Boosting Liver and Lesion Segmentation from CT Scans by Mask Mining

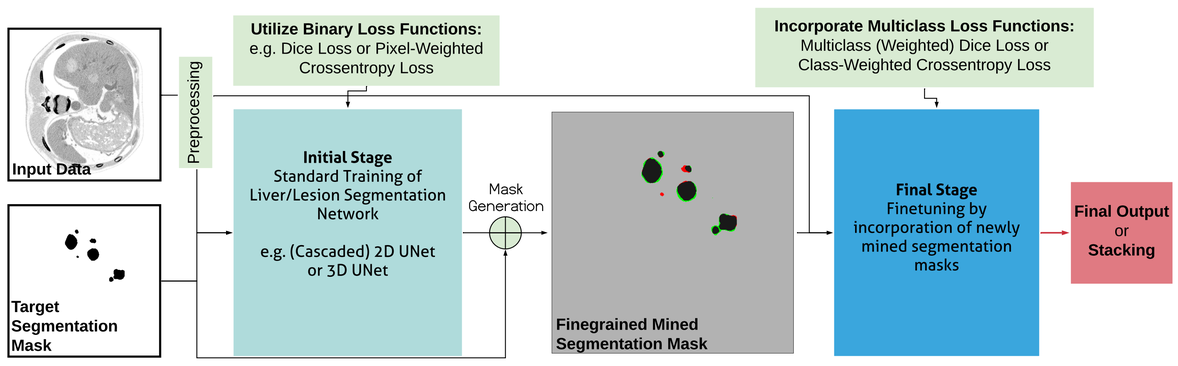
Karsten Roth[a,b], Jürgen Hesser[a,b], Tomasz Konopczyński[a,b,c]

[a] Department of Radiation Oncology, University Medical Center Mannheim,
[b] IWR, ZITI, Heidelberg University [c] Tooploox

Introduction

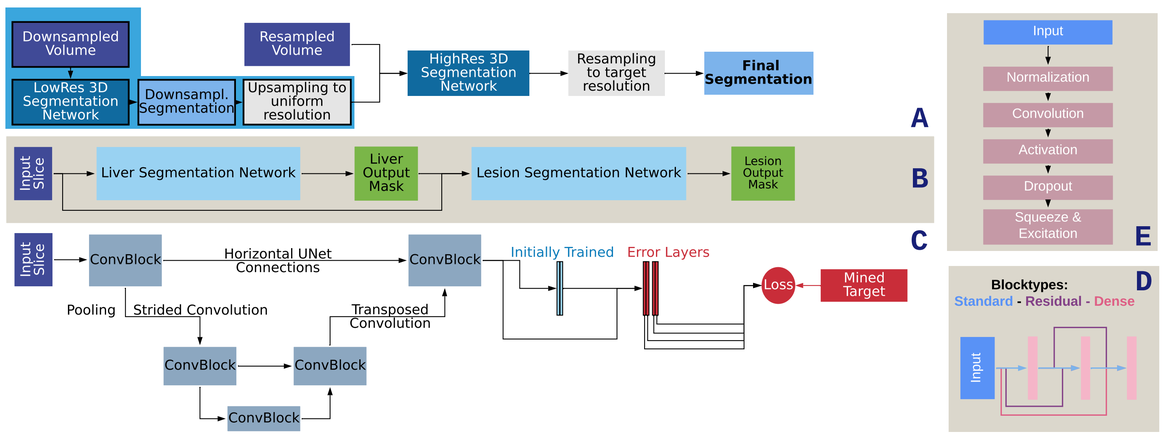
Objective: Improve fine-grained 2D and 3D Segmentation of liver and liver lesions from CT data.
Challenge: Propose an architecture-independent extension to U-Net based segmentation methods and handle fine-grained segmentation errors.
Contribution: Through a secondary refinement step fitting the current network to previous segmentation errors (Mask Mining), we achieve improved target pixel recovery and control over segmentation properties of the network.

Pipeline Setup



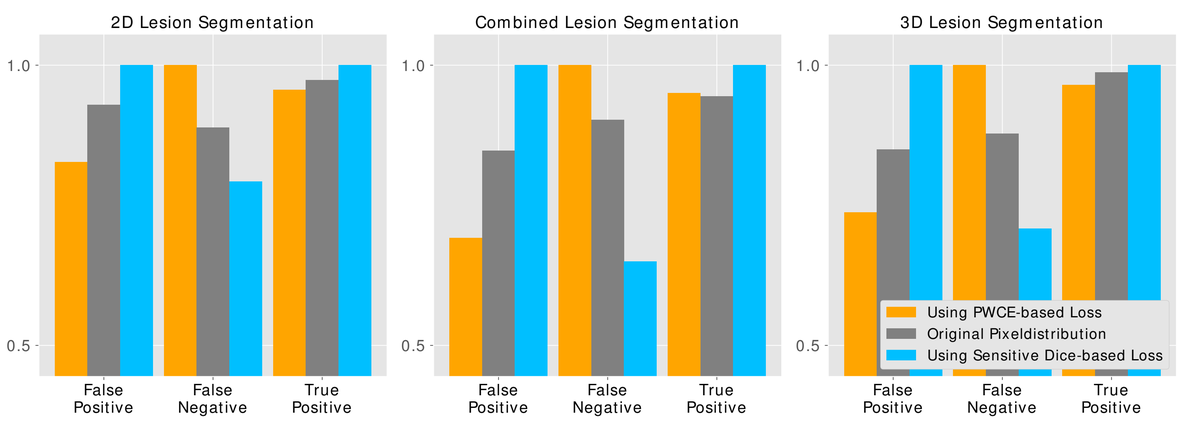
Our proposed setup is build on a two-level training stack: After a first standard training run, learned segmentation masks are compared with the ground truth masks to generate new fine-grained training masks. These masks containing previously made segmentation errors s.a. false positives and false negatives. Retraining on these errors allows the network to learn to explain away mistakes.

Evaluated Architectures



Evaluated 2D and 3D architectures: **(A)** (two-stage) 2D and 3D U-Net pipeline with extension elements shown in (D) and (E). **(B)** Simultaneous training of liver and lesion segmentation networks. **(C)** showcases the minor adjustments in the final convolutional layer to allow for multiclass error prediction.

Control Study



Control of produced segmentation error types: the distribution of segmentation error types before and after running a mask mining step is compared over all proposed architectures. A clear shift in false-positive and false-negative pixels (loss-dependent) can be seen. The network segmentation behaviour changes for different loss functions, with generally improved segmentations.

Loss Functions and Binary Mask Generation

The retraining on mined masks is performed using either or a smooth dice loss:
A smooth dice loss

$$L^{dice}(x^k, \phi) = \frac{1}{K} \sum_{k=c=0}^{K,C} \frac{\sum_{i=j=0}^{H,W} \phi_{ij}^{c}(x^k) \cdot t_{ij}^{c,k}}{\sum_{i=j=0}^{H,W} \phi_{ij}^{c}(x^k) + \sum_{i=j=0}^{H,W} t_{ij}^{c,k} + \epsilon} \tag{1}$$

or *a multiclass pixelweighted crossentropy loss*

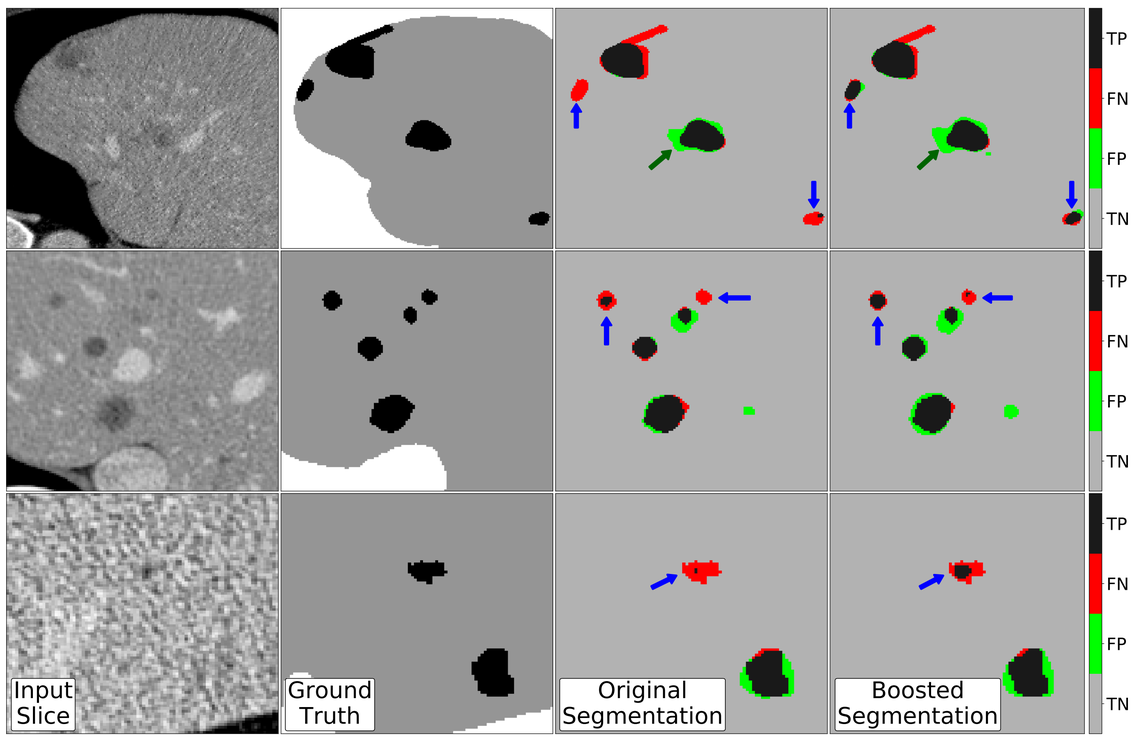
$$L^{pwce}(x^k, \phi) = \frac{1}{KC} \sum_{k=i=j=c=0}^{K,C,H,W} t_{ij}^{c,k} \cdot \log\left(\phi_{ij}^{c}(x^k)\right) \tag{2}$$

with images $\{x^k\}_{k\in[1,K]}$ in minibatch K, network ϕ, number of target classes C and the target mask $t^k \in \mathcal{N}^{W\times H}$.
The multiclass output is combined to the final binary segmentation mask as follows:

$$O_{ijm}^{k}(x^k) = \lfloor \frac{\text{argmax}_{c\in[0,..,C-1]}\ \phi_{ijm,c}^{multi}(x^k)}{2} \rfloor \tag{3}$$

where each class index from 0 to 3 denote *true negatives*, *false positives*, *false negatives* and *true positives*.

Qualitative Evaluation



Examples: When comparing to the initial segmentation, we see improved target pixel recovery of important false negative predictions (red area reduction, see blue arrows). While this comes at the cost of slightly more false positive predictions (green area, green arrows), the overall performance is boosted.

Quantitative Results

Averaged dice score per volume before and after method application: For all architectures, relative improvement matters, with results in the table below. These show consistent gains over initial models.

Setup	Training Dice		Validation Dice		Online Test Dice	
	LIV	LES	LIV	LES	LIC	LES
2D	96.9	71.9	95.9	63.5	95.3	62.9
Inc.	**97.0**	**73.7**	**96.3**	**64.9**	**95.5**	**63.5**
3D	92.2	63.0	91.4	56.8	91.2	55.5
Inc.	**94.2**	**66.1**	**91.8**	**57.7**	**92.0**	**56.5**
Cmb	94.5	70.1	92.9	61.6	93.4	61.9
Inc.	**96.2**	**72.3**	**94.0**	**63.4**	**94.7**	**63.0**

The inclusions of mined trained masks into the training process specifically benefits validation performance. This is rooted in the splitting procedure, as training and validation set are drawn from the same sample set. Due to different sources contributing to the dataset, the test set samples therefore differ much stronger from the training set. Newly mined features are therefore more expressive on the validation set.

karsten.roth@stud.uni-heidelberg.de