How precise are performance estimates for typical medical image segmentation tasks?

Rosana El Jurdi, Olivier Colliot

HOW PRECISE ARE PERFORMANCE ESTIMATES FOR TYPICAL MEDICAL IMAGE SEGMENTATION TASKS?

Rosana El Jurdi, Olivier Colliot

Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France

ABSTRACT

An important issue in medical image processing is to be able to estimate not only the performances of algorithms but also the precision of the estimation of these performances. Reporting precision typically amounts to reporting standard-error of the mean (SEM) or equivalently confidence intervals. However, this is rarely done in medical image segmentation studies. In this paper, we aim to estimate what is the typical confidence that can be expected in such studies. To that end, we first perform experiments for Dice metric estimation using a standard deep learning model (U-net) and a classical task from the Medical Segmentation Decathlon. We extensively study precision estimation using both Gaussian assumption and bootstrapping (which does not require any assumption on the distribution). We then perform simulations for other test set sizes and performance spreads. Overall, our work shows that small test sets lead to wide confidence intervals (e.g. $\sim$8 points of Dice for 20 samples with $\sigma \simeq 10$).

***Index Terms*—** Segmentation, Performance, Validation, Statistical analysis, Confidence interval, Standard error.

1. INTRODUCTION

In medical imaging, it is not uncommon that sample sizes are in the order of dozens of subjects, at best hundreds or thousands. In 3D medical image segmentation, the size of the set used to evaluate the performance may be even smaller than for other medical imaging tasks as obtaining the ground truth requires voxel-wise annotation by trained raters.

Intuitively, the precision of the estimation of the performance depends on two factors: the variability of the performance among the test set (the more variable, the less precise) and the size of the test set (smaller sets will lead to lower precision and therefore larger confidence intervals). However, papers usually report the average performance for different metrics (e.g. average Dice) but not the precision [1] with which this average performance is estimated. Such precision can be provided in the form of confidence intervals or equivalently standard error of the mean (SEM) which are not often reported. What is more often reported is the empirical standard deviation over different folds of a cross-validation. While this may qualitatively characterize the variability of the learning procedure when the training and testing set change, it should never be used to compute the SEM, since here n would be the number of folds or splits, which is arbitrary and can be made as large as one wants, thereby making the confidence interval arbitrarily narrow. It is not even an unbiased estimate of the standard deviation of the performance metric [1].

Quantifying the precision of the estimation of the performances thus requires an independent test set, on which confidence intervals or SEM are reported. Since this is not typically done in medical image segmentation papers, one may ask the following question. What precision can be expected for a typical sample size? How trustworthy are the average performance estimates (for instance Dice coefficients) reported in medical image segmentation papers?

Surprisingly, this question has been little studied in medical imaging. In the case of a different task, namely image classification, it is necessary to have large sample sizes for a precise estimation of the accuracy (typically 10,000 samples to achieve a 1%-wide confidence interval given an accuracy of about $90\% - 95\%$) [2, 3]. However, to the best of our knowledge, this is not widely known in the case of segmentation. We hypothesize that the test size needed to achieve a given precision is lower than for classification due to the continuous nature of performance measures [4].

Our objective is to study the precision that can be expected in 3D medical image segmentation for typical test set sizes. We first conduct experiments using a standard deep learning network applied to a classical segmentation task from the Segmentation Decathlon Challenge [5] in order to estimate confidence intervals which are obtained for variable test set sizes. We then perform simulations for other sizes and spreads. We insist that the aim of the present paper is not to propose a new segmentation methodology. Instead, the main aims are to provide information regarding the confidence intervals that can typically be expected in medical image segmentation and to raise awareness of the community on this important issue.

[1] Throughout the paper, precision means how precise are the estimates of the performance. It has nothing to do with the performance metric Precision also known as Positive Predictive Value.

2. OVERVIEW

Our aim is to provide confidence intervals (or standard-errors) for the mean of a given performance metric for different test set sizes and different spreads (standard deviation) of the performance. If the performance metric follows a Gaussian distribution, one can obtain those values as follows:

$$\begin{aligned} \text{SEM} &= \frac{\sigma}{\sqrt{n}} \\ \text{CI} &= [\mu - 1.96 \times \text{SEM}, \mu + 1.96 \times \text{SEM}] \end{aligned} \tag{1}$$

where μ denotes the mean, σ the standard-deviation, n the test set size, SEM the standard error of the mean and CI the 95% confidence interval. In the following, for conciseness, we will also denote the width of CI $= [a, b]$ as $w = b - a$.

One does not know a priori if using Equations 1 is valid for a given performance metric. On the other hand, in the absence of assumption on the distribution of the metric, one can use bootstrapping to estimate SEM and CI [6]. In Section 3, we will perform experiments in order: i) to verify the validity of using Equations 1 by comparing the estimates obtained using these equations to those using bootstrapping; ii) obtain empirical estimates of μ and σ; iii) study the effect of using subsamples of reduced size.

In Section 4, we will perform simulations using Equations 1 to study how the precision varies when varying the test set size and performance variability (as measured by σ).

3. EXPERIMENTS

3.1. Dataset and segmentation method

We used the Hippocampus dataset from the Medical Decathlon challenge [5], composed of 260 3D MR images. The task is to segment the anterior and posterior parts of the hippocampus. For evaluation, we merged the two regions (in prediction and ground-truth respectively) and considered the hippocampus as a whole. From the 260 samples, 100 patients were randomly selected for training, 50 for validation and the remaining 110 samples constituted the test set.

For the segmentation, we used a U-net type network [7]. Note that our aim is not to achieve the highest possible segmentation scores or to propose a novelty in the segmentation method but rather to obtain *typical* performances. We thus relied on a standard approach. We treat a 3D MRI as a sequence of 2D images and used a 2D architecture. At inference, we predict for each slice independently then stack the slices belonging to the same patient together to form a 3D-volume prediction. The architecture is a 3-stage structure composed of convolutional and de-convolutional blocks, bottleneck and skip connections. An ensemble of convolutional and batch normalization layers constitutes the encoder part. Each stage within the decoder path is composed of 2 consecutive convolutional blocks followed by an upsampling layer. The bottleneck is composed of 2 convolutional blocks separated by a residual block [8]. The architecture has been used in previous publications [9, 10]. The optimizer was Adam, the learning rate was 0.001 and the batch size was 8. The learning rate was halved if the validation performances did not improve over 20 epochs as proposed by [9]. Note that we used the standard generalized Dice loss [11]. The model was trained over 500 epochs. We performed a three-fold cross-validation with the 150 patients of the training/validation sets and selected the best model across these three folds. Note that the test set was left untouched and was never used at any stage for training, model selection or architecture/parameter optimization. To evaluate the performance, we computed the Dice coefficient in % [2] for the whole hippocampus (thus merging anterior and posterior parts before computing the metric).

The code used to generate the results is available online [3]

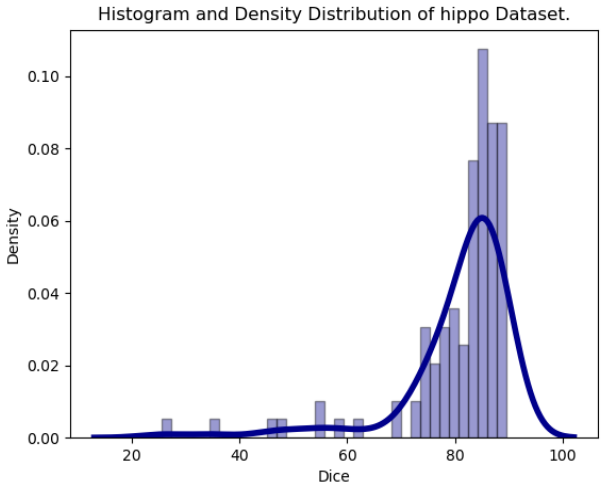

Fig. 1: Histogram of Dice accuracy over the entire test set. It is shown together with a kernel density estimation (KDE) which smoothes the observations with a Gaussian kernel.

3.2. Precision on the whole test set

We first studied the precision of performance estimates using the maximum test set size. The distribution of Dice values over the test set is shown in Figure 1. One can observe that the Gaussian assumption is not unreasonable despite underlying outliers and skewness. More importantly, we will now compare estimates based on this assumptions to corresponding non-parametric bootstrap estimates.

We first compute μ, σ, SEM and w using Equations 1. We then compute their bootstrap counterparts μ^*, SEM*, w^* as follows [4]. Given a test set of size n, $M = 15000$

[2] Performance metric was computed using this code: `https://github.com/deepmind/surface-distance`

[3] `https://github.com/rosanajurdi/SegVal/tree/ISBI2023`

[4] Throughout the paper, the bootstrap estimate of a given x is always denoted as x^*

n	μ	σ	SEM	w	μ^*	SEM^*	w^*
n = 110	80.70	10.75	1.02	4.02	80.70	1.02	3.99

Table 1: **Results on the full test set** ($n = 110$). μ and σ are the empirical mean and standard deviation of the Dice coefficient across all patients in the test set. SEM is the standard error of the mean and w is the width of the 95% confidence interval calculated via Equations 1. SEM^*, μ^* and w^* are the values obtained via Bootstrapping.

Subsample size k	μ_k	σ_k	SEM_k	w_k	μ_k^*	SEM_k^*	w_k^*
k = 10	81.01 ± 3.04	8.17 ± 4.75	2.58 ± 1.5	10.13 ± 5.88	81.01 ± 3.04	2.59 ± 1.51	9.86 ± 5.62
k= 20	80.61 ± 2.16	9.96 ± 3.74	2.23 ± 0.84	8.73 ± 3.28	80.61 ± 2.16	2.23 ± 0.84	8.61 ± 3.21
k= 30	80.63 ± 1.6	10.36 ± 2.94	1.89 ± 0.54	7.41 ± 2.1	80.64 ± 1.6	1.89 ± 0.54	7.34 ± 2.08
k= 50	80.95 ± 1.14	9.99 ± 2.1	1.41 ± 0.3	5.54 ± 1.16	80.95 ± 1.14	1.41 ± 0.3	5.51 ± 1.15
k= 100	80.64 ± 0.32	10.78 ± 0.53	1.08 ± 0.05	4.22 ± 0.21	80.64 ± 0.31	1.08 ± 0.05	4.21 ± 0.21
k= 110	80.70 ± 0.0	10.75 ± 0.0	1.02 ± 0.0	4.02 ± 0.0	80.71 ± 0.01	1.02 ± 0.01	3.99 ± 0.03

Table 2: **Results on subsamples of size $k \leq 110$.** Results are shown as $mean \pm std$ where $mean$ and std are the mean and standard-deviation over all the subsamples $S_{k,j}$ of a given size k (k is fixed and $j \in \{1, \dots, 100\}$).

bootstrap samples of size n are drawn with replacement. We denote a given bootstrap sample as S_m^* and its mean as μ_m^* where $m \in \{1, \dots, M\}$. The bootstrap mean μ^* is the mean of the bootstrap sample means μ_m^*. The standard error of the mean μ^* obtained via bootstrapping (SEM^*) is the standard deviation of the means of all bootstrap samples: $\text{SEM}^* = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\mu_m^* - \mu^*\right)^2}$. The 95% confidence interval $\text{CI}^* = [a^*, b^*]$ is the set of values between the 2.5% and 97.5% percentiles of the sorted bootstrap means $\{\mu_1^*, \mu_2^*, \dots, \mu_m^*, \dots \mu_M^*\}$. We finally define the width of the confidence interval as $w^* = b^* - a^*$. The estimates using the Gaussian assumption and the bootstrap are very close as can be seen on Table 1.

3.3. Precision on subsamples of size $k \leq n = 110$

We now study experimentally the relationship existing between the test set size and the precision of the estimation of the segmentation performance. To that end, we draw subsamples of variable size $k \in K = \{10, 20, 30, 50, 100, 110\}$ from the whole test set of size n. In order not to depend on a particular drawing (which may be lucky or unlucky), we repeat the procedure 100 times for each k. We denote the subsamples as $(S_{k,j})$ where k is the subsample size and $j \in \{1, \dots, 100\}$ is the index of a particular drawing.

We then proceed with the computations of the different estimates based either on the Gaussian assumption or on the bootstrap.

For the Gaussian assumption, we denote as $\mu_{k,j} = \frac{1}{k}\sum_{l=1}^{k} D_{j,k,l}$ and $\sigma_{k,j} = \sqrt{\frac{1}{k}\sum_{l=1}^{k}\left(D_{j,k,l} - \mu_{k,j}\right)^2}$ the empirical mean and standard deviation for the subsample $S_{k,j}$ where $D_{j,k,l}$ is the Dice coefficient of a given subject in the subsample $S_{k,j}$. Similarly, we use the notations $\text{SEM}_{k,j} = \frac{\sigma_{k,j}}{\sqrt{k}}$ and $w_{k,j} = 2 * 1.96 * \text{SEM}_{k,j}$. We can then study how these values vary across the 100 subsamples of a given size k. To that end, we compute the average and standard-deviation of $\mu_{k,j}$, $\sigma_{k,j}$, $\text{SEM}_{k,j}$ and $w_{k,j}$ across the different subsamples $S_{k,j}$ for k fixed and $j \in \{1, \dots, 100\}$. This provides the following estimates $\mu_k \pm \sigma_{\mu_k}$, $\sigma_k \pm \sigma_{\sigma_k}$, $\text{SEM}_k \pm \sigma_{\text{SEM}_k}$ and $w_k \pm \sigma_{w_k}$. The values are displayed in Table 2. One can gather that, as the sample size increases, the standard deviation and the standard error decrease.

Bootstrap estimations are performed as follows. For a given subsample $S_{k,j}$ of size k and index j, $M = 15000$ bootstrap samples of size k are drawn with replacement. We denote a given bootstrap sample as $S_{k,j,m}^*$ and its mean as $\mu_{k,j,m}^*$ where $m \in \{1, \dots, M\}$ is the index of the m^{th} bootstrap sample of subsample $S_{k,j}$. The bootstrap mean $\mu_{k,j}^*$ of $S_{k,j}$ is the mean of the bootstrap sample means $\mu_{k,j,m}^*$: $\mu_{k,j}^* = \frac{1}{M}\sum_{m=1}\mu_{k,j,m}^*$. The standard error of the mean $\mu_{k,j}^*$ (denoted as $\text{SEM}_{k,j}^*$) obtained via bootstrapping is the standard deviation of the means of all bootstrap samples of subsample $S_{k,j}$: $\text{SEM}_{k,j}^* = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\mu_{k,j,m}^* - \mu_{k,j}^*\right)^2}$. The 95% confidence interval of the sample $S_{k,j}$ is denoted as $[a_{k,j}^*, b_{k,j}^*]$ and is the set of values between the 2.5% and 97.5% percentile of the sorted bootstrap means of subsample $S_{k,j}$. We finally define the width of the confidence interval via bootstrapping as $w_{k,j}^* = b_{k,j}^* - a_{k,j}^*$. We study how these values vary across the 100 samples of a given size k. To that end, we compute the averages and the standard deviations of $\mu_{k,j}^*$, $\text{SEM}_{k,j}^*$ and $w_{k,j}^*$ across the different subsamples $S_{k,j}$ for k fixed and $j \in \{1, ..., 100\}$. This provides the following estimates $\mu_k^* \pm \sigma_{\mu_k^*}$, $\text{SEM}_k^* \pm \sigma_{\text{SEM}_k^*}$ and $w_k^* \pm \sigma_{w_k^*}$. Results are shown in Table 2.

As for the whole test set, estimates using Equation 1 and bootstrapping are very close across different subsample sizes. As expected, precision decreases with the sample size.

σ	2		5		8		10.75		12		15		18	
	SEM	w_k	SEM	w_k	SEM	w_k	SEM	w_k	SEM	w_k	SEM	w_k	SEM	w_k
$k = 10$	0.63	2.48	1.58	6.2	2.53	9.92	3.4	13.33	3.79	14.88	4.74	18.59	5.69	22.31
$k = 20$	0.45	1.75	1.12	4.38	1.79	7.01	2.4	9.43	2.68	10.52	3.35	13.15	4.02	15.78
$k = 30$	0.37	1.43	0.91	3.58	1.46	5.73	1.96	7.7	2.19	8.59	2.74	10.74	3.29	12.88
$k = 50$	0.28	1.11	0.71	2.77	1.13	4.43	1.52	5.96	1.7	6.65	2.12	8.32	2.55	9.98
$k = 100$	0.2	0.78	0.5	1.96	0.8	3.14	1.08	4.22	1.2	4.7	1.5	5.88	1.8	7.06
$k = 200$	0.14	0.55	0.35	1.39	0.57	2.22	0.76	2.98	0.85	3.33	1.06	4.16	1.27	4.99
$k = 300$	0.12	0.45	0.29	1.13	0.46	1.81	0.62	2.43	0.69	2.72	0.87	3.39	1.04	4.07
$k = 500$	0.09	0.35	0.22	0.88	0.36	1.4	0.48	1.89	0.54	2.1	0.67	2.63	0.8	3.16
$k = 1000$	0.06	0.25	0.16	0.62	0.25	0.99	0.34	1.33	0.38	1.49	0.47	1.86	0.57	2.23

Table 3: Simulation of SEM_k and the width w_k for different sizes k of the test set and for different values of σ. The gray column with $\sigma = 10.75$ corresponds to the standard deviation found in the experimental section.

4. SIMULATIONS WITH A GAUSSIAN DISTRIBUTION

In the previous section, we showed that estimates of SEM and confidence intervals computed using either Equations 1 or the bootstrap are in accordance. Given this, we now perform simulations using Equations 1 for more values of k (including in particular larger test sets) and for different values of σ. Note that this is independent of μ which, in itself, has no impact on the SEM nor on the width of the CI (even though it is usually observed that lower performing models, thus associated with a lower value of μ, also have a more variable performance and thus a larger value of σ). Results are displayed in Table 3.

Comparing the gray column in Table 3 to Table 1, one can observe that both the standard error and the confidence interval width are close to the previously obtained experimental values (for $k \leq 100$)). Nevertheless, the experimental values are slightly lower than those of the simulation. As the value of k increases, the gap between experimental and simulated values decreases.

5. DISCUSSION

In this paper, we have provided elements regarding the precision with which segmentation performance can be estimated in typical medical imaging studies. As hypothesized, the test set size needed to obtain a given confidence interval is smaller than in image classification. Typically, with 10,000 samples for classification (for a high performing model with over 90% accuracy), one obtains a 1.2%-wide CI. For our segmentation experiments, such width is obtain with only about 1000 samples. For a 4%-wide CI, one needs about 1000 samples for the aforementioned classification and about 100 samples for segmentation. Of course, one needs to keep in mind that this depends not only on the test sample size but also on σ of the performance. For example for σ = 5, the 1%-wide CI would then require between 300 and 500 samples.

Confidence intervals are rarely reported in medical image segmentation papers. To illustrate this, we have conducted a search for papers published in 2022 in *IEEE Transactions on Medical Imaging (IEEE TMI)* dealing with segmentation of 3D images. We found 51 papers containing 107 experiments (a given paper may include several experiments). Only 71 (66%) of these experiments (corresponding to 21 papers) were done on an independent test set, the rest reporting cross-validation results on training/validation sets. This is problematic for two reasons. First, the performance on cross-validation results can lead to optimistically biased performances [12]. Moreover, they cannot be used to estimate standard errors. For the experiments that had an independent test set, the median size of the test set was 21 (minimum: 4, mean: 61.5, maximum: 697). In light of our experiments, it is not unreasonable that the associated confidence intervals will often be wide (typically between 9 and 10 for a test set with 20 samples). Finally, of the 21 papers that adopted an independent test set, only 3 of them (amounting to 6% of the 51 surveyed papers) reported confidence intervals or standard-error [13, 14, 15]. Note that 11 other papers used statistical testing to compare different approaches even though they did not provide confidence intervals [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26].

Our study has the following limitations. First, we studied only one dataset. Second, we used only one segmentation model. Future work would need to assess other models and datasets, in particular to see which is the typical range of values that can be expected for σ. Second, it was restricted to the Dice performance metric and it would be interesting to study if the same observations hold for other metrics (e.g. Hausdorff distance, volume error...).

Overall, the experiments presented in our paper show the importance of reporting confidence intervals on independent test sets and that, in general, studies with small test sets cannot claim accurate estimation of the performances. We believe that it is an important issue on which the community should focus.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by the Medical Segmentation Decathlon challenge [5, 27]. Ethical approval was not required as confirmed by the license attached with the open access data.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6).

8. REFERENCES

[1] Yoshua Bengio and Yves Grandvalet, "No unbiased estimator of the variance of K-fold cross-validation," *Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, 2004.

[2] Gaël Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *NeuroImage*, vol. 180, pp. 68–77, 2018.

[3] Gaël Varoquaux and Olivier Colliot, "Evaluating machine learning models and their diagnostic value," *HAL preprint*, vol. hal-03682454, 2022.

[4] Douglas G Altman and Patrick Royston, "The cost of dichotomising continuous variables," *BMJ*, vol. 332, no. 7549, pp. 1080, May 2006.

[5] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, et al., "The medical segmentation decathlon," *Nature Communications*, vol. 13, no. 1, pp. 4128, 2022.

[6] Bradley Efron and Robert J Tibshirani, *An introduction to the bootstrap*, CRC press, 1993.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[8] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, May 2018.

[9] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed, "Boundary loss for highly unbalanced segmentation," in *Medical Imaging with Deep Learning*, July 2019, vol. 102, pp. 285–296.

[10] Rosana EL Jurdi, Caroline Petitjean, Paul Honeine, Veronika Cheplygina, and Fahed Abdallah, "A surprisingly effective perimeter-based loss for medical image segmentation," in *Medical Imaging with Deep Learning*, 2021.

[11] Carole H. Sudre, Wenqi Li, Tom Kamiel Magda Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *MICCAI DLMIA Worshop*, 2017, vol. 2017, pp. 240–248.

[12] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, et al., "Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation," *Medical image analysis*, vol. 63, pp. 101694, 2020.

[13] Jingliang Zhao, Jie Zhao, Shumao Pang, and Qianjin Feng, "Segmentation of the true lumen of aorta dissection via morphology-constrained stepwise deep mesh regression," *IEEE TMI*, vol. 41, no. 7, pp. 1826–1836, 2022.

[14] Jiawei Chen, Ziqi Zhang, Xinpeng Xie, Yuexiang Li, Tao Xu, Kai Ma, and Yefeng Zheng, "Beyond mutual information: Generative adversarial network for domain adaptation using information bottleneck constraint," *IEEE TMI*, vol. 41, no. 3, pp. 595–607, 2022.

[15] Xiaoting Han, Lei Qi, Qian Yu, Ziqi Zhou, Yefeng Zheng, Yinghuan Shi, and Yang Gao, "Deep symmetric adaptation network for cross-modality medical image segmentation," *IEEE TMI*, vol. 41, no. 1, pp. 121–132, 2022.

[16] Jiahuan Song, Xinjian Chen, Qianlong Zhu, Fei Shi, Dehui Xiang, Zhongyue Chen, Ying Fan, Lingjiao Pan, and Weifang Zhu, "Global and local feature reconstruction for medical image segmentation," *IEEE TMI*, vol. 41, no. 9, pp. 2273–2284, 2022.

[17] Ying Chen, Darui Jin, Bin Guo, and Xiangzhi Bai, "Attention-assisted adversarial model for cerebrovascular segmentation in 3D TOF-MRA volumes," *IEEE TMI*, pp. 1–1, 2022.

[18] Jue Jiang, Andreas Rimner, Joseph O. Deasy, and Harini Veeraraghavan, "Unpaired cross-modality educed distillation (CMEDL) for medical image segmentation," *IEEE TMI*, vol. 41, no. 5, pp. 1057–1068, 2022.

[19] Alex Ling Yu Hung, Haoxin Zheng, Qi Miao, Steven S. Raman, Demetri Terzopoulos, and Kyunghyun Sung, "CAT-Net: A cross-slice attention transformer model for prostate zonal segmentation in MRI," *IEEE TMI*, pp. 1–1, 2022.

[20] Raghav Mehta, Thomas Christinck, Tanya Nair, Aurélie Bussy, Swapna Premasiri, Manuela Costantino, M. Mallar Chakravarthy, Douglas L. Arnold, Yarin Gal, and Tal Arbel, "Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference," *IEEE TMI*, vol. 41, no. 2, pp. 360–373, 2022.

[21] Bin Huang, Yufeng Ye, Ziyue Xu, Zongyou Cai, Yan He, Zhangnan Zhong, Lingxiang Liu, Xin Chen, Hanwei Chen, and Bingsheng Huang, "3D lightweight network for simultaneous registration and segmentation of organs-at-risk in CT images of head and neck cancer," *IEEE TMI*, vol. 41, no. 4, pp. 951–964, 2022.

[22] Jingkun Chen, Jianguo Zhang, Kurt Debattista, and Jungong Han, "Semi-supervised unpaired medical image segmentation through task-affinity consistency," *IEEE TMI*, pp. 1–1, 2022.

[23] Jeya Maria Jose Valanarasu, Vishwanath A. Sindagi, Ilker Hacihaliloglu, and Vishal M. Patel, "KiU-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation," *IEEE TMI*, vol. 41, no. 4, pp. 965–976, 2022.

[24] Rencheng Zheng, Qidong Wang, Shuangzhi Lv, Cuiping Li, Chengyan Wang, Weibo Chen, and He Wang, "Automatic liver tumor segmentation on dynamic contrast enhanced MRI using 4D information: Deep learning model based on 3D convolution and convolutional LSTM," *IEEE TMI*, vol. 41, no. 10, pp. 2965–2976, 2022.

[25] Han Wang, Fasheng Yi, Jingling Wang, Zhang Yi, and Haixian Zhang, "RECISTSup: Weakly-supervised lesion volume segmentation using RECIST measurement," *IEEE TMI*, vol. 41, no. 7, pp. 1849–1861, 2022.

[26] Xiahan Chen, Zihao Chen, Jun Li, Yu-Dong Zhang, Xiaozhu Lin, and Xiaohua Qian, "Model-Driven deep learning method for pancreatic cancer segmentation based on spiral-transformation," *IEEE TMI*, vol. 41, no. 1, pp. 75–87, 2022.

[27] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.