Mitochondria Segmentation From EM Images via Hierarchical Structured Contextual Forest

Jialin Peng and Zhimin Yuan

***Abstract*—Delineation of mitochondria from electron microscopy (EM) images is crucial to investigate its morphology and distribution, which are directly linked to neural dysfunction. However, it is a challenging task due to the varied appearances, sizes and shapes of mitochondria, and complicated surrounding structures. Exploiting sufficient contextual information about interactions in extended neighborhood is crucial to address the challenges. To this end, we introduce a novel class of contextual features, namely local patch pattern (LPP), to eliminate the ambiguity of local appearance and texture features. To achieve accurate segmentation, we propose an automatic method by iterative learning of hierarchical structured contextual forest. With a novel median fusion strategy, the probability predictions from long history iterations are augmented to encode spatial and temporal contexts and suppress false detections. Moreover, the LPP features are extracted on both images and history predictions, resulting in a hierarchy of contextual features with increasing receptive fields. Other than using computationally demanding graph based methods, we perform joint label prediction using structured random forest. In addition to direct 3D segmentation of EM volumes, we introduce a 2D variant without sacrificing accuracy using a novel hierarchical multi-view fusion strategy. We evaluated our proposed methods on public EPFL Hippocampus benchmark, achieving state-of-the-art performance of 90.9% in Dice. Quantitative comparison showed the effectiveness of the proposed features and strategies.**

***Index Terms*—Segmentation, electron microscopy image, contextual features, hierarchical learning, multi-view fusion.**

I. Introduction

CONNECTOMICS is an emerging domain of neuroscience, seeking to understand connections between a brain's functions and its neuronal structure. At nano-scale, electron microscope (EM) is one of the state-of-the-art imaging devices for the reconstruction of neural circuits. It is capable of investigating the ultrastructure of cell, where mitochondria are subcellular organelle and have been directly linked to aging, cancer, and neurodegenerative diseases [1]. Reconstruction of neuron structures from EM image enables a better understanding the basic cognitive functions of the brain. However, the high resolution of EM leads to image data sets at huge scale. Being a critical step for neuron structure reconstruction, volume EM image segmentation is extremely laborious, time-consuming and also subjective, requiring months of effort of annotators for manual labeling [2].

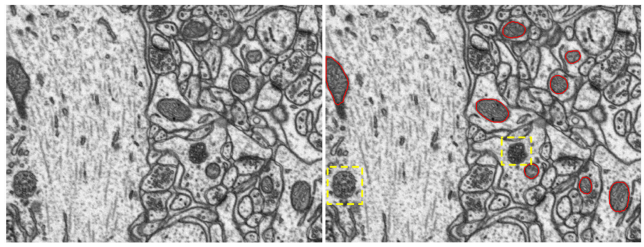


Fig. 1. Illustration of challenges in mitochondria segmentation (Red contours: manual segmentation). Many sub-structures in EM image share similar intensities and shapes with mitochondria, which are highlighted with boxes.

Automation of image segmentation is crucial and highly desirable, but challenging, as illustrated in Fig. 1. In EM images, mitochondria show large variability in density, location, size and shape. The anatomical details in cells such as membranes and intracellular structures make the segmentation more complicated. For example, many sub-structures in EM image share similar intensities with mitochondria, which are highlighted with boxes in Fig. 1; strong gradients do not necessarily correspond to the semantic boundaries of mitochondria. As a result, it can be quite difficult to distinguish mitochondria from other structures based solely on local intensities and textures. Therefore, extraction of highly discriminating features is crucial for a machine learning based method to automatically determine the presence of mitochondria at a given position and delineate their boundaries accurately.

Detection and segmentation of mitochondria from EM images have recently attracted a variety of studies [3]–[10]. Extraction of highly discriminative features is the key step to address this problem [11]–[13]. Early works [7], [11], [14], [15] were based on general texture features in the field of computer vision, e.g., filter-banks, texton, Haar-like descriptors [16], local binary pattern (LBP) [17]. Since EM images are cluttered with structures exhibiting similar intensities and textures, it is difficult to distinguish structures solely based on local image statistics.

Manuscript received September 16, 2019; revised November 23, 2019; accepted December 19, 2019. Date of publication December 23, 2019; date of current version August 5, 2020. This work was supported in part by the NSFC under Grant 11771160, in part by STPF (2019H0016), and in part by the Fund of HQU (ZQN-PY411). *(Corresponding authors: Jialin Peng; Zhimin Yuan.)*

The authors are with the College of Computer Science and Technology, Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen 361021, China (e-mail: 2004pjl@163.com; yuanzhimin2018@163.com).

Digital Object Identifier 10.1109/JBHI.2019.2961792

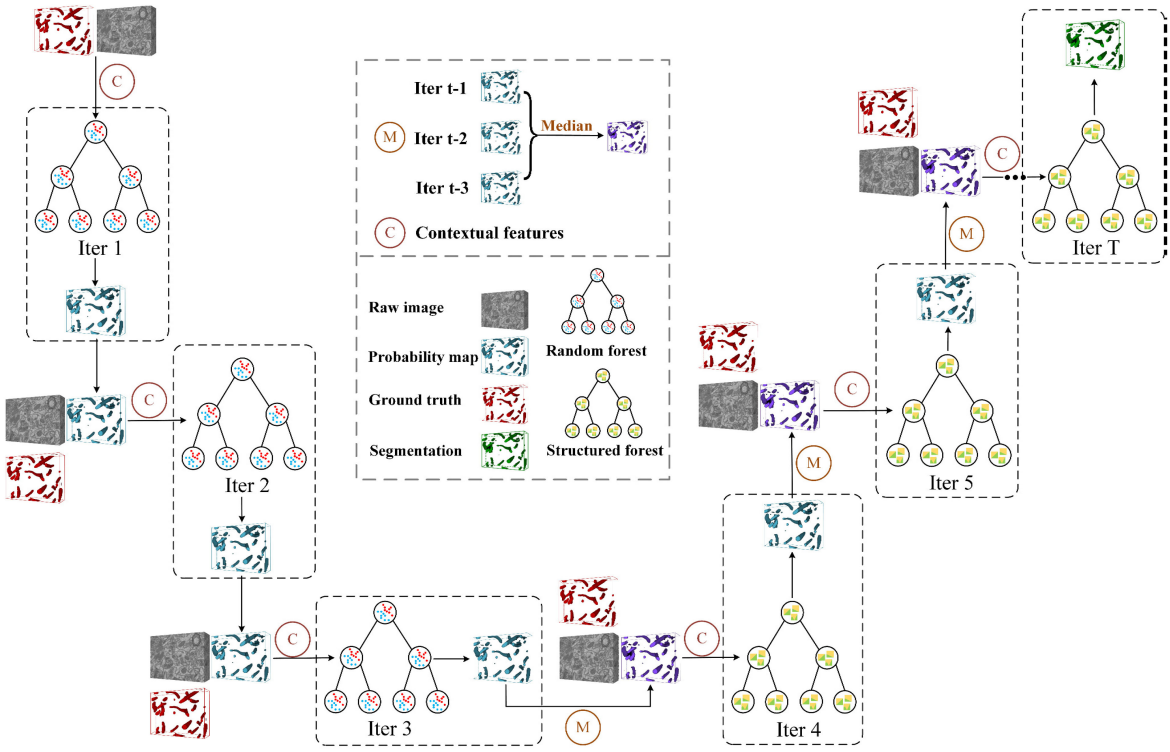


Fig. 2. Overview of our proposed method. To capture contextual information, a novel class of features named LPP are introduced. The proposed pipeline uses an iterative learning strategy to progressively obtain finer segmentation and extract hierarchical LPP features. At each iteration, except for original images, probabilistic predictions from long history (e.g., last three iterations) are incorporated using a median fusion strategy to encode spatial and temporal contexts and suppress false detections. Since local image labels are highly interdependent, structured random forest is employed to perform local structured prediction.

Recent studies introduced specifically designed geometrical and contextual features for mitochondria segmentation, such as Ray feature [3], Radon-Like features [6], etc. Similar to the findings in computer vision, with contextual cues, these methods have shown greatly improved discriminative ability. However, these geometrical/spatial contextual features heavily rely on edge detectors or strong gradients, which may not necessarily correspond to semantic boundaries. Thus, it is desirable to introduce informative contextual features with computational efficiency and robustness.

Data-driven label prediction typically treats image segmentation as a pixel-wise classification problem with a variety of powerful classifiers, e.g., random forest [18], AdaBoost [19], support vector machine (SVM) [20]. However, these methods usually neglect the high interdependency among local image labels. Conditional random fields (CRF) [21] with hand-crafted or learned parameters have been usually employed to enforce structural dependencies between neighboring pixels/voxels of EM images [13], [22]–[25]. However, not only the parameter learning [13] but also the optimization of CRF [21] are complicated inference problems and computationally demanding for volume data.

To overcome the drawbacks mentioned above, we propose a cascade of structured contextual forest to address the 3D detection and delineation of mitochondria. To capture contextual information, we introduce a novel class of computationally efficient features, namely *local patch pattern* (LPP), which is an extension of LBP features and Haar-like features. The overview of our approach is shown in Fig. 2 and also Fig. 3. In particular, we follow the iterative learning strategy of iterated contextual classification [26] and auto-context [27] to progressively obtain finer segmentation. With a novel median fusion strategy, the label predictions from long history iterations are taken into account

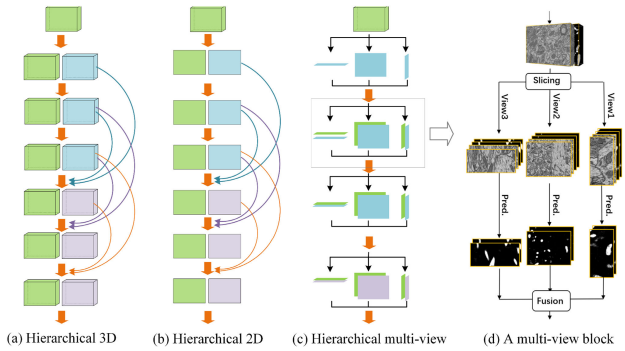


Fig. 3. Illustration of the proposed hierarchical models and its multi-view fusion variant. (a) The proposed hierarchical 3D segmentation with median fusion of long history predictions; (b) hierarchical 2D segmentation using single view; (c) the proposed hierarchical multi-view segmentation; (d) a multi-view fusion block in (c). For both 2D and 3D methods, as well as the pseudo 3D variant (c), iterative learning strategy, where classifiers of each step utilize features not only from input images (green boxes) but also from the predictions (blue and purple boxes) of previous or/and multi-view classifiers, is applied. Purple boxes denote the predictions using three previous predictions.

in conjunction with the image itself for subsequent prediction. Note that the LPP features are extracted not only from raw images but also from predictions of previous iterations, resulting in a class of hierarchical context features with the ability to encode both middle-range and long-range spatial contexts. In addition, we capture structural label dependencies between neighboring voxels with a structured random forest [28], which is much more computationally efficiency than graph based CRF methods for 3D segmentation. As a novel extension of our 3D method, we also introduce a novel hierarchical multi-view method shown in Fig. 3 to trade off memory efficiency and segmentation accuracy.

A. Related Work

Early works for EM image segmentation were based on general texture features in the field of computer vision. For example, Narasimha *et al.* [14] explored a filter-bank consist of Gaussian filters and Laplacian of Gaussian filters, as well as texton-based region features as texture feature encoders. Kaynig *et al.* [15] used Haar-like descriptors as well as histograms of intensities as features. For mitochondria segmentation, Neila *et al.* [7] utilized a set of linear operators (zero, first and second order derivatives) of the image at several scales as visual features. The same class of features were also used in [4] for synapses and mitochondria segmentation. In [11], Cetina *et al.* compared a set of popular features for mitochondria segmentation, including histogram of intensity, histogram of oriented gradients (HOG), LBP, Gaussian filters at different scales, difference of Gaussians, Laplacian of smoothed image, eigenvalues of the structure tensor, and so on. To take spatial context into account for synapse segmentation from EM images, Becker *et al.* [12] firstly run large amount of Gaussian filters on various location inside an extended neighborhood of the voxel to be classified, and then used AdaBoost to select the relevant filter channels as well as the relevant spatial locations as context cues.

Specifically designed features describing geometrical and spatial context have also introduced. To capture deformable or irregular shapes like mitochondria, Smith *et al.* [9] proposed a class of geometrical features, namely Ray features, which can provide a description of the local shape relative to a given location. A 3D extension of Ray features based on supervoxels were used in [3]. For mitochondria segmentation, Kumar *et al.* [6] proposed a novel class of hybrid features, called Radon-Like features, that aimed to aggregate both textural cues (e.g. image intensities) and geometric information (e.g. cell boundaries). Seyedhosseini *et al.* [29] extracted shape features together with texture features by fitting algebraic curves, of different degrees, on local image patches of different sizes. However, these contextual features heavily rely on edges or strong gradients and are not robust due to the presence of complex intracellular structures and nosies.

As for segmentation of EM images, Seyedhosseini *et al.* [29] used random forest for pixel-wise labeling. Cetina *et al.* [4] used boosting to label of each voxel separately, the results of which were refined using CRF. Rigamonti *et al.* [30] proposed to use a cascade of an improved KernelBoost classifier to progressively focus more and more on difficult-to-classify samples. Additionally, the detection of mitochondria membranes were included to boost the segmentation. Since enforcing the dependencies between neighboring pixels is extremely useful, Lucchi *et al.* [3] used 3D pairwise CRF on super-voxels with boundary and appearance terms learned by SVM. The 3D CRF with pairwise potential was computed with graph cut algorithm, which is memory demanding for large volume data. A novel random field with higher-order potential was proposed in [24] to encourage pixels in a local patch to take a joint labeling. Since a general higher-order CRF is difficult to solve, they converted the higher-order potentials to a pair-wise form and solved it using traditional inference algorithms. CRF typically involves a large number of parameters that must be carefully chosen to achieve good performance. In [8], [13], [23], Lucchi *et al.* employed Structured SVM (SSVM) to learn the parameters of CRF from data with another complicated inference problem. Deep fully convolutional neural networks have also been used for EM segmentation [10], [31], which requires large amount of training data and data augmentation. Although the graph based CRF methods have shown state-of-the-art results, they are limited by the high computational cost and memory demanding. As a result, we adopt the fully data-driven random forest to conduct structured label prediction.

II. Methods

An overview of our hierarchical structured learning method is illustrated in Fig. 2 and Fig. 3. Basically, we take 3D volume data as input to take advantage of 3D spatial contexts. However, the proposed method is flexible to apply on both 2D slices and 3D volumes. In this section, we present the details of our approach, including a class of novel contextual features, techniques used to perform hierarchical structured learning, and also a pseudo 3D extension with a hierarchical multi-view fusion strategy.

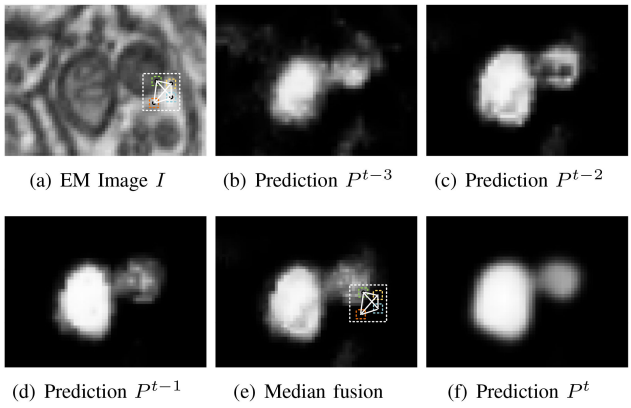
A. Hierarchically Structured Learning Framework

3D image segmentation can be defined as a voxel labelling problem with possible structural dependencies between nearby voxels. Let I be an image composed of m voxels $\{x_i, i = 1, 2, \ldots, m\}$. The aim is to infer a label image Y such that each voxel x_i is assigned a corresponding label $y_i = c \in \{0, 1\}$. Voxels taking the same label belong to the same segment of the image. Finding the optimal configuration Y^* out of all possible segmentations is defined in terms of a maximization of the posterior probability given the observed image, i.e.,

$$Y^* = \arg\max_Y \Pr(Y|I). \tag{1}$$

When treating image segmentation as an voxel-wise independent classification problem, the labels $y_i, i = 1, 2, \ldots, m$ are typically assumed independent. In the case of this scenario, the label y of voxel x is conditioned only on a local patch $N(x)$ of image voxels centred around the voxel x, which results in the probability model $\Pr(y|I_{N(x)})$.

Structured prediction. In this study, we seek the joint prediction of the label y and the labels $Y_{N'(x)}$ in its extended neighborhood $N'(x)$, which is modeled as $\Pr(Y_{N'(x)}|I_{N(x)})$



(a) EM Image I (b) Prediction P^{t-3} (c) Prediction P^{t-2}

(d) Prediction P^{t-1} (e) Median fusion (f) Prediction P^{t}

Fig. 4. Hierarchical learning with a median fusion strategy: (a) the EM image I; (b), (c) and (d) are probability predictions of the last three steps; (e) the fused prediction of (b), (c) and (d); (f) the new prediction based on (a) and (e). Appearance and contextual features are extracted on the image (a) and the fused prediction (e).

or $\Pr(y|I_{N(x)}, Y_{N'(x)\setminus x})$. Please note that the size of the local patch $N'(x)$ may differ from the size of $N(x)$. In other words, whereas traditional methods associate only the center label y to an image patch $I_{N(x)}$, we seek to incorporate the topology of the neighboring labels and perform structured prediction. The final segmentation is obtained by aggregating all predictions with voting. Other than utilizing CRF to enforce spatial consistency constraints on labeling, we employ structured random forest (SRF) [28], which is a simple and effective way to integrate structural information in random forest. The more detailed description of SRF is postponed to Section II-D.

Hierarchical learning with median fused long history predictions. We use an iterative learning strategy to progressively obtain finer segmentation and extract higher level features. We accomplish this by introducing an iterative time-step t and updating segmentations by reclassifying each voxel based on features from the original images and additionally, the probability maps P predicted by previous classifiers, as shown in Fig. 3(a) and Fig. 4. At each iteration, we use SRF as the discriminating classifier to perform structured prediction.

Formally, the probability prediction P_x^t of the voxel x at iteration t is based on previous predictions, $P_x^t = \Pr(y|I_{N(x)}, Y_{N'(x)\setminus x}, P_{N(x)}^{\tau}, 1 \le \tau < t)$. Other than using only the prediction at time step $t-1$, we use the last three predictions with a *voxel-wise median fusion* strategy, as illustrated in Fig. 4. Specifically, the proposed hierarchical learning is as follows,

$$P_x^t = \Pr\Big(y|I_{N(x)}, Y_{N'(x)\setminus x}, median\left\{P_{N(x)}^{\tau}, t-3 \le \tau < t\right\}\Big). \quad (2)$$

For $1 < t < 3$, we include all one or/and all zero predictions to perform median fusion.

Incorporating history prediction can help capture structured contextual information about the interactions between neighboring predictions. The advantages of the median fusion strategy are twofold: 1) we can incorporate longer history predictions with extremely low cost; 2) it can help suppress false detections, as shown in Fig. 4, since iterative refinement using only prediction from one previous step prediction of single step may re-enforce an incorrect segmentation.

Suppose the procedure stops after T iterations, at testing stage a novel image goes through this hierarchical iterative learning process using all predictions $\{P^t, 1 \le t < T\}$. Therefore, the final segmentation Y^* can be obtained through the thresholding of the Tth prediction P^T.

B. Hierarchical Multi-View Fusion: A Pseudo 3D Variant

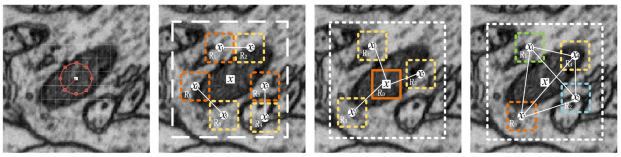
We extend our 3D segmentation method (Fig. 3(a)) to *a pseudo 3D method* with a hierarchical multi-view fusion strategy shown in Fig. 3(c), which takes advantage of both the memory efficiency of 2D methods and rich 3D spatial context of 3D methods. It is based on the observation that 3D segmentation methods are more memory demanding than 2D counterpart due to high dimensional features, whereas slice-by-slice 2D segmentation (Fig. 3(b)) neglects the inter-slice continuity in medical volumes. Formally, at each iteration t, we perform 2D segmentations on slices from three orthogonal views independently, as shown in Fig. 3(d). Then, the resulted three probability predictions $\{P_{\text{view}_i}^t, i = 1, 2, 3\}$ are fused into a single prediction with voxel-wise median fusion $P^t = median\{P_{\text{view}_i}^t, i = 1, 2, 3\}$, the output of which is used to help the subsequent refinements. This hierarchichal strategy is significantly different from the multi-view fusion strategy which just performs final fusion of several 2D predictions.

C. Hierarchical Feature Extraction

Extraction of expressive features is crucial for the discriminating of mitochondria from background with complicated contents. In addition to intensity and gradient based local appearance features, a novel class of informative, fast and robust context features, namely local patch pattern (LPP) features are introduced to encode interactions in local window patch. Formally, given a $n \times n \times n$ local window $N(x)$ centered at the voxel x, the feature vector $\mathbf{v}(x)$ of a 3D voxel x is a collection of appearance and LPP features detailed as follows.

Appearance features. Since structures in EM images are directly characterized by different intensity values and edges, it makes sense to include intensity and gradient statistics over the local patch window to construct visual features. Formally, we use the *grayscale intensity* at center voxel x, *average, median, variance, skew, kurtosis* of *intensity, gradient, and gradient magnitude* over the window $N(x)$, which are simple, robust and computationally efficient.

Proposed LPP features. The LPP features are extensions of the popular Haar-like and LBP features. They focus on encoding interaction and association between each voxel of the image and its extended neighborhood, which provide rich spatial contextual information to eliminate the ambiguity of the local appearance and texture features. The LPP features contain three subtypes, i.e. LPP-I, LPP-II and LPP-III, which are calculated by comparing sparsely distributed cubical sub-regions. Given a $n \times n \times n$ local window $N(x)$ centered at the voxel x, a set of $S \times (Q+1)$ cuboidal sub-regions $\{R_q^s, q = 0, 1, \ldots, Q\}$ with different sizes



(a) LBP (b) Haar (c) LPP-I (d) LPP-II

Fig. 5. 2D Illustration of LBP feature, Haar feature and the proposed LPP features. These features can be extracted on image data of any dimension.

($S = 2$, i.e., $3 \times 3 \times 3$ and $5 \times 5 \times 5$) and different offsets are extracted. Particularly, R_0^s is the one centered at x.

a) **LPP-I features** are computed by comparing average intensities of the sub-regions $\{R_q^s, q = 1, \ldots, Q\}$ and R_0^s, as demonstrated in Fig. 5(c). We record both the real differences and binary comparison results, resulting in LPP-Ir and LPP-Ib features,

$$\text{LPP-Ir}(x, I, s)[q] = \frac{1}{|R_q^s|} \sum_{u \in R_q^s} I_x(u) - \frac{1}{|R_0^s|} \sum_{u \in R_0^s} I_x(u), \quad (3)$$

$$\text{LPP-Ib}(x, I, s) = \sum_{q=1}^{Q} 2^{q-1} \delta \left(\sum_{u \in R_q^s} I_x(u) - \sum_{u \in R_0^s} I_x(u) \right), \quad (4)$$

where $\delta(z) = 1$, if $z \geq 1$; $\delta(z) = 0$ otherwise. Repeating the computation of LPP-Ib in different orders of $\{R_q^s, q = 1, \ldots, Q\}$ generates a rich set of codes that account for rotated patterns. Similar to the widely used texture encoder LBP, LPP-I features can describe textural context. However, this LPP-I feature is obviously different from LBP feature, which is computed by binary test comparing the centering voxel with its L neighbor voxels, as shown in Fig. 5(a). Specifically, 1) using average intensity in sub-regions instead of single pixel values makes the LPP features be robust to noises; 2) larger and irregular patterns can be described with LPP-I features.

b) **LPP-II features** illustrated in Fig. 5(d) are extension of traditional Haar features (Fig. 5(b)) to compare randomly selected pairs of nonadjacent sub-region R_p and R_q from the local image patch (such as R_1 and R_2 that are shown in Fig. 5(d)).

$$\text{LPP-II}(x, I, s)[p, q] = \frac{1}{|R_p^s|} \sum_{u \in R_p^s} I_x(u) - \frac{1}{|R_q^s|} \sum_{u \in R_q^s} I_x(u). \quad (5)$$

Repeating this operation can generate a rich set of features that are robust against the strong noises in EM images. This LPP-II features can describe spatial contextual information, including pairwise interactions of neighbouring structures. In contrast, the classical Haar features only calculate the difference of sub-regions of adjacent positions, resulting in features only encoding local contrasts as shown in Fig. 5(b).

c) **LPP-III features** calculate the average intensities over the sub-regions $\{R_q^s, q = 1, \ldots, Q, s = 1, 2\}$.

$$\text{LPP-III}(x, I, s)[q] = \frac{1}{|R_q^s|} \sum_{u \in R_q^s} I_x(u). \quad (6)$$

All of the LPP features can be computed with efficiency using integral image [16] as Haar features.

Hierarchical contextual features. To incorporate longer-range spatial contexts and the information on score maps coming from previous iterations, we extract the same class of LPP features, which are different from the context features used in [27], on score maps of the fused history predictions. As illustrated in Fig. 4, given a fused probability prediction P (Fig. 4(e)) from previous steps (Fig. 4(b)–(d)), we extracted [LPP-Ir(x, P),LPP-Ib(x, P),LPP-II(x, P),LPP-III(x, P)] from each voxel, which are augmented to image features for subsequent refinements. In this way, we obtain a hierarchy of contextual features with increasing receptive fields.

D. Structured Prediction With Random Forest

Since local labels are highly interdependent, we perform structured prediction using structured random forest [28].

Random forest is an ensemble of K independent tree-structured classifiers $\{f_k(\mathbf{v}), k = 1, 2, \ldots, K\}$. Each classifier $f_k(\mathbf{v})$ classifies a sample in feature vector $\mathbf{v}$ by recursively branching left or right down the tree until reaching a leaf node. Each splitting node j has an associated test function $h(\mathbf{v}, \boldsymbol{\theta}_j) \in \{0, 1\}$, such as simple decision stump used in our experiments, which sends the sample $\mathbf{v}$ to its left or right child node. As the data sample reaches a leaf node, the leaf node casts a class prediction $\Pr(c|\mathbf{v}, f_k)$ to the sample. The final prediction is achieved through aggregating all the predictions,

$$y^* = \arg\max_y \left\{ \Pr(y|\mathbf{v}) = \frac{1}{K} \sum_{k=1}^{K} \Pr(y|\mathbf{v}, f_k) \right\}. \quad (7)$$

Structured forest. To perform structured label prediction using random forest, it stores all the labels $Y_{N'(x)}$ of a local patch $N'(x)$(smaller than $N(x)$) at leaf node [28]. At each splitting node, label from a random position in $N'(x)$ is used in the Gini impurity criterion, resulting in the same parameter learning procedure as random forest. Let $\mathcal{Y}_k$ be the label patch set that present at leaf node reached by $N(x)$ in kth tree. The prediction of $Y_{N'(x)}^k$ on the leaf node for the patch $I_{N(x)}$ is achieved by maximizing the joint probability over $N'(x)$ with pixel independence assumption $\Pr(Y_{N'(x)}|\mathcal{Y}_k) = \prod_{u \in N'(x)} \Pr(y_u|\mathcal{Y}_k)$. Following the similar idea, the patch predictions gathered from all the K trees can be combined into a single label patch,

$$Y_{N'(x)}^* = \arg\max_{Y_{N'(x)}} \left\{ \prod_{u \in N'(x)} \Pr\left(y_u | \left\{ Y_{N'(x)}^k \right\}_{k=1}^{K} \right) \right\}. \quad (8)$$

The final segmentation for whole image is computed by ensembling all the structured predictions with voting. Therefore, except for the label fusion step for whole image, the structured forest shares similar computational complexity as random forest. In our implements, we use classical random forest for the first three iterative refinements and only use the structured random forest for later iterations to reduce computational costs, which is shown in Fig. 2 and Fig. 3 with purple cubes.

III. Experimental Results

In this section, we first describe the datasets, the evaluation metrics, and experimental settings. Then, we compare our approaches with state-of-the-art methods.

A. Dataset

We evaluate mitochondria segmentations on two public benchmarks. The first one (EPFL Dataset) is the EPFL Hippocampus Data (https://cvlab.epfl.ch/data/em), which is the de facto standard benchmark for evaluation and has been used by many studied. The image stack, which is acquired using focused ion beam scanning electron microscope (FIB-SEM), represents a $5 \times 5 \times 5$ μm section taken from the CA1 hippocampus region of mouse brain, corresponding to a $2048 \times 1536 \times 1065$ volume. The voxel resolution is approximately $5 \times 5 \times 5$ nm. The mitochondria were annotated on two neighboring sub-stacks (each $1024 \times 768 \times 165$) of the first 165 slices of the $2048 \times 1536 \times 1065$ image stack. The two sub-volumes [3] are used for training and testing, separately.

The second dataset (Kasthuri++ Dataset) is the dataset released by Kasthuri *et al.* [33] with refined annotations (https://casser.io/connectomics/). The image stack, which is acquired using focused serial section electron microscope (ssEM), is taken form a 3-cylinder mouse cortex volume. The size of the training and testing stacks are $1463 \times 1613 \times 85$ (training set) and $1334 \times 1553 \times 75$ (testing set) with an anisotropic resolution of $3 \times 3 \times 30$ nm per voxel. The drawback of the annotations on this dataset is the inclusion of membrane as part of mitochondria.

B. Evaluation Metrics

We mainly focus on the segmentation accuracy of the foreground mitochondria. Given a ground truth (Y) and a segmentation (X), it is evaluated by comparing the segmentation to manual segmentation using Dice similarity coefficient (DSC)($2 \times |X \cap Y|/(|X| + |Y|)$), Jaccard index (Jaccard)($|X \cap Y|/(|X \cup Y|)$) and VOC score defined as the average of the Jaccard Index of the forground (mitochondria) and back-ground (non-mitochondria), i.e., $\text{VOC}(X, Y) = \frac{1}{2}\{\text{Jaccard}(X, Y) + \text{Jaccard}(\overline{X}, \overline{Y})\}$. These measures are in accordance with related works in this area, making our results more comparable with several existing works in this area. Note that the VOC score is not an accurate assessment for our task, as the segmentation accuracy of background is included.

C. Experimental Settings

All experiments have been carried out with the following settings. For computational efficiency, we resized all the original $1024 \times 768 \times 165$ volume images to $512 \times 384 \times 165$. The patch size for feature extraction was fixed to $11 \times 11 \times 11$, and the effective patch size for structured label prediction is $9 \times 9 \times 9$. For both random forest and structured forest, we used $K = 15$ trees both for our 3D segmentation method and hierarchical multi-view fusion method. The maximum depth for each tree is 50, and the minimum training set size for growing a leaf node is 15. For the iterative refinement, our 3D method conducted 7 iterations and our multi-view method 4 iterations. For both methods, we used random forest for the first 3 iterations and structured forest for remaining iterations.

D. Segmentation Results

EPFL Dataset. On the testing data, the accuracy of our 3D method reached 90.9%, 83.3%, and 90.9% in terms of DSC, Jaccard and VOC, respectively. The performance of our hierarchical multi-view fusion method was slightly lower than the 3D counterpart, and achieved 90.5%, 82.6% and 90.8% in terms of DSC, Jaccard and VOC, respectively. Since the hierarchical multi-view fusion method is building on parallel 2D segmentations, it requires lower memory for each forest, and is ready for parallel or distributed computation.

Figure 6 shows some representative results obtained by Lucchi-CVPR [22], U-Net [32], our method with hierarchical multi-view fusion, our 3D method, and the ground truth segmentation. Overall, we can clearly see that our segmentation results (Fig. 6(c) and (d)) are more consistent with ground truth (Fig. 6(e)). In particular, the method of Lucchi-CVPR shows sound detection accuracy but with obvious more false positive regions than U-Net and our methods. Some typical false positive detected regions segmented by Lucchi-CVPR and U-Net have been highlighted with yellow boxes, and false negative detected regions are highlighted with blue boxes. These visual observations also indicate the discriminative ability of the proposed contextual features.

Kasthuri++ Dataset. To validate the the applicability of our approaches on diverse data, we report the results on Kasthuri++ dataset in Table II. The results show that the proposed models also show promising performances on ssEM images similar to that on FIB-SEM images of EPFL Dataset. For this data with anisotropic resolution, our hierarchical multi-view fusion method shows slightly better performance (81.4% in Jaccard) than our 3D method (80.6% in Jaccard).

E. Comparison With State-of-the-Art Methods

In this experiment, we perform a comparison with state-of-the-art approaches for mitochondria segmentation including methods based on hand-crafted features and deep learning method. Below, we briefly outline these methods,

- **Seyedhosseini [34]**: A hierarchy of logistic disjunctive normal networks using features including Haar, HOG, SIFT, Gabor, as well as position and its higher orders.
- **Lucchi-TMI** [3]: 3D CRF using super-voxels and learned shape cues and boundary appearance. Ray descriptor and intensity histogram were extracted as features.
- **Lucchi-CVPR** [22]: 3D segmentation method using learned CRF. Ray descriptor and intensity histogram on super-voxels were used as features.
- **Lucchi-MICCAI** [23]: an improved method of [22] via explicitly exploiting enclosing membranes cues.
- **Márquez** [24]: 3D CRF using higher-order potentials to capture structural and geometric information.
- **Rigamonti** [30]: a recursion of a set of classifiers designed to progressively focus on difficult-to-classify locations.

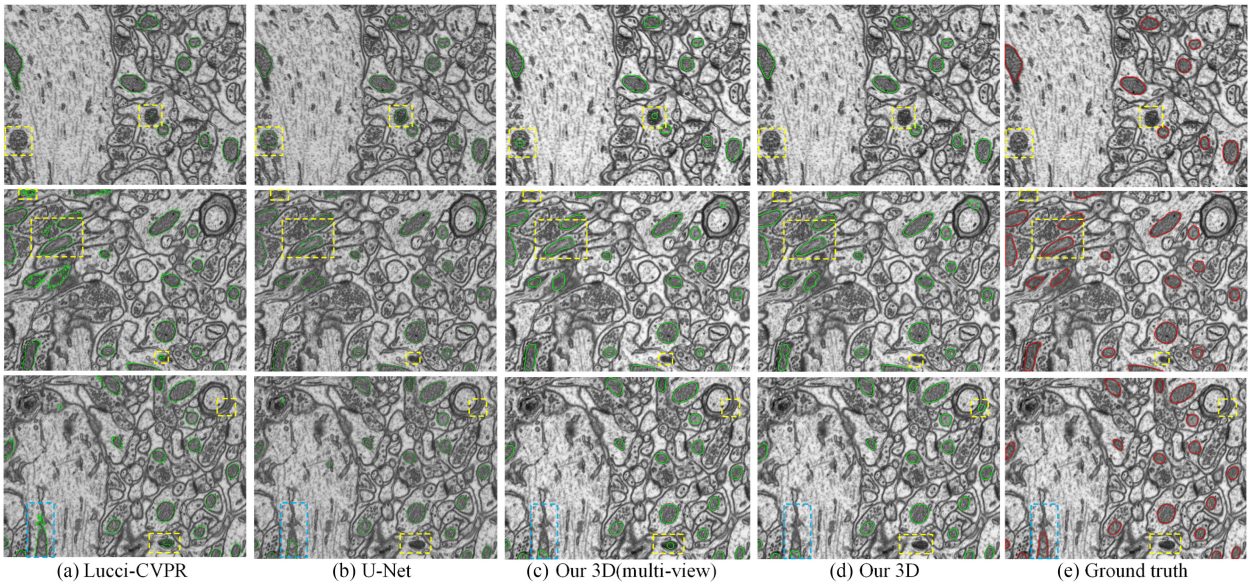


Fig. 6. Visual comparison of (a) Lucchi-CVPR [22], (b) U-Net [32], (c) our method using hierarchical multi-view fusion, (d) our 3D, and (e) ground truth.

An improved kernelboost with learned convolutional discriminative features was used as the classifier.

- **Cetina** [4]: Partially Informative Boosting was used. Multi-scale Gaussian and elliptic torus-like kernel filters with learned scales were used to extract features.
- **U-Net** [32]: a fully convolutional deep neural network with an encoder-decoder architecture and skip connections, which have shown strikingly superior results on many medical image segmentation tasks. Data augmentation was used to compensate the limited data.

We also compared several variants of our method,

- **Our 3D method**: it is our default method, which follows the pipeline shown in Fig. 2 and Fig. 3(a). It uses features extracted from local 3D patches.
- **Our 2D method**: it is the same as our 3D method but using features extracted from 2D slices on a single view, as shown in Fig. 3(b). Obviously, this strategy neglects inter-slice continuity.
- **Our method with simple multi-view fusion of final 2D segmentations**: it improves the 2D method by additional average of 2D segmentations from other two orthogonal views of a 3D volume.
- **Our method with hierarchical multi-view fusion**: it is an alternative of our 3D method using hierarchical multi-view fusion, as shown in Fig. 3(c) and (d). It uses features extracted from local 2D patches.

From the results in Table I, we can see that our 3D method and it 2D variant using hierarchical multi-view fusion are on par with the U-Net but without the need of complicated data augmentation and long-time training on GPUs. Besides, our 3D method yields significantly better results than all of the completive methods that are not based on deep learning. Even only using information from 2D slices, our 2D method also has shown comparable results with the method of Rigamonti [30] and Lucchi-CVPR [22]. Using the proposed hierarchical multi-view fusion, our 2D method obtains sound performance similar to the 3D method. Beside, the hierarchical multi-view fusion strategy obviously outperforms simple fusion strategy, i.e., only final average of multi-view 2D results. Table II shows evaluation results of our methods against the state-of-the-art methods [22], [34]. The results suggest that the proposed methods performs better on diverse dataset. Even our 2D variant shows competing perfromance.

Although our methods are based on hand-crafted features, it has shown significantly improved performance compared with classical methods based on manual crafted features, and also produced comparable performance compared with

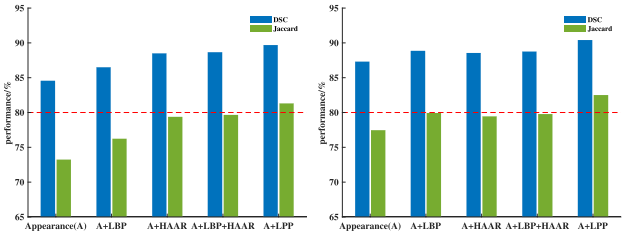
TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON EPFL DATASET

Methods	DSC(%)	Jaccard(%)	VOC(%)
Seyedhosseini [34]	83.5	71.7	84.8
Lucchi-TMI [3]	-	-	84.0
Lucchi-CVPR[a] [22]	86.0	75.5	86.8
Lucchi-MICCAI [23]	85.1	74.1	-
Márquez [24]	86.5	76.2	-
Rigamonti [30]	87.4	77.6	-
Cetina [4]	86.4	76.0	-
U-Net [32]	91.2	83.8	91.4
Ours (2D)	86.6	76.4	87.5
Ours (final multi-view fusion)	87.9	78.5	89.3
Ours (hierarchical multi-view fusion)	90.5	82.6	90.8
Ours (3D)	90.9	83.3	90.9

[a] The results were evaluated using segmentations provided by the authors.

TABLE II
PERFORMANCE COMPARISON ON KASTHURI++ DATASET

Methods	DSC(%)	Jaccard(%)	VOC(%)
Seyedhosseini [34]	83.9	72.3	85.6
Lucchi-CVPR [22]	86.2	75.8	87.5
Ours (2D)	87.1	77.2	88.2
Ours (hierarchical multi-view fusion)	89.7	81.4	90.4
Ours (3D)	89.3	80.6	90.0



(a) without structured prediction (b) with structured prediction

Fig. 7. The effectiveness of different features and their combinations. The segmentation performances are based on (a) our 3D method using standard random forest, and (b) our 3D model using structured random forest.

state-of-the-art deep learning methods. These comparisons can also indicate the discriminative ability of the proposed features.

IV. Discussion

Feature effectiveness study. To validate the effectiveness of the proposed features, we compare the performances using different features, i.e., appearance features (A) (defined in Section II-C), LBP features (LBP) [17], Haar features (HAAR) [16], and the proposed LPP features, and their combinations. We use the proposed 3D method, and its simplified version without structured prediction as the classifiers. Figure 7 reports the results under different settings. With different choices of classifiers, the proposed LPP feature consistently outperforms the counterpart features, including LBP, HAAR and also their combinations. LBP features have been extensively utilized to encode textures [17], thanks to their robustness against illumination changes. However, due to the strong background noises and complex structures in EM images, LBP features show limited performance with random forest (Fig. 7(a)), which is improved by using structured random forest (Fig. 7(b)). HAAR features showed robust segmentation results with respect to classifiers. As an extension of LBP and HAAR features, the proposed LPP features showed substantial performance gains over A, A+LBP, A+HAAR, and A+LBP+HAAR, which indicates the effectiveness of proposed LPP features.

Ablation study. To validate the benefits of using *structured prediction* and *incorporating longer history predictions with median fusion* in iterative refinement, we disable median fusion and structured prediction respectively, and evaluate the performances. Specifically, the ablation experiment is conducted under four different settings: a) our 3D model without using structured prediction and median fusion, b) our 3D model without using median fusion, c) our 3D model without using structured prediction, and d) our full 3D model. The results have been summarized in Table III. Specifically, without using median fused history predictions (Experiment b), the model shows a performance loss of 1.0% in Jaccard; without structured prediction (Experiment c), the model shows a performance loss of 1.2% in Jaccard; without using both of them (Experiment a), there is a performance loss of 2.0% in Jaccard. The ablation study verifies the effectiveness of structured prediction and median fusion of history predictions.

Sensitivity analysis of parameters. We report comprehensive experimental results about the robustness of the proposed method with respect to the choice of parameters.

TABLE III
Ablation Study of Our 3D Method

Experiments		a	b	c	d
Our 3D method					
	Structured prediction	✗	✓	✗	✓
	Median fusion	✗	✗	✓	✓
Measures					
	Jaccard (%)	81.3	82.5	82.3	83.3
	DSC (%)	89.7	90.4	90.3	90.9

a) Effect of the patch size for feature extraction. Figure 8 plots the scores of DSC and Jaccard on testing data as functions of patch size. We run our algorithm using values over 7, 9, 11, 13, 15. As we can see that the performance of our model increases as the increase of patch size, and reaches the best performance when using patch size 13. The default choice of patch size in our method is 11, which is a trade off of performance and computational efficiency.

b) Effect of the number of trees in random forest. Figure 8 also plots the scores of DSC and Jaccard as functions of number of tree. We run our algorithm over values over {5, 10, 15, 20, 30, 40}. From Fig. 8, we can see that our model is robust to the number of trees. Thus, we choose 15 as a balance of computational efficiency and accuracy.

c) Effect of the iterative refinement. For computational efficiency, we use random forest in the first three steps of refinements, and structured random forest in the remaining steps. As shown in Fig. 8, we obtained twice significant improvements of performance, that are the first refinement with random forest and the third refinement with structured random forest. After 5 times of refinements, the performance gains are marginal. In summary, both the iterative refinement and structured prediction can yield improved performances.

Weakness of the proposed method. Overall, the proposed 3D model observes a significant performance gain under different settings. One limitation is its high memory cost compared to the 2D counterpart. However, the proposed hierarchical multi-view variant is also building on 2D segmentation, but showing similar performance as the 3D method, reduced memory overhead and convenience for distributed learning. Another limitation of all our methods is that they are not end-to-end learning using image-level input as fully convolutional neural networks [32] but stage-wise learning methods using small patch input and hand-crafted features, which may limit the performance on complex tasks.

V. Conclusion

This study investigated automatic mitochondria segmentation from EM images. We developed two cascade approaches with novel architectures. While the first one performed direct 3D segmentation with superior performance, the second one was essentially a 2D model but with a novel hierarchical multi-view fusion strategy. Moreover, we introduced a new class of contextual features, which are computational efficient and robust. Structured random forest, retaining the low computation times of ensemble classifiers, was used to conduct joint label prediction. Comprehensive experiments have shown the state-of-the-art performance of our methods and the robustness of our model

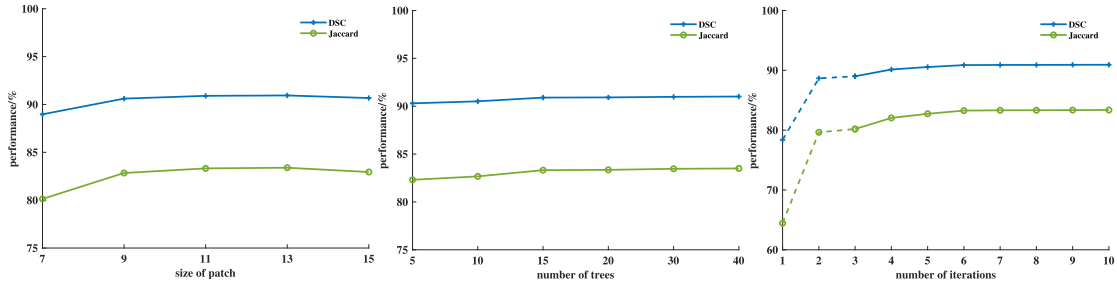


Fig. 8. The influences of patch size for feature extraction, number of trees and the step of iterative refinement on our 3D model.

with respect to parameters. In future work, we will investigate how to incorporate nonlocal similarity of patches over the whole EM images and structured sparsity [35] to enhance the label prediction.

REFERENCES

[1] J. Nunnari and A. Suomalainen, "Mitochondria: in sickness and in health," *Cell*, vol. 148, no. 6, pp. 1145–1159, 2012.

[2] A. J. Perez *et al.*, "A workflow for the automatic segmentation of organelles in electron microscopy image stacks," *Front. Neuroanatomy*, vol. 8, no. 126, p. 126, 2014.

[3] L. Aurlien, S. Kevin, A. Radhakrishna, K. Graham, and F. Pascal, "Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 474–486, Feb. 2012.

[4] K. Cetina, J. M. Buenaposada, and L. Baumela, "Multi-class segmentation of neuronal structures in electron microscopy images," *BMC Bioinform.*, vol. 19, no. 1, 2018, Art. no. 298.

[5] R. J. Giuly, "Method: Automatic segmentation of mitochondria utilizing patch classification, contour pair classification, and automatically seeded level sets," *BMC Bioinform.*, vol. 13, no. 1, 2012, Art. no. 29.

[6] R. Kumar, A. Reina, and H. Pfister, "Radon-like features and their application to connectomics," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.-Workshops*, 2010, pp. 186–193.

[7] P. Neila, L. Baumela, J. González-Soriano, J. Rodríguez, J. Defelipe, and Á. Merchán-Pérez, "A fast method for the segmentation of synaptic junctions and mitochondria in serial electron microscopic images of the brain," *Neuroinformatics*, vol. 14, no. 2, pp. 235–250, 2016.

[8] A. Lucchi *et al.*, "Learning structured models for segmentation of 2-D and 3-D imagery," *IEEE Trans. Med. Imag.*, vol. 34, no. 5, pp. 1096–1110, May 2015.

[9] K. Smith, A. Carleton, and V. Lepetit, "Fast ray features for learning irregular shapes," in *Proc. IEEE Int. Conf. Comput. Vision*, 2009, pp. 397–404.

[10] T. Zeng, B. Wu, and S. Ji, "DeepEM3D: Approaching human-level performance on 3D anisotropic EM image segmentation," *Bioinf.*, vol. 33, no. 16, pp. 2555–2562, 2017.

[11] K. Cetina, P. Márquez-Neila, and L. Baumela, "A comparative study of feature descriptors for mitochondria and synapse segmentation," in *Proc. 22nd Int. Conf. Pattern Recognit.*, 2014, pp. 3215–3220.

[12] C. Becker, K. Ali, G. Knott, and P. Fua, "Learning context cues for synapse segmentation in EM volumes," *IEEE Trans. Med. Imag.*, vol. 32, no. 10, pp. 1864–1877, Oct. 2013.

[13] A. Lucchi, Y. Li, K. Smith, and P. Fua, "Structured image segmentation using kernelized features," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 400–413.

[14] R. Narasimha, O. Hua, A. Gray, S. W. Mclaughlin, and S. Subramaniam, "Automatic joint classification and segmentation of whole cell 3D images," *Pattern Recognit.*, vol. 42, no. 6, pp. 1067–1079, 2009.

[15] V. Kaynig, T. J. Fuchs, and J. M. Buhmann, "Neuron geometry extraction by perceptual grouping in ssTEM images," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2010, pp. 2902–2909.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2001, pp. 511–518.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[18] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[19] R. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Mach. learn.*, vol. 37, no. 3, pp. 297–336, 1999.

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[21] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[22] A. Lucchi, Y. Li, and P. Fua, "Learning for structured prediction using approximate subgradient descent with working sets," in *Proc. IEEE Conf Comput. Vision Pattern Recognit.*, 2013, pp. 1987–1994.

[23] A. Lucchi, C. Becker, P. Mrquez Neila, and P. Fua, "Exploiting enclosing membranes and contextual cues for mitochondria segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2014, pp. 65–72.

[24] P. Márquez-Neila, P. Kohli, C. Rother, and L. Baumela, "Non-parametric higher-order random fields for image segmentation," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 269–284.

[25] M. G. Uzunbas, C. Chen, and D. Metaxas, "An efficient conditional random field approach for automatic and interactive neuron segmentation," *Med. Image Anal.*, vol. 27, pp. 31–44, 2016.

[26] M. Loog and B. Ginneken, "Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification," *IEEE Trans. Med. Imag.*, vol. 25, no. 5, pp. 602–611, May 2006.

[27] Z. Tu, "Auto-context and its application to high-level vision tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1744–1757, Oct. 2010.

[28] P. Kontschieder, S. R. Bul, H. Bischof, and M. Pelillo, "Structured class-labels in random forests for semantic image labelling," in *Proc. Int. Conf. Comput. Vision*, 2011, pp. 2190–2197.

[29] M. Seyedhosseini, M. Ellisman, and T. Tasdizen, "Segmentation of mitochondria in electron microscopy images using algebraic curves," *Int. Symp. Biomed. Imag.*, 2013, pp. 860–863.

[30] R. Rigamonti, V. Lepetit, and P. Fua, "Beyond kernelboost," 2014, *arXiv:1407.8518*.

[31] J. Funke *et al.*, "Large scale image segmentation with structured loss based deep learning for connectome reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1669–1680, Jul. 2019.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[33] N. Kasthuri *et al.*, "Saturated reconstruction of a volume of neocortex," *Cell*, vol. 162, no. 3, pp. 648–661, 2015.

[34] M. Seyedhosseini and T. Tasdizen, "Semantic image segmentation with contextual hierarchical models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 951–964, May 2015.

[35] J. Peng, X. Zhu, Y. Wang, L. An, and D. Shen, "Structured sparsity regularized multiple kernel learning for Alzheimers disease diagnosis," *Pattern Recognit.*, vol. 88, pp. 370–382, 2019.