This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Synergies and Conflicts in Multi-Task Surgical Vision: Stereo Matching, Segmentation, and Classification of Stereoendoscopic Video Frames in Mitral Valve Surgery

Julia Mellert[1,2,3][0009−0004−9919−0058]⋆, Julius Kaemmel[5][0000−0002−5640−0543], Joerg Kempfert[5,6][0000−0002−2009−6375], Volker Markl[3,7][0009−0009−0964−026X], Simon Sündermann[5,6][0000−0003−4927−1584], Steffen Zeuch,[3,7][0000−0002−4082−7788], Olaf Hellwich[1][0000−0002−2871−9266], and Anja Hennemuth[2,3,4,6][0000−0002−0737−7375]

[1] Department of Computer Vision and Remote Sensing, TU Berlin, Germany
[2] Institute for Cardiovascular Computer-assisted Medicine, DHZC, Germany
[3] Berlin Institute for the Foundations of Learning and Data, Berlin, Germany
[4] Fraunhofer Institute for Digital Medicine MEVIS, Berlin, Germany
[5] Department of Cardiothoracic and Vascular Surgery, DHZC, Germany
[6] DZHK (German Centre for Cardiovascular Research), Berlin, Germany
[7] Department of Database Systems and Information Management, TU Berlin, Germany

Abstract. In minimally invasive and robot-assisted mitral valve surgery, direct vision is replaced by endoscopic imaging, which limits depth perception and increases the cognitive demands on surgeons. AI-based assistance systems could address this issue by allowing a holistic surgical understanding. We introduce MiTriSurv, a multi-task framework that simultaneously performs the classification and segmentation of surgical instruments, the segmentation of the mitral valve leaflets, and stereo-based depth estimation meeting near real time intraoperative requirements. We present an annotated stereo-endoscopic dataset from 20 patients. We perform a systematic analysis of the interactions between tasks, which shows that tissue segmentation benefits from geometric depth cues and that instrument classification benefits from joint training with instrument segmentation. However, combining mitral valve leaflet segmentation with instrument classification is counterproductive, demonstrating that task compatibility is a critical factor in multitask surgical vision.

Keywords: Multi-Task · Surgical Vision · Endoscopy · Mitral Valve.

1 Introduction

Minimally invasive procedures have increasingly become the standard in mitral valve surgery, offering less tissue damage, faster patient recovery, and smaller

⋆ Corresponding author: julia.mellert@dhzc-charite.de

surgical incisions [14, 17]. Robotic platforms further improve instrument handling, visibility, and surgical precision. However, the surgeon's direct view of the surgical field is replaced by an endoscopic image stream, significantly limiting depth perception [19]. This loss is particularly severe in mitral valve surgeries, where complex cardiac anatomy, limited surgical space, and the need for precise instrument coordination place high demands on surgeons, especially those with limited experience [14, 25, 5].

AI-based assistance systems could provide valuable support [8, 2, 9]. They should provide a comprehensive scene understanding from endoscopic videos [11] — simultaneously localizing and recognizing objects, and determining their geometric relations.

Multi-task learning offers several advantages that are well-suited to understand surgical scenes: Related tasks can reinforce each other through shared representations, overfitting is mitigated in the small datasets typical of clinical environments, and joint inference reduces memory and computational requirements — a critical constraint in real-time endoscopic applications [1]. The combination of object segmentation and depth estimation is well established [15, 22]. Recent work on colonoscopy data has incorporated depth priors into foundation model training [3] or combines instrument and tissue segmentation with monocular depth estimation, although without quantitative depth evaluation [7]. Joint training in instrument segmentation and classification is equally common and has been applied in the following competitions: HeiCo [10] and AutoLaparo [21] (surgery phase detection, instrument and tissue segmentation). A work [6] focuses on segmentation of instruments and the mitral valve. In addition, previous publications rarely report inference latency or memory footprint, making it difficult to assess their suitability for real-time intraoperative use.

We intend to simultaneously address the classification and segmentation of surgical instruments, tissue segmentation, and stereo-based depth estimation to analyze how these tasks interact to enable a holistic understanding of the surgical scene. We address this with a unified framework that combines all three tasks, explicitly analyzes their interactions, and evaluates computational efficiency in the context of mitral valve surgery. Our approach also differs from previous work on laparoscopy due to the low-contrast soft tissue boundaries and the subtle anatomy of the valve, as well as a quantitative depth assessment enabled by ruler-based ground truth. We make the following contributions:

- **MiTriSurv** (Mitral Valve Tri-task Surgical Vision): a multi-task architecture extending a state-of-the-art stereo-matching model with semantic segmentation of instruments and mitral valve leaflets, and instrument classification with close to real-time capability (20 fps).
- **A stereo-endoscopic dataset** with annotations of surgical instruments and mitral valve leaflets from 20 patients, created as part of this work.
- **A systematic analysis of task interactions**, revealing mutual benefits between stereo matching and tissue segmentation and between instrument recognition and segmentation, while characterizing conditions under which joint training is counterproductive.

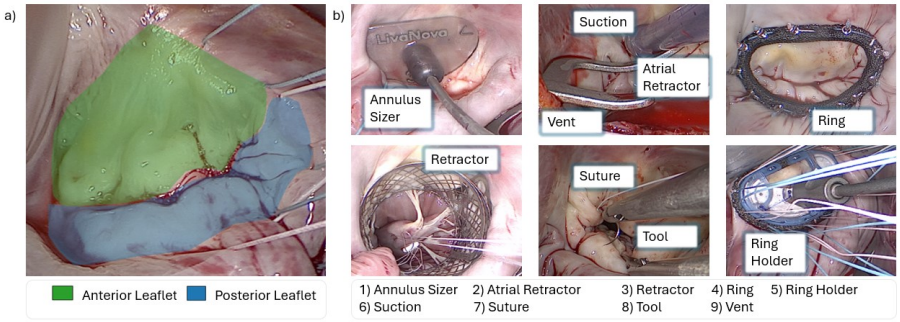

Fig. 1. Endoscopic view of the mitral valve and its leaflets with segmentation overlay (a) and the surgical instruments used in mitral valve repair (b).

2 Method

2.1 Dataset

Stereo endoscopic data were captured using the EinsteinVision 2.0 system (Aesculap) at 1920×1080 resolution and 30 frames per second (fps). The stereo streams were deinterlaced during preprocessing, followed by downsampling by a factor of 4. Stereo calibration was performed before stereo matching using a standard checkerboard-based approach [13, 4].

Frames were manually annotated by medical experts with pixel-wise masks for surgical instruments and the anterior and posterior mitral valve leaflets see Fig. 1a. Nine instrument classes (Fig. 1b) are included: Vent, Atrial Retractor, Tool, Annulus Sizer, Ring, Suction, Suture, Ring Holder, and Retractor, with class occurrences reported in Table 2. For training we use the S2M2-M model [12] to generate pseudo ground truth disparity maps.

The dataset covers 20 patient recordings and was approved by the local institutional ethics committee (EA4/130/22). It is split into training (528 frames, 14 patients), validation (71 frames, 3 patients), and test (111 frames, 3 patients), with a split at the patient-level to avoid data leakage. For depth evaluation, 17 frames with a ruler were selected and the ruler coordinates annotated, comprising $n = 624$ coordinates with distances in mm ($\mu = 2.68$, $\delta = 3.62$).

2.2 Network Architecture

An overview of the proposed architecture is shown in Fig. 2.

Shared Encoder We adopt the encoder from S2M2-S[12] to jointly process the left and right stereo images. Each image is first passed through a convolutional feature extraction module, producing feature maps at two spatial resolutions. These are fed into a Feature Pyramid Network, whose multi-scale outputs are subsequently processed by a Multi-Resolution Transformer (MRT). The MRT

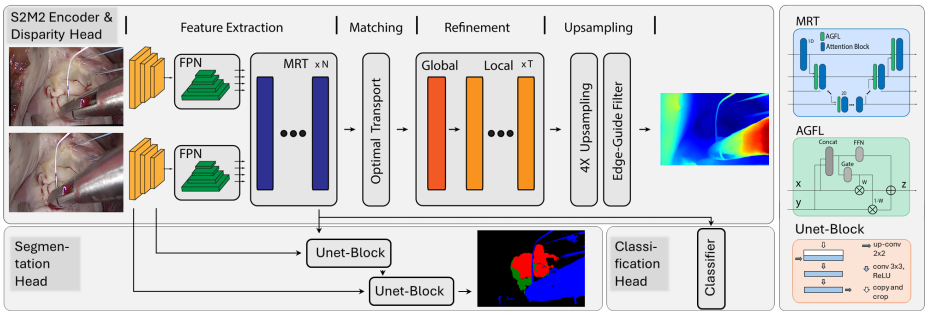

Fig. 2. Overview of the proposed multi-task architecture.

applies efficient 1D horizontal attention along epipolar lines at higher resolutions and full self- and cross-attention between the two stereo views at the lowest resolution, yielding a set of multi-resolution stereo feature maps. The highest-resolution feature map is forwarded to each of the task-specific decoder heads.

Segmentation Head The segmentation head follows a UNet-style [16] decoder design. The encoder feature map is upsampled by a factor of 2 via a transposed convolution, while the corresponding skip connection is channel-projected through a convolution. The two are concatenated and refined by two successive convolutional blocks with batch normalization, ReLU activation, and intermediate spatial dropout ($p = 0.1$). A final convolution projects the features to 4 class logits, producing a pixel-wise segmentation map.

Classification Head We apply a simple classification head design of global average pooling, layer normalization, and linear projection to produce class logits.

Stereo Matching Head To benefit from the superior performance of S2M2, we retain its decoder head unchanged. It consists of four stages: (1) an Adaptive Gated Feature Layer (AGFL) for efficient feature alignment, (2) optimal transport-based global matching via all-pairs correlation at 1/4 resolution, estimating disparity, occlusion, and confidence, (3) iterative disparity refinement, and (4) upsampling to the full output resolution.

2.3 Loss Formulation

For segmentation, cross-entropy loss $\mathcal{L}_{\text{CE}}$ is used. For stereo matching, Smooth-ℓ_1 loss $\mathcal{L}_{\text{SL1}}$ is applied to predicted disparity $\hat{d}_i$ and ground-truth disparity d_i. Instrument recognition uses binary cross-entropy $\mathcal{L}_{\text{BCE}}$.

The total training loss is a weighted sum of the segmentation, stereo matching, and instrument recognition objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{seg}}\mathcal{L}_{\text{CE}} + \lambda_{\text{stereo}}\mathcal{L}_{\text{SL1}} + \lambda_{\text{class}}\mathcal{L}_{\text{BCE}}. \tag{1}$$

The weights λ_{seg}, λ_{stereo}, and λ_{cls} were selected via grid search over $\Lambda = 0.01, 0.1, 1.0, 10.0$, evaluating 63 configurations. The best-performing setting was

$$\lambda_{\text{seg}} = 0.1, \qquad \lambda_{\text{stereo}} = 0.01, \qquad \lambda_{\text{cls}} = 0.1. \tag{2}$$

2.4 Training Strategy

For all experiments, we use the pretrained S2M2-S model [12] as our baseline checkpoint and use an identical training shcedule. The only exception is the ablation study on decoder capacity, where the M and XL variants are initialized from their respective pretrained checkpoints. Multi-task training including segmentation follows a two-stage schedule. In the first stage, we freeze the pretrained encoder as well as the stereo and classification heads and train the segmentation head in isolation. In the second stage, we unfreeze all components and perform joint multi-task training on top of this initialization.

3 Experiments

3.1 Evaluation Metrics

Segmentation performance is evaluated using Dice coefficient and Intersection over Union (IoU), reported per class and as background included macro averages.

Instrument recognition is evaluated with macro, micro and weighted F1 score.

The stereo matching is evaluated using supervised and unsupervised metrics. As a supervised depth evaluation we predict the disparity maps of the ruler images and combine them with camera-specific intrinsic parameters and the reprojection matrix Q from the calibration to project pixels into 3D world coordinates. For each pair of ruler coordinates, we calculate the Euclidean distance in the reconstructed 3D space and the mean squared error relative to the ground truth annotated distance. We use the median to summarize the ruler error. As an unsupervised metric, we calculate the photometric error and the SSIM error between the left image and the warped right image for the test dataset.

3.2 Tasks

Segmentation The segmentation results are shown in Table 1a and visualized in Fig. 3a. Single-task segmentation achieves a mDice of 0.774 and a mIoU of 0.639 and serves as a baseline for measuring the impact of additional tasks. Adding stereo matching as a task leads to consistent improvements (mDice:+2.33%), with the gains being particularly pronounced in the anterior and posterior leaflets (Anterior Dice: +4.18%; Posterior Dice: +4.40%). However, including classification as an additional task has the opposite effect: segmentation performance falls below baseline (mDice:−0.90%), with the leaflet classes being the most affected. The complete three-task model achieves the highest mDice value of 0.805 (+4.01%), which is due to the improved performance of instrument segmentation

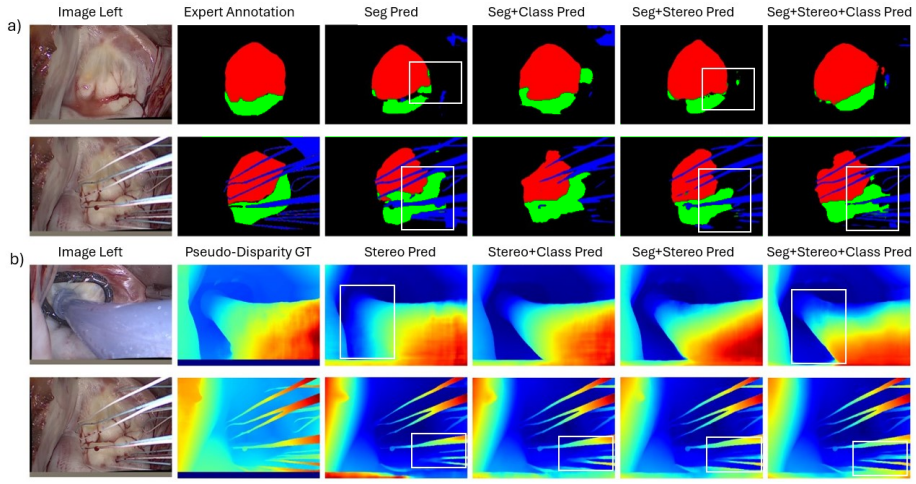

Fig. 3. Qualitative evaluation of the segmentation (a) and stereo matching (b) tasks. The Seg+Stereo combination improves the segmentation of leaves compared to Seg or Seg+Stereo+Class (see first row), while the segmentation of instruments benefits from the combination of all three tasks (see second row). Stereo matching does not derive any measurable benefit from joint training, and its performance deteriorates (see second row); however, advantages can be observed in some examples (see first row).

(Instrument Dice: +2.23%). Compared to the two-task combination with stereo matching, leaflet segmentation benefits less (Anterior Dice: −0.83%; Posterior Dice: −6.25%). When the segmentation task is divided into just targeting instrument or leaflet segmentation (Table 1b), the same pattern emerges: Leaflet segmentation benefits from the stereo task and instrument segmentation benefits slightly from the stereo combination, but most from the three-task combination.

Classification The classification results (Table 1a) for a single task yield an F1-macro score of 0.181, an F1-micro score of 0.701, and a class-averaged F1 score of 0.641, reflecting the difficulty of the task given a significant class imbalance. Joint training with segmentation improves performance (F1-macro: +32.04%; F1-micro: +10.98%; class-average F1: +27.30%), with the most improvements observed in rare classes (Table 2). This is particularly evident for Ring Holder (+214.29%), Retractor (+87.50%), and Vent (+325.00%), while the frequently occurring class Atrial Retractor shows a moderate increase (+7.40%). The classes Annulus Sizer, Suction, and Ring remain difficult (F1 $\leq$ 0.063). Adding stereo matching, does not provide a measurable benefit for classification (F1-macro: −0.55%; F1-micro: −0.43%; F1 class average: −0.78%). The complete three-task model achieves an average classification performance (F1 macro: +26.52%; F1-micro: +13.41%; F1-class average: +24.49%), in which the micro-F1 score continues to rise compared to Seg+Class, but the macro-score declines, indicating poorer classification of rare instruments.

Stereo-Matching Stereo matching (Table 1a and Fig. 3b) achieves the best ruler-based accuracy in the single-task setting, and adding auxiliary tasks consistently degrades it, although the extent varies: classification interferes less than segmentation, and the combination of all three tasks yields the highest relative median error (5.55%). The self-supervised metrics rank the multi-task configurations in the same order, but place the single-task model last. The two metric families therefore agree on the relative cost of each auxiliary task and disagree only on the single-task baseline.

Multi-Task Efficiency The proposed model processes video at 19.58 fps (on an NVIDIA A100 PCIe 40GB GPU) with only 26.7M parameters, demonstrating a compact and close to real-time-capable architecture.

3.3 Ablation Studies

Decoder Architecture Replacing our proposed decoder with a SegFormer decoder [23] yields comparable but slightly lower performance across tasks (Table 1c), suggesting that gains are not primarily driven by the decoder design.

Encoder Capacity To assess whether performance limitations stem from insufficient model capacity, we compare three backbone scales: S, M, and XL (Table 1c). Performance is largely consistent across scales. It indicates once sufficient capacity is reached, conflicting instrument classification and leaflet segmentation limitations in a three-task set-up are not due to encoder size.

Orthoganl Gradient Projection To exclude conflicting task gradients hindering optimization we applied gradient projection to the normal plane of the conflicting gradient [24, 18]. The experiments (Table 1c) show no clear improvement indicating no gradient conflict in optimzation.

4 Discussion and Conclusion

This paper presents a multi-task learning framework for understanding surgical scenes that jointly addresses semantic segmentation, instrument recognition, and stereo-based depth estimation; its efficiency and accuracy make it well-suited for use under resource-constrained intraoperative conditions.

The effects of multi-task learning on segmentation yields different results depending on the task. For instrument segmentation, combining tasks is generally helpful—especially when using the full multi-task setup. For tissue segmentation, adding a classification task does not improve performance, likely because the semantic categories do not align well with the morphology of the heart valve. However, adding depth information from the stereo image helps with fine-grained tissue structures, demonstrating that geometric information is particularly useful for homogeneous areas with low contrast.

Instrument classification benefits most from joint training, in fact when combined with semantic segmentation suggesting that segmentation-based representations provide additional semantic context particularly helpful for rare classes. This could be further investigated as a valuable transfer learning task.

In stereo matching, joint training with additional tasks leads to a decrease in performance, which underscores the strong one-shot capability of the original S2M2 model. In contrast to the results reported in other studies [15, 22], we do not observe any measurable benefit from segmentation-guided stereo training, possibly because pseudo-ground-truth disparity labels introduce misleading training signals and thus prevent the stereo branch from explicitly utilizing the additional edge information from segmentation. It is noteworthy that, for the single-task baseline, the supervised and unsupervised metrics diverge. Endoscopic scenes contain many textureless, reflective, and occluded regions that violate the assumptions underlying self-supervised metrics; therefore, we consider the ruler-based error as the primary criterion. The model demonstrates competitive stereo performance with a sparse depth error in the submillimeter range, thereby surpassing previous results from the literature in the millimeter range [20], which, however, often report dense depth metrics.

Overall, the benefits of multi-task learning depend on task compatibility. Tissue segmentation profits most from joint training with stereo-based depth estimation, where geometric cues sharpen low-contrast leaflet boundaries. Instrument-related tasks similarly benefit from shared representations, with classification gaining the most. However, combining leaflet segmentation with instrument classification proves incompatible regardless of backbone size or gradient conflict mitigation, suggesting differing representation requirements. These findings extend previous instrument-centric work [1] to a realistic intraoperative setting, underscoring that task compatibility is the decisive factor in multi-task surgical scene understanding. Nevertheless, these findings require validation in a larger patient cohort, given the small sample size and limited availability of instruments for certain classes.

Acknowledgments. This study was funded by the German Ministry for Education and Research (BMBF) under Grant BIFOLD26B.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Alabi, O., Vercauteren, T., Shi, M.: Multitask learning in minimally invasive surgical vision: A review. Medical Image Analysis **101**, 103480 (Apr 2025). https://doi.org/10.1016/j.media.2025.103480
2. van Amsterdam, B., Funke, I., Edwards, E., Speidel, S., Collins, J., Sridhar, A., Kelly, J., Clarkson, M.J., Stoyanov, D.: Gesture Recognition in Robotic Surgery With Multimodal Attention. IEEE Transactions on Medical Imaging **41**(7), 1677–1687 (Jul 2022). https://doi.org/10.1109/TMI.2022.3147640

3. Han, J.J., Schmidt, A., Jamal, M.A., Nwoye, C., Rau, A., Wu, J.Y., Mohareri, O.: On the Role of Depth in Surgical Vision Foundation Models: An Empirical Study of RGB-D Pre-training (2026). https://doi.org/10.48550/ARXIV.2601.18929
4. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge, 2 edn. (2004). https://doi.org/10.1017/CBO9780511811685
5. Hayashi, Y., Nakamura, Y., Hirano, T., Ito, Y., Watanabe, T.: Cumulative sum analysis for the learning curve of minimally invasive mitral valve repair. Heart and Vessels **36**(10), 1584–1590 (Oct 2021). https://doi.org/10.1007/s00380-021-01838-7
6. Ivantsits, M., Tautz, L., Sündermann, S., Wamala, I., Kempfert, J., Kuehne, T., Falk, V., Hennemuth, A.: DL-based segmentation of endoscopic scenes for mitral valve repair. Current Directions in Biomedical Engineering **6**(1) (May 2020). https://doi.org/10.1515/cdbme-2020-0017
7. Jiang, X., Lu, D., Cao, S., Shao, L., Fan, J., Yang, J.: Multi-Task Learning for Optical Flow-Guided Self-Supervised Depth Estimation and Semantic Segmentation in Endoscopic Surgery. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 5463–5469 (Dec 2025). https://doi.org/10.1109/BIBM66473.2025.11356430
8. Lanfranco, A.R., Castellanos, A.E., Desai, J.P., Meyers, W.C.: Robotic Surgery: A Current Perspective. Annals of Surgery **239**(1), 14 (Jan 2004). https://doi.org/10.1097/01.sla.0000103020.19595.7d
9. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., Hashizume, M., Katic, D., Kenngott, H., Kranzfelder, M., Malpani, A., März, K., Neumuth, T., Padoy, N., Pugh, C., Schoch, N., Stoyanov, D., Taylor, R., Wagner, M., Hager, G.D., Jannin, P.: Surgical data science for next-generation interventions. Nature Biomedical Engineering **1**(9), 691–696 (Sep 2017). https://doi.org/10.1038/s41551-017-0132-7
10. Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P.M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., Bruno, P., Kisilenko, A., Müller, B., Davitashvili, T., Capek, M., Tizabi, M.D., Eisenmann, M., Adler, T.J., Gröhl, J., Schellenberg, M., Seidlitz, S., Lai, T.Y.E., Pekdemir, B., Roethlingshoefer, V., Both, F., Bittel, S., Mengler, M., Mündermann, L., Apitz, M., Kopp-Schneider, A., Speidel, S., Nickel, F., Probst, P., Kenngott, H.G., Müller-Stich, B.P.: Heidelberg colorectal data set for surgical data science in the sensor operating room. Scientific Data **8**(1), 101 (Apr 2021). https://doi.org/10.1038/s41597-021-00882-2
11. Mascagni, P., Alapatt, D., Sestini, L., Altieri, M.S., Madani, A., Watanabe, Y., Alseidi, A., Redan, J.A., Alfieri, S., Costamagna, G., Boškoski, I., Padoy, N., Hashimoto, D.A.: Computer vision in surgery: from potential to clinical value. npj Digital Medicine **5**(1), 163 (Oct 2022). https://doi.org/10.1038/s41746-022-00707-5
12. Min, J., Jeon, Y., Kim, J., Choi, M.: S^2M^2: Scalable stereo matching model for reliable depth estimation (Oct 2025). https://doi.org/10.48550/arXiv.2507.13229
13. Papadimitriou, D.V., Dennis, T.J.: Epipolar line estimation and rectification for stereo image pairs. Trans. Img. Proc. **5**(4), 672–676 (Apr 1996). https://doi.org/10.1109/83.491345
14. Praz, F., Borger, M.A., Lanz, J., Marin-Cuartas, M., Abreu, A., Adamo, M., Ajmone Marsan, N., Barili, F., Bonaros, N., Cosyns, B., De Paulis, R., Gamra, H., Jahangiri, M., Jeppsson, A., Klautz, R.J.M., Mores, B., Pérez-David, E., Pöss, J., Prendergast, B.D., Rocca, B., Rossello, X., Suzuki, M.,

Thiele, H., Tribouilloy, C.M., Wojakowski, W., ESC/EACTS Scientific Document Group: 2025 ESC/EACTS Guidelines for the management of valvular heart disease. European Heart Journal **46**(44), 4635–4736 (Nov 2025). https://doi.org/10.1093/eurheartj/ehaf194
15. Psychogyios, D., Mazomenos, E., Vasconcelos, F., Stoyanov, D.: MSDESIS: Multitask Stereo Disparity Estimation and Surgical Instrument Segmentation. IEEE Transactions on Medical Imaging **41**(11), 3218–3230 (Nov 2022). https://doi.org/10.1109/TMI.2022.3181229
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation (2015)
17. Sündermann, S.H., Czerny, M., Falk, V.: Open vs. Minimally Invasive Mitral Valve Surgery: Surgical Technique, Indications and Results. Cardiovascular Engineering and Technology **6**(2), 160–166 (Jun 2015). https://doi.org/10.1007/s13239-015-0210-5
18. Tseng, W.C.: (2020), https://github.com/WeiChengTseng/Pytorch-PCGrad
19. Vajsbaher, T., Schultheis, H., Francis, N.K.: Spatial cognition in minimally invasive surgery: a systematic review. BMC surgery **18**(1), 94 (Nov 2018). https://doi.org/10.1186/s12893-018-0416-1
20. Wang, X., Yang, B., Wei, M., Liu, L., Zhang, J., Nie, Y.: Deep learning for endoscopic depth estimation: A review. Displays **90**, 103086 (Dec 2025). https://doi.org/10.1016/j.displa.2025.103086
21. Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y.: AutoLaparo: A New Dataset of Integrated Multi-tasks for Image-guided Surgical Automation in Laparoscopic Hysterectomy (Aug 2022). https://doi.org/10.48550/arXiv.2208.02049
22. Wu, R., He, C., Liang, P., Liu, Y., Huang, Y., Liu, W., Shu, B., Xu, P., Chang, Q.: MCF-SMSIS: Multi-tasking with complementary functions for stereo matching and surgical instrument segmentation. Computers in Biology and Medicine **179**, 108923 (Sep 2024). https://doi.org/10.1016/j.compbiomed.2024.108923
23. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers (Oct 2021). https://doi.org/10.48550/arXiv.2105.15203
24. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning (2020)
25. Zhong, L., Huang, H.: Application and development of totally thoracoscopic mitral valve plasty. Chinese Journal of Clinical Thoracic and Cardiovascular Surgery **30**(3), 458–463 (2023). https://doi.org/10.7507/1007-4848.202109005

Table 1. Comparison of task combinations, target-specific analysis, and ablation studies.

	Segmentation					Classification			Stereo Matching				Efficiency		
Section	**mDice**	**mIoU**	**Instr.**	**Ant.**	**Post.**	**F1-mac**	**F1-mic**	**F1-cls**	**Rel.Err [%]**	**Abs.Err [cm]**	**SSIM**	**Photo.**	**FPS**	**Params**	**Mem [MB]**
(a) All-task combinations															
Seg	0.774	0.639	0.851	0.813	0.659	–	–	–	–	–	–	–	–	–	–
Stereo	–	–	–	–	–	–	–	–	3.837	0.006	0.227	0.347	–	–	–
Class	–	–	–	–	–	0.181	0.701	0.641	–	–	–	–	–	–	–
Seg + Stereo	0.792	0.663	0.842	0.847	0.688	–	–	–	4.757	0.009	0.249	0.335	–	–	–
Seg + Class	0.767	0.630	0.839	0.815	0.657	0.239	0.778	0.816	–	–	–	–	–	–	–
Stereo + Class	–	–	–	–	–	0.180	0.698	0.636	3.973	0.007	0.250	0.335	–	–	–
Seg + Stereo + Class	0.805	0.653	0.870	0.840	0.645	0.229	0.795	0.798	5.554	0.010	0.249	0.337	19.58	26.7M	1169.91
(b) Tissue & instrument segmentation (target-specific)															
Instrument															
Seg	0.941	0.899	0.822	–	–	–	–	–	–	–	–	–	–	–	–
Seg + Stereo	0.946	0.907	0.837	–	–	–	–	–	5.317	0.01	0.250	0.337	–	–	–
Seg + Stereo + Class	0.957	0.924	0.871	–	–	0.240	0.881	0.838	6.367	0.011	0.250	0.337	–	–	–
Valve Leaflet															
Seg	0.837	0.65	–	0.840	0.670	–	–	–	–	–	–	–	–	–	–
Seg + Stereo	0.847	0.757	–	0.851	0.691	–	–	–	4.385	0.009	0.247	0.337	–	–	–
Seg + Stereo + Class	0.822	0.725	–	0.825	0.640	0.200	0.747	0.726	5.378	0.009	0.249	0.337	–	–	–
(c) Ablation															
encoder size M	0.791	0.660	0.826	0.705	0.843	0.213	0.790	0.735	5.553	0.010	0.248	0.337	11.94	80.81M	2118.23
encoder size XL	0.810	0.690	0.889	0.686	0.856	0.227	0.789	0.776	4.434	0.010	0.249	0.337	8.75	405.1M	7684.08
segformer decoder	0.762	0.630	0.864	0.608	0.813	0.224	0.781	0.787	4.667	0.008	0.250	0.337	18.25	26.7M	1170.9
orthogonal gradient	0.781	0.650	0.854	0.652	0.838	0.228	0.800	0.783	5.149	0.009	0.249	0.337	19.23	26.7M	1171.60

Table 2. Per-class F1-score for different task combinations. Class occurrence denotes the percentage of samples in which the respective instrument appears.

	Annulus Sizer	**Atrial Retractor**	**Retractor Ring**	**Ring Holder**	**Suction**	**Suture**	**Tool**	**Vent**
Occurrence (%)	4	85	6	20	1	15	59	38	31
Class	0.000	0.730	0.072	0.063	0.063	0.000	0.378	0.288	0.036
Class + Seg	0.009	0.784	0.135	0.063	0.198	0.000	0.514	0.297	0.153
Class + Stereo	0.000	0.910	0.000	0.000	0.090	0.000	0.333	0.207	0.081
Class + Seg + Stereo	0.009	0.856	0.081	0.063	0.171	0.000	0.405	0.288	0.126