Lesion-Harvester: Iteratively Mining Unlabeled Lesions and Hard-Negative Examples at Scale

Jinzheng Cai, Adam P. Harrison, Youjing Zheng, Ke Yan, Yuankai Huo, *Member, IEEE*, Jing Xiao, Lin Yang, and Le Lu

***Abstract*— The acquisition of large-scale medical image data, necessary for training machine learning algorithms, is hampered by associated expert-driven annotation costs. Mining hospital archives can address this problem, but labels often incomplete or noisy, *e.g.*, 50% of the lesions in DeepLesion are left unlabeled. Thus, effective label harvesting methods are critical. This is the goal of our work, where we introduce Lesion-Harvester—a powerful system to harvest missing annotations from lesion datasets at high precision. Accepting the need for some degree of expert labor, we use a small fully-labeled image subset to intelligently mine annotations from the remainder. To do this, we chain together a highly sensitive lesion proposal generator (LPG) and a very selective lesion proposal classifier (LPC). Using a new hard negative suppression loss, the resulting harvested and hard-negative proposals are then employed to iteratively finetune our LPG. While our framework is generic, we optimize our performance by proposing a new 3D contextual LPG and by using a global-local multi-view LPC. Experiments on DeepLesion demonstrate that Lesion-Harvester can discover an additional 9,805 lesions at a precision of 90%. We publicly release the harvested lesions, along with a new test set of *completely* annotated DeepLesion volumes. We also present a pseudo 3D IoU evaluation metric that corresponds much better to the real 3D IoU than current DeepLesion evaluation metrics. To quantify the downstream benefits of Lesion-Harvester we show that augmenting the DeepLesion annotations with our harvested lesions allows state-of-the-art detectors to boost their average precision by 7 to 10%.**

***Index Terms*— Lesion harvesting, lesion detection, hard negative mining, pseudo 3D IoU.**

Manuscript received July 23, 2020; accepted August 31, 2020. Date of publication September 7, 2020; date of current version December 29, 2020. *(Corresponding author: Jinzheng Cai.)*

Jinzheng Cai, Adam P. Harrison, Ke Yan, and Le Lu are with PAII Inc., Bethesda, MD 20817 USA (e-mail: caijinzheng883@paii-labs.com; adampharrison070@paii-labs.com; yanke383@paii-labs.com; le.lu@paii-labs.com).

Youjing Zheng is with the Department of Biomedical Sciences and Pathobiology, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 USA (e-mail: zhengyoujing@vt.edu).

Yuankai Huo is with the Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37212 USA (e-mail: yuankai.huo@vanderbilt.edu).

Jing Xiao is with Ping An Insurance (Group) Company of China, Ltd., Shenzhen 518001, China (e-mail: xiaojing661@pingan.com.cn).

Lin Yang is with the J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: lin.yang@bme.ufl.edu).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

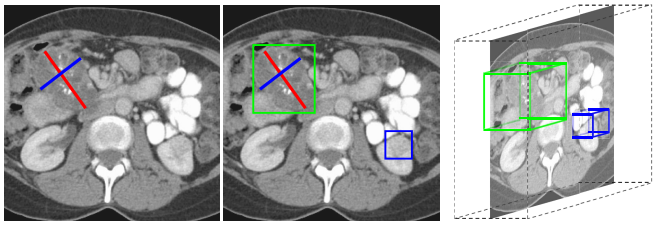
Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.3022034

I. Introduction

PARALLELING developments in computer vision, recent years have seen the emergence of large-scale medical image databases [1]–[5]. These are seminal milestones in medical imaging analysis research that help address the data-hungry needs of deep learning and other machine learning technologies. Yet, most of these databases are collected retrospectively from hospital picture archiving and communication systems (PACSs), which house the medical image and text reports from daily radiological workflows. While harvesting PACSs will likely be essential toward truly obtaining large-scale medical imaging data [6], their data are entirely ill-suited for training machine learning systems [7] as they are not curated from a machine learning perspective. As a result, popular large-scale medical imaging datasets suffer from uncertainties, mislabelings [3], [8], [9] and incomplete annotations [5], a trend that promises to increase as more and more PACS data is exploited. Correspondingly, there is a great need for effective data curation, but, unlike in computer vision, these problems cannot be addressed by crowd-sourcing approaches [10], [11]. Instead this need calls for alternative methods tailored to the demanding medical image domain. This is the focus of our work, where we articulate a powerful and effective label completion framework for lesion datasets, applying it to harvest unlabeled lesions from the recent DeepLesion dataset [5].

DeepLesion [4], [5] is a recent publicly released medical image database of CT sub-volumes along with localizations of lesions. These were mined from computed tomography (CT) scans from the US National Institutes of Health Clinical Center PACS. The mined lesions were extracted from response evaluation criteria in solid tumours (RECIST) [12] marks performed by clinicians to measure tumors in their daily workflow. See Fig. 1(a) for an example of a RECIST marked lesion. In total, DeepLesion contains 32, 735 retrospectively clinically annotated lesions from 10, 594 CT scans of 4, 427 unique patients. A variety of lesion types and subtypes have been included in this database, such as lung nodules, liver tumors, and enlarged lymph nodes. As such, the DeepLesion dataset is an important source of data for medical imaging analysis tasks, including training and characterizing lesion detectors and for developing radiomics-based biomarkers for tumor assessment and tracking. However, due to the RECIST guidelines [12] and workload limits, physicians typically marked only a small amount of lesions per CT scan as the *finding(s) of interest*.



a. RECIST mark b. Harvested lesion c. 3D box

Fig. 1. (a) depicts an example of a RECIST marked CT slice. RECIST marks may be incomplete by both not including co-exsiting lesions, *e.g.*, the blue 2D box in (b), or by not covering the 3D extent of lesions in other slices, *e.g.*, the green and blue *3D* boxes in (c). We aim to complete lesion annotations in both senses of (b) and (c).

Yet, as shown in Fig. 1(b), more often than not CT images exhibit multiple co-existing lesions per patient. Indeed, based on a recent empirical study [13], and our own results presented later, there are about the same quantity of missing findings compared to reported ones. Moreover, as Fig. 1(c) illustrates, RECIST marks do not indicate the 3D extent, leaving tumor regions of the same instance in adjoining slices unmarked. This severely challenges the development of high-fidelity disease detection algorithms and artificially limits the dataset's usefulness for biomarker development. Nevertheless, it is highly impractical and infeasible to recruit physicians to manually revise and add back annotations for the entire database.

To address this issue, we aim to reliably discover and harvest unlabeled lesions. Given the expert-driven nature of annotations, our approach, named Lesion-Harvester, accepts the need for a small amount of supplementary physician labor. It integrates three processes: (1) a highly sensitive detection-based lesion proposal generator (LPG) to generate lesion candidates, (2) manual verification of a small amount of the lesion proposals, and (3) a very selective lesion proposal classifier (LPC) that uses the verified proposals to automatically harvest prospective positives and hard negatives from the rest. These processes are tied together in an iterative fashion to strengthen the lesion harvesting at each round. Importantly, for (1) and (3) our framework can accept any state-of-the-art detector and classifier, allowing it to benefit from future improvements in these two domains. Even so, in this work, we develop our own LPG, called contextual-enhanced CenterNet (CECN), that combines the recent innovations seen in CenterNet [14] and the multitask universal lesion analysis network (MULAN) [15]. We also propose a hard negative suppression loss (HNSL) to boost our LPG with harvested hard negative cases. Among choices of LPCs, we use a global-local classifier with multi-view input (GLC-MV) to further reduce the false positive rate of produced lesion proposals. With our framework, we are able to harvest an additional 9, 805 lesions from DeepLesion, while keeping the label precision above 90%, at a cost of only fully annotating 5% of the data. Compared to the original dataset this is a boost of 11.2% in recall rates. Thus, our lesion harvesting framework, along with the introduced CECN LPG and HNSL, represent our main contributions.

However, we also provide additional important contributions. For one, we completely annotate and publicly release 1 915 of the DeepLesion subvolumes with RECIST marks, in addition to our harvested lesions. Second, we introduce and validate a new pseudo 3D (P3D) evaluation metric, designed for completely annotated data, that serves as a much better measurement of 3D detection performance than current practices.[1] Since all DeepLesion test data, up to this point, is incompletely annotated, these contributions allow for more accurate evaluations of lesion detection systems. Finally, we report concrete benefits of harvesting unlabeled lesions. To do this, we train several state-of-the-art detection systems [14], [16] using data augmented with our harvested prospective lesions and hard negative examples. We show that with even the best published method to date [15], the average precision (AP) can be improved by *10* percent. We also show that our CECN LPG can also be used as an extremely effective detector, outperforming the state-of-the-art and providing additional methodological contributions to the lesion detection topic.

II. Related Work

A. Detection With Incomplete Ground-Truth

The problem of missing annotations [17], [18] is related to, but differs from scenarios where the labeled data is independent from the unlabeled data. Thus, approaches to address the latter, such as deep growing learning [19], which uses a small-sized but fully-annotated training set to initialize the model and then gradually expand training examples with pseudo labels, would not fully exploit DeepLesion labels. Instead, labeled and unlabeled lesions in DeepLesion often co-exist in the same training images, which changes the nature of the problem.

Ren *et al.*, [17] addressed the partial label problem and reduced the effect of missing labels by only using true positives and hard negatives for training detectors, where the latter were defined as region proposals with at least an overlap of 0.1 with existing ground truth boxes. This setup, which we denote overlap-based hard sampling (OBHS), should help mitigate false negatives; however, it inevitably sacrifices a large amount of informative true negatives that have $\leq$0.1 overlap with ground truth boxes. Wu *et al.*, [18] proposed overlap-based soft sampling (OBSS) to improve object detectors, which weights contributions of region proposals proportional to their overlap with the ground truth. These strategies reduce the impact of true negatives not overlapping with the ground truth. Yet, true negatives in the background body structures are usually informative for training a robust lesion detector. By explicitly attempting to harvest prospective positives, Lesion-Harvester minimizes false negatives without suppressing informative background regions.

Focusing specifically on DeepLesion, Wang *et al.*, [20] first applied missing ground-truth mining (MGTM) to mine unlabeled lesions on the RECIST-marked CT slices. Complementary with MGTM, they also apply slice-level label propagation (SLLP) to propagate bounding boxes from the

[1]We open source our evaluation code, annotations, and results at https://github.com/JimmyCai91/DeepLesionAnnotation.

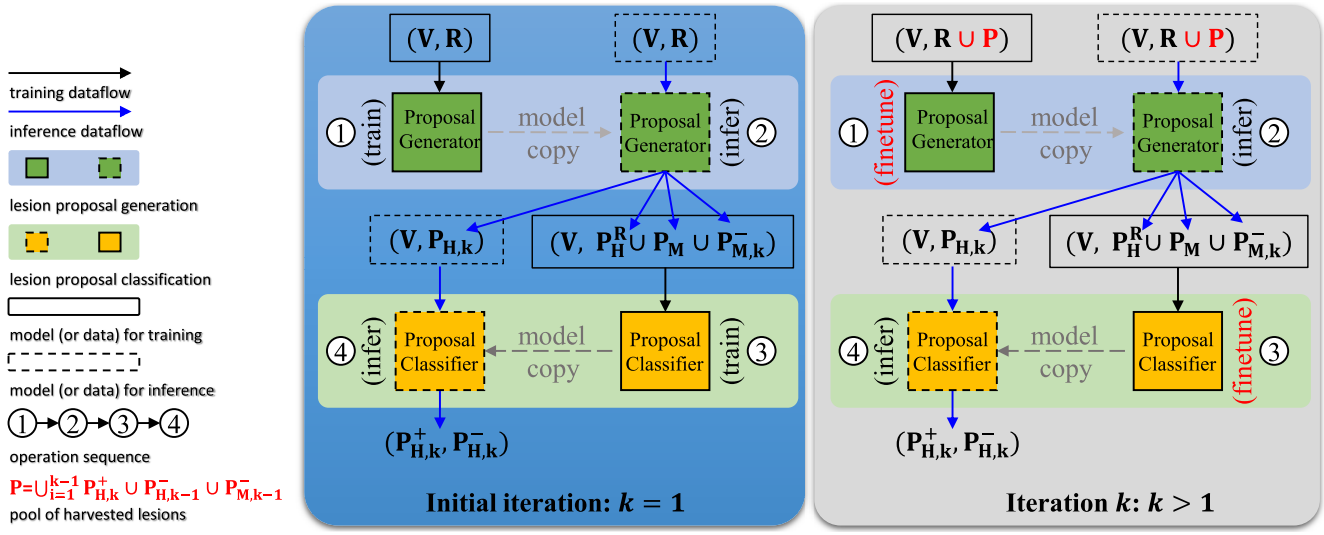


Fig. 2. The flowchart of Lesion-Harvester. The second and last columns depict the initial iteration and follow-up iterations, respectively, where we use ①, ②, ③, and ④ to indicate the sequence of operations. In step ①, we train (or finetune) the lesion proposal generator (LPG), and then, in step ②, we apply it on CT volumes $\mathbf{V}$ to generate 3D proposals. In step ③, we train/finetune the lesion proposal classifier (LPC), and apply it in step ④ to automatically separate proposals into positive and negative groups as $\mathbf{P}^{+}_{\mathbf{H},\mathbf{k}}$ and $\mathbf{P}^{-}_{\mathbf{H},\mathbf{k}}$, respectively.

RECIST-marked slices to adjacent slices to reconstruct lesions in 3D. Different from MGTM-SLLP, Lesion-Harvester conducts lesion mining on the whole CT volume. It also alternately mines and then updates the deep learning models, to iteratively strengthen the harvesting process. Finally, it also identifies and then integrates hard negative examples using hard negative suppression loss (HNSL).

B. Label Propagation From Partial Labels

Our work also relates to efforts on knowledge distillation and self label propagation. Radosavovic *et al.*, [21] proposed a data distillation method to ensemble predictions and automatically generate new annotations for unlabeled data from internet-scale data sources. Gao *et al.*, [22] investigated propagating labels from fully-supervised interstitial lung disease masks to unlabeled slices using convolutional neural networks (CNNs) and conditional random fields. Cai *et al.*, [23] recovered 3D segmentation masks from 2D RECIST marks in DeepLesion by integrating a CNN and GrabCut [24]. We also tackle the large-scale and noisy DeepLesion dataset, but we process each CT volume as a whole, rather than focus on post-processing a given region of interest.

C. Lesion Detection

Recently, Yan *et al.*, [4] introduced the DeepLesion dataset, which is a large-scale clinical database for whole body lesion detection, with follow-up work focusing on incorporating 3D context into 2D two-stage region-proposal CNNs [13], [15]. These studies demonstrate the importance of incorporating 3D context in lesion detection and currently represent the state-of-the-art performance on DeepLesion. Some recent works also investigated one-stage detectors [25], [26]. Compared with two-stage detectors, one-stage detectors are more flexible, straightforward, and computationally efficient. Recent work has also focused on false-positive detection. For instance, Ding *et al.*, [27] used a 3D-CNN for this task, while Dou *et al.*, [28] used a multi-scale context which ensembles three 3D-CNNs with small, medium, and large input patches. Varghese *et al.*, [29] proposed an interesting novelty detector (ND) for false positive reduction and applied it to brain tumor detection. More study is required to determine if their single layer denoising autoencoder has enough capacity to handle lesions with large appearance variabilities. Finally, Tang *et al.*, [30] showed that hard negative mining can increase detection sensitivity on DeepLesion.

Unlike the above works, our main focus is on harvesting missing *annotations* using a small amount of physician labor. To do so, we articulate the Lesion-Harvester pipeline that integrates state-of-the-art detection and classification solutions as LPG and LPC, respectively. We also introduce a one-stage multi-slice LPG that performs better than prior work. Like Tang *et al.*, [30], we also demonstrate the importance of hard negative mining. Finally, with our lesion harvesting task completed, we then show how the complete labels can be used to train the same LPG detection frameworks to better localize lesions on *unseen* data.

III. Methods

Fig. 2 overviews our proposed Lesion-Harvester. As motivated above, we aim to harvest missing annotations from the incomplete DeepLesion dataset [4], [5]. Additionally, as Fig. 1(c) demonstrates, another important aim is to fully localize the 3D extent of lesions, which means we aim to also generate 3D proposals for both RECIST-marked and unlabeled lesions. In this section, we first overview our method in Sec. III-A and then detail each component in Sec.s III-B–III-D. This is followed by Sec. III-E, which outlines our proposed pseudo 3D (P3D) evaluation metric.

A. Overview

For the problem setup, we assume we are given a set of CT volumes, $\mathbf{V} = \{v_i\}$. Within each volume, there is a set of RECIST-marked lesions which are denoted as

$$\mathbf{L_i^R} = \{\ell_{i,j}^R\}, \tag{1}$$

where $\ell_{i,j}^R$ is the j-th RECIST-marked lesion in v_i. We would like to use 3D bounding boxes to represent each lesion or lesion proposal. However, because a RECIST mark is only applied on the key slice, they only define a set of 2D boxes:

$$\mathbf{R}_i = \{r_{i,j}^R\}. \tag{2}$$

Because volumes are incompletely labeled, if we denote all lesions within volume v_i as $\mathbf{L_i}$, then $\mathbf{L_i^R}$ is a subset of $\mathbf{L_i}$. Our goal is to both determine the 3D extent of unlabeled lesions:

$$\mathbf{L_i^U} = \mathbf{L_i} \setminus \mathbf{L_i^R}, \tag{3}$$

and also recover the full 3D extent of any RECIST-marked lesion, *i.e.*, convert $\mathbf{R}_i$ to 3D bounding boxes. To do this, we first construct a *completely* annotated set of CT volumes, $\mathbf{V_M}$, by augmenting the original RECIST marks for these volumes with additional **manual** annotations for $\mathbf{L_M^U}$. We denote the supplementary and complete RECIST marks as $\mathbf{R_M^U}$ and $\mathbf{R_M}$.

The remainder of volumes we wish to harvest from are denoted as $\mathbf{V_H} = \mathbf{V} \setminus \mathbf{V_M}$, which are accompanied by their *incomplete* set of RECIST marks, $\mathbf{R_H}$. By exploiting $\mathbf{R_M}$ and $\mathbf{R_H}$, we attempt to discover all unlabeled lesions in $\mathbf{V_H}$ and the full 3D extent of both marked and unmarked lesions. Importantly, we constrain the size of $\mathbf{V_M}$ to be much smaller than $\mathbf{V_H}$, *e.g.*, 5%, to keep labor costs low.

In the initial round, we train a lesion proposal generator (LPG) using only the RECIST-derived 2D bounding boxes $\mathbf{R_M}$ and $\mathbf{R_H}$. To ensure flexibility, any state-of-the-art lesion detector can be used, either an off-the-shelf variant or the customized contextual-enhanced CenterNet (CECN) approach we elaborate in Sec. III-B. After convergence, we then execute the trained LPG on $\mathbf{V}$, producing a set of 3D lesion proposals, $\mathbf{P}$. These likely cover a large number of lesions but they may suffer from high false positive rates. To correct this, we divide $\mathbf{P}$ into $\mathbf{P_M}$ and $\mathbf{P_H}$ as proposals generated from $\mathbf{V_M}$ and $\mathbf{V_H}$, respectively. By comparing $\mathbf{P_M}$ with $\mathbf{R_M}$, we can further divide it into true and false positives:

$$\mathbf{P_M^R} = \mathbf{P_M} \cap \mathbf{R_M}, \quad \text{and } \mathbf{P_M^-} = \mathbf{P_M} \setminus \mathbf{R_M}, \tag{4}$$

where we use a pseudo-3D metric described in Sec. III-E to determine whether a 3D proposal and a RECIST-derived 2D bounding box intersect. Similarly, we can create a set of generated lesion proposals that intersect with the RECIST marks from $\mathbf{V_H}$, which we denote $\mathbf{P_H^R}$. We then train a lesion proposal classifier (LPC) by using $\mathbf{P_H^R}$ and $\mathbf{P_M^R}$ for positive training examples and $\mathbf{P_M^-}$ for negative examples. Like the LPG, any well-performing classification method can be used; however, we show that global-local classifier with multi-view input (GLC-MV) is particularly useful. The trained LPC is then used to classify the remaining proposals from $\mathbf{P_H}$ into $\mathbf{P_H^+}$ and $\mathbf{P_H^-}$, which are the harvested positive and negative lesion proposals, respectively.

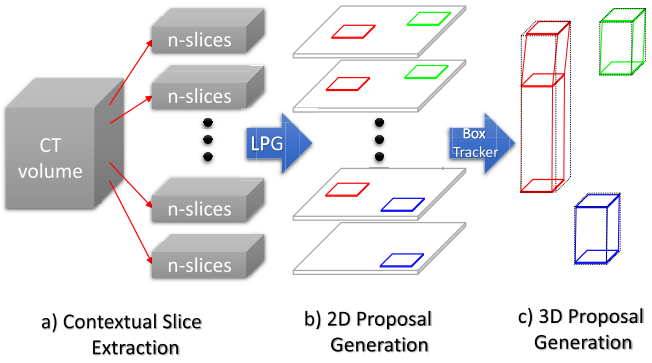


Fig. 3. Lesion proposal generation.

In subsequent rounds, we harvest positive and negative 3D proposals to finetune the LPG and begin the process anew. Yet, when fine tuning the proposed CECN we employ a hard negative suppression loss (HNSL) using $\mathbf{P_H^-}$ and $\mathbf{P_M^-}$ as mined hard negatives. We index algorithm iterations with subscript k and each iteration will provide harvested lesion proposals. Accordingly, the complete pool of all harvested lesions is iteratively updated as

$$\mathbf{P_H^+} = \mathbf{P_H^+} \cup \mathbf{P_{H,k}^+}, \tag{5}$$

where we abuse notation here and use $\cup$ as an operator that will fuse lesion proposals of the same lesion by simply keeping the one with the highest detection confidence. For $\mathbf{P_H^R}$ we only keep proposals intersecting with the RECIST marks. As for hard negatives, these are *reset* after each iteration. Unless needed, we drop the round index k for clarity. Below, we elaborate further on the individual system components.

B. Lesion Proposal Generation

The lesion proposal generator (LPG) uses a detection framework to generate lesion proposals. Following the state-of-the-art on DeepLesion [15], our LPG relies on a 2.5D lesion detection model to process CT images slice by slice; thus, 2D proposals must be aggregated together to produce 3D proposals. We outline each consideration below.

1) CECNs: The task of our LPG is to produce as high-quality lesion candidates as possible. While any state-of-the-art detection system can serve as LPG, there are attributes which are beneficial. An LPG with high sensitivity will help recover unlabeled lesions. Meanwhile, if it retains reasonable specificity, it will make downstream classification of proposals into true- and false-positives much more feasible. Computational efficiency is also important, to not only make training scalable, but also to be efficient in processing large amounts of CT images. Finally, simplicity and efficiency are also crucial virtues, as the LPG will be one component in a larger system.

While 3D LPGs can be powerful, there is no straightforward approach to apply them on PACS-mined data, like DeepLesion, which only provide 2D RECIST-derived annotations. The *de facto* standard in lesion detection for DeepLesion

are 2D/2.5D approaches [4], [13], [15], [25], [26], [30], which also avoid the prohibitive computational and memory demands of 3D LPGs. Thus, our LPG of choice is a CECN, which combines state-of-the-art one-stage anchor-free 2D proposal generation [14] with 3D context fusion [15]. Others have articulated the benefits of dense pixel-wise supervision [16], [30] and one-stage approaches provide a more straightforward means to aggregate such signals. Additionally, the choice of an anchor-free approach avoids the need to tune anchor-related hyper-parameters. Because lesions have convex shapes which have centroids located inside lesions, the center-based loss of CenterNet [14] is a natural choice. A final advantage to the one-stage anchor-free approach, which we will show in Sec. III-D, is that it allows for a natural incorporation of hard negative examples, which significantly improves performance.

We then follow the same pipeline and hyper-parameter settings described by Zhou *et al.*, [14]. Namely, we create ground-truth heat-maps centered at each lesion using Gaussian kernels $Y \in [0,1]^{\tilde{W}\times\tilde{H}}$. The training objective is to then produce a heatmap, $\hat{Y}_{xy}$, using a penalty-reduced pixel-wise logistic regression with focal loss [31]:

$$E_{\text{LPG}} = \frac{-1}{m}\sum_{xy}\begin{cases}(1-\hat{Y}_{xy})^{\alpha}\log(\hat{Y}_{xy}) & \text{if } Y_{xy}=1\\ (1-Y_{xy})^{\beta}(\hat{Y}_{xy})^{\alpha} & \\ \log(1-\hat{Y}_{xy}) & \text{otherwise,}\end{cases} \tag{6}$$

where m is the number of objects in the slice and $\alpha = 2$ and $\beta = 4$ are hyper-parameters of the center loss. At every output pixel, the width, height, and offset of lesions are also regressed, but they are only supervised where $Y_{xy} = 1$. The lesion proposals are produced by combining center points with regressed width and height. See Zhou *et al.*, [14] for more details.

To incorporate 3D context, which Yan *et al.*, [15] demonstrated can benefit lesion detection, we use a 2.5D DenseNet-121 backbone [32]. This backbone is associated with the highest performance for DeepLesion detection to-date [15] and functions by including consecutive CT slices as input channels. Based on a balance between performance and computational efficiency, we follow Yan *et al.*, [15] choose to include the 4 adjoining slices above and below.

2) 3D Proposal Generation: Regardless of the LPG used, if it operates slice-wise, like the CECN, then post-processing is required to generate 3D proposals, $\mathbf{P}$. To do this, we first apply the LPG to scan over CT volumes generating detection results on every axial slice. This produces a set of 2D proposals, each with a detection score. Next, we stack proposals in consecutive slices using the same Kalman filter-based bounding box tracker as Yang *et al.*, [33]. More specifically, we first select 2D proposals whose detection score is greater than a threshold t_G. 2D proposals from adjoining slices are then stacked together if their intersection over union (IoU) is $\geq$0.8. Finally, in case the LPG misses lesions in intermediate slices, we extend each 3D box up and down by one slice, and if any two 3D boxes become connected with $\geq$0.8 overlap on the connecting slices, then they will be fused as one 3D proposal. When lesion harvesting, we choose 0.1 as the value for t_G which helps keep the number of proposals manageable. However, our experience indicates that results are not sensitive to significant deviations from our chosen threshold value.

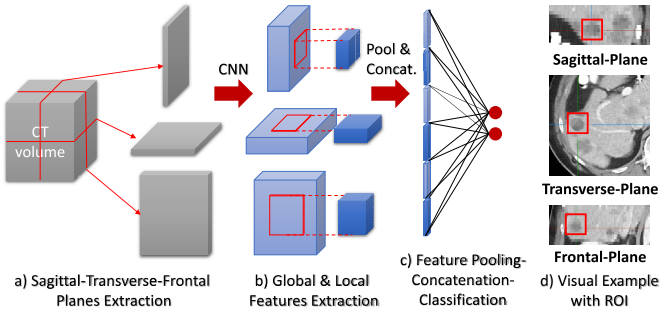


Fig. 4. Lesion proposal classification using global-local classifier with multi-view input (GLC-MV).

The next step in our process is to separate lesion candidates, $\mathbf{P}$, into true- and false-positives. Because we have access to a small subset of fully-annotated volumes, $\mathbf{V_M}$, we can identify the 3D proposals that overlap (see Sec. III-E) with the RECIST marks to be true-positives and denote them as $\mathbf{P}^{\mathbf{R}}_{\mathbf{M}}$. The remaining false-positive proposals are denoted $\mathbf{P}^{-}_{\mathbf{M}}$. We can also identify true-positives $\mathbf{P}^{\mathbf{R}}_{\mathbf{H}}$ that overlap with existing RECIST marks in $\mathbf{V_H}$. Because $\mathbf{V_H}$ is only partially labeled, the remainder of proposals must be filtered somehow into true and false positives.

C. Lesion Proposal Classification

With the manually verified proposals in hand, namely $\mathbf{P}^{\mathbf{R}}_{\mathbf{M}}$, $\mathbf{P}^{-}_{\mathbf{M}}$, and $\mathbf{P}^{\mathbf{R}}_{\mathbf{H}}$, the aim is to identify proposals in $\mathbf{P_H} \setminus \mathbf{P}^{\mathbf{R}}_{\mathbf{H}}$. To do this, we use the verified proposals to train a binary lesion proposal classifier (LPC). In principle, any classifier can be used, but we opt for a global-local classifier with multi-view input (GLC-MV), which combines two main concepts. The first concept is that 3D context is necessary for differentiating true positive lesion proposals from false positives [27], [28], whether for machines or for clinicians. A fully 3D classifier can satisfy this need, but, as we show in the results, a multi-view approach that operates on transverse, sagittal, and coronal planes centered at each proposal can perform better. This matches prior practices [34] and has the virtue of offering a much more computational and memory efficient means to encode 3D context compared to true 3D networks.

The second concept is multi-scale learning. As reported by Yan *et al.*, [5], a "global" context can aid in lesion characterization. This is based on the intuition that the surrounding anatomy can help place a prior on how lesions should appear. We use the same recommendation as Yan *et al.*, and use a $128 \times 128 \times 32$ region centered at each lesion to extract global features. ROI pooling based on the lesion proposal is then used to extract more local features. The local and global features are concatenated together before being processed by a fully-connected layer. The GLC-MV is shown in Fig. 4.

We choose to use a ResNet-18 [35] as our backbone because of its proven usefulness and availability of pre-trained weights. The LPC is trained with the manually verified

proposals using cross-entropy loss. We expect $\mathbf{P}_{\mathbf{M}}^{\mathbf{R}}$, $\mathbf{P}_{\mathbf{M}}^{-}$, and $\mathbf{P}_{\mathbf{H}}^{\mathbf{R}}$ to be representative of the actual distribution of lesions in DeepLesion. In particular, although the negative samples $\mathbf{P}_{\mathbf{M}}^{-}$ are only generated from $\mathbf{V}_{\mathbf{M}}$, they should also be representative to the dataset-wide distribution of hard negatives since *the hard negatives are typically healthy body structures, which are common across patients*.

With the LPC trained, we then apply it to the proposals needing harvesting: $\mathbf{P}_{\mathbf{H}} \setminus \mathbf{P}_{\mathbf{H}}^{\mathbf{R}}$. Since the LPG and LPC are independently trained, we make an assumption, for simplicity, that their pseudo-probability outputs are independent as well. Thus, the final score of a 3D proposal can be calculated as

$$s_{\{p|g,c\}} = s_g \cdot s_c, \tag{7}$$

where $s_{\{p|g,c\}}$ is called the lesion score and s_g and s_c are the LPG detection score and LPC classification probability, respectively. We obtain the former by taking the max detection score across all 2D boxes in the proposal. Based on $s_{\{p|g,c\}}$, we generate prospective positive proposals, $\mathbf{P}_{\mathbf{H}}^{+}$, by choosing a threshold for $s_{\{p|g,c\}}$ that corresponds to a precision above 95% on the completely annotated set $\mathbf{V}_{\mathbf{M}}$. From the remainder, we select proposals whose detection score satisfy $s_g \geq 0.5$ as the hard negative examples. Then from each volume we choose up to five negative examples with the top detection scores to construct $\mathbf{P}_{\mathbf{H}}^{-}$.

D. Iterative Updating

After a round of harvesting, we repeat the process by fine-tuning the LPG, but with important differences. First, we now have prospective positive proposals corresponding to unlabeled lesions, *i.e.*, $\mathbf{P}_{\mathbf{H}}^{+}$, to feed into training. In addition, for all proposals, even those corresponding to RECIST-marked lesions, we now have 3D proposals. To keep computational demands reasonable, only the 2D slices with the highest detection score within each proposal are used as additional samples to augment the original RECIST slices.

Secondly, to incorporate harvested hard negative proposals, we use the same procedure in Sec. III-B, but replace the center-loss in Eq. 6 with our proposed hard negative suppression loss (HNSL). To do this, we create separate heat maps for positive (RECIST-marked or prospective positive) and hard-negative lesions. We denote these heat maps as Y_{xy}^{p} and Y_{xy}^{n}, respectively. We then create a master ground truth heat map, Y_{xy}, by overwriting Y_{xy}^{p} with Y_{xy}^{n}:

$$Y_{xy} = \begin{cases} -Y_{xy}^{n} & \text{if } Y_{xy}^{n} > 0 \\ Y_{xy}^{p} & \text{otherwise.} \end{cases} \tag{8}$$

The result is a ground truth map that can now range from $[-1, 1]$. When used in the loss of (6), the effect is that positive predictions in hard negative regions are penalized much heavier than standard negative regions (16 times heavier when $\beta = 4$). This simple modification works surprisingly well for further reducing false positive rates. We visually depict example ground truth heatmaps in Fig. 5.

E. Pseudo-3D Evaluation

Apart from the lesion completion framework, introduced above, another important aspect to discuss is evaluation.

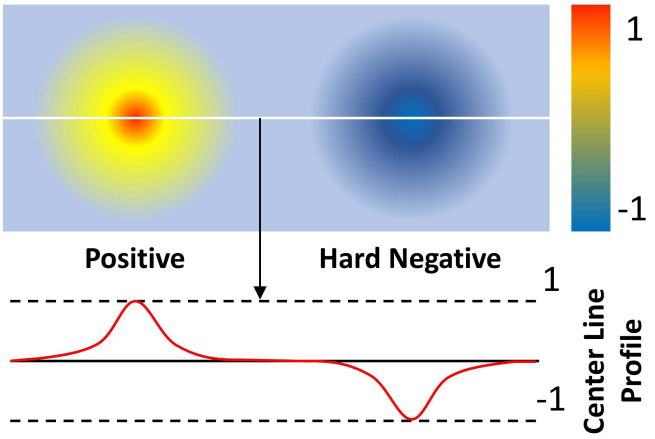


Fig. 5. Ground-truth heatmaps for positives and hard-negative examples are shown in the left and right, respectively. We define both RECIST-marked lesions and mined lesions to be positive examples.

Current DeepLesion works [4], [13], [15], [25], [26], [30] operate and evaluate only based on the 2D RECIST marks on selected 2D slices that happen to contain said marks. This is problematic, as RECIST-based evaluation will not reflect actual performance: it will miscount *true positives* on unmarked lesions or on adjoining slices as *false positives*. Moreover, automated methods should process the whole image volume, meaning precision should be correlated to false positives per *volume* rather than per selected *slice*. In this way, automated methods can be more effective on holistically describing and recording tumor existence, complimentary to human efforts to better achieve precision medicine.

Because we aim to harvest 3D bounding boxes that cover all lesions, we must evaluate, by definition, on completely annotated test data. Yet, it is not realistic to assume data will be fully annotated with 3D bounding boxes. Instead, a more realistic prospect is that test data will be completely annotated with 2D RECIST marks, especially by clinicians who are more accustomed to this. Thus, assuming this is the test data available, we propose a pseudo 3D (P3D) IoU metric. For each RECIST mark, we can generate 2D bounding boxes based off of their extent, as in [4]. This we denote $(x_1, x_2, y_1, y_2, z, z)$, where z is the slice containing the mark. Given a 3D bounding box proposal, $(x_1', x_2', y_1', y_2', z_1', z_2')$, our P3D IoU metric will be counted as a true positive if and only if $z_1' \leq z \leq z_2'$ and $IoU\left[(x_1, x_2, y_1, y_2), (x_1', x_2', y_1', y_2')\right] \geq 0.5$. Otherwise, it is considered a false positive. Because we publicly release complete RECIST marks of 1915 volumes, the P3D IoU metric can also be used to benchmark DeepLesion detection performance, replacing the one currently used. As we show in the results, the P3D IoU metric is a much more accurate performance measure.

IV. Experiments

A. Dataset

To harvest lesions from the DeepLesion dataset, we randomly selected 844 volumes from the original

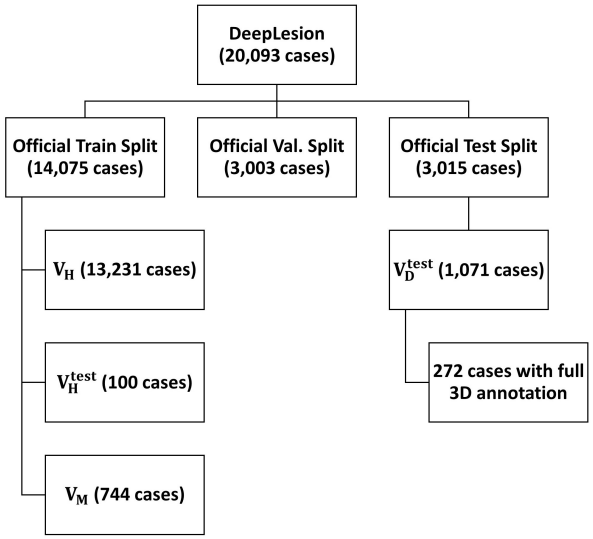


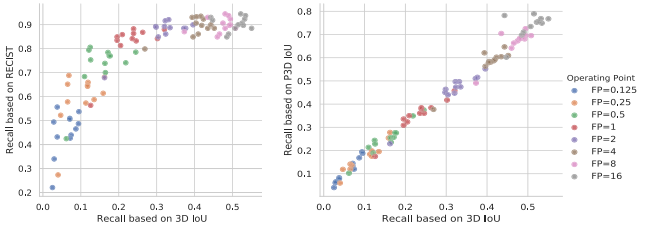
Fig. 6. **Dataset Splits:** we follow DeepLesion's official data split and further define $\mathbf{V_H}$, $\mathbf{V_H^{test}}$, $\mathbf{V_M}$, and $\mathbf{V_D^{test}}$ for the needs of lesion harvesting.

14 075 training CTs.[2] These are then annotated by a board-certified radiologist. Of these, we select 744 as $\mathbf{V_M}$ (5.3%) and leave another 100 as an evaluation set for lesion harvesting. This latter subset, denoted $\mathbf{V_H^{test}}$, is treated identically at $\mathbf{V_H}$, meaning the algorithm only sees the original DeepLesion RECIST marks. After convergence, we can measure the precision and recall of the harvested lesions. In addition, we later measure detection performance on systems trained on our harvested lesions by also fully annotating, with RECIST marks, 1,071 of the testing CT volumes. These volumes, denoted $\mathbf{V_D^{test}}$, are never seen in our harvesting framework. To clarify the dataset split in our experiments, we show it hierarchically in Fig. 6.

B. P3D IoU Evaluation Metric

Before validating our lesion harvesting framework, we first validate our proposed P3D metric. To do this, we used *3D bounding boxes* to annotate a small set of 272 CT test volumes, randomly selected from $\mathbf{V_D^{test}}$. From these, we can calculate a "gold-standard" 3D IoU metric, and analyze the concordance of different proxies. Accordingly, we trained state-of-the art detection methods (the same 12 outlined in Table IV's later experiments) on the DeepLesion dataset and measured their performance using 3D IoU, incomplete RECIST [4], and the proposed P3D IoU metrics. We use a 3D IoU threshold of 0.3, instead of 0.5 commonly used in 2D applications, to help compensate for the severity of 3D IoU. As shown in Fig. 7a and Fig. 7b, we measured free-response ROC curve (FROC) curves and compare detection recalls of these methods at operating points varying from false positive (FP) rates of 0.125 to 16 per volume. As can be seen, our P3D metric has much higher concordance with the true 3D IoU than does the incomplete 2D RECIST metric. Moreover, the latter exhibits a relationship that is much noisier and non-monotonic, making

[2]https://nihcc.app.box.com/v/DeepLesion



(a) 0.838, $p \ll 10^{-5}$. (b) 0.993, $p \ll 10^{-5}$;

Fig. 7. Comparing the concordance of the incomplete 2D RECIST and our P3D metric compared to the gold-standard 3D IoU metric. For each metric, lesion detection recalls at operation points from FP = 0.125 to FP = 16 are collected from the FROC curves of 12 lesion detection methods. Pearson coefficients are shown below each chart.

TABLE I
LESION HARVESTING PERFORMANCE EVALUATED ON $\mathbf{V_H^{TEST}}$. DETECTION RECALLS (R) AT PRECISIONS (P) FROM 80% TO 95% ARE REPORTED AFTER SIX HARVESTING ROUNDS

Training Label Set	R@80P	R@85P	R@90P	R@95P	Avg.
$\mathbf{R_H}$ (RECIST)			36.7@100P		
MGTM-SLLP [20]	40.0	38.8	38.6	37.6	38.7
$\mathbf{P_H^R} \cup \mathbf{P_{H,1}}$	46.7	42.4	40.7	39.9	42.4
$\mathbf{P_H^R} \cup \mathbf{P_{H,1}^+}$	46.3	45.7	44.0	41.7	44.5
$\mathbf{P_H^R} \cup \bigcup_{i=1}^{2} \mathbf{P_{H,i}^+}$	52.4	50.0	44.7	41.3	47.1
$\mathbf{P_H^R} \cup \bigcup_{i=1}^{3} \mathbf{P_{H,i}^+}$	51.4	**50.5**	**47.9**	41.2	47.7
$\mathbf{P_H^R} \cup \bigcup_{i=1}^{4} \mathbf{P_{H,i}^+}$	**53.3**	50.0	46.1	41.9	**47.8**
$\mathbf{P_H^R} \cup \bigcup_{i=1}^{5} \mathbf{P_{H,i}^+}$	52.9	48.6	45.8	**42.5**	47.4
$\mathbf{P_H^R} \cup \bigcup_{i=1}^{6} \mathbf{P_{H,i}^+}$	**53.3**	48.6	43.8	40.2	46.5

it likely any ranking of methods does not correspond to their true ranking. Thus, for the remainder of this work we report lesion harvesting and lesion detection performance using only the P3D metric. Moreover, we advocate using the P3D metric, and the fully RECIST-annotated test sets we publicly release, to evaluate DeepLesion detection systems going forward.

We also evaluated whether the Kalman filtering method in Sec. III-B can produce accurate 3D lesion proposals from 2D detections. Regardless of the detection framework used, median 3D IoUs are 0.4, which is a high overlap for 3D detection. All of the first quartiles are above 0.2 3D IoU, indicating that the most of the reconstructed 3D boxes are of high quality. Violin plots can be found in our supplementary material.

C. Main Result: Lesion Harvesting

We validate our lesion harvesting by running it for 6 iterations. As can be seen in Table I, the set of original RECIST-marked lesions, $\mathbf{R_H}$, only has a recall of 36.7% for the lesions in $\mathbf{V_H^{test}}$, with an assumed precision of 100%. After one iteration, the initial lesion proposals generated by the CECN-based LPG, denoted as $\mathbf{P_{H,1}}$, can boost the recall to 40.7%, while keeping the precision at 90%. However, after filtering with our GLC-MV-based LPC, which selects $\mathbf{P_{H,1}^+}$ from $\mathbf{P_{H,1}}$, the recall is boosted to 44.0%, representing 8% increase in recall over $\mathbf{R_H}$. This demonstrates the power

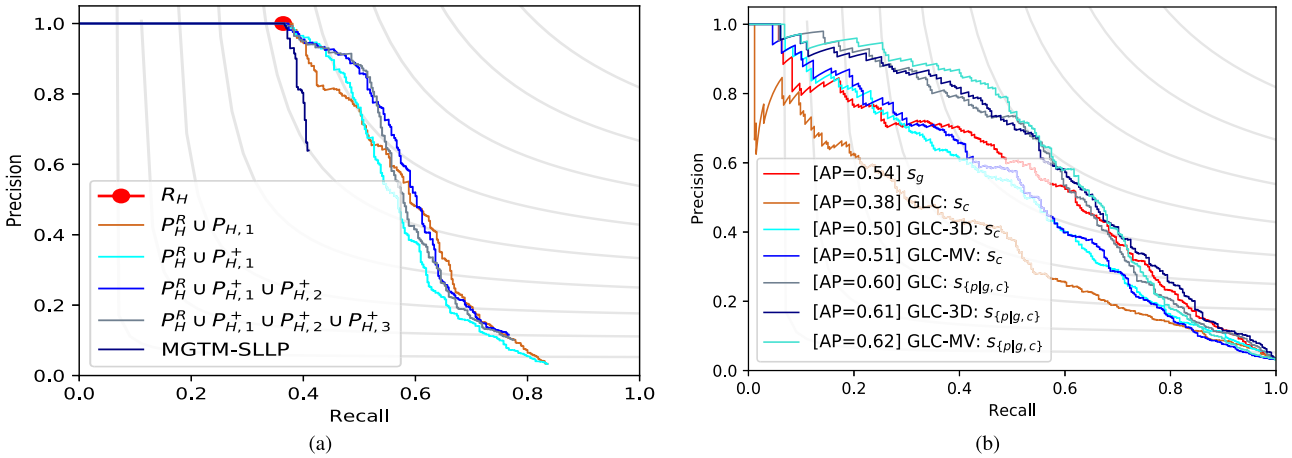


Fig. 8. (a) PR curve evaluating our iterative lesion harvesting procedure. The original RECIST marks only have a recall of 36.7%, which is shown as the red dot. $\mathbf{P_{H,1}}$ is the set of lesion proposals from LPG, which are then filtered by the LPC to generate $\mathbf{P^+_{H,1}}$. $\mathbf{P^+_{H,2}}$ and $\mathbf{P^+_{H,3}}$ are harvested lesions from the second and third iterations, respectively. MGTM-SLLP [20] is a baseline method for mining missing lesions. (b) PR curves of different LPCs in the first iteration, where GLC, GLC-3D, and GLC-MV denote classifiers using the transverse-plane, the multi-view sagittal-transverse-frontal plane, and 3D sub-volume inputs, respectively.

TABLE II
EVALUATIONS OF LPG TRAINED WITH DIFFERENT LABEL SETS INCLUDING THE ORIGINAL RECIST MARKS, $\mathbf{R}$, RECOVERED 3D RECIST-MARKED LESIONS $\mathbf{P^R_H}$ AND $\mathbf{P^R_M}$, MINED LESIONS $\bigcup_{i=1}^{k-1} \mathbf{P^+_{H,I}}$, AND MINED HARD NEGATIVES $\mathbf{P^-_H}$ AND $\mathbf{P^-_M}$. THE RECALL NUMBERS ARE EXTRACTED FROM FROC CURVES AT OPERATION POINTS FROM FP = 0.125 TO FP = 8 PER VOLUME. HIGHER RECALL DEMONSTRATES BETTER DETECTION PERFORMANCE

Exp.	iter. k	$\mathbf{R}$	$\mathbf{P^R_M}$	$\mathbf{P^-_{M,k-1}}$	$\mathbf{P^R_H}$	$\bigcup_{i=1}^{k-1} \mathbf{P^+_{H,k}}$	$\mathbf{P^-_{H,k-1}}$	Recall (%) @ FPs=[0.125,0.25,0.5,1,2,4,8] per volume						
(a)	1	✓						12.0	21.0	31.7	45.7	53.1	61.0	66.9
(b)	1	✓	✓					16.0	21.7	32.9	44.3	55.7	62.6	69.5
(c)	2	✓	✓	✓				19.5	27.4	36.9	48.8	55.5	64.8	70.5
(d)	2	✓	✓	✓	✓			17.9	31.2	45.5	52.4	**60.5**	**66.2**	**72.6**
(e)	2	✓	✓	✓	✓	✓		28.9	35.5	45.5	53.1	57.6	65.2	70.7
(f)	2	✓	✓	✓	✓	✓	✓	**30.4**	38.6	**47.1**	53.6	58.6	62.6	68.3
(g)	3	✓	✓	✓	✓	✓	✓	25.5	**38.8**	**47.1**	**54.3**	59.5	64.1	71.7

and usefulness of our LPG and LPC duo. After 3 rounds of our system, the performance increases further, topping out at 47.9% recall at 90% precision. This corresponds to harvesting **9 805** more lesions from the $\mathbf{V_H}$ CT volumes. As Table I also indicates, running Lesion-Harvester for more rounds after 3 does not notably contribute to performance improvements. Fig. 8a, depicts PR curves for the first three rounds and it can be seen that PR curve has begun to top out after the third round. Finally, we also compare against the MGTM-SLLP lesion mining algorithm [20]. As can be seen, while MGTM-SLLP can also improve recall, the Lesion-Harvester outperforms it by 9.3% in recall at 90% precision.

Importantly all 2D lesion bounding boxes are now also converted to 3D. It should be stressed that these results are obtained by annotating 744 volumes, which represents only 5.3% of the original training split of DeepLesion. In terms of the distribution of harvested lesions, they match the original DeepLesion distribution which is weighted toward lung, liver, kidney lesions as well as enlarged lymph nodes. In our supplementary material, we illustrate the similarity of the distributions between the original DeepLesion lesions and the harvested ones in details. Fig. 10 provides some visual examples of the harvested lesions. As can be seen, lesions missing from the original RECIST marks can be harvested. These examples, coupled with the quantitative boosts in recall (seen in Fig. 8a), demonstrate the utility and power of our lesion harvesting approach.

In our implementation, we trained the LPG and LPC using 3 NVIDIA RTX6000 GPUs, which took a few hours to converge for each round. However, producing the lesion proposals after each round is time consuming (∼12 hours) since it requires the LPG to scan every CT slice in DeepLesion. In total, the Lesion-Harvester takes 3 days to converge on DeepLesion.

D. Ablation Study: LPG

Table II presents the performance of the CECN-based LPG when trained with different combinations of harvested lesions. Please note, these results only measure the LPG performance, and do not include the effect of the LPC filtering. First, as expected, when including the additional labeled proposals, $\mathbf{R^U_M}$, the performance does not improve much over simply using the original RECIST marks. This reflects the relatively small size of $\mathbf{R^U_M}$ compared to the entire dataset. However, larger impacts can be seen when including the hard negatives,

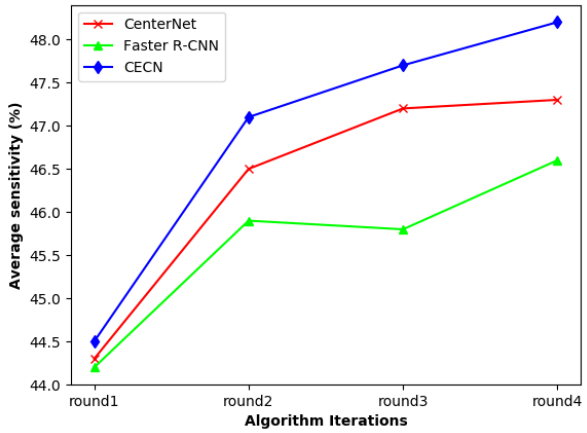


Fig. 9. Lesion-Harvester using different LPGs, including CenterNet [14], Faster R-CNN [17], and our proposed CECN. Mean value of recalls at precision $= [80\%, 85\%, 90\%, 95\%]$ is measured with each ablation study.

$\mathbf{P}^{-}_{\mathbf{M,i}}$, from the fully-labeled subset. When including hard negatives from our volumes needing harvesting, *i.e.*, $\mathbf{P}^{-}_{\mathbf{H,i}}$, performance boosts are even greater at the high precision operation points where FPs ≤ 1. This validates our HNSL approach of using hard-negative cases. Meanwhile, the addition of extra positive samples, $\mathbf{P}^{+}_{\mathbf{H,i}}$ and $\mathbf{P}^{\mathbf{R}}_{\mathbf{H}}$, contribute much to the recall when FPs ≥ 2 per volume, as the trained LPG becomes much more sensitive. In summary, these results indicate that the harvested lesions and hard negatives can significantly boost how many lesions can be recovered from the DeepLesion dataset.

Because of the simple one-stage anchor-free architecture, our proposed CECN processes CT slices at the speed of 26.6 frames per second (FPS). It runs two times faster than Faster R-CNN [17]. It largely speeds up the lesion mining pipeline and will satisfy clinical needs better.

In addition to measuring the impact of the different types of harvested lesions, we also ran the Lesion-Harvester pipeline with different LPGs by replacing CECN with CenterNet [14] and Faster R-CNN [17]. We compare different LPGs in Fig. 9 and demonstrate that our proposed pipeline generalizes well across choices of LPG. We monitor the mean recall on the validation set and stop algorithm updates after this value has converged. We plot the mean recalls of the first four iterations in Fig. 9. Compared with the baseline 36.7% recall provided by RECIST marks, Lesion-Harvesters using Faster R-CNN [17], CenterNet [14], and our proposed CECN have achieved 9.9%, 10.6%, and 11.5% improvements, respectively. In spite of the used LPG, our proposed pipeline can roughly recover 10% of unlabeled lesions. We also note that more powerful LPGs could result in faster convergence rates as well as better recovery rates.

E. Ablation Study: LPC

We validate our choice of LPC by first comparing the performance of different global-local classifiers evaluated on at the first iteration of our method. We compare our multi-view GLC-MV with two alternatives: a 2D variant that only accepts the axial view as input and a 3D version of ResNet-18 [36]. We measure results when using the raw LPG detection score (s_g), LPC classification probability (s_c), and the final lesion score ($s_{\{p|g,c\}}$) of Eq. 7. As can be seen in Fig. 8b, not all LPCs outperform the raw detection scores (s_g). However, they all benefit from the re-scoring of Eq. 7 using detection scores. Out of all options, the multi-view approach works the best. In addition to its high performance, it also has the virtue of being much simpler and faster than a full 3D approach.

TABLE III
MEAN RECALLS (%) FROM RECALL AT 80% PRECISION (R@80P) TO RECALL AT 95% PRECISION (R@95P) OF DIFFERENT LESION PROPOSAL CLASSIFIERS

Method	R@ 80P	R@ 85P	R@ 90P	R@ 95P	R@ Avg.
MLC3D [28]	42.4	41.2	39.8	38.0	40.3
ResNet-3D	44.8	43.3	42.1	40.0	42.6
GLC-3D	46.0	43.4	42.6	**40.7**	43.2
ResNet-MV	**47.9**	46.2	43.1	40.1	44.3
GLC-MV	47.4	**46.3**	**44.7**	40.4	**44.7**

In addition, we also validate the global-local feature. To do this, we compare against ResNet classifiers, both multi-view and 3D, that forego the global-local feature concatenation. We also test against a recent multi-scale approach called multi-level contextual 3-D (MLC3D) [28], which is designed for lesion FP filtering. We also implemented the 3D classifier presented in [27]; however, it did not converge with DeepLesion as it uses a much shallower network than ResNet18. For all, we evaluated using different input sizes centered at the lesion: $(24\times24\times4)$, $(48\times48\times8)$, $(64\times64\times16)$, and $(128\times128\times32)$. Here we only report results using the best scale for each, but our supplementary material contains more complete results. As can be seen in Table III, MLC3D [28] does not deliver any improvement over standard ResNet-3D. In contrast the global-local variants outperform their standard counterparts demonstrating the effectiveness of our multi-scale approach design. Again, the GLC-MV delivers the best classification performance.

F. Main Result: Detectors Trained on Harvested Lesions

While the above demonstrated that we can successfully harvest missing lesions with high precision, it remains to be demonstrated how beneficial this is. To this end, we train state-of-the-art detection systems with and without our harvested lesions and also compare against some alternative approaches to manage missing labels.

1) Using Harvested Lesions and Hard Negative Examples: After our method converged, we fused mined lesions and hard negatives, *i.e.*, $\mathbf{P}^{+}_{\mathbf{H}} = \bigcup_{i=1}^{3} \mathbf{P}^{+}_{\mathbf{H,i}}$ and $\mathbf{P}^{-} = \mathbf{P}^{-}_{\mathbf{H,3}} \cup \mathbf{P}^{-}_{\mathbf{M,3}}$, respectively. We tested CenterNet [14], Faster R-CNN [16], and our CECN (used now as a detector instead of an LPG), trained both on the original DeepLesion RECIST marks and on the data augmented with our harvested lesions. Since MULAN requires tags, which are not available for harvested lesions, we only test it using the publicly-released model. As well,

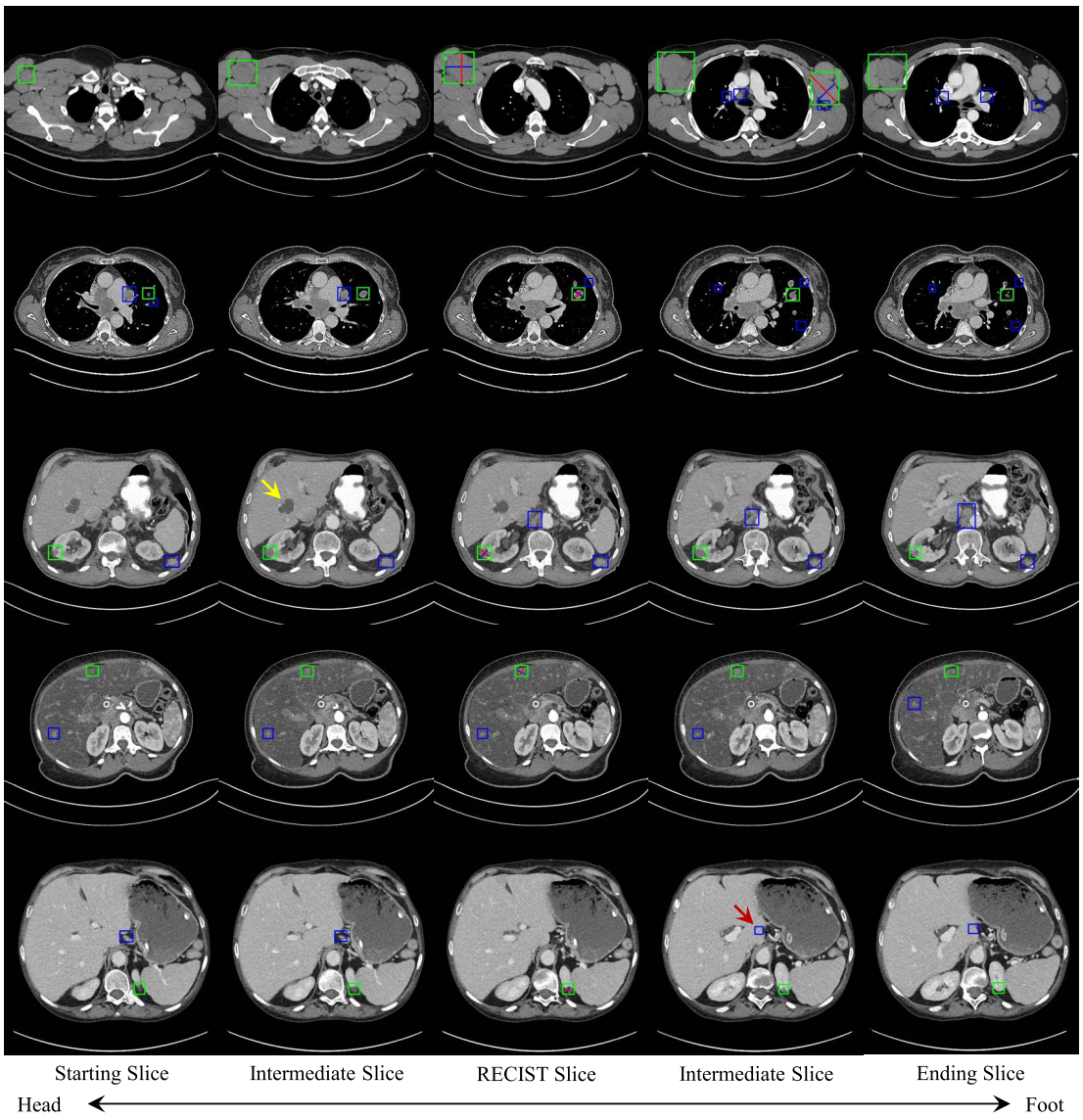


Fig. 10. Examples of 3D detection results and mined positive lesions from the harvesting set $\mathbf{V_H}$. We use green and blue boxes to show RECIST-marked and mined lesions, respectively. Each 3D detection consists of multiple axial slices and we show 5 typical slices: the starting slice, the RECIST slice, the ending slice, and two intermediate slices. We show RECIST marks as crosses with red and blue lines. We also show one failure case at the bottom row indicated by the red arrow and a lesion still remains unlabeled at the 3rd row indicated by the yellow arrow.

we also test the impact of our hard negatives ($\mathbf{L}^-$), with our proposed HNSL on CenterNet and CECN. We do not test the HNSL with Faster R-CNN, since it is not compatible with two-stage anchor-based systems. Instead, we follow Tang *et al.*,'s (ULDor) [30], which is a hard-negative approach designed for Faster R-CNN that defines an additional non-lesion class containing all hard negative examples. We test all detector variants on the unseen fully labeled $\mathbf{V_D^{test}}$ data.

As Table IV demonstrates, using the harvested lesions to train detectors can provide significant boosts in recall and precision for all methods. For instance, the extra mined lesions $\mathbf{P_H^+}$ boosts Faster R-CNN's detection performance by 4.3% in average precision. When incorporating hard negatives using the HNSL, CenterNet and CECN both benefited even more, with additional boosts of $5-8\%$ AP. In total, the harvested prospective positive and hard negative lesions are responsible for a boost of 7.2% and 8.9% in AP, for CenterNet and CECN, respectively, representing a dramatic boost in performance. Despite the great differences in architecture, incorporating harvested hard negatives as an extra non-lesion class [30] also significantly boosts Faster R-CNN performance, further demonstrating the broad impact of the Lesion-Harvester proposals. Finally, we also note that our CECN outperforms the state-of-the-art lesion detection model, MULAN [15], even when no mined lesions are used. The addition of mined lesions, which is not directly applicable with MULAN, further boosts the performance gap to 10.1%. This further validates our LPG design choices and represents an additional contribution of this work, in addition to our main focus of lesion harvesting.

2) Alternative Missing Label Approaches: We also evaluated other strategies for managing missing labels. These include

TABLE IV
EVALUATION OF DETECTORS TRAINED WITH AND WITHOUT MINED LESIONS ON $\mathbf{V}_{\mathbf{D}}^{\mathbf{TEST}}$

Method	Backbone	Input	R	$\mathbf{P}_{\mathbf{H}}^{+}$	$\mathbf{P}^{-}$	Recall (%) @ FPs=[0.125,0.25,0.5,1,2,4,8] per volume							AP
MULAN [15]	2.5D DenseNet-121	9	✓			11.43	18.69	26.98	38.99	50.15	60.38	69.71	41.8
Faster R-CNN [17]	2.5D DenseNet-121	9	✓			07.20	13.21	20.97	31.27	43.87	54.92	64.20	34.2
w/ OBHS [17]	2.5D DenseNet-121	9	✓			02.51	04.77	08.40	17.46	22.65	34.24	47.01	17.3
w/ OBSS [18]	2.5D DenseNet-121	9	✓			06.85	12.58	21.92	32.39	44.53	57.08	67.91	36.3
w/ M-OBHS	2.5D DenseNet-121	9	✓			08.53	13.67	22.89	34.46	46.85	58.31	68.08	38.3
Faster R-CNN	2.5D DenseNet-121	9	✓	✓		10.35	15.98	25.02	35.63	47.15	57.74	66.77	38.5
w/ ULDor [30]	2.5D DenseNet-121	9	✓	✓	✓	**22.95**	**30.15**	**38.09**	**46.90**	55.23	63.05	70.48	51.3
CenterNet	DenseNet-121	3	✓			12.91	19.89	26.19	35.77	45.32	56.94	67.83	41.0
CenterNet	DenseNet-121	3	✓	[20]		12.47	18.12	25.36	35.03	46.52	58.20	68.95	41.3
CenterNet	DenseNet-121	3	✓	✓		14.38	19.75	28.04	36.89	46.82	58.94	68.70	42.8
CenterNet w/ HNSL	DenseNet-121	3	✓	✓	✓	19.26	25.87	34.54	43.17	53.34	63.08	71.68	48.3
CECN	2.5D DenseNet-121	9	✓			11.92	18.42	27.54	38.91	50.15	60.76	69.82	43.0
CECN	2.5D DenseNet-121	9	✓	[20]		15.88	20.98	29.40	38.36	47.89	58.14	67.89	43.2
CECN	2.5D DenseNet-121	9	✓	✓		13.40	19.16	27.34	37.54	49.33	60.52	70.18	43.6
CECN w/ HNSL	2.5D DenseNet-121	9	✓	✓	✓	19.86	27.11	36.21	46.82	**56.89**	**66.82**	**74.73**	**51.9**

overlap-based hard sampling (OBHS) [17] and overlap-based soft sampling (OBSS) [18], both of which are designed only for two-stage anchor-based detection networks, meaning they are incompatible with our one-stage CECN. Thus, we use Faster R-CNN as baseline and compared OBHS and OBSS to training using our harvested prospective positive lesions, $\mathbf{P}_{\mathbf{H}}^{+}$. OBHS only uses proposals with small overlap to a true positive as negative examples for the second stage classifier. As Table IV demonstrates, the baseline Faster R-CNN performed with 34.2% AP, whereas using OBHS reduced the AP to 17.3%, demonstrating that simply ignoring the predominant background negatives causes large performance degradation. Thus, we also re-trained Faster R-CNN with a modified OBHS (M-OBHS), which preserves all background samples but raises weights of overlapping proposals to be twice as much as positive and standard background cases. This variant achieved 38.3% AP. As a different strategy, OBSS reduces the contributions of proposals which have small overlaps with the ground truth boxes but keeps all background cases. This strategy increased the AP to 36.3%. Finally, our method that trains Faster R-CNN with harvested prospective positive lesions achieved the best performance at 38.5% AP, with markedly higher recalls at lower tolerated FPs. This demonstrates that completing the label set with harvested lesions $\mathbf{P}_{\mathbf{H}}^{+}$ can provide greater boosts in performance than these alternative strategies.

Moving on to a more general approach, we also compare against the MGTM-SLLP [20] lesion mining algorithm. As shown in Table IV, MGTM-SLLP boosts the AP of CenterNet from 41.0% to 41.3%, while training with $\mathbf{P}_{\mathbf{H}}^{+}$ results in a much larger boost to 42.8%. Similarly, training with $\mathbf{P}_{\mathbf{H}}^{+}$ garners higher boosts in AP when using our CECN detector. Unlike MGTM-SLLP, Lesion-Harvester explores the whole CT volume, iteratively improving performance after each round, and integrates hard negatives within the training of LPG. We surmise these differences explain the increased performance even when comparisons are limited to only using the harvested positive proposals of $\mathbf{P}_{\mathbf{H}}^{+}$. Finally, as we demonstrate, when using our harvested hard-negative proposals performance can be increased even further.

V. CONCLUSION

We present an effective framework to harvest lesions from incompletely labeled datasets. Leveraging a very small subset of fully-labeled data, we chain together an LPG and LPC to iteratively discover and harvest unlabeled lesions. We test our system on the DeepLesion dataset and show that after only annotating 5% of the volumes we can successfully harvest 9,805 additional lesions, which corresponds to 47.9% recall at 90% precision, which is a boost of 11.2% in recall over the original RECIST marks. Since, our proposed method is an open framework, it can accept any state-of-the-art LPG and LPC, allowing it to benefit from future improvements in these two domains.

Our work's impact has several facets. For one, in terms of DeepLesion specifically, the lesions we harvest and publicly release enhance the utility of an already invaluable dataset. As we demonstrated, training off-the-shelf detectors on our harvested lesions allows them to outperform the current best performance on the DeepLesion dataset by margins as high as 10% AP, which is a significant boost in performance. Furthermore, we expect our harvested lesions will prove useful to many applications beyond detection, *e.g.*, radionomics studies. More broadly, our results indicate that the lesion harvesting framework is a powerful means to complete PACS-derived datasets, which we anticipate will be an increasingly important topic. Thus, this approach may help further expand the scale of data for the medical imaging analysis field.

Important contributions also include our proposed CECN LPG, which outperforms the current state-of-the-art MULAN detector and helps push forward the lesion detection topic. In addition, the introduced P3D IoU metric acts as a much better evaluation metric for detection performance than current practices. As such, the adoption of the P3D metric as a standard for DeepLesion evaluation should better rank

methods going forward. Future work should include investigating different LPG paradigms, *e.g.*, full 3D approaches, executing user studies, *e.g.*, to assess and compare with clinician performance, and exploring active learning to more efficiently choose which volumes to label. In addition, it is also crucial to measure the impact of differing labor budgets on harvesting performance. Finally, we estimate that possibly 40% of the lesions still remain unlabeled in DeepLesion. It is possible to calibrate our proposed LPG to detect lesions with above 90% recall, however, the challenge is to separate true lesions from thousands of false positives. Therefore, one other promising avenue is to model the relationship between lesions proposals for false positive reduction, which may lead to non-parametric matching and graph-based connections between instances within the dataset. We hope the public annotations and benchmarks we release will spur further solutions to this important problem.

REFERENCES

[1] P. Rajpurkar *et al.*, "MURA dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs," in *Proc. MIDL*, 2018, pp. 1–10.

[2] X. Wang *et al.*, "Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 998–1007.

[3] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3462–3471.

[4] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *J. Med. Imag.*, vol. 5, no. 3, p. 1, Jul. 2018.

[5] K. Yan *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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9261–9270.

[6] M. D. Kohli, R. M. Summers, and J. R. Geis, "Medical image data and datasets in the era of machine learning—Whitepaper from the 2016 C-MIMI meeting dataset session," *J. Digit. Imag.*, vol. 30, no. 4, pp. 392–399, Aug. 2017.

[7] H. Harvey and B. Glocker, "A standardised approach for preparing imaging data for machine learning tasks in radiology," in *Artificial Intelligence in Medical Imaging*. Cham, Switzerland: Springer Nature, 2019, pp. 61–72.

[8] J. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI*, 2019, pp. 590–597.

[9] E. Calli, E. Sogancioglu, E. T. Scholten, K. Murphy, and B. van Ginneken, "Handling label noise through model confidence and uncertainty: Application to chest radiograph classification," *Proc. SPIE*, vol. 10950, Mar. 2019, p. 1095016.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[11] T. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, vol. 8693, 2014, pp. 740–755.

[12] E. A. Eisenhauer *et al.*, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *Eur. J. Cancer*, vol. 45, no. 2, pp. 228–247, Jan. 2009.

[13] K. Yan, M. Bagheri, and R. M. Summers, "3D context enhanced region-based convolutional neural network for end-to-end lesion detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 511–519.

[14] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850

[15] K. Yan *et al.*, "MULAN: Multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 194–202.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[18] Z. Wu, N. Bodla, B. Singh, M. Najibi, R. Chellappa, and L. S. Davis, "Soft sampling for robust object detection," in *Proc. BMVC*, 2019, p. 225.

[19] G. Wang, X. Xie, J. Lai, and J. Zhuo, "Deep growing learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2831–2839.

[20] Z. Wang, Z. Li, S. Zhang, J. Zhang, and K. Huang, "Semi-supervised lesion detection with reliable label propagation and missing label mining," in *Proc. Pattern Recognit. Comput. Vis.*, 2019, pp. 291–302.

[21] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4119–4128.

[22] M. Gao *et al.*, "Segmentation label propagation using deep convolutional neural networks and dense conditional random field," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 1265–1268.

[23] J. Cai *et al.*, "Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3D mask generation from 2d RECIST," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 396–404.

[24] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut' interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[25] Q. Shao, L. Gong, K. Ma, H. Liu, and Y. Zheng, "Attentive CT lesion detection using deep pyramid inference with multi-scale booster," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 301–309.

[26] M. Zlocha, Q. Dou, and B. Glocker, "Improving RetinaNet for CT lesion detection with dense masks from weak recist labels," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 402–410.

[27] J. Ding, A. Li, Z. Hu, and L. Wang, "Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 559–567.

[28] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, Jul. 2017.

[29] V. Alex, K. Vaidhya, S. Thirunavukkarasu, C. Kesavdas, and G. Krishnamurthi, "Semi-supervised learning using denoising autoencoders for brain lesion detection and segmentation," 2016, *arXiv:1611.08664*. [Online]. Available: http://arxiv.org/abs/1611.08664

[30] Y.-B. Tang, K. Yan, Y.-X. Tang, J. Liu, J. Xiao, and R. M. Summers, "Uldor: A universal lesion detector for CT scans with pseudo masks and hard negative example mining," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Venice, Italy, Apr. 2019, pp. 833–836.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[32] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[33] F. Yang, H. Chen, J. Li, F. Li, L. Wang, and X. Yan, "Single shot multibox detector with Kalman filter for online pedestrian detection in video," *IEEE Access*, vol. 7, pp. 15478–15488, 2019.

[34] H. R. Roth *et al.*, "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2014, pp. 520–527.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[36] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.