Transferring Relative Monocular Depth to Surgical Vision with Temporal Consistency

Charlie Budd[1] and Tom Vercauteren[1,2]

[1] King's College London, Biomedical Engineering & Imaging Science, London
[2] Hypervision Surgical Limited, 1st Floor 85 Great Portland Street, London

Abstract. Relative monocular depth, inferring depth correct up to a shift and scale from a single image, is an active research topic. Recent deep learning models, trained on large and varied meta-datasets, now provide excellent performance in the domain of natural images. However, few datasets exist which provide ground truth depth for endoscopic images, making training such models from scratch unfeasible. This work investigates the transfer of these models into the surgical domain, and presents an effective and simple way to improve on standard supervision through the use of temporal consistency self-supervision. We show temporal consistency significantly improves supervised training alone when transferring to the low-data regime of endoscopy, and outperforms the prevalent self-supervision technique for this task. In addition we show our method drastically outperforms the state-of-the-art method from within the domain of endoscopy. We also release our code, models, and ensembled meta-dataset, `Meta-MED`, establishing a strong benchmark for future work.

Keywords: Monocular Depth · Transfer Learning · Self-supervision · Surgical Vision · Computer Assisted Intervention.

1 Introduction

The task of monocular depth in computer vision focuses on estimation from a single viewpoint, as opposed to using multi-view geometry. Example depth maps are shown in Fig. 1. Throughout this work, we focus on dense depth, wherein the depth is estimated for every pixel in the image, and relative depth, meaning the estimated depth values are only sought up to scale and shift. Targeting relative depth, as proposed in [20], allows training on data from multiple sources without concern for calibration or large expected range differences between applications. State-of-the-art methods for natural images have exploited this property to leverage huge *meta-datasets*, composed of multiple existing datasets, to train large transformer based models. The most recent MiDaS models [4], the culmination of a series of recent works [19, 20], are trained on 1.4 million labelled images. Depth Anything [30], released during the later stages of our work, iterates on MiDaS and trains on 1.5 million labelled images and 62 million unlabelled images.

An adequate solution to monocular depth in surgery would act as a key enabler for many surgical vision tasks. It would be directly useful for tasks such

Fig. 1. Example images from the natural image and endoscopic domains with corresponding inverse relative depth-maps. The depth-maps are generated using a MiDaS model pre-trained on natural images. The ability of the model to transfer to the endoscopic domain is notable and speaks to the fundamental nature of the depth estimation task. However, closer inspection reveals significant flaws in the estimated depth-map.

as intraoperative autofocus [6], AR overlays [18], and surgical site mapping [21]. More generally, though, the ability of a model to correctly discern depth from an endoscopic image demonstrates a nuanced understanding of the geometry of a surgical scene, which may be key for better solving more complex vision tasks. However, as very few surgical datasets with ground truth depth exist, following the same approach as state-of-the-art natural image models is unfeasible. Experimenting with these models, we observed that one of the main failure modes when inferring outside their training domain, and specifically in surgery, seems to be an uncertainty in the depth ordering of semantic elements. This leads to temporal inconsistent predictions, an aspect noticed by other works [28]. We therefore choose to investigate the transfer of these models to the surgical domain while following the paradigm of ensembled meta-datasets. We focus our attention on temporal consistency which provides a path for self-supervised training.

In the natural image domain, prior to the advent of current large datasets and models, structure from motion (SFM) approaches had been used to generate pseudo ground truth depth [33, 15]. This line of self-supervision approaches still dominates monocular depth in endoscopy [14, 24, 13, 12]. Due to the dynamic nature of surgery and the highly deformable nature of tissue and articulated surgical tools, these methods are restricted to training on videos specifically collected for such tasks, featuring static tissues with little to no tool interaction. To improve results in gastro-endoscopy, geometric prior knowledge have been exploited [31]. While elegant, the tailored nature of such solutions typically reduces the generalisability to other surgical domains. To the best of our knowledge, the state-of-the-art in general surgical vision with released models is AF-SFMLearner [24]. An incremental improvement over this method has been reported with WS-SFMLearner [14] but this model has not been publicly released. Beyond SFM-based methods, in [7] the authors proposed to use large pretrained

natural image models as a starting point for endoscopy. They fine-tune DinoV2's depth estimation model [17] to various individual endoscopy depth datasets. Due to the small sizes and low variability of the datasets used, generalisability is a natural concern. As they don't release the models, and present results for depth up to scale only rather than scale and shift, benchmarking against their reported results is not possible.

In this work, we start by compiling `Meta-MED`, the first meta-dataset intended to train and evaluate relative monocular depth models in the endoscopic domain. We then demonstrate the ability of large state-of-the-art relative depth models for natural images to generalise reasonably to our domain and that standard supervised fine tuning on our meta-dataset provides a strong boost to performance. We also experiment with adding self-supervision methods to the training pipeline to great effect, and demonstrate that our novel temporal consistency self-supervision outperforms a typical self-supervision approach. We show that many of our models outperform our domain-specific state-of-the-art baseline AF-SFMLearner. Finally, we release our code[3], models, and dataset, establishing a strong benchmark for future research.

2 Materials and Methods

Datasets Two high-quality datasets for dense depth estimation in minimally invasive surgery are the SCARED [2] and SERV-CT [8] datasets. The ground truth for these two datasets are collected using two different measuring techniques. SCARED uses structured light projection with a stereo camera to create sparse depth maps, whereas SERV-CT creates dense depth maps using a pre-operative CT which is registered to image space. Both of these datasets are small, consisting of just 45 and 16 images captured in 9 and 2 ex-vivo porcine samples, respectively. While SCARED also has approximate depth maps for more frames, these are made by reconstruction between the ground truth frames. We ignore these to reduce errors and redundancy in the data. Due to the small size of these datasets, we choose to combine these to form a holdout testing dataset.

Our first fine-tuning approach relies on supervision with pseudo ground truth. We make use of an ensemble of stereo endoscopic datasets with depth maps calculated using stereo disparity, a full list of which can be seen in Table 1. The stereo data totals a little over 3 hours of footage and is drawn from 57 surgeries, most of which were in-vivo porcine. The videos are first sampled to one frame per second to reduce redundancy in the data. The left and right images are then stereo rectified and cropped to remove borders where needed. RAFT [26], a capable and easily available optical flow model which has seen previous usage in surgical videos [11, 23] , is then used to calculate the bidirectional optical flow maps $F_{a\to b}$ and $F_{b\to a}$. In order to only provide supervision for regions of the image with a confident ground truth, we mask out correspondences where the bidirectional flow does not form a closed loop. We define a warping operator

[3] https://github.com/charliebudd/transferring-relative-monocular-depth-to-surgical-vision

Table 1. A tabulation of all the datasets used in this work. The domain column summarises the type of data, specifically, how many samples/surgeries, whether they were in-vivo or ex-vivo, and whether they were porcine, human, or mixed. The usage column indicates in what way the data was used.

Source dataset	Ground truth	Format	Domain	Usage
RobutsMIS [22]	None	27h, 25 fps	30 in-vivo human	Self-supervision
Cholec80 [27]		51h, 25 fps	80 in-vivo human	
EndoVis2017 [3]	Stereo disparity	50m, 1 fps	10 in-vivo porcine	Supervision
EndoVis2018 [1]		95m, 1 fps	19 in-vivo porcine	
KidneyBoundary [9]		12m, 1 fps	15 in-vivo porcine	
StereoMIS [10]		45m, 1 fps	6 in-vivo mixed	
Hamlyn [32]	Stereo disparity	36m, 1 fps	7 in-vivo mixed	Validation
SERV-CT [8]	Registered CT	16 images	2 ex-vivo porcine	Testing
SCARED [2]	Structured light	45 images	9 ex-vivo porcine	
SCARED Clips [2]	None	23m, 25 fps	9 ex-vivo porcine	

$F \cdot I$ as the image I warped by the flow F. We then define a correspondence mask as

$$C_{a\to b} = \mathbf{1}(|F_{a\to b} + F_{a\to b} \cdot F_{b\to a}| < \varepsilon) \tag{1}$$

where $\mathbf{1}$ is the indicator function and ε is set to 2 pixels for our experiments. As the images are rectified, the horizontal component of $F_{a\to b}$ gives us the stereo disparity and the vertical component of the flow should be zero. We use a threshold of 2 pixel displacement to allow some leniency but mask out the deviations.

The data described so far is combined to form, `Meta-MED`, the meta monocular endoscopic depth dataset. Due to combining disparate data sources, and maintaining strict splits, we believe this to be much more appropriate for training generalisable depth models than what is currently available. However, the amount of training data is still relatively tiny, totalling just 14,310 images with pseudo ground truth depth, compared to 1.4 million used to train the MiDaS models. Self supervision allows us to leverage data that has no ground truth, thereby learning from a much broader domain during training. In addition to `Meta-MED`, we choose to combine RobutsMIS [22] and Cholec80 [27] giving us a total of 78 hours of 25 fps footage (117k images) from 110 in-vivo human surgeries. The videos are highly dynamic and feature extensive tissue-tool interaction. Most of the videos in these datasets feature prominent endoscopic content areas, which are detected and cropped out using the method in [5].

Fine-tuning Losses Here, we introduce the three losses we work with in our experiments: standard supervision ($\mathcal{L}_{sup}$), and our self-supervision losses, including augmentation consistency ($\mathcal{L}_{aug}$) and temporal consistency ($\mathcal{L}_{temp}$). During self-supervision, by comparing the output of the model with an output of the same model for a different input, it is possible for the model to *collapse* to a trivial solution that maximises the objective function, for example, outputting all zeros no matter the input. A common way to help prevent collapse is to have

the two branches use slightly different models.This can be achieved by having one branch use a *slow/teacher* model M_{slow}, which does not receive gradient updates and instead use weights calculated via an exponential moving average of the *fast/student* model M_{fast} which is used for the other branch. We adopt this approach for both our self-supervision losses.

Standard Supervision In order to construct loss functions and evaluation metrics for relative depth, we must be able to compare two depth maps in a way that is invariant to any scale and shift between the two. MiDaS and Depth Anything take the absolute difference between the two depth maps, each normalised by subtracting the median and dividing by the standard deviation, with some minor modifications to account for outliers. We found it produced better results to perform a linear least squares fitting of the predicted depth $\boldsymbol{d}$ to the ground truth depth $\boldsymbol{d^*}$, the latter being normalised as above to provide a consistent loss magnitude between different depth sources. We refer to this as Scale and Shift Invariant Mean Absolute Error (SSIMAE):

$$\text{SSIMAE}(\boldsymbol{d}, \boldsymbol{d^*}) = \frac{1}{N} \sum_{i=0}^{N} |(\alpha \boldsymbol{d} + \beta) - \hat{\boldsymbol{d}}^*|_i \tag{2}$$

where N is the number of unmasked pixels in $\boldsymbol{d^*}$; $\hat{\boldsymbol{d}}^*$ is the normalised version of $\boldsymbol{d^*}$; and α and β are found on a per-sample basis, by a linear least squares fit of $\boldsymbol{d}$ to $\hat{\boldsymbol{d}}^*$ ignoring any masked pixels. We then define our standard supervision loss $\mathcal{L}_{sup}$ as the SSIMAE between the depth map inferred by our fast model and the ground truth:

$$\mathcal{L}_{sup} = \text{SSIMAE}\left(M_{fast}(\boldsymbol{I}), \boldsymbol{d^*}\right) \tag{3}$$

Augmentation Consistency Self-supervision A common method of self supervision is to provide the model with a weakly and strongly augmented image. The output from the weakly augmented image may then be used as pseudo ground truth for the output from the strongly augmented image. This was originally proposed in FixMatch [25] for classification tasks and has been adapted to segmentation in endoscopy [29]. We use our slow model to infer a depth map for an image from our unlabeled datasets. We apply an augmentation transform which uses colour jittering A^c for the image and affine spatial transformations A^s applied jointly to the image and inferred depth map to keep them aligned. The fast model is then used to infer the depth map from the strongly augmented image. Our augmentation consistency loss $\mathcal{L}_{aug}$ is defined as

$$\mathcal{L}_{aug} = \text{SSIMAE}\left(M_{fast}\big(A^c(A^s(\boldsymbol{I}))\big), A^s\big(M_{slow}(\boldsymbol{I})\big)\right) \tag{4}$$

Temporal Consistency Self-supervision To address the temporal inconsistency we observed with pre-trained models, we construct a loss which encourages

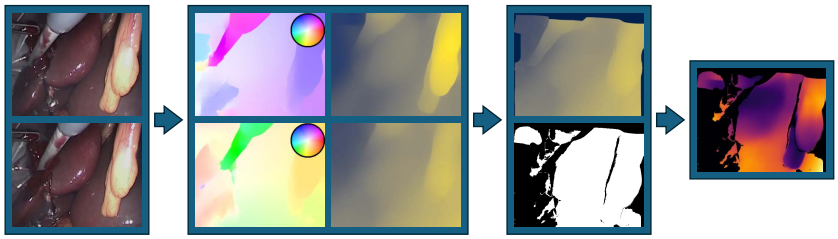

Fig. 2. Schematic of our temporal consistency loss. Starting from two temporally close input images at time t and $t + \delta t$ (panel one), the optical flows and depth-maps are inferred (panel two). The flow is then used to calculate a correspondence mask, and to warp the inferred depth map at time $t + \delta t$ to align it with the image at time t (panel three). The pixel-wise error between the warped and original depth map is masked (panel four) and averaged over to provide a final loss.

the model to provide consistent results for temporally close frames. We sample a pair of frames, I_a and I_b, randomly from a clip ensuring the time between the frames does not exceed one tenth of a second. The fast model and slow model are used to infer depth maps from I_a and I_b respectively. We then compute bidirectional optical flow giving $F_{a\to b}$ and $F_{b\to a}$. The depth map from I_b is warped to align it with I_a and masked with the correspondence mask calculated as in (1) to provide a ground truth. This process is illustrated in Fig. 2. The final temporal consistency loss may then be written as

$$\mathcal{L}_{temp} = \text{SSIMAE}\left(M_{fast}(\boldsymbol{I_a}), F_{a\to b} \cdot M_{slow}(\boldsymbol{I_b})\right) \tag{5}$$

We note that this construction ignores the effects of camera or subject motion along the optical axis. We hypothesis that enough positive supervision signal can be extracted during training before these errors become significant.

Evaluation We evaluate performance using two criteria. We first re-purpose our SSIMAE metric to assess accuracy of the depth estimation on individual images. The advantage of this metric is that it allows for direct comparison between datasets without worrying about changes in the shift or scale of the depth maps. Then, focusing on the errors which inspired our approach, we consider the temporal smoothness throughout clips of endoscopic footage.

Using our correspondence masking, we identify sections of the SCARED Clip videos for which the majority of pixels in the start frame can be tracked to every frame. We then use stereo disparity, and the models to be evaluated, to infer depth maps for all frames. Depth maps from the monocular models are fitted to the start frame, as in (2), to account for shift and scale. This allows us to build a "depth trajectory" of each tracked pixel throughout the clip. We then subtract the stereo disparity trajectory from the inferred trajectories and calculate the standard deviation of the result. This standard deviation is averaged over tracked pixels from all identified clips to provide a final metric.

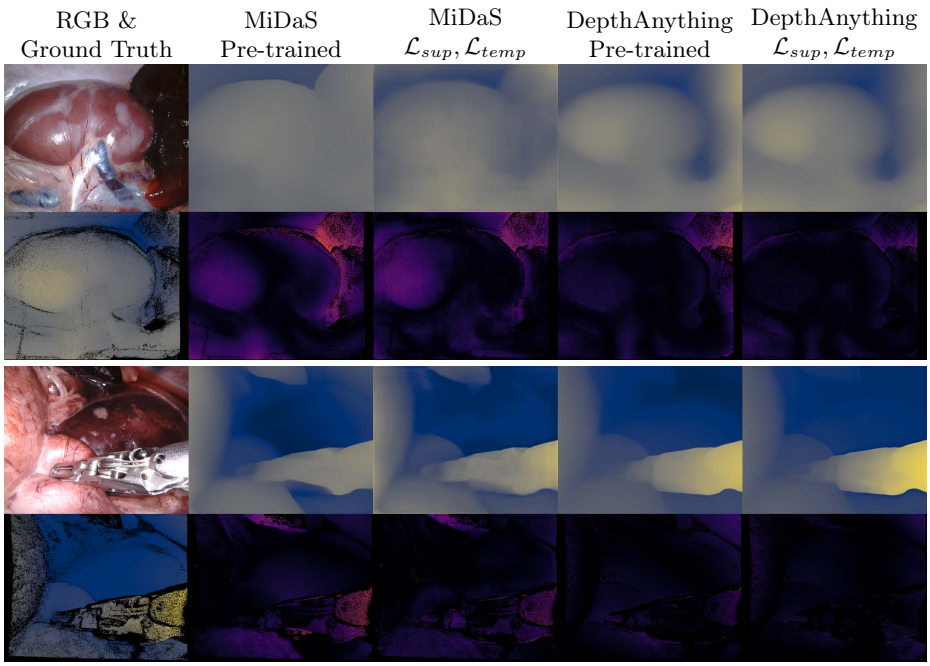

Fig. 3. Examples from our testing dataset. The left most column features the RGB (top) and the sparse ground truth depth (bottom), with the subsequent columns containing the predicted depth (top), which has been fitted to the ground truth as in (2), and the error (bottom) for a selection of models on a unified colour scale.

Table 2. Our scale and shift invariant MAE, and temporal inconsistency metrics for all our inference methods. The values shown for our trained models represent the mean and standard deviations, calculated across three training attempts. RAFT displays n/a for temporal inconsistency, as it is used in the construction of the metric.

Method		Average SSIMAE			Temporal Inconsistency
		SCARED	SERV-CT	Both	SCARED Clips
AF-SFMLearner [24]		0.319*	0.590	0.390*	0.823*
MiDaS		0.370	0.418	0.382	1.266
MiDaS	$\mathcal{L}_{sup}$	0.355±.001	**0.407±.001**	0.369±.001	1.215±.003
	$\mathcal{L}_{sup}, \mathcal{L}_{temp}$	**0.316±.001**	0.417±.001	**0.343±.001**	**1.094±.003**
	$\mathcal{L}_{sup}, \mathcal{L}_{aug}$	0.317±.001	0.436±.002	0.349±.001	1.104±.002
	$\mathcal{L}_{sup}, \mathcal{L}_{temp}, \mathcal{L}_{aug}$	0.333±.004	0.436±.009	0.360±.001	1.169±.012
DepthAnything		0.309	0.342	0.318	1.102
Depth Anything	$\mathcal{L}_{sup}$	0.280±.001	0.332±.002	0.294±.001	0.966±.005
	$\mathcal{L}_{sup}, \mathcal{L}_{temp}$	**0.265±.001**	0.282±.004	**0.269±.002**	**0.890±.014**
	$\mathcal{L}_{sup}, \mathcal{L}_{aug}$	0.281±.004	**0.280±.001**	0.280±.003	0.924±.014
	$\mathcal{L}_{sup}, \mathcal{L}_{temp}, \mathcal{L}_{aug}$	0.268±.003	0.299±.015	0.276±.002	0.892±.009
RAFT Stereo Disparity		0.080	0.165	0.102	n/a

*Probable overfitting: Evaluation data observed during training.

3 Experiments

For our fine-tuning experiments, we choose the best performing models from MiDaS and Depth Anything to act as our base models, specifically the MiDaS v3.1 dpt_large_512 and the Depth-Anything-Large models. These transformer based models have very similar architectures and parameter counts (345 and 335 million respectively), the main difference being the inclusion of more data, augmentation consistency self-supervision, and jointly learning a segmentation task in the training of the Depth Anything model. We fine-tune these base models using standard supervision, as well as semi-supervised learning using temporal consistency, augmentation consistency, and a combination of the two. Images were prepared to match the preparations used in the original training of the base models, color jittered, and loaded in batches of 15. Models were optimised using SGD with a learning rate of 1×10^{-6}, and gradient norm clipping of 10 to prevent collapse from unlucky batches during self-supervision. Multiple objective functions are optimised for by using interleaved gradient updates for each loss [16]. We take 100 batches to be an epoch and use our validation dataset for early stopping with a patience of 50 epochs.

We present our error and smoothness metrics for all models in Table 2. Example outputs for selected models are shown in Fig. 3 and videos are available in the supplementary material. The results show significant improvement over the base models from all fine-tuning methods. Adding self-supervision offers a significant benefit to the training, more than doubling the reduction in error of standard supervision alone in almost all cases. Temporal consistency proves to be the best self-supervision method for both base models, both in terms of SSIMAE and temporal consistency. Finally we show clear superiority over AF-SFMLearner, the domain-specific baseline. Many of our models outperform this domain-specific baseline on SCARED despite the dataset being used to train the baseline. When comparing performance on SERV-CT, data which is unseen for both methods, our best models show a reduction in error of over 50%.

4 Discussion and Conclusion

In this work, we have demonstrated the ability of state-of-the-art transformer based relative monocular depth models, trained on huge natural image meta-datasets, to generalise reasonably to the surgical domain, and shown that careful fine-tuning can significantly improve the performance. In line with the concurrent work of Depth Anything, we show that self-supervision can aid monocular depth models significantly. Following the method used for Depth Anything, we also experimented with jointly learning binary tool segmentation. However, we found this to dramatically reduce performance, so full experiments were not run. Furthermore, we show that our method of temporal consistency self-supervision significantly surpasses the predominantly used self-supervision method, augmentation consistency, when transferring to the surgical domain. We have also demonstrate that our transferred models drastically outperform the state-of-the-art method developed specifically for endoscopy, demonstrating the potential of

transferring natural image models to endoscopy. Finally, to aid in further research on this topic, we release our code, models, and dataset, `Meta-MED`, the meta monocular endoscopic depth dataset.

Acknowledgments. This study/project is funded by the NIHR [NIHR202114]. This work was supported by core funding from the Wellcome/EPSRC [WT203148/Z/16/Z; NS/A000049/1]. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101016985 (FAROS project). For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Disclosure of Interests. TV is a co-founder and shareholder of Hypervision Surgical.

References

1. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., *et al.*: 2018 robotic scene segmentation challenge, (2020). arXiv: 2001.11190 [cs.CV].
2. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., *et al.*: Stereo correspondence and reconstruction of endoscopic data challenge, (2021). arXiv: 2101.01133 [cs.CV].
3. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.-H., *et al.*: 2017 robotic instrument segmentation challenge, (2019). arXiv: 1902.06426 [cs.CV].
4. Birkl, R., Wofk, D., Müller, M.: MiDaS v3.1 – a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023)
5. Budd, C., Garcia-Peraza Herrera, L.C., Huber, M., Ourselin, S., Vercauteren, T.: Rapid and robust endoscopic content area estimation: a lean GPU-based pipeline and curated benchmark dataset. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**(4), 1215–1224 (2023)
6. Budd, C., Qiu, J., MacCormac, O., Huber, M., Mower, C., Janatka, M., *et al.*: "Deep reinforcement learning based system for intraoperative hyperspectral video autofocusing". In: Medical image computing and computer assisted intervention – miccai 2023. Springer Nature Switzerland, 2023, pp. 658–667. ISBN: 9783031439964.
7. Cui, B., Islam, M., Bai, L., Ren, H.: Surgical-DINO: adapter learning of foundation models for depth estimation in endoscopic surgery, (2024). arXiv: 2401.06013 [cs.CV].
8. Edwards, P.E., Psychogyios, D., Speidel, S., Maier-Hein, L., Stoyanov, D.: SERV-CT: a disparity dataset from cone-beam CT for validation of endoscopic 3D reconstruction. Medical Image Analysis **76**, 102302 (2022)
9. Hattab, G., Arnold, M., Strenger, L., Allan, M., Arsentjeva, D., Gold, O., *et al.*: Kidney edge detection in laparoscopic image data for computer-assisted surgery: kidney edge detection. Int. J. Comput. Assist. Radiol. Surg. **15**(3), 379–387 (2020)
10. Hayoz, M., Hahne, C., Gallardo, M., Candinas, D., Kurmann, T., Allan, M., *et al.*: Learning how to robustly estimate camera pose in endoscopic videos. International journal of computer assisted radiology and surgery **18**, 1185–1192 (2023)

11. Kiyasseh, D., Ma, R., Haque, T.F., Miles, B.J., Wagner, C., Donoho, D.A., *et al.*: A vision transformer for decoding surgeon activity from surgical videos. Nat. Biomed. Eng. **7**(6), 780–796 (2023)
12. Li, L., Li, X., Yang, S., Ding, S., Jolfaei, A., Zheng, X.: Unsupervised-learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery. IEEE Transactions on Industrial Informatics **17**(6), 3920–3928 (2021)
13. Liu, X., Sinha, A., Ishii, M., Hager, G.D., Reiter, A., Taylor, R.H., *et al.*: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE Transactions on Medical Imaging **39**(5), 1438–1447 (2020)
14. Lou, A., Noble, J.: WS-SfMLearner: self-supervised monocular depth and ego-motion estimation on surgical videos with unknown camera parameters, (2024). arXiv: 2308.11776 [cs.CV].
15. Luo, X., Huang, J.-B., Szeliski, R., Matzen, K., Kopf, J.: Consistent video depth estimation. ACM Transactions on Graphics (ToG) **39**(4), 71–1 (2020)
16. Mayo, D., Scott, T.R., Ren, M., Elsayed, G., Hermann, K., Jones, M., *et al.*: Multitask learning via interleaving: a neural network investigation. In: Proceedings of the Annual Meeting of the Cognitive Science Society (2023)
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., *et al.*: DINOv2: learning robust visual features without supervision, (2024). arXiv: 2304.07193 [cs.CV].
18. Ramalhinho, J., Yoo, S., Dowrick, T., Koo, B., Somasundaram, M., Gurusamy, K., *et al.*: The value of augmented reality in surgery — a usability study on laparoscopic liver surgery. Medical Image Analysis **90**, 102943 (2023)
19. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ICCV (2021)
20. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3) (2022)
21. Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J.M.M., Civera, J.: Endo-Depth-and-Motion: reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. IEEE Robotics and Automation Letters **6**(4), 7225–7232 (2021)
22. Ross, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., *et al.*: Robust medical instrument segmentation challenge 2019, (2020). arXiv: 2003.10299 [cs.CV].
23. Sestini, L., Rosa, B., De Momi, E., Ferrigno, G., Padoy, N.: FUN-SIS: a fully unsupervised approach for surgical instrument segmentation. Medical Image Analysis **85**, 102751 (2023)
24. Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., *et al.*: Self-supervised monocular depth and ego-motion estimation in endoscopy: appearance flow to the rescue. Medical image analysis **77**, 102338 (2022)
25. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., *et al.*: FixMatch: simplifying semi-supervised learning with consistency and confidence. In: Advances in Neural Information Processing Systems, pp. 596–608 (2020)
26. Teed, Z., Deng, J.: RAFT: recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pp. 402–419 (2020)

27. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
28. Wang, Y., Shi, M., Li, J., Huang, Z., Cao, Z., Zhang, J., *et al.*: Neural video depth stabilizer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9466–9476 (2023)
29. Wei, M., Budd, C., Garcia-Peraza-Herrera, L.C., Dorent, R., Shi, M., Vercauteren, T.: SegMatch: a semi-supervised learning method for surgical instrument segmentation. arXiv preprint arXiv:2308.05232 (2023)
30. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything: unleashing the power of large-scale unlabeled data, (2024). arXiv: 2401.10891 [cs.CV].
31. Yang, Y., Shao, S., Yang, T., Wang, P., Yang, Z., Wu, C., *et al.*: A geometry-aware deep network for depth estimation in monocular endoscopy. Engineering Applications of Artificial Intelligence **122**, 105989 (2023)
32. Ye, M., Johns, E., Handa, A., Zhang, L., Pratt, P., Yang, G.-Z.: Self-supervised Siamese learning on stereo image pairs for depth estimation in robotic surgery, (2017). arXiv: 1705.08260 [cs.CV].
33. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)