):201 – 202, 2017.

[4] A. Blanié, P. Roulleau, C. Mengelle, and D. Benhamou. Comparison of learning outcomes between learning roles (spectator and actor) during an immersive simulation. *Anaesthesia, Critical Care and Pain Medicine*, 36(4):243 – 244, 2017.

[5] M. Bleyer, C. Rhemann, and C. Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011.

[6] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015.

[7] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[8] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[9] W. Hartmann, S. Galliani, M. Havlena, L. Van Gool, and K. Schindler. Learned multi-patch similarity. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1586–1594, 2017.

[10] V. H. Hiep, R. Keriven, P. Labatut, and J.-P. Pons. Towards high-resolution large-scale multi-view stereo. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1430–1437. IEEE, 2009.

[11] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

[12] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR 2011*, pages 3121–3128. IEEE, 2011.

[13] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014.

[14] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

[15] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[16] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016.

[18] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017.

[19] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006.

[20] S. Shen. Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes. *IEEE transactions on image processing*, 22(5):1901–1914, 2013.

[21] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. Ieee, 2008.

[22] K. Wang and S. Shen. Mvdepthnet: real-time multiview depth estimation neural network. In *2018 International Conference on 3D Vision (3DV)*, pages 248–257. IEEE, 2018.

[23] Q. Xu and W. Tao. Multi-view stereo with asymmetric checkerboard propagation and multi-hypothesis joint view selection. *arXiv preprint arXiv:1805.07920*, 2018.

[24] Q. Xu and W. Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5483–5492, 2019.

[25] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*, 2018.

[26] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5534, 2019.

[27] J. Zbontar, Y. LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016.

[28] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm. Patchmatch based joint view selection and depthmap estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1517, 2014.