HALOS: Hallucination-Free Organ Segmentation After Organ Resection Surgery

Anne-Marie Rickmann[1,2](✉), Murong Xu[2], Tom Nuno Wolf[2], Oksana Kovalenko[2], and Christian Wachinger[1,2]

[1] Lab for Artificial Intelligence in Medical Imaging, Ludwig Maximilians University, Munich, Germany
[2] Department of Radiology, Technical University Munich, Munich, Germany
arickman@med.lmu.de

Abstract. The wide range of research in deep learning-based medical image segmentation pushed the boundaries in a multitude of applications. A clinically relevant problem that received less attention is the handling of scans with irregular anatomy, e.g., after organ resection. State-of-the-art segmentation models often lead to *organ hallucinations*, i.e., false-positive predictions of organs, which cannot be alleviated by oversampling or post-processing. Motivated by the increasing need to develop robust deep learning models, we propose HALOS for abdominal organ segmentation in MR images that handles cases after organ resection surgery. To this end, we combine missing organ classification and multi-organ segmentation tasks into a multi-task model, yielding a classification-assisted segmentation pipeline. The segmentation network learns to incorporate knowledge about organ existence via feature fusion modules. Extensive experiments on a small labeled test set and large-scale UK Biobank data demonstrate the effectiveness of our approach in terms of higher segmentation Dice scores and near-to-zero false positive prediction rate.

1 Introduction

Deep learning methods have become state-of-the-art for many medical image segmentation tasks, e.g. structural brain segmentation [18], tumor segmentation [12] or abdominal organ segmentation [4,5,8,17]. A challenge that remains is the generalization to unseen data, where a domain shift between training and testing data often leads to performance degradation. Research on robustness and domain adaptation [6] introduced new methods for handling domain shift, where the focus has mainly been on a shift in the intensity distribution of image data due to different imaging protocols, different scanner types or different modalities.

In contrast, a domain shift in the anatomy itself, e.g., by missing organs due to surgical organ resection has received less attention. In comparison to natural images, which can show image compositions with arbitrary objects, medical

A.-M. Rickmann and M. Xu—The authors contributed equally.

A. Frangi et al. (Eds.): IPMI 2023, LNCS 13939, pp. 667–678, 2023.
https://doi.org/10.1007/978-3-031-34048-2_51

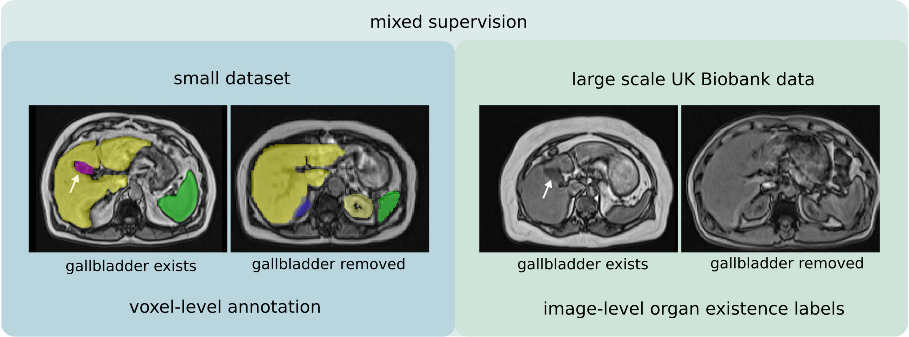

Fig. 1. Mixed supervision in HALOS using a small dataset with voxel-level annotations of multiple organs and a large-scale dataset with image-level binary labels of organ existence. The white arrow points to the gallbladder.

images of the human abdomen usually contain the same organs in the same ordering. This constraint of the human anatomy is beneficial for training networks and has, for instance, been explicitly used by incorporating shape priors [11,15,24]. However, as we move to clinical translation or to large-scale population studies, we will also encounter cases that do not follow the normal anatomy, which will yield a degradation in segmentation accuracy.

In this work, we mainly focus on gallbladder resection (cholecystectomy), as it is one of the most commonly performed abdominal surgeries. The indication for gallbladder removal is usually gallstones, which most of the time has no effect on other organs and the overall anatomy. We further evaluate our method on the cases of nephrectomy (kidney resection), where the indication can be more severe, e.g., kidney tumors, which could come with anatomical changes in other organs, like metastases. Further, kidneys are much larger than gallbladders, so that their removal can lead to post-surgical organ shift [19].

As we will demonstrate in our experiments, state-of-the-art segmentation networks often identify organs in the images, although they were removed. A phenomenon that we refer to as *organ hallucination.* We believe that organ hallucinations have so far not received more attention because publicly available segmentation datasets rarely contain cases after organ resection. This is probably due to the relatively small sample size of most segmentation datasets, as manual segmentation is time-consuming and costly. Fortunately, large-scale population imaging studies like the UK Biobank (UKB) Imaging study [10] with a targeted 100,000 subjects are becoming available that provide representative data of the population. The prevalence of cholecystectomies (gallbladder resection) in our sample of UK biobank is 3.7%, which provides enough data for studying this research question.

We introduce HALOS for the HALlucination-free Organ Segmentation after organ resection surgery. HALOS is a multi-task network that simultaneously learns classification of organ existence and segmentation of six abdominal organs (liver, spleen, kidneys, pancreas, gallbladder). HALOS is trained using mixed supervision, which accounts for the fact that we only have voxel-level annotations for a small dataset but image-level labels of organ removal on a large dataset, see Fig. 1. A key component of HALOS is a feature fusion module [22] that integrates the knowledge of organ existence into the segmentation branch. The key contributions are:

- a robust and flexible multi-task segmentation and classification model that predicts near-to-zero false positive cases on the UKB dataset
- the multi-scale feature fusion with the dynamic affine feature map transform [22] of the classification output into the segmentation branch
- a demonstration of the relevance of the missing organ problem by comparing to state-of-the-art segmentation models.

1.1 Related Work

Abdominal Multi-organ Segmentation. Nowadays, convolutional neural networks are state-of-the-art for abdominal organ segmentation in CT and MRI scans [3–5,17,21]. One method to point out is nnU-Net [8], which is an automatic pipeline to configure a U-Net to a given dataset. nnU-Net has won several medical image segmentation challenges, and has proven to be a robust and generic method. Therefore we consider nnU-Net as a baseline in our experiments.

Missing Organ Segmentation. To our best knowledge, the missing organ problem has so far only been studied for CT scans in [19], where an atlas-based approach is used. It trains a Gaussian Mixture Model on normal images and detects missing organs by analyzing fitting errors. However, this method inevitably relies on heavy simulation for parameter tuning and is therefore vulnerable to distribution shift. In a more recent method [20], the Dice loss was studied and it was argued that setting the reduction dimension over the complete batch would help to predict images with missing organs. However, the method was not tested on cases after organ resection. We compare to this approach in our experiments.

Classification-Assisted Segmentation. As image-level labels are easier to obtain than voxel-wise annotations, prior work has considered including these additional labels by extending the segmentation network with a classification branch [13,14,23]. In [14], the two branches are trained jointly using both fully-annotated and weakly-annotated data with shared layers at the beginning, for 2D brain tumor segmentation and classification of tumor existence. They showed that the additional classification significantly improved segmentation performance compared to standard supervised learning. We compare to this approach in our experiments.

Feature Fusion. Some approaches for classification-assisted segmentation use feature fusion, i.e., the interweaving of segmentation and classification branches. For example, separate segmentation and classification models are trained in [23] for Covid-19 diagnosis. Feature maps of the classification and segmentation model are merged with Squeeze-and-Excitation (SE) blocks [7]. After the feature fusion, the enhanced feature map is fed into the decoder for segmentation. An alternative for feature fusion is the combination with metadata, such as age, gender, or measurements of biomarkers. The Dynamic Affine Feature Map Transform (DAFT) [22] predicts the scales and shifts to excite or repress feature maps on a channel-level from such metadata, as seen in Fig. 2.

2 Methods

Figure 2 illustrates the dual-branch classification-assisted segmentation pipeline of HALOS that combines Multitask Learning and Feature Fusion to handle missing organs. In the following, we describe each part of our pipeline in detail.

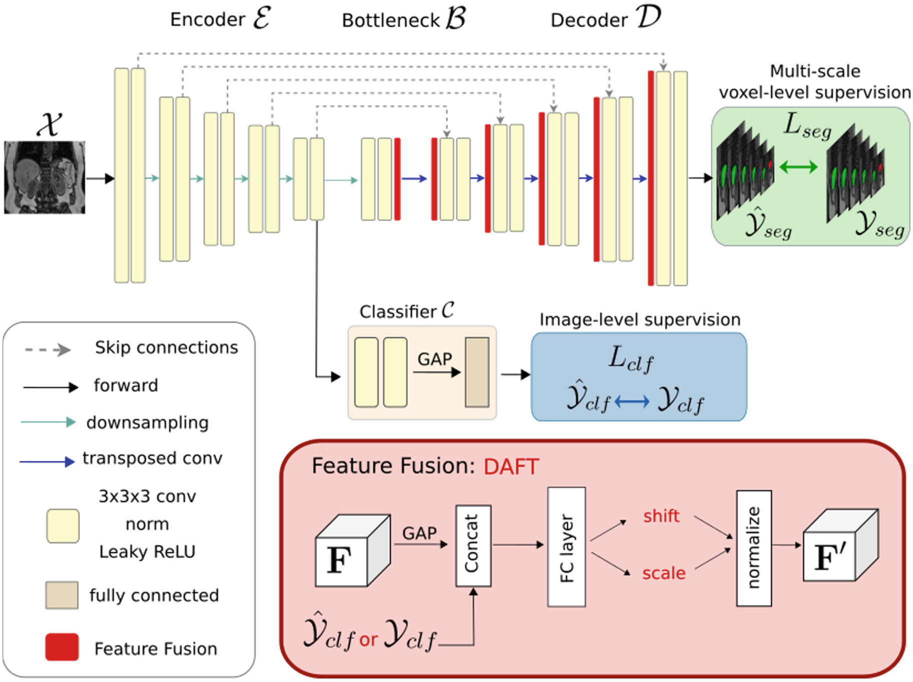

Fig. 2. Overview of the HALOS multi-task pipeline.

Segmentation Branch. In the segmentation branch, we use a U-Net architecture, based on nnU-Net [8] as the segmentation network which consists of an encoder $\mathcal{E}$, bottleneck $\mathcal{B}$, and a decoder $\mathcal{D}$. As previously mentioned, nnU-Net [8] is one of the most generic and well-performing medical image segmentation models. The nnU-Net pipeline automatically determines the best U-Net architecture and data augmentation for the given data. Therefore, we fed our segmentation dataset into the nnU-Net pipeline and took the architecture of the best-performing nnU-Net model and the data augmentation scheme as our baseline. The resulting model is a 3D U-Net with 32 starting channels and 5 downsampling levels.

The advantage of using encoder-decoder structured networks is that intermediate representations can be obtained at different scales. The U-Net is trained on input MR images $\mathcal{X}$ and voxel-level annotations $\mathcal{Y}_{seg}$ to output segmentation predictions $\hat{\mathcal{Y}}_{seg}$ under full supervision. The segmentation loss is defined as an average of Dice and Cross-Entropy loss L_{seg} with enabling of deep supervision at each feature map scale and dynamic class weights for individual images:

$$\begin{aligned} L_{seg} &= L_{CE} + L_{Dice}, \qquad L_{CE} = -\frac{1}{N}\sum_{i}^{C} y_i \log(\hat{y}_i), \\ L_{Dice} &= 1 - \frac{2 \cdot |\hat{\mathcal{Y}}_{seg} \cap \mathcal{Y}_{seg}| + \epsilon}{|\hat{\mathcal{Y}}_{seg}| + |\mathcal{Y}_{seg}| + \epsilon}, \end{aligned} \tag{1}$$

where we denote the class-wise ground truth y_i, class-wise predictions $\hat{y}_i$, the number of classes C and samples N, a smoothing term ϵ. Note that some implementations of the Dice loss only add ϵ to the denominator, to avoid division by zero. In our case, it is important to add ϵ to numerator and denominator, as we want to ensure a Dice loss of 0, rather than 1, for true negative predictions of gallbladders.

Classification Branch. Compared to manual voxel-level annotations, the global image-level labels are less informative but can be obtained at a substantially lower cost. Hence, we incorporate the classification task into the pipeline to study the impact of the low-dimensional prior knowledge on the final predicted segmentation. In the classification branch, classifier $\mathcal{C}$ is built on top of the encoder $\mathcal{E}$ and takes a feature map from a specific encoder block as input. The precise location of the classifier can be tuned as a hyper-parameter, but we found encoder blocks 4 and 5 promising for most models. Compared to training a standalone classification model, such a shared feature structure between $\mathcal{C}$ and $\mathcal{E}$ enables a more lightweight classification model and thus saves redundant computation. $\mathcal{C}$ consists of a convolutional block with the same structure as an encoder block, a 3D global average pooling step, and a fully connected layer for producing the final classification. The classifier is trained on MR scans with image-level surgery labels $\mathcal{Y}_{clf}$. The classification loss L_{clf} is the average cross-entropy weighted by the actual class ratio in the training set.

Feature Fusion. A key component of HALOS is the feature fusion module. The prior information about the resection of the gallbladder is fused with the feature maps of the segmentation branch at multiple locations. As shown in Fig. 2, these locations are the bottleneck and each stage of the decoder. Importantly, we can either use the ground truth image-level labels $\mathcal{Y}_{clf}$ or the classifier's prediction $\hat{\mathcal{Y}}_{clf}$ as input to the feature fusion, depending on whether the information about previous surgeries is available at test time.

We use DAFT [22] to perform feature fusion, which was originally designed to combine 3D images with low-dimensional tabular information, and can be conveniently integrated into any type of CNN. In our case, the tabular data to be concatenated is the binary classification result or ground truth label about gallbladder resection. To the best of our knowledge, DAFT has not yet been used in segmentation models or in a multi-scale fashion. We expect that information sharing at multiple scales of the decoder will emphasize the prior knowledge about the organ's presence and conduce the decoder to produce fewer false positive predictions of non-existing classes. The exact position of integrating feature fusion modules into the U-Net architecture is illustrated in Fig. 2. The classification labels are fused to the bottleneck feature map, which contains the highest-level information. Then the fused version will be forwarded to the decoder where we repeat the feature fusion blocks after each transpose convolution. We place feature fusion via DAFT before each decoder block, which avoids interaction with other normalization layers. Formally, for each item in a batch, let $\hat{\mathbf{y}} \in \mathbb{R}$, be the predicted output from the classifier, and $\mathbf{F}_{d,c} \in \mathbb{R}^{D \times H \times W}$, where D, H, W denote the depth, height, and width of the feature map, the c-th channel of the input feature map of block $d \in \{0, \ldots, 5\}$ in the decoder, as illustrated in Fig. 2. DAFT [22] learns to predict scale $\alpha_{d,c}$ and offset $\beta_{d,c}$

$$\mathbf{F}'_{d,c} = \alpha_{d,c}\mathbf{F}_{d,c} + \beta_{d,c}, \tag{2}$$

$$\alpha_{d,c} = f_c(\mathbf{F}_{d,c}, \hat{\mathbf{y}}_d), \qquad \beta_{d,c} = g_c(\mathbf{F}_{d,c}, \hat{\mathbf{y}}_d), \tag{3}$$

where f_c, g_c are arbitrary mappings from image and tabular As proposed in [22], a single fully connected neural network h_c models f_c, g_c and outputs a single α-β-pair.

During training, we randomly sample MR images with voxel-level and image-level labels to form batches and use them to update the segmentation model and classifier respectively. With the previously defined L_{seg} and L_{clf}, the final loss of HALOS is

$$L = \alpha \cdot L_{seg} + (1 - \alpha) \cdot L_{clf}, \tag{4}$$

where α indicates the weight assigned to the segmentation loss.

3 Results and Discussion

3.1 Experiment Setup

Segmentation Data. We use whole-body MRI scans with voxel-level annotations from three different sources: the German National Cohort (NAKO) [2], the Cooperative Health Research in the Region of Augsburg (KORA) [1], and UKB [10]. The samples cover a general population from Germany and the UK. All three studies acquired abdominal images with a two-point Dixon sequence, where we use the oppose-phase scans in this work. For pre-processing, we follow guidelines of other work [9,16]. The scans were manually segmented by a medical expert. The dataset contains 63 scans in total (16 NAKO, 15 KORA, 32 UKB), of which 18 are patients after gallbladder resection. We have split this data into 42(9) scans for training, 7(3) for validation and 11(6) scans for testing, the count of missing gallbladder cases is given in parentheses.

UKB Data. The UK Biobank dataset is much larger than the segmentation data, but only contains image-level annotations indicating organ presence. We use it for training the organ existence classifier in our multi-task pipeline. It can also be used for evaluating the model robustness since we can count the false positive segmentations of non-existing gallbladders. We used the information about past surgeries from the UKB database and our medical expert verified the labels for correctness. Out of 19,000 images we requested from UK Biobank, we counted 701 after gallbladder removal. We additionally randomly selected normal subjects. We split the data into two subsets, one for training and validation of models (899 scans with and 349 without gallbladder), and one which serves as an unseen test set (952 scans with and 352 without gallbladder). The ratio of no-gallbladder cases in each subset is set to be roughly 0.4.

Implementation Details and Hyperparameter Tuning. In this work, we use GPUs DGX A100 for running our experiments. The implementation is based on Python, PyTorch and MONAI. We perform hyperparameter tuning for the loss weight α, weight decay, learning rates for the segmentation model and classifier, normalization type (instance or batch normalization), batch size, and the location of the classifier using Ray Tune. We train our models using the automated mixed precision of PyTorch. Our code is publicly available at https://github.com/ai-med/HALOS.

Metrics. We evaluate our models by comparing Dice scores for all organs and false positive rate (FPR) of gallbladder segmentations. We define a sample as false positive if one or more voxels have been segmented as non-existing gallbladder. As the Dice score is not defined for non-existing organs, we define it to be 1 for true negative cases and 0 for false positive cases. Therefore, we can observe large changes in the Dice score when reducing the false positive rate.

Baselines. Apart from the nnU-Net baseline as described in Sect. 2, we further choose two alternative baselines, i.e., oversampling and post-processing. For oversampling, we oversample the cases without a gallbladder in training to achieve a balance in class frequency. Note that we are already weighting the loss functions by class frequency. The post-processing baseline is another method, where we use the prior information about gallbladder resection to remove false positives as a direct post-processing step.

Table 1. Comparison of HALOS with baseline nnU-Net, with oversampling, post-processing, Dice loss with batch reduction [20] (Dice batch red.), and multi-task model [14]. FF: feature fusion, gt: ground truth labels at test-time. We list Dice scores for all organs and false positive rate (FPR) for removed gallbladders. We provide mean and standard deviation over 5-fold cross-validation. *: best architecture for our data proposed by the nnU-Net pipeline was re-implemented.

	Dice Scores ↑						FPR ↓
Method	Mean	liver	spleen	r kidney	l kidney	pancreas	gallbl.	
nnU-Net* [8]	0.823±0.014	0.938±0.004	0.891±0.006	0.898±0.003	0.894±0.002	0.643±0.016	0.674±0.076	0.267±0.149
+ oversampling	0.832±0.008	0.940±0.006	0.894±0.005	0.901±0.005	0.891±0.005	0.655±0.011	0.712±0.052	0.233±0.091
+ post-proc. (gt)	0.847±0.005	0.938±0.004	0.891±0.006	0.898±0.003	0.894±0.002	0.643±0.016	0.819±0.009	0±0
+ batch red. [20]	0.818±0.010	0.945±0.002	0.895±0.002	0.901±0.005	0.894±0.006	0.663±0.014	0.610±0.045	0.400±0.091
multi-task [14]	0.822±0.010	0.930±0.006	0.879±0.004	0.895±0.003	0.885±0.002	0.625±0.016	0.716±0.054	0.233±0.091
HALOS w/o FF	0.825±0.010	0.941±0.002	0.892±0.009	0.898±0.004	0.892±0.005	0.657±0.013	0.668±0.073	0.3±0.139
HALOS (pred, gt)	0.853±0.002	0.939±0.003	0.899±0.005	0.899±0.003	0.893±0.004	0.649±0.021	0.840±0.015	0±0

3.2 Experiments on Cholecystectomy Cases

We train the baseline nnU-Net, oversampling and post-processing baselines, state-of-the-art methods [14,20] and HALOS using 5-fold cross-validation and report the average results over all folds on the segmentation test set in Table 1 and on the UKB test set in Table 2. The average FPR is quite high for the baseline nnU-Net on both datasets, which leads to a low gallbladder Dice score of 0.674. The segmentation performance on pancreas is also low 0.643, but the pancreas is very hard to segment, due to its shape variability. The oversampling only slightly improves performance, so we can assume that the reason for the high FPR is not only caused by class imbalance. As expected, the post-processing leads to higher gallbladder Dice scores and zero FPR, since it uses the ground truth information about cholecystectomy. A shortcoming of the post-processing is that the model's false positive prediction may appear in neighboring organs, which will result in a hole in the segmentation. The gallbladder usually lies in fossa vesicae biliaris, which is a depression on the visceral surface of the liver anteriorly, between the quadrate and the right lobes. Since the location is closely connected to the liver, we found many mistakes produced by our baseline that are either localized inside the liver or partly in the liver and partly in other tissues like visceral fat. Examples of typical organ hallucinations are shown in Fig. 3C–D, where the gallbladder is predicted in the fossa vesicae biliaris (C), inside the liver (D) and in the intestine (E). The recent work [20] proposes to set ϵ in the Dice loss to a low value, e.g. 10^{-7}, the batch size higher than 1 and to reduce the Dice loss over the batch dimension. We set the batch size to 8, which reached the limit of our GPU memory. Note that in our baseline model the batch size is set to 2, ϵ is 1 and we also reduce over the batch dimension. Interestingly, we observe an increase in FPR for both datasets. In preliminary experiments, we have removed the batch reduction in the Dice loss, but we have observed no significant difference in performance. The multi-task model proposed in [14] includes a classifier right before the segmentation output of the decoder. We use our nnU-Net model and extend it with a classifier, following the architecture

Table 2. Comparison of HALOS with baseline nnU-Net, oversampling, post-processing, Dice loss with batch reduction [20] (+ batch red.) and multi-task model [14] on the UKB dataset. FF: feature fusion, gt: ground truth labels for FF at test-time, pred: classification predictions for FF at test-time. We provide false positive (FP), false negative (FN), true positive (TP), true negative (TN), false positive rate (FPR) and F1 score for removed gallbladders, and the balanced accuracy (BAcc) of all classifiers. All values are mean and standard deviation over 5-fold cross-validation. *: best architecture for our data proposed by the nnU-Net pipeline was re-implemented.

Method	FP ↓	TN ↑	TP ↑	FN ↓	FPR ↓	F1 ↑	BAcc ↑
nnU-Net* [8]	91.2 ±30.62	260.8 ±30.62	537.2 ±16.62	62.8 ±16.62	0.259 ±0.087	0.875 ±0.009	
+ oversampling	66.6 ±6.633	285.4 ±6.633	522.6 ±13.18	77.4 ±13.18	0.189 ±0.028	0.879 ±0.011	
+ post-proc. (gt)	0 ±0	352 ±0	537.2 ±16.62	62.8 ±16.62	0 ±0	0.945 ±0.015	
+ batch red. [20]	135.2 ±57.15	216.8 ±57.15	530.2 ±24.39	69.8 ±24.39	0.384 ±0.162	0.838 ±0.017	
multi-task [14]	100.2 ±16.48	251.8 ±16.48	578.2 ±3.701	21.8 ±3.701	0.285 ±0.047	0.905 ±0.054	0.874 ±0.045
HALOS w/o FF	52.6 ±17.67	299.4 ±17.67	547.4 ±22.39	52.6 ±22.39	0.149 ±0.050	0.869 ±0.056	0.896 ±0.047
HALOS (gt)	2 ±2.550	350 ±2.550	564.8 ±14.74	35.2 ±14.74	0.006 ±0.007	0.968 ±0.010	0.933 ±0.005
HALOS (pred)	11 ±5.339	341 ±5.339	541.8 ±14.20	58.2 ±14.20	0.031 ±0.015	0.940 ±0.010	0.933 ±0.005

proposed in [14] at decoder block 5. This model leads to a slight decrease in FPR on the segmentation data, but to a higher FPR on the UKB data. To analyze the impact of multi-task learning and the feature fusion models, we train HALOS without the feature fusion, which interestingly leads to a slight increase in FPR and an decrease in gallbladder Dice score and slight decrease in FPR on UKB, compared to nnU-Net, even though the balanced accuracy of the classifier is 0.896. Therefore we argue, that multi-task training alone is not sufficient to reduce organ hallucinations. HALOS was trained using the ground truth labels $\mathcal{Y}_{clf}$ as input for feature fusion, and leads to an impressive reduction of the FPR to 0 on the segmentation data and 0.006 on the UKB data. The multi-task classifier achieves a balanced accuracy of 0.93. When we use the classifier's prediction for feature fusion at test time, we observe a slight increase in FPR over using the ground truth labels to 0.03. This shows, that our method is flexible and depending if prior information about gallbladder resection is available at test time or not, one can either fuse the ground truth labels or the classifier's predictions.

3.3 Experiments on Nephrectomy Cases

To evaluate if HALOS can be applied to other organ resection cases, we validate the effectiveness of HALOS on cases after nephrectomy. Note that we did not do any further hyper-parameter tuning in this experiment. We create a kidney segmentation dataset that contains 46(6/2) scans for training and 10(2/1) for testing, the count of missing kidney cases is given in parentheses with the format left/right. For UKB data, we split the available subjects into one training and validation set (200 scans with 17/5 missing left/right kidneys), and another hold-out test set (55 scans with 4/2 no-kidneys). Similar to gallbladder experiments, we train the baseline nnU-Net and HALOS using 5-fold cross-validation. Note

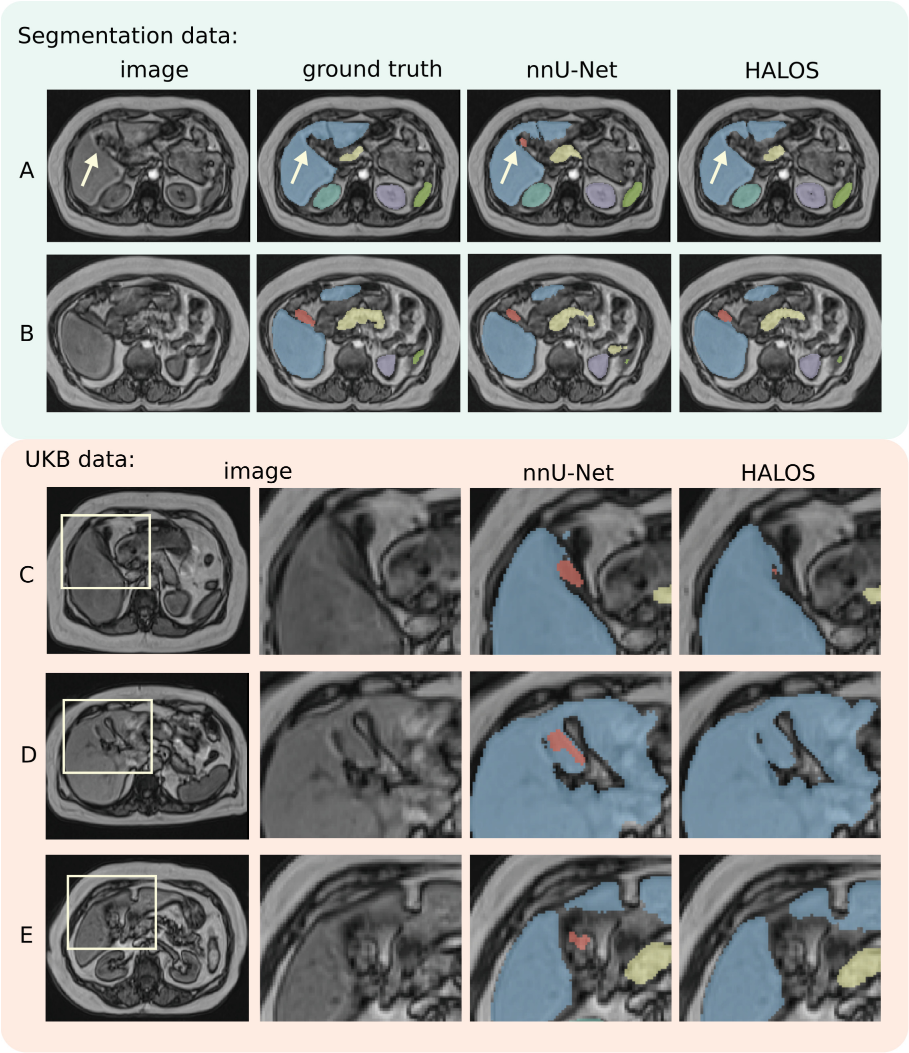

Fig. 3. Segmentation results on the segmentation data (top) and UKB (bottom). Comparison of nnU-Net and HALOS. A: scan with a resected gallbladder, nnU-Net produces a false positice. B: scan with an existing gallbladder. C: both models predict a false positive in the location where the gallbladder was resected. D: nnU-Net produces a false positive inside the liver. E: nnU-Net produces a false positive in the intestine.

that the classifier learns a multi-class classification, in contrast to the binary classification in the gallbladder experiments. We report the results of the nephrectomy experiment in the following. The baseline nnU-Net achieved an FPR of 0.2 for left kidney and 1 for right kidney on the UKB data. We observe that HALOS achieves a lower FPR of 0 for the left kidney and still high FPR of 0.7

for the right kidney, while having a significantly reduced voxel-level FP count of 16.5, compared to 129 for nnU-Net. The left kidney Dice of HALOS (0.9024) is higher than of nnU-Net (0.8413) while having no improvement on the right kidney 0.864 vs 0.867. A possible reason might be the small dataset size with severe class imbalance, of having only two cases with missing right kidneys in the training set. The balanced accuracy of the HALOS classifier is 0.93 for left kidney and 0.58 for right kidney, which also suggests that the class imbalance has more impact in this setting.

4 Conclusion

In this work, we introduced HALOS, a multi-task classification and segmentation model for hallucination-free organ segmentation. We propose to use multi-scale feature fusion, via the dynamic affine feature-map transform, to enrich the feature maps of the segmentation branch with prior information on organ existence. We have shown on cases after cholecystectom0,y and nephrectomy, that HALOS significantly reduces false positive predictions on a large scale UK Biobank test set, and increases gallbladder and left kidney Dice scores on a smaller segmentation test set, compared to nnU-Net and several additional baselines and multi-task approaches. HALOS is flexible to use ground truth organ existence labels at test-time or the prediction of the classifier, depending on the availability of such labels. In future work we would like to extend HALOS to additional cases of organ resection, e.g. hysterectomy (removal of uterus) or splenectomy (removal of spleen).

Acknowledgment. This research was partially supported by the Bavarian State Ministry of Science and the Arts and coordinated by the bidt, the BMBF (DeepMentia, 031L0200A), the DFG and the LRZ.

References

1. Bamberg, F., et al.: Subclinical disease burden as assessed by whole-body MRI in subjects with prediabetes, subjects with diabetes, and normal control subjects from the general population: the KORA-MRI study. Diabetes **66**(1), 158–169 (2017)
2. Bamberg, F., et al.: Whole-body MR imaging in the German national cohort: rationale, design, and technical background. Radiology **277**(1), 206–220 (2015)
3. Bobo, M.F., et al.: Fully convolutional neural networks improve abdominal organ segmentation. In: Medical Imaging 2018: Image Processing, vol. 10574, p. 105742V. International Society for Optics and Photonics (2018)
4. Chen, Y., et al.: Fully automated multi-organ segmentation in abdominal magnetic resonance imaging with deep neural networks. Med. Phys. **47**(10), 4971 (2020)
5. Gibson, E., et al.: Automatic multi-organ segmentation on abdominal CT with dense V-networks. IEEE Trans. Med. Imaging **37**(8), 1822–1834 (2018)
6. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. **69**(3), 1173–1185 (2021)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)

8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
9. Kart, T., et al.: Automated imaging-based abdominal organ segmentation and quality control in 20,000 participants of the UK Biobank and German national cohort studies. Sci. Rep. **12**(1), 1–11 (2022)
10. Littlejohns, T.J., et al.: The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat. Commun. **11**(1), 1–12 (2020)
11. Liu, L., Wolterink, J.M., Brune, C., Veldhuis, R.N.: Anatomy-aided deep learning for medical image segmentation: a review. Phys. Med. Biol. **66**(11), 11TR01 (2021)
12. Liu, Z., et al.: Deep learning based brain tumor segmentation: a survey. Complex Intell. Syst. **9**(1), 1001–1026 (2023)
13. Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L.: Y-Net: joint segmentation and classification for diagnosis of breast biopsy images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 893–901. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_99
14. Mlynarski, P., Delingette, H., Criminisi, A., Ayache, N.: Deep learning with mixed supervision for brain tumor segmentation. J. Med. Imaging **6**(3), 034002 (2019)
15. Oktay, O., et al.: Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE Trans. Med. Imaging **37**(2), 384–395 (2017)
16. Rickmann, A.M., Senapati, J., Kovalenko, O., Peters, A., Bamberg, F., Wachinger, C.: AbdomenNet: deep neural network for abdominal organ segmentation in epidemiologic imaging studies. BMC Med. Imaging **22**(1), 1–11 (2022)
17. Roth, H.R., et al.: Hierarchical 3D fully convolutional networks for multi-organ segmentation. arXiv preprint arXiv:1704.06382 (2017)
18. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N.: QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. Neuroimage **186**, 713–727 (2019)
19. Suzuki, M., Linguraru, M.G., Okada, K.: Multi-organ segmentation with missing organs in abdominal CT images. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7512, pp. 418–425. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33454-2_52
20. Tilborghs, S., Bertels, J., Robben, D., Vandermeulen, D., Maes, F.: The dice loss in the context of missing or empty labels: introducing Φ and ϵ. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention (MICCAI 2022). LNCS, vol. 13435, pp. 527–537. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_51
21. Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L.: Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Med. Image Anal. **55**, 88–102 (2019)
22. Wolf, T.N., Pölsterl, S., Wachinger, C., Initiative, A.D.N., et al.: DAFT: a universal module to interweave tabular data and 3D images in CNNs. Neuroimage **260**, 119505 (2022)
23. Wu, Y.H., et al.: JCS: an explainable COVID-19 diagnosis system by joint classification and segmentation. IEEE Trans. Image Process. **30**, 3113–3126 (2021)
24. Zhou, Y., et al.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: ICCV (2019)