Accurate Weakly-Supervised Deep Lesion Segmentation using Large-Scale Clinical Annotations: Slice-Propagated 3D Mask Generation from 2D RECIST

Jinzheng Cai[1,2⋆], Youbao Tang[1⋆], Le Lu[1], Adam P. Harrison[1], Ke Yan[1], Jing Xiao[3], Lin Yang[2], and Ronald M. Summers[1]

[1] National Institutes of Health, Bethesda, MD, 20892, USA
[2] University of Florida, Gainesville, FL, 32611, USA
[3] Ping An Insurance (Group) Company of China, Ltd., Shenzhen, 510852, PRC
jimmycai@ufl.edu, {youbao.tang, le.lu, adam.harrison, ke.yan}@nih.gov, xiaojing661@pingan.com.cn, lin.yang@bme.ufl.edu, rms@nih.gov

Abstract. Volumetric lesion segmentation from computed tomography (CT) images is a powerful means to precisely assess multiple time-point lesion/tumor changes. However, because manual 3D segmentation is prohibitively time consuming, current practices rely on an imprecise surrogate called response evaluation criteria in solid tumors (RECIST). Despite their coarseness, RECIST markers are commonly found in current hospital picture and archiving systems (PACS), meaning they can provide a potentially powerful, yet extraordinarily challenging, source of weak supervision for full 3D segmentation. Toward this end, we introduce a convolutional neural network (CNN) based weakly supervised slice-propagated segmentation (WSSS) method to 1) generate the initial lesion segmentation on the axial RECIST-slice; 2) learn the data distribution on RECIST-slices; 3) extrapolate to segment the whole lesion slice by slice to finally obtain a volumetric segmentation. To validate the proposed method, we first test its performance on a fully annotated lymph node dataset, where WSSS performs comparably to its fully supervised counterparts. We then test on a comprehensive lesion dataset with 32,735 RECIST marks, where we report a mean Dice score of 92% on RECIST-marked slices and 76% on the entire 3D volumes.

1 Introduction

Given the prevailing clinical adoption of the response evaluation criteria in solid tumors (RECIST) [4] for cancer patient monitoring, many modern hospitals' picture archiving and communication systems (PACS) store tremendous amounts of lesion diameter measurements linked to computed tomography (CT) images. In this paper, we tackle the challenging problem of leveraging existing RECIST diameters to produce fully volumetric lesion segmentations in 3D. From any input CT image with the RECIST diameters, we first segment the lesion on the RECIST-marked image (RECIST-slice) in a weakly supervised manner, followed by generalizing the process into other successive slices to obtain the lesion's full volume segmentation.

⋆ indicates equal contribution

Inspired by related work [6,8,10,3] of weakly supervised segmentation in computer vision, we design our lesion segmentation in an iteratively slice-wise propagated fashion. More specifically, with the bookmarked long and short diameters on the RECIST-slice, we initialize the segmentation using unsupervised learning methods, *e.g.*, GrabCut [13]. Afterward, we iteratively refine the segmentation using a supervised convolutional neural network (CNN), which can accurately segment the lesion on RECIST-slices. Importantly, the resulting CNN model, trained from all RECIST-slices, can capture the appearance of lesions in CT slices. Thus, the model is capable of detecting lesion regions from images other than the RECIST-slices. With more slices segmented, more image data can be extracted and used to further fine-tune the model. As such, the proposed weakly supervised segmentation model is a slice-wise label-map propagation process, from the RECIST-slice to the whole lesion volume. Therefore, we leverage a large amount of retrospective (yet clinically annotated) imaging data to automatically achieve the final 3D lesion volume measurement and segmentation.

To compare the proposed weakly supervised slice-propagated segmentation (WSSS) against a fully-supervised upper performance limit, we first validate on a publicly-available lymph node (LN) dataset [12], consisting of 984 LNs with full pixel-wise annotations. After demonstrating comparable performance to fully-supervised approaches, we then evaluate WSSS on the DeepLesion dataset [16], achieving mean DICE scores of 92% and 76% on the RECIST-slices and lesion volumes, respectively.

2 Method

In the DeepLesion dataset [16], each CT volume contains an axial slice marked with RECIST diameters that represent the longest lesion axis and its perpendicular counterpart. RECIST diameters can act as a means of weakly supervised training data. Thus, we leverage weakly supervised principles to learn a CNN model using CT slices with no extra pixel-wise manual annotations. Formally, we denote elements in DeepLesion as $\{(V^i, R^i)\}$ for $i \in \{1, \ldots, N\}$, where N is the number of lesions, V^i is the CT volume of interest, and R^i is the corresponding RECIST diameter. To create the 2D training data for the segmentation model, the RECIST-slice and label pairs, X^i and Y^i, respectively, must be generated, and $X^i = V^i_r$ is simply the RECIST-slice, *i.e.*, the axial slice at index r that contains R. For notational clarity, we drop the superscript i for the remainder of this discussion.

2.1 Initial RECIST-Slice Segmentation

We adopt GrabCut [13] to produce the initial lesion segmentation on RECIST-slices. GrabCut is initialized with image foreground and background seeds, Y^s, and produces a segmentation using iterative energy minimization. The resulting mask is calculated to minimize an objective energy function conditioned on the input CT image and seeds:

$$Y = \arg\min_{\tilde{Y}} E_{gc}(\tilde{Y}, Y^s, X), \tag{1}$$

where we follow the original definition of the energy function E_{gc} in [13].

Given the fact that the quality of GrabCut's initialization will largely affect the final result, we propose to use the spatial prior information, provided by R, to compute high quality initial seeds, $Y^s = S(R)$, where $S(R)$ produces four categories: regions of background (BG), foreground (FG), *probable* background (PBG), and *probable* foreground (PFG). More specifically, if the lesion bounding box tightly around the RECIST axes is $[w, h]$, a $[2w, 2h]$ region of interest (ROI) is cropped from the RECIST-slice. The outer 50% of the ROI is assigned to BG whereas 10% of the image region, obtained from a dilation around R is assigned to FG. The remaining 40% is divided between PFG and PBG based on the distances to FG and BG. Fig. 1 visually depicts the training mask generation process (see the "RECIST to Mask" part). We use FG and BG as GrabCut seed regions, leaving the rest as regions where the initial mask is estimated.

2.2 RECIST-Slice Segmentation

We represent our CNN model as a mapping function $\hat{Y} = f(X; \theta)$, where θ represents the model parameters. Our goal is to minimize the differences between $\hat{Y}$ and the imperfect GrabCut mask Y, which contains 3 groups, namely the RECIST pixel indices $\mathcal{R}$, the estimated lesion (foreground) pixel indices $\mathcal{F}$, and the estimated background pixel indices set $\mathcal{B}$. Formally, the indices sets are defined to satisfy the constraints as $Y = Y_{\mathcal{R}} \cup Y_{\mathcal{F}} \cup Y_{\mathcal{B}}$, and $\mathcal{R} \cap \mathcal{F} = \mathcal{R} \cap \mathcal{B} = \mathcal{F} \cap \mathcal{B} = \emptyset$. Thus, we define CNN's training objective containing 3 loss parts as,

$$L = L_{\mathcal{R}} + \alpha L_{\mathcal{F}} + \beta L_{\mathcal{B}}, \tag{2}$$

$$= \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} -\log \hat{y}_i + \alpha \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} -\log \hat{y}_i + \beta \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} -\log (1 - \hat{y}_i), \tag{3}$$

where $\hat{y}_i$ is the i^{th} pixel in $\hat{Y}$, $|\cdot|$ represents the set cardinality and α, β are positive weights to balance the losses. Empirically, we set α, and β to small values at the start of model training when $\mathcal{F}, \mathcal{B}$ regions are estimated with low confidence. Afterwards, we set α, and β to larger values, *e.g.*, 1, when training converges.

2.3 Weakly Supervised Slice-Propagated Segmentation

To obtain volumetric measurements, we follow a similar strategy as with the RECIST-slices, except in this slice-propagated case, we must infer R for off-RECIST-slices and also incorporate inference results $\hat{Y}$ from the CNN model. These two priors are used together for slice-propagated CNN training.

RECIST Propagation: A simple way to generate off-RECIST-slice diameters $\hat{R}$ is to take advantage of the fact that RECIST-slice R lies on the maximal cross-sectional area of the lesion. The rate of reduction of off-RECIST-slice endpoints is then calculated by their relative offset distance to the RECIST-slice. Propagated RECIST endpoints are then projected from the actual RECIST endpoints by the Pythagorean theorem using physical Euclidean distance. The "3D RECIST Propagation" part in Fig. 1 depicts the propagation across CT slices. Given the actual RECIST on the r^{th} slice, $\hat{R}_{r-1}$ and $\hat{R}_{r-2}$ are the estimated RECISTs on the first and second off-RECIST-slices, respectively.

Off-RECIST-Slice Segmentation: For slice r, offset from the RECIST-slice, we update the seed generation function from Sec. 2.1 to now take both the inference from the RECIST-slice trained CNN, $\hat{Y}$, and the estimated RECIST, $\hat{R}$:

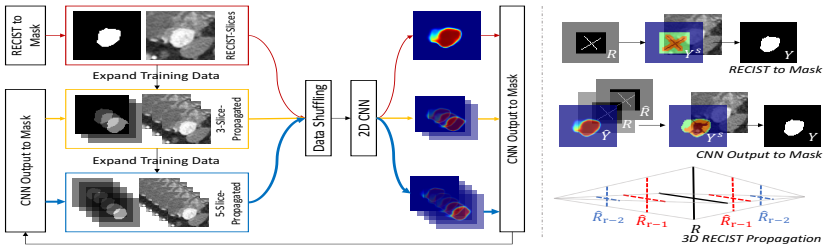

Fig. 1: Overview of the proposed method. **Right:** we use CNN outputs to gradually generate extra training data for lesion segmentation. Arrows colored in red, orange, and blue indicate slice-propagated training at its 1^{st}, 2^{nd}, and 3^{rd} steps, respectively. **Left:** regions colored with red, orange, green, and blue inside the initial segmentation mask Y present FG, PFG, PBG, and BG, respectively. Best viewed in color.

$Y^s = S(\hat{Y}, \hat{R}, R)$. More specifically, $\hat{Y}$ is first binarized by adjusting the threshold so that it covers at least 50% of R's pixels. Regions in $\hat{Y}$ that associate with high foreground probability values, *i.e.*, > 0.8, and overlap with $\hat{R}$ will be set as FG together with $\hat{R}$. Similarly, regions with high background probabilities and that have no overlap with $\hat{R}$ will be assigned as BG. The remaining pixels are left as uncertain using the same distance criteria as in the 2D mask generation case and fed into GrabCut for lesion segmentation. In the limited cases where the CNN fails to detect any foreground regions, we fall back to seed generation in Sec. 2.1, except we use $\hat{R}$ as input. The GrabCut mask is then generated using Equation (1) as before. This procedure is also visually depicted in Fig. 1 (see the "CNN Output to Mask" part).

Slice-Propagated CNN Training: To generate lesion segmentations in all CT slices from 2D RECIST annotations, we train the CNN model in a slice-propagated manner. The CNN first learns lesion appearances based on the RECIST-slices. After the model converges, we then apply this CNN model to slices $[V_{r-1}, V_{r+1}]$ from the entire training set to compute initial predicted probability maps $[\hat{Y}_{r-1}, \hat{Y}_{r+1}]$. Given these probability maps, we create initial lesion segmentations $[Y_{r-1}, Y_{r+1}]$ using GrabCut and the seed generation explained above. These segmentations are employed as training labels for the CNN model on the $[V_{r-1}, V_{r+1}]$ slices, ultimately producing the finally updated segmentations $[\hat{Y}_{r-1}, \hat{Y}_{r+1}]$ once the model converges. As this procedure proceeds iteratively, we can gradually obtain the converged lesion segmentation result across CT slices, and then stack the slice-wise segmentations $[\ldots, \hat{Y}_{r-1}, \hat{Y}_r, \hat{Y}_{r+1}, \ldots]$ to produce a volumetric segmentation. We visually depict this process in Fig. 1 from RECIST-slice to 5 successive slices.

3 Materials & Results

Datasets: The DeepLesion dataset [16] is composed of 32,735 bookmarked CT lesion instances (with RECIST measurements) from 10,594 studies of 4,459 patients. Lesions have been categorized into 8 subtypes: lung, mediastinum (MD), liver, soft-tissue (ST), abdomen (AB), kidney, pelvis, and bone. For quantitative evaluation, we segmented 1,000 testing lesion RECIST-slices manually. Out of these 1000, 200 lesions ($\sim$ 3,500 annotated slices) are fully segmented in 3D

Table 1: Performance in generating Y, the initial RECIST-slice segmentation. Mean DICE scores are reported with standard deviation for methods that defined in Sec. 3.1.

	Lymph Node			DeepLesion (on RECIST-Slice)		
Method	Recall	Precision	mDICE	Recall	Precision	mDICE
RECIST-D	0.35±0.09	**0.99±0.05**	0.51±0.09	0.39±0.13	**0.92±0.14**	0.53±0.14
DCRF	0.29±0.20	0.98±0.05	0.41±0.21	0.72±0.26	0.90±0.15	0.77±0.20
GrabCut	0.10±0.25	0.32±0.37	0.11±0.26	0.62±0.46	0.68±0.44	0.62±0.46
GrabCut^i	0.53±0.24	0.92±0.10	0.63±0.17	0.94±0.11	0.81±0.16	0.86±0.11
GrabCut-R	**0.83±0.11**	0.86±0.11	**0.83±0.06**	**0.94±0.10**	0.89±0.10	**0.91±0.08**

as well. Additionally, we also employ the lymph node (LN) dataset [12], which consists of 176 CT scans with complete pixel-wise annotations. Enlarged LN is a lesion subtype and producing accurate segmentation is quite challenging even with fully supervised learning [9]. Importantly, the LN dataset can be used to evaluate our WSSS method against an upper-performance limit, by comparing results with a fully supervised approach [9].

Pre-processing: For the LN dataset, annotation masks are converted into RECIST diameters by measuring its major and minor axes. For robustness, up to 20% random noise is injected into the RECIST diameter lengths to mimic the uncertainty of manual annotation by radiologists. For both datasets, based on the location of RECIST bookmarks, CT ROIs are cropped at two times the extent of the lesion's longest diameters so that sufficient visual context is preserved. The dynamic range of each lesion ROI is then intensity-windowed properly using the CT windowing meta-information in [16]. The LN dataset is separated at the patient level, using a split of 80% and 20% for training and testing, respectively. For the DeepLesion [16] dataset, we randomly select $28,000$ lesions for training.

Evaluation: The mean DICE similarity coefficient (mDICE) and the pixel-wise precision and recall are used to evaluate the quantitative segmentation accuracy.

3.1 Initial RECIST-Slice Segmentation

We denote the proposed GrabCut generation approach in Sec. 2.1 as GrabCut-R. To demonstrate our modifications in GrabCut-R are effective, we have compared it with general initialization methods, *i.e.*, densely connected conditional random fields (DCRF) [7], GrabCut, and GrabCut^i [6]. First, we define a bounding box (bbox) which is tightly covering the extent of RECIST marks. To initialize GrabCut, we set areas inside and outside the bbox as BG and PFG, respectively. To initialize GrabCut^i, we set the central 20% bbox region as FG, regions outside the bbox as BG, and the rest as PFG, which is similar to the setting of bbox^i in [6]. We then test DCRF [7] using the same bbox^i as the unary potentials and intensities to compute pairwise potentials. Since the DCRF is moderately sensitive to parameter variations, we record the best configuration we found and have it reported in Table 1. Finally, we measure results as we directly use the RECIST diameters, but dilated to 20% of bbox area, to generate the initial segmentation. We denote this approach RECIST-D, which produces the best precision, but at the cost of very low recall. From Table 1, we observe that GrabCut-R significantly outperforms all its counterparts on both of the Lymph Node and the DeepLesion datasets, demonstrating the validity of our mask initialization process.

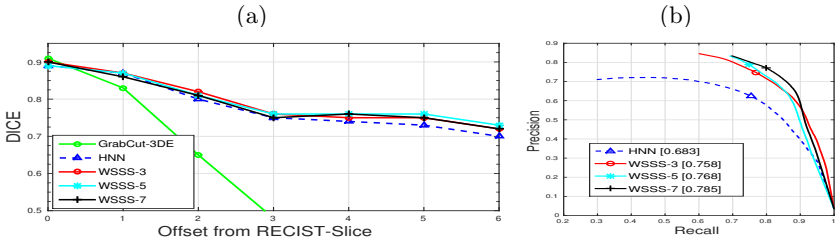

Fig. 2: WSSS on DeepLesion. (a) depicts mean Dice scores on *2D slices* as a function of offsets with respect to the RECIST-slice. (b) depicts *volumetric* precision-recall curves.

3.2 CNN based RECIST-Slice Segmentation

We use holistically nested networks (HNNs) [15] as our baseline CNN model, which has been adapted successfully for lymph node [9], pancreas [2], and lung segmentation [5]. In all experiments, deep learning is implemented in Tensorflow [1] and Tensorpack [14].The initial learning rate is 5×10^{-5}, dropping to 1×10^{-5} when the model training-validation plot plateaus. Given the results of Y, *i.e.*, >90% mDICE, we simply set the balance weights in Equation (2) as $\alpha = \beta = 1$.

Following Sec. 3.1, we select three ways to generate training masks on the RECIST-slice: the RECIST-D, GrabCut-R and the fully annotated ground truth (GT). As Table 2 demonstrates, on the LN dataset[12], HNNs trained using masks Y generated from RECIST-D, GrabCut-R, and GT achieve 61%, 70%, and 71% mDICE scores, respectively. This observation demonstrates the robustness and effectiveness of using GrabCut-R labels, which only performs slightly worse than using the GT. On the DeepLesion [16] testset of $1,000$ annotated RECIST-slices, HNN trained on GrabCut-R outperforms the deep model learned from RECIST-D by a margin of 25% in mean DICE (90.6% versus 64.4%). GrabCut post-processing, denoted with the suffix "-GC", further improves the results from 90.6% to 91.5%.

We demonstrate our weakly supervised approach, trained on a large quantity of "imperfectly-labeled" object masks, can outperform fully-supervised models trained on fewer data. To do this, we separated the $1,000$ annotated testing images into five folds and report the mean DICE scores using fully-supervised HNN [15] and UNet [11] models on this smaller dataset. Impressively, the 90.6% DICE score of the weakly supervised approach considerably outperforms the fully supervised HNN and UNet mDICE of 83.7% and 72.8%, respectively. Coupled with an approach like ours, this demonstrates the potential in exploiting large-scale, but "imperfectly-labeled", datasets.

3.3 Weakly Supervised Slice-Propagated Segmentation

In Fig. 2a, we show the segmentation results on 2D CT slices arranged in the order of offsets with respect to the RECIST-slice. GrabCut with 3D RECIST estimation (GrabCut-3DE), which is generated from RECIST propagation, produces good segmentations (∼91%) on the RECIST-slice but degrades to 55% mDICE when the offset rises to 4. This is mainly because 3D RECIST approx-

Table 2: Results of using different training masks, where GT refers to the manual segmentations. All results report mDICE ± std. GT results for the DeepLesion dataset are trained on the subset of 1,000 annotated slices. See Sec. 3.2 for method details.

	Lymph Node		DeepLesion (on RECIST-Slice)	
Method	CNN	CNN-GC	CNN	CNN-GC
UNet + GT	**0.729±0.08**	0.838±0.07	0.728±0.18	0.838±0.16
HNN + GT	0.710±0.18	**0.845±0.06**	0.837±0.16	0.909±0.10
HNN + RECIST-D	0.614±0.17	0.844±0.06	0.644±0.14	0.801±0.12
HNN + GrabCut-R	0.702±0.17	0.844±0.06	**0.906±0.09**	**0.915±0.10**

Table 3: Mean DICE scores for lesion volumes. "HNN" is the HNN [15] trained on GrabCut-R from RECIST slices and "WSSS-7" is the proposed approach trained on 7 successive CT slices. See Sec. 3.3 for method details.

Method	Bone	AB	MD	Liver	Lung	Kidney	ST	Pelvis	Mean
GrabCut-3DE	0.654	0.628	0.693	0.697	0.667	0.747	0.726	0.580	0.675
HNN	0.666	0.766	0.745	0.768	0.742	0.777	**0.791**	**0.736**	0.756
WSSS-7	**0.685**	0.766	**0.776**	**0.773**	0.757	**0.800**	0.780	0.728	0.762
WSSS-7-GC	0.683	**0.774**	0.771	0.765	**0.773**	**0.800**	0.787	0.722	**0.764**

imation often is not a robust estimation across slices. In contrast, the HNN trained with only RECIST slices, *i.e.*, the model from Sec. 3.2, generalizes well with large slice offsets, achieving mean DICE scores of $> 70\%$ even when the offset distance ranges to 6. However, performance is further improved at higher slice offsets when using the proposed slice-propagated approach with 3 axial slices, *i.e.*, WSSS-3, and even further when using slice-propagated learning with 5 and 7 axial slices, *i.e.*, WSSS-5, and WSSS-7, respectively. This propagation procedure is stopped at 7 slices as we observed the performance had converged. The current results demonstrate the value of using the proposed WSSS approach to generalize beyond 2D RECIST-slices into full 3D segmentation. We observe that improvements in mean DSC are not readily apparent, given the normalizing effect of that metric. However, when we measure F1-scores aggregated over the entire dataset (Fig. 2b, WSSS-7 improves over HNN from 0.683 to 0.785 (i.e., a lot of more voxels have been correctly segmented).

Finally, we reported the categorized 3D segmentation results. As demonstrated in Table 3, WSSS-7 propagates the learned lesion segmentation from the RECIST-slice to the off-RECIST-slices improving the 3D segmentation results from baseline 0.68 Dice score to 0.76. From the segmentation results of WSSS-7, we observe that the Dice score varies from 0.68 to 0.80 on different lesion categories, where the kidney is the easiest one and bone is the most challenging one. This suggests future investigation of category-specific lesion segmentation may yield further improvements.

4 Conclusion

We present a simple yet effective weakly supervised segmentation approach that converts massive amounts of RECIST-based lesion diameter measurements (retrospectively stored in hospitals' digital repositories) into full 3D lesion volume segmentation and measurements. Importantly, our approach does not require pre-existing RECIST measurement on processing new cases. The lesion seg-

mentation results are validated quantitatively, *i.e.*, 91.5% mean DICE score on RECIST-slices and 76.4% for lesion volumes. We demonstrate that our slice-propagated learning improves performance over state-of-the-art CNNs. Moreover, we demonstrate how leveraging the weakly supervised, but large-scale data, allows us to outperform fully-supervised approaches that can only be trained on subsets where full masks are available. Our work is potentially of high importance for automated and large-scale tumor volume measurement and management in the domain of precision quantitative radiology imaging.

Acknowledgement: This research was supported by the Intramural Research Program of the National Institutes of Health Clinical Center and by the Ping An Insurance Company through a Cooperative Research and Development Agreement. We thank Nvidia for GPU card donation.

References

1. Abadi, M., Agarwal, A., Barham, P., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), `https://www.tensorflow.org/`
2. Cai, J., Lu, L., Xie, Y., Xing, F., Yang, L.: Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function. MICCAI (2017)
3. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: IEEE ICCV. pp. 1635–1643 (2015)
4. Eisenhauer, E., Therasse, P., et al.: New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). Eur. J. Cancer pp. 228–247 (2009)
5. 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: MICCAI. pp. 621–629 (2017)
6. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: IEEE CVPR (2017)
7. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS. pp. 1–9 (2012)
8. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: IEEE CVPR. pp. 3159–3167 (2016)
9. Nogues, I., Lu, L., Wang, X., Roth, H., Bertasius, G., Lay, N., et al.: Automatic lymph node cluster segmentation using holistically-nested neural networks and structured optimization in ct images. In: MICCAI. pp. 388–397 (2016)
10. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: IEEE ICCV. pp. 1742–1750 (2015)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
12. Roth, H., Lu, L., Seff, A., Cherry, K., Hoffman, J., Liu, J., et al.: A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations. In: MICCAI. pp. 520–527 (2014)
13. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM TOG. vol. 23(3), pp. 309–314 (2004)
14. Wu, Y., et al.: Tensorpack. `https://github.com/tensorpack/` (2016)
15. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV. pp. 1395–1403 (2015)
16. Yan, K., Wang, X., Lu, L., Zhang, L., Harrison, A., et al.: Deep lesion graphs in the wild: Relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In: CVPR (2018)