User-Guided Domain Adaptation for Rapid Annotation from User Interactions: A Study on Pathological Liver Segmentation

Ashwin Raju[1,2], Zhanghexuan Ji[1,3], Chi Tung Cheng[4], Jinzheng Cai[1], Junzhou Huang[2], Jing Xiao[5], Le Lu[1], ChienHung Liao[4], and Adam P. Harrison[1]

[1] PAII Inc., Bethesda MD, USA
[2] The University of Texas at Arlington, Arlington TX, USA
[3] University at Buffalo, Buffalo NY, USA
[4] Chang Gung Memorial Hospital, Linkou, Taiwan, ROC
[5] PingAn Technology, Shenzhen, China

Abstract. Mask-based annotation of medical images, especially for 3D data, is a bottleneck in developing reliable machine learning models. Using minimal-labor user interactions (UIs) to guide the annotation is promising, but challenges remain on best harmonizing the mask prediction with the UIs. To address this, we propose the user-guided domain adaptation (UGDA) framework, which uses prediction-based adversarial domain adaptation (PADA) to model the combined distribution of UIs and mask predictions. The UIs are then used as anchors to guide and align the mask prediction. Importantly, UGDA can both learn from unlabelled data and also model the high-level semantic meaning behind different UIs. We test UGDA on annotating pathological livers using a clinically comprehensive dataset of 927 patient studies. Using only extreme-point UIs, we achieve a mean (worst-case) performance of 96.1% (94.9%), compared to 93.0% (87.0%) for deep extreme points (DEXTR). Furthermore, we also show UGDA can retain this state-of-the-art performance even when only seeing a fraction of available UIs, demonstrating an ability for robust and reliable UI-guided segmentation with extremely minimal labor demands.

Keywords: Liver Segmentation · Interactive Segmentation · User-guided Domain Adaptation

1 Introduction

Reliable computer-assisted segmentation of anatomical structures from medical images can allow for quantitative biomarkers for disease diagnosis, prognosis, and progression. Given the extreme labor to fully annotate data, especially for 3D volumes, a considerable body of work focuses on weakly-supervised segmentation solutions [24]. Solutions that can leverage user interactions (UIs), *e.g.*, extreme-points, scribbles, and boundary marks, are an important such category.

The main challenge is effectively leveraging UIs to constrain or guide the mask generation. Classic approaches, like the random walker (RW) algorithm [5], do

so via propagating seed regions using intensity similarities. Later approaches add additional constraints, *e.g.*, based on presegmentations [6] or learned probabilities [7]. With the advent of deep-learning, harmonizing mask predictions with the UIs continues to be a challenge. Deep extreme points (DEXTR) [17], which requires the user to click on the extreme boundary points of an object, is a popular and effective approach. But DEXTR only adds the extreme point annotations as an additional channel when training the segmentor, meaning the predicted mask may not agree with the UIs. Later work uses expectation-maximization strategies that alternate between network training and then regularization via RW [22] or dense conditional random fields (CRFs) [20,27,2]. However, over and above their computational demands, the intensity-based RW or CRF regularization may not capture high-level semantics and any guidance on the mask predictions still remains highly indirect. DeepIGeoS [28] offers an alternative that uses a deep-CRF optimizer to propagate scribbles. However, DeepIGeos would only allow boundary annotations or extreme points to be treated as simple seed regions, neglecting their rich semantic meaning.

To address these issues, we propose user-guided domain adaptation (UGDA). Our new method uses prediction-based adversarial domain adaptation (PADA) [26] to guide mask predictions by the UIs. UGDA's advantage is that it is equipped to model the high-level meaning behind different types of UIs and how they should impact the ultimate mask prediction. Importantly, the UIs are used as anchors when adapting the mask. Another advantage is that, like PADA, UGDA can learn from and exploit completely unlabelled data, in addition to those accompanied by UIs. Without loss of generality, we focus on using DEXTR-style extreme point UIs because of their intuitiveness and effectiveness [19,22]. But other types, *e.g.*, boundary corrections, are equally possible in addition to, or instead of, extreme points. The only constraint is that we assume a fully-supervised dataset is available in order to model the interplay between mask and UIs. But such data can originate from sources other than the target dataset, *e.g.*, from public data. To the best of our knowledge, we are the first to use domain adaptation as a mechanism to drive UI-based segmentation.

We test UGDA on an extremely challenging pathological liver segmentation dataset collected directly from the picture archiving and communication system (PACS) of Chang Gung Memorial hospital (CGMH). This dataset comprises 927 patient studies, all with hepatocellular carcinoma (HCC), intrahepatic cholangiocellular carcinoma (ICC), metastasized, or benign lesions and is *one of the most challenging datasets to date for pathological liver segmentation*. Liver volumetry analysis is a crucial pre-requisite prior to many hepatic procedures, such as liver transplantation [18,25] or resection [15,14]. Despite success with fully-automated solutions, modern deep solutions [29,12,11] are often trained only on public datasets, *e.g.*, LiTS [1], which only represents HCC and metastasized lesions and does not fully represent patient/co-morbidity distributions. Thus, they may not generalize to all datasets, such as ours. Using *only extreme-point UIs*, we achieve state-of-the-art mean (worst-case) Dice-Sørensen coefficient (DSC) scores of 96.1% (94.5%) on our dataset, compared to 93.0% (79.0%) and 93.1%

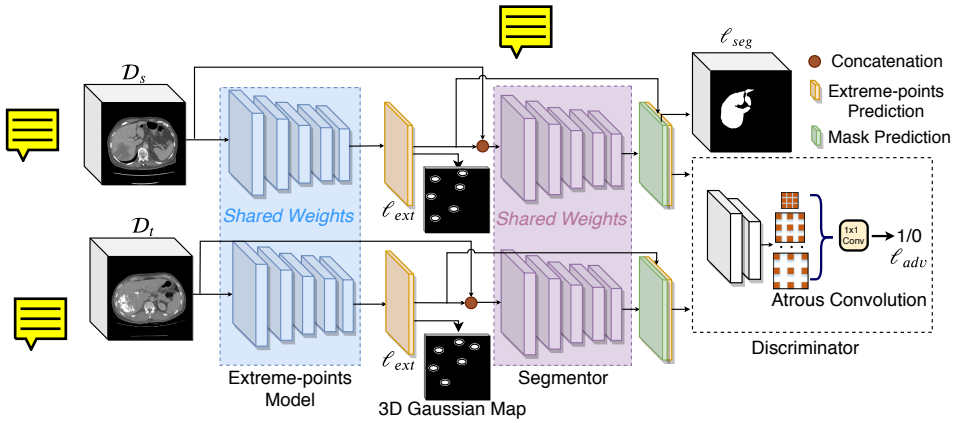

Fig. 1: **Overview.** UGDA chains together (1) an initial FCN that predicts the object's extreme points, and (2) a second FCN that accepts the first's predictions to predict a mask. For source data, $\mathcal{D}_s$, where the mask label is present, we compute fully-supervised loss on both the extreme-point and mask predictions. For target data, $\mathcal{D}_t$, we compute a fully-supervised loss when UIs are available. For all $\mathcal{D}_t$ volumes, whether UI-labelled or completely unlabelled, we use PADA to guide the mask predictions based on the extreme-point anchors.

(87.0%) for a strong a fully-supervised baseline and DEXTR [17], respectively. We also show that UGDA can improve over PADA [26] by 1.3% DSC and that UGDA can even perform robustly when only shown incomplete sets of UIs. Finally, we demonstrate predicted masks align extraordinarily well with UIs, allowing the user to interact with confidence and with minimal frustration.

2 Methods

We aim to produce reliable mask predictions on a target dataset or deployment scenario, given only minimal UIs. More formally, we assume we are given a dataset composed of both UI-labelled and completely unlabelled volumes, $\mathcal{D}_t = \{X_i, E_i\}_{i=1}^{N_w} \bigcup \{X_i\}_{i=1}^{N_u}$, with X_i and E_i denoting the images and extreme points UIs, respectively. In addition, we also assume a fully-supervised source dataset with masks is also available, $\mathcal{D}_s = \{\mathcal{X}_i, Y_i\}_{i=1}^{N_s}$. As long as the masks and extreme points describe the same anatomical structure, $\mathcal{D}_s$ may originate from entirely different sources, *e.g.*, public data. The goal is to use only the extreme point UIs to robustly annotate $\mathcal{D}_t$. Fig. 1 outlines our workflow, which uses user-guided domain adaptation (UGDA) to efficiently and effectively exploit the extreme-point UIs.

2.1 Supervised Workflow

The backbone of UGDA is two 3D FCNs chained together, where the first FCN predicts extreme-points while the second predicts a full mask. Working back-

ward, the second FCN acts very similarly to DEXTR [17], where the latter predicts a mask given an input image along with the extreme-point UIs:

$$\hat{Y} = s(X, E), \tag{1}$$

where we have used $s(.)$ to represent the segmentation FCN. Each of the 6 extreme points are represented by a 3D Gaussian heat map centered on the user clicks and rendered into an additional input channel, E_i. We find the method is not sensitive to the size of the Gaussian heat maps, and we use a kernel with 5-pixel standard deviation. However, unlike DEXTR, we do not assume all training volumes come with extreme-points, as only some target volumes in $\mathcal{D}_t$ may have UIs.

To loosen these restrictions, we use the first FCN to predict extreme-point heatmaps for each volume. Following heatmap regression conventions [30], the extreme-point FCN, $h(.)$, outputs 6 3D Gaussian heatmaps, each corresponding to one extreme point. These are then summed together into one channel prior to being inputted into our segmentation FCN:

$$\hat{Y} = s(X, \hat{E}), \tag{2}$$

$$\hat{E} = h(X), \tag{3}$$

where for convenience we have skipped showing the summation of 6 heatmaps into one channel. By relying on predictions, this allows our system to operate even with unlabelled data. Ground-truth extreme-points can be generated from any UIs performed on $\mathcal{D}_t$. Extreme-points for $\mathcal{D}_s$ can be deterministically generated from the full masks, simulating the UIs.

In terms of loss, if we use $\mathcal{D}_e$ to denote any input volume associated with extreme-point UIs, whether from $\mathcal{D}_s$ or $\mathcal{D}_t$, then a supervised loss can be formulated:

$$\mathcal{L}_{sup} = \mathcal{L}_{seg} + \mathcal{L}_{ext}, \tag{4}$$

$$\mathcal{L}_{ext} = \frac{1}{N_e} \sum_{X,E \in \mathcal{D}_e} \ell_{ext} \left(h(X), E \right), \tag{5}$$

$$\mathcal{L}_{seg} = \frac{1}{N_s} \sum_{X,Y \in \mathcal{D}_s} \ell_{seg} \left(s(X, h(X)), Y \right), \tag{6}$$

where N_e denotes the cardinality of $\mathcal{D}_e$. Following heat-map regression practices [30], we implement ℓ_{ext} using mean-squared error. For ℓ_{seg} we use a summation of cross entropy and DSC losses, which has experienced success in segmentation tasks [11]. While we focus on extreme points for this work, other types of UIs, such as boundary corrections, can be readily incorporated in this framework.

2.2 User-Guided Domain Adaptation

Similar to DEXTR [17], (6) indirectly guides mask predictions by using extreme-point heat maps as an additional input channel for the segmentor. Additionally,

for UI-labelled volumes in $\mathcal{D}_t$, the supervised loss in (5) encourages the extreme-point predictions to actually match the UIs. However, the mask prediction may contradict the UIs because there is no penalty for disagreement between the two. Thus, an additional mechanism is needed to align the mask with UIs. We opt for an adversarial domain adaption approach to penalize discordant mask predictions. While image translation-based adversarial domain adaptation methods show excellent results [10,13], these are unsuited to our task because we are concerned with adapting the *prediction-space* to produce a mask well-aligned with the UIs. Thus, we use prediction-based adversarial domain adaptation (PADA) [26].

More specifically, we use a discriminator, $d(.)$, to learn the distribution and interplay between masks and extreme-points. Treating samples from $\mathcal{D}_s$ as the "correct" distribution, the discriminator loss can be expressed as

$$\mathcal{L}_d = \frac{1}{N_s} \sum_{\mathcal{D}_s} \ell_{bce}(d(\{\hat{Y}, \hat{E}\}, \mathbf{1})) + \frac{1}{N_t} \sum_{\mathcal{D}_t} \ell_{bce}(d(\{\hat{Y}, \hat{E}\}, \mathbf{0})), \tag{7}$$

where ℓ_{bce} denotes the cross-entropy loss. Importantly, to model their combined distribution, the discriminator accepts both the UI and mask predictions. Following standard adversarial training, here gradients only flow through the discriminator. UGDA then attempts to fool the discriminator by predicting extreme-point/mask pairs for $\mathcal{D}_t$ that match $\mathcal{D}_s$'s distribution. More formally, an adversarial loss is set up for volumes in $\mathcal{D}_t$:

$$\mathcal{L}_{adv} = \frac{1}{N_t} \sum_{\mathcal{D}_t} \ell_{bce}(d(\{\hat{Y}, \hat{E}\}, \mathbf{1})). \tag{8}$$

Note that compared to (7), the "label" for $\mathcal{D}_t$ has been switched from $\mathbf{0}$ to $\mathbf{1}$. Like standard PADA setups, gradients do not flow through the discriminator weights in (8). Importantly, *gradients also do not flow through the extreme-point predictions when the UIs are present*. Consequently, when UIs are available, extreme-point predictions are only influenced by the supervised loss in (5) to match the UIs. While there is no strict guarantee, our results demonstrate almost perfect matching. Thus, the extreme-point predictions act as anchors, while the adversarial loss in (8) guides the mask predictions to properly align with the UIs. This alignment is more than simply making mask predictions agree with the UIs, as by modelling their interplay, PADA also guides mask regions far away from UIs. Finally, the use of PADA provides another important benefit, as completely unlabelled volumes in $\mathcal{D}_t$ can seamlessly contribute to the learning process in (7) and (8). In fact, UGDA can be seen as integrating domain adaptation learning processes [26] in addition to DEXTR-style guidance [17] from UIs. Thus, the overall training objective for UGDA is to minimize the following total loss:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_{adv} \mathcal{L}_{adv}, \tag{9}$$

where we have purposely kept loss weighting to only the adversarial component to reduce hyper-parameter tuning.

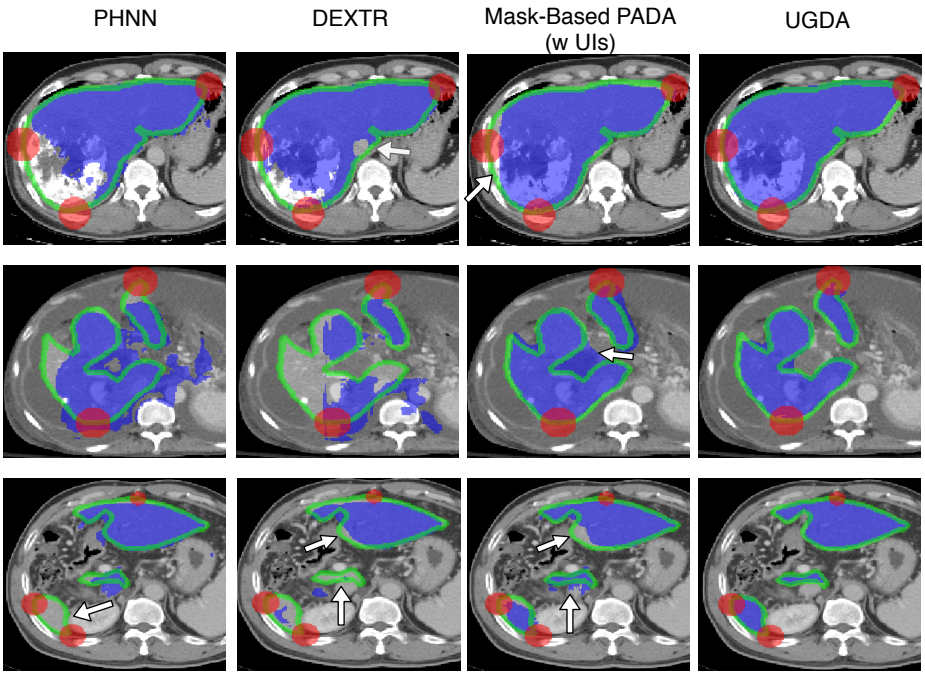

Fig. 2: **Qualitative results.** Liver mask ground truth and predictions are rendered in green contour and blue mask, respectively, and Guassian heatmaps centered on the extreme point UIs are shown in red. As can be seen, UGDA can much better align the masks with the UIs. Arrows highlight selected baseline prediction errors that UGDA corrects. (Best viewed in color)

3 Experiments

Dataset description. We test UGDA on segmenting pathological livers, using a target dataset of 927 venous-phase computed tomography (CT) studies from the PACS of CGMH. The only selection criterion was of patients with biospied or resecuted liver lesions, with CT scans taken within one month before the procedure. Patients directly reflect clinical distributions and represent ICC, HCC, benign or metastasized lesions, along with co-occuring maladies, such as liver fibrosis, splenomegaly, or embolized lesions. This serves as $\mathcal{D}_t$. From these, we selected 47 and 100 studies as validation and test sets, respectively, and delineated the patient livers. These 147 CTs are called *evaluation volumes*. We annotated the remainder using only extreme point UIs. For $\mathcal{D}_s$ we collected 235 fully-labelled venous-phase CT studies from public datasets [1,4,9,3], which, unlike $\mathcal{D}_t$, comprises both healthy and pathological livers and only represents HCC and metastasized tumors. Corresponding extreme-point "UIs" were generated from the full masks following [17]. For internal validation, we also split $\mathcal{D}_s$ into 70%, 20%, and 10% for training, testing and validation, respectively.

Implementation Details. For UGDA's two FCN architectures, we use a 3D version of the deeply-supervised progressive holistically nested network (PHNN) [8], which provides an efficient and decoder-free pipeline. As backbone, we use a 3D generalization of VGG-16 [23]. We first train a fully-supervised baseline on $\mathcal{D}_s$ using (4), then we finetune after convergence using (9). This dual-PHNN baseline is very strong on our public data, achieving a DSC score of 96.9% on the $\mathcal{D}_s$ test set. For discriminator, we use a 3D version of a popular architecture [16] using atrous convolution, which has proved a useful discriminator for liver masks [21]. We do not use the multi-level discriminator variant proposed by Tsai *et al.* [26], as its added complexity does not seem to result in noticeable improvements for liver-based PADA [21]. Specific details on the choice of hyper parameters are listed in the supplementary material.

Evaluation protocols. We evaluate UGDA on how well it can annotate $\mathcal{D}_t$ using only extreme-point UIs. To do this, we include the evaluation volumes and their extreme point UIs within the training procedure, *but hide their masks*. For evaluation, we measure DSC scores and also the mask-extreme-point agreement (MXA). The latter measures the average distance between all six of a predicted *mask's* extreme-points vs. the ground-truth extreme points. This directly measures how well the model can produce a mask prediction that actually matches the extreme-point UIs.

Comparisons. We test against the user-interactive DEXTR [17] using the same PHNN backbone. DEXTR can only be trained on $\mathcal{D}_s$ because it requires fully-labelled training data, but during inference it sees $\mathcal{D}_t$'s extreme-point UIs. The DEXTR authors claim their approach can generalize well to new data. We also test against two mask-based PADA [26] variants. The first variant, called "mask-based PADA (no UIs)", matches published practices [26] and just uses a single network (PHNN in our case) to directly predict masks, using a discriminator to penalize ill-behaving mask predictions on unlabeled data. It does not incorporate UIs. The second variant, simply called "mask-based PADA (w UIs)", is almost identical to UGDA, meaning the mask and extreme-point predictions are still trained with the supervised losses of (6) and (5), respectively, but the discriminator in (7) and (8) only models mask predictions without seeing the extreme-point predictions. Comparing this variant to UGDA reveals the impact of using extreme-points as anchors to guide the domain adaptation of the masks.

Finally, we also compare against UGDA variants trained with only a fraction of the UIs available in $\mathcal{D}_t$, with the remainder being left unlabelled. This reveals how well UGDA can operate in scenarios when only a fraction of target volumes have UI-labels. When doing so, we remove the same percentage of UI-labels from both non-evaluation and evaluation volumes.

Results. Tab. 2 outlines the performance of all variants in annotating $\mathcal{D}_t$. As can be seen, compared to its performance on $\mathcal{D}_s$, the fully-supervised dual PHNN's performance drops from 96.9% to 93.0% due to the major differences between public liver datasets and our PACS-based clinical target dataset. This suggests that other strategies are required, *e.g.*, exploiting minimal-labor UIs.

Table 1: DSC and MXA mean and standard deviation scores. In parentheses are the fraction of UI-labelled $\mathcal{D}_t$ volumes used for training (or inference for DEXTR).

Model	**% UIs**	**Mean DSC**	**Mean MXA (mm)**
Dual PHNN [8]	n/a	93.0 ± 3.2	4.3 ± 1.2
DEXTR [17]	100%	93.1 ± 2.4	3.9 ± 1.2
Mask-based PADA (no UIs) [26]	0%	94.8 ± 1.8	3.4 ± 1.6
Mask-based PADA (w UIs) [26]	100%	95.5 ± 1.0	2.5 ± 1.0
UGDA	25%	95.8 ± 0.8	1.7 ± 0.8
UGDA	50%	96.0 ± 0.9	1.4 ± 0.9
UGDA	100%	$\mathbf{96.1 \pm 0.8}$	$\mathbf{1.1 \pm 0.9}$

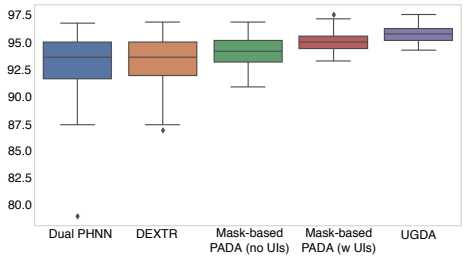

Fig. 3: **Box and whisker plot** of pathological liver segmentation DSC scores across the 100 $\mathcal{D}_t$ test set volumes.

As the table also demonstrates, by exploiting UIs, DEXTR can significantly boost the MXA, but its lackluster DSC scores suggest that the resulting masks, while aligning better with the extreme-points, still do not properly capture the liver extent. On the other hand, both mask-based PADA variants perform better, indicating that modelling mask distributions on top of DEXTR-style UI guidance can more robustly annotate $\mathcal{D}_t$. Finally, UGDA performs best, demonstrating that modelling the *interplay* between the UIs and mask predictions can boost performance even further. Importantly, UGDA's MXA is extremely good, which shows that the mask predictions well match the UIs. These mean scores are bolstered by Fig. 3's box-and-whisker plots, which demonstrate that UGDA provides important boosts in reliability, with an extremely robust worst-case performance of 94.9% DSC, compared to 93.2% for the mask-based PADA (w UIs) variant. Viewing the qualitative examples in Fig. 2 reinforces these quantitative improvements. In particular, UGDA is able to ensure that masks both agree with extreme-points and provide robust predictions away from the UIs.

Finally, as Tab. 2 also demonstrates, UGDA can perform almost as well when only a fraction of $\mathcal{D}_t$ is UI-labelled, outperforming both DEXTR and the mask-based PADA, both of which see all 100% of the UIs. These results indicate that

UGDA can operate well even in scenarios with *extremely minimal* UI annotation. providing further evidence of its very high versatility.

4 Conclusion

We presented user-guided domain adaptation (UGDA), an effective approach to use UIs to rapidly annotate 3D medical volumes. We tested UGDA on arguably the most challenging and comprehensive pathological liver segmentation dataset to date, demonstrating that by only using extreme-point UIs we can achieve DSC scores of 96.1%, outperforming both DEXTR [17] and conventional mask-based PADA [26]. Future work includes testing on other anatomical structures, incorporating additional types of UIs, and adapting UGDA for real-time interaction.

Supplementary material

Hyper parameters

We set λ_{adv} to 0.0001. The dual-PHNN learning rates was set to 0.003 and reduced by a factor of 0.1 when validation DSCs do not improve after 15 epochs, whereas the discriminator has a constant learning rate of 0.0003. We use the Adam optimizer [**?**] for both.

Table 2: Some of the pertinent details we followed for all our experiments.

Input volume resolution	$256 \times 256 \times 48$
Standard deviation for 3D Gaussian map	5
GPU system	Quadro RTX 8000
Cuda version	10.2
Number of GPUs and memory	$4 \times 48GB$
Deep learning framework and version	pytorch 1.4.0

References

1. Bilic, P., et al.: The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056 (2019)
2. Can, Y.B., Chaitanya, K., Mustafa, B., Koch, L.M., Konukoglu, E., Baumgartner, C.F.: Learning to segment medical images with scribble-supervision alone. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 236–244. Springer International Publishing, Cham (2018)
3. chaos: Chaos - combined (ct-mr) healthy abdominal organ segmentation (2019), `https://chaos.grand-challenge.org/Combined_Healthy_Abdominal_Organ_Segmentation`
4. Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C.: Multi-organ Abdominal CT Reference Standard Segmentations (Feb 2018), `https://doi.org/10.5281/zenodo.1169361`
5. Grady, L.: Random walks for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(11), 1768–1783 (Nov 2006)
6. Grady, L., Funka-Lea, G.: An energy minimization approach to the data driven editing of presegmented images/volumes. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006. pp. 888–895. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
7. Harrison, A.P., Birkbeck, N., Sofka, M.: Intelledits: Intelligent learning-based editor of segmentations. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. pp. 235–242. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
8. Harrison, A.P., Xu, Z., George, K., Lu, L., Summers, R.M., Mollura, D.J.: Progressive and multi-path holistically nested neural networks for pathological lung segmentation from ct images. In: International conference on medical image computing and computer-assisted intervention. pp. 621–629. Springer (2017)
9. Heimann, T., et al.: Comparison and evaluation of methods for liver segmentation from ct datasets. IEEE Transactions on Medical Imaging **28**(8), 1251–1265 (Aug 2009)
10. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1989–1998. PMLR, Stockholmsmssan, Stockholm Sweden (10–15 Jul 2018)
11. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)
12. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)
13. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 6936–6945. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00710

14. Lim, M., Tan, C., Cai, J., Zheng, J., Kow, A.: Ct volumetry of the liver: where does it stand in clinical practice? Clinical radiology **69**(9), 887–895 (2014)
15. Lodewick, T.M., Arnoldussen, C.W., Lahaye, M.J., van Mierlo, K.M., Neumann, U.P., Beets-Tan, R.G., Dejong, C.H., van Dam, R.M.: Fast and accurate liver volumetry prior to hepatectomy. HPB **18**(9), 764–772 (2016)
16. Lv, F., Lian, Q., Yang, G., Lin, G., Jialin Pan, S., Duan, L.: Domain adaptive semantic segmentation through structure enhancement. In: The European Conference on Computer Vision (ECCV) Workshops (September 2018)
17. Maninis, K.., Caelles, S., Pont-Tuset, J., Van Gool, L.: Deep extreme cut: From extreme points to object segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 616–625 (June 2018)
18. Nakayama, Y., Li, Q., Katsuragawa, S., Ikeda, R., Hiai, Y., Awai, K., Kusunoki, S., Yamashita, Y., Okajima, H., Inomata, Y., et al.: Automated hepatic volumetry for living related liver transplantation at multisection ct. Radiology **240**(3), 743–748 (2006)
19. Papadopoulos, D.P., Uijlings, J.R.R., Keller, F., Ferrari, V.: Extreme clicking for efficient object annotation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 4940–4949 (Oct 2017)
20. Rajchl, M., Lee, M.C.H., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., Rueckert, D.: Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE Transactions on Medical Imaging **36**(2), 674–683 (Feb 2017)
21. Raju, A., Cheng, C.T., Huo, Y., Cai, J., Huang, J., Xiao, J., Lu, L., Liao, C., Harrison, A.P.: Co-heterogeneous and adaptive segmentation from multi-source and multi-phase ct imaging data: A study on pathological liver and lesion segmentation. In: ECCV 2020 (2020)
22. Roth, H., Zhang, L., Yang, D., Milletari, F., Xu, Z., Wang, X., Xu, D.: Weakly supervised segmentation from extreme points. In: Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention, pp. 42–50. Springer (2019)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
24. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation (2019)
25. Taner, C.B., Dayangac, M., Akin, B., Balci, D., Uraz, S., Duran, C., Killi, R., Ayanoglu, O., Yuzer, Y., Tokat, Y.: Donor safety and remnant liver volume in living donor liver transplantation. Liver transplantation **14**(8), 1174–1179 (2008)
26. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472–7481 (2018)
27. Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T.: Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Transactions on Medical Imaging **37**(7), 1562–1573 (July 2018)
28. Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T.: Deepigeos: A deep interactive geodesic framework for medical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(7), 1559–1572 (July 2019)

29. Zhang, J., Xie, Y., Zhang, P., Chen, H., Xia, Y., Shen, C.: Light-weight hybrid convolutional network for liver tumour segmentation. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China. pp. 10–16 (2019)
30. Zhou, X., Zhuo, J., Krähenbühl, P.: Bottom-up object detection by grouping extreme and center points. In: CVPR (2019)