Accurate 3D Bone Segmentation in Challenging CT Images: Bottom-up Parsing and Contextualized Optimization

Le Lu Dijia Wu Nathan Lay David Liu Isabella Nogues Ronald M. Summers
National Institutes of Health le.lu@nih.gov Google Inc. Siemens Healthcare

Abstract

In full or arbitrary field-of-view (FOV) 3D CT imaging, obtaining an accurate per-voxel segmentation for complete large and small bones remains an unsolved and challenging problem. The difficulty lies in the notable variation in appearance and position observed among cortical bones, marrow and pathologies. To approach this problem, several studies have employed active shape models and atlas models. In this paper, we argue that a bottom-up approach, defined by classifying and grouping supervoxels, is another viable technique. Moreover, it can be integrated into a conditional random field (CRF) representation. Our approach consists of the following steps: first, an input CT volume is decomposed into supervoxels, in order to ensure very high bone boundary recall. Supervoxels are generated via a robust process of conservative region partitioning and recursive region merging. In order to maximize sparsity and classification efficiency, we use a Bayesian sparse linear classifier to compute and optimize middle-level image features. Next, we disambiguate the CRF unary potentials via contextualized optimization by pooling over selective supervoxel pairs. Finally, we adopt a pairwise support vector machine (SVM) model to learn the CRF pairwise potential in a fully supervised manner. We evaluate our method quantitatively on 137 low-resolution, low-contrast CT volumes with severe imaging noise, among which various bone pathologies are represented. Our system proves to be efficient; it achieves a clinically significant segmentation accuracy level (Dice Coefficient 98.2%*).*

1 Introduction

The human skeletal system comprises a multitude of primary bones (e.g. skull, vertebrae, ribs, pelvis), secondary bones (e.g. upper and lower limbs, hands) and joints to support the body's soft tissue. 3D bone segmentation in CT images is a prerequisite for masking bone tissue voxels and later performing quantitative pathological diagnoses. Examples of important imaging-based diagnostic assessments include bone mineral density assessment [40], bone lesion and fracture detection [37], and the analysis of abnormal shapes caused by ankylosing spondylitis [35, 42]. However, it is very difficult to obtain an accurate mask for all bone voxels, from full or arbitrary field-of-view (FOV) 3D CT images. Different bone structures may belong to a wide range of body regions, and their complex, inhomogeneous appearance needs to be precisely modeled. In CT voxel intensity, the bone marrow and other cancerous, spongy bone structures overlap significantly with other non-bone soft tissues and organs. Hence, simple intensity thresholding is not sufficient. Furthermore, high intensity "non-bone outliers", such as metal implants, partially tagged stools and IV-contrasted vessels, can be confused with cortical bones. Several examples are illustrated in Fig. 1. In practice, one is not previously aware when such ambiguities occur in a full-body CT scan. Most previous works have focused on a specific bone organ, such as the spine [30, 39, 31] or the hip [31]). To the best of our knowledge, our work is the first to address 3D bone segmentation in full-body or arbitrary FOV CT scans.

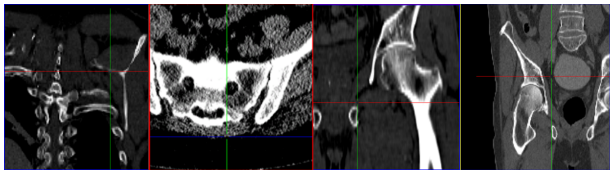


Figure 1. *Detailed bone appearance patterns in CT images. Bone structures are widely distributed in the full body region, unlike spatially-concentrated organs, such as the heart and the liver. They also show high inhomogeneity on intensity patterns from cortex to marrow and very different 3D shapes for various bones.*

Recent studies [38, 22] show that the current PET-CT imaging protocol of applying CT intensity values for attenuation correction may degrade the quality of obtained PET images, due to bone inhomogeneities. As a result, the standardized uptake values (SUVs) of spine/bone lesions [22] are considerably underestimated. To compute more accurate SUVs, one may replace the CT numbers of segmented bone voxels with a single patient-specific (e.g., mean intensity) or generic value. This new protocol has the potential to effectively correct the previous bias, which leads to the misinterpretation of clinical images. In fact, enabling semantic

bone-aware attenuation correction (AC) for PET-CT imaging is the second primary goal of this work. Nevertheless, PET-CT scans face many more difficulties than do normal CT scans. CT volumes from PET-CT are of non-diagnostic quality and are generated with lower radiation doses. They present weaker contrast, lower resolution and heavier imaging noise levels. Furthermore, the patient population has much higher rates of various pathologies (e.g., lesion, tumor, scoliosis, osteoporosis), because PET-CT is used very late in the clinical imaging diagnosis timeline.

Among the numerous studies on generic organ segmentation, most recent works employ **top-down** methods: statistical/active shape model (ASM) [52, 27] for heart/liver segmentation, and atlas registration label fusion framework [21, 50] for brain and cardiac segmentation. However, these methods are not suitable for full-body 3D CT skeleton segmentation problem for the following reasons: **1)**, The human skeleton system contains up to 206 bones. It is computationally expensive, if not utterly infeasible, to build a *statistical shape model* [52] for each bone, assuming a sufficient number of individually segmented and aligned bone meshes are given. **2)**, The methods used in [52, 21, 50] require the computation of either *3D initialization poses* to align ASM models or *3D dense non-rigid registration fields* to warp atlases to CT images. The bone skeleton is highly articulated, and upper limb poses are often freely distributed during scanning. Some bones have extremely complex, thin, or singular 3D shape parts, which makes them hard to detect and localize. No prior work has applied a non-rigid volume registration scheme to full-body or arbitrary FOV bone skeletons, which have a high degree of anatomical variability. Furthermore, registration based organ localization has a gross failure rate of $\sim 30\%$ [13]. **3)**, ASM needs *precise organ boundary detection* for the 3D surface model to fit and converge [52, 27]. Bones often spatially correlate to each other in the skeleton. In other cases, multiple adjacent bone surfaces may interact (e.g., successive vertebrae). Hence, it is difficult to perform high quality bone boundary delineation, even without considering bone fractures. **4)**, Both ASM and atlas-based **top-down** methods [52, 21, 50] rely on the validity of statistical models learned from a population of normal patients. It is unclear how to generalize and handle pathological cases (studied here) under such frameworks. **5)**, Per-voxel classification approaches [16, 5] are too slow to handle large volumes. They often generate very noisy class-conditional response maps, requiring non-trivial spatial regularization/smoothing.

We argue that **bottom-up** representations, such as those presented in [14, 10, 6, 44], should be employed, instead of **top-down** approaches [52, 21, 50]. The region (i.e., superpixel) proposal based techniques have attained the highest ranks in both the PASCAL semantic segmentation [10, 6] and detection [44, 17] challenges, even in the era of deep learning [17, 19]. Defining an object as a flexible, spatial composite of middle-level image regions [14] (superpixels in 2D, supervoxels in 3D) provides a good balance. It allows for effectively capturing very large variations in holistic object-level models [52, 21, 50] and better disambiguates noisy, local image appearance than per-voxel detectors [5].

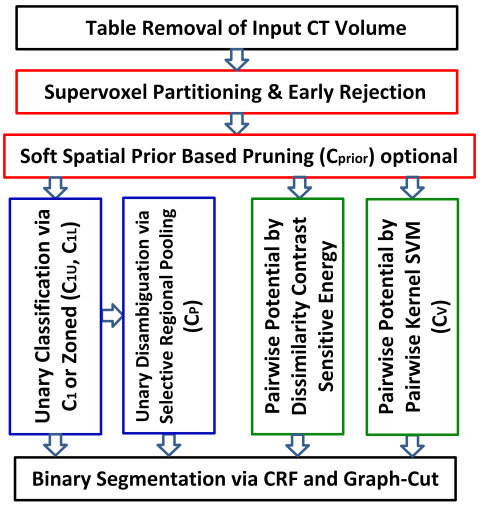


Figure 2. *The system flowchart of 3D bone segmentation in CT.*

We provide three major contributions: **1)**, We present a bottom-up bone segmentation approach with middle-level image representation (i.e.,supervoxels) and sparse regional image features. **2)**, We propose and exploit three types of contextualized optimization: a novel soft-weighted bone spatial prior pruning, selective regional pooling (SRP) to disambiguate unary CRF energies, and supervised pairwise SVM based second-order potential learning (adapted from biological network inference). All components can be integrated into a CRF representation, in which the supervoxels represent the graph nodes. **3)**, We evaluate our system quantitatively on 137 challenging CT volumes from PET-CT imaging collected from two hospitals. This is the largest existing dataset for bone segmentation.

We report accurate 3D bone segmentation results (Dice score 98.2%, Precision 97.9% and Recall 97.9%) that compare favorably to previous results [39, 31] and to a commercial bone removal software. The **Full** 3D bone segmentation system flowchart is provided in Fig. 2.

2 High-recall Region Proposals

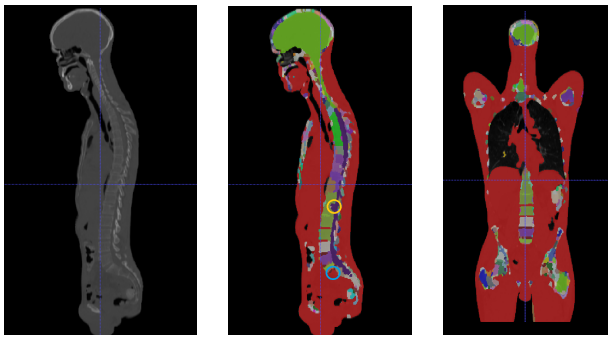
Several previous studies describe superpixel generation using graph partitioning [15, 3, 12] or gradient-ascent-based techniques [48]. The methods presented in [15, 48] are relatively fast. They have $O(NlogN)$ complexity, where N is the number of pixels or voxels in the image. The computational complexities of other methods are too heavy (e.g., mean-shift [11]: $O(N^2)$, normalized cut [12]: $O(N^{\frac{3}{2}})$) for 3D medical volumes ($N = 50 \sim 120$ million voxels). Particularly, the excellent yet expensive contour and region detector of [3] (used in the winning PASCAL method [10])

is not a viable tool. In our study, we adopt the 3D watershed algorithm [48], which preserves the semantic bone surface boundary with good sensitivity (by the enhancement of p-percentile gradient filtering) and generates significantly larger, but fewer, supervoxels (SVs) in the non-bone regions (i.e., background) shown in Fig. 3 (b,c). This partitioning scheme is due to the lack of complete strong gradient surfaces separating soft-tissue and fat into small zones. Consequently, these large background supervoxels can be effectively rejected (i.e. not classified as bone voxels) by simple volume and mean intensity based thresholding. This greatly enhances the overall system efficiency. Simple linear iterative clustering (SLIC) [1] is also fast and memory efficient. However, in our case, it generates too many SVs ($\sim 10^6$ versus 10^3 [48]), which increase the computational cost of the later optimization stages. In this paper, we use the publicly available ITK implementation of MorphologicalWatershedImageFilter( ) [23, 48] with the following modifications and enhancements.

First, CT volumes are smoothed with Gaussian filters. Next, we perform a simple ρ-percentile gradient filtering on the gradient domain where the 3D watershed algorithm [48] is implemented, in order to fill potential bone surface holes. For each voxel in the gradient image, the filter value is set to $G = \max(G_0, G_\rho)$. G_0 is the original gradient magnitude, and G_ρ is the ρ-percentile gradient within a local window of $(5 \times 5 \times 5)$ voxels ($\rho = 85\%$) centered at G. This filter only runs once, in order to restrict iterative propagation. To achieve a high recall value on splitting precisely at bone boundary surfaces, it is critical to set a low pre-flooding parameter as $itk :: MorphologicalWatershedImageFilter :: SetLevel(6)$. This allows us to achieve a mean voxel-level bone boundary recall value as high as $\Re = 98.2\%$ without ρ-percentile filtering. These parameters are calibrated in training. We use the same parameter setting in all of our experiments. In Fig. 3 (b), the breakage of bone boundaries indicated by a yellow circle (bone osteoporosis) can be corrected by this filtering scheme. For quantitative assessment, ρ-percentile filtering improves $\Re$ to 99.5%.

Given the annotated CT volumes (bone voxels labeled as 1, others as 0), the largest supervoxel bone volume V_b^{max} in mm^3 can be counted or calibrated from 57 training datasets. After this, any supervoxels with a volume $\geq \kappa \times Vol_b^{max}$ are eliminated (κ is a small constant set to 3). Supervoxels with mean intensity ≤ 450 (with respect to soft-tissue CT intensity values ≥ 700) are also removed. The main objective of this pruning process is to discard all supervoxels associated with lung tissue. With this process, $> 85\%$ of absolute volumes inside the scanned 3D CT body region can be safely excluded from further processing.

Recursive Region/Supervoxel Merging: After partitioning and thresholding the CT volumes, we build a region adjacency graph (RAG) and nearest neighbor graph (NNG) from the remaining supervoxel label maps. RAG contains the list of all supervoxels sharing spatial boundaries. NNG preserves its nearest neighbor selected from RAG. Information-theoretic measures [9], namely the area-weighted *Bhattacharyya* coefficients, are used to compute the log-negative dissimilarity score between any pair of supervoxels $\{V_i, V_j\}$.



(a) (b) (c)

Figure 3. *Examples of supervoxel generation results using ITK watershed and recursive region merging, without the gradient enhancement: (a) the original data volume in sagittal view; (b) a supervoxel label map overlaid on the data volume in the same sagittal view. Yellow and blue circles indicate the breakage of bone boundaries (thus partial marrow regions are connected with background in red; (c) supervoxel label map in coronal view.*

$$S(V_i, V_j) = -\min(N_i, N_j) log(\sum_k \sqrt{H_i^k H_j^k}) \quad (1)$$

, where N_i is the volume normalized voxel number ($8mm^3$/voxel), H_i^k is the intensity histogram (binned at CT numbers of $[0, 450, 750, 1250, 1600, 4095]$) according to approximate tissue types of air, different levels of soft tissue, bone, metal impants) for supervoxel V_i, and $k \in [1, 5]$ is the index of the histogram. In each iteration, we choose the nearest-neighbor pair $\{V_i, V_j\}$ with the minimal dissimilarity cost to merge (sorted by NNG) and efficiently update RAG and NNG only for the first-hop neighbors of $\{V_i, V_j\}$. The stopping criteria are the number of remaining supervoxels ≤ 2000 and the minimum dissimilarity > 1.35, as estimated by cross validation. Instead of generating the whole region hierarchy via merging [44], we aim to reduce the number of supervoxels (after merging) for efficiency purposes and to eliminate very small or similar segments. The resulting supervoxels are more likely to contain sufficient spatial supports and to facilitate the computation of statistically stronger middle-level image features.

3 Bottom-up Parsing & Contextualized Optimization

We first formulate the robust unary potentials to achieve high accuracy: efficient computation of regional image features, sparse linear feature classification and a novel se-

lective regional pooling (SRP) scheme for spatial context based disambiguation. Then the supervised pairwise SVM is exploited for the second-order CRF potential learning.

3.1 Regional Image Features

Until the past two years, using hard and soft quantized dense image descriptors such as SIFT, HOG, pyramid HOG, and SSIM (self-similarity) and spatially aggregating them into histograms for superpixels [6, 10] was the main stream method in the VOC challenge [14]. The 3D extensions of these descriptors are not trivial for dense volume patch labeling. Their discriminative power is unclear for modeling densely grid-sampled 3D patches in CT images, due to the wide variety of local image statistics. For example, volumetric image patches in bone marrow or soft-tissue classes are nearly indistinguishable. Due to their computational complexities, the dense descriptor computing and encoding schemes may not be able to process a full-body CT volume under a $2 \sim 3$ minute time limit. Deep Convolutional Neural Network (CNN) [26] based per-voxel (3D image patch) bone classification would be too slow, as segmentation incurs a high computational cost [36]. Applying a 3D regional CNN [17, 19] to our dataset is infeasible, due to a lack of training data and of a good quality pre-trained model for adaption and fine-tuning. For problems differing from ImageNet classification (with a much smaller variance and scale but a high accuracy requirement, such as face or pedestrian detection), deep learning does not necessarily yield the best results, as observed in [32]. We will leave the extension of classifying 3D subvolumes via deep CNN [36] for future work.

Recently, [4] has found that efficient image region features are nearly as effective as densely encoded SIFT and HOG descriptors for object recognition. Image region features [2, 29] describe and summarize the voxel intensity, gradient distributions and shape and size statistics of supervoxels, and are scale-invariant.

Treating all voxels in a supervoxel V_i as an order-less distribution, we compute the following features as empirical statistical measurements per supervoxel.

Intensity Features: mean, median, 25%, 75%-percentile and the intensity histogram $H_i^{k=0\sim4}$.

Size/Shape Features: size in metric volume (mm^3), three eigenvalues of the spatial occupancy mass of V_i defined as R_1, R_2, R_3, their ratios $R_2/R_1, R_3/R_1$ and "plateness", "stickness" and "ballness" features [18].

Boundary Features: 1st through 4th order statistical moments of three empirical feature distributions: normalized distance $\{D_N(v_b)\}$, normalized gradient $\{G_N(v_b)\}$ and orientation $\{O(v_b)\}$ for all boundary voxels $\{v_b\}$ in V_i.

H is binned at CT numbers $[0, 450, 750, 1250, 1600, 4095])$, which roughly correspond to the tissue types of air, mixed, lower-, and upper-range of soft-tissue and bone. Tissue type binning can produce compact and representative histograms to capture the supervoxels' tissue composition ratios. Likewise, [29] finds that a simple supervoxel intensity histogram outperforms several other local texture and intensity descriptors, e.g., LBP and DAISY [43]. The boundary features are computed:

$$D_N(v_b) = \|[x_b, y_b, z_b] - [\bar{x}, \bar{y}, \bar{z}]\| \,/\boldsymbol{D} \tag{2}$$

$$G_N(v_b) = \|\nabla(x_b, y_b, z_b)\| \,/\, \boldsymbol{G} \tag{3}$$

$$O(v_b) = \frac{\nabla(x_b, y_b, z_b)}{\|\nabla(x_b, y_b, z_b)\|} \circ \frac{[x_b, y_b, z_b] - [\bar{x}, \bar{y}, \bar{z}]}{\|[x_b, y_b, z_b] - [\bar{x}, \bar{y}, \bar{z}]\|}, \tag{4}$$

where $[x_b, y_b, z_b]$ records the volumetric coordinates of boundary voxel $v_b \in V_i$, and $[\bar{x}, \bar{y}, \bar{z}]$ is the centroid location of V_i. $\boldsymbol{D}$ is the maximum of $\|[x_b, y_b, z_b] - [\bar{x}, \bar{y}, \bar{z}]\|$ for all $\{v_b\}$ in V_i, and $\boldsymbol{G}$ is the maximum $\|\nabla(x_b, y_b, z_b)\|$ from training. Therefore, $D_N(v_b)$ and $G_N(v_b)$ are normalized to values in $[0, 1]$ and are scale- invariant. They measure implicit shape regularity and boundary gradient contrast respectively. $O(v_b)$ is the dot-product ($\in [-1, +1]$) of the centroid-to-boundary direction and its local gradient direction, which encodes boundary shape information. Our boundary features are related to the 2D and 3D ray features described in [29], which describe irregular shapes but are more compact and are fully rotation invariant for robustness (without concatenating dozens of sampling directions as in [29]). In total, we compute 30 features $X_i \in \boldsymbol{R}^{30}$ per V_i in $O(N)$ running time, where N is the number of voxels.

3.2 Sparse Linear Unary Classification

The nonlinear support vector machine has performed better than linear kernels in PASCAL VOC [6, 45], except when deep neural features [17] or sophisticated high-dimensional feature encoding and pooling [41, 20] were incorporated in the classification scheme. We present a linear yet high performing probabilistic classifier. To find the optimal linearly weighted decision hyperplane w^T that separates $\{X_i\}$ according to the labels $\{Y_i = 0, 1\}$, the objective is:

$$Y_i = 1, if \;\; w^T X_i > b; \quad Y_i = 0, otherwise, \tag{5}$$

where b is the bias to be estimated. The decision margin can be converted to the pseudo-probability of class $(Y_i = 1)$ through the logistic sigmoid function as

$$p(X_i|w) = \phi(w^T X_i), \tag{6}$$

where $\phi(z) = 1/(1 + e^{-z})$. The overall data log-likelihood for all training samples is

$$logp(\mathbf{X}|w) = \sum_{X_i} Y_i p(X_i|w) + (1 - Y_i)(1 - p(X_i|w)). \tag{7}$$

A zero mean Gaussian prior $G(w|0, \Sigma)$ is further assumed for the parameter w, with diagonal covariance matrix $\Sigma = diag(1/\alpha_1, 1/\alpha_2, ..., 1/\alpha_d)$ and $d = 30$. $\alpha_1, \alpha_2, ..., \alpha_d$ are hyper-parameters that are later used for feature selection as

$1/\alpha_j \rightarrow 0$. By Bayes' rule, we can obtain the *maximum a-posterior* estimate of w as

$$\hat{w}_{MAP} = arg\max_{w}[logp(\mathbf{X}|w) + logp(w)], \quad (8)$$

or

$$\hat{w}_{MAP} = arg\max_{w}\left[logp(\mathbf{X}|w) - \frac{w^T\Sigma^{-1}w}{2}\right]. \quad (9)$$

Eq. 9 can be solved via an iterative gradient based optimization method, such as the Newton-Raphson update, $w^{t+1} = w^t - \lambda\mathbf{H}^{-1}g$.$\mathbf{H}$ is the Hessian matrix, g is the gradient vector and λ is the step length.

The hyper-parameters in Σ can also be optimized by maximizing the marginal likelihood $logp(\mathbf{X}|\Sigma)$ approximated by the Taylor series expansion (known as the type-II maximum likelihood method in Bayesian statistics [34]). A closed-form solution can be obtained by equating the first derivative to zero and simplifying the resulting expression.

$$\alpha_j = 1/(\hat{w}_j^2 + h_{jj}); j = 1, 2, ..., 30, \quad (10)$$

where $\hat{w}$ is the optimal, converged $\hat{w}_{MAP}$ and h_{jj} is the jth diagonal element in $\mathbf{H}^{-1}(\hat{w}_{MAP}, \Sigma)$, given the previous hyper-parameters (all α_j are initialized at 1). The overall optimization scheme has two iteration layers: an inner loop for optimizing $\hat{w}_{MAP}$ and an outer loop for updating $\{\alpha_j\}$. For all $\alpha_j > 10^{10}$, we set $\hat{w}_j = 0$ and remove the jth feature from the vector X_i, in order to achieve sparseness. This thresholding scheme is numerically stable for feature sparsity optimization. For further details, we refer the reader to [34]. Note that a Laplace prior [49] may be used to replace the Gaussian prior (no thresholding is needed) for sparsity.

All X features are first normalized with $G(w|0, 1)$. Hence, the magnitudes of the standardized coefficients $\{\hat{w}_j\}$ in the final classifier represent their weight importance. Directly interpretable feature importance is desirable for medical diagnosis. (*Our classification scheme surpasses others such as SVM or random forest in this regard*). In Fig. 4 (b), four features are eliminated, and four have negligible weights. Feature priors Σ and weights $\hat{w}$ are also captured. Using all three types of features yields the best validation ROC curve (shown in Fig. 4 (a)), as opposed to using individual feature groups. The relative group importance ranks are *Intensity*, *Boundary* and *Shape* features, in descending order. Classifier training and testing are both very efficient. Training on 78997 samples takes less than 1 minute. Once the regional image features X_i are computed in $O(N_i)$ time, the classifier evaluation (Eq. 6) requires minimal time due to its linear sparsity. We denote the learned classifier as $\mathbf{C}_1$.

3.3 Disambiguating Unaries by SRP

The classifier $\mathbf{C}_1$ attains high accuracy levels. Its ROC curve on validation in Fig. 4 (a) gives an AUC (Area-under-Curve) ($= 0.9765$) , a92.6% recall, and a $\sim 10\%$ false positive (FP) rate. Nevertheless, the segmentation accuracy is not yet sufficient for our tasks. Certain portions of bone supervoxels (e.g., bone marrow: Y=1) cannot be easily distinguished from some background supervoxels (e.g., soft-tissue: Y=0), if only visual cues are used. Their appearances, and thus their derived descriptive features, can be very ambiguous. We propose a spatial *selective regional pooling* (SRP) algorithm to disambiguate such regions. Our main idea is that bone marrow supervoxels are more likely to have bone SV neighbors with high $p(X|\mathbf{C}_1)$ scores represented in the RAG graph than soft-tissue supervoxels.

All supervoxels with $p(X_i|\mathbf{C}_1) \in [0.4, 0.6]$ are considered to have "high classification uncertainty". To disambiguate them, we feed them into $\{X'_i\}$ to train a second classifier $\mathbf{C}_p$. The positive (+) training set consists of supervoxel pairs: a bone supervoxel $\in \{X'_i\}$ and an adjacent bone supervoxel (available in its RAG graph built in Sec. 2). The negative (-) set contains the following pairs: **1)**, a bone supervoxel $\in \{X'_i\}$ with any of its non-bone neighbors and **2)**, a non-bone supervoxel $\in \{X'_i\}$ with any of its adjacent non-bone neighbors. The regional features of paired supervoxels are concatenated in an ordered fashion to form a pairwise feature vector $\bar{\chi}$. The feature X with the highest $p(X|\mathbf{C}_1)$ score is placed first in $\bar{\chi}$. Their $p(\cdot|\mathbf{C}_1)$ probability values are also added into $\bar{\chi}$, thus $\bar{\chi} \in \boldsymbol{R}^{62}$ to train $\mathbf{C}_p$. $p(\cdot|\mathbf{C}_1)$ plays an important role since high-confident bone supervoxels in the pairs are already indicated in $\bar{\chi}$. After $\mathbf{C}_p$ classification, each X'_i receives multiple scores depending on the dimension of its RAG list. The final bone class probability is computed:

$$p(X'_i|\mathbf{C}_p) = \sum_{X_j \in RAG(X'_i)} \gamma_{(X'_i, X_j)} p(\bar{\chi}_{(X'_i, X_j)}|\mathbf{C}_p), \quad (11)$$

where $\bar{\chi}_{(X'_i, X_j)}$ is the feature vector formed by the pair of (X'_i, X_j), and $\gamma_{(X'_i, X_j)}$ is the ratio of the shared boundary surface of X'_i and X_j divided by all boundary surface area of X'_i . The probability $p(X'_i|\mathbf{C}_p)$ is a weighted average of local spatial pooling based on RAG groups, related to but significantly different from the pooling strategies in [7, 10]. Supervoxels $\notin \{X'_i\}$ keep their original probabilities $p(X_i|\mathbf{C}_1)$. Overall, this forms a shallow cascade decision tree for learning unary probabilities (Fig. 2). Finally, their negative log-probability transforms are used as CRF unary energy terms.

3.4 Pairwise SVM 2nd-Order Potential Learning

Our 3D bone segmentation problem is treated as a binary (foreground-background) CRF optimization problem over a graph of supervoxel nodes. Supervoxels are classified by $\mathbf{C}_1$ and $\mathbf{C}_p$ as unary CRF terms and regularized by pairwise constraints. The typical pairwise potential functions are the Potts model or an intensity contrast-sensitive term [8]. To achieve semantically high bone segmentation accuracy, we adapted a supervised second-order potential function learning method using a pairwise support vector machine (SVM) from a scheme used to predict inter-protein

connections in biological network inference [46, 47]. We employ it to identify pairs of supervoxels that share bone and non-bone boundaries, and thus incur a low cost to be assigned different labels in CRF optimization.

Similar pairs of nodes in the feature space are always allocated with high costs to be classified under different labels [8].The pairwise potential learning is more critical for **dissimilar pairs**: *supervoxels are separated at a low cost, if they are separated by an object-level boundary, and at a high cost otherwise (e.g., a pair of cortical and marrow bone supervoxels)*. Dissimilar supervoxel pairs are defined as (V_i, V_j) in RAG graphs with $S(V_i, V_j) > \tau$. The constant τ equals $\mu(\{S\}) - 2\sigma(\{S\})$, where $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation of the dissimilarity score (Eq. 1) computed from all adjacent bone and non-bone supervoxel pairs. Empirically, $-2\sigma(\{S\})$ guarantees that most bone and non-bone pairs can be possibly learned. We label the **positive** supervoxel pairs as {**bone and non-bone**} pairs and the others (especially cortical-marrow bone pairs) as negative. Most pairs satisfy $S(V_i, V_j) \leq \tau$ and follow a *contrast-sensitive energy term* [8], using χ^2 intensity histogram distances. Overall, the CRF pairwise potentials are formulated as a switchable model (see Fig. 2).

The *metric learning pairwise kernel* (MLPK) for SVM is employed, as presented in [46]:

$$K((X_1, X_2), (X_3, X_4)) = (K(X_1, X_3) - K(X_1, X_4) - K(X_2, X_3) + K(X_2, X_4))^2 \quad (12)$$

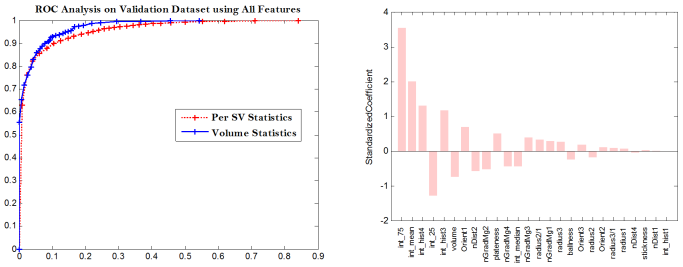
or, given $\Phi(X_1)$ as Hilbert feature mapping,

$$K((X_1, X_2), (X_3, X_4)) = ((\Phi(X_1) - \Phi(X_2))^T (\Phi(X_3) - \Phi(X_4)))^2. \quad (13)$$

The pairwise potential function should be symmetric for undirectional image grids/graphs satisfied by the MLPK kernel. Just as in Sec. 3.3, X_i is a vector in $\boldsymbol{R}^{31}$, with $p(\cdot|\mathbf{C}_1)$ added. Both linear and Gaussian kernels are evaluated to train the classifier $\mathbf{C}_v$. Finally, for each pair satisfying $S(V_i, V_j) > \tau$, the *supervised pairwise energy term* is $-log(p((X_i', X_j')|\mathbf{C}_v))$. The SVM confidence of the classifier $\mathbf{C}_v$ is converted into a pseudo-probability by a *Sigmoid* function. If trained properly, particular cortical-marrow bone supervoxel pairs will obtain low probability values from $\mathbf{C}_v$ (as a negative class) but high energy penalties to be split in the CRF. CRFs with higher order cliques, hierarchical CRFs [6, 25] or holistic context [51, 33] will be investigated in future work.

4 Experiments & Discussion

Data: 137 CT volumes in PET-CT are collected from two hospitals in Europe and the US with low voltage and dose protocols. They are randomly separated into sets of 57 and 80 scans for training and validation respectively. The in-slice resolution range is $1.2 - 2.0$ mm, and the inter-slice resolution is $R_z \in [1.5, 5mm]$. The majority of images are full-body scans, and the remaining 26 are skull or thoracic scans. 8 pediatric scans are present, and the remainder are adult scans. All volumes are resampled as $[2mm, 2mm, max(2mm, R_z)]$ in $[x, y, z]$ coordinates.



(a) (b)

Figure 4. *Plots of linear classification on supervoxels: (a) validation ROC curves of per-V_i or volume mm^3 statistics; (b) standardized coefficients of selected features.*

Implementation Details: Our **full** system runs under $2 \sim 3$ minutes per 3D CT volume. Based on the annotated ground-truth, the bone volume statistics $(mean, std)$ are 3.58 ± 0.86 liters. After the volume thresholding in Sec. 2, there are on average 8.17 ± 2.28 liters of CT volume remaining. Our quantitative evaluation only counts **these remaining** $193,618$ **supervoxels** in 137 CT scans. Using the middle-level feature optimization in Sec. 3, the AUC of $\mathbf{C}_1$ ROC in validation is 0.9765 (i.e., 92.6% recall at 10% FP rate or the precision of 0.8789). By comparing (a,c) in Fig. 4, one can see the advantage of fusing multiple channels of features over using only the boundary features. We compare this result (obtained by $\mathbf{C}_1$) to the results obtained by using SVM with linear kernels and RBF kernels and random forest (TreeBagger() in Matlab). On the validation datasets, $\mathbf{C}_1$ slightly outperforms other classifiers ($3 \sim 7\%$ higher recall at at 10% FP rate) with higher computational efficiency. We use $\mathbf{C}_1$ in an under-sampled learning scenario and note that our classifier improves generalization for the segmentation problem. We also propose learning a probabilistic spatial prior model via a sparse linear classifier ($\mathbf{C}_{prior}$ to remove non-bone voxels prior to computing and optimizing CRF appearance features. Due to space restrictions, we have placed the description of this method in the supplementary material section.

Zoned Piecewise Linear Classification: We test the zoned classification scheme by dividing the CT body range into three regions: **R1**: head, neck, and arms above the cervical vertebra C-6; **R2**: torso above the sacrum; **R3**: pelvis and legs below the sacrum, or two zones: R1 and R2+R3. We train a separate linear classifier in each zone. In the divide-and-conquer sense, zoned processing provides an overall piecewise linear classification framework. Empirically, the two-zoned approach yields a slightly better ROC curve than the three-zoned or one-zoned settings. The validation AUC increases to 0.9846 when $\mathbf{C}_1$ is replaced by $\mathbf{C}_{1U}$ and $\mathbf{C}_{1L}$. These results occur, as R1 has the dens-

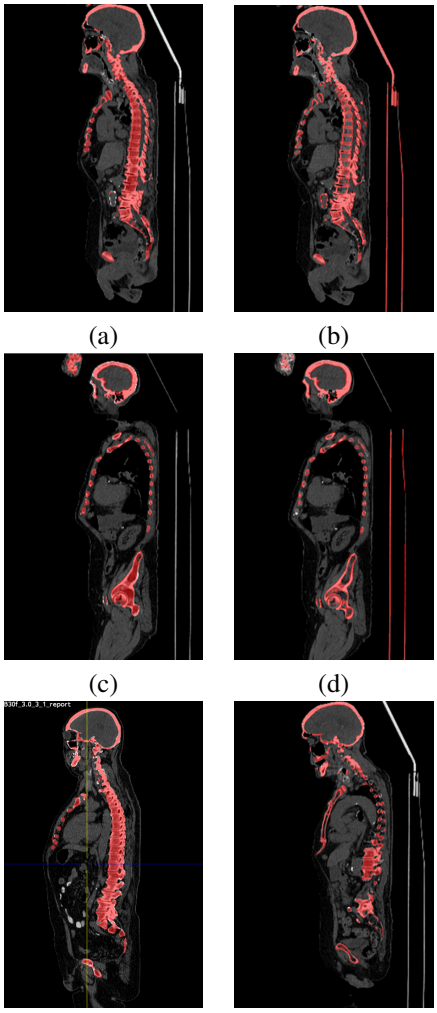


Figure 5. *Bone segmentation examples in sagittal view: (a,c,e,f) illustrate results obtained by our method in four CT scans, and (b,d) show results generated by a bone-removal commercial software. The images in b) and d) are identical to those in a) and c) respectively. The bone segmentation masks in (c,d) include the patient's hand wrist (very flexible, fine bone structure). Several bone pathologies are observable, such as degenerative disc disease, osteoarthritis, spinal stenosis (a,b,e), and scoliosis (f).*

est (and thus brightest in CT) bones in the skull and upper vertebral bodies. This causes $\mathbf{C}_{1U}$ and $\mathbf{C}_{1L}$ to weight X features differently. Sparse spine landmark detection and labeling are generally available but are optional. Without spine labeling, our workflow runs with only the unary and pairwise classifiers of [$\mathbf{C}_1$, $\mathbf{C}_p$, $\mathbf{C}_v$] and graph-cut as the **base** system. Adding selective pairwise pooling $\mathbf{C}_p$ to [$\mathbf{C}_1$, $\mathbf{C}_v$] increases the Dice coefficient by $\sim 2.5\%$. We summarize the primary approach using piecewise linear classification ($\mathbf{C}_{prior}$, $\mathbf{C}_1$, $\mathbf{C}_{1U}$, $\mathbf{C}_{1L}$ and $\mathbf{C}_p$) with added refinements $\mathbf{C}_v$ and graph-cut. The precision and recall results of different system configurations are reported in Table 1. C_{prior} does not influence sensitivity much but improves the precision scores from **Base** to **Base+**, and from **Base++** to **Full**. Zoned piecewise linear classification ($\mathbf{C}_{1U}$ and $\mathbf{C}_{1L}$) in **Base++** only outperforms $\mathbf{C}_1$ in **Base**.

Quantitative Evaluation & Comparison: Previous bone segmentation work focuses on specific bone categories, e.g., the spine [39, 30] and the hip [31]. Our method is directly targeted to the full-body skeleton in any volumetric images. Since it requires no 3D mesh based statistical shape models, the mesh surface-to-surface distance metric [30] on segmentation accuracy measurement is not applicable. Instead, we compare our results with those from [39, 31], which use overlapping precision, recall and Dice coefficients between the computed and ground-truth 3D bone masks. For our two targeted applications, namely bone pathology detection and attenuation correction, volumetric segmentation accuracy is more meaningful than surface-distance error metrics, such as the Hausdorff distance. [39] tackles spine column segmentation using 20 CT scans, and [31] employs 12 CT volumes on hip segmentation via voxel labeling and graph-cut [8]. The available full-body solution is a commercial bone-removal software (BRS) from MeVisLab that relies on fuzzy voxel intensity thresholding and region growing, and thus severely undersegments the vertebra body and bone marrow ($\sim 60\%$ overall sensitivity). Our method significantly outperforms the techniques from all three previous works in segmentation accuracy. (We do not have access to the data in [39, 31], but we evaluate BRS on our datasets). The reader may confirm this by referring to Table 1 and may also compare figures (a,c) to figures (b,d) in Fig. 5. Classifying 2D, 2.5D (RGB-D), and 3D regional candidates (supervoxels in this case) using middle-level image features can yield signicantly better results than per-pixel/voxel based MRF and CRF methods [2].

A major source of error is that ITK 3D-watershed surfaces are not always [48, 23] well-aligned with the ground-truth boundary. This may be addressed by running a fast narrow-band level-set surface evolution [24] to locally optimize the boundary. Also, large calcified vessel lesions residing close to bone structures (e.g., aorta near spine) are sometimes inaccurately segmented. Qualitative examples of bone segmentation in different views are shown in Fig 6.

5 Conclusion

In this paper, we present an efficient bottom-up approach for accurate bone segmentation in challenging 3D CT images. Our method exploits the roles of high boundary-recall regions, middle-level image features and CRF contextualized parsing and optimization. The main applications of this method are usage as a prerequisite segmentation step for full-body bone pathology diagnosis and semantic bone-aware PET-CT attenuation correction (to make new provisional imaging protocols possible in the future). Our method is applicable for bottom-up abdomen CT organ

	Unary-1	**Unary-2**	**Base**	**Base+**	**Base++**	**Full**	Spine [39]	Hip [31]	BRS
Precision	92.6%	94.3%	96.7%	98.1%	95.3%	98.6%	–	–	93.5%
Recall	87.9%	90.4%	92.8%	92.8%	97.9%	97.9%	–	$73.1 - 86.4\%$	57.2%
Dice	90.2%	92.3%	94.7%	95.4%	96.6%	98.2%	93.4%	$65 - 92\%$	71.0%

Table 1. *Precisions (1st row), recalls (2nd row) and Dice coefficients under different algorithm configuration settings and other methods [39, 31], including a bone-removal software (BRS).* **Unary-1***: only* $\boldsymbol{C}_1$ *is used;* **Unary-2***: two unary energy terms, (*$\boldsymbol{C}_{1U}$ *and* $\boldsymbol{C}_{1L}$*), are used;* **Base+***:* **Base** *+* C_{prior}*;* **Base++** *:* **Base** *with* $\boldsymbol{C}_1$ *replaced by* $\boldsymbol{C}_{1U}$ *and* $\boldsymbol{C}_{1L}$*. [31] achieves sensitivities of* $73.1 - 86.4\%$ *and Dice coefficients of* $65 - 92\%$ *for Fibrotic, Trabecular and Cortical bone tissue voxels. We have no access to data in [39, 31], but the BRS is evaluated on our datasets.*

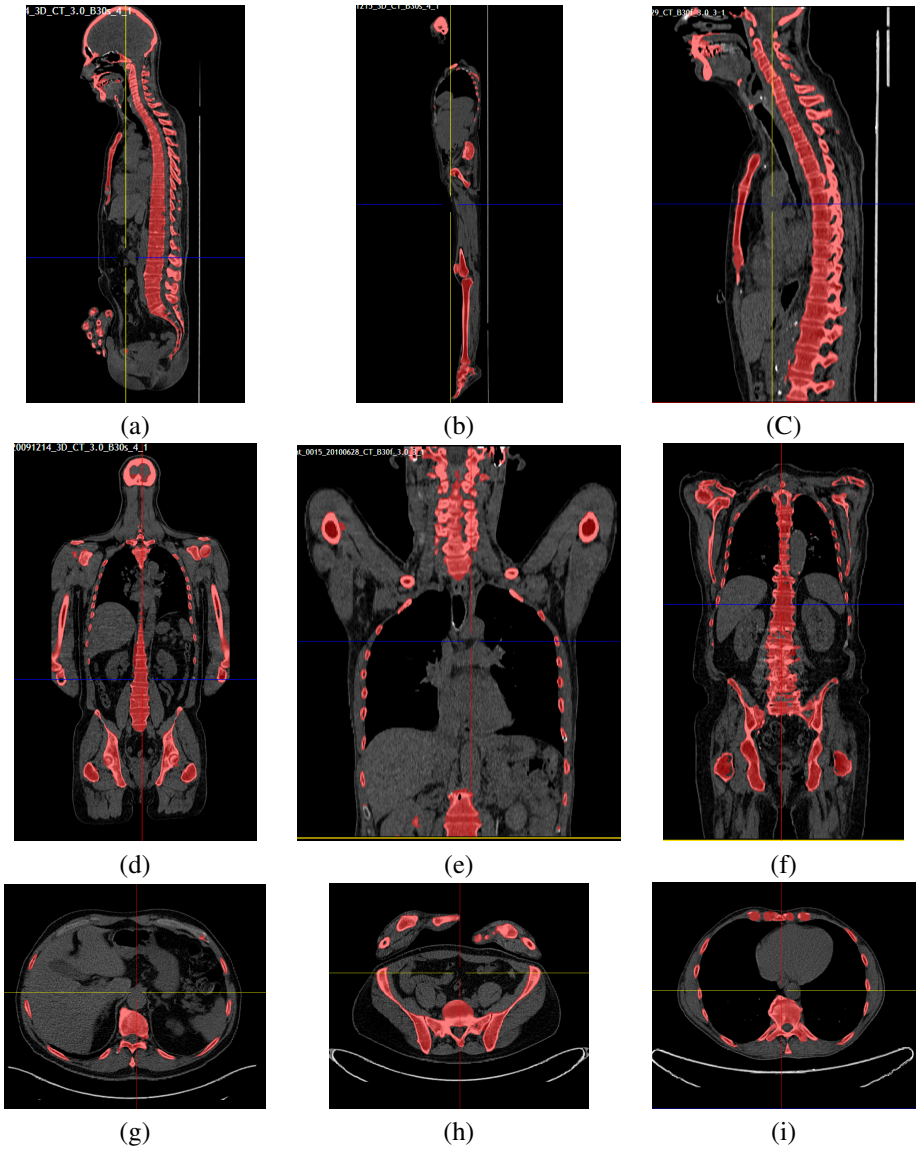


Figure 6. *Qualitative examples of skeleton masking in different views, field-of -views (FOV) and body postures. (a,b,c): sagittal views of typical full-body, head-to-toe and thoracic FOVs are presented; (d,e,f): coronal views illustrate different arm poses, (f): this example yields a slightly lower recall due to severe metal artifacts/implants on vertebral bodies; (g,h,i): axial views. Images (g,i) demonstrate accurate segmentation over vertebral diseases; (h) shows an unusual body pose with the hands in the front of torso.*

and pathology segmentation due to its representation flexibility. Different supervoxel generation schemes (e.g., SLIC, Entropy-Rate Clustering [1, 28] preserving soft-tissue organ boundaries) may be necessary.

Acknowledgements

We thank Dr. Timo Kohlberger (Google), Dr. Neil Birkbeck (Google) and Dr. Shaohua Zhou (Siemens Healthcare) for useful and insightful discussions. We also thank Dr.

Adam P. Harrison (University of Alberta) for making the supervoxel visualization & annotation software tool. This research was supported in part by the Intramural Research Program of the NIH, Clinical Center.

References

[1] A. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–82, 2012. 3, 8

[2] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. *CVPR*, pages 3378–85, 2012. 4, 7

[3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):989–916, 2011. 2

[4] P. Arbelez, J. Pont-Tuset, J. Barron, F. Marqus, and J. Malik. Multiscale combinatorial grouping. *CVPR*, 2014. 4

[5] J. Barron, P. Arbelez, S. Kernen, M. Biggin, D. Knowles, and J. Malik. Volumetric semantic segmentation using pyramid context features. *ICCV*, 2013. 2

[6] X. Boix and et al. Harmony potentials - fusing global and local scale for semantic image segmentation. *Int. J. Comp. Vis.*, pages 83–102, 2012. 2, 4, 6

[7] Y. Boureau, N. LeRoux, F. Bach, J. Ponce, and Y. LeCun. Ask the locals: multi-way local pooling for image recognition. *ICCV*, pages 2651–58, 2011. 5

[8] Y. Boykov and G. Funka-Lea. Graph cuts and efcient nd image segmentation. *IJCV*, 17:109–131, 2006. 5, 6, 7

[9] F. Calderero and F. Marques. Region merging techniques using information theory statistical measures. *IEEE Trans. Image Proc.*, 19(6), 2010. 3

[10] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. *ECCV*, pages 430–443, 2012. 2, 4, 5

[11] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–19, 2002. 2

[12] T. Cour, F. Benezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. *CVPR*, 2005. 2

[13] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui. Regression forests for efficient anatomy detection and localization in computed tomography scans. *Medical Image Analysis*, 17(8):1293–1303, 2013. 2

[14] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge (voc2012) results. 2012. 2, 4

[15] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004. 2

[16] J. Folkesson, E. B. Dam, O. F. Olsen, P. C. Pettersen, and C. Christiansen. Segmenting articular cartilage automatically using a voxel classification approach. *Medical Imaging, IEEE Trans. on*, 26(1):106–115, 2007. 2

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, 2014. 2, 4

[18] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29:2247–53, 2007. 4

[19] B. Hariharan, P. Arbelez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. *ECCV*, 2014. 2, 4

[20] Y. Huang, Z. Wu, L. Wang, and T. Tan. Feature coding in image classification: A comprehensive study. *TPAMI*, 36(3):493–506, 2014. 4

[21] I. Isgum, M. Staring, and et al. Multi-atlas-based segmentation with local decision fusionapplication to cardiac and aortic segmentation in ct scans. *IEEE Trans. Med. Imaging*, 28(7):1000–1010, 2009. 2

[22] J. Kim, J. Lee, and D. Lee. Comparison of segmentation-based attenuation correction methods for pet/mri: evaluation of bone and liver standardized uptake value with oncologic pet/ct data. *J. Nucl. Med.*, 53(12):1878–82, 2012. 1

[23] Kitware. Insight segmentation and registration toolkit (itk). 2013. 3, 7

[24] T. Kohlberger, M. Sofka, and et al. Automatic multi-organ segmentation using learning-based segmentation and level set optimization. *MICCAI*, 2011. 7

[25] A. Kolesnikov, M. Guillaumin, V. Ferrari, and C. Lampert. Closed-form training of conditional random fields for large scale image segmentation. *CoRR abs/1403.7057*, 2014. 6

[26] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012. 4

[27] H. Ling and et al. Hierarchical, learning-based automatic liver segmentation. *CVPR*, 2008. 2

[28] M. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy-rate clustering: Cluster analysis via maximizing a submodular function subject to a matroid constrain. *TPAMI*, 36(1):99–112, 2014. 8

[29] A. Lucchi, K. Smith, A. Achanta, G. Knott, and P. Fua. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE Trans. Med. Imaging*, 31(2):474–486, 2012. 4

[30] J. Ma and L. Lu. Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model. *CVIU*, 117(9):1072–83, 2013. 1, 7

[31] D. Malan, C. Botha, and E. Valstar. Voxel classication and graph cuts for automated segmentation of pathological periprosthetic hip anatomy. *Int. J. CARS*, 8:63–74, 2013. 1, 2, 7, 8

[32] M. Mathias, R. Benenson, M. Pedersoli, and L. V. Gool. Face detection without bells and whistles. *ECCV*, 2014. 4

[33] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. *CVPR*, 2014. 6

[34] V. Raykar, B. Krishnapuram, J. Bi, M. Dundar, and R. Rao. Bayesian multiple instance learning: automatic feature selection and inductive transfer. *ICML*, pages 808–815, 2008. 5

[35] H. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. Summers. Improving computer-aided detection using

convolutional neural networks and random view aggregation. *IEEE Trans. on Medical Imaging*, 2015. 1

[36] H. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. M. Cherry, E. Turkbey, and R. M. Summers. Improving computer-aided detection using convolutional neural networks and random view aggregation. In *IEEE Trans. on Medical Imaging, to appear*. 2015. 4

[37] H. Roth, J. Yao, L. Lu, J. Burns, and R. Summers. Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications. *MICCAI Spine Imaging Workshop*, 2014. 1

[38] P. Schleyer, T. Schaeffter, and P. Marsden. The effect of inaccurate bone attenuation coefficient and segmentation on reconstructed pet images. *Nucl. Med. Commun.*, 31(8):708–716, 2010. 1

[39] M. Schwier, T. Chitiboi, T. Hulnhagen, and H. Kahn. Automated spine and vertebrae detection in ct images using object-based image analysis. *Int. J. Numer. Meth. Biomed. Engng.*, 29(9):938–63, 2013. 1, 2, 7, 8

[40] E. Siris and et al. Identification and fracture outcomes of undiagnosed low bone mineral density in postmenopausal womenresults from the national osteoporosis risk assessment. *Journal of American Medical Association*, 286(22):2815–2822, 2001. 1

[41] J. Snchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013. 4

[42] S. Tan, J. Yao, and et al. Computer aided evaluation of ankylosing spondylitis using high-resolution ct. *IEEE Trans. Med. Imaging*, 27(9):1252–67, 2008. 1

[43] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide baseline stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(5), 2009. 4

[44] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–71, 2013. 2, 3

[45] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *TPAMI*, 34(3):480–492, 2012. 4

[46] J. Vert, J. Qiu, and W. Noble. A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics*, 8(10), 2007. 6

[47] J. Vert and Y. Yamanishi. Supervised graph inference. *NIPS*, 2005. 6

[48] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(6):583–598, 1991. 2, 3, 7

[49] P. Williams. Bayesian regularisation and pruning using a laplace prior. *Neural Comput.*, 2:117–143, 1994. 5

[50] R. Wolz, C. Chengwen, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert. Automated abdominal multi-organ segmentation with subject-specific atlas generation. *IEEE Trans. Med. Imaging*, 32(9):1723–1730, 2013. 2

[51] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. *CVPR*, 2012. 6

[52] Y. Zheng, A. Barbu, and et al. Four-chamber heart modeling and automatic segmentation for 3-d cardiac ct volumes using marginal space learning and steerable features. *IEEE Trans. Med. Imaging*, 27(11):1668–1681, 2008. 2