Shi *et al. Art Int Surg* 2024;4:247-57
DOI: 10.20517/ais.2024.17

Artificial Intelligence Surgery

Original Article

Open Access

Check for updates

Long-term reprojection loss for self-supervised monocular depth estimation in endoscopic surgery

Xiaowei Shi[1], Beilei Cui[2], Matthew J. Clarkson[1], Mobarakol Islam[1]

[1]Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) and Department of Medical Physics and Biomedical Engineering, University College London, London WC1E 6BT, UK.
[2]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong 999077, China.

Correspondence to: Xiaowei Shi, Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) and Department of Medical Physics and Biomedical Engineering, University College London, Gower Street, London WC1E 6BT, UK. E-mail: xiaowei.shi.22@alumni.ucl.ac.uk

How to cite this article: Shi X, Cui B, Clarkson MJ, Islam M. Long-term reprojection loss for self-supervised monocular depth estimation in endoscopic surgery. *Art Int Surg* 2024;4:247-57. https://dx.doi.org/10.20517/ais.2024.17

Received: 1 Mar 2024 **First Decision:** 12 Jul 2024 **Revised:** 6 Aug 2024 **Accepted:** 2 Sep 2024 **Published:** 10 Sep 2024

Academic Editors: Luca Milone, Andrew A. Gumbs **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Aim: Depth information plays a key role in enhanced perception and interaction in image-guided surgery. However, it is di icult to obtain depth information with monocular endoscopic surgery due to a lack of reliable cues for perceiving depth. Although there are reprojection loss-based self-supervised learning techniques to estimate depth and pose, the temporal information from the adjacent frames is not e iciently utilized to handle occlusion in surgery.

Methods: We design long-term reprojection loss (LT-RL) self-supervised monocular depth estimation techniques by integrating longer temporal sequences into reprojection to learn better perception and to address occlusion artifacts in image-guided laparoscopic and robotic surgery. For this purpose, we exploit four temporally adjacent source frames before and after the target frame, where conventional reprojection loss uses two adjacent frames. The pixels that are visible in the target frame but occluded in the immediate two adjacent frames will produce the inaccurate depth but a higher chance to appear in the four adjacent frames during the calculation of minimum reprojection loss.

Results: We validate LT-RL on the benchmark surgical datasets of Stereo correspondence and reconstruction of endoscopic data (SCARED) and Hamlyn to compare the performance with other state-of-the-art depth estimation methods. The experimental results show that our proposed technique yields 2%-4% better root-mean-squared

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

error (RMSE) over the baselines of vanilla reprojection loss.

Conclusion: Our LT-RL self-supervised depth and pose estimation technique is a simple yet effective method to tackle occlusion artifacts in monocular surgical video. It does not add any training parameters, making it flexible for integration with any network architecture and improving the performance significantly.

Keywords: Monocular depth estimation, self-supervised learning, reprojection loss, robotic surgery

INTRODUCTION

Depth estimation in robotic surgery is vital for surgical field mapping, instrument tracking, 3D modeling for surgical training, and lesion inspection in virtual and augmented reality. However, traditional stereo cameras consist of better depth cues with stereo correspondences and multiview images, and monocular endoscopes are unable to obtain the depth information. However, in image-guided surgery, such as robotic and laparoscopic surgery, the monocular endoscope is more popular due to better accessibility and smaller incisions. Recently, there have been a couple of reprojection loss-based self-supervised depth estimation techniques using monocular videos for both computer vision and surgical vision[1-3]. Nevertheless, the small camera pose changes in the narrow surgical environment requires long-term dependency on the monocular video frames to address the occlusion artifacts during depth estimation in the surgical environment. In this work, we propose a long-term reprojection loss (LT-RL) by considering longer temporal adjacent frames before and after the target frame in self-supervised depth estimation.

There are several works in improving reprojection loss for self-supervised depth estimation. Garg *et al.* pioneered self-supervised depth estimation with the proxy task of stereo view synthesis based on a given camera model using an L1 loss[4]. Monodepth[2] refined this via differentiable bilinear synthesis[5] and a weight of SSIM and L1 loss[6]. SfM-Learner[7] proposed the first fully monocular self-supervised depth-pose framework by substituting the stereo transform (fixed stereo baseline) with another regression network to predict the ego-motion of the camera. Monodepth2[1] optimized this work through the introduction of a minimum reprojection loss and edge-aware smoothness loss. The minimum reprojection loss attempts to address the occlusion artifacts by selecting minimum reprojection loss or photometric error between the target frame and the first adjacent frames before and after it. However, we argue that selecting minimum loss by only comparing with the first adjacent frames is not sufficient in the surgical environment where changes in camera pose are very small.

In this work, we design a LT-RL by considering the second adjacent or four frames before and after the target frame to select the minimum reprojection loss. In the surgical domain, small camera pose changes limit the reprojection error to project the pixels that are visible in the target image and are not visible in the immediate source images before and after the target image. Hence, LT-RL with four adjacent temporally frames increases the chances of tackling the occlusion artifacts. Our contributions and findings can be summarized as:

- Design a LT-RL to address occlusion artifact integrating longer temporal information during self-supervised depth estimation using monocular video in endoscopic surgery.
- Demonstrate the flexibility of the proposed LT-RL by plugging into Monodepth2 network architecture.
- Validate the proposed method with the benchmark surgical depth estimation dataset of Stereo correspondence and reconstruction of endoscopic data (SCARED) and compare it with state-of-the-art self-supervised baselines. The results suggest the effectiveness of our LT-RL in both depth estimation and 3D reconstruction.

METHODS

Preliminaries

Vanilla reprojection loss

Vanilla reprojection loss calculates the photometric errors between the target frame and two temporally adjacent frames (first adjacent) and finally chooses the minimum error between them. For example, if I_t is the target frame at time t, I_{t-1} and I_{t+1} the two source frames can be denoted as I_s for simplicity, D_t depth estimation for target frame, p_t the homogeneous coordinates and K intrinsic matrix, and then view synthesis from I_s to I_t can be formulated as:

$$p_{s\to t} = K\hat{T}_{t\to s}\hat{D}_t K^{-1} p_t \tag{1}$$

Then view synthesis $p_{s\to t}$ can be used to obtain synthesized source frame of $I_{s\to t}$:

$$I_{s\to t} = I_s\langle p_{s\to t}\rangle \tag{2}$$

Finally, the pixel-wise Vanilla reprojection loss $Loss_{rl}$ from a source frame I_s to target frame I_t can be calculated using synthesized target frame from source $I_{s\to t}$:

$$Loss_{rl}(I_t, I_{s\to t}) = \min_s RL(I_t, I_{s\to t}) = (1 - \mathrm{SSIM}(I_t, I_{s\to t})) + \mathrm{L1}(I_t, I_{s\to t}) \tag{3}$$

Where RL is the reprojection loss with combined $SSIM$ and $L1$ losses. For the two source frames of I_{t-1}, I_{t+1}, the Vanilla reprojection loss can be formulated as:

$$Loss_{rl} = \min_s(RL(I_t, I_{t-1\to t}), RL(I_t, I_{t+1\to t})) \tag{4}$$

Figure 1 demonstrates the reprojection loss with source and target view synthesis in the point cloud and then projection back to synthesize the target frame from a source. In the view synthesis approach, back-projection and reprojection are crucial steps. First, back-projection converts 2D source image pixels into a 3D point cloud usingdepth information and camera intrinsics. This point cloud is then transformed into the target point cloud using thepredicted camera pose. Next, reprojection projects the 3D target point cloud back onto the 2D image plane of thetarget camera using its intrinsic parameters.

Smoothness loss

To mitigate the issue of smoothing over edges, we integrate an edge-aware smoothness loss, as employed in[8], into our approach. This loss function effectively reduces the weight in regions with strong intensity gradients, thereby promoting local smoothness in the predicted depth map. This enables our model to preserve sharp edges and fine details in the depth estimation process. The smoothness loss aims to optimize the predicted depth by considering image gradients. It encourages depth estimation to adhere to local smoothness patterns based on the intensity gradients present in the input images. By incorporating this loss function, the model can effectively capture and preserve the continuity and smoothness of depth variations across the image, resulting in more visually coherent and accurate depth predictions, which has demonstrated success in[7].

$$Loss_{smooth} = |\nabla_x D_s| e^{-|\nabla_x I_s|} + |\nabla_y D_s| e^{-|\nabla_y I_s|} \tag{5}$$

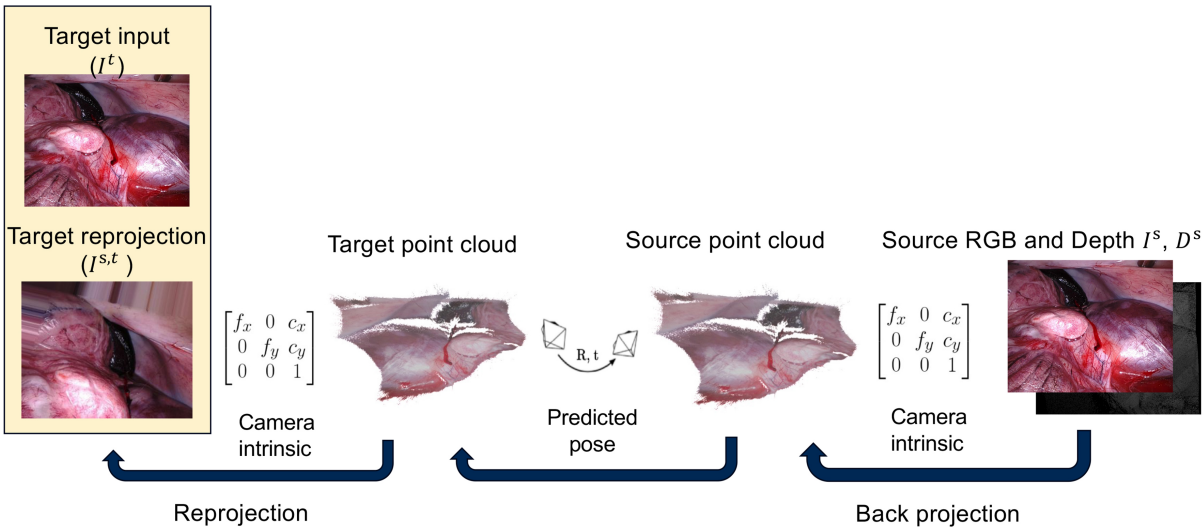

Figure 1. Demonstration of back-projection and reprojection serves as two crucial steps in the view synthesis approach for depth and pose estimation with monocular endoscopy. Given camera intrinsics, the source image is projected onto the target image using predicted depth and pose. The reprojection loss then quantifies the dissimilarity between the target image and the reprojected image using L1 and SSIM losses.

Tihkonov regularizer

To refine the generated depth map, Tihkonov regularizer is used in the AF-SfM learner[9]. It consists of three losses of residual-based smoothness loss $\mathcal{L}_{rs}$, auxiliary loss $\mathcal{L}_{ax}$, edge-aware smoothness loss $\mathcal{L}_{es}$. Overall, Tihkonov Regularizer $\mathcal{R}$(p) can be formulated as:

$$\mathcal{R}(\mathrm{p}) = \mathcal{L}_{\mathrm{rs}} + \mathcal{L}_{\mathrm{ax}} + \mathcal{L}_{\mathrm{es}} \tag{6}$$

Proposed method

LT-RL

To tackle occlusion artifact, we design LT-RL by considering four adjacent source frames temporally for a target frame. In the surgical environment, due to small camera pose changes, two adjacent frames are not sufficient to avoid the occlusion artifact. The nature of rotations poses a significant challenge in the task of pose estimation, as they are well-suited for the purpose of motion in a car while driving along a baseline. However, when it comes to endoscopy, where the endoscope is inserted into the patient's body through a small incision during surgery, it undergoes complex three-dimensional rotational movements with a restricted translation motion range. This intricate behavior of the endoscope makes the estimation of poses a more difficult and demanding task. Occlusion is no longer visible in long-span frames in comparison to the target picture as a result of the motions of the endoscope in a back-and-forth motion. This helps address inaccuracies in depth caused by occlusion artifacts, as pixels occluded in the immediate frames have a higher chance of appearing in the four adjacent frames during minimum reprojection loss calculation.

Thus, we train the network and calculate reprojection loss with scenes temporally a little further apart. In our proposed LT-RL approach, we choose individual frames from longer spans to use as the source pictures. Following Equation (4), we can consider 4 adjacent source frames of I_{t-2}, I_{t-1}, I_{t+1}, I_{t+2} and a target frame of I_t at time t. Therefore, our LT-RL can be expressed as:

$$Loss_{lt-rl} = \min_{s}(RL(I_t, I_{t-2\to t}), RL(I_t, I_{t-1\to t}), RL(I_t, I_{t+1\to t}), RL(I_t, I_{t+2\to t})) \quad (7)$$

Overall network architecture

The structure of the depth estimation network and pose network is shown in Figure 2. It consists of a regression network for pose and an encoder-decoder network of depth estimation. By following Monodepth2, we adopt the depth estimation network of commonly used UNet architecture[10] with ResNet18[11] encoder and corresponding decoder blocks. On the other hand, the pose network is also another separate ResNet18 regressor network. Our goal is to demonstrate the effectiveness and flexibility of the proposed loss function utilizing existing networking architectures.

There are a total of five input frames during training, one target and four source frames. The self-supervised optimization is performed using a combined loss of our LT-RL and smoothness loss following the baseline model of Monodept2[1] and AF-SfM learner[9]. The combined loss can be expressed as:

$$Loss = Loss_{lt-rl} + Loss_{smooth} \quad (8)$$

To enhance the depth map, we adopted Tihkonov regularizer $\mathcal{R}$(p) during training by following Equation (6) as AFSfMLearner[9].

Dataset

SCARED dataset

SCARED is the sub-challenge of the MICCAI EndoVis 2019 challenge[12]. It contains 7 endoscopic videos of seven different scenes, and each scene was captured from a stereo viewpoint, providing two perspectives for depth perception, but only the left view was used. The data were collected from the internal abdominal anatomy of fresh pig cadavers using a da Vinci Xi surgical system and a projector. We downscaled the images to 320 × 256 pixels (width × height), which was a quarter of their original size. Bi-linear interpolation was used during the down-sampling process to preserve as much visual information as possible. The depth capping (CAP) was set to 150 mm followed by[13], which means that the depth range was scaled within this threshold. The experiment was conducted with 15,351 images used for training, 1,705 images for validation, and 551 images for testing. Following previous work[13], our data split strategy follows established methodologies: training set (keyframe1 and keyframe2) from datasets 1-9 and keyframe 3-4 from dataset 8-9, validation set (keyframe3) from datasets 2-7, and test set (keyframe4 from datasets 2-7 and keyframe3 from dataset 1) with no overlap. This approach ensures a robust model evaluation across datasets, aligning with field practices.

Hamlyn dataset

Hamlyn (https://hamlyn.doc.ic.ac.uk/vision/) dataset consists of 21 videos from various surgical procedures and contains complex surgical scenes with deformations, reflections, and occlusions. All 21 videos are used for external validation to investigate the depth prediction with occlusion for the proposed method following[14].

Implementation details

We adopt the official implementation (https://github.com/ShuweiShao/AF-SfMLearner) of the AF-SfMLearner[9] as our backbone network and base optimizer. The network is trained for 20 epochs, employing the Adam optimizer with a batch size of 40 and a learning rate of 10^{-4}. The overall network and training script are implemented using the Pytorch framework. The optimization is performed in a self-supervised manner using our proposed LT-RL loss formulated in the Equation (8). To compare the

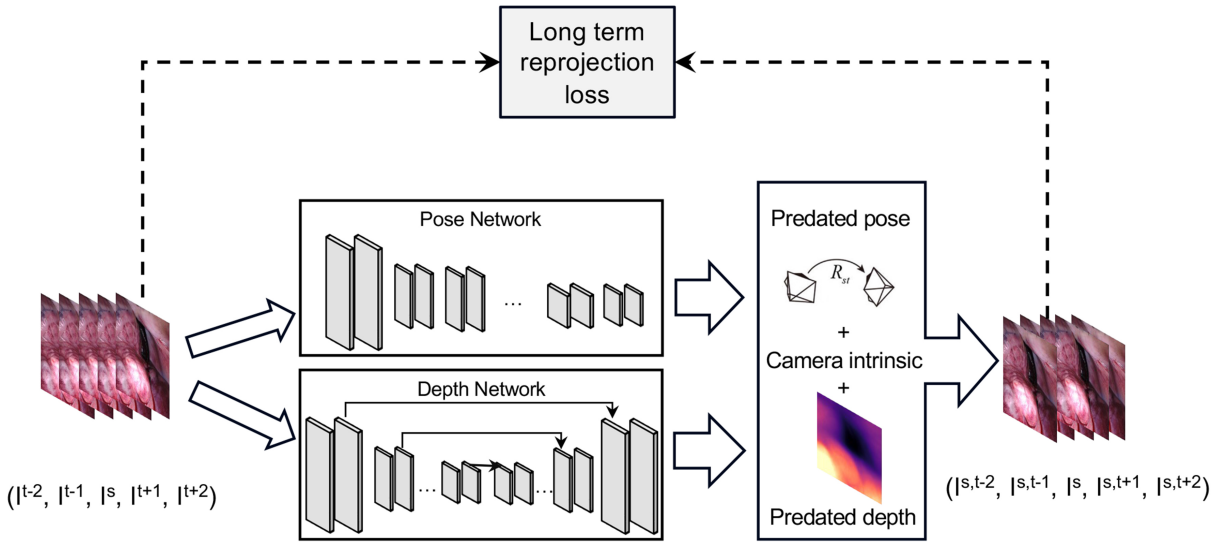

Figure 2. Illustration of the pipeline of depth and pose estimation, from input, outputs to loss calculation.

performance, we choose baselines of AF-SfMLearner[9], Monodepth2[1], HR-Depth[3], AJ-Depth[15], Lite-Mono[16] and MonoViT[17]. For a fair comparison, we retrain all the baselines following their official code repositories. Due to different versions of Python libraries and the graphics processing unit (GPU) settings, some of the baselines obtain different performances than the reported results in the corresponding papers. To tackle the issue of scale ambiguity in the predicted depth maps, wherein the depth values are subject to an unknown scaling factor, we utilized a single median scaling method following SfMLearner[7] as shown in Equation (9), similar to the baseline, enabling better comparison and analysis of the depth estimations. A range spanning from 0 to 150 mm is sufficiently broad to cover nearly all possible depth values.

$$\mathbf{D}_{\text{scaled}} = \left(\mathbf{D}_{\text{pred}} \times \left(\text{median } \left(\mathbf{D}_{\text{gt}}\right) / \text{median } \left(\mathbf{D}_{\text{pred}}\right)\right)\right) \tag{9}$$

RESULTS

Evaluation metrics

The model performance evaluation of the depth estimation method employed multiple indicators to assess its effectiveness. For measuring the quality of depth estimation, the square relative error (Sq Rel), the absolute relative error (Abs Rel), the root-mean-squared error (RMSE), the root-mean-square logarithmic error (RMSE Log) are utilized. Evaluations were conducted by capping (CAP) or restricting the depth values to 150 millimeters as described in Equation (9).

Quantitative results

The quantitative results of the experiments are presented in Table 1. The performance of our proposed method is compared against the state-of-the-art (SOTA) models of Monodepth2[1], HR-Depth[3], AJ-Depth[15], Lite-Mono[16], MonoViT[17] and Depth anything[18]. Table 1 demonstrates the superior performance of our method with the metrics of Abs Rel and RMSE Log and obtains competitive metrics of Sq Rel and RMSE. Table 1 demonstrates the superior performance of our method with the metrics of Abs Rel and RMSE Log and obtains competitive metrics of Sq Rel and RMSE. We also investigate the generalization and robustness of our model by validating it on an external dataset of Hamlyn. For this external validation, we utilized the models trained on the SCARED dataset and validated them on Hamlyn. Table 2 shows the prediction results in comparison with SOTA models of AF-SfMLearner[9] and Endo-Depth-and-Motion[14].

Table 1. Quantitative results with the SCARED dataset

	CAP	Abs Rel (% ↓)	Sq Rel (% ↓)	RMSE (mm ↓)	RMSE Log (mm ↓)
HR-Depth[3]	150	0.080	0.938	7.943	0.104
MonoViT[17]		0.074	0.865	7.517	0.097
Lite-Mono[16]		0.073	0.803	11.684	0.107
AJ-Depth[15]		0.078	0.896	7.578	0.101
Monodepth2[1]		0.083	0.994	8.167	0.107
AF-SfMLearner[9]		0.062	0.513	5.289	0.087
Depth anything (Zero-shot)[18]		0.106	1.376	8.695	0.146
Ours		**0.058**	**0.452**	**5.014**	**0.083**

The unit of % and millimeter (mm) of each metric is indicated in the bracket. The best results are in bold. SCARED: Stereo correspondence and reconstruction of endoscopic data; CAP: the capping or restriction of the depth value; Abs Rel: absolute relative error; Sq Rel: square relative error; RMSE: root-mean-squared error; RMSE Log: root-mean-square logarithmic error.

Table 2. Quantitative results with the Hamlyn dataset

	Abs Rel (% ↓)	Sq Rel(% ↓)	RMSE (↓)	RMSE Log (↓)
Endo-Depth-and-Motion[14]	0.185	5.424	16.1	0.225
AF-SfMLearner[9]	0.175	4.589	14.21	0.209
Ours	**0.165**	**4.081**	**13.497**	**0.201**

The downward arrow represents the lower, the better, and the upward arrow represents the higher, the better. Each metric's unit of % and millimeter (mm) is indicated in the bracket. The best results are in bold. Abs Rel: Absolute relative error; Sq Rel: square relative error; RMSE: root-mean-squared error; RMSE Log: root-mean-square logarithmic error.

The superior performance of our model demonstrates the better generalization and robustness of the proposed LT-RL loss. Overall, our solution is simple yet effective, easy to integrate with conventional reprojection loss, and delivers superior performance in monocular depth estimation. Extending the method to four temporally adjacent frames improves the accuracy and robustness of depth estimation by providing more temporal context. This additional information helps better capture the motion and structural details of the scene, leading to more accurate and consistent depth maps. We have conducted an external evaluation on the Hamlyn dataset, where our method marginally outperformed existing methods in depth estimation in Table 1. While the improvement in depth estimation may seem small, such enhancements can be significant for subsequent reconstruction tasks. This demonstrates the robustness and practical value of our approach.

Qualitative results

The qualitative performance of the experiments is presented in Figures 3-5. The depth prediction of our method is compared with the closely related works Lite-Mono[16] and ground-truth in Figure 3. The quantitative results demonstrate the superiority of our model over all competing methods. It is worth noting that our model excels not only in generating more continuous depth values and performing better on anatomical structures, especially in less textured and reflective regions, but also in areas with complex structures and substantial depth variations.

Figure 4 plots the pose trajectory for a testing video. We compare the predicted trajectory of the Lite-Mono pose prediction over the ground-truth (GT) pose with ours. The ground-truth trajectory is represented by a grey dashed line, while the trajectory predicted by the model is shown as a black solid line. The trajectories demonstrate the accuracy of our model prediction, which is almost similar to GT, where Lite-Mono shows a large deviation.

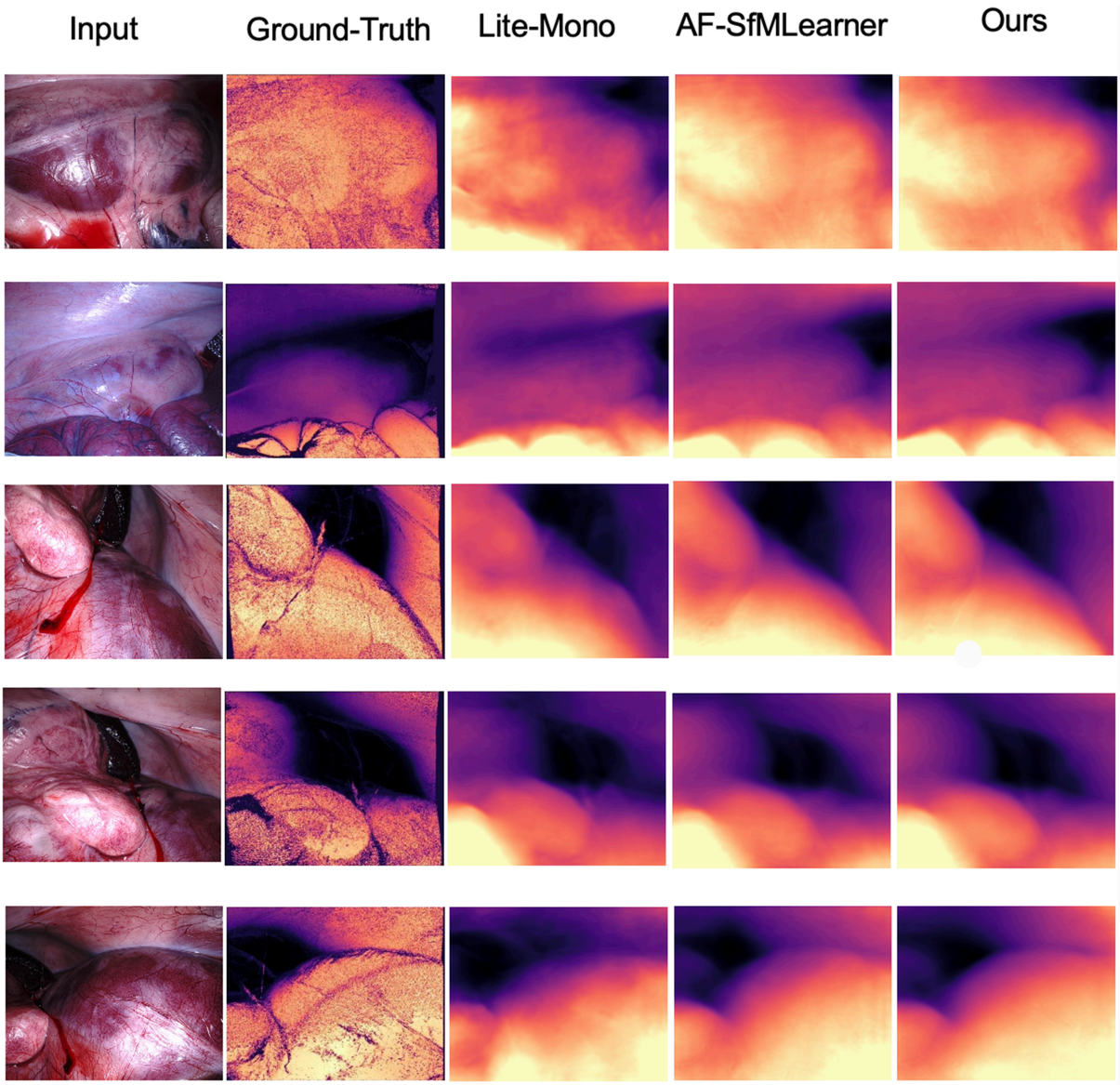

Figure 3. Qualitative comparison of predicted depth map on the SCARED dataset between our method with SOTA depth estimation methods. For visualization and evaluation purposes, the ground-truth depth values are scaled. SCARED: Stereo correspondence and reconstruction of endoscopic data; SOTA: state-of-the-art.

Figure 5 illustrates an example of 3D surface reconstruction using the predicted depth of our method over Lite-Mono and AF-SfMLearner. When analyzing the reconstruction in Figure 5, it is evident that visual distortions or errors exist at the edges of objects, making it challenging to accurately represent spatial relationships between objects. This difficulty is particularly pronounced at the edges or transitions between different regions, but in the depth map visualization shown in Figure 3, these interpolation artifacts may not be visible. Our model results provided a greatly superior surface prediction outcome, which is crucial because the reconstruction outcomes reflect a major true objective of monocular depth estimation.

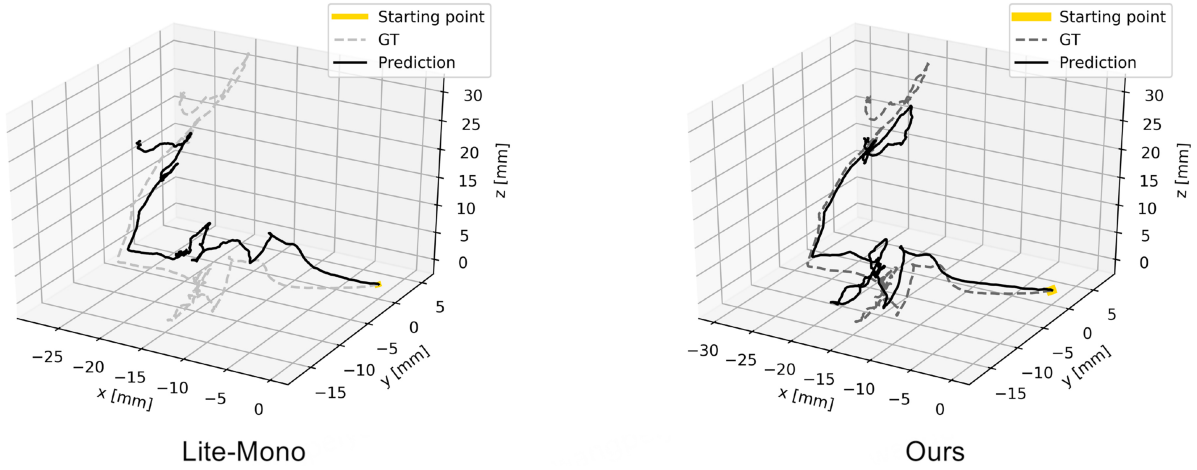

Figure 4. The predicted pose trajectory of a video from the SCARED dataset. The ground-truth trajectory is represented by a grey dashed line, while the trajectory predicted by the model is shown as a black solid line. SCARED: Stereo correspondence and reconstruction of endoscopic data.

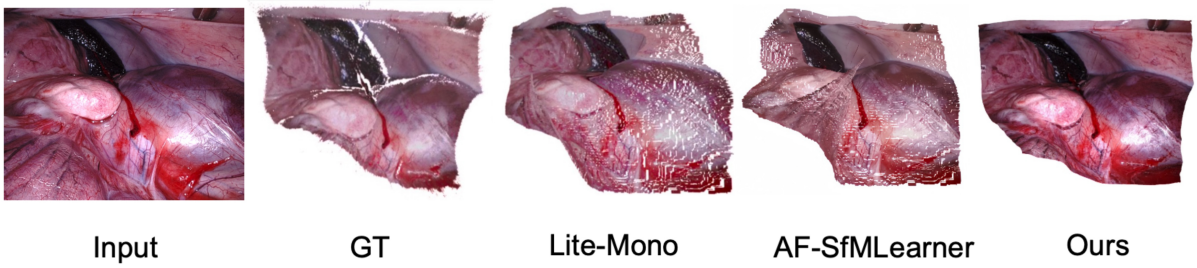

Figure 5. An example of 3D reconstruction of our model compared to the SOTA models of Lite-Mono and AF-SfMLearner. SOTA: State-of-the-art.

Influence in various frames number

To investigate the efficacy of the length of temporal information, we conducted an ablation study with different numbers of consecutive frames in Table 3. The total consecutive frame numbers of the source frame to calculate the reprojection loss are indicated in the column of "Frames". For example, Frames of "2" means 2 consecutive frames (one forward and one backward) from the source frame are utilized in the training. Our experiments found that total consecutive frames of 4 (2 forward and 2 backward frames) yielded the best performance in estimating the depth while addressing occlusion challenges.

DISCUSSION

Our comprehensive analysis encompassed quantitative evaluations, qualitative assessments, and detailed ablation studies, all of which underscore the effectiveness and innovation of our method. The proposed method demonstrates significant improvements in monocular depth estimation within endoscopic surgery contexts. Quantitative results highlight superior performance compared to SOTA models, particularly in Abs Rel and RMSE Log metrics. While the improvement of 2%-4% in RMSE compared to baselines using the standard reprojection loss may seem modest, we believe it is significant in the context of our study. In the field of self-supervised depth estimation, even small improvements can be critical in transitioning a prototype into a viable technology. These enhancements can lead to meaningful differences in real-world applications, especially when considering the cumulative effect of multiple incremental improvements. Qualitative analysis reveals LT-RL's ability to produce more continuous depth maps, excel in less textured

Table 3. Ablation study on less or more frames

Frames	Abs Rel (% ↓)	Sq Rel(% ↓)	RMSE (↓)	RMSE Log (↓)
2	0.062	0.513	5.289	0.094
4	**0.058**	**0.452**	**5.014**	**0.083**
6	0.611	0.448	5.209	0.091

Quantitative comparison with 2, 4 and 6 consecutive frames conducted on the SCARED dataset. The unit of % and millimeter (mm) of each metric is indicated in the bracket. The best results are in black bold. Abs Rel: Absolute relative error; Sq Rel: square relative error; RMSE: root-mean-squared error; RMSE Log: root-mean-square logarithmic error; SCARED: stereo correspondence and reconstruction of endoscopic data.

regions, and handle complex anatomical structures. Ablation studies underscore the importance of utilizing an optimal number of consecutive frames (in this case, 4) to maximize depth estimation performance while mitigating occlusion. While LT-RL does not affect the inference phase, its requirement for additional frames during training increases the computational overhead. Additionally, although our method demonstrates excellent generalization on the Hamlyn dataset, the specificity of our validation datasets suggests that further research is needed to fully understand LT-RL's performance across a broader range of endoscopic and surgical scenarios.

In conclusion, we present LT-RL by integrating longer temporal information to tackle occlusion artifacts in endoscopic surgery. Our extensive validation and comparison demonstrate the evidence that it is crucial to consider small camera pose changes in endoscopic surgery, and the proposed LT-RL addressed the issue successfully. The external validation of the Hamlyn dataset demonstrates the better robustness and generalization of the proposed method. Although LT-RL requires extra computation for the additional frames during training, there is no effect in the inference phase as there is no need for loss calculation in deployment. Our self-supervised loss is simple, flexible and easy to adapt to any network architecture of convolution and recent transformer-based models. The excellent 3D reconstruction reflects the better depth and pose learning and prediction of our LT-RL over other models. Future work should consider investigating the reliability of the LT-RL over vanilla reprojection loss. Computational efficiency can also be improved by using a shared encoder and an equal number of input frames for both depth and pose estimation tasks.

DECLARATIONS

Authors' contributions

Conceptualization, investigation, methodology, validation, visualization, writing - original draft, writing - review and editing: Shi X

Conceptualization, methodology, visualization, writing - original draft, writing - review and editing: Islam M

Conceptualization, methodology, writing - review and editing: Clarkson MJ

Conceptualization, validation, visualization, writing - review and editing: Cui B

Availability of data and materials

Our code is available at https://github.com/xiaowshi/Long-Term_Reprojection_Loss.

Financial support and sponsorship

This work was part-funded by the EPSRC grant [EP/W00805X/1].

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2024.

REFERENCES

1. Godard C, Mac Aodha O, Firman M, Brostow G. Digging into self-supervised monocular depth estimation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27 - Nov 02; Seoul, Korea. IEEE; 2019. pp. 3827-37. DOI
2. Godard C, Mac Aodha O, Brostow GJ. Unsupervised monocular depth estimation with left-right consistency. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honolulu, USA. IEEE; 2017. pp. 6602-11. DOI
3. Lyu X, Liu L, Wang M, et al. Hr-depth: high resolution self-supervised monocular depth estimation. *AAAI Conf Artif Intell* 2021;35:2294301. DOI
4. Garg R, Bg VK, Carneiro G, Reid I. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In: Computer Vision - ECCV 2016: 14th European Conference; 2016 Oct 11-14; Amsterdam, the Netherlands. Springer; 2016. pp. 740-56. DOI
5. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial transformer networks. 2015. Available from: https://proceedings.neurips.cc/paper_files/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf. [Last accessed on 5 Sep 2024]
6. Wang Z, Simoncelli EP, Bovik AC. Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003; 2023 Nov 09-12; Pacific Grove, USA. IEEE; 2003. pp. 1398-402. DOI
7. Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised learning of depth and ego-motion from video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honululu, USA. IEEE; 2017. pp. 6612-9. DOI
8. Ranjan A, Jampani V, Balles L, et al. Competitive collaboration: joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15-20; Long Beach, USA. IEEE; 2019. pp. 12232-41. DOI
9. Shao S, Pei Z, Chen W, et al. Self-supervised monocular depth and ego-motion estimation in endoscopy: appearance flow to the rescue. *Med Image Anal* 2022;77:102338. DOI
10. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015: 18th International Conference; 2015 Oct 5-9; Munich, Germany. Springer; 2015. pp. 234-41. DOI
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, USA. IEEE; 2016. pp. 770-8. DOI
12. Allan M, Mcleod J, Wang C, et al. Stereo correspondence and reconstruction of endoscopic data challenge. arXiv. [Preprint.] Jan 28, 2021 [accessed on 2024 Sep 5]. Available from: https://doi.org/10.48550/arXiv.2101.01133.
13. Shao S, Pei Z, Chen W, et al. Self-supervised learning for monocular depth estimation on minimally invasive surgery scenes. In: 2021 IEEE International Conference on Robotics and Automation (ICRA); 2021 May 30 - Jun 05; Xi'an, China. IEEE; 2021. pp. 7159-65. DOI
14. Recasens D, Lamarca J, Fácil JM, Montiel JMM, Civera J. Endo-depth-and-motion: reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robot Autom Lett* 2021;6:7225-32. DOI
15. Li W, Hayashi Y, Oda M, Kitasaka T, Misawa K, Mori K. Context encoder guided self-supervised siamese depth estimation based on stereo laparoscopic images. In: Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling. 2021. pp. 77-82. DOI
16. Zhang N, Nex F, Vosselman G, Kerle N. Lite-mono: a lightweight CNN and transformer architecture for self-supervised monocular depth estimation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17-24; Vancouver, Canada. IEEE; 2023. pp. 18537-46. DOI
17. Zhao C, Zhang Y, Poggi M, et al. Monovit: self-supervised monocular depth estimation with a vision transformer. In: 2022 International Conference on 3D Vision (3DV); 2022 Sep 12-16; Prague, Czech Pepublic. IEEE; 2022. pp. 668-78. DOI
18. Yang L, Kang B, Huang Z, Xu X, Feng J, Zhao H. Depth anything: unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. pp. 10371-81. Available from: https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Depth_Anything_Unleashing_the_Power_of_Large-Scale_Unlabeled_Data_CVPR_2024_paper.html. [Last accessed on 5 Sep 2024].