Recognizing Focal Liver Lesions in CEUS With Dynamically Trained Latent Structured Models

Xiaodan Liang, Liang Lin*, Qingxing Cao, Rui Huang, and Yongtian Wang

***Abstract*—This work investigates how to automatically classify Focal Liver Lesions (FLLs) into three specific benign or malignant types in Contrast-Enhanced Ultrasound (CEUS) videos, and aims at providing a computational framework to assist clinicians in FLL diagnosis. The main challenge for this task is that FLLs in CEUS videos often show diverse enhancement patterns at different temporal phases. To handle these diverse patterns, we propose a novel structured model, which detects a number of discriminative Regions of Interest (ROIs) for the FLL and recognize the FLL based on these ROIs. Our model incorporates an ensemble of local classifiers in the attempt to identify different enhancement patterns of ROIs, and in particular, we make the model reconfigurable by introducing switch variables to adaptively select appropriate classifiers during inference. We formulate the model learning as a non-convex optimization problem, and present a principled optimization method to solve it in a dynamic manner: the latent structures (e.g. the selections of local classifiers, and the sizes and locations of ROIs) are iteratively determined along with the parameter learning. Given the updated model parameters in each step, the data-driven inference is also proposed to efficiently determine the latent structures by using the sequential pruning and dynamic programming method. In the experiments, we demonstrate superior performances over the state-of-the-art approaches. We also release hundreds of CEUS FLLs videos used to quantitatively evaluate this work, which to the best of our knowledge forms the largest dataset in the literature. Please find more information at "http://vision.sysu.edu.cn/projects/fllrecog/".**

***Index Terms*—Cancer recognition, CEUS, computer-aided diagnosis, focal liver lesions.**

I. Introduction

LIVER cancer is known as the fifth most common cancer and second cause of cancer-related death reported by World Health Organization (WHO) [1]. Focal Liver Lesions (FLLs) are or cystic masses that are identified as an abnormal part of the liver [2]. Hepatocellular carcinoma (HCC) is the most common type of liver cancer [3]. Early diagnosis of FLLs plays a key role in successful cancer treatment. Ultrasound imaging is often used for cancer diagnosis due to its low cost, efficiency and non-invasiveness. However, conventional ultrasound may produce vague images and fail in detecting small masses due to its low sensitivity and signal-to-noise ratio [4]. Thus, further imaging is required, using Tomography (CT) or Magnetic Resonance Imaging (MRI) techniques, while their major disadvantages include the cost of examination, the cumbersome equipment and the exposure to the ionizing radiation used in CT [5]. Recently, Contrast-Enhanced Ultrasound (CEUS) was proposed to study the enhancement dynamics of FLLs in real time, by assessing the FLL enhancement patterns, i.e. the intensity changes of the FLL areas relative to that of their adjacent healthy liver tissues (parenchyma) [6]. CEUS has since markedly improved the accurate diagnosis of the FLLs [7], [8].

In this paper, our method focuses on three specific types of FLLs: one malignant FLL (i.e. Hepatocellular Carcinoma (HCC)) and two benign FLLs (i.e. Hemangioma (HEM) and Focal Nodular Hyperplasia (FNH)). We use the *classification* term to indicate the distinction between benign or malignant, and the *characterization* term to indicate the distinction between the specific FLL types (i.e. HCC, HEM, FNH). As reported in the medical guidelines [6], radiologists recognize FLLs by observing the enhancement patterns during all vascular phases (arterial, portal venous, late) [7]. With regard to classification problem, as shown in Fig. 1, portal venous and late phase may assist in differentiating malignant FLL (HCC) and benign ones (HEM and FNH). HCC is hypo-enhancing (e.g. the wash out phenomenon) while HEM and FNH are iso- or hyper-enhancing in the portal venous and late phase. Arterial phase, on the other hand, provides essential information to distinct between the specific FLL types. For example, as the malignant lesions on our dataset, more than 97% HCC cases often show homogeneous hyper-enhancement (shown as "HCC1" in Fig. 1), and the remaining 1–3% cases may be inhomogeneous or rim enhancement in larger nodules (>5 cm), which contain regions of necrosis ("HCC2"). Among the benign FLLs, typical HEM cases show peripheral nodular enhancement ("HEM1") while high flow HEM cases show rapid homogeneous hyper-enhancement ("HEM2"). FNH is often visualized with spoke-wheel vascular pattern hyper-enhancement ("FNH1") or homogeneous enhancement ("FNH2") of the whole lesion. The clinical guidelines also advise the radiologists to consider these diverse characteristics of FLLs in CEUS for diagnoses [6]. These complicated enhancement patterns of different FLL types (i.e. HCC, HEM and FNH) make accurate diagnosis extremely difficult.

Manuscript received July 12, 2015; revised October 09, 2015; accepted October 13, 2015. Date of publication October 26, 2015; date of current version March 01, 2016. This work was supported in part by the China 863 program under Grant.2013AA013801, in part by the Guangdong Natural Science Foundation under Grant S2013050014548 and Grant 2014A030313201, in part by the Program of Guangzhou Zhujiang Star of Science and Technology under Grant 2013J2200067, in part by the Special Project on the Integration of Industry, Education and Research of Guangdong Province under Grant 2012B091000101, in part by the National Natural Science Foundation of China under Grant 61370185, and in part by the Fundamental Research Funds for the Central Universities under Grant 14lgjc11. *Asterisk indicates corresponding author.*

X. Liang and Q. Cao are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China.

*L. Lin is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China and also with the SYSU-CMU Shunde International Joint Research Institute, Shunde 528300, China (e-mail: linliang@ieee.org).

R. Huang is with the NEC Laboratories China, Beijing 100084, China..

Y. Wang is with the School of Optoelectronics, Beijing Institute of Technology, Beijing 100081, China.

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

Digital Object Identifier 10.1109/TMI.2015.2492618

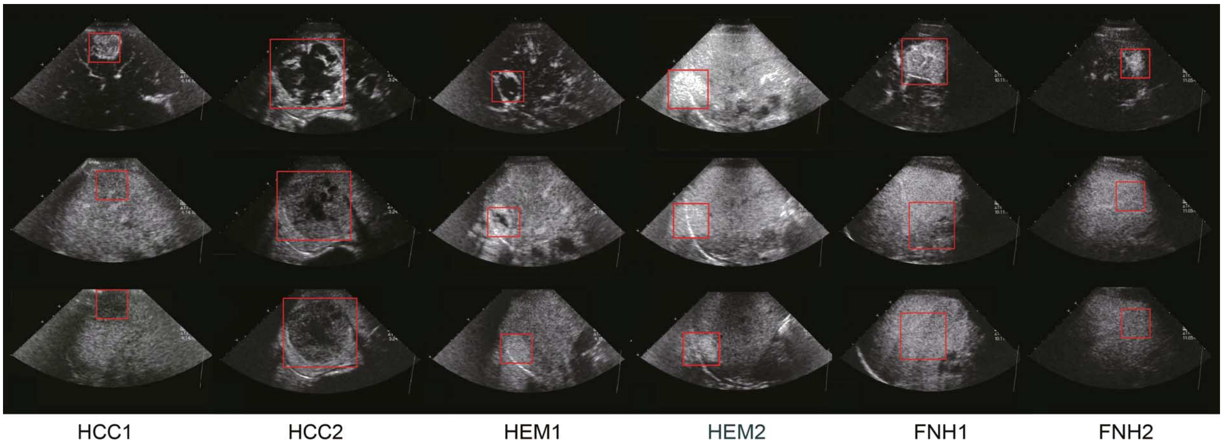

Fig. 1. Diverse enhancement patterns of FLLs in CEUS videos. We consider three common types: Hepatocellular carcinoma (HCC), Hemangioma (HEM) and Focal Nodular Hyperplasia (FNH). Each FLL can be recognized by the enhancement patterns of Regions of interest (ROIs, indicated by the red boxes) in the arterial phase (first row), portal venous phase (second row) and late phase (third row). Two instances of each FLL type (i.e. HCC, HEM, FNH) are shown to visualize the diverse enhancement patterns of FLLs.

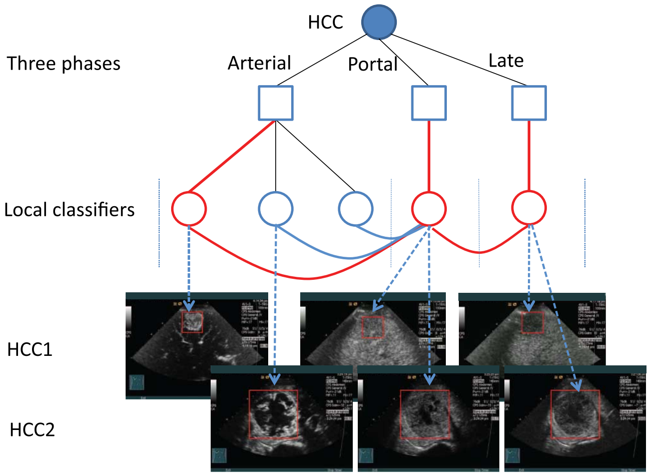

Fig. 2. An example of latent structured model generated by our approach. From bottom to top, the local classifiers (denoted by the solid circles) at the bottom for localizing the candidate ROIs in each phase; then the most appropriate classifiers are selectively combined to conduct the FLL recognition. We utilize multiple local classifiers to capture the diverse enhancement patterns in the arterial phase. The curves between the local classifiers are used to incorporate temporal transitions between pairwise ROIs. The bold red circles and curves represent the selection of local classifiers during the inference.

Currently, the accuracy of diagnosis highly depends on the expertise of the radiologists. The experts often make a lot effort on reviewing the whole CEUS video back and forth to find the lesions and their patterns, based on numerous diagnosis guidelines [3], [6]. Thus, Computer-Aided Diagnosis (CAD) systems are proposed as a “second reader” to characterize various types of tumors [9], [10], and in particular, FLLs using ultrasound images [11], [12]. However, CAD systems for analyzing FLLs in CEUS videos are still rarely developed, mainly due to the large variations of enhancement patterns in the three vascular phases. Some preliminary CAD attempts [13]–[19] rely on manually pre-defined regions of FLLs, with the exception of Bakas *et al.* [16] who proposed a method to automatically identify the frame where an FLL is best distinguished.

Different from the prior works, we propose to recognize the FLL types by identifying Regions-of-Interest (ROIs) in CEUS videos, i.e. detecting one ROI for each phase. The ROIs in all temporal phases of a CEUS video are also generated for further assisting the clinicians. We develop a novel latent structured model that incorporates an ensemble of local classifiers to capture diverse enhancement patterns of ROIs. In particular, we make the model reconfigurable by introducing latent switch variables to select appropriate local classifiers for optimal combination, inspired by recently proposed And-Or graph models [20], [21]. That is, we deploy a set of local classifiers for a phase, and each local classifier localizes one candidate ROI. Thus, the set of discriminative ROIs can be produced by searching for the optimal ROI in each phase. We treat the temporal and spatial locations, area size and classifier selection of each ROI as latent structures of our model, which are automatically determined during inference for different CEUS videos.

To be accommodated with the real circumstances in FLL diagnosis, such as the instability of operation and the motion of organs, we propose an effective feature representation to describe the appearance of ROIs, which characterizes the appearances of ROIs from different aspects: (i) the region inside the lesion, (ii) the morphology of the lesion, and (iii) the tissue area surrounding the lesion. In addition, the appearance similarity between two ROIs of different temporal phases are incorporated as a constraint in our model.

Training our latent structured model is another innovation of this work. There are several challenges for this task. First, the training data are not manually labeled (e.g. layouts of ROIs), and the number of local classifiers for different FLL types is unknown, which is related to the intrinsic pattern variants for each FLL type. It is difficult to automatically generate the latent structures without any supervision. In previous works, researchers utilized elaborative annotations to manually determine the model configurations [22], [23]. Second, simultaneously predicting the ROI layouts and FLL types is difficult because these two tasks depend on each other. In particular, recognizing FLL types is founded on the extracted features of the predicted ROIs and in turn the predicted ROI layouts are generated by maximizing the recognition accuracy. In our approach, we overcome these problems by proposing a novel principled optimization algorithm, inspired by the structural learning method [21], [24]. The proposed learning algorithm iteratively optimizes the latent structures along with the model parameter learning. To determine the latent structures, we propose a two-step procedure. First, we apply the current updated model on all FLL training videos, in which each local classifier localizes one ROI by maximizing the detection score. This step can generate a batch of candidate ROIs for each training video. Second, we reconfigure the model structures by reproducing local classifiers for each temporal phase, based on jointly evaluating the similarities of ROIs from different training videos and the weightings of ROIs contributing to the classification. In this way, we re-associate the ROIs of different training videos to the local classifiers, whose parameters can be updated accordingly. The number of local classifiers can thus be adaptively adjusted by considering the number of training samples that select this classifier.

The main contributions of our method are two-fold. First, we propose a reconfigurable compositional model to recognize FLLs in CEUS videos, which is shown to handle well the variations of the three FLL types (i.e. HCC, HEM, FNH). Second, we study a novel non-convex optimization algorithm to dynamically generate the model structures along with the parameter learning. We apply our method on the SYSU-FLL-CEUS dataset collected from clinics, which contains in total 353 CEUS video sequences of three types of FLLs (HCC, HEM, and FNH). Our method has shown very promising results in characterizing the FLL types (i.e. distinguishing between HCC, HEM and FNH), and classifying the FLLs as benign or malignant. We also evaluate different learning algorithms and show that our learning algorithm outperforms the latent structural SVM [25] and the latent max-margin clustering algorithm [26]. Moreover, we extensively investigate how the individual components of our system contribute to its overall performance.

II. Related Work

In medical imaging, the application of CEUS for differentiating the FLLs is still a relatively new field [13]–[19], [27]. Early results from [27] confirmed that quantitative parametric curve analysis could help in differentiating FNH from the others. Cascades of Artificial Neural Networks [13] have also been employed to classify FLLs based on manually segmented lesion regions. Anaye *et al.* [14] analyzes the Dynamic Vascular Patterns (DVPs) of FLLs with respect to surrounding healthy parenchyma to differentiate between benign and malignant FLLs. Rognin *et al.* [19] developed the parametric imaging technique for mapping the DVP signatures into a single image. Bakas *et al.* [15] developed a histogram-based method to track a manually initialized FLL and its surrounding parenchyma to classify it as either benign or malignant based on its vascular signature. In their recent work [16], a fully automatic method for selecting the optimal frame for initialization of the FLL candidates is proposed. Additionally, their other tracking methods [17], [18] were proposed to track the FLL and a healthy liver region for assisting the differentiation between benign and malignant FLLs.

In all these works, varying degrees of manual interactions are required to identify the ROIs of FLLs or the normal parenchyma area. The manual annotations are highly dependent on the skills and knowledge of the experts, leading to large variations in inter-/intra-observer variability, the median value of which can reach 24% according to [28]. Furthermore, the ever-increasing amount of CEUS data acquired and processed nowadays demands automatic systems that can save the radiologists' time and effort.

On the other hand, automatic detection and segmentation of other tumors (e.g. breast tumor, prostate cancer, obstetrics) using conventional ultrasound have been well studied, as surveyed in [29], [30]. From the computational point, various methods [31], [32] were proposed to segment the suspicious lesions by using the intensity or edge information. Hessian analysis were also explored to segment common geometrical structures for all kinds of tumors [33]–[35] with different modalities (e.g. ultrasound, CT, MRI). In addition, some works transformed the detection problem into classification task using the user-defined features [36], [37].

This paper is an extended version of our previous work in [38], and provides further background, description, insight, analysis, and evaluation. Compared with the previous version, our improved model is more effective and flexible in capturing the diverse patterns of ROIs in each phase. In addition, the previous model can be directly solved by latent SVM while our extended model is formulated as a non-convex problem. We thus propose a novel concave-convex optimization algorithm to dynamically generate the model structures along with the parameter learning.

In computer vision area, image parsing was proposed to parse the natural images into their constituent visual patterns [22], [23], e.g. object parts, scenes or skeletons, in a manner similar to parsing sentences in speech and natural language. This topic has since drawn much attention [39], [40]. Compared with the traditional classification method, image parsing aims at seamlessly unifying segmentation, detection and recognition. However, the uncertainty of hierarchical representations for the images and the well-known complexity of segmentation and recognition make it extremely hard to design effective and efficient models. Zhu and Mumford [23] employed the conceptual stochastic grammar in And-Or graph model to represent complex visual patterns and their relationships. More specifically, the Or-node is used for alternative configurations of structural variations for each component and the And-node points to the composition of a number of components, which can be generally exhibited as the "selective" and "compositional" concepts in our model, respectively. This idea has been extended to other tasks, e.g. action recognition [41], background modeling [42] and trajectory analysis [43]. Our approach was partially motivated by these works, and we investigate a unified selective compositional model for FLL recognition in CEUS that can be discriminatively trained with a novel non-convex optimization method.

III. Our Model

In this section, we present a latent structured model to capture large variations of FLLs in CEUS videos. Given a CEUS video sequence $\mathbf{x}$, y is the corresponding type of the FLL, ranging over a finite set $\mathcal{Y}$ (i.e. HCC, HEM and FNH). We assume the FLLs can be represented by a number of ROIs in three vascular phases: arterial, portal venous, and late phases. We define an

ROI as the minimum enclosing box of an FLL, which is depicted by the corresponding layout in the video. Thus, the objective of our model is to locate the most discriminative ROIs, $\{R_1, R_2, \ldots, R_m\}$, in the CEUS video sequence, and predict the FLL types (including the characterization and classification tasks). To capture the diverse variations of FLLs, we train a set of local classifiers to detect ROIs for each FLL type. Each local classifier is used to detect one candidate ROI, and the most appropriate classifiers (associating with ROIs) are selectively combined to conduct the classification. Meanwhile, the scale and location information of each ROI are also determined.

A. Latent Structures

The detection results of discriminative ROIs are expressed with a set of hidden variables $\mathbf{h} = \{h_1, h_2, \ldots, h_m\}$ (associated with the ROIs $\{R_1, R_2, \ldots, R_m\}$), where h_i takes value from a finite set $\mathcal{H}_i$ of all possible hypotheses about R_i. More precisely, $h_i = (z_i, v_i)$, includes two terms: the layout z_i (i.e. the location and scale) of R_i and the local classifier selection v_i. The layout $z_i = (p_i^x, p_i^y, t_i, s_i)$ specifies the spatial coordinates (p_i^x, p_i^y), the temporal location t_i (i.e. the frame number in the video sequence), and the scale s_i of R_i. For each ROI R_i, we define a set of local classifiers to capture diverse enhancement patterns. The exact number of local classifiers of each phase for each FLL type is automatically learned, and limited to be smaller than the pre-defined maximum number L_i. We denote $N_v = \sum_{i=1}^{m} L_i$ as the maximum number of local classifiers. The classifier selection variable $v_i \in [1, L_i]$ for R_i is used to indicate the most appropriate local classifier after performing inference algorithm. As illustrated in Fig. 1, by adaptively combining different local classifiers in three phases, the diverse enhancement patterns can be captured and our learned model for each FLL type is reconfigurable.

B. Modeling Observations

The ultrasonic characteristics (e.g. internal echo, morphology, edge, echogenicity and posterior echo enhancement) of each ROI often show large variations among different frames in each video and different patients. To capture all these variations of the lesions, we represent the feature of each ROI from the following aspects: the region inside the lesion, denoted as R^-, is used to capture the internal echo of the FLL; the lesion region R to observe the boundary and the morphology of the FLL; and the tissue area surrounding the lesions, denoted as R^+, to measure the posterior echo enhancement. In addition, the echogenicity of the lesion can be measured by comparing the intensities of above regions. Given an ROI R, the region R^- is obtained by shrinking R by a small factor. The region R^+ is the annular region obtained by enlarging the region R by a small factor and then subtracting it from the original region R. We then propose an effective region representation, which consists of 5 components as follows:

$$f(R) = \left[f^t(R^-), f^t(R), f^t(R^+), f^d(R^-, R), f^d(R, R^+)\right] \tag{1}$$

where f^t extracts the appearance features of each region, such as Grey Level Co-occurrence Matrix (GLCM) and Local Phase (LP); f^d calculates the mean intensity difference of two regions. Consequently, the concatenation of all these features, $f(R)$, captures all the desired ultrasonic characteristics of region R.

C. Model Definition

Given a CEUS video $\mathbf{x}$, its FLL type y, and hidden variables $\mathbf{h}$, the conditional probability of the whole recognition problem is defined as,

$$\begin{aligned} p(y|\mathbf{x};\omega) &= \max_{\mathbf{h}\in\mathcal{H}} p(y, \mathbf{h}|\mathbf{x};\omega) \\ &= \frac{\max_{\mathbf{h}\in\mathcal{H}} \exp\left(\omega \cdot \psi(\mathbf{x}, y, \mathbf{h})\right)}{\sum_{\tilde{y}\in\mathcal{Y}} \max_{\mathbf{h}\in\mathcal{H}} \exp\left(\omega \cdot \psi(\mathbf{x}, \tilde{y}, \mathbf{h})\right)} \end{aligned} \tag{2}$$

where ω is the model parameter vector, $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2 \times \cdots \times \mathcal{H}_m$, and $\psi(\mathbf{x}, \mathbf{h}, y)$ is a feature vector depending on the video sequence $\mathbf{x}$, the FLL type y, and the hidden variables $\mathbf{h}$. We define the formulation of $\omega \cdot \psi(\mathbf{x}, \mathbf{h}, y)$ as the following, including two factors: the appearance potential and temporal potential,

$$\begin{aligned} \omega \cdot \psi(\mathbf{x}, y, \mathbf{h}) = &\sum_{i\in m} \alpha_i \cdot \phi^s(\mathbf{x}, y, h_i) \\ &+ \sum_{(i,j)\in\mathcal{E}} \beta_{i,j} \cdot \phi^t(\mathbf{x}, y, h_i, h_j) \end{aligned} \tag{3}$$

where $\phi^s(\cdot)$ is the appearance potential function of variable h_i and $\phi^t(\cdot)$ is the temporal potential function of (h_i, h_j). $\mathcal{E}$ is the set of neighboring hidden variables (defined for the pairwise temporally adjacent ROIs).

1) Appearance Potential $\alpha_i \cdot \phi^s(\mathbf{x}, y, h_i)$: Given the hidden variable $h_i = (z_i, v_i)$, the singleton potential function $\phi^s(\cdot)$ is conditioned on the FLL type y, classifier selection v_i and appearance feature of R_i.

$$\begin{aligned} \alpha_i \cdot \phi^s(\mathbf{x}, y, h_i) &= \alpha_i \cdot \phi^s(\mathbf{x}, y, z_i, v_i) \\ &= \sum_{a\in\mathcal{Y}} \mathbb{1}(y=a) \sum_{b=1}^{L_i} \alpha_{iab} \cdot \mathbb{1}(v_i = b) \cdot f\left(\mathbf{x}(z_i)\right) \end{aligned} \tag{4}$$

where $f(\mathbf{x}(z_i))$ is the feature vector describing the appearance of the region (Section III-B). The indicator function $\mathbb{1}(y = a)$ is equal to one if $y = a$, zero otherwise. We denote the weight parameter of each local classifier as α_{iab}. Intuitively, the different optimal regions are selected with different local classifiers to capture the variations. The whole parameter α_i is simply the concatenation of all local classifiers for all FLL types.

2) Temporal Potential $\beta_{i,j} \cdot \phi^t(\mathbf{x}, y, h_i, h_j)$: The potential function $\phi^t(\cdot)$ models the appearance similarity between FLL type y and the temporal transition of a pair of temporally neighboring regions (h_i, h_j),

$$\beta_{i,j} \cdot \phi^t(\mathbf{x}, y, h_i, h_j) = \sum_{a\in\mathcal{Y}} \beta_{ija} \cdot F(\mathbf{x}, z_i, z_j) \tag{5}$$

where $F(\cdot)$ is concatenated by two features: appearance difference feature, computed by the difference of $f(\mathbf{x}(z_i))$ and $f(\mathbf{x}(z_j))$, and spatial displacement feature, i.e. Euclidean distance between the spatial coordinates of z_i and z_j. Intuitively, the temporal potential is used to model the enhancement changes of a certain FLL between two phases (e.g. hyper-, iso-, or hypo-enhancement), but not diverse enhancement patterns

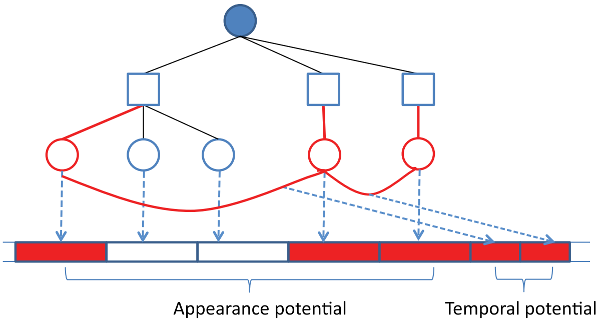

Fig. 3. Mapping the latent structured model with the potentials defined in Eqn. (3). The squares denote the three different phases and the circles represent the ROIs detected by local classifiers in each phase. The bottom bar indicates the feature vectors in different potentials. These ROIs are described by the feature vectors in the potential, which are connected by the dashed blue arrow lines. The feature vectors for each FLL consists of seven components (the bottom bar): three for describing the ROIs detected by different local classifiers in the arterial phase, two for the ROIs in the portal venous phase and the late phase, and two for the temporal potential. The feature vectors for the selected ROIs are highlighted by red, and other feature bins are set into zeros.

of a ROI, we thus use the same weight vector $\beta_{i,j}$ for different pairs of classifier selections (v_i, v_j), i.e. the variable v_i and v_j in the temporal potential are ignored for simplicity. The parameter $\beta_{i,j}$ is then simply the concatenation of all β_{ija}. Fig. 3 illustrates the global feature assignments of our model.

The overall parameter ω of our model is summarized as,

$$\omega = [\alpha_1, \ldots, \alpha_m, \beta_{1,2}, \beta_{2,3}, \ldots, \beta_{m-1,m}]. \tag{6}$$

IV. Data Driven Inference

To provide more information for diagnosis, the task is formulated as the following joint inference problem of both the FLL type and the hidden variables:

$$\begin{aligned}(y^*, \mathbf{z}^*, \mathbf{v}^*) &= \arg\max_{y,\mathbf{z},\mathbf{v}} p(y, \mathbf{z}, \mathbf{v} | \mathbf{x}; \omega) \\ &= \arg\max_{y,\mathbf{z},\mathbf{v}} \omega \cdot \psi(\mathbf{x}, y, \mathbf{z}, \mathbf{v}). \end{aligned} \tag{7}$$

The inference is hard primarily because the state space of the ROI layout variable $\mathbf{z}$ is huge. Here we propose an efficient inference algorithm by adopting several data-driven pruning steps into the dynamic programming. For each possible type $\tilde{y}$ and local classifier selections $\tilde{\mathbf{v}}$, we first search for the optimal ROI layout $\tilde{\mathbf{z}}$. Finally the best detection result including the optimal y^*, $\mathbf{z}^*$ and $\mathbf{v}^*$ is determined by finding maximum detection score after exhausting searching over all possible types $\tilde{y}$ and local classifier selections $\tilde{\mathbf{v}}$.

More precisely, given the FLL type $\tilde{y}$ and the local classifier selections $\tilde{\mathbf{v}}$, our model is simplified into a standard chain structure with m nodes that can be effectively solved by dynamic programming. However, the searching space for the optimal ROIs is huge if we check every location and scale in every frame. Considering many of these candidate regions are redundant, we propose an efficient data-driven inference algorithm, which combines spatial and temporal pruning techniques to disregard those less discriminative frames and regions. The algorithm includes the following three components:

1) Temporal Pruning: In a typical CEUS video, the appearance of ultrasound frames often varies slowly and smoothly according to the hemodynamic, and the most discriminative frames are usually those with the largest contrast changes compared with neighboring frames. Thus, a small set of candidate frames, which have local maximum of the contrast change, are automatically selected. Formally, for each frame I in a video $\mathbf{x}$, we compute the contrast feature from the co-occurrence distribution defined over I [44]. The contrast vector $\mathbf{q}$ for all frames is then $(q_1, q_2, \ldots, q_T)$. Let $\Delta\mathbf{q}$ be the difference vector of $\mathbf{q}$, the candidate frame set B is formed by finding the frames at the local maximum of $\Delta\mathbf{q}$. The duration time for each phase may vary due to individual hemodynamic or different site of injection, and the average duration time for each phase is reported in [6]. We use the most stable duration time for each phase, that is, 10 s–30 s as the arterial phase, 45 s–120 s as the portal venous phase and $\geq$120 s as the late phase. During these stable duration periods, the explicit FLL enhancement patterns for each phase can be observed. In addition, since we assume that the enhancement pattern appears at least in one selected frame of each phase, our proposed method can automatically locate the FLL regions by our inference algorithm.

2) Spatial Pruning: After temporal pruning, we also prune the less important regions in each frame considering the following two priors: saliency prior and location prior. First, we believe that *salient* regions (e.g. having higher contrast or containing typical structures) have more discriminative information, and thus are more likely to be the candidates of ROIs. Second, we observe that FLLs often appear in or close to the center of the images, probably because a skilled ultrasound operator usually places the liver area in the middle of the display. According to these two observations, we evaluate all the regions with different scales in each candidate frame $I \in B$ (sliding window protocol), and only select the regions with prior probability larger than a threshold τ as the ROI candidates. The threshold τ determines the number of candidate ROIs by the following dynamic programming algorithm, and whether the candidate ROIs can cover the main parts of FLLs. The larger τ means the larger number of candidate ROIs that leads to longer training and testing time, and higher possibility to cover the complete FLLs. In our experiments, we set $\tau = 0.6$ to balance the learning efficiency and classification accuracy, which is decided by cross-validation on a small validation set. The prior probability of a region R being an ROI is,

$$p(R) = \mathcal{S}(R)\mathcal{G}(C^R | C^I, \sigma), \tag{8}$$

where $\mathcal{S}(R)$ is the normalized mean saliency of the region R in the saliency map $\mathcal{S}$, computed by the quaternion-based spectral saliency method [45]. C^R and C^I are the centers of region r and the whole image (not just the conical area in the ultrasound image), respectively. $\mathcal{G}(C^R | C^I, \sigma)$ is a Gaussian distribution. The σ represents the confidence that the FLL is close to the center of each frame. Due to the difference of radiologists' operation or patients' hemodynamic, the locations of FLL region may vary in different CEUS videos. In our experiments, we set $\sigma = 0.5$ to balance the efficiency and accuracy by cross-validation. As shown in Fig. 4, by combining these two priors, our

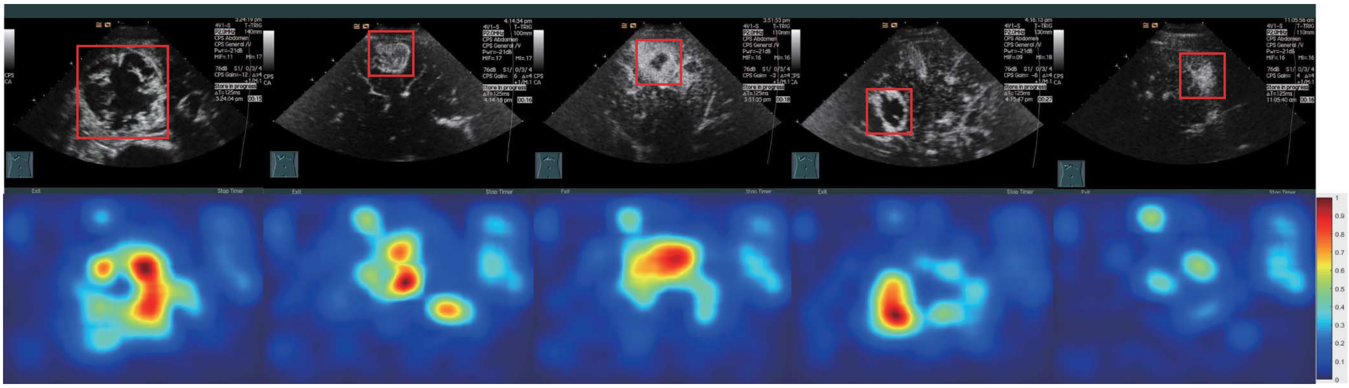

Fig. 4. Results in the spatial pruning stage. The first row presents the FLLs in the arterial phase for different examples and the location of each FLL is annotated by a red rectangle. The second row shows the probability map for each pixel. We denote the red colored locations have the highest probability to be a suspicious FLL.

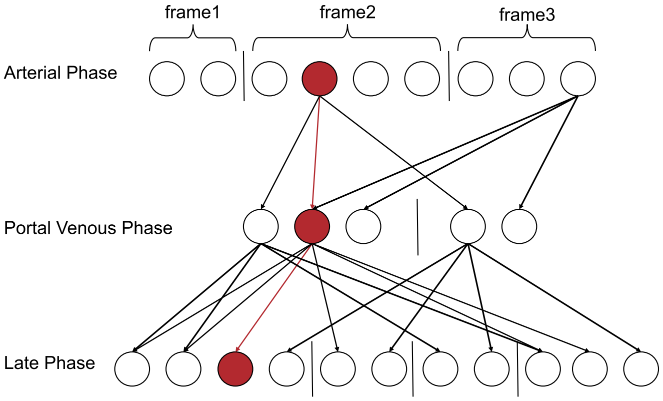

Fig. 5. Example of the dynamic programming inference. We represent each candidate ROI of each phase as a circle in each row, and the whole hypotheses of ROIs consist of all candidate ROIs in all candidate frames. Only the ROIs in the small spatial neighborhood are linked. The optimal locations of ROIs in three phases (denoted as the red circles) are determined by the efficient dynamic programming.

predicted probability map has high consistency with the locations of FLLs. The few wrongly predicted locations will most likely be further pruned by the following global optimization. It is often observed that the locations of FLLs do not change much after the arterial phase [6], and the spatial pruning in the arterial phase can be of no effect. Thus, in the last two phases, we only search the regions in a spatial neighborhood (e.g. 50 × 50 pixels) around the locations of the ROI candidates found in the arterial phase. Note that the size of the spatial neighborhood is empirically chosen according to the resolution and zoom factor of the input video.

3) Dynamic Programming: Given the FLL type $\tilde{y}$ and the local component selection $\tilde{\mathbf{v}}$, the hidden variables $\mathbf{z} = \{z_1, z_2, \ldots, z_m\}$ forms a Markov chain. This model is composed of the appearance potential $\alpha_i \cdot \phi^s(\cdot)$ for each z_i and the temporal transition potential $\beta_{i,j} \cdot \phi^t(\cdot)$ for each pair of neighboring variables (z_i, z_j). The possible layout of each ROI, z_i, ranges in the hypothesis set $\mathcal{Z}_i$, after above mentioned temporal and spatial pruning. As shows in Fig. 5, each candidate ROI is only connected with its spatial neighboring candidates in the next phase. Thus the optimal locations for $\tilde{y}$ and $\tilde{\mathbf{v}}$, $\tilde{z}$, can be calculated by the Viterbi algorithm [46].

```
Algorithm 1 Data-driven inference algorithm
Input:
    Given the model parameter $\omega$ and a CEUS video $\mathbf{x}$.
Output:
    The optimal detection results: $\mathbf{z}^*, \mathbf{v}^*$ and $y^*$ for the optimal ROI
    layouts, classifier selections and predicted FLL types.
Initialization:
    Partition $x$ as $m$ vascular segments according to the duration times.
Inference:
    for all possible FLL type $\tilde{y}$ and classifier selections $\tilde{v}$ do
    1. Temporal pruning
        (a) Compute the contrast statistic feature for each frame $I$.
        (b) Calculate the gradient of contrast vector $\Delta\mathbf{q}$ of all frames.
        (c) find the local minimums of $\Delta\mathbf{q}$ and thus form the candidate
            set $B$.
    2. Spatial pruning
        (a) For $I \in B$ in the arterial phase, select the candidate regions
            with three scales whose prior probability $p(R) \geq \tau$.
        (b) For $I \in B$ in the last two phases, select the regions in a small
            neighborhood of candidate regions in the arterial phase.
    3. Find the optimal locations $\tilde{\mathbf{z}}$ by Viterbi algorithm[46].
    end for
    Calculate $y^*, \mathbf{v}^*, \mathbf{z}^*$ for solving the problem.
```

Therefore, for each pair of $\tilde{y}$ and $\tilde{\mathbf{v}}$, we can find its optimal ROI layout $\tilde{\mathbf{z}}$, and the final detection results $(y^*, \mathbf{v}^*, \mathbf{z}^*)$ can be determined by exhaustively searching over all $\tilde{y}$ and $\tilde{\mathbf{v}}$ ((7)), which is related to a small space of possibility. The entire inference procedure is outlined in Algorithm 1.

V. Model Learning

In this section we introduce a novel non-convex optimization algorithm for jointly generating the latent structure and learning parameters. Inspired by the existing non-convex optimization methods [20], [47], our algorithm trains the model in a dynamic manner: iteratively determining the model structures (i.e. re-producing the local classifiers) along with learning the model parameters. Specifically, the new local classifier is activated to better capture variations within training data. One example is illustrated in Fig. 6: from (a) to (b), a new local classifier is created to capture the additional variations.

Given a CEUS video $\mathbf{x}$, we are interested in obtaining the accurate FLL type y as well as the layout $\mathbf{z}$ of ROIs and the classifier selection $\mathbf{v}$ for the FLL. During the manual diagnosis procedure, the radiologists often annotate a reference ROI of the FLL

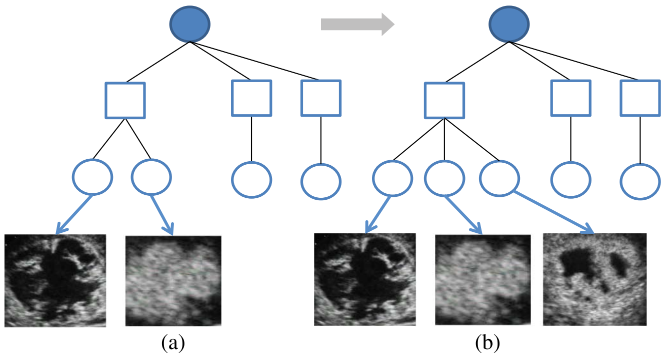

Fig. 6. An illustration of structural model learning. Our learned models for the HCC class are illustrated in different iterations. a) The learned local classifiers after the first iteration; b) a new local classifier is created to capture another pattern variance of FLL.

with maximum contrast and sharpness in the arterial phase, and then make further decisions by checking the rest of the video but without labeling the FLL in subsequent frames. Accordingly, we introduce one reference ROI layout $r = (p^x, p^y, t, s)$ from the annotated ground-truth, which represents the expert knowledge of diagnosis. Let $D = \{(\mathbf{x}_1, y_1, r_1), \ldots, (\mathbf{x}_n, y_n, r_n)\}$ be a set of labeled training samples. Given the true label y_k, $k = 1, \cdots, n$ and the reference ROI r_k, we denote the user-specified risk of prediction $(\hat{y}_k, \hat{\mathbf{h}}_k)$ as $\Delta(y_k, \hat{y}_k, \hat{\mathbf{h}}_k, r_k)$. The details of our optimization method are presented in the following section.

A. Optimization

Beyond the assessment of the predicted label $\hat{y}_k$, our loss function $\Delta(y_k, \hat{y}_k, \hat{\mathbf{h}}_k, r_k)$ also measures how the output hidden variables $\hat{\mathbf{h}}_k$ are compatible with the expert knowledge. Note that the annotated ROI r_k can be used to guide our localization of ROIs of the same FLL in the subsequent frames which should be discriminative as well as informative for recognizing the FLLs. Thus, only the located ROIs in the spatial neighborhood of r_k (i.e. 50*50 neighborhood) can be regarded as correct. We use this rule to define the loss function,

$$\Delta(y_k, \hat{y}_k, \hat{\mathbf{h}}_k, r_k) = \begin{cases} 1 & \text{if } y_k = \hat{y}_k, \neg\text{neighbor}(\hat{\mathbf{h}}_k, r_k) \\ 0 & \text{if } y_k = \hat{y}_k, \text{neighbor}(\hat{\mathbf{h}}_k, r_k) \\ 1 & \text{if } y_k \neq \hat{y}_k \end{cases} \tag{9}$$

This loss pushes down the score of ROIs that are not in the neighborhood of the annotated ROI r_k, denoted as $\neg\text{neighbor}(\hat{\mathbf{h}}_k, r_k)$. However, Δ is typically discontinuous, which is very difficult to minimize. In our method we optimize the model by minimizing a regularized upper bound on the risk,

$$\begin{aligned} &\min_{\omega, \xi_k \geq 0} \quad \frac{1}{2}\|\omega\|^2 + \frac{C}{n}\sum_{k=1}^{n}\xi_k \\ &\text{s.t.} \quad \mathcal{S}_a - \sum_{k=1}^{n} \max_{(\hat{y}_k, \hat{\mathbf{h}}_k)} \omega \cdot \psi(\mathbf{x}_k, \hat{y}_k, \hat{\mathbf{h}}_k) \\ &\qquad \geq \sum_{k=1}^{n} \Delta(y_k, \hat{y}_k, \hat{\mathbf{h}}_k, r_k) - \sum_{k=1}^{n} \xi_k \end{aligned} \tag{10}$$

where $\mathcal{S}_a$ represents the aggregated response of all examples with hidden variables $\{\mathbf{h}_k\}_1^n$, given the true label y_k. C is the penalty parameter for the training loss. ξ_k represents the loss for each training sample, and it is subject to the Maximizing Margin constraints (i.e. the condition term in (10). Many previous methods, e.g. latent structure SVM [25] and And-Or graph learning [20], separately optimized the hidden variable $\mathbf{h}_k$ for each sample by maximizing the response $\omega \cdot \psi(\mathbf{x}_k, y_k, \mathbf{h}_k)$, and we optimize $\mathbf{h}_k$ over all training samples jointly while producing local classifiers. Intuitively, for a local classifier, we encourage its detections (i.e. ROIs) to be coherent over all examples, so that the detected ROIs sharing similar enhancement patterns tend to be grouped together during training. Thus, we discover the similarity between training samples together with solving $\mathbf{h}_k$. In this way, we can generate different local classifiers to capture the variations of FLLs over all training instances. We now present how to encompass these intuitions into our model optimization.

We define the response $\mathcal{S}_a$ over all training samples with three terms. First, the unary term is used to verify the discriminative capabilities of ROIs and we represent it by the model response; second, the pairwise term is introduced to measure the appearance similarity among the ROIs of the different samples; finally, the regularization term prevents the cluster imbalance problem. The global optimal hidden variables of all examples can be thus optimized by maximizing the summation of a unary term φ^u for the model response, a pairwise term φ^p for the appearance similarity, and a regularization term φ^r,

$$\begin{aligned} \mathcal{S}_a = \max_{\{\mathbf{h}_k\}_1^n} \Bigg\{ &\sum_{k=1}^{n} \varphi^u(\mathbf{x}_k, y_k, \mathbf{h_k}) + \sum_{j=1}^{L_1} g_j \varphi^r\left(\{\mathbf{h}_k\}_1^n, j\right) \\ &+ \alpha \sum_{(k,k') \in \mathcal{N}} \varphi^p(\mathbf{x}_k, x_{k'}, \mathbf{h}_k, \mathbf{h}_{k'}) \Bigg\}. \end{aligned} \tag{11}$$

where $\mathcal{N}$ is the neighborhood set of the examples and in our case, it contains all pairwise samples in the training data; we weight the pairwise term by α. The regularization term restricts the number of valid local classifiers in our learned structure. g_j is a non-negative score of each classifier and $\varphi^r(\cdot)$ is the corresponding indicator function. We define $\varphi^u(\cdot)$, $\varphi^p(\cdot)$, $\varphi^r(\cdot)$ and g_j as follows, respectively,

$$\begin{aligned} \varphi^u(\mathbf{x}_k, y_k, \mathbf{h}_k) &= \omega \cdot \psi(\mathbf{x}_k, y_k, \mathbf{h}_k) \\ \varphi^p(\mathbf{x}_k, \mathbf{x}_{k'}, \mathbf{h}_k, \mathbf{h}_{k'}) &= -\,\mathbb{1}(\mathbf{v}_{k,1} = \mathbf{v}_{k',1}) \\ &\quad \mathbf{d}\left(f\left(\mathbf{x}_k(\mathbf{z}_{k,1})\right), f\left(\mathbf{x}_{k'}(\mathbf{z}_{k',1})\right)\right) \\ \varphi^r\left(\{\mathbf{h}_k\}_1^n, j\right) &= \begin{cases} 0 & \exists k : j = \mathbf{v}_{k,1}; \\ 1 & \text{otherwise} \end{cases} \end{aligned} \tag{12}$$

where $\mathbb{1}(\mathbf{v}_{k,1} = \mathbf{v}_{k',1})$ is the indicator function; In particular, because the maximum number L_i of local classifier in last two phases is set to 1, we only need one local classifier for these phases. In this way, we only measure the appearance similarities of ROIs in the arterial phase and the specific classifier selection is denoted as $\mathbf{v}_{k,1}$. $\mathbf{d}(f(\mathbf{x}_k(\mathbf{z}_{k,1})), f(\mathbf{x}_{k'}(\mathbf{z}_{k',1}))$ is the squared Euclidean distance between appearance feature of the ROIs of samples $(\mathbf{x}_k, \mathbf{x}_{k'})$. $\varphi^r(\{\mathbf{h}_k\}_1^n, j)$ is set to 1 only if there exists at least one example which selects the j-th classifier.

By substituting (11) into the optimizing problem (10), the global optimization function can be rewritten as,

$$\begin{aligned}\min_{\omega}\Bigg\{&\frac{1}{2}\|\omega\|^2\\&+\frac{C}{n}\sum_{k=1}^{n}\max_{(\hat{y}_k,\hat{\mathbf{h}}_k)}\Big(\omega\cdot\psi(\mathbf{x}_k,\hat{y}_k,\hat{\mathbf{h}}_k)+\Delta(y_k,\hat{y}_k,\hat{\mathbf{h}}_k,r_k)\Big)\Bigg\}\\&-\frac{C}{n}\left\{\max_{\{\mathbf{h}_k\}_1^n}\left(\sum_{k=1}^{n}\varphi^u(\mathbf{x}_k,y_k,\mathbf{h}_k)+\sum_{j=1}^{L_1}g_j\varphi^r\left(\{\mathbf{h}_k\}_1^n,j\right)\right.\right.\\&\left.\left.+\,\alpha\sum_{(k,k')\in\mathcal{N}}\varphi^p(\mathbf{x}_k,\mathbf{x}_{k'},\mathbf{h}_k,\mathbf{h}_{k'})\right)\right\}\end{aligned}\tag{13}$$

The term $\|\omega\|^2$ is convex and the pointwise maximum of convex function is convex. The objective in (13) is a summation of a convex part and concave part. A local minimum or saddle point solution can be thus found using the concave-convex procedure (CCCP) and it can guarantee that our algorithm always converges [47]. To solve this non-convex optimization problem, the proposed algorithm iteratively updates the latent structure and optimizes the model parameters. Given the updated model parameters in each iteration, we generate the latent structures with two steps: first, each local classifier locates one candidate ROI by maximizing the detection score by applying the current model; second, we select the optimal local classifiers for each example by transforming this task into a graph-based labeling problem. According to the classifier selections over examples, we re-associate the ROIs of all examples with these classifiers, and reconfigure the model structure. In the next iteration, the parameters of local classifiers will be trained on the newly partitioned data. The number of local classifiers will be adjusted according to the selections of examples. Our whole learning procedure is presented as follows.

Step 1: The parameter ω_t is fixed. We compute the hidden variables for all training data, which corresponds to approximating the concave function by a linear upper bound,

$$-\frac{C}{n}(\mathcal{S}_a)\leq-\frac{C}{n}\sum_{k=1}^{n}\omega_t\cdot\psi\left(\mathbf{x}_k,y_k,\mathbf{h}_k^*\right)\tag{14}$$

The optimal hidden variables $\mathbf{h}_k^*=(\mathbf{z}_k^*,\mathbf{v}_k^*)$ are computed by maximizing $\mathcal{S}_a$ with two following steps: find the best locations $\tilde{\mathbf{z}}_k$ of ROIs for each possible classifier selection $\mathbf{v}_k$, and then given all $\tilde{\mathbf{z}}_k$ of all possible $\mathbf{v}_k$, we select the best local classifiers $\mathbf{v}_k^*$ and the corresponding best ROIs $\mathbf{z}_k^*$, i.e. we determine the latent structures.

a) For each example $\mathbf{x}_k$, we compute the optimal locations $\tilde{\mathbf{z}}_k$ for each classifier selection hypothesis $\mathbf{v}_k$ by,

$$\tilde{\mathbf{z}}_k=\arg\max\nolimits_{\mathbf{z}_k}\omega_t\cdot\psi(\mathbf{x}_k,y_k,\mathbf{z}_k,\mathbf{v}_k).\tag{15}$$

Eqn. (15) can be solved effectively by our proposed inference algorithm, detailed in the Section IV.

b) Given the candidate $\tilde{\mathbf{z}}_k$ for each possible $\mathbf{v}_k$, we find the optimal classifier configurations $\{\mathbf{v}_k^*\}_1^n$ by solving,

$$\begin{aligned}\mathcal{S}_a=\max_{\{\mathbf{v}_k\}_1^n}&\left\{\sum_{k=1}^{n}\varphi^u(\mathbf{x}_k,y_k,\tilde{\mathbf{z}}_k,\mathbf{v}_k)+\sum_{j=1}^{L_1}g_j^t\varphi^r\left(\{\mathbf{h}_k\}_1^n,j\right)\right.\\&\left.+\alpha^t\sum_{(k,k')\in\mathcal{N}}\varphi^p(\mathbf{x}_k,x_{k'},\tilde{\mathbf{z}}_k,\tilde{\mathbf{z}}_{k'},\mathbf{v}_k,\mathbf{v}_{k'})\right\}.\end{aligned}\tag{16}$$

And this optimization problem in (16) can be equivalent to the minimization of energy $E(\{v_k\}_1^n)$,

$$\begin{aligned}E\left(\{v_k\}_1^n\right)=&\sum_{k=1}^{n}\left(1-\varphi^u(\mathbf{x}_k,y_k,\hat{\mathbf{z}}_k,\mathbf{v}_k)\right)\\&+\sum_{j=1}^{L_1}g_j^t\left(1-\varphi^r\left(\{\mathbf{h}_k\}_1^n,j\right)\right)\\&+\sum_{(k,k')\in\mathcal{N}}\alpha^t\left(1-\varphi^p(\mathbf{x}_k,\mathbf{x}_{k'},\tilde{\mathbf{z}}_k,\tilde{\mathbf{z}}_{k'},\mathbf{v}_k,\mathbf{v}_{k'})\right)\end{aligned}\tag{17}$$

where α^t determines the weight of appearance similarities of examples within the same classifier. If α^t is large enough, this optimization problem can be approximately transferred into a standard clustering algorithm as [20]; if α^t approaches 0, then E can be minimized by only optimizing the unary term of all training data using the latent structural SVM [48]. In our framework, we integrate these two kinds of learning frameworks. A growing parameter $\lambda>1$ is introduced to iteratively increase the weight α^t, which increases the weight of the appearance similarity constraints. The weight g_j^t is also updated to adjust the score of each classifier according to the model structure in the last iteration. Let α^o be the initialized weight parameter, we iteratively adjust the weight α^t and the score g_j^t by,

$$\begin{aligned}\alpha^t&=\alpha^o\cdot\lambda^t,\\g_j^t&=1-\frac{\left|\sum_{k=1}^{n}\mathbb{1}(j=\mathbf{v}_{k,1})\right|^2}{n^2}.\end{aligned}\tag{18}$$

where λ is set as 1.6 empirically. According to our reported results in Section VI-D, the growing parameter λ effectively avoids the premature convergence in model structure optimization. Note that $\sum_{k=1}^{n}\mathbb{1}(j=\mathbf{v}_{k,1})$ represents the number of examples which select the j-th classifier, g_j^t is thus the negative of square of the number of samples which select the j-th classifier. Intuitively, g_j^t encourages the model select the local classifiers that are able to handle more training instances. In this way, our regularization term eliminates solitude local classifiers iteratively.

By substituting $\varphi^p(\cdot)$ from (12) into (17), the pairwise term is a submodular function because its value would range from [0,1] if $\mathbf{v}_{k,1}\neq\mathbf{v}_{k',1}$. Besides, the regularization term can be transformed into the label cost in the graph-cut problem [49]. Then, the optimization of (17) is transfered as a tractable graphical labeling problem. We solve this optimization problem by the well-studied α-expansion method [50]. The classifier selections for all samples $\{\mathbf{v}_k^*\}_1^n$ can be effectively determined and

```
Algorithm 2 Model learning algorithm
Input:
    Training samples, $D = \{(\mathbf{x}_1, y_1, r_1), (\mathbf{x}_k, y_k, r_k), \ldots, (\mathbf{x}_n, y_n, r_n)\}$,
    where $\mathbf{x}_k$ and $y_k$ denote the CEUS video and its cooresponding FLL
    type, respectively. $r_k$ is the provided reference ROI layout for the
    FLL in each video.
Output:
    The learned model parameter $\omega$ and the latent structures (i.e. the local
    classifiers) for each FLL type.
Initialization:
    1. Initialize the number of local classifiers and $\mathbf{v}_k$ for all samples
       by performing spectral clustering over $\{r_k\}_1^n$ .
    2. Initialize the hidden ROI layouts $\mathbf{z}$ for all training data and model
       parameters $\omega$.
Repeat
    1. Given the updated parameter $\omega^t$, estimate the hidden variable
       $\mathbf{h}_k^* = (\mathbf{v}_k^*, \mathbf{z}_k^*)$ for each sample $(\mathbf{x}_k, y_k)$.
       (a) Compute the optimal locations $\tilde{\mathbf{z}}_k$ for ROIs in terms of each
           local classifier by solving Eqn. (15).
       (b) Update the parameter $\alpha^t$ and $g_j^t$ in Eqn. (18).
       (c) Find the optimal local classifier selection $\mathbf{v}_k^*$ for each video
           by optimizing the energy (17) using $\alpha$-expansion algorithm.
       (d) Obtain the best location $\mathbf{z}_k^*$ associated with $\mathbf{v}_k^*$.
    2. Update the model parameters $\omega_{t+1}$ by solving the problem in
       Eqn. (13).
until The objective in Eqn. (10) converges.
```

$\mathbf{z}_k^*$ for each sample is conveniently obtained by $\mathbf{v}_k^*$. The optimal hidden variable can be thus calculated. The number of local classifiers L_1 can thus be automatically determined according to the selections $\{\mathbf{v}_k^*\}_1^n$. Specifically, the local classifiers none of examples has selected will be deleted and the rest local classifiers are the finally used in the model.

Step 2: By fixing the hidden variable $\mathbf{h}_k^*$ for all training data, the model parameter can be updated by minimizing a convex upper bound of the objective in (13),

$$\omega_{t+1} = \arg\min_{\omega} \frac{1}{2}\|\omega\|^2 + C\sum_{k=1}^{n} \max_{(\hat{y}_k, \hat{\mathbf{h}}_k)} \Big\{\omega \cdot \psi(\mathbf{x}_k, \hat{y}_k, \hat{\mathbf{h}}_k) + \Delta(y_k, \hat{y}_k, \hat{\mathbf{h}}_k, r_k)\Big\} - \omega \cdot \psi\left(\mathbf{x}_k, y_k, \mathbf{h}_k^*\right) \quad (19)$$

This is a standard structural SVM problem, which can be solved by the cutting plane method [25]. Thus, we can keep the optimization objective decreasing in each iteration. We adopt an usual one-vs-one binary classification approach and output the predicted FLL type and best hidden variables $\mathbf{h}$ for each sample.

B. Initialization

To obtain a better local minimal, we initialize the number of local classifiers and classifier selections $\mathbf{v}_k$ of each sample by performing clustering of the reference ROIs r_k. Intuitively, we partition the samples according to the confidential ROIs from the radiologists and indirectly give a good initialization for capturing the variance of FLLs in CEUS video. The spectral clustering method is adopted due to its fast implementation. The hidden variable $\mathbf{z}_k$ for determining the location of each ROI is initialized as the center of image in the middle frame of each vascular phase. The sketch of our learning algorithm is presented in Algorithm 2.

VI. Experiments

To evaluate our performance, we conducted a series of challenging tests on a large dataset, and the empirical results are presented with the analysis in this section.

A. Dataset

Since CEUS is a relatively new technique, especially in the CAD field, there are not many public datasets available. To advance research in this area, we build the SYSU-FLL-CEUS dataset from the CEUS data collected at the First Affiliated Hospital, Sun Yat-sen University, which has been made publicly available.[1] Consent was obtained from all patients for using this dataset. The equipment used was Aplio SSA-770A (Toshiba Medical System), and all videos included in the dataset are collected from pre-operative scans. The dataset consists of CEUS data of FLLs in three types: 186 HCC, 109 HEM and 58 FNH instances (i.e. 186 malignant and 167 benign instances). We use 10 CEUS videos of each type as the validation set. The spatial resolution of each CEUS video is 768 × 576, and the video length varies from 3 to 4 minutes with 15 fps. All videos are selected based on the assumption that the FLLs can be observed in all three phases and duration time for each phase is similar among different videos (e.g. $\geq$120 s in the video as the late phase). These CEUS videos are collected by starting from the arterial phase, and the previous frames during injection are excluded. The challenges in this dataset are summarized as follows. First, no manual temporal segmentation for different phases is provided. Second, the FLL instances have large variations in size, location, enhancement patterns. Third, the regions of FLLs may be invisible in several intermediate frames. Besides the specific FLL types (i.e. HCC, HEM, FNH), we also provided, in the arterial phase of each video, an ROI which is annotated by a doctor to assist diagnosing the FLL. These videos were taken by experts with more than ten years of experience.

B. Implementation Details

In our implementation, we use the popular GLCM feature [44] to describe the appearance (i.e. enhancement pattern) of each ROI. Specifically, we extract four statistics (i.e. Contrast, Correlation, Energy, Homogeneity) of GLCM with four orientations ($\theta = 0°, 45°, 90°, 135°$). Therefore the feature dimension of f^t is $4 \times 4 = 16$, and the overall feature dimension of $f(R)$ for each region in (1) is thus 52, which contains three f^t and two f^d. During the data-driven inference, three scales of regions (i.e. 64 × 64, 128 × 128, 200 × 200) and one step length 20 are used for the sliding window search. $\tau = 0.6$ and $\sigma = 0.5$ are empirically set for spatial pruning. Intuitively, the ROIs of different FLLs often show diverse enhancement patterns (e.g. hyper-, iso- or hypo-enhancement) in the arterial phase, while the ROIs in the portal and late phases mostly appear less diverse for malignant or benign tumors. Therefore the maximum number of local classifiers L_i of the ROIs in the first phase is set to 3, and L_i is set equal to 1 for portal venous phase, but also $L_i = 1$ for late phase, respectively. In the learning stage, we empirically set the

[1] https://drive.google.com/folderview?id=0B5LimsUgYY7ifjRfLUtxb1FR-Z2ZXcHN0a0oyeFFUaXdyT2xBMDRpclZES0dTMS1uTXk3VjA&usp=sharing

TABLE I
COMPARISONS OF DIFFERENCE LEARNING ALGORITHM SETTINGS. WE REPORT SENSITIVITIES AND MEAN ACCURACIES FOR CHARACTERIZATION OF HCC, HEM AND FNH TYPES, AND CLASSIFICATION OF BENIGN AND MALIGNANT. ACC1 AND ACC2 ARE THE MEAN ACCURACIES FOR THESE TWO CLASSIFICATION PROBLEMS, RESPECTIVELY. SENS. MEANS THE SENSITIVITY OF EACH CLASS

	Sens. HCC(%)	Sens. HEM(%)	Sens. FNH(%)	ACC1(%)	Sens.Benign(%)	Sens.Malignant(%)	ACC2(%)
Ours1 (unary+pairwise)	84.2%	85.7%	**81.8%**	84.3%	**99.0%**	84.3%	91.2%
Ours2 (unary+regularization)	86.5%	**89.6%**	54.5%	83.6%	93.9%	86.1%	90.3%
Ours3 (one local classifier)	88.9%	80.9%	63.6%	82.4%	85.7%	93.4%	89.7%
Ours4 ($\lambda = 1$)	**92.9%**	75.8%	63.6%	83.6%	97.9%	86.7%	92.4%
Ours	88.5%	86.2%	63.6%	**84.8%**	88.6%	**97.3%**	**92.7%**
LMMC [26]	84.4%	**89.6%**	54.5%	82.3%	50.5%	93.9%	83.3%
LSSVM [25]	85.9%	85.7%	54.5%	82.0%	78.0%	96.1%	85.3%

initialized weight parameter $\alpha^0 = 0.5$, the growing parameter $\lambda = 1.6$ and the penalty parameter $C = 100$.

In the experiments, we adopt the standard setup of randomized 5-fold cross-validation. The sensitivity for each class and mean accuracy are used as the evaluation criteria, which are commonly used by others [14]. The experiments are carried out on a PC with Core I7 3.4 GHz CPU, 12 GB RAM and conventional hard drive, and the typical processing time for testing a 4-min CEUS video is about 200 seconds.

C. Results and Comparisons

We first report the sensitivities and mean accuracies of our method in differentiating benign and malignant FLLs. The average accuracy (92.7%) on 353 FLLs is comparable to the results reported in previous studies on smaller datasets: 97% [14] for 146 FLLs, 82.4% [15] for 17 FLLs, 91.6% [17] for 107 FLLs and 92.8% [18] for 14 FLLs. Note that for all these previous works, a reference ROI or a region of healthy tissues must be annotated manually. However, our method automatically produces the classification results (reported in Table I) without any manual labeling. In addition, our results compare favorably with those medical diagnoses in medical literature, which reported sensitivity ranging from 85%–97% [4], [5], [19], [51]. Moreover, we achieve the promising mean accuracy 84.8% for characterizing the specific HCC, HEM and FNH types and the sensitivities for each type are 88.5%, 86.2% and 63.6%, respectively. In contrast, the sensitivities of HCC and HEM reported in [13] are 86.9% and 93.8% with the manually labeled contours of FLLs.

Table II reports the detection accuracies for the lesion regions of three FLL types. The reference ROI provided by the radiologists in the arterial phase is treated as ground-truth. Our detected ROI in the arterial phase is evaluated by comparing with the ground-truth ROI, and the ROI with $\geq$0.5 Jaccard similarity coefficient with ground-truth is regarded as correct. Our method achieves the promising performance on automatically detecting the lesion regions of all the three FLL types. This effectiveness may give rise to a computer-aided system assisting clinicians in diagnosis of such lesions.

We also visualize the results of our model for the three FLL types. Our model outputs three most discriminative ROIs in the three vascular phases for each FLL. In addition, we believe that the ROIs in the arterial phase which select the same local classifier for different CEUS videos, form a specific subtype, related with the different pattern variants of FLLs. Fig. 7 illustrates three discovered subtypes of HCC and two examples within each subtype are shown. Fig. 8 and Fig. 9 show the discovered subtypes and results of HEM and FNH types, respectively. The first column shows the annotated reference ROI provided by the radiologists. Obviously, the frame number of the ROI in the arterial phase detected by our algorithm (i.e. the inferred value for hidden variable t_1) could be different from the frame number of the annotated ROI picked by the radiologists, because our learning algorithm does not simply simulate what the radiologists would do, as in [16], but tries to find the most discriminative regions in terms of classification. Our algorithm does tend to pick up those visually discriminative ROIs, such as the edge-like regions and high-contrast regions, which play the most important role in recognizing the FLLs. These results clearly demonstrate that our model can automatically predict the locations of ROIs of FLLs, as well as the meaningful subtypes (i.e. the intrinsic variants of enhancement patterns within each type).

TABLE II
THE SENSITIVITIES (SENS. FOR SHORT) FOR ROI DETECTION OF HCC, HEM AND FNH TYPES, ACC IS THE MEAN ACCURACY FOR ALL FLL TYPES

	Sens^{HCC}	Sens^{HEM}	Sens^{FNH}	ACC
Ours	86.3%	84.5%	62.7%	**77.9%**
LMMC [26]	81.3%	82.5%	59.3%	**74.3%**
LSSVM [25]	82.1%	76.2%	58.7%	**72.3%**

To understand the advantages of our framework, we perform another three sets of experiments to further investigate the effectiveness and efficiency of our framework in terms of learning, inference and feature representation, respectively.

D. Discussions on Learning

First, we evaluate the effectiveness of different components of our model in Table I. By eliminating the regularization term and pairwise term, our model can be simplified as “Ours1” and “Ours2”, respectively. By comparing “Ours1” with our full version (“Ours”) on the accuracy of characterization, we can observe that the pairwise term makes the accuracy increase by 1.2% on average, especially by 9.1% for the FNH. The regularization term improves the average accuracy by 0.5% and the sensitivity of HCC by 4.3%, which shows that the unary score and pairwise similarity measures should be combined together when targeting on learning rich and flexible models. Our extended model also demonstrates superior performance over the previous method [38]. We denote the previous method [38] as “Ours3”, where the maximum number of local classifiers for

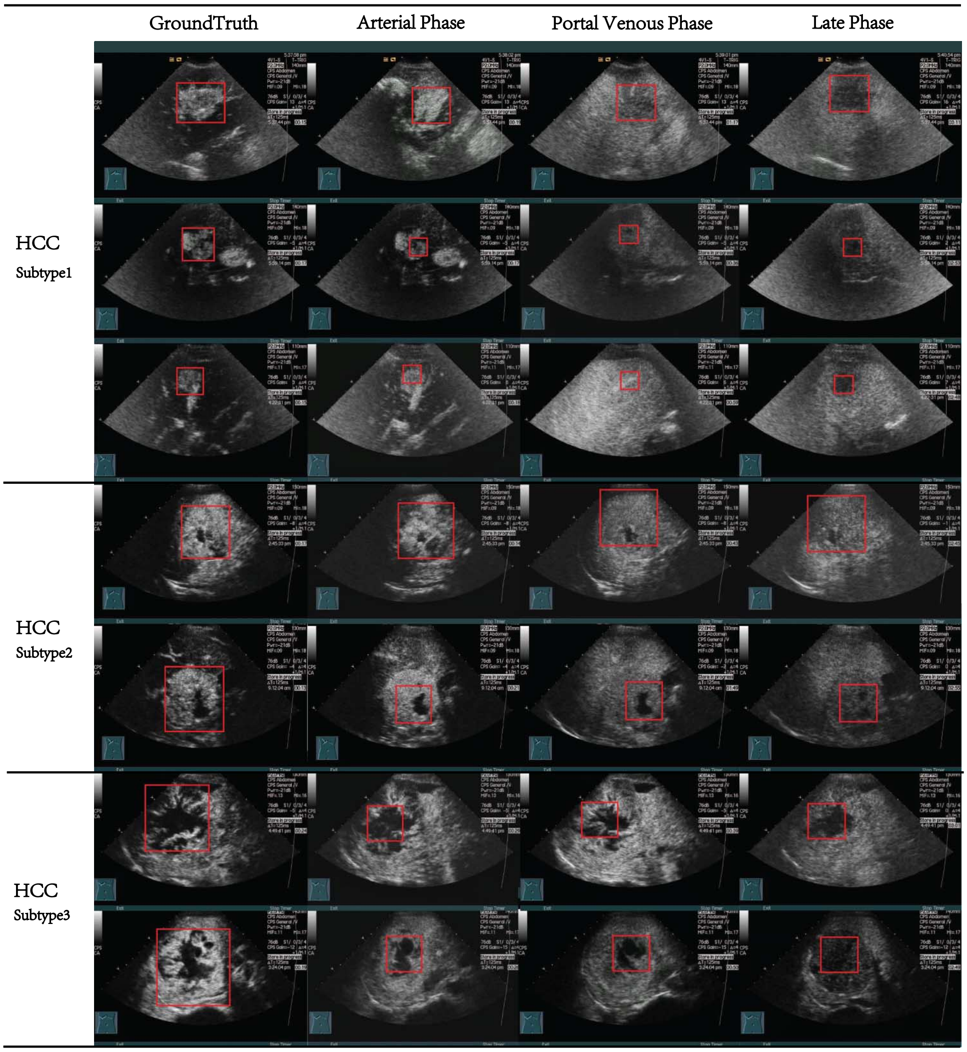

Fig. 7. Example results of HCC type. Our model has identified 3 subtypes (i.e. variations) for HCC, and the results of two examples for each subtype are illustrated. The ground truth ROIs annotated by the radiologists are displayed in the first column. The red boxes of different phases represent the localized discriminative ROIs.

ROIs in each phase is set as 1 to limit the model's capability of capturing variations. The mean accuracy of "Ours3" decreases by 2.4% compared to our proposed model "Ours". This demonstrates the effectiveness of the selective compositional characteristics of our model.

We test different learning algorithms for training our model. All classification results are listed in Table I, which shows that our proposed learning algorithm (named as "Ours") consistently outperforms the standard Latent Structural SVM (LSSVM) [25] and the Latent Max-Margin Clustering (LMMC) [26]. In particular, for LSSVM [25] used in our experiment, we optimize all the hidden variables by only maximizing the model score, that is, the second and third term defined in (11) are eliminated; for LMMC, the classifier selections are treated as the labeling assignments in [26]. Following [26], we first optimize all ROI layouts and then find the optimal label assignment. According to the results, our method improves LSSVM by 7.4% and LMMC by 9.4% on average for classifying malignant and benign FLLs. For multi-class recognition, the average accuracy of our method is higher than LSSVM by 2.8% and LMMC by 2.5%. In particular, the sensitivity of FNH of our method is superior to both competitors by 9.1% and sensitivity of benign is increased by 10.6% and 38.1%, respectively. The confusion matrices are also presented in Fig. 10. Our method distinguishes the malignant HCC from the benign HEM and FNH much better than the LSSVM and LMMC frameworks. This demonstrates well that learning by integrating together the unary score of each example and appearance similarities between examples can help exploit rich and more discriminative representation of videos. As reported in Table II, our method also achieves superior performance on detecting ROIs from lesion regions, i.e. 3.6% over LMMC and 5.6% over LSSVM.

We also investigated the effect of the growing parameter λ on the performance, which is used to gradually increase the weight of pairwise term with iterations. As displayed in Fig. 11(a), without the growing parameter, our model structure (i.e. the number of examples selecting each local classifier) will be converging immediately after 2 iterations. However,by empirically setting the growing parameter as $\lambda = 1.6$, the examples that select each local classifier will gradually adjusted with iterations. Furthermore, the accuracies reported as "Ours4" $(\lambda = 1)$ in Table I decrease by 1.2% than those obtained with $\lambda = 1.6$ for multi-class classification. Besides, the sensitivities of each FLL type in "Ours4" become imbalanced. In particular, sensi-

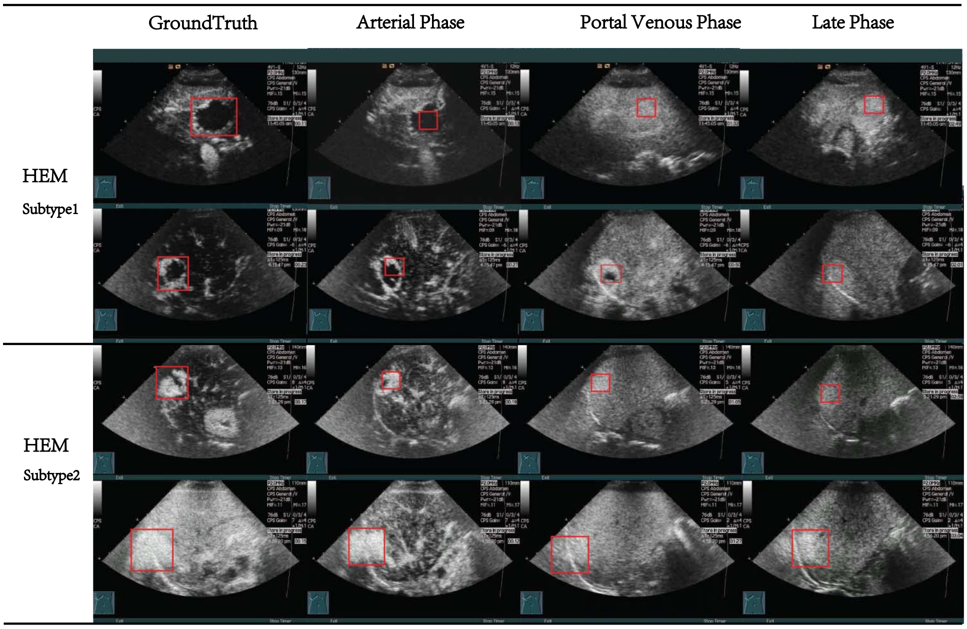

Fig. 8. Example results of HEM type. Our model has identified 2 subtypes (i.e. variations) for HEM, and the results of two examples are illustrated. The ground truth ROIs annotated by the radiologists are displayed in the first column. The red boxes of different phases represent the localised discriminative ROIs.

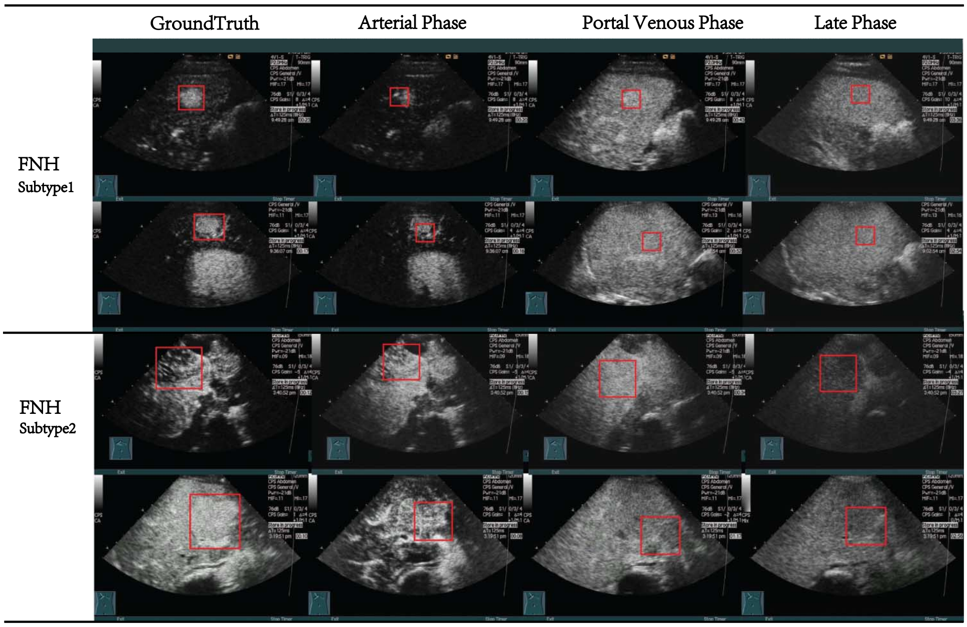

Fig. 9. Example results of FNH type. Our model has identified 2 subtypes (i.e. variations) for FNH, and the results of two examples are illustrated. The ground truth ROIs annotated by the radiologists are displayed in the first column. The red boxes of different phases represent the localized discriminative ROIs.

tivity of HEM decreases by 10.4% and that of HCC increases by 4.5%, which are more sensitive to the imbalance of data sizes for different types. Note that the larger value λ will lead to the longer convergence time for learning model structure, while has less impact on the accuracies. Thus we empirically set $\lambda = 1.6$, which is a trade-off between the accuracy and learning time.

E. Discussions on Inference

In this experiment, the performance of our data-driven inference algorithm is tested by altering the procedure of determining the ROIs, as shown in Table III. Our Data-Driven Inference ("DDI") algorithm is compared with 1) "manual1": the ROIs in the arterial phase are fixed by the annotation and the inference is only performed in the portal and late phases; 2) "manual2": the maximum number of local classifiers is limited as 1, and other settings are same as "manual1"; 3) "bruteforce": the liver region is labeled and the optimal ROIs are searched in the entire region of liver, without spatial-temporal pruning. The results demonstrate that our fully automatic inference algorithm achieves comparable performance to the "manual1" method, and performs better than "manual2" by 1.1% and "bruteforce" by 9.8%. Based on the candidate ROIs after the temporal and spatial pruning, the dynamic programming algorithm takes 7.3 seconds on average to infer the optimal locations of ROIs. The

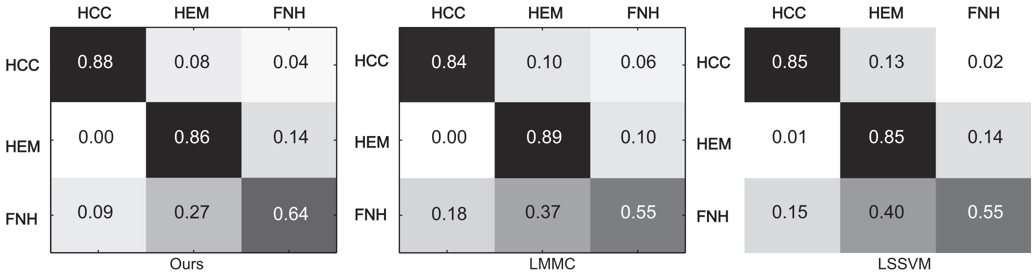

Fig. 10. Three confusion matrices on characterizing the FNH, HCC and HEM types by three different learning algorithms. FNH and HEM types among the benign lesions can be easily confused with each other due to the small inter-type differences. Our learning algorithm achieves better classification results than LMMC and LSSVM for the HCC and FNH.

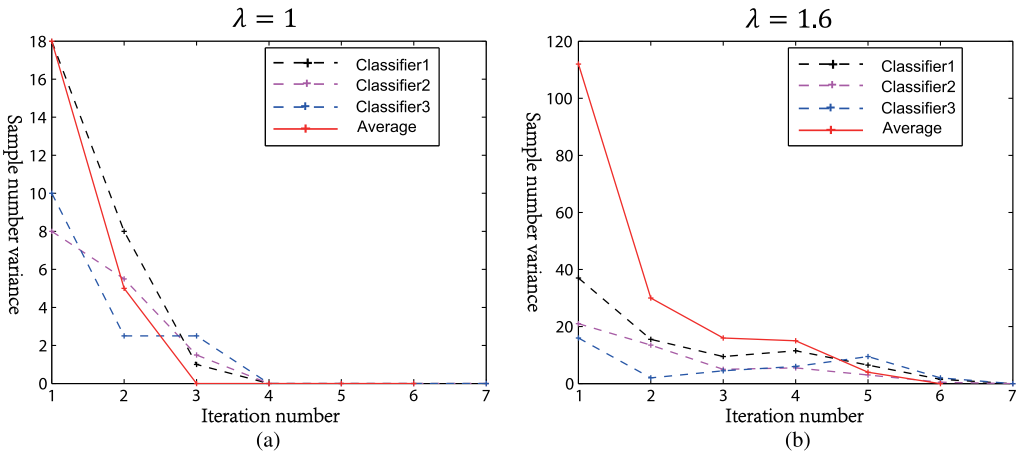

Fig. 11. Evaluation of the effectiveness of the growing parameter λ. We display the changes of the number of samples that select each local classifier in each iteration, denoted as the dashed line. The red solid line indicates the average change of all local classifiers. (a): for $\lambda = 1$, which means the weight of the pairwise terms has not iteratively adjusted; (b) for $\lambda = 1.6$, same setting as our complete model. The model structure will be converging gradually.

TABLE III
SENSITIVITIES AND MEAN ACCURACIES BY USING THE DIFFERENT INFERENCE STRATEGIES

	Sens^{HCC}	Sens^{HEM}	Sens^{FNH}	ACC
DDI	88.5%	86.2%	63.6%	84.8%
manual1	**90.4%**	82.8%	**72.7%**	**85.8%**
manual2	86.1%	**85.7%**	**72.7%**	83.7%
bruteforce	83.3%	80.1%	36.4%	75.0%

"manual1" takes about 5 seconds on average. Without using any pruning, the "bruteforce" method spends about 150 seconds, which is time-consuming.

As shown in Fig. 12, we also extensively evaluate how our algorithm performs under the setting of different step length in the spatial pruning. Specifically, this evaluation is conducted with two aspects: the mean accuracy for the benign/malignant classification and average testing time for processing a 4-min video. We set 11 different step lengths in the spatial pruning. Based on the results, we observe that larger step length leads to the decreased accuracy and shorter testing time. In this paper, we set the step length as 20.

F. Discussions on Feature Representation

Finally, we compare the region representation of our framework with other state-of-the-art methods: Multiple-ROI [12], ROI^{posterior} [52] and ROI^{out} [53]. Each region representation is tested with three most popular low-level features used for ultrasound images: GLCM [44], Law's texture [54], and Local Phase (LP) [55]. In particular, we extract 16 dimensions of GLCM features, exactly the same as described in the Section VI-B, 30 dimensions of Law's texture features, and 256 of Local Phase features for each single ROI. The whole feature vector used to represent a CEUS video is extracted and concatenated from three manually labeled ROIs in three phases. Note that we ignore the shape features since all FLLs often show circular and rounded shapes or even have unclear boundaries due to the low contrast. We manually select ROIs in three phases as required in previous works [12], [52], [53] (note here we do not consider the performance of the inference algorithm), and use linear SVM as the classifier. Table IV shows that our region representation achieves superior performance in general.

VII. CONCLUSION

In this work, we first propose a novel structured model to capture the large variations of FLLs in CEUS videos. A novel non-convex optimization algorithm is then proposed to iteratively optimize the model structure along with the parameter learning. An efficient data-driven inference method is presented for recognizing FLLs in videos with the trained model. The experimental results show very promising classification accuracies and we also demonstrate how the system components contribute to the overall performance.

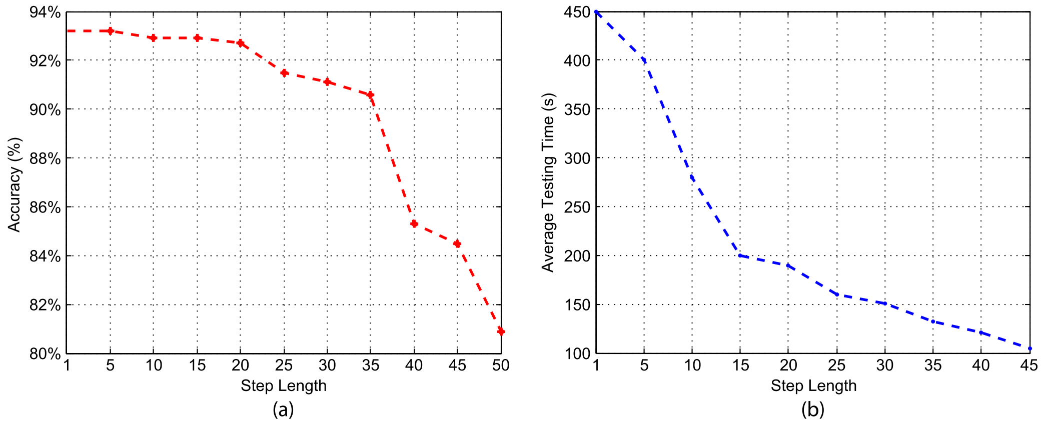

Fig. 12. Evaluation of mean accuracies for benign/malignant classification (a) and the average testing times (b) with different step length settings during the data-driven inference.

TABLE IV
COMPARISONS OF REGION REPRESENTATION METHODS BY APPLYING DIFFERENT FEATURE DESCRIPTORS

	Sens. HCC(%)	Sens. HEM(%)	Sens. FNH(%)	ACC1(%)	Sens.Benign(%)	Sens.Malignant(%)	ACC2(%)
Multiple ROI(GLCM)	85.9 %	75.9%	36.2%	74.7%	15.2%	**98.4 %**	59.1%
ROI^{posterior}(GLCM)	**88.1%**	67.5%	51.7%	75.8%	79.3%	80.5 %	80.0%
ROI^{out}(GLCM)	82.1%	61.1%	34.4%	67.8%	13.9%	96.2 %	57.4%
Ours(GLCM)	87.2%	**83.5%**	**67.2%**	**82.7%**	**84.2%**	89.7 %	**87.1%**
Multiple ROI(Law's)	82.7 %	75.9%	**72.4%**	78.9%	79.3%	**82.7%**	81.1%
ROI^{posterior}(Law's)	75.6%	77.7%	62.0%	74.2%	76.3%	73.5%	74.8%
ROI^{out}(Law's)	69.7%	72.2%	56.9%	68.4%	75.1%	74.3 %	74.8%
Ours(Law's)	**84.3%**	**87.2%**	67.6%	**82.4%**	**86.7%**	82.1 %	**84.2%**
Multiple ROI(LP)	85.9 %	67.5%	63.7%	76.5%	77.5%	83.7 %	80.8%
ROI^{posterior}(LP)	80.0%	69.4%	55.1%	72.7%	75.1%	77.8%	76.5%
ROI^{out}(LP)	78.9%	50.9%	48.2%	65.2%	69.7%	71.5 %	70.5%
Ours(LP)	**86.1%**	**73.4%**	**63.8%**s	**78.3%**	**79.4%**	**83.8 %**	**81.7%**

There are several directions in which we intend to extend this work. The first is to develop an interactive system based on our algorithm, which enables radiologists to revise the diagnosis according to the detected discriminative ROIs of FLLs in CEUS videos (e.g. the locations of ROIs and the reference frames). Second, we plan to integrate deep learning techniques (e.g. convolutional neural networks) into our framework, instead of using hand-crafted features. Moreover, our learning framework is very general to be applied to other pattern recognition tasks including large intraclass variations, e.g. activity analysis, object modeling, and scene understanding.

REFERENCES

[1] W. H. Organisation, "Fact sheets by population-globocan 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012," [Online]. Available: http://www.who.int/mediacentre/factsheets/fs297/en/

[2] J. A. Marrero, J. Ahn, and K. R. Reddy, "ACG clinical guideline: The diagnosis and management of focal liver lesions," *Am. J. Gastroenterol.*, 2014.

[3] J. Llovet, A. Burroughs, and J. Bruix, "Hepatocellular carcinoma," *Lancet*, vol. 362, no. 9399, pp. 1907–1917, 2003.

[4] G.-J. Liu *et al.*, "Real-time contrast-enhanced ultrasound imaging of focal liver lesions in fatty liver," *Clin. Imag.*, vol. 34, no. 3, pp. 211–221, 2010.

[5] D. Strobel *et al.*, "Contrast-enhanced ultrasound for the characterization of focal liver lesions-diagnostic accuracy in clinical practice," *Ultraschall Med.*, vol. 29, no. 5, pp. 499–505, 2008.

[6] M. Claudon *et al.*, "Guidelines and good clinical practice recommendations for contrast enhanced ultrasound (CEUS) in the Liver-pdate 2012," *Ultrasound Med. Biol.*, vol. 39, no. 2, pp. 187–210, 2013.

[7] S. R. Wilson and P. N. Burns, "Microbubble-enhanced US in body imaging: What role?," *Radiology*, vol. 257, no. 1, pp. 24–39, 2010.

[8] F. Piscaglia *et al.*, "Characterization of focal liver lesions with contrast-enhanced ultrasound," *Ultrasound Med. Biol.*, vol. 36, no. 4, pp. 531–550, 2010.

[9] T. Tan *et al.*, "Computer-aided lesion diagnosis in automated 3-D breast ultrasound using coronal spiculation," *IEEE Trans. Med. Imag.*, vol. 31, no. 5, pp. 1034–1042, May 2012.

[10] Y. Song, W. Cai, Y. Zhou, and D. Feng, "Feature-based image patch approximation for lung tissue classification," *IEEE Trans. Med. Imag.*, vol. 32, no. 4, pp. 797–808, Apr. 2013.

[11] J. A. Noble, "Ultrasound image segmentation and tissue characterization," *Proc. Inst. Mechan. Eng.*, vol. 224, no. 2, pp. 307–316, 2010.

[12] J. Jae Hyun, J. Y. Choi, S. Lee, and Y. M. Ro, "Multiple ROI selection based focal liver lesion classification in ultrasound images," *Expert Syst. Appl.*, vol. 40, no. 2, pp. 450–457, 2013.

[13] S. Junji, S. Katsutoshi, M. Fuminori, K. Naohisa, and D. Kunio, "Computer-aided diagnosis for the classification of focal liver lesions by use of contrast-enhanced ultrasonography," *Med. Phys.*, vol. 35, no. 5, pp. 1734–1746, 2008.

[14] A. Anaya *et al.*, "Differentiation of focal liver lesions: Usefulness of parametric imaging with contrast-enhanced US," *Radiology*, vol. 261, no. 1, pp. 300–310, 2011.

[15] S. Bakas *et al.*, "Histogram-based motion segmentation and characterisation of focal liver lesions in CEUS," *Ann. BMVA.*, vol. 7, pp. 1–14, 2012.

[16] S. Bakas, G. Hunter, D. Makris, and C. Thiebaud, "Spot the best frame: Towards intelligent automated selection of the optimal frame for initialisation of focal liver lesion candidates in contrast-enhanced ultrasound video sequences," *Intell. Environ.*, pp. 196–203, 2013.

[17] S. Bakas *et al.*, "Non-invasive offline characterisation of contrast-enhanced ultrasound evaluations of focal liver lesions: Dynamic assessment using a new tracking method," *Eur. Congr. Radiol.* 2014 [Online]. Available: http://dx.doi.org/10.1594/ecr2014/C-1378

[18] S. Bakas *et al.*, "Focal liver lesion tracking in CEUS for characterisation based on dynamic behaviour," *Adv. Vis. Comput.*, pp. 32–41, 2012.

[19] N. G. Rognin *et al.*, "Parametric imaging for characterizing focal liver lesions in contrast-enhanced ultrasound," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 57, no. 11, pp. 2503–2511, Nov. 2010.

[20] X. Wang and L. Lin, "Dynamical and-or graph learning for object shape modeling and detection," *Adv. Neural Inf. Process. Syst.*, pp. 242–250, 2012.

[21] L. Lin, X. Wang, W. Yang, and J.-H. Lai, "Discriminatively trained and-or graph models for object shape detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 959–972, May 2015.

[22] L. Lin, T. Wu, J. Porway, and Z. Xu, "A stochastic graph grammar for compositional object representation and recognition," *Pattern Recognit.*, vol. 42, no. 7, pp. 1297–1307, 2009.

[23] S.-C. Zhu and D. Mumford, "A stochastic grammar of images," *Foundat. Trends Comput. Graph Vis.*, vol. 2, no. 4, pp. 259–362, 2006.

[24] X. Liang, L. Lin, and L. Cao, "Learning latent spatio-temporal compositional model for human action recognition," in *ACM Int. Conf. Multimedia*, 2013.

[25] C.-N. J. Yu and T. Joachims, "Learning structural SVMs with latent variables," in *Int. Conf. Mach. Learn.*, 2009, pp. 1169–1176.

[26] Z. Guang-Tong, L. Tian, V. Arash, and M. Greg, "Latent maximum margin clustering," *Adv. Neural Inf. Process. Syst.*, pp. 28–36, 2013.

[27] C. Huang-Wei *et al.*, "Differential diagnosis of focal nodular hyperplasia with quantitative parametric analysis in contrast-enhanced sonography," *Invest. Radiol.*, vol. 41, no. 3, pp. 363–368, 2006.

[28] S. Bakas *et al.*, "Fast semi-automatic segmentation of focal liver lesions in contrast-enhanced ultrasound, based on a probabilistic model," *Comput. Methods Biomechan. Biomed. Eng., Imag. Visualizat.*, pp. 1–10, 2015.

[29] H. D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognit.*, vol. 43, no. 1, pp. 299–317, 2010.

[30] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: A survey," *IEEE Trans. Med. Imag.*, vol. 25, no. 8, pp. 987–1010, Aug. 2006.

[31] K. Drukker, M. L. Giger, and E. B. Mendelson, "Computerized analysis of shadowing on breast ultrasound for improved lesion detection," *Med. Phys.*, vol. 30, p. 1833, 2003.

[32] R.-F. Chang *et al.*, "Rapid image stitching and computer-aided detection for multipass automated breast ultrasound," *Med. Phys.*, vol. 37, no. 5, pp. 2063–2073, 2010.

[33] P. Vos, J. Barentsz, N. Karssemeijer, and H. Huisman, "Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis," *Phys. Med. Biol.*, vol. 57, no. 6, p. 1527, 2012.

[34] X. Ye, X. Lin, J. Dehmeshki, G. Slabaugh, and G. Beddoe, "Shape-based computer-aided detection of lung nodules in thoracic CT images," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 7, pp. 1810–1820, Jul. 2009.

[35] W. K. Moon *et al.*, "Computer-aided tumor detection based on multi-scale blob detection algorithm in automated breast ultrasound images," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1191–1200, Jul. 2013.

[36] B. Liu *et al.*, "Fully automatic and segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images," *Pattern Recognit.*, vol. 43, no. 1, pp. 280–298, 2010.

[37] J. Ding, H. Cheng, J. Huang, J. Liu, and Y. Zhang, "Breast ultrasound image classification based on multiple-instance learning," *J. Dig. Imag.*, vol. 25, no. 5, pp. 620–627, 2012.

[38] X. Liang, Q. Cao, R. Huang, and L. Lin, "Recognizing focal liver lesions in contrast-enhanced ultrasound with discriminatively trained spatio-temporal model," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2014, pp. 1184–1187.

[39] X. Liang *et al.*, "Deep human parsing with active template regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2402–2414, Dec. 2015.

[40] X. Liang *et al.*, "Human parsing with contextualized convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1419–1427.

[41] X. Liang, L. Lin, and L. Cao, "Learning latent spatio-temporal compositional model for human action recognition," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 263–272.

[42] L. Lin, Y. Xu, X. Liang, and J. Lai, "Complex background subtraction by pursuing dynamic spatio-temporal models," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3191–3202, Jul. 2014.

[43] L. Lin, Y. Lu, Y. Pan, and X. Chen, "Integrating graph partitioning and matching for trajectory analysis in video surveillance," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4844–4857, Dec. 2012.

[44] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[45] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Eur. Conf. Comput. Vis.*, 2012, pp. 116–129.

[46] D. Koller and N. Friedman*, Probabilistic Graphical Models-Principles and Techniques*. Cambridge, MA: MIT Press, 2009.

[47] A. L. Yuille and A. Rangarajan, "The concave-convex procedure (CCCP)," *Adv. Neural Inf. Process. Syst.*, pp. 1033–1040, 2002.

[48] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[49] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov, "Fast approximate energy minimization with label costs," *Int. J. Comput. Vis.*, vol. 96, no. 1, pp. 1–27, 2012.

[50] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Graph cut based inference with co-occurrence statistics," in *Eur. Conf. Comput. Vis.*, 2010, pp. 239–253.

[51] E. Quaia *et al.*, "Characterization of focal liver lesions with contrast-specific US modes and a sulfur hexafluoride-filled microbubble contrast agent: Diagnostic performance and confidence," *Radiology*, vol. 232, no. 2, pp. 420–430, 2004.

[52] S. Kim *et al.*, "Computer-aided image analysis of focal hepatic lesions in ultrasonography: Preliminary results," *Abdominal Imag.*, vol. 34, no. 2, pp. 183–191, 2009.

[53] G. Xian, "An identification method of malignant and benign liver tumors from ultrasonography based on GLCM texture features and fuzzy SVM," *Expert Syst. Appl.*, vol. 37, pp. 6737–6741, 2010.

[54] K. I. Laws, "Rapid texture identification," in *24th Annu. Tech. Symp.*, 1980, pp. 376–381.

[55] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Trans. Signal Process.*, vol. 49, no. 12, pp. 3136–3144, Dec. 2001.