This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

# A Randomized Ensemble Approach to Industrial CT Segmentation

Hyojin Kim * Jayaraman J. Thiagarajan Peer-Timo Bremer
Lawrence Livermore National Laboratory
7000 East Avenue, Livermore, CA, USA
{kim63, jayaramanthi1, bremer5}@llnl.gov

## Abstract

*Tuning the models and parameters of common segmentation approaches is challenging especially in the presence of noise and artifacts. Ensemble-based techniques attempt to compensate by randomly varying models and/or parameters to create a diverse set of hypotheses, which are subsequently ranked to arrive at the best solution. However, these methods have been restricted to cases where the underlying models are well established, e.g. natural images. In practice, it is difficult to determine a suitable base-model and the amount of randomization required. Furthermore, for multi-object scenes no single hypothesis may perform well for all objects, reducing the overall quality of the results.*

*This paper presents a new ensemble-based segmentation framework for industrial CT images demonstrating that comparatively simple models and randomization strategies can significantly improve the result over existing techniques. Furthermore, we introduce a per-object based ranking, followed by a consensus inference that can outperform even the best case scenario of existing hypothesis ranking approaches. We demonstrate the effectiveness of our approach using a set of noise and artifact rich CT images from baggage security and show that it significantly outperforms existing solutions in this area.*

## 1. Introduction

Broadly speaking, the goal of image segmentation is to use the low level information at each voxel to infer high level semantics such as objects. However, in applications where the source data contains large amounts of noise and artifacts low level features such as edges or voxel values become unreliable. A common solution is to require additional domain knowledge, *i.e*. the shape of an organ in medical CT segmentation, to constrain the segmentation [14, 19]. However, many applications do not readily admit a parameterized model, for example due to the sheer variety of objects to consider. In such cases the additional information is typically given in form of training data, providing examples of objects of interest. Traditionally, the training data is used to tune parameters of the segmentation algorithm, *e.g*. thresholds, energy functionals, etc. However, this process can be labor intensive, is difficult to control, and the results typically do not generalize gracefully. Despite sophisticated optimization tools for inference with segmentation models (*e.g*., Markov Random Fields), the underlying model, learned from a finite training set, is often insufficient to produce accurate results. In order to bridge this gap, approaches that learn multiple hypotheses to produce an overall more accurate solution have been developed [9, 24]. In these techniques both model and/or parameters are varied to create an ensemble of possible solutions. These are then ranked using information from the training data to choose the *best* hypothesis from the ensemble.

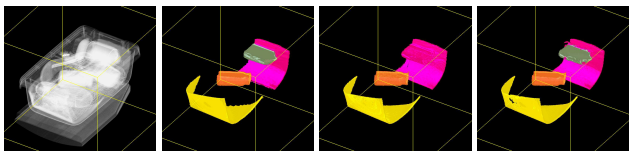

Figure 1. The difficulty in choosing the optimal parameters limits the performance of industrial CT segmentation. From *left* to *right*, the original CT image, ground-truth segmentation, region-growing method which wrongly merges two objects even after careful parameter tuning, and the proposed segmentation.

Ensemble approaches are attractive since their randomized nature can compensate for some level of noise and artifacts. However, adapting these ideas to new applications can be challenging. In particular, picking an appropriate base-model to vary is difficult as is understanding the amount of diversity required to produce good results. Furthermore, for complex, multi-object segmentations no sin-

*This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-PROC-677352. This material is based upon work supported by the U.S. Department of Homeland Security, Science and Technology Directorate, Office of University Programs, under Grant Award 2013-ST-061-ED0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

gle hypothesis may be accurate for all objects equally. Finally, creating optimal ranking criteria is challenging and it is well known that for many metrics the ground truth solution can be ranked lower than inferior results [20].

Instead, we introduce a new ensemble based segmentation framework that uses a simple bottom-up hierarchical segmentation with a randomized merge order to create multiple hypotheses, similar to the approach in [11]. We demonstrate that with a reasonable degree of randomness, our method can generate hypotheses that are significantly better than the greedy solution and can compensate for a large amount of noise and artifacts. Furthermore, rather than ranking different hypotheses as a whole, we explore the space of all potential objects in the ensemble and use the training data to identify likely matches. The different variations of a (potential) object are then combined in per-object consensus segmentation to arrive at the final result. This enables us to exploit locally accurate segmentations from globally sub-optimal hypotheses. Additionally, this strategy reduces the dependency on an optimal ranking criterion as we no longer use it to evaluate the final result but only to extract likely candidate objects. We shown that by locally combining information from multiple hypothesis in this manner our approach can outperform even the best case scenarios in existing ranked hypothesis approaches, making our results qualitatively different from the one in [24]. We demonstrate the effectiveness of our system using a challenging set of CT scans from a baggage security system. This data is well known to contain noise as well as severe artifacts which makes many existing segmentation methods ineffective. In particular, as shown in Figure 1 and discussed in Section 5, our results are significantly better than even hand-tuned versions of existing methods.

## 2. Related Work and Contributions

**Finding multiple hypotheses**: The idea of identifying multiple hypotheses has been explored in a variety of computer vision problems. In particular, a class of methods collectively referred as "M-Best Map" have been successfully used to generate multiple configurations for image segmentation [16, 25, 9]. However, these methods produce solutions that tend to be very similar to the Maximum a Posteriori (MAP) solution and each other. Batra *et al.* [4] developed a sequential model selection technique that emphasized the diversity of the solutions and showed it can produce significantly better results. An alternative approach to generating multiple hypotheses is to use sampling strategies that perturb the parameters of a segmentation algorithm [6, 18, 17]. However, refining these solutions can be challenging if the data is sensitive to the parameter settings. Recently, Kim *et al.* [11] proposed an ensemble creation strategy that randomized the merge order in bottom-up hierarchical segmentation for foreground-background separation.

**Industrial CT segmentation**: Three dimensional CT image segmentation is a well-studied problem, and used in a wide variety of applications [23]. One of the most challenging aspects of industrial CT segmentation is the presence of severe metal artifacts in form of streaks, blooming, or cupping (see Section 4). In several of these applications, it is typical to start with a prior knowledge (or parametric model) of the objects present in the image, *e.g.* mechanical part, and use accurate segmentation results to compensate for metal artifacts and other sources of noise, so that interesting anomalies, such as defects, can be easily identified. For example, Li et al. [12] adopt a non-parametric estimation method to estimate the spatial probability distribution of gray-level intensities, and use the minimum cross entropy technique to segment an object of interest. In [2], the authors address a more challenging problem of detecting metal features of varying thickness, and showed that region-growing is very effective in such cases. Nevertheless, these techniques are targeted specifically to metal objects and do not perform well for other materials.

**Transportation security**: Finally we review the application considered in this paper, where the goal is to identify potentially suspicious material signatures from baggage scans. In [15], the authors adopt a fuzzy connectedness technique to obtain an initial object segmentation for detecting potential threats. However, this system requires extensive parameter-tuning, and cannot easily generalize to a broader class of objects or materials. In order to improve robustness against artifacts, Stratovan Tumbler, a medical image segmentation framework, has been adapted for delineating objects in baggage scans [22] and has shown to be effective in partitioning some heterogeneous objects (for example, parts of a laptop). Since the performance of bottom-up hierarchical segmentation depends heavily on the merging order, the authors in [10] proposed a reverse approach which begins by separating the set of object voxels from the background, and then creates candidate splits into individual objects based on global criteria. However, the process of identifying object voxels typically needs rigorous training and appears very sensitive to image artifacts.

### 2.1. Contributions

In this paper, we propose to build an ensemble of hierarchical segmentations for industrial CT volumes, and leverage semantic information to perform localized consensus inference. Our contributions in detail are

**Segmentation Ensembles**: Following the work in [11], we build an algorithm to create randomized ensembles for CT volumes, and empirically determine the required degree of randomness to compensate for the inherent uncertainties;

**Semantic Candidate Selection**: We develop a novel discriminative feature for regions in CT images, and design a simple reference-based scheme for identifying potential

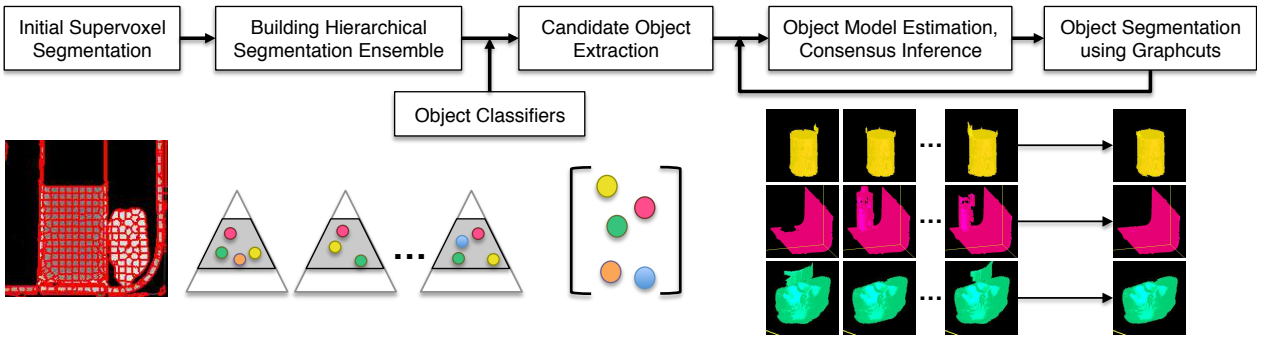


Figure 2. An overview of the proposed approach for CT segmentation. Starting with an initial oversegmentation of the volume, it builds an ensemble of hierarchical segmentations and exploits the semantic information from supervisory data to identify candidate segments, that are likely to contain objects of interest. Finally, consensus segmentation with graphcuts provides the overall partitioning.

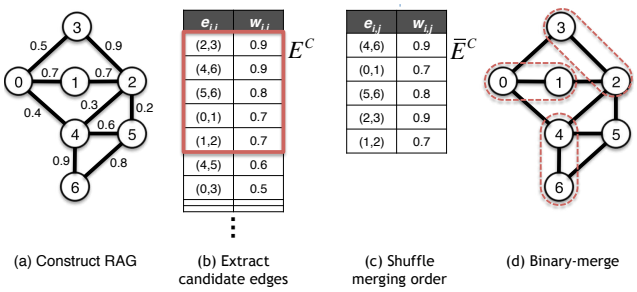


Figure 3. Hierarchy construction through randomization. In each level of a hierarchy, candidate edges are randomly shuffled and these edges are incrementally merged.

segments of interest;

**Consensus Inference**: Instead of ranking the hypotheses in an ensemble, we propose to obtain a localized consensus inference using graph-cuts for each object of interest; and

**Application to Airport Security**: We use the proposed method in a transportation security application and demonstrate its effectiveness in comparison to the popularly adopted region-growing methods, using a challenging dataset provided by the Awareness and Localization of Explosives-Related Threats (ALERT) Center of Excellence.

## 3. Proposed Approach

As illustrated in Figure 2, our method is comprised of four stages: (a) initial oversegmentation, (b) building segmentation ensembles, (c) semantic candidate selection, and (d) localized consensus inference.

### 3.1. Initial Oversegmentation

Similar to several existing approaches, we begin by oversegmenting the CT volume to create perceptually meaningful atomic groups, referred to as supervoxels. In addition to providing a non-uniform partitioning, supervoxels can capture the image redundancy, and greatly reduce the computational complexity of subsequent stages. In this paper, we generate the initial supervoxels (each of size $10-12^3$ voxels) using the SLIC algorithm [1].

### 3.2. Building Segmentation Ensembles

Though region-growing methods have produced state-of-the-art results in industrial CT segmentation, we observed that a simple bottom-up hierarchical segmentation can generate greedy solutions of reasonable quality. Motivated by its flexibility and simplicity, as discussed in [11], we adopted a bottom-up approach for creating segmentation ensembles with industrial CT images. Note that, all hierarchies in the ensemble start with the same set of supervoxels. Each hierarchy incrementally merges regions from the previous level. The edge affinity, $w^{\ell}_{i,j}$, between two regions $r^{\ell}_i$ and $r^{\ell}_j$ in level $\ell$ is measured as the similarity between their intensity histograms:

$$w^{\ell}_{i,j} = \exp\big(-\sigma_1 \chi^2(H(r^{\ell}_i), H(r^{\ell}_j))\big). \qquad (1)$$

Here, $H(r^{\ell}_i)$ is the intensity histogram of region $r^{\ell}_i$, $\chi^2$ measures the chi-square distance between two histograms, and $\sigma_1$ is the parameter for the Gaussian radial basis function.

We now generate multiple independent segmentations from the same set of supervoxels by randomizing the merging order of candidate edges, which allows us to explore as many aggregations as possible. At level $\ell$, we sort the edges in the descending order based on their edge weights. We then extract the candidate edge set, $E^{\ell}_C = \{e^{\ell}_{i,j} | w^{\ell}_{i,j} \geq \delta\}$, where $\delta$ is a predefined threshold. From the candidate set, we randomly choose edges sequentially, merge the regions corresponding to that edge if either of the regions have not been merged previously. For this random sampling, we can use a simple uniform distribution for all candidates or create a discrete distribution that is proportional to the edge similarity. Figure 3 illustrates this randomization procedure. Note that, in contrast to the approach in [11], we can control the degree of randomness through the parameter $\delta$. Low degree of randomness will produce hypotheses that are

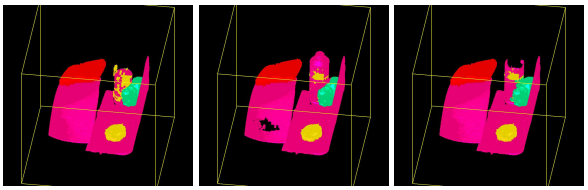

Figure 4. Candidate object regions of different hierarchical segmentations, each of which shows a different object configuration.

very similar to the greedy solution. Whereas, arbitrarily increasing the degree of randomness can result in inaccurate segmentations. In Section 5, we will empirically study the effect of randomness on the solution diversity, and show that the greedy solution can be significantly improved by compensating for artifacts with a low degree of randomness in the merging order. Note that, this process is computationally very simple when compared to solving a complex, discrete optimization problem as in [4].

### 3.3. Semantic Candidate Selection

Given a set of hypotheses, it is typical to adopt a supervisory approach that predictively ranks the solutions based on object plausibility [6] or other criterion on the segmentation quality [24]. These approaches assume that for an image with multiple objects, the *best* hypothesis can provide accurate segmentations for all objects equally. However, in industrial CT images with severe non-uniform artifacts, finding such an optimal hypothesis is very difficult. The proposed approach addresses this challenge by using supervisory data to filter the large pool of segments from the ensemble and identify a small set of candidate segments, that can potentially contain the objects of interest. Instead of identifying the *best* hypothesis, this approach identifies multiple configurations of the same object from different hypotheses, and obtains a weighted consensus inference.

In order to identify candidate segments, we propose to build discriminative features for the segments in the hierarchy. We begin by extracting the following set of features for each region in the ensemble: (a) Intensity statistics (mean, standard deviation, and percentiles); (b) Histogram of number of voxels in each radii bin from the center of mass [3]; (c) Area; (d) Volume-to-surface area ratio. Following this, we use local discriminant embedding (LDE) [8], a supervised graph embedding approach, to build a semantic descriptor for each region. Note that, this step can be replaced with any semantic feature learning technique.

Similar to existing supervisory approaches, we assume that the total number of ground truth objects, $N_c$, is known apriori. The features for the ground truth segments in the training data are stored in the matrix $\mathbf{X} = [\mathbf{x}_i]_{i=1}^{T}$ and their class labels are denoted as $\{y_i | y_i \in \{1, 2, \ldots, N_c\}\}_{i=1}^{T}$. The goal of LDE is to exploit both the supervisory label information, and the local structure in data to create a subspace representation that can discriminate between different classes of objects. We construct the undirected, intra-class and inter-class graphs $G$ and $G'$ respectively, and the edges between the samples are coded in the affinity matrices $\mathbf{W}$ and $\mathbf{W}'$. The affinities are defined as follows:

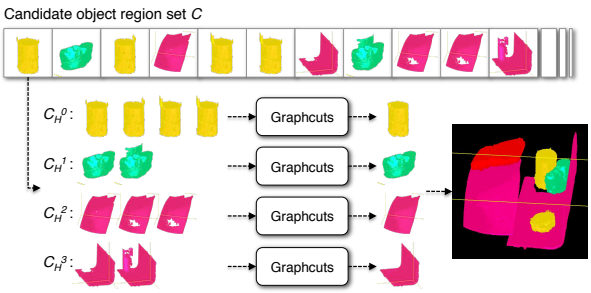


Figure 5. Localized consensus inference for each potential object.

$$w_{ij} = \begin{cases} 1 & \text{if } y_i = y_j \text{ AND } [i \in \mathcal{N}_k(j) \text{ OR } j \in \mathcal{N}_k(i)], \\ 0 & \text{otherwise.} \end{cases}$$

$$w'_{ij} = \begin{cases} 1 & \text{if } y_i \neq y_j \text{ AND } [i \in \mathcal{N}'_k(j) \text{ OR } j \in \mathcal{N}'_k(i)], \\ 0 & \text{otherwise.} \end{cases}$$

Here $\mathcal{N}_k(i)$ and $\mathcal{N}'_k(i)$ denote the intra-class and inter-class neighborhood for the sample $\mathbf{x}_i$. Following this, we build the intra-class graph Laplacian as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a degree matrix with each diagonal element containing the sum of the corresponding row or column of $\mathbf{L}$. Similarly, we construct the inter-class graph Laplacian $\mathbf{L}'$. The $d$ projection directions for LDE, $\mathbf{V}$, is computed by optimizing

$$\max_{\mathbf{V}} \frac{\text{Tr}[\mathbf{V}^T\mathbf{X}^T\mathbf{L}'\mathbf{X}\mathbf{V}]}{\text{Tr}[\mathbf{V}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{V}]}. \tag{2}$$

Instead of finding the global solution to the trace-ratio maximization problem in (2), a greedy solution can be obtained by converting it to an equivalent ratio-trace maximization, $\max_{\mathbf{V}} \text{Tr}[(\mathbf{V}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{V})^{-1}\mathbf{V}^T\mathbf{X}^T\mathbf{L}'\mathbf{X}\mathbf{V}]$. The solution to this problem can be obtained using the generalized eigen value decomposition.

Given the semantic descriptor for a segment, $\mathbf{V}^T\mathbf{x}$, our goal is to estimate the likelihood of that segment containing each of the $N_c$ objects. Though any non-parametric modeling technique can be used to obtain the likelihood estimates, we observed that a simple reference-based scheme [13] was sufficient for this task. By computing the average similarity of the semantic descriptor for a segment to each class of ground truth data, we measure the relevance of each class. We use the following similarity metric:

$$S(r, g_i^k) = 1 - \frac{\gamma\left(\frac{k}{2}, \frac{d(r, g_i^k)}{2}\right)}{\Gamma(\frac{t}{2})}, \tag{3}$$

where $d(r, g_i^k)$ denotes the $\chi^2$ distance between the semantic descriptor of a region (segment) $r$ and that of the $i^{\text{th}}$ training sample in class $k$, $\gamma(.)$ is the lower incomplete gamma function, $\Gamma$ denotes the gamma function, and $t$ is a positive integer that specifies the number of degrees of freedom (set to 4 in our experiments). The second term in the expression is the cumulative distribution for chi-squared distribution. For a ground truth class $k$, we use the average similarity of the segment with respect to all samples in that class to define the likelihood,

$$L_k(r) = \frac{1}{n_k} \sum_{i=1}^{n_k} S(r, g_i^k). \tag{4}$$

For each hierarchy $m$, we evaluate the likelihood $L$ for all regions whose volume is higher than a domain-specific threshold, and assign the region to one of the $N_c$ classes, if the corresponding likelihood (referred to as the confidence measure) is greater than a confidence threshold $\kappa$. We explore all levels in the hierarchy and retain only the most likely regions. Note that, there can be multiple candidate segments for an object within the same hierarchy. The overall candidate segment set, $C$, is created by merging the individual sets from all hierarchies (Figure 4).

### 3.4. Consensus Inference using Graphcuts

As discussed earlier, for each potential object, we obtain a consensus inference using candidates from multiple hypotheses. As shown in Figure 5, we begin by sorting all candidate regions from $C$ in the decreasing order of their confidence measures. Following this, we pick the candidate region with the highest confidence, $c_0 \in C$, and collect the set of regions ($C_H$) that have a high volume overlap ratio with $c_0$ (set to 0.6 in our framework). We propose to perform consensus graphcut segmentation on the union of regions from $C_H$. Let us define the set of supervoxels in the union of $C_H$ as $V_0$, and the corresponding set of edges by $E_0$. The segmentation indicator set $A = \{\alpha_i\}$ defines a binary label $\alpha_i$ (foreground/background) for each supervoxel in $V_0$. The Markov Random Field (MRF) formulation for graphcuts [5] can be expressed as

$$F(\mathbf{A}) = \sum_{r_i^0 \in V_0} F_d(\alpha_i) + \lambda \sum_{e_{i,j}^0 \in E_0} F_s(\alpha_i, \alpha_j). \tag{5}$$

To define the data penalty, $F_d$, we build 256-bin intensity histograms for all supervoxels, and build parametric models for foreground/background regions. We experimented with few modeling choices, including the popularly adopted Gaussian Mixture Models, and found that a simple K-Means clustering with $K = 2$ produced very similar results. Given the two cluster centroids $H_0$ and $H_1$, the data cost for a supervoxel $r_i^0$, $F_d(\alpha_i)$, becomes $\exp\big(-\gamma\chi^2(H(r_i^0), H_0)\big)$, and $\exp\big(-\gamma\chi^2(H(r_i^0), H_1)\big)$ for the foreground and background labels respectively. To define $F_s$, we perform a consensus inference on the supervoxel composition from all candidate regions in $C_H$. We first initialize all entries of the consensus matrix $\mathbf{M}$ to 1. For any pair of supervoxels, we count the number of candidate regions where the two supervoxels are not merged ($\eta(i,j)$), and we update the entry $m_{i,j}$ as $1 - (\eta(i,j)/N_H)$, where $N_H$ is the total number of elements in $C_H$. The resulting segmentation is stored, all candidate regions used in the current iteration ($C_H$) are removed from $C$ and this procedure is repeated until $C = \emptyset$.

## 4. Application to Transportation Security

CT imaging has found wide-spread use in applications, beyond medical diagnosis, such as material characterization and transportation security. While each area uses different imaging modalities, the overarching problem is to find objects or other prominent features in noisy, cluttered images.

In this paper, we apply our technique to the problem to the CT scans of checked luggage. This is a pressing problem of national interest as checked luggage represents a significant risk factor for commercial air traffic. The dataset has been generated by the DHS ALERT Center of Excellence at Northeastern University [7] to develop and test Automatic Threat Recognition (ATR) systems. The data set contains 188 bags (100 bags for training and the rest for testing) and each bag's reconstructed volume varies between $512 \times 512 \times 180$ and $512 \times 512 \times 420$, depending on the size of the bag. All bags contain a variety of everyday objects, *e.g.* clothes, food, electronics, etc., alongside one or more simulated threats. There are 3 different targets of interest: saline solution in bottles or bags, rubber sheets of varying thickness, and modeling clay in different shapes and quantities. The targets of interest are hand-labeled and the ultimate goal is to accurately identify all targets. However, as shown in Section 5 the data contains severe noise and artifacts and suffers from partial volume effects due to insufficient spatial resolution. Note that, having an accurate estimation of the volume and configuration of target materials is essential in assessing a potential threat.

As described in Section 2, existing frameworks for analyzing airport scans are designed for specific tasks, and it is not straightforward to adopt them to a general ATR problem. Hence, we resort to comparing our method to the popularly adopted region-growing technique. In addition to demonstrating the effectiveness of the proposed method, we study empirically the effectiveness of randomness and ensemble size on the segmentation performance.

## 5. Results and Discussion

The crucial parameters in our setup are the ensemble size $M$ and the level of randomness in the ensemble creation, for

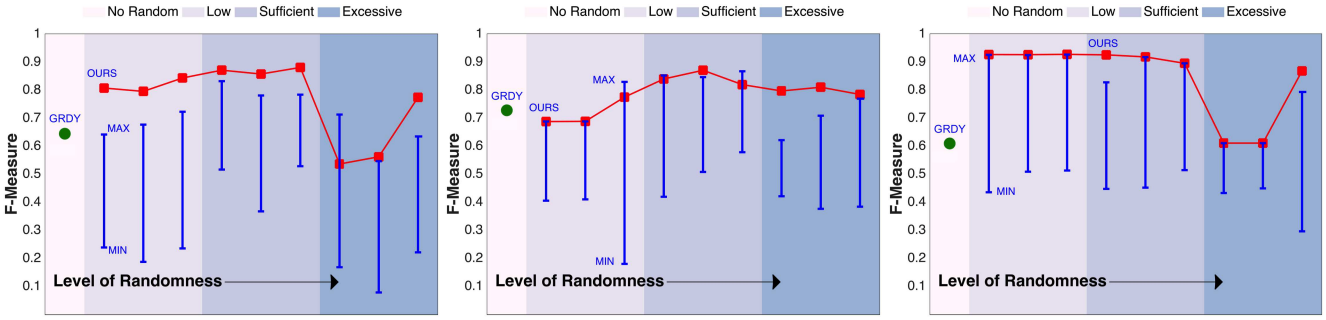


Figure 6. Effect of randomness on the segmentation performance. For three example images, we show the F-measures for the most and the least accurate hypotheses from the ensemble at increasing levels of randomness. For comparison, we include the accuracies for the greedy solution and the proposed consensus inference. At low degrees of randomness, the method fails to compensate for the image uncertainties, while with excessive randomness is unable to produce highly accurate hypotheses, especially for small ensemble sizes.

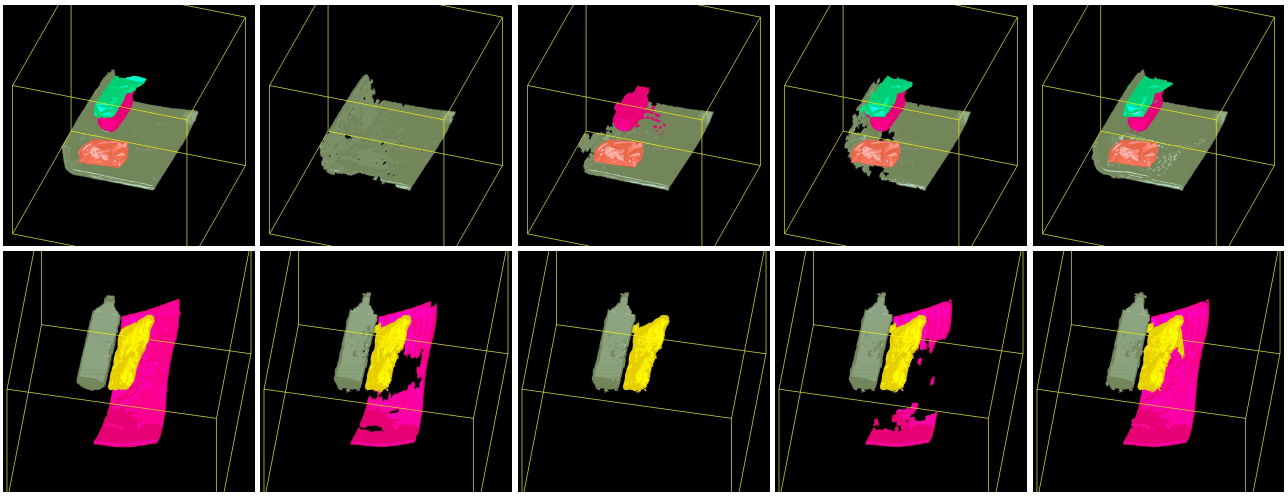

Figure 7. Effect of ensemble size on the segmentation performance. For two example images, we show the segmentation results obtained using our algorithm at varying ensemble sizes. From *left* to *right* in each row, ground-truth labels, greedy solution, proposed segmentation at $M = 1$, $M = 5$, and $M = 20$ respectively. For a sufficient level of randomness during ensemble construction, we do not observe any significant performance improvement beyond $M = 20$, with this dataset.

which we provide a detailed empirical analysis. To evaluate segmentation performance with respect to the ground-truth labels, we compute the F-measure [21], $\frac{2PR}{P+R}$, where $P$ and $R$ are the precision and recall respectively. For all results reported this section we fixed the LDE dimension $d = 5$, confidence threshold $\kappa = 0.65$, the penalty in (5) at $\lambda = 0.4$

**Effect of randomness:** We proposed to improve the greedy solution by creating an ensemble of segmentations, based on randomized merging during hierarchy construction. The desired level of randomness is provided through the threshold parameter $\delta$. The degree of randomness directly controls the both the segmentation quality and the required ensemble size. At one end, insufficient randomization will produce hypotheses that are mere perturbations of the greedy solution, and hence cannot resolve uncertainties in detecting object boundaries arising due to image artifacts. Whereas, excessive randomization can produce highly inaccurate, albeit diverse, hypotheses. Furthermore, at a higher degree of randomness, the number of hypotheses required to produce high quality solutions can be quite large. As a result, we need to determine a sufficient level of randomness to compensate for the artifacts, while keeping the ensemble size reasonable. To this end, we ran our segmentation algorithm at increasing levels of randomness ($M$ fixed at 10), and computed the F-measure for each of the hypotheses. We used three levels of randomness, each in turn containing 3 settings for $\delta$, $\{low; 0.98, 0.975, 0.98\}$, $\{sufficient; 0.95, 0.9, 0.85\}$, and $\{excessive; 0.65, 0.6, 0.55\}$. Figure 6 shows the F-measures for the least and the most accurate hypotheses in an ensemble for 3 example images, along with the results of the greedy solution, and the proposed method. As expected, increasing the randomness tends to produce more accurate hypotheses when compared to the greedy solution. Equivalently, the minimum F-measure of an ensemble is significantly lower at high degree of randomness, indicat-

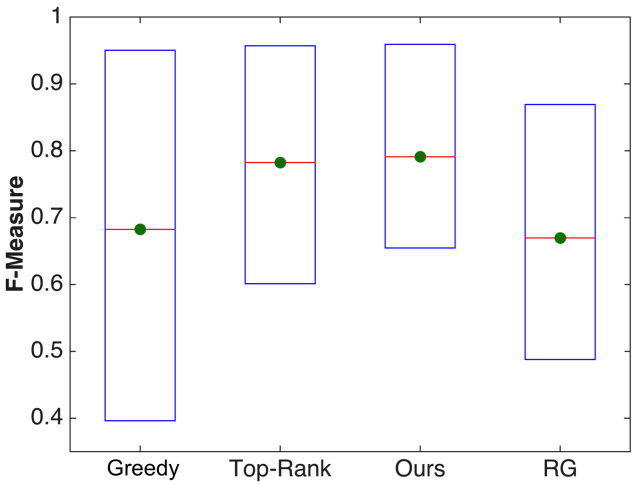


Figure 8. Segmentation performance evaluation - In each case, the green circle corresponds to the mean performance, while the bottom and top edges of the box show the performance at 20 (worst-case behavior) and 80 (optimistic behavior) percentiles.

ing that the ensemble size must be much larger to guarantee the inclusion of good quality hypothesis. For example in 6(a), the method produces poor quality hypotheses at both low and excessive levels of randomness. The former is due to inability of the hierarchical merging process to compensate for the artifacts, while the latter is because the method allows very low probability merges during hierarchy reconstruction. Furthermore, in most cases, our per-object consensus produces more accurate results than the *best* hypothesis, which is the upper-bound on the performance of any hypothesis ranking technique [24]. Note that our approach identifies *potentially correct* hypotheses for each object based on discriminative features and obtains a weighted consensus from them. Consequently, inaccuracies in our feature matching strategy can sometimes make the consensus solution slightly inferior to the *best* hypothesis in the ensemble. From our empirical analysis, we find that sufficient amount of randomness that can provide robust hypotheses can be achieved at $0.85 \leq \delta \geq 0.95$.

**Effect of ensemble size ($M$):** In addition to impacting the segmentation performance, the choice of $M$ determines the computational complexity of our approach. In order to demonstrate its effect on segmentation accuracy, we applied our technique at different ensemble sizes of $M = \{1, 5, 10, 20\}$ for $\delta$ fixed at 0.9. We observed a consistent improvement in performance with increasing ensemble size and no significant improvements beyond $M = 20$. The segmentation results for two example images, at different number of hierarchies, are shown in Figure 7. Although segmentation with larger $M$ is computationally expensive, the construction of the ensemble can be easily parallelized.

**Performance Evaluation:** In order to evaluate the proposed method on the ALERT airport security dataset, we fixed the number of hierarchies, $M = 20$ and the threshold $\delta = 0.9$. Furthermore, we evaluated the performances of the greedy solution and the *best* hypothesis in the ensemble respectively. Note that, the latter denotes the maximum achievable performance of a hypothesis ranking method. Together with these two approaches, we considered a region-growing method, similar to the one proposed in [22], that often produces state-of-the-results in this area. In particular, we ran the region-growing method (RG) at multiple, hand-tuned parameter settings, and report the results from the best-performing parameter setting.

We computed the F-measure of the segmentation for each baggage, using all four methods. Figure 8 shows the mean performance for each case, along with the F-measure at the 20 (worst-case behavior), and 80 (optimistic behavior) percentiles of the total number of test images. First, not so surprisingly, the average performance of the region-growing method is very similar to that of the greedy solution. The optimistic behavior of the greedy solution is significantly better showing that a simple, hierarchical segmentation method is effective for the CT volumes. However, the worst-case performance is quite inferior in comparison to that of RG, implying that this greedy merging process can be very sensitive to the inherent uncertainties arising due to complex object interactions. Furthermore, it is important to note that the performance of region-growing methods is very sensitive to the parameter choices, and hence finding an optimal set of parameters that can generalize is extremely challenging. Overall, when compared to these two benchmarks, the ensemble approach produces qualitatively superior segmentations, with a relatively simpler parameter-tuning process. Furthermore, our proposed per-object consensus inference outperforms the top-rank performance, and in particular its worst-case performance ($\sim 0.65$) is significantly better than that of the latter ($\sim 0.6$). Segmentation results illustrated in Figure 9 evidences the ability of our method to robustly handle complex object configurations.

## 6. Conclusions

We propose a novel segmentation approach using an ensemble of randomized hierarchical segmentations to compensate for noise and artifacts as well as complex object configurations. Our approach leverages supervisory information, in the form of semantic labels, to build discriminative features that enable us to identify potential object segments from the hypotheses in an ensemble. Furthermore, by adopting a per-object consensus inference strategy, we can improve upon the upper-bound performance of hypothesis reranking techniques. Our approach successfully segments threat objects from baggage CT scans, compared to benchmark segmentation approaches for industrial CT segmentation. Empirical studies were carried out to understand

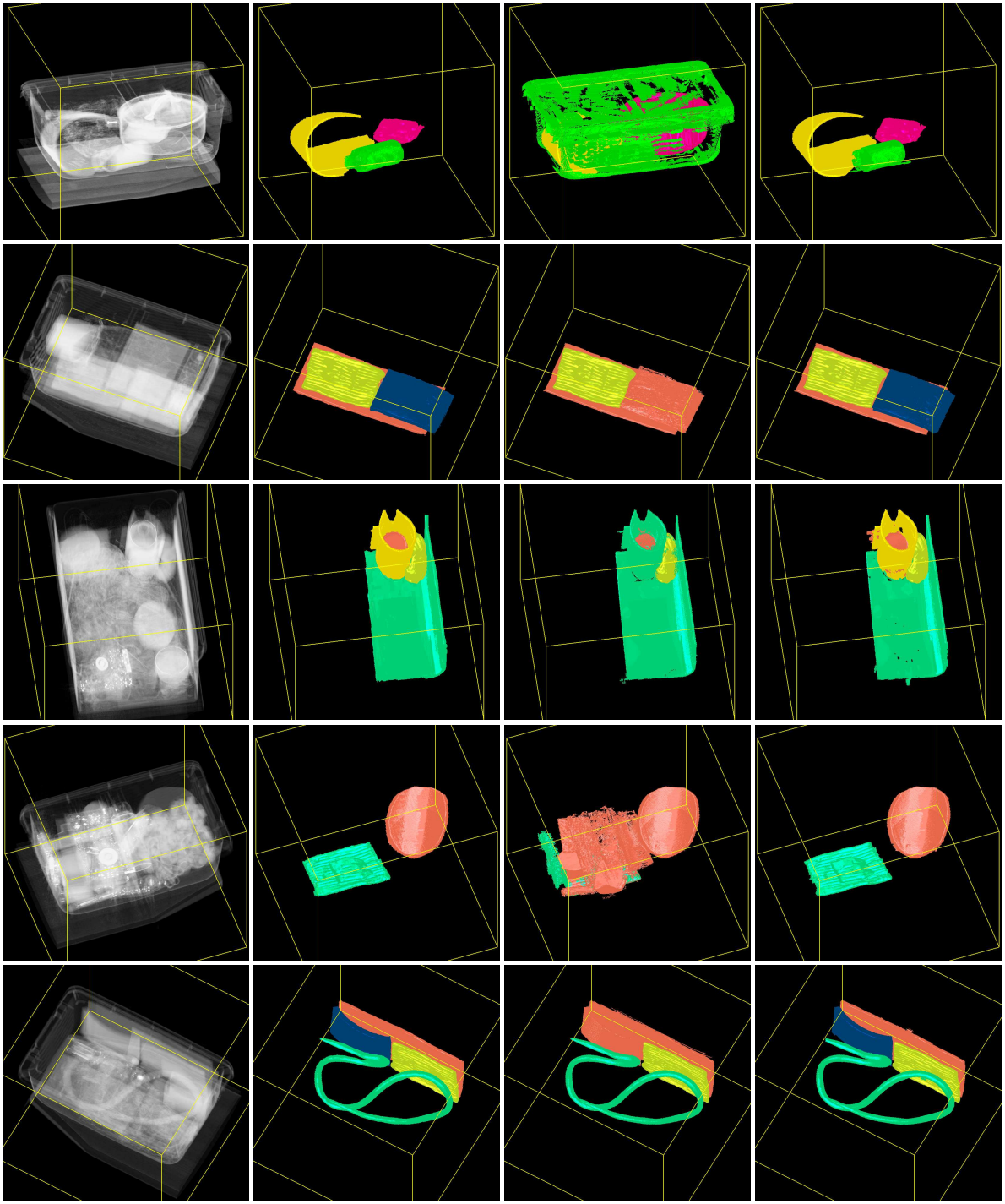

Figure 9. Segmentation results of the proposed algorithms with other methods. From *left* to *right* in each row, the original CT image, ground-truth labels, region growing (RG), and our segmentation.

the effect of randomness during ensemble creation and the ensemble size on segmentation performance. Results show that even with a simple base-model, when coupled with appropriate ensemble construction strategies, can significantly improve the greedy solution, and produce segmentations that are robust against the inherent image artifacts. Future directions of work include designing sophisticated model averaging strategies in lieu of consensus inference, and creating a GPU-parallelized implementation of our ensemble construction.

## References


[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:2274–2282, 2012. 4323

[2] A. Amirkhanov, C. Heinzl, M. Reiter, J. Kastner, and M. E. Groller. Projection-based metal-artifact reduction for industrial 3D X-ray computed tomography. *IEEE Transactions on Visualization and Computer Graphics*, 17:2193–2202, 2011. 4322

[3] M. Ankerst, G. Kastenmuller, H.-P. Kriegel, and T. Seidl. 3D shape histograms for similarity search and classification in spatial databases. In *Proc. of International Symposium on Advances in Spatial Databases*, pages 207–226, 1999. 4324

[4] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse m-best solutions in markov random fields. In *Computer Vision ECCV 2012*, volume 7576 of *Lecture Notes in Computer Science*, pages 1–16, 2012. 4322, 4324

[5] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *Proc. of IEEE ICCV*, pages 105–112, 2001. 4325

[6] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3241–3248, June 2010. 4322, 4324

[7] D. Center of Excellence at Northeastern University. ALERT TO4 datasets for automated threat recognition. Website, 2014. available at http://www.northeastern.edu/alert. 4325

[8] H.-T. Chen, H.-W. Chang, and T.-L. Liu. Local discriminant embedding and its variants. In *Proc. of IEEE CVPR*, pages 846–853, 2005. 4324

[9] M. Fromer and A. Globerson. An LP view of the m-best map problem. In *In Advances in Neural Information Processing Systems 22*, pages 567–575, 2009. 4321, 4322

[10] L. Grady, V. Singh, T. Kohlberger, C. Alvino, and C. Bahlmann. Automatic segmentation of unknown objects, with application to baggage security. In *Proc. of ECCV*, pages 430–444, 2012. 4322

[11] H. Kim, J. Thiagarajan, and P.-T. Bremer. Image segmentation using consensus from hierarchical segmentation ensembles. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 3272–3276, Oct 2014. 4322, 4323

[12] J. Li, L. Wang, and P. Bao. An industrial CT image segmentation algorithm based on non-parameter estimation. *Journal of Computational Information Systems*, pages 3103–3109, 2010. 4322

[13] Q. Li, H. Zhang, J. Guo, B. Bhanu, and L. An. Reference-based scheme combined with k-svd for scene image categorization. *IEEE Signal Process. Lett.*, 20(1):67–70, 2013. 4324

[14] H. Ling, S. K. Zhou, Y. Zheng, B. Georgescu, M. Suehling, and D. Comaniciu. Hierarchical, learning-based automatic liver segmentation. In *Proc. of IEEE CVPR*, pages 1–8, 2008. 4321

[15] N. Megherbi, G. T. Flitton, and T. P. Breckon. A classifier based approach for the detection of potential threats in CT based baggage screening. In *Proc. of IEEE ICIP*, pages 1833–1836, 2010. 4322

[16] D. Nilsson. An efficient algorithm for finding the m most probable configurations in probabilistic expert systems. *Statistics and Computing*, 8:159–173, 1998. 4322

[17] G. Papandreou and A. Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 193–200, Nov 2011. 4322

[18] J. Porway and S.-C. Zhu. $c^4$: Exploring multiple solutions in graphical models by cluster sampling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(9):1713–1727, Sept 2011. 4322

[19] Y. Shi, S. Liao, Y. Gao, D. Zhang, Y. Gao, and D. Shen. Prostate segmentation in CT images via spatial-constrained transductive lasso. In *Proc. of IEEE CVPR*, pages 2227–2234, 2013. 4321

[20] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(6):1068–1080, June 2008. 4322

[21] C. J. van Rijsbergen. *Information Retireval*. Butterworth-Heinemann, London, United Kingdom, second edition edition, 1979. 4326

[22] D. F. Wiley, D. Ghosh, and C. Woodhouse. Automatic segmentation of CT scans of checked baggage. In *Proc. of Image Formation in X-ray CT*, pages 310–313, 2012. 4322, 4327

[23] O. Wirjadi. Survey of 3D image segmentation methods. *ITWM 123 (Technical Report, Fraunhofer ITWM)*, 2007. 4322

[24] P. Yadollahpour, D. Batra, and G. Shakhnarovich. Discriminative re-ranking of diverse segmentations. In *CVPR*, pages 1923–1930, 2013. 4321, 4322, 4324, 4327

[25] C. Yanover and Y. Weiss. Finding the m most probable configurations using loopy belief propagation. In *In Advances in Neural Information Processing Systems 16*. MIT Press, 2004. 4322