Harvesting, Detecting, and Characterizing Liver Lesions from Large-scale Multi-phase CT Data via Deep Dynamic Texture Learning

Yuankai Huo, Jinzheng Cai, Chi-Tung Cheng, Ashwin Raju, Ke Yan, Bennett A. Landman, Jing Xiao, Le Lu, Chien-Hung Liao, Adam Harrison

Abstract—Effective and non-invasive radiological imaging based tumor/lesion characterization (e.g., subtype classification) has long been a major aim in the oncology diagnosis and treatment procedures, with the hope of reducing needs for invasive surgical biopsies. Prior work are generally very restricted to a limited patient sample size, especially using patient studies with confirmed pathological reports as ground truth. In this work, we curate a patient cohort of 1305 dynamic contrast CT studies (i.e., 5220 multi-phase 3D volumes) with pathology confirmed ground truth. A novel fully-automated and multi-stage liver tumor characterization framework is proposed, comprising four steps of tumor proposal detection, tumor harvesting, primary tumor site selection, and deep texture-based characterization. More specifically, (1) we propose a 3D non-isotropic anchor-free lesion detection method; (2) we present and validate the use of multi-phase deep texture learning for precise liver lesion tissue characterization, named spatially adaptive deep texture (SaDT); (3) we leverage small-sized public datasets to semi-automatically curate our large-scale clinical dataset of 1305 patients where four main liver tumor subtypes of primary, secondary, metastasized and benign are presented. Extensive evaluations demonstrate that our new data curation strategy, combined with the SaDT deep dynamic texture analysis, can effectively improve the mean F1 scores by > 8.6% compared with baselines, in differentiating four major liver lesion types. This is a significant step towards the clinical goal.

Index Terms—liver tumor, texture learning, deep learning

I. INTRODUCTION

LIVER cancer is the one of the most highly fatal cancer types in the world. Identifying lesions is crucial, as treatment strategies vary widely between different categories. Most liver lesions can be diagnosed non-invasively, using clinical information coupled with image modalities (e.g. ultrasound, multi-parametric magnetic resonance imagery (MRI), or multi-phase dynamic contrast computed tomography (CT)). CT is the most common choice due to its cost-effectiveness and fidelity [1], [2]. However, patients with uncertain diagnoses require invasive procedures, such as biopsies or surgery, which might lead to unpredictable hemorrhage, infections, and even death [3]. Thus, improved non-invasive tumor characterization is a crucial aim within medical imaging analysis [4], [5], [6], [7], [8], [9], [10], [11], [12]. However, challenges in data collection and analysis remain to be an unsolved problem. We aim to address these problems, proposing an end-to-end liver lesion characterization solution that incorporates effective data-curation, 3D detection, primary tumor selection, and principled deep texture learning to tackle some of the more challenging lesion characterization tasks within the clinic.

Why is this a difficult problem? Although clinical diagnoses from multi-phase CTs are usually performed through consensus between radiologists and clinicians, misdiagnoses remain (Fig 1). For instance, benign and intrahepatic cholangiocarcinomas (ICC) lesions can be misinterpreted as hepatocellular carcinoma (HCC) [13], [14], with separating HCC from hemangioma being a particularly pernicious difficulty [15], [16]. Solitary liver metastases are also difficult to differentiate from ICC [17] (Fig. 1). Underscoring this, a 2006 retrospective study [13], investigating patients receiving a liver transplant to treat image-determined HCC, discovered that, 20% had benign lesions. Even with new scoring systems, inter-reader variability is still a problem [18]. We aim to research computer-aided approaches to alleviate these issues, with many of the above challenging cases. For methodological challenges, apart from Xiao et al. [12], all previous approaches rely on manually drawn tumor region of interests (ROIs). Most importantly, there are few, if any, prior deep learning based
methods have been investigated on more than 5000 CT scans with pathology confirmed liver tumors.

Our approach. Our work is distinct from prior art in that we not only (1) propose a fully automated end-to-end framework, but also (2) test our approach on a highly challenging partially annotated dataset of 1305 multi-phase CT studies (5220 3D volumes). To meet this challenge, we develop a fully automated liver lesion classifier using multi-phase CT, articulating several innovations along the way. A key engine of our approach is a 3D generalization of CenterNet detection [19]. During inference this is crucial to localize ROIs for later downstream classification. Unique to our work, we employ 3D detection as part of a principled and minimal-labor data curation workflow, allowing us to harvest and leverage large-scale hospital data from a small seed dataset. To train our detector, we show how to employ outside single-phase (single-channel) data sources [20], [21] to boost multi-phase (multi-channel) performance. With the curated data in hand, we then execute classification in two stages. The first stage, called PTS, uses the aforementioned 3D detector and a primary tumor classifier, trained using our curated data, to filter proposals into primary tumor slices. The second stage then applies a deep texture analysis [22] on these primary tumor slices, providing a determination of lesion type. To push performance higher, we use a principled lesion-based pooling method in the texture analysis. As we demonstrate, deep texture analysis is key to achieving good performance, and, to the best of our knowledge we are the first to apply such techniques to the demanding problem of liver lesion characterization.

Our data. 1305 patients (5220 scans with four volumes per patient) were selected comprehensively represent various lesion types, if they had either a liver resection or biopsy, producing gold-standard, and not image-derived, labels. Importantly, benign lesions represents some of the most challenging cases, as except for rare cases, these are all lesions that ideally should not have been biopsied or resected in the first place, meaning they are very hard to differentiate via CT. Validating our approach, we demonstrate that our data curation pipeline and deep texture learning can each boost the four class classification accuracy by 7.4 and 3.7 percent, respectively. With the complete pipeline in place, we achieve a mean F1 score of 0.681 in classifying four tumor subtypes, compared to the top physicians’ performance of F1 = 0.690 when focusing on the binary HCC vs. non-malignant differentiation of 1-2 centimeter large tumors in CT (or 0.765 under MRI) [23]. As such, this work represents an step forward toward clinically useful computer-aided tools for liver lesion characterization.

A. Related Work

Automatic liver lesion characterization. Several previous works have proposed automatic lesion classifiers. Some only focus on distinguishing metastasis, cysts and hemangiomas apart [4], [5], [6], not yet addressing the much more difficult separation of ICC, HCC, metastasized, and hemangioma lesions. Other work tackle this more challenging scenario, e.g. using manually specified semantic features [7], texture [8], [9], and deep networks [10], [11], [12] apart from Xiao et al [12], all previous approaches rely on manually drawn tumor ROIs. In this paper, we develop an effective and automatic 3D primary tumor identification pipeline, and propose a principled multi-phase deep texture modeling scheme for precision tumor characterization.

Liver Lesion Detection. Within the deep learning era [24], liver tumor detection approaches can be categorized to segmentation-based detection [24], [25] or pure detection-based methods [20], [26], with recent work merging these two into a multi-task design [27]. Most prior work use two-stage based detection methods, e.g. based on Faster-RCNN [28] or Mask-RCNN [29]. Yet, recent advanced one-stage methods [19], [30], [31], [32], [33] have shown excellent performance while retaining simpler formulations. Among these, CenterNet [19] provides a solution well balanced in terms of
complexity and the performance. However, 2D methods can suffer from inconsistency across slices and high false positive rates. Therefore, we employ CenterNet as our detection backbone, but extend it to 3D and tailor it to handle the non-isotropic nature of medical images.

Deep Texture Learning. Texture analysis has been a canonical computer vision task for several decades. Traditional designs consist of three major steps: feature extraction [34], [35], [36], dictionary based feature encoding [37], [38], [39], and classification. Recently, classic methods, e.g., Bag-of-words [39], VLAD [40], and probabilistic version Fisher vectors (FV) [41], have been supplanted or augmented by deep texture learning [42], [22], [43], [44]. Unlike the traditional texture analysis has been widely used in medical imaging field [45], deep texture learning is not as well studied, especially for tumor analysis. Given that texture is one of the most important visual features for the radiologists [46], texture analysis should be a prominent focus for liver lesion characterization. To do this, we adapt the DeepTEN network [22], which provides an end-to-end CNN solution for material texture classification. However, SaDT manages to combine the three loss functions (Sec. II-B) and for primary tumor selection (PTS) (Sec. II-C). For the second stage, SaDT classifies ROIs extracted by PTS (Sec. II-D).

II. METHODS

Fig. 2 illustrates our overall workflow. First, we register all CT scans of the same patient together using DEEDS [47]. Then, 3D detection (Sec. II-A) is the first stage, both for data harvesting (Sec. II-B) and for primary tumor selection (PTS) (Sec. II-C). For the second stage, SaDT classifies ROIs extracted by PTS (Sec. II-D).

A. 3D Detection

As depicted in Fig. 3, we opt for a 3D extension of 2D CenterNet [19]. We choose the one-stage and anchor-free CenterNet implementation as it possesses an ideal combination of high performance and simplicity, without many of the complicating hyper-parameters required for two-stage and/or anchor-based approaches. We choose a 3D implementation, as it proved to be more reliable and natural than the 2D variant. Throughout, we follow the definition of the terms from Zhou et al. [19]. Briefly, because CenterNet follows a fully convolutional network (FCN) pipeline, the primary outcome is a 3D heatmap, \( Y \in [0, 1]^{W \times H \times D \times C} \), where \( C \) indicates the number of candidate classes while \( R \) is the downsampling factor of the prediction. The heatmap should equal to 1 at lesion centers and 0 otherwise. As FCN backbone, we extend the stacked Hourglass-104 network [48] in the original CenterNet formulation [19] to a 3D variant. Following standard practice [30], [19], the ground truth of the 3D target center point can be modeled as a 3D Gaussian kernel. However, because medical images are typically non-isotropic (the physical voxel spacing in \( D \), e.g., 5 mm, is larger than that in \( W \) and \( H \), e.g., 1 mm), we propose to utilize a non-isotropic Gaussian kernel:

\[
Y_{\hat{x} \hat{y} \hat{z}} = \exp \left( -\frac{(\hat{x} - \tilde{p}_x)^2}{2\sigma^2_p} + \frac{(\hat{y} - \tilde{p}_y)^2}{\gamma^2} + \frac{\hat{z} - \tilde{p}_z)^2}{\gamma^2} \right),
\]

where \( \tilde{p}_x, \tilde{p}_y, \) and \( \tilde{p}_z \) are the downsampled target center points \( \hat{p} = \left( \frac{p}{R} \right) \) and \( \sigma_p \) is the kernel standard deviation. \( \gamma_x, \gamma_y, \) and \( \gamma_z \) are the resolution coefficients to compensate for resolution differences. The corresponding pixel regression loss \( L_{k} \), and the \( \ell_1 \)-norm offset prediction loss \( L_{off} \), are formulated identically as Zhou et al. [19].

Given a 3D bounding box \((x_1, y_1, z_1, x_2, y_2, z_2)\), the center point is modeled as \( p = \left( \frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}, \frac{z_1 + z_2}{2} \right) \) [19]. The true bounding box size is computed as \( s = (x_2 - x_1, y_2 - y_1, z_2 - z_1) \). For a predicted bounding box \( \hat{s} \), the same \( L1 \) loss at the center point is calculated:

\[
L_{size} = \frac{1}{N} \sum_{k=1}^{N} \| \hat{s}_k - s_k \|_1, \tag{2}
\]

We used same hyper-parameter settings as Zhou et al. [19] to combine the three loss functions \( (L_k, L_{off}, L_{size}) \) and to set \( \sigma_p \). Finally, to fit volumes into memory, we first apply an FCN-based liver segmentation, trained from public data [21], [49], [50], [51], [52]. We then crop around this liver mask and resample the resulting region to \( 176 \times 256 \times 48 \).
To implement a fully automatic pipeline, we rely on localization as a first stage. While we can directly use the 3D detection of Sec. II-A, the data curation of Sec. II-B demonstrates that additional filtering is required to extract reliable ROIs. For this reason, we implement a primary tumor selection (PTS) pipeline for more precise ROI extraction.

Fig. 2 illustrates the PTS workflow. As can be seen, the first step is to apply the 3D CenterNet detection to generate primary lesion candidates. We train this implementation using public datasets, our labeled dataset $D_\ell$, and the new harvested lesions in $D_u$ from Sec. II-B. From this detector, we extract the 10 leading candidate 3D bounding boxes per volume. One or more of these candidates ideally overlaps with primary lesion(s); however, many will likely also overlap with false positives, e.g., blood vessels, or non-significant lesions, e.g., cysts (Fig. 5). These must be filtered out.

For reasons Sec. II-D will explain, and our results support, downstream classification is performed in 2D. Thus, any lesion candidate filtering must both select key 2D slices and, from these, select the most likely one belonging to a primary tumor. We call this key slice filtering (KSF). As Fig. 5 illustrates, we perform KSF by first applying a binary lesion segmentation network, trained using $D_\ell$, on each whole slice of the volume. Then for each 3D candidate, we choose the slice corresponding to the greatest prediction area as the key slice. While in principle any segmentation network can be incorporated, we choose to finetune the 2.5D segmentation head of the MULAN lesion detector [27]. We opt for this model for two reasons: 1) because the segmentation head is pretrained on DeepLesion [20], it should possess features with high lesion affinity; and 2) the 2.5D segmentation allows us to incorporate sufficient background context without downsampling and compromising the resolution.

With key slices selected for each candidate, these must then be separated into primary and non-primary tumors. To do this, we perform classification on the multi-phase key slices (Fig. 5). Fortunately, thanks to the data curation pipeline in Sec. II-B, we have ample true- and false-positive candidates to train such a classifier. In this way, the data curation and KSF steps can be viewed as hard-negative mining and hard-negative filtering, respectively. Any high-performing classifier can be used for this step, e.g., well-established appearance based convolutional neural networks (CNNs) or even the texture-based classifier we outline next in Sec. II-D.
The visual descriptors, counting, or soft-counting, the codewords found in a set of 2D models also allows us to use pretrained models, which is recent DeepTEN model [22] to create a SaDT network. Using deep texture learning workflow, adapting and enhancing the high inter-slice thickness of most CTs, practical limitations hamper such an approach. Namely, it is theoretically possible to perform 3D texture modelling, particularly prior to deep learning. For this reason, we but texture-based approaches have experienced success [8], [9], particularly prior to deep learning. For this reason, we

Fig. 6. The proposed SaDT deep texture learning network.

D. Texture Classification

With lesion ROIs localized and filtered using the PTS of Sec. II-C, the next step is to differentiate them into HCC, ICC, metastasis, or benign (hemangioma, focal nodular hyperplasia, and adenoma). Standard CNNs can serve this function, but texture-based approaches have experienced success [8], [9], particularly prior to deep learning. For this reason, we investigate the impact deep texture based approaches. While it is theoretically possible to perform 3D texture modelling, practical limitations hamper such an approach. Namely, the high inter-slice thickness of most CTs, e.g. 5 mm in our data, is much too course to capture texture. Instead, we apply a 2D deep texture learning workflow, adapting and enhancing the recent DeepTEN model [22] to create a SaDT network. Using 2D models also allows us to use pretrained models, which is another key benefit. Fig. 6 depicts our SaDT model.

Following Zhang et al. [22], texture modeling relies on counting, or soft-counting, the codewords found in a set of visual descriptors. The visual descriptors, \( F \) = \{\( f_1, \ldots, f_M \)\}, are generated from the features of an encoding FCN, where \( M \) is the number of spatial locations in the activation map. The residuals of each feature compared to a set of \( K \) codewords \( C = \{c_1, \ldots, c_K\}\) are then computed:

\[
r_{ik} = f_i - c_k,
\]

where \( i = 1, \ldots, M \) and \( k = 1, \ldots, K \). Classically these codewords have been fixed, but recent approaches allow these to be learned [22]. This is the approach we take.

The set of all \( M \times K \) residuals must now be aggregated into a global size \( K \) feature describing the overall texture. Before aggregation, the weights of each encoded residual are computed based on a “soft-assignment” [22]. Traditionally, only a single weight is assigned to each descriptor \( x_i \), called “hard-assignment”, but this makes the process non-differentiable. Therefore, the weight of each \( r_{ik} \) is calculated using a softmax:

\[
a_{ik} = \frac{\exp(-s_k r_{ik})}{\sum_{j=1}^{K} \exp(-s_j r_{ij})},
\]

where \( s_k \) is the smoothing factor for each cluster center \( c_k \). The smoothing factors are also designed as learnable parameters, which are updated during the training procedure.

Given the residuals as well as their weights, any arbitrary set of features, \( F \), can be encoded into a fixed-size \( K \) global vector [22]:

\[
e_k = \sum_{i=1}^{M} a_{ik} r_{ik},
\]

where all spatial locations are aggregated together. However, different from the canonical scenarios in deep texture learning, the sizes of the tumors are heterogeneous across different cases (Fig. 6). A simple resize operation is not optimal, since it alters the intrinsic physical resolutions to medical imaging. To address this spatial variations, we propose a spatial adaptive aggregation that modifies Eq. 5 to

\[
e_k = \sum_{i=1}^{M} a_{ik} r_{ik} \delta_i,
\]

where each \( \delta_i \) is a 0-1 binary value, indicating if the corresponding visual descriptor should be aggregated. In our implementation, \( \delta_i \) is the tumor mask generated using the segmentation network of Sec. II-C. Despite its straightforwardness, this process, denoted SaDT modelling, can result in significant performance improvements. Finally, following standard texture analysis practices [22], [41], [54], the final output vector is normalized using an \( l_2 \)-norm. Since the training cohorts are highly unbalanced, for all experiments, the same weighted focal-loss (\( \gamma = 2 \)) [31] is used as the classification loss function to reconcile the overfitting to the dominating class. The weights were set to 5 for HCC, 2 for metastasis, and 1 for remaining classes. The same stochastic gradient descent (SGD) optimizer with learning rate = 0.004 [22] is used for all experiments.

III. EXPERIMENTS

A. Data

We collected multi-phase dynamic CT scans from the archives of Anonymized from patients with liver lesions that received surgical resection or percutaneous biopsy in the period between June 2003 to April 2018. CT scans were acquired from the hospital archives within one month before the invasive procedure. This resulting in 1305 studies, with all CT scans having standard 5 mm slice thickness. Most patients underwent resection (979, 72.2%), with the remainder receiving biopsy (377, 27.8%) for pathological confirmation. The dataset was then split into training (1197) and testing (108) studies, keeping the distribution of lesion types the same. From the training set, the lesions of 196 studies were manually segmented under the supervision of a trained clinician (Patho-200), with the remainder only having image-level labels (Patho-1000). This corresponds to \( D_t \) and \( D_u \), respectively. Because of the paucity of certain lesion types, during model development both Patho-200 and Patho-1000 were split into 5-folds for cross validation for more stable assessments. Tab. I provides more data details. The in-plane resolution of all multi-phase scans...
TABLE I
THE DATA USED IN THIS STUDY INCLUDE THE PUBLICLY AVAILABLE LITS AND DEEPLesION DATASET. PATHO IS OUR IN-HOUSE MULTI-PHASE DYNAMIC CT DATASET, WHICH CONTAINS NON-CONTRAST (NC), ARTERIAL (A), VENOUS (V), AND DELAYED (D) PHASES. PATHO-200, PATHO-VAL, AND PATHO-TEST ARE MANUALLY ANNOTATED, WHILE PATHO-1000 IS ANNOTATED USING OUR SEMI-AUTOMATED DATA HARVESTING. ALL TRAINING DATA ARE USED IN DETECTION AND TUMOR HARVESTING, WHILE ONLY PATHO-200 AND PATHO-1000 ARE USED IN TUMOR CHARACTERIZATION.

| Dataset          | Annotation | Studies | Phases | Categories          |
|------------------|------------|---------|--------|---------------------|
| LITS             | 3D Seg.    | 131     | V      | Liver Tumor (131)   |
| Train & Validation | DeepLesion (DL) | 2D RECIST | 2372 | Unknown | Liver Lesion (2372) |
| Patho-200        | 3D Seg.    | 196     | NC, A, V, D | HCC(52), ICC(40), Benign(55), Meta(49) |
| Patho-1000       | 3D B.Box   | 1001    | NC, A, V, D | HCC(921), ICC(4), Benign(2), Meta(74) |
| Test             | Patho-test | 3D B.Box | 108 | NC, A, V, D | HCC(60), ICC(8), Benign(16), Meta(24) |

![Fig. 7. Labor time using different curation strategies.](image)

are 512 × 512 with averaging 42 slices on z directions. The average voxel size is 0.69 mm × 0.69 mm × 5 mm.

Since only Patho-200 has localizations, we use the data curation pipeline from Sec. II-B to extract tumor ROIs from Patho-1000. In total, this required employing our mouse-click QA on 1064 lesion candidates (∼1 min/study) and executing 3D bounding box annotations (∼15 min/study) on 193 volumes whose primary tumors were not captured by the proposals. The labor savings provided by the curation is presented in Fig. 7.

### B. Tumor Characterization

We first validate our choice of SaDT model for lesion characterization, assuming for now that manually drawn ROIs are available. As comparison, we test against both 3D and 2D classification networks. For a fair comparison, we use an ImageNet pretrained ResNet-50 backbone for all 2D networks. For 3D networks, we test ResNet3D [55], [56], while for 2D networks we test standard ResNet [56] and also the DeepTEN network [22]. The same single top-1 most confident tumor (largest volume size) for a given scan is used to train both 3D and 2D classifiers. For 2D, the top-5 largest area slices are employed to train the network, while the top-1 largest slices are used during testing. For 3D, the entire top-1 mask center cropped tumor volume is used for both training and testing. Such design ensures only one prediction for each patient, which enables the validation using image level ground truth.

For DeepTEN, the tumor region within the bounding box is resized to 256 × 256 following typical texture analysis practices [22]. For the SaDT method, the bounding box location is encoded as a binary mask channel for spatial adaptive aggregation (Fig. 6). When compared against ResNet, DeepTEN reveals the impact of applying texture analysis vs. standard appearance-based CNNs. Comparisons against SaDT reveal the impact of our spatially adaptive approach to deep texture learning.

We also perform ensembling experiments, which applies all five models trained in cross validation to the test set with majority voting. We do not perform ensembling using 3D methods due to the heavy computational burdens. However, when using single-model inference, we ensure that the same train-validation fold is used across all methods. We evaluate the methods using overall accuracy (Acc.) as well as one-vs.all F1 scores ($F_1$).

Tab. II presents the results. Some immediate conclusions can be drawn. First, the inclusion of Patho-1000 significantly boosts performance for most models, validating our semi-automatic data curation pipeline. Second, 2D models tend to perform better than 3D variants. Third, from the 2D models, texture-based variants outperformed appearance-based ones, demonstrating the value of applying multi-phase deep texture learning for lesion characterization. Fourth, unsurprisingly, ensembling aided all models. Finally, SaDT achieves the highest, or nearly highest performance, across all metrics except the F1 benign. While 2D ResNet exhibits the highest performance for this metric, its other metrics are much worse, whereas SaDT provides much more stable performance across lesion types. Thus, these results demonstrate the benefits of applying SaDT to lesion analysis. Given that these include some of the most difficult cases to differentiate, these metrics demonstrate the promise of our lesion characterization strategy.

### C. Tumor Detection

While Sec. III-B demonstrates the value of our texture learning approach, SaDT alone does not provide a complete pipeline. For this reason, we also validate our detection network. To do this, we test both the 2D CenterNet [30] as well as our proposed CenterNet-3D variant. For 2D detection, CT slices are first normalized to 0.8 mm × 0.8 mm pixel size, and then either center cropped or zero padded to 512 × 512. Three consecutive slices (with 2 mm slice thickness) are used as three input channels. We use the Hourglass [48] network, pretrained on MS-COCO [57], as backbone. Thus the 2D CenterNet represents a strong baseline. For the CenterNet-3D, liver segmentation, cropping, and resizing is performed as described in Sec. II-A. The hyper parameters of both CenterNet-2D [30] and the CenterNet-3D are optimized based on the validation dataset. Since the goal is to have the detector capture the primary tumor in its top 10 candidates, with the KSF process then filtering these, we measure performance by the percentage of patients with at least one primary tumor...
detected (P1TD). This metric can be divided into whether the first 1 (P1TD-1) or first 10 (P1TD-10) candidates captured a primary tumor. P1TD-10 measures whether we are able to even capture a primary tumor in our candidate cases. But since we break ties in KSF using detection scores, it is also important to measure P1TD. We test each variant when trained only on (A) Patho-200, (B) Patho-200 + public data [19, 20], and (C) Patho-200 + Patho-1000 + public data, with latter measuring the impact of our data curation on detection performance.

Tab. [IV] presents the detection results. As can be seen, incorporating the harvested lesions benefits all models, further validating our data curation pipeline. Comparing the two variants, CenterNet-3D exhibits lower P1TD-10 accuracy while producing higher P1TD-1 accuracy. Given the complex interaction, it is difficult to assess which balance between the two metrics is best. However, as Sec. [III-F] will demonstrate, when incorporated within our complete automatic pipeline, CenterNet-3D corresponds to higher overall performance, making it our preferred choice.

D. Key Slice Filtering

The first stage of our automatic pipeline, PTS, consists of both detection and KSF, where the latter selects primary tumor slices from the detection candidates. A key component of KSF, in turn, is a classification network to filter out spurious detections. As such, it is important to characterize this classifier performance. To do this, we measure classifier performance on slices extracted from the test set, measuring a standard ResNet vs. our SaDT. As can be seen in Tab. [IV], both models achieve near 90% accuracy. However, the SaDT wins out, further validating the model. For this reason, we employ the SaDT for both KSF and for downstream lesion classification.

E. Unified Automatic Framework

Finally, we test the contribution of each component to a unified and fully automatic pipeline. We compare against an upper bound where lesion ROIs are manually drawn. Note that, with the PTS model, we now select the top-1 tumor and top-1 slice using the detection scores and KSF predictions since the manual segmentation is not available for testing data as Sec. [III-B]. We also compare against variants without KSF, meaning we just select the lesion candidate with the highest detection score. As Tab. [IV] shows, both CenterNet-3D and KSF contribute significant performance improvements, validating our choices. While performance gaps remain compared to a manual approach, these results demonstrate that our lesion characterization pipeline can provide an effective fully-automated solution. Given the challenging nature of our dataset, the results are highly encouraging.

F. Comparing with Human Physicians

Even not a direct comparison, the proposed SaDT achieves a mean $F_1 = 0.681$ in classifying four tumor subtypes, compared to the top physicians’ $F_1 = 0.690$ when focusing
on the binary HCC vs. non-malignant differentiation of 1-2 centimeter large tumors in CT (or 0.765 under MRI) [23], which is a well-cited clinical study to comprehensively evaluate physician performance on liver tumor characterization using CT, MRI, and CEUS. Both [23] and we are using 4-phase CT, so the basic imaging protocol is comparable in both studies. (1) However our patient population selection is more challenging by using only pathologically confirmed cases (indicates our samples are biased towards patients with high diagnosis uncertainty by radiology), while [23] only sampled from a more general population (cirrhotic patients with nodules). (2) In this work, we are performing fine-grained four class diagnosis (HCC vs. ICC vs. Meta vs. Benign), where [23] only does binary classification (HCC vs. others). (3) When evaluating F1 (HCC vs. remaining) in Table 2, our methods achieved F1 = 0.797 (Manual+SaDT) and 0.763 (Fully automated), which is considerably higher than F1 = 0.690 using CT in [23], and even comparable with F1 = 0.765 with an advanced MRI protocol in [23].

IV. DISCUSSION

To the best of our knowledge, we are the first to use deep dynamic CT texture learning for liver tumor characterization. While performance has been pushed on our challenging dataset, there is room for improvement. Increased data collection to handle imbalanced training classes (e.g. ∼77% are HCC), would certainly be of benefit. Multi-center data collection could be a viable strategy, with our curation pipeline being a key component. Another important direction is incorporating other radiomic features (e.g. shape and size) [46], [53]. Finally, it is critical to model prediction confidences [59], empowering our system to more reliably provide diagnoses when it is safe to do so. These directions should push progress even further toward minimizing risky and painful invasive biopsy and surgical procedures.

Our study is by far one of the largest liver tumor characterization work ever reported on 1305 pathologically-confirmed patients. The patient-wise ground truth labels are converted to tumor-wise labels using our proposed data harvesting approach. Our novel SaDT model is able to adaptively aggregate over tumor size, location, and deep dynamic texture appearance (in multiphase imaging) as a whole framework of tumor detection/localization/diagnosis, while [22] mainly exploited texture cues for the material based 2D image recognition.

In this study, approximately 10% of the cases were used for testing, which may seem small. However, in absolute numbers our test set consists of 108 patients (432 multi-phase CT scans), which is already considerably larger than previous well known clinical study [23], which studied a total of 101 nodules...found in 84 patients. Collecting more was hindered by the labor costs to obtain clinical-quality pixel-level gold standard annotations, requiring our collaborating physicians to confirm/annotate each tumor in 3D/4D from both radiological and pathological reports. On the other hand, our training dataset is weakly annotated, which allowed for its comparatively larger size.

We present a fully-automated and unified framework for liver tumor characterization from multi-phase dynamic CT. We use a 3D one-stage anisotropic detection network, and employ it for highly effective semi-automated tumor harvesting and also tumor ROI extraction during inference. We investigate liver tumor characterization from a deep texture learning perspective and propose the spatially adaptive deep texture (SaDT) network. To bridge 3D detection and 2D classification, a key slice filtering (KSF) process is proposed that cascades segmentation and classification. Experiments used a dataset of 1305 multi-phase CTs, all with pathology-proven diagnoses, that represent some of the most challenging lesion characterization cases. Results demonstrate that our 3D detection, KSF, and SaDT model all provide significant performance contributions.

V. CONCLUSION

We present a novel fully-automated multi-stage deep texture learning framework for liver tumor characterization under multi-phase dynamic CT. To the best of our knowledge, we are the first to use deep dynamic CT texture representation for liver tumor characterization. We employ a 3D anchor-free anisotropic lesion detection network for highly effective semi-automated tumor harvesting and tumor ROI extraction during inference. We exploit the liver tumor characterization from a deep texture learning perspective by the proposed the spatially adaptive deep texture (SaDT) network. To bridge 3D detection and 2D classification, a key slice filtering (KSF) process is proposed to filter out hard-negatives. Experiments are conducted on a curated dataset of 1305 multi-phase CT studies (with pathology-proven diagnoses) that represent some of the most challenging lesion characterization cases. Results validate that our 3D lesion detection, KSF, and SaDT models all provide significant performance contributions.

REFERENCES

[1] M. R. Oliva and S. Saini, “Liver cancer imaging: role of ct, mri, us and pet,” Cancer imaging, vol. 4, no. Spec No A, p. S42, 2004.
[2] D. P. Burrowses, A. Medellin, A. C. Harris, L. Milot, and S. R. Wilson, “Contrast-enhanced us approach to the diagnosis of focal liver masses,” Radiographics, vol. 37, no. 5, pp. 1388–1400, 2017.
[3] A. Grant and J. Neuberger, “Guidelines on the use of liver biopsy in clinical practice,” Gut, vol. 45, no. suppl 4, pp. IV1–IV11, 1999.
classification,” *International Journal of Computer Vision*, vol. 127, no. 1, pp. 74–109, 2019.

[45] G. Castellano, L. Bonilha, L. Li, and F. Cendes, “Texture analysis of medical images,” *Clinical radiology*, vol. 59, no. 12, pp. 1061–1069, 2004.

[46] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn *et al.*, “Artificial intelligence in cancer imaging: clinical challenges and applications,” *CA: a cancer journal for clinicians*, vol. 69, no. 2, pp. 127–157, 2019.

[47] M. P. Heinrich, M. Jenkinson, M. Brady, and J. A. Schnabel, “Mrf-based deformable registration and ventilation estimation of lung ct,” *IEEE transactions on medical imaging*, vol. 32, no. 7, pp. 1239–1248, 2013.

[48] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[49] E. Gibson, F. Giganti, Y. Hu, E. Bon-Mati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, “Multi-organ abdominal ct reference standard segmentations,” *This data set was developed as part of independent research supported by Cancer Research UK (Multidisciplinary C28070/A19985) and the National Institute for Health Research UCL/UCL Hospitals Biomedical Research Centre*, 2018.

[50] A. E. Kavur, N. S. Gezer, M. Barış, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkam, B. Baydar *et al.*, “Chaos challenge–combined (ct-mr) healthy abdominal organ segmentation,” *arXiv preprint arXiv:2001.06535*, 2020.

[51] T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes *et al.*, “Comparison and evaluation of methods for liver segmentation from ct datasets,” *IEEE transactions on medical imaging*, vol. 28, no. 8, pp. 1251–1265, 2009.

[52] L. Soler, A. Hostettler, V. Agnus, A. Charnoz, J. Fasquet, J. Moreau, A. Osswald, M. Bouhadjar, and J. Marescaux, “3d image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database,” *IRCAD, Strasbourg, France, Tech. Rep*, 2010.

[53] K. Yan, Y. Peng, V. Sandfort, M. Bagheri, Z. Lu, and R. M. Summers, “Holistic and comprehensive annotation of clinically significant findings on diverse ct images: learning from radiology reports and label ontology,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8523–8532.

[54] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “Netvlad: Cnn architecture for weakly supervised place recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[55] K. Hara, H. Kataoka, and Y. Satoh, “Learning spatio-temporal features with 3d residual networks for action recognition,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3154–3160.

[56] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[57] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[58] R. L. Siegel, A. Jemal, R. C. Wender, T. Gansler, J. Ma, and O. W. Brawley, “An assessment of progress in cancer control,” *CA: a cancer journal for clinicians*, vol. 68, no. 5, pp. 329–339, 2018.

[59] F. C. Ghesu, B. Georgescu, E. Gibson, S. Guendel, M. K. Kalra, R. Singh, S. R. Digumarthy, S. Grbic, and D. Comaniciu, “Quantifying and leveraging classification uncertainty for chest radiograph assessment,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 676–684.