Semantic Segmentation of Surgical Hyperspectral Images Under Geometric Domain Shifts

Jan Sellner[1,2,3,4(✉)], Silvia Seidlitz[1,2,3,4], Alexander Studier-Fischer[4,5,6], Alessandro Motta[1], Berkin Özdemir[5,6], Beat Peter Müller-Stich[5,6], Felix Nickel[2,5,6], and Lena Maier-Hein[1,2,3,4,6]

[1] Division of Intelligent Medical Systems (IMSY), German Cancer Research Center (DKFZ), Heidelberg, Germany
j.sellner@dkfz-heidelberg.de
[2] Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[3] Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany
[4] National Center for Tumor Diseases (NCT), NCT Heidelberg, a Partnership Between DKFZ and University Medical Center Heidelberg, Heidelberg, Germany
[5] Department of General, Visceral, and Transplantation Surgery, Heidelberg University Hospital, Heidelberg, Germany
[6] Medical Faculty, Heidelberg University, Heidelberg, Germany

Abstract. Robust semantic segmentation of intraoperative image data could pave the way for automatic surgical scene understanding and autonomous robotic surgery. Geometric domain shifts, however – although common in real-world open surgeries due to variations in surgical procedures or situs occlusions – remain a topic largely unaddressed in the field. To address this gap in the literature, we (1) present the first analysis of state-of-the-art (SOA) semantic segmentation networks in the presence of geometric out-of-distribution (OOD) data, and (2) address generalizability with a dedicated augmentation technique termed 'Organ Transplantation' that we adapted from the general computer vision community. According to a comprehensive validation on six different OOD data sets comprising 600 RGB and yperspectral imaging (HSI) cubes from 33 pigs semantically annotated with 19 classes, we demonstrate a large performance drop of SOA organ segmentation networks applied to geometric OOD data. Surprisingly, this holds true not only for conventional RGB data (drop of Dice similarity coefficient (DSC) by 46 %) but also for HSI data (drop by 45 %), despite the latter's rich information content per pixel. Using our augmentation scheme improves on the SOA DSC by up to 67% (RGB) and 90% (HSI)) and renders performance

J. Sellner and S. Seidlitz—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_59.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 618–627, 2023.
https://doi.org/10.1007/978-3-031-43996-4_59

on par with in-distribution performance on real OOD test data. The simplicity and effectiveness of our augmentation scheme makes it a valuable network-independent tool for addressing geometric domain shifts in semantic scene segmentation of intraoperative data. Our code and pretrained models are available at https://github.com/IMSY-DKFZ/htc.

Keywords: deep learning · domain generalization · geometrical domain shifts · semantic organ segmentation · hyperspectral imaging · surgical data science

1 Introduction

Automated surgical scene segmentation is an important prerequisite for context-aware assistance and autonomous robotic surgery. Recent work showed that deep learning-based surgical scene segmentation can be achieved with high accuracy [7,14] and even reach human performance levels if using hyperspectral imaging (HSI) instead of RGB data, with the additional benefit of providing functional tissue information [15]. However, to our knowledge, the important topic of geometric domain shifts commonly present in real-world surgical scenes (e.g., situs occlusions, cf. Fig. 1) so far remains unaddressed in literature. It is questionable whether the state-of-the-art (SOA) image-based segmentation networks in [15] are able to generalize towards an out-of-distribution (OOD) context. The only related work by Kitaguchi et al. [10] showed that surgical instrument segmentation algorithms fail to generalize towards unseen surgery types that involve known instruments in an unknown context. We are not aware of any investigation or methodological contribution on geometric domain shifts in the context of surgical scene segmentation.

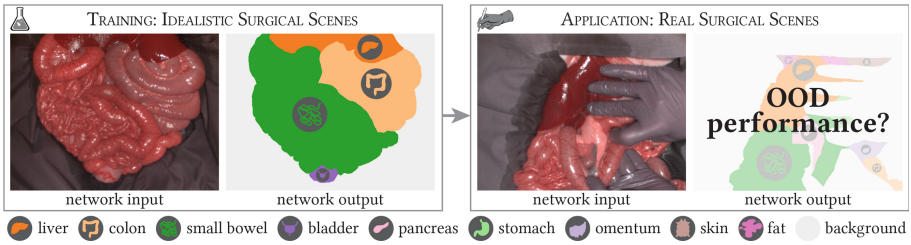

Fig. 1. State-of-the-art (SOA) surgical scene segmentation networks show promising results on idealistic datasets. However, in real-world surgeries, geometric domain shifts such as occlusions of the situs by operating staff are common. The generalizability of SOA algorithms towards geometric out-of-distribution (OOD) has not yet been addressed.

Generalizability in the presence of domain shifts is being intensively studied by the general machine learning community. Here, data augmentation evolved as

a simple, yet powerful technique [1,16]. In deep learning-based semantic image segmentation, geometric transformations are most common [8]. This holds particularly true for surgical applications. Our analysis of the SOA (35 publications on tissue or instrument segmentation) exclusively found geometric (e.g., rotating), photometric (e.g., color jittering) and kernel (e.g., Gaussian blur) transformations and only in a single case elastic transformations and Random Erasing (within an image, a rectangular area is blacked out) [22] being applied. Similarly, augmentations in HSI-based tissue classification are so far limited to geometric transformations. To our knowledge, the potential benefit of complementary transformations proposed for image classification and object detection, such as Hide-and-Seek (an image is divided into a grid of patches that are randomly blacked out) [17], Jigsaw (images are divided into a grid of patches and patches are randomly exchanged between images) [2], CutMix (a rectangular area is copied from one image onto another image) [21] and CutPas (an object is placed onto a random background scene) [4] (cf. Fig. 2), remains unexplored.

Given these gaps in the literature, the contribution of this paper is twofold:

1. We show that geometric domain shifts have disastrous effects on SOA surgical scene segmentation networks for both conventional RGB and HSI data.
2. We demonstrate that topology-altering augmentation techniques adapted from the general computer vision community are capable of addressing these domain shifts.

2 Materials and Methods

The following sections describe the network architecture, training setup and augmentation methods (Sect. 2.1), and our experimental design, including an overview of our acquired datasets and validation pipeline (Sect. 2.2).

2.1 Deep Learning-Based Surgical Scene Segmentation

Our contribution is based on the assumption that application-specific data augmentation can potentially address geometric domain shifts. Rather than changing the network architecture of previously successful segmentation methods, we adapt the data augmentation.

Surgery-Inspired Augmentation: Our Organ Transplantation augmentation illustrated in Fig. 2 has been inspired by the image-mixing augmentation CutPas that was originally proposed for object detection [4] and recently adapted for instance segmentation [5] and low-cost dataset generation via image synthesis from few real-world images in surgical instrument segmentation [19]. It is based on placing an organ into an unusual context while keeping shape and texture consistent. This is achieved by transplanting all pixels belonging to one object class (e.g., an organ class or background) into a different surgical scene. Our selection of further computer vision augmentation methods that could potentially improve

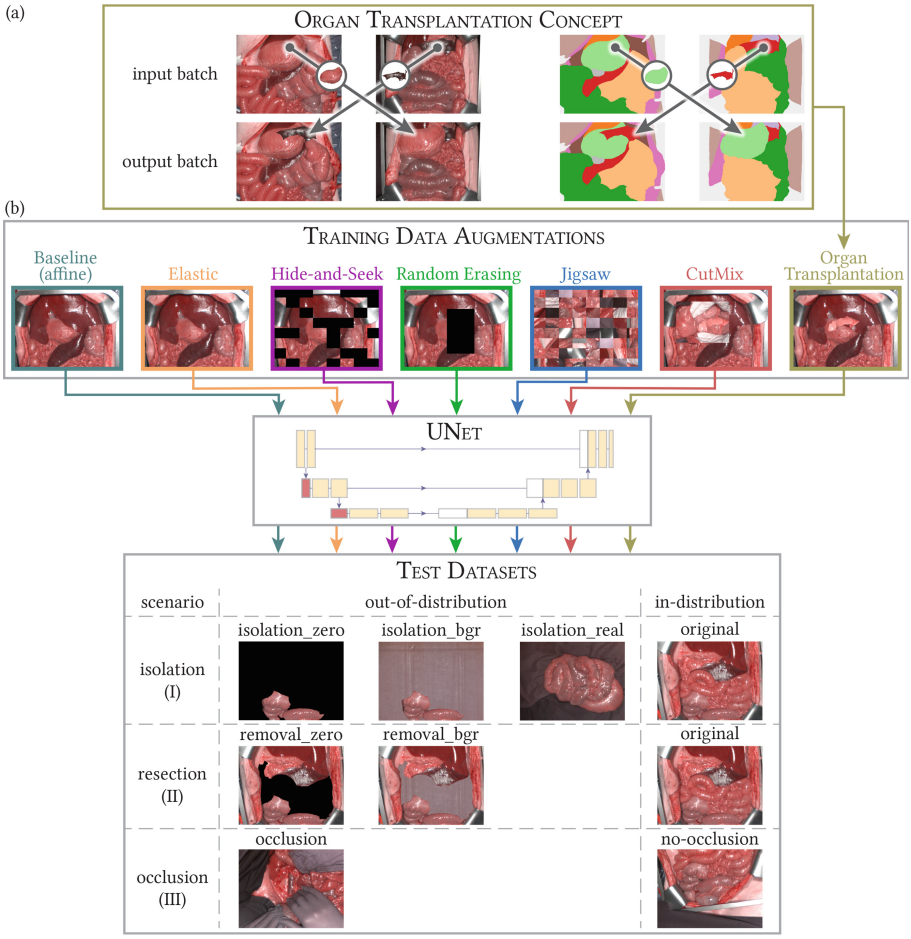

Fig. 2. (a) *Organ Transplantation* augmentation concept inspired from [4]. Image features and corresponding segmentations of randomly selected organs are transferred between images in one batch (in the example, the stomach is transferred from the left to the right and the spleen from the right to the left image). (b) Illustration of our validation experiments. We assess the generalizability under geometric domain shifts of seven different data augmentation techniques in deep learning-based organ segmentation. We validate the model performance on a range of out-of-distribution (OOD) scenarios, namely (1) organs in isolation (*isolation_zero*, *isolation_bgr* and *isolation_real*), (2) organ resections (*removal_zero* and *removal_bgr*), and (3) situs occlusions (*occlusion*), in addition to in-distribution data (*original* and *no-occlusion* (subset of *original* without occlusions)).

geometric OOD performance (cf. Fig. 2) was motivated by the specific conditions encountered in surgical procedures (cf. Sect. 2.2 for an overview). The noise augmentations Hide-and-Seek and Random Erasing black out all pixels inside

rectangular regions within an image, thereby generating artificial situs occlusions. Instead of blacking out, the image-mixing techniques Jigsaw and CutMix copy all pixels inside rectangular regions within an image into a different surgical scene. We adapted the image-mixing augmentations to our segmentation task by also copying and pasting the corresponding segmentations. Hence, apart from occluding the underlying situs, image parts/organs occur in an unusual neighborhood.

Network Architecture and Training: We used a U-Net architecture [13] with an efficientnet-b5 encoder [18] pre-trained on ImageNet data and using stochastic weight averaging [6] for both RGB and HSI data as it achieved human performance level in recent work [15]. As a pre-processing step, the HSI data was calibrated with white and dark reference images and ℓ^1-normalized to remove the influence of multiplicative illumination changes. Dice and cross-entropy loss were equally weighted to compute the loss function. The Adam optimization algorithm [9] was used with an exponential learning rate scheduler. Training was performed for 100 epochs with a batch size of five images.

2.2 Experiments

To study the performance of SOA surgical scene segmentation networks under geometric domain shifts and investigate the generalizability improvements offered by augmentation techniques, we covered the following OOD scenarios:

(I) *Organs in isolation*: Abdominal linens are commonly used to protect soft tissue and organs, counteract excessive bleeding, and absorb blood and secretion. Some surgeries (e.g., enteroenterostomy), even require covering all but a single organ. In such cases, an organ needs to be robustly identified without any information on neighboring organs.
(II) *Organ resections*: In resection procedures, parts or even the entirety of an organ are removed and surrounding organs thus need to be identified despite the absence of a common neighbor.
(III) *Occlusions*: Large parts of the situs can be occluded by the surgical procedure itself, introducing OOD neighbors (e.g., gloved hands). The non-occluded parts of the situs need to be correctly identified.

Real-World Datasets: In total, we acquired 600 intraoperative HSI cubes from 33 pigs using the HSI system Tivita® Tissue (Diaspective Vision GmbH, Am Salzhaff, Germany). These were semantically annotated with background and 18 tissue classes, namely heart, lung, stomach, small intestine, colon, liver, gallbladder, pancreas, kidney with and without Gerota's fascia, spleen, bladder, subcutaneous fat, skin, muscle, omentum, peritoneum, and major veins. Each HSI cube captures 100 spectral channels in the range between 500nm and 1000nm at an image resolution of 640×480 pixels. RGB images were reconstructed by aggregating spectral channels in the blue, green, and red ranges. To study organs in isolation, we acquired 94 images from 25 pigs in which all but a specific organ

were covered by abdominal linen for all 18 different organ classes (dataset *isolation_real*). To study the effect of occlusions, we acquired 142 images of 20 pigs with real-world situs occlusions (dataset *occlusion*), and 364 occlusion-free images (dataset *no-occlusion*). Example images are shown in Fig. 2.

Manipulated Data: We complemented our real-world datasets with four manipulated datasets. To simulate organs in isolation, we replaced every pixel in an image I that does not belong to the target label l either with zeros or spectra copied from a background image. We applied this transformation to all images in the dataset *original* and all target labels l, yielding the datasets *isolation_zero* and *isolation_bgr*. Similarly, we simulated organ resections by replacing all pixels belonging to the target label l either with zeros or background spectra, yielding the datasets *removal_zero* and *removal_bgr*. Example images are shown in Fig. 2.

Train-Test Split and Hyperparameter Tuning: The SOA surgical scene segmentation algorithms are based on a union of the datasets *occlusion* and *no-occlusion*, termed dataset *original*, which was split into a hold-out test set (166 images from 5 pigs) and a training set (340 images from 15 pigs). To enable a fair comparison, the same train-test split on pig level was used across all networks and scenarios. This also holds for the occlusion scenario, in which the dataset *no-occlusion* was used instead of *original* for training. All networks used the geometric transformations shift, scale, rotate, and flip from the SOA prior to applying the augmentation under examination. All hyperparameters were set according to the SOA. Only hyperparameters related to the augmentation under examination, namely the probability p of applying the augmentation, were optimized through a grid search with $p \in \{0.2, 0.4, 0.6, 0.8, 1\}$. We used five-fold-cross-validation on the datasets *original*, *isolation_zero*, and *isolation_bgr* to tune p such that good segmentation performance was achieved on both in-distribution and OOD data.

Validation Strategy: Following the recommendations of the Metrics Reloaded framework [11], we combined the Dice similarity coefficient (DSC) [3] as an overlap-based metric with the boundary-based metric ormalized surface distance (NSD) [12] for validation for each class l. To respect the hierarchical test set structure, metric aggregation was performed by first macro-averaging the class-level metric value M_l ($M \in \{\text{DSC}, \text{NSD}\}$) across all images of one pig and subsequently across pigs. The organ removal experiment required special attention in this context, as multiple M_l values per image could be generated corresponding to all the possible neighbour organs that could be removed. In this case, we selected for each l the minimum of all M_l values, which corresponds to the segmentation performance obtained after removing the most important neighbour of l. The same class-specific NSD thresholds as in the SOA were used.

3 Results

Effects of Geometric Domain Shifts: When applying a SOA segmentation network to geometric OOD data, the performance drops radically (cf. Fig. 3). Starting from a high DSC for in-distribution data (RBG: 0.83 (standard deviation

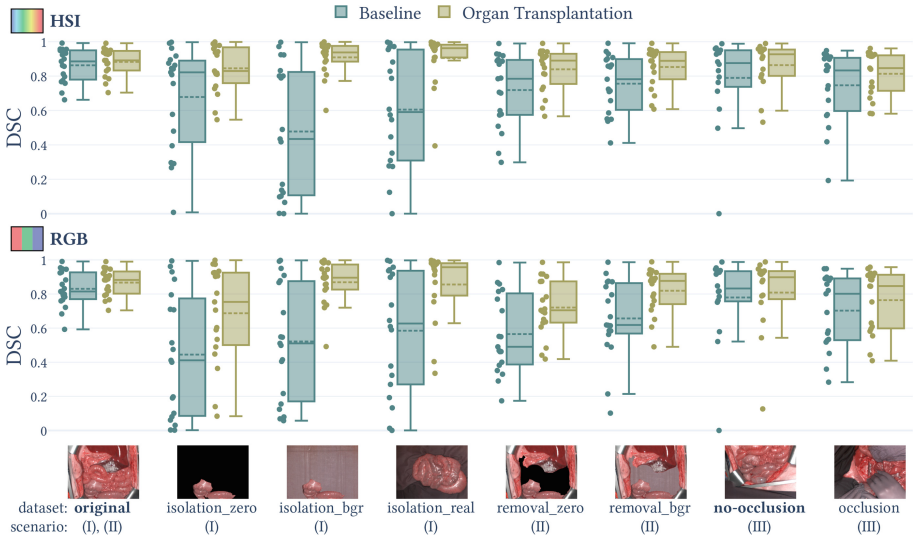

Fig. 3. Segmentation performance of the hyperspectral imaging (HSI) and RGB modality for all eight test datasets (six out-of-distribution (OOD) and two in-distribution datasets (bold)) comparing the baseline network with the Organ Transplantation network. Each point denotes one out of 19 class-level Dice similarity coefficient (DSC) values after hierarchical aggregation across images and subjects. The boxplots show the quartiles of the class-level DSC. The whiskers extend up to 1.5 times the interquartile range and the median and mean are represented as a solid and dashed line, respectively.

(SD) 0.10); HSI: 0.86 (SD 0.10)), the performance drops by 10%-46% for RGB and by 5 %-45% for HSI, depending on the experiment. In the organ resection scenario, the largest drop in performance of 63% occurs for the gallbladder upon liver removal (cf. Suppl. Fig. 1). Similar trends can be observed for the boundary-based metric NSD, as shown in Suppl. Fig. 2.

Performance of Our Method: Figure 3 and Suppl. Fig. 2 show that the Organ Transplantation augmentation (gold) can address geometric domain shifts for both the RGB and HSI modality. The latter yields consistently better results, indicating that the spectral information is crucial in situations with limited context. The performance improvement compared to the baseline ranges from 9 %-67% (DSC) and 15 %-79% (NSD) for RGB, and from 9%-90% (DSC) and 16 %-96% (NSD) for HSI, with the benefit on OOD data being largest for organs in isolation and smallest for situs occlusions. The Organ Transplantation augmentation even slightly improves performance on in-distribution data (*original* and *no-occlusion*). Upon encountering situs occlusions, the largest DSC improvement is obtained for the organ classes pancreas (283 %) and stomach (69 %). For organs in isolation, the performance improvement on manipulated data (DSC increased by 57% (HSI) and 61% (RGB) on average) is comparable to that on real data (DSC increased by 50% (HSI) and 46% (RGB)).

Comparison to SOA Augmentations: There is no consistent ranking across all six OOD datasets except for Organ Transplantation always ranking first and baseline usually ranking last (cf. Fig. 4 for DSC- and Suppl. Fig. 3 for NSD-based ranking). Overall, image-mixing augmentations outperform noise augmentations. Augmentations that randomly sample rectangles usually rank better than comparable augmentations using a grid structure (e.g., CutMix vs. Jigsaw).

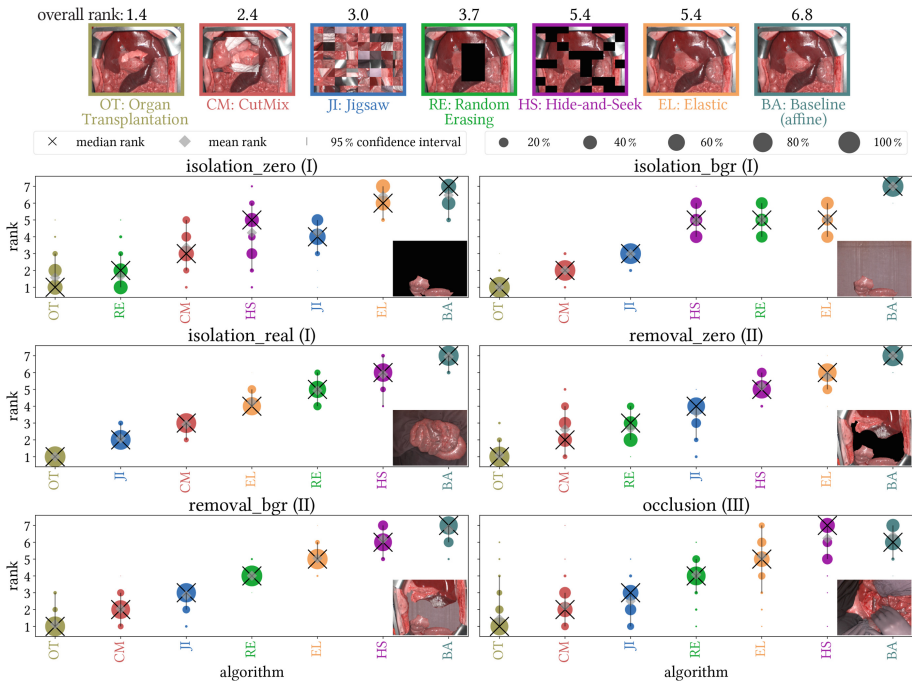

Fig. 4. Uncertainty-aware ranking of the seven augmentation methods for all six geometric out-of-distribution (OOD) test datasets. Organ Transplantation consistently ranks first and baseline last. The area of each blob for one rank and algorithm is proportional to the relative frequency of that algorithm achieving the respective rank across 1000 bootstrap samples consisting of 19 hierarchically aggregated class-level Dice similarity coefficient (DSC) values each (concept from [20]). The numbers above the example images denote the overall ranking across datasets (mean of all mean ranks).

4 Discussion

To our knowledge, we are the first to show that SOA surgical scene segmentation networks fail under geometric domain shifts. We were particularly surprised by the large performance drop for HSI data, rich in spectral information. Our results clearly indicate that SOA segmentation models rely on context information.

Aiming to address the lack of robustness to geometric variations, we adapted so far unexplored topology-altering data augmentation schemes to our target

application and analyzed their generalizability on a range of six geometric OOD datasets specifically designed for this study. The Organ Transplantation augmentation outperformed all other augmentations and resulted in similar performance to in-distribution performance on real OOD data. Besides its effectiveness and computational efficiency, we see a key advantage in its potential to reduce the amount of real OOD data required in network training. Our augmentation networks were optimized on simulated OOD data, indicating that image manipulations are a powerful tool for judging geometric OOD performance if real data is unavailable, such as in our resection scenario, which would have required an unfeasible number of animals. With laparoscopic HSI systems only recently becoming available, the investigation and compensation of geometric domain shifts in minimally-invasive surgery could become a key direction for future research. Our proposed augmentation is model-independent, computationally efficient and effective, and thus a valuable tool for addressing geometric domain shifts in semantic scene segmentation of intraoperative HSI and RGB data. Our implementation and models will be made publicly available.

Acknowledgements and Data Usage. This project was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (NEURAL SPICING, 101002198), the National Center for Tumor Diseases (NCT) Heidelberg's Surgical Oncology Program, the German Cancer Research Center (DKFZ), and the Helmholtz Association under the joint research school HIDSS4Health (Helmholtz Information and Data Science School for Health). The private HSI data was acquired at Heidelberg University Hospital after approval by the Committee on Animal Experimentation (G-161/18 and G-262/19).

References

1. Alomar, K., Aysel, H.I., Cai, X.: Data augmentation in classification and segmentation: a survey and new strategies. J. Imaging **9**(2), 46 (2023)
2. Chen, Z., Fu, Y., Chen, K., Jiang, Y.G.: Image block augmentation for one-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 3379–3386 (2019)
3. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
4. Dwibedi, D., Misra, I., Hebert, M.: Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection (2017)
5. Ghiasi, G., et al.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 2917–2927. IEEE (2021)
6. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. In: Proceedings of the International Conference on Uncertainty in Artificial Intelligence (2018)
7. Kadkhodamohammadi, A., Luengo, I., Barbarisi, S., Taleb, H., Flouty, E., Stoyanov, D.: Feature aggregation decoder for segmenting laparoscopic scenes. In: Zhou, L., et al. (eds.) OR 2.0/MLCN -2019. LNCS, vol. 11796, pp. 3–11. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32695-1_1

8. Kar, M.K., Nath, M.K., Neog, D.R.: A review on progress in semantic image segmentation and its application to medical images. SN Comput. Sci. **2**(5), 397 (2021)
9. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv:1412.6980 (2017)
10. Kitaguchi, D., Fujino, T., Takeshita, N., Hasegawa, H., Mori, K., Ito, M.: Limited generalizability of single deep neural network for surgical instrument segmentation in different surgical environments. Sci. Rep. **12**(1), 12575 (2022)
11. Maier-Hein, L., et al.: Metrics reloaded: pitfalls and recommendations for image analysis validation (2023)
12. Nikolov, S., et al.: Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. J. Med. Internet Res. **23**(7) (2021)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Scheikl, P., et al.: Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery. Curr. Dir. Biomed. Eng. **6**, 20200016 (2020)
15. Seidlitz, S., et al.: Robust deep learning-based semantic organ segmentation in hyperspectral images. Med. Image Anal. **80**, 102488 (2022)
16. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 60 (2019)
17. Singh, K.K., Lee, Y.J.: Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3544–3553 (2017)
18. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019)
19. Wang, A., Islam, M., Xu, M., Ren, H.: Rethinking surgical instrument segmentation: a background image can be all you need. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13437, pp. 355–364. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_34
20. Wiesenfarth, M., et al.: Methods and open-source toolkit for analyzing and visualizing challenge results. Sci. Rep. **11**(1), 2369 (2021)
21. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: CutMix: regularization strategy to train strong classifiers with localizable features. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 6022–6031. IEEE (2019)
22. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 13001–13008 (2020)