Strategies for Training Deep Learning Models in Medical Domains with Small Reference Datasets

Gerald A. Zwettler
RG Advanced Information Systems and Technology (AIST), Department of Software Engineering, University of Applied Sciences Upper Austria
Softwarepark 11
4232 Hagenberg, Austria
gerald.zwettler@fh-hagenberg.at

David R. Holmes III
Biomedical Analytics and Computational Engineering Lab, Dept. of Physiology and Biomedical Engineering, Mayo Clinic College of Medicine
200 1st St SW,
MN 55905, Rochester, USA
holmes.david3@mayo.edu

Werner Backfrieder
Medical Informatics, Department of Software Engineering, University of Applied Sciences Upper Austria
Softwarepark 11
4232 Hagenberg, Austria
werner.backfrieder@fh-hagenberg.at

ABSTRACT

With the steady progress of Deep Learning (DL), powerful tools are now present for sophisticated segmentation tasks. Nevertheless, the generally very high demand for training data and precise reference segmentations often cannot be met in medical domains when processing small and individual studies or acquisition protocols. As common strategies, reinforcement learning or transfer learning are applicable but coherent with immense effort due to domain-specific adjustment. In this work the applicability of a U-net cascade for training on a very low amount of abdominal MRI datasets of the parenchyma is evaluated and strategies to compensate for the lack of training data are discussed. Although the model accuracy when training on 13 MRI volumes with achievable JI=89.41 is rather low, results are still good enough for manual post-processing utilizing a Graph cut (GC) approach with medium demand for user interaction. This way, the DL models are retrained, when additional test data sets become available to subsequently improve the classification accuracy. With only 2 additional GC post-processed datasets, the accuracy after model re-training is increased to JI= 89.87. Besides, the applicability of Generative Adversial Networks (GAN) in the medical domain is evaluated discussing to synthesize axial CT slices together with perfect ground truth reference segmentations. It is shown for abdominal CT slices of the parenchyma, that in case of lack of training data, synthesized slices, that can be derived at arbitrary number, help to significantly improve the DL training process when only an insufficient amount of data is available. While training on 2,200 real images only leads to accuracy JI=88.75, the enrichment with 2,200 additional images synthesized from a GAN trained on 5,000 datasets only leads to an increase up to JI=92.02. Even if the DL model is exclusively trained on 4,400 computer-generated images, the classification accuracy on real-world data is notable with JI=90.81.

Keywords

Medical Image Segmentation, Deep Learning, Generative Adversial Networks, Graph cut.

1. INTRODUCTION

The research and development on preferably automated, generic and precise segmentation strategies for processing medical image datasets has been of high importance since the very first computed tomographic image acquisition devices in the early 1970s. Since then many semi-automatic segmentation concepts have been proposed, but most of them conserve the medical diagnostician and its experience-based evaluation as central criterion for the final decision. While there are some few off-the-shelf applications available for specific diagnostic domains [Chr18], in general computer-aided diagnostics is still achieved in a user-centric process utilizing tools and frameworks for semi-automated image processing [Str15].

1.1 Field of Medical Application

Whenever quantitative evaluation is required for computer-based planning of a surgery, evaluating the success of the therapy or progress of a degenerative disease in a follow up study [Agg11], a precise segmentation of the target structures is inevitable. With emerging progress in 3D visualization with respect to the mixed reality continuum and the 3D print of anatomical structures as 3D models for surgery planning, training and education [Squ18] the

https://doi.org/10.24132/JWSCG.2020.28.5

fields of application for medical image processing are steadily growing.

A broad range of semi-automated image processing tools is available for user-centric processing in particular diagnostic domains. Nevertheless, evaluation of datasets with these tools requires a high level of experience in both, the medical and the technical domain to prevent from misapplication, errors and subjective bias of the results. To close this gap, the application of standardized image processing chains for processing medical tomographic datasets is recommended in radiographer training [Zwe13].

1.2 State of the Art

Since the beginnings of computer-based medical image analysis, besides semi-automatic and user-oriented methods such as Region Growing, Graph cut or Live-Wire [Son13], shape-modeling approaches have been a strategy for automation of the segmentation process in specific medical domains. Rudimental a priori knowledge regarding shape is incorporated for deformable models [McI96] or Level Set segmentation [Set99] while statistical shape models [Coo92] are trained on a large dataset of reference shapes representing the expected anatomical variability. While statistical shape models lead to a domain-specific compact representation of shape due to PCA, the generic use is still hard to achieve as extraction of corresponding landmarks is very domain-specific and difficult for non-prominent structure shapes. Incorporating the expected intensity profile besides shape makes Active Appearance Models [Coo98] relevant in several diagnostic domains but their most significant field of application is still human face recognition.

Advances in available GPU hardware and machine learning frameworks such as Tensorflow/Keras lead to an immense boost in Deep Learning (DL) applications, facilitating practical use of formerly only theoretically specified concepts. Conventional Feed Forward networks, already applied in medical domains such as multi-modal image fusion [Zha11] are steadily enriched by an increasing number of hidden layers thus increasing the overall number of trainable parameters. All common Deep Learning Concepts found their way and application in specific medical domain. For instance self-organizing maps neural networks [Koh97] that are due to their grid/graph nature good for complex clustering tasks, were successfully applied for classification of renal diseases too [Van98]. Advanced semantic interpretation of input signals is of high relevance in natural langue, optical character or audio processing and was significantly boosted by recurrent neural networks introduced as long/short-term-memories (LSTM) [Hoc97] allowing incorporation of historical contextual aspects for an increased classification accuracy. A key trigger for the technological progress of Deep Learning is the development of convolutional neural networks (CNN). Instead of expert or machine driven feature selection in the machine learning domain, with convolutional networks the search for domain-specific and adequate features is now handled within the training process. The possible multi-resolution convolution pyramids and depth of 1200 layer for some CNN structures thereby significantly outperform classic convolution approaches such as Haar Cascades [Vio01] of sequentially applied weak classifiers. Another key development in Deep Learning is generative adversial networks (GAN) [Goo14] opening up totally new domains of application. GANs thereby are composed of a classification (discriminator) and a convolution network (generator) both alternatingly trained. While the generator tries mimicking the given reference samples with synthesized data, the discriminators loss indicates if the fake and real data are differentiable. As fields of application, the synthesis of handwritings, paintings, and medical data [Yi19] or general enrichment of the training data to prevent from over-fitting [Fri18] are of relevance.

1.3 Related Work

With few datasets available, i.e. around 10-20 volumes only, common Deep Learning models generally cannot be trained to highest accuracy. Applying drop-out and data augmentation strategies, the risk for over-fitting can be significantly reduced [Gao19]. Besides data augmentation, the generation of synthesized data is another strategy for enriching the reference database as successfully shown for liver segmentation [Yan17]. Nevertheless, due to the black-box nature of Deep Learning models, dealing with results below required accuracy is difficult.

A marker-based U-net model was proposed in [Sak19] that is simultaneously trained on both, the automatic classification of the target anatomical structure and to thereby obey markers optionally placed by the user for a post-processing classification run. While training a correction mechanism together with the net unmasks much of accept as is nature of Deep learning models, the problem of insufficient data for proper training still remains.

Another approach to incorporate the human expert in the computer-based diagnostics is to interpret the DL computer outcome as a “third eye” [Fou19], i.e. using deep learning recognition algorithms to solely improve visual diagnostics in medicine.

1.4 Training Deep Learning Models on Small Training Datasets

In this work the applicability of Graph cuts, a segmentation concept known since two decades, is evaluated in a totally different context, namely as post-processing for segmentations of an insufficiently trained Deep Learning model. That way, the black box nature of DL models is overcome and adequate correction of the results by experts becomes possible in a semi-automated way. The test data classifications significantly improved and validated that way by human experts in image processing are then used to enrich the data basis for model retraining. Thereby, a generic strategy for allowing expert-centered post-processing of DL models is introduced allowing to correct the seldom cases where obvious segmentation errors are introduced.

Furthermore, a strategy for synthesizing artificial medical image datasets together with highly correlating reference segmentations is evaluated. The medical fake images are mixed with real data at different mixing ratios and used for training a U-net cascade. As GANs generally require a huge amount of input data for training, a chicken-egg problem would arise. Thus, in this work it is shown that even GANs trained on a low amount of reference images can synthesize images that help to train DL models.

2. MATERIAL

As diagnostic domain in this research work, the parenchyma from CT and MRI imaging modality is chosen. The 131 CT datasets from the Medical Segmentation Decathlon [Sim19] with provided ground truth expert reference segmentations are used for training a GAN as well as for evaluations of real and synthesized data on the training process of a U-net cascade [Zwe20] with the class *liver* and *tumor* being merged. Graph cut based post-processing of classification results from a weakly trained U-net cascade are evaluated on 20 parenchyma MRI datasets in axial CAIPIRINHA non-contrast, breath-holding fat-suppressed AX CAIPI VIBE FS protocol [Mor15]. From the MRI datasets, 10 are without and 10 with Hepatocellular Liver Lesions (HCC). In case of a lesion, this class is merged with the parenchyma region leading to a binary classification problem. The MRI datasets are segmented using spline tracing (live-wire) and auto-tracing (region growing) available from Analyze software [Rob98].

2.1 Data Preparation and Pre-processing

For both, CT and MRI datasets, the same pre-processing strategies are applied. To balance the varying slice thicknesses, the z-spacing is adjusted to the x/y inter-slice spacing using cubic interpolation for the intensity dataset and binary shape interpolation [Raj03] for the reference segmentations. To limit the extent of the input slices, an area of 352×288 pixels, referring to an average axial width-to-height aspect-ratio, is extracted based on the segmentation ROI. To allow for data augmentation during the training process, a safety margin of 10px along the borders is applied, leading to original slice extent of 372×308 pixels.

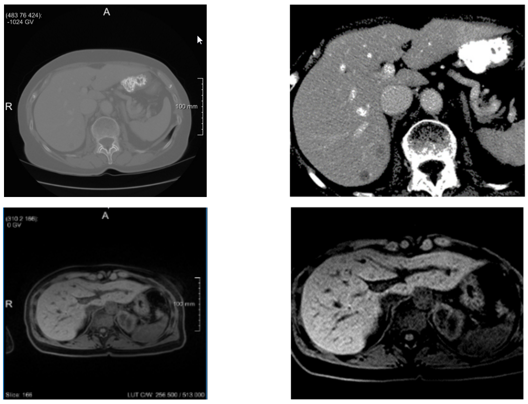

Figure 1. Slice Pre-processing for CT (top row) and MRI (bottom row).

To normalize the intensity profile w.r.t the target anatomical structure, μ_{liver} and σ_{liver} are evaluated per dataset based on the segmentation mask. An intensity transform similar to windowing is applied to shift the mean object intensity to 127.0 per dataset applying a scale factor $s = \frac{115}{3 \cdot \sigma_{liver}}$, see Eqn. 1.

$$\mathrm{T}(a_i) = \begin{cases} MAX(127 - |a_i - \mu_{liver}| \cdot s, 0) & a_i \leq \mu_{liver} \\ MIN(127 + |a_i - \mu_{liver}| \cdot s, 255) & a_i > \mu_{liver} \end{cases} \quad (1)$$

Preprocessing on slice #100 for the first CT and MRI volume respectively with size restriction to 372×308 pixel and intensity transformation is shown in Fig. 1.

3. METHODOLOGY

The Deep Learning Network to be used in this paper for validation on low or synthesized data is a U-net [Ron15] applied as cascade with combining axial, sagittal and coronal views [Zwe20], see section 3.1. Although the Deep Learning model processes the input in a 2D slice-wise manner, results of Jaccard Index JI=.9529 and Dice Coefficient DC=.9759 are achievable with slice-mini-batch optimization and training on solid 22,000 images from 100 volumes.

3.1 U-net Cascade for Liver Segmentation

The input axial images are transformed to sagittal and coronal view with the 2D slices then each classified by a separately trained U-net, see Fig. 2.

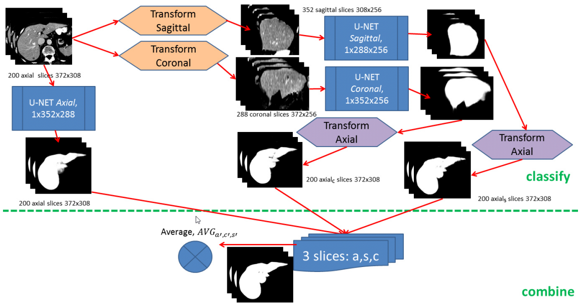

Figure 2. Slice-wise processing in axial, sagittal and coronal view with final results as average.

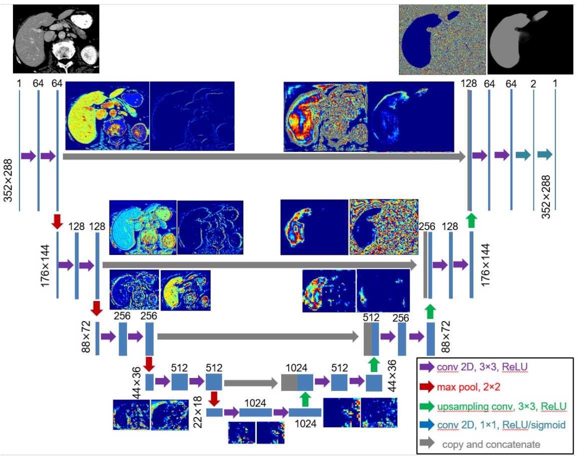

Figure 3. U-net adapted from [Ron15] with exemplary kernel output per hierarchy level.

The U-nets thereby expect an unsigned char single channel 352×288 image as input and incorporate 176×144, 88×72, 44×36 and 22×18 resolution hierarchies for the classic encoder/decoder pattern, see Fig. 3. For matter of illustration, in Fig. 1 kernel output feature images are visualized at different levels of hierarchy.

For training, an Adam Optimizer [Kin14] with learning rate $lr = 5 \cdot 10^{-6}$ using *cross-entropy* as loss is utilized to train the model at a $batchSize = 32$ and $patience = 12$ up to *200* epochs.

3.2 Data Augmentation

A general key strategy for training is application of data augmentation, thereby slightly varying the input data. With data augmentation over-fitting can be reduced not only if training on small datasets but in general. For this research work, the following geometric and intensity transformations are applied on the input data at random scale:

- *transX* and *transY*: translation in x-direction and y-direction of the current slice
- *rot*: rotation around the image center
- *intMul*: linear scale of the image intensities leading to brighter or darker pixel values
- *intAdd*: additive manipulation of the intensities within the window, leading to a uniform shift for full scalar range

For training of the U-net cascade as delineated in section 3.1 and [Zwe20], the augmentation scale is randomly varied within range $[16, 16, 10, .1, 30]$ for *transX, transY, rot, intMul* and *intAdd* respectively.

3.3 GAN for Medical Image Synthesis

A GAN architecture with Generator as convolutional neural network fed by random input and a Discriminator as binary classificatory (fake or real) is chosen.

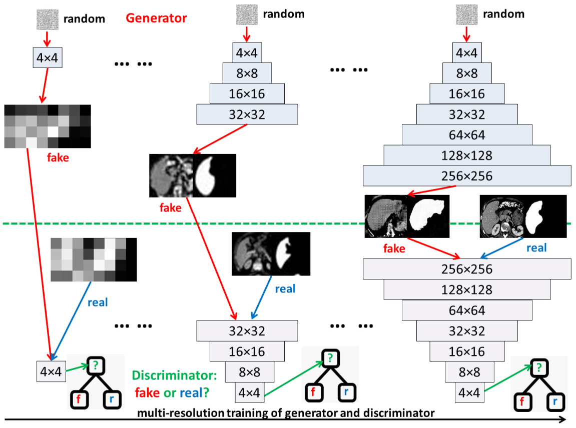

Figure 4. Adapted GAN architecture [Ker17].

The GAN is trained in a progressive way using 7 levels of resolution from 4×4 to 256×256, see Fig. 4 for illustrated levels 1 (4×4), 4 (32×32) and 7 (256×256). For training, an *AdamOptimizer* with a learning rate of 0.00015 and an iteration count *iter* increasing with the hierarchy level l as $iter = 20{,}000 + l * 32{,}000 + l^2 * 3{,}200$ is chosen, thus leading to iterations between [55.200;400.800]. As input for the Discriminator 5,000 CT slices non-overlapping with the test-set are randomly arranged in batches of size 2 with data augmentation applied. To resample the input size between 256×256 used for the GAN and 352×288 used for the U-net cascade, bilinear interpolation is applied. An important aspect is the synthesis of an accurate ground-truth reference mask besides the fake medical images. Thus, the single-channel input tensor of size 2×256×256×1 with values in [0.0;1.0] is extended to 2×256×256×2 with the associated reference mask as additional channel.

3.4 Graph cut Post-Processing

To allow for user-guided post-processing of the DL results, a fitness function for N_4 Graph cut processing combining both, original image properties and DL segmentation results is required, namely:

- *ORIG*: horizontal (H) and vertical (V) edges of the original intensity profile after applying intensity shift, c.f. Equ. 1.
- *EXP*: ORIG damped / amplified by a difference image from the expected intensity level after smoothing (median $r = 1$ and Gauss $r = 5, \sigma = 2.5$).
- *S1* and *S4*: H/V edges from the binary segmentation results from axial, sagittal, coronal and combined with 1 and 4 hits per voxel respectively.

Besides S1 and S4, the 2- and 3-hit cases are omitted due to lack in entropy. The cumulated fitness function is composed by applying Equ. 2 with function *s()* scaling to $[0; w_i]$ and the weights optimized via Evolution Strategy.

$$\max\begin{pmatrix} s(ORIG_H, w_0), s(EXP_H, w_1), \\ s(S1_H, w_2), s(S4_H, w_3) \end{pmatrix} \tag{2}$$

For Evolution Strategy optimization of the weights, recombination ($\mu/p+, \lambda$) with *epochs=100*, *batchSize=8*, *populationSize=8*, children *λ=32, mutationChance=0.4, mutationRate=0.25* dropping by 4% each epoch is applied.

To reduce the required user-interaction to a minimum, the FG and BG seeds are derived from the combined DL segmentation via skeleton calculation.

4. IMPLEMENTATION

Model training and testing is implemented in Python version 3.7.3 using Tensorflow 2.0 beta together with Keras. The Python image processing is built upon OpenCV and numpy for fast matrix operations. To provide the model with training data, a `DataGenerator` class is derived from `Sequence` base class providing data augmentation functionality.

5. RESULTS

For evaluation, the *Sørensen-Dice coefficient* (DSC) and the Jaccard index (JI) are calculated from reference mask R and result foreground region S of image I with $R \subseteq I, S \subseteq I$ and pixel $(x, y) \in R \cup S$. Additionally, the normalized surface distance (NSD) [Lap19] evaluates the FP and FN error based on the 3D distance-map based proximity to the next correct border position, see Equ. 3-5.

$$DSC(R,S) = \frac{2 \cdot |R \cap S|}{|R| + |S|} \tag{3}$$

$$NSD(R,I) = 1 - \frac{\sum_{x,y} [\![R_{x,y} \neq I_{x,y}]\!] \cdot D(R)_{x,y}}{\sum_{x,y} D(R)_{x,y}}, \quad D_{x,y}(R) = dist_{Euc}(surf(R)) \tag{4}$$

$$JI(R,S) = \frac{|R \cap S|}{|R| + |S| - |R \cap S|} \tag{5}$$

The process steps discussed in this paper, namely data preparation, pre-processing, GAN and U-net cascade model training and validation/test are performed on a *Colfax SX9600 GPU Rack* with 2×I*ntel Xeon Gold 6148 2.4GHZ* processors and 768GB of DDR4 memory with 2667MHZ clock frequency split into 24 partitions of 32GB each. The system runs *CentosOS 7.6* operating system and provides for fast tensor calculation 8 GPU cores, namely 4× *NVIDIA Volta Titan V 12G* and 4× *NVIDIA Tesla V100 32G*. For training and evaluation, parallelization was omitted and only one single Tesla core used in a sequential manner.

5.1 GAN Synthesis of Medical Images

Utilizing the GAN structure trained on 5,000 CT slices, 10,000 synthesized images are generated that meet the range criterion for cumulated discriminator and generator loss as $|D_{loss}| + |G_{loss}| \leq 160.0$, see Fig. 5 for images 10 (upper left), 20 (upper right), 30 (lower left) and 40 (lower right).

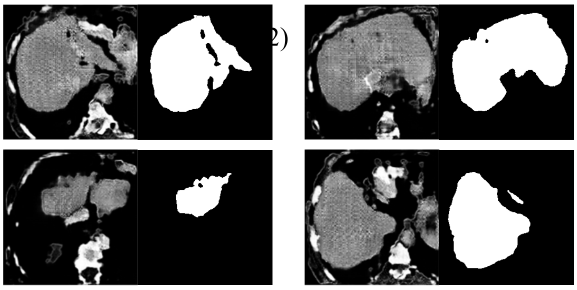

Figure 5. Synthesized Images with Reference.

The synthesized images show a slightly different intensity profile with μ_{real}=68.268 [26.52;106.70] and σ_{real}=16.420 for the 640 real and μ_{fake}=69.341 [26.827;104.022] and σ_{fake}=15.872 for the 640 fake images according to HYBRID_T in Tab. 1. The generated binary regions are of average size $\mu_{sizeRefREAL}$=3,836,278 [6,156;7,809,799] and $\mu_{sizeRefFAKE}$=3,678,637 [0;8.235,480] respectively. It seems that the loss-check for the data synthesis leads to an increase in caudal and cranial slices with smaller parenchyma cross-sections, areas that generally are weaker in axial U-nets and necessitate for additional training [Zwe20].

5.1.1 GAN Synthesis

For testing the applicability of GAN-synthesized data on training the U-net cascade, real and fake CT slices of the parenchyma are utilized for training and test as enlisted in Tab. 1. Training and validation of the U-net utilizes different real CT slices (#9000-15000) as used for the GAN training (first 5000) while then for test, the CT slices #22500-27000 are used for extraction of 640 slices. With respect to the generated fake data, the synthesized slices for U-net training and test are separated datasets too.

It is further evaluated, how different the images synthesized for FAKE_T according to Tab. 1 and the ones used for training are, i.e. that the random number generator utilized by Tensorflow does not lead to redundancy, see Fig. 6 for the first three generated images and their most similar images from the fake training database evaluating the difference images (mid) and the reference mask match (right).

purpose	dataset	#real	#fake
TRAIN	REAL_SMALL	2,200	0
	REAL	4,400	0
	FAKE	0	4,400
	HYBRID	2,200	2,200
	HYBRID_LARGE	2,200	6,600
TEST	REAL_T	640	0
	FAKE_T	0	640
	HYBRID_T	640	640

Table 1. Small train and test ensembles for validation on the axial U-net.

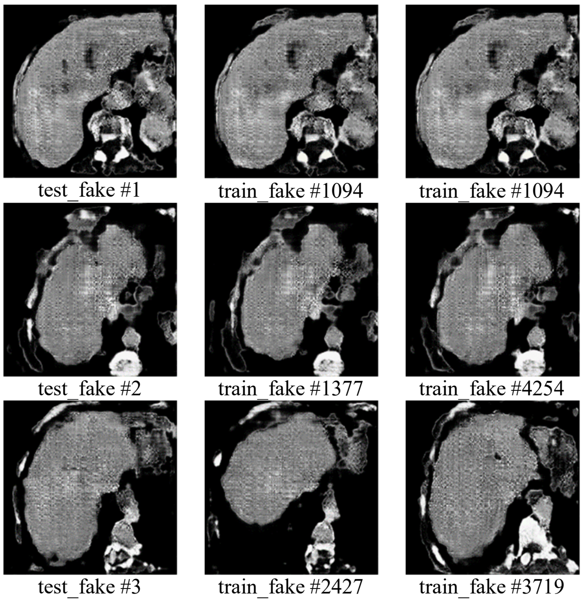

Figure 6. GAN synthesized image similarity.

As the same GAN model is utilized, depending on the random seeds the similarity at least implies, the slices might result from a same virtual volume.

Nevertheless, the accuracy of a potential model on the synthesized data is not as relevant as the difference in the real data from the synthesized samples, see Fig. 7. Here the GAN being trained on a distinctive dataset prevents from potential bias. For the real-data, separation into slices for train/validation/test that do not originate from the same volume is important. It ensures that not only the slices but the entire tomographic information is distinctive, very relevant to prevent from bias due to high z-resolution.

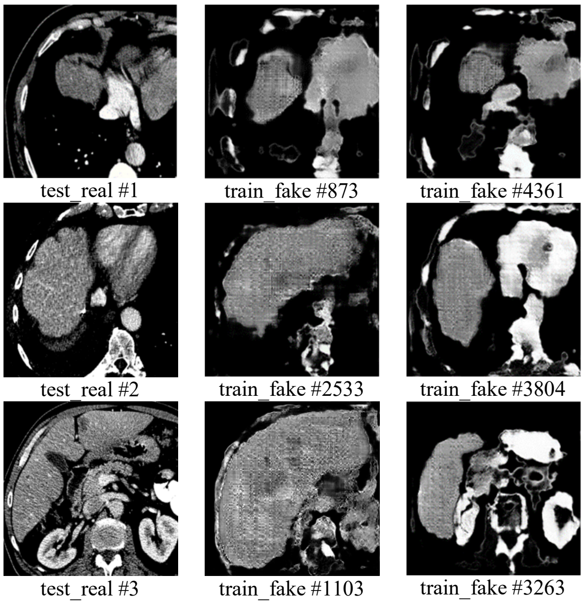

Figure 7. Similarity of real images with GAN-synthesized images in train and test.

model	dataset		
	REAL_T	FAKE_T	HYBRID_T
REAL	92.3572	89.9044	91.1111
REAL_SMALL	88.7514	87.7616	88.3140
FAKE	90.8081	95.7848	93.0345
HYBRID	92.0237	94.9955	93.4568
HYBRID_LARGE	92.9145	97.6974	95.1660

Table 2. JI on axial U-net with Synthesized Data from 5k GAN.

5.1.2 Training with synthesized data

With the proposed datasets, 5 axial U-net models are trained with the mean JI evaluated as percentage-accuracy on the 3 test datasets, see Tab. 2. With 4,400 compared to 2,200 datasets used for training, REAL significantly outperforms REAL_SMALL.

Comparing the performance on real data with the models HYBRID and HYBRID_LARGE it becomes obvious that synthesized data significantly boosts the training process. The model FAKE only trained on 4,400 synthesized data achieves solid JI=90.808 on the REAL_T test data. In Fig. 8 results of REAL_SMALL model (second row) and HYBRID_LARGE model (third row) on REAL_T test data for slices #1-3 (first row) with color-encoded FN (red) and FP (blue) are presented.

The training process for 97 epochs on the HYBRID_LARGE model is shown in Fig. 9. Due to heterogeneity of the data and applied data augmentation, the validation accuracy follows the trend of the test accuracy indicating no over-fitting. The loss is reduced from 0.7 to 0.18 within the first epoch and then drops approximately indirect proportional to the loss with a sink at epoch 46.

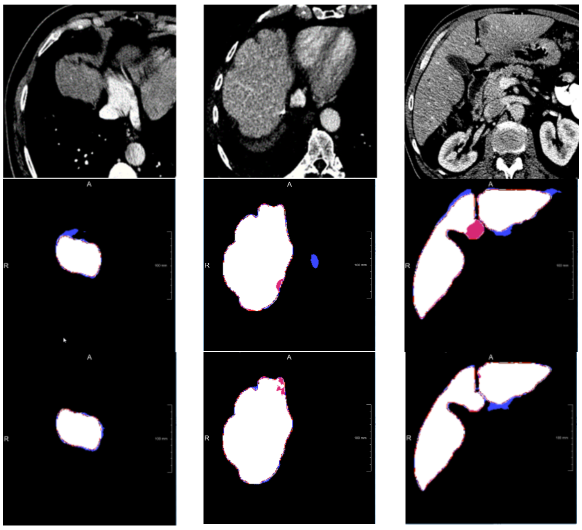

Figure 8. FN, FP and correct classification with HYBRID model on REAL_T test data.

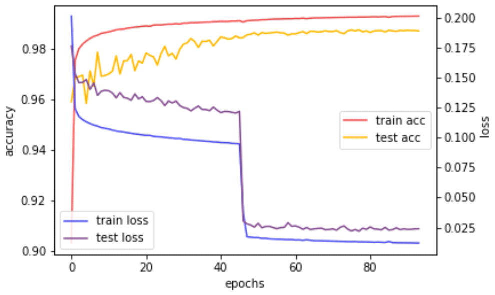

Figure 9. Loss and accuracy of train/validation per epoch when training HYBRID_LARGE.

5.2 Graph cut Post-processing

In this section, the applicability of Graph cut as post-processing tool for DL classifications as well the influence of enriched datasets on the training process is evaluated.

5.2.1 Training U-net cascade on MRI data

Training on 13 test datasets (#2-14, 7 without and 6 with HPC; 3,132 slices) gives solid results for the axial, sagittal and coronal models with significant boots when combining with 2/3 majority voting ($AVG_{a,c,s}$) for tests on 420 slices of datasets #1 and #14 (1 with and 1 without HPC), see Tab. 3.

model	DSC	JI	NSD
ax	92.1154	85.3834	92.3271
ax_s	91.9219	85.0513	88.6966
ax_c	92.5392	86.1144	93.2273
$AVG_{a,c,s}$	94.4096	89.4111	97.0472

Table 3. U-net cascade trained with 13 datasets in axial, sagittal, coronal and combined $AVG_{a,c,s}$.

5.2.2 Graph cut for post-processing

Utilizing evolution strategy, after 100 epochs evaluating on the CT parenchyma data with automatic application of the Graph cut post-processing, the fitness-function weights get optimized to w_0=0.287, w_1=0.217, w_2=0.419, w_3=0.641, c.f. Fig. 10.

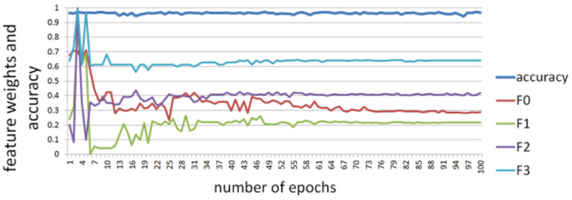

Figure 10. Evolution Strategy-based optimization of the Graph cut fitness function.

With only the weight ratio relevant, w_3 for the cumulated DL result is of highest importance, while the original image aspects (w_0, w_1) still allow for adaption to the image profile, see Fig. 11 for original images, combined DL segmentation and the horizontal edges derived from the fitness function for slices 30, 100 and 190 of MRI dataset #1.

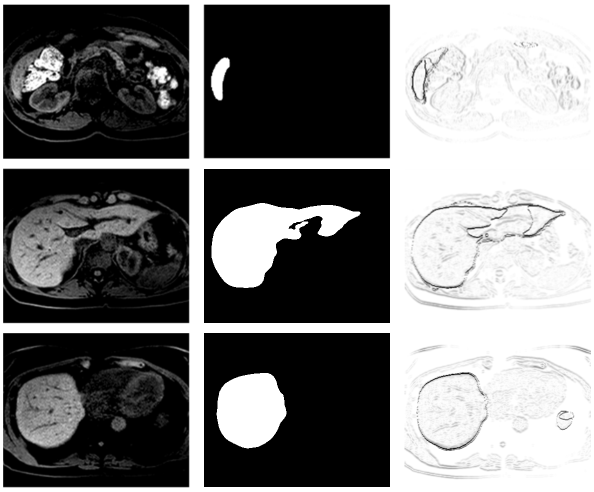

Figure 11. Horizontal GC fitness image.

With the FG and BG seeds automatically derived from the combined segmentations, the required amount of user interaction is kept to a minimum; see Fig. 12 for skeleton, GC result and mismatch for slices 30, 100 and 190 of MRI dataset #1.

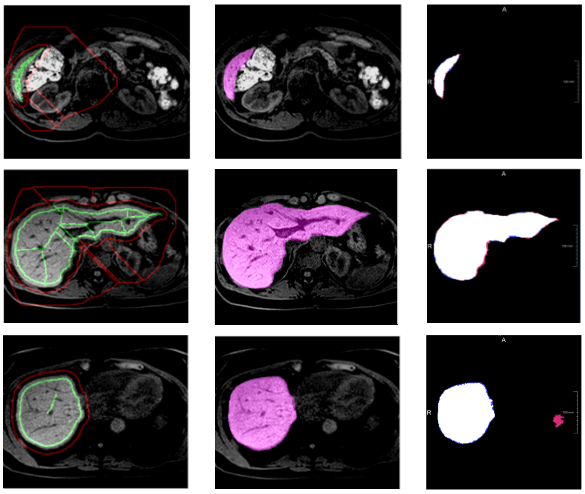

Figure 12. GC segmentation based on skeletons.

To allow for tests with enriched datasets, the remaining 5 MRI datasets (2 without and 3 with HPC; 1,238 slices in total) are segmented by a medical expert utilizing Live-wire (LW) to provide rough reference segmentations. From these 5 datasets, two (484 slices) are classified by the U-net cascade ($AVG_{a,c,s}$) as described in sections 3.1 and 5.2.1. The two datasets thereby only achieve JI=82.1644, DSC=90.2091 and NSD=91.5809. After the GC post-processing, the accuracy is increased to JI=88.8186, DSC=94.0782 and NSD=98.1452 with the semi-automated processing lasting for 152min (18.8sec per slice on average). To evaluate robustness of the Graph cut post-processing step, three experts evaluate the same subset of n=30 randomly selected slices. Although the FG skeletons placed by the experts in a manual way vary, the segmentation outcome is very stable, see Fig. 13.

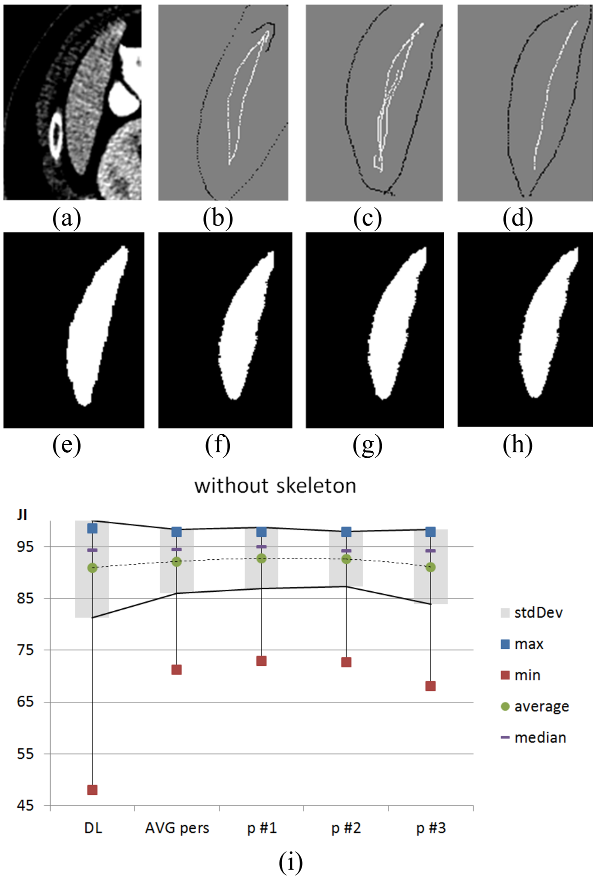

Figure 13. Slice #28 robustly segmented by three medical experts utilizing GC with different FG and BG skeletons.

Slice #28 without skeleton support (a) and expected ground truth (e) shows suboptimal DL result JI=.877 and can be improved by all test persons (f-h) in range [.911;.920] even with very different GC skeleton interpretations (b-d). Utilizing GC for manual DL post-processing, potential invalid DL results get improved by test persons p#1-#3 and variability of the JI accuracy is generally decreased, see Fig. 13 (i).

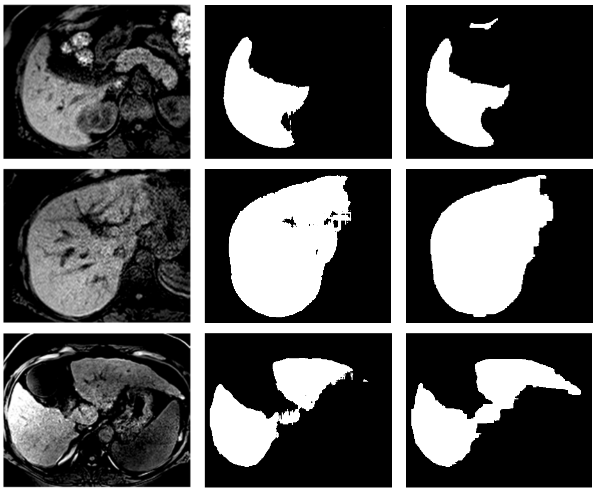

Figure 14. MRI add-on data with original, result from $AVG_{a,c,s}$ and result from Graph cut.

As a rough Live-wire segmentation with immanent inaccuracies is used as base for comparison, the accuracies before and after GC correction allow for relative comparison only, see Fig. 14 comparing original slice, initial result from $AVG_{a,c,s}$ and after GC correction for add-on slices #100, 150 and 1100 respectively.

5.2.3 Re-training the U-net on enriched data

The training of the 4 models for the U-net cascade, as described in section 5.2.1, is re-run with enriched data. The initial 13 datasets thereby get enriched by 484 slices (1 dataset with and 1 without HPC) previously corrected by user-guided Graph cut post-processing, see Tab. 4. Furthermore, all 5 add-on datasets (1,238 slices) are used to directly train the U-net cascade based on the rough Live-wire segmentation provided, see Tab. 5.

model	DSC	JI	NSD
ax	93.2937	87.4303	96.0364
ax_s	92.1216	85.3940	85.8697
ax_c	92.9323	87.1230	93.4942
$AVG_{a,c,s}$	94.6671	89.8742	97.4724

Table 4. Results on Models trained with 13 datasets enriched by 2GC datasets

model	DSC	JI	NSD
ax	92.6168	86.2488	94.3139
ax_s	92.1746	85.4850	87.9913
ax_c	92.1341	85.4155	91.9131
$AVG_{a,c,s}$	94.5163	89.6028	97.0967

Table 5. Results on Models trained with 13 datasets enriched by 5 LW datasets.

A comparison of the achievable test accuracy for the models trained on the 13, the 13+2GC datasets and the 13+5 LW datasets is provided in Fig. 15.

A visual representation of the tomographic segmentation achievable by the $AVG_{a,c,s}$ model trained on 13 + 2GC is given in Fig. 16 with FP (blue) and FN (red) visualized for test dataset #1 with errors in cranial and caudal direction and high accuracy in the mid slices.

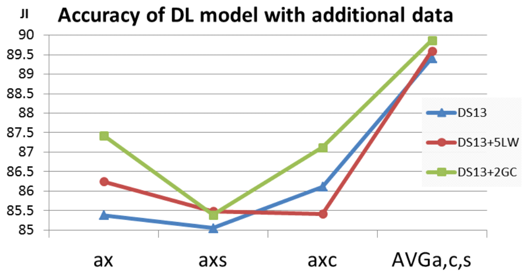

Figure 15. Achievable JI accuracy for DS13, 2GC add-on and 5LW add-on, c.f. Fig.12-14.

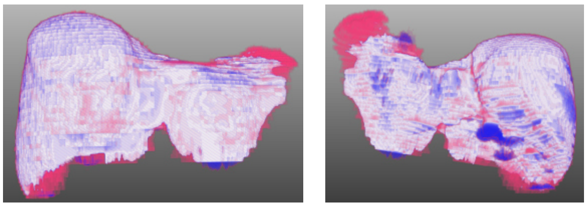

Figure 16. 3D segmentation of test dataset #1.

6. DISCUSSION AND CONCLUSION

Although the texture of the GANs at increased mask size does not look natural for a human inspector, yet the synthesized data is highly applicable for training of DL Models with an abstraction from pure visual input to implicit features anyway. The results in Tab. 2 indicate that enriching the data with synthesized data boosts the training process similar to data augmentation. In case of lack of real-world data, the gap can be closed. As models REAL (4,400 real) and HYBRID_LARGE (2,200 real, 6,600 fake) perform at comparable accuracy, the impact of synthetic data seems of course not equal to enriching with real data but leading to an improvement with increasing number as compared to models REAL_SMALL and HYBRID. As synthesized data can be generated in arbitrary numbers at very high variability it is a good option to compensate for lack of real-world testing data. It has to be stated, that training the GAN itself at a high accuracy necessitates for a sufficient amount of real data too.

For the GAN used in this paper, 5,000 input slices from 24 CT volumes were used. Thus, for application domains with a very low number of samples, a kind of chicken-egg-problem arises as the GAN needs to be trained first in a sufficient way. Here the application of data augmentation as applied for the U-net training helps to significantly reduce the demand for training data. Furthermore, a reasonable batch size, e.g. 16 or 32, which was not possible due to resource limits of the DL HW, will help to reduce the data demand for GAN training. The presented encoding of intensity profile and binary reference segmentation as a 2-channel tensor allows for generation of medical data together with very precise ground-truth. To reduce the memory demand by a factor of two, both the intensity and the reference channel could be encoded in a single channel.

The applicability of GC as post-processing tool is given due to fitness function weights optimized with Evolution strategy to balance between original image intensity and segmentation edges. Using two additional input data first weakly classified by the U-net cascade at only JI=82.1644 and then post-processed by GC allows to iteratively enrich the training dataset. As the reference segmentations generally originate from different tools and experts, this heterogeneity affects the training process too. It is shown, that a small set of segmentations prepared with region growing, Graph cut and Live-wire trains well with the U-net cascade. As shown in Fig. 16, adding data of adequate quality can help to improve the training and test quality.

Thus, if in specific diagnostic domains or for particular studies the initially available amount of test data is insufficient to train DL models at highest accuracy, an incremental approach for data revision and model re-training was presented. In future, training of GANs with less input slices and the combination with GC will be evaluated.

7. ACKNOWLEDGMENTS

Thanks to the BIR research team at Mayo Clinic, Rochester, MN for valuable discussion, support and the provided GPU infrastructure.

8. REFERENCES

[Agg11] Aggarwal, A., Vig, R., Bhadoria, S., and Dethe, C.G., Role of Segmentation in Medical Imaging: A Comparative Study. Int. Journal of Comp. Applic. 29(1), 2011.

[Chr18] Christensen, A., and Wake, N. Wohler Report: Medical image processing software. http://www.wohlersassociates.com, 2018.

[Coo92] Cootes, T.F., Taylor, C.J., Cooper, D.H., and Graham, J., 1992. Training Models of Shape from Sets of Examples. Proc. of the British Machine Vision Conference, 1992.

[Coo98] Cootes, T.F., Edwards, G.J., and Taylor, C.J., 1998. Active Appearance Models. Proc. of the 5th Europ. Conf. on Computer Vision, 1998.

[Fou19] Fourcade, A., and Khonsari, R.H. Deep learning in medical image analysis: A third eye for doctors. J Stomatol Oral Maxillofac Surg 120 (2019) 279–288, 2019.

[Fri18] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 2018.

[Gao19] Gao, H., Pei, J., and Huang, H. Demystifying Dropout. Proc. of the 36th Inter. Conf. on Machine Learning, PMLR, 2019.

[Goo14] Goodfellow , I.J., Pouget Abadie , J., Mirza, M., Xu, B., Warde Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. Proc. of the 27th Int. Conf. on Neural Information Proc. Sys., vol. 2, 2014.

[Hoc97] Hochreiter, S., and Schmidhuber, J. Long Short-Term Memory. Neural Comp. 9(8), 1997.

[Kin14] Kingma, D.P., and Ba, J.L. Adam: A method for stochastic optimization. Int. Conf. on Learning Representations (ICLR), 2014.

[Koh95] Kohonen, T. Self-Organizing Maps. 3rd ed. Springer, 1997.

[Lap19] Laplante, P.A. (ed.). Encyclopedia of Image Processing. CRC Press/Taylor & Francis Publishing, 2019.

[McI96] McInerney, T., and Terzopoulos, D. Deformable Models in Medical Image Analysis : A Survey. Medical Image Analysis 1(2), 1996.

[Mor15] Morani, A.C., Vicens, R.A., Wei, W., Gupta, S., Vikram, R., Balachandran, A., Reed, B.J., Ma, J., Qayyum, A., and Szklaruk, J. T1-weighted (T1W) gradient recall echo volumetric interpolated breath-hold examination. J Comput Assist Tomogr. 39(2), 2015.

[Raj03] Rajagopalan, S., Karwoski, R. A., Robb, R. A., Ellis, R. E., and Peters, T. M. Shape-Based Interpolation of Porous and Tortuous Binary Objects. MICCAI 2003, 2003.

[Rob98] Robb, R.A., Hanson, D.P., Karwoski, R.A., Larson, A.G., Workman, E.L. and Stacy, M.C., 1989. Analyze: a comprehensive, operator-interactive software package for multidimensional medical image display and analysis. Comput Med Imaging Graph 13(6), 1998.

[Ron15] Ronneberg, O., Fischer, P., and Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI, 2015.

[Sak19] Sakinis, T., Milletari, F., Roth, H., Korfiatis, P., Kostandy, P., Philbrick, K., Akkus, Z., Xu, Z., Xu, D., and Erickson, B.J. Interactive segmentation of medical images through fully convolutional neural networks. MICCAI, 2019.

[Set99] Sethian, J.A. Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science. Cambridge University Press, 1999.

[Sim19] Simpson, A., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Ginneken, B., Kopp-Schneider, A., Landman, B., Litjens, G., Menze, B., Ronneberger, O., Summers, R., Bilic, P., Christ, P., Do, R., Gollub, M., Golia-Pernicka, J., Heckers, S., Jarnagin, W. and Cardoso, M.J., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. CoRR, 2019.

[Son13] Sonka, M., Hlavac, V., and Boyle, R. Image Processing, Analysis, and Machine Vision. CENAGE Learning, 4th ed., 2014.

[Squ18] Squelch, A. 3D printing and medical imaging. Journal Med Radiat Sci. 65(3), 2018.

[Str15] Strakos, P., Jaros, M., Karasek, T., Kozubek, T., Vavra, P., and Jonszta, T. Review of the Software Used for 3D Volumetric Reconstruction of the Liver. Int. Journal of Computer and Information Engineering 9(2), 2015.

[Van98] Van Biesen, W., Sieben, G., Lameire, N., and Vanholder, R. Application of Kohonen neural networks for the non-morphological distinction between glomerular and tubular renal disease. Nephrol Dial Transplant 13(1), 1998.

[Vio01] Viola, P., and Jones, M. Rapid Object Detection using a Boosted Cascade of Simple Features. Conf. on Computer Vision and Pattern Recognition ‘01, 2001.

[Yan17] Yang, D., Xu, D., Zhou, S.K., Georgescu, B., Chen, M., Grbic, S., Metaxas, D., and Comaniciu, D. Automatic Liver Segmentation Using an Adversarial Image-to-Image Network. MICCAI, 2017.

[Yi19] Yi, X., Walia, E., and Babyn, P. Generative Adversarial Network in Medical Imaging: A Review. Medical Image Analysis vol. 58, 2019.

[Zha11] Zhang, J., and Wang, X.W. The application of feed forward neural network for the X ray image fusion. J. Phys., 2011.

[Zwe13] Zwettler, G., and Backfrieder, W. Generic Model-Based Application of Modular Image Processing Chains for Medical 3D Data Analysis in Clinical Research and Radiographer Training. Proc. of IWISH, 2013.

[Zwe20] Zwettler, G.A., Backfrieder, W., and Holmes III, D.R. Pre- and Post-processing Strategies for Generic Slice-wise Segmentation of Tomographic 3D datasets Utilizing U-Net Deep Learning Models Trained for Specific Diagnostic Domains. Proc. of VISAPP, 2020.