Lalith Sharan*, Lukas Burger*, Georgii Kostiuchik, Ivo Wolf, Matthias Karck,
Raffaele De Simone and Sandy Engelhardt*

Domain gap in adapting self-supervised depth estimation methods for stereo-endoscopy

<https://doi.org/10.1515/cdbme-2020-0004>

Abstract: In endoscopy, depth estimation is a task that potentially helps in quantifying visual information for better scene understanding. A plethora of depth estimation algorithms have been proposed in the computer vision community. The endoscopic domain however, differs from the typical depth estimation scenario due to differences in the setup and nature of the scene. Furthermore, it is unfeasible to obtain ground truth depth information owing to an unsuitable detection range of off-the-shelf depth sensors and difficulties in setting up a depth-sensor in a surgical environment. In this paper, an existing self-supervised approach, called *Monodepth* [1], from the field of autonomous driving is applied to a novel dataset of stereo-endoscopic images from reconstructive mitral valve surgery. While it is already known that endoscopic scenes are more challenging than outdoor driving scenes, the paper performs experiments to quantify the comparison, and describe the domain gap and challenges involved in the transfer of these methods.

Keywords: computer assisted surgery; depth estimation; endoscopy.

Introduction

The task of depth estimation is a commonly encountered problem in computer vision. Beyond prevalent applications in the field of autonomous driving and robotic navigation, depth estimation also finds use for endoscopy in a

surgical setup. Particularly in the case of minimally invasive surgeries, depth estimation shows promise in assisting navigation, quantification of visual information, and facilitating an improved understanding of the surface anatomy. 2D endoscopes, which are more commonly used in a surgical scenario, compromise depth perception. On the other hand, 3D endoscopes provide on-screen depth perception through stereo-imaging, from which the data can be captured intra-operatively and further analysed. Reconstructive mitral valve surgery is one such application, that is used to treat mitral valve insufficiency, a valvular disease of the heart [2]. Modern mitral valve repair is commonly done in a minimally invasive setup. In this case, intra-operative depth estimation could be extremely valuable to support decision making based on quantification [3] of endoscopic data.

Depth estimation has been tackled through various approaches in the literature. Sparse depth estimation methods focus on identifying matching feature points, or matching image patches [4]. In the case of dense depth estimation approaches, large datasets can be leveraged along with convolutional neural networks to predict depth for unseen examples. The datasets comprise of depth information acquired by depth sensors such as infrared or LiDAR cameras [5] as the ground truth to supervise the learning. In the case where the ground truth information is not available, the supervision comes from motion or binocular parallax, in other words additional information from the temporal or spatial domain [6]. In mitral valve surgery, the acquisition of ground truth depth information is unfeasible due to logistical and safety considerations. Furthermore, the off-the-shelf depth sensors typically operate in a range much wider than suitable for surgery. Besides, endoscopic scenes in the case of mitral valve repair are prone to specularities, reflection and occlusion artefacts. Occlusions occur due to tissue or instruments partially obstructing the endoscopic field of view, and may persist for a major part of the surgery. On the other hand, artefacts like specularities and reflection occur sporadically depending on the lighting conditions of the scene. The paper examines how existing self-supervised depth estimation approaches address this domain gap in endoscopy, in particular for mitral valve repair.

***Corresponding authors: Lalith Sharan and Sandy Engelhardt**, WG Artificial Intelligence in Cardiovascular Medicine (AICM), University Hospital Heidelberg, Heidelberg, Germany; and Informatics for Life, Heidelberg, Germany, E-mail: lalithnag.sharangururaj@med.uni-heidelberg.de (L. Sharan); and **Lukas Burger**, WG Artificial Intelligence in Cardiovascular Medicine (AICM), University Hospital Heidelberg, Heidelberg, Germany

Georgii Kostiuchik and Ivo Wolf, Department of Computer Science, Mannheim University of Applied Sciences, Mannheim, Germany

Matthias Karck and Raffaele De Simone, Department of Cardiac Surgery, University Hospital Heidelberg, Heidelberg, Germany

Related work

Dense depth estimation approaches aim to predict a depth value for each pixel of the input image. The input can either be a single frame or a sequence of frames from a monocular or stereo-camera. Learning based approaches compare patches of stereo-images to estimate correspondences and in turn the disparity, which can be used to compute the depth [7]. In endoscopy, [8] used *Hybrid Recursive Matching* to estimate the depth and model intra-operative surface deformation. More recently, the introduction of large datasets such as [5] enabled supervised training of convolutional neural networks for depth estimation. Some such approaches include comparing image patches to estimate the disparities using a *Siamese* architecture [9], and using an encoder-decoder architecture for end-to-end learning of depth from ground truth information [10]. While ground truth information is not easy to obtain in endoscopy, [11] demonstrated the use of a synthetic dataset created from CT images to train a network for depth estimation.

Self-supervised learning approaches

The approaches pertaining to monocular depth estimation without ground truth data, typically rely on a sequence of frames as a form of self-supervision. A survey of such monocular depth estimation techniques is provided by [6]. In the case of depth estimation using multi-view data, the shift information of the pixels between the views can be used to estimate the depth. [12] used a Structure-from-Motion (SfM) approach to predict depth from a monocular endoscopic video sequence. The approach uses sparse depth information as a form of supervision to generate a dense depth map.

Godard et al. [13] provided a self-supervised network that imposed left-right consistency from a stereo-pair of images during training time, to predict depth from monocular images during test time. *Monodepth* [1], the approach considered in this paper, further improved upon [13] by using a minimum reprojection loss, pixel auto-masking and a multi-scale encoder-decoder architecture. The network is capable of learning from both temporal (sequence of frames) and spatial (stereo-image pair) information without requiring the use of ground truth data. Therefore, this approach is chosen as a relevant candidate for depth prediction of stereo-endoscopic images from mitral valve repair. In this paper, three variants of the *Monodepth* [1] model (Figure 1) are trained on a combination of the KITTI [5] and Mitral Valve datasets.

Figure 1: An overview of the different models used in [1].

Methods

Monodepth architecture

The *Monodepth* [1] network predicts a dense depth map from an RGB image as input. During training, the network takes as input either a sequence of temporal frames or the corresponding stereo frame, and learns to reproject them to the frame at $t=0$ with depth as an intermediate variable. In the case of temporal frames, a pose network is trained along with the depth network to learn the pose of the camera relative to the frames. In the case of a stereo-pair, the pose that is known from camera calibration is used to reproject the image. A reprojection error is minimised along with an additional smoothness constraint. The depth prediction network uses a multi-scale encoder-decoder architecture, with the depth maps at each scale of the decoder being upsampled to the original input resolution. For further details, the reader is referred to [1].

Preparation of the mitral valve dataset

The dataset comprises stereo-images captured intra-operatively from mitral valve reconstruction surgeries. The images are acquired at a resolution of 1920×540 as a video at $25fps$. The stereo-endoscope used for the data acquisition is an *Image1S* 3D Full HD camera (Karl Storz, Tuttlingen, Germany). The frames from the video are extracted, shuffled and split into train (64%), validation (16%) and test (20%) datasets. The intrinsics of the stereo-endoscope is computed from calibration on a checkerboard. The calibrated camera parameters are then used to undistort and row-rectify the images. The images are normalised in the range $[0,1]$.

Experiments

Three variants of the *Monodepth* [1] model are used for the experiments. The *mono* model uses temporal frames as the input, the *stereo* model uses the corresponding stereo frame as the input, and the *mixed* model uses both the temporal and the corresponding stereo frames as input (Figure 1). Firstly, the three variants of the model are trained on the Mitral Valve dataset. Secondly, the models that are pre-trained on the KITTI [5] dataset are further fine-tuned on the Mitral Valve dataset. Finally, the models that are trained on the KITTI [5] dataset are directly used for prediction on the Mitral Valve dataset. In order to facilitate quantitative analysis, a test set comprising 299 pairs of images is prepared by identifying sparse corresponding points between the left and right images of the stereo-pair. The results of the nine experiments on this test set is presented in Table 1.

Training with temporal frames is performed on the KITTI [5] dataset using the split of Zhou et al. [14] and the complete Eigen et al.

Table 1: Error metrics computed from the predicted and ground truth depth of the points identified in the evaluation dataset. The three model variants used are as shown in Figure 1. K, M and M + K refer to models trained on the KITTI [5], Mitral Valve and both datasets respectively.

(Lower is better)				
Metric	Data	Mono	Stereo	Mixed
MRE	K	0.370	0.479	0.494
	M	0.649	1.935	2.752
	K + M	0.651	1.200	0.543
RMSE	K	0.184	0.239	0.241
	M	0.324	1.025	1.518
	K + M	0.354	0.598	0.343

[15] split is used for stereo-only training. The network is trained on a total of 39,810 sets of images and validated on 4,424 sets of images (a set is the image plus the preceding and succeeding frames). The Mitral Valve dataset used for training comprises 2,212 pairs of images from mitral valve repair surgeries. The images are resized and cropped to 352×192 , maintaining the aspect ratio. The depth predictions refer to the metric depths in case of stereo-input and relative depths in the case of monocular input. For monocular input, median scaling is performed as described in [1]. For all the experiments, the network is trained for 20 epochs with a learning rate of 0.0001 and *Adam* optimiser.

Results

Figure 2 shows sample predictions on the Mitral Valve dataset of three different network variants trained on the KITTI [5] dataset. An overview of the three different models used, namely *mono*, *stereo* and *mixed*, is provided in Figure 1. Similarly, Figure 3 shows the three variants of the model, first trained on the KITTI [5] dataset and further finetuned on the Mitral Valve dataset. Since the model is trained to converge for a baseline of 0.1, the outputs

obtained from the stereo training are scaled by a factor of 0.391, to adjust for the baseline of the stereo-endoscope.

The depth map predictions on the Mitral Valve Dataset are further evaluated on the test set to provide a quantitative comparison. The metrics of *Mean Relative Error (MRE)* and the *Root Mean Square Error (RMSE)* are computed for each image in the test set and then averaged over the test set. Table 1 shows the results of this evaluation and presents the comparison of the results obtained from the different experiments. Table 1 indicates that the *mono* model trained on the KITTI [5] dataset is the best performing model, and there is a decrease in performance when trained on the Mitral Valve dataset.

Discussion

From Figure 2, it can be seen that while predictions on the Mitral Valve dataset results in a smooth depth map, the predicted depth map does not completely align with the objects in the scene. Further training on the Mitral Valve dataset introduces texture copy artefacts in *stereo* and *mixed* training, as can be noticed from Figure 3. Additionally, while training on both datasets, the network finds it harder to adapt to the new endoscopic dataset, due to differences in the nature of the scene and the calibration parameters. In the case of stereo-endoscopy, especially in mitral valve repair, irregular lighting conditions give rise to specularities that are not constant between the views of a stereo-camera. This is in contrast to the constant light source assumption made in autonomous driving. Moreover, in autonomous driving and in other endoscopic applications such as colonoscopy, a moving camera enables the use of temporal supervision for depth estimation. The

Figure 2: Depth prediction results on the stereo-endoscopic images from reconstructive mitral valve surgery. The network is trained on the KITTI dataset [5]. A *min-max* normalisation is performed on the depth map for visual representation and a reverse *magma* colormap is applied. Lighter areas represent regions closer to the camera and darker regions represent distant regions. B, C and D represent mono, stereo and mixed models respectively.

Figure 3: Depth prediction results of the network trained on the stereo-endoscopic images from mitral valve repair surgery. A *min-max* normalisation is performed on the depth map for visual representation and a reverse *magma* colormap is applied. Lighter areas represent regions closer to the camera and darker regions represent distant regions. B, C and D represent mono, stereo and mixed models respectively.

setup in mitral valve repair involves a static camera with partially moving objects in the scene.

In conclusion, it is a non-trivial task to transfer the self-supervised depth estimation methods as such from the field of autonomous driving to the stereo-endoscopic domain. It is indeed known that endoscopic scenes are more challenging than outdoor driving scenes. However, the paper performs experiments with *Monodepth* [1] to compare quantitatively the domain gap and describe the challenges involved in the transfer of these methods.

Research funding: The research was supported by the German Research Foundation DFG Project 398787259, DE 2131/2-1 and EN 1197/2-1 and by Informatics for Life funded by the Klaus Tschira Foundation.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Authors state no conflict of interest.

Informed consent: Informed consent has been obtained from all individuals included in this study.

Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

- Godard C, Aodha OM, Firman M, Brostow GJ. Digging into self-supervised monocular depth estimation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South); 2019:3827–37 pp. <https://doi.org/10.1109/ICCV.2019.00393>.
- Carpentier A, Deloche A, Dauptain J, Soyer R, Blondeau P, Piwnica A, et al. A new reconstructive operation for correction of mitral and tricuspid insufficiency. *J Thorac Cardiovasc Surg* 1971;61:1–13.
- Engelhardt S, De Simone R, Al-Maisary S, Kolb S, Karck M, Meinzer HP, et al. Accuracy evaluation of a mitral valve surgery assistance system based on optical tracking. *Int J Comput Assist Radiol Surg* 2016;11:1891–904.
- Usenko V, Engel J, Stückler J, Cremers D. Reconstructing street-scenes in real-time from a driving car. In: 2015 International Conference on 3D Vision, Lyon; 2015:607–14. <https://doi.org/10.1109/3DV.2015.75>.
- Menze M, Heipke C, Geiger A. Joint 3D estimation of vehicles and scene flow. In: ISPRS annals of the photogrammetry, remote sensing and spatial information sciences; 2015:427–34 pp.
- Bhoi A. Monocular depth estimation: a survey. *arXiv* 2019, arXiv: 1901.09402.
- Hirschmüller H. Semi-global matching–motivation, developments and applications. In: Fritsch D, editor. *Photogrammetric Week' 11*. Wichmann Verlag, ISBN 978-3-87907-507-2; 2011: 173–84 pp.
- Röhl S, Bodenstedt S, Suwelack S, Kenngott H, Müller-Stich BP, Dillmann R, et al. Dense GPU-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration. *Med Phys* 2012;39:1632–45.
- Žbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches. *J Mach Learn Res* 2016;17:2287–318.
- Kendall A, Martirosyan H, Dasgupta S, Henry P, Kennedy R, Bachrach A, et al. End-to-end learning of geometry and context for deep stereo regression. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice; 2017:66–75 pp. <https://doi.org/10.1109/ICCV.2017.17>.
- Mahmood F, Chen R, Durr NJ. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE Trans Med Imag* 2018;37:2572–81.
- Liu X, Sinha A, Ishii M, Hager GD, Reiter A, Taylor RH, et al. Dense depth estimation in monocular endoscopy with self-supervised learning methods. In: *IEEE transactions on medical imaging*, vol. 39; 2020:1438–47 pp. <https://doi.org/10.1109/TMI.2019.2950936>.

13. Godard C, Aodha OM, Brostow GJ. Unsupervised monocular depth estimation with left-right consistency. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI; 2017:6602–11 pp. <https://doi.org/10.1109/CVPR.2017.699>
14. 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), Honolulu, HI; 2017:6612–19 pp. <https://doi.org/10.1109/CVPR.2017.700>.
15. Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago; 2015:2650–58 pp. <https://doi.org/10.1109/ICCV.2015.304>.