Fully-Unsupervised Image Segmentation of CT Scans

Tian Xiang Du

Full credit goes to Asako Kanezaki, the author of the original papers that inspired this project

Introduction:

Image segmentation has been a core research topic in computer vision for several decades, with wide applications in image compression and object detection/recognition. With the advent of deep learning and its application within medicine, image segmentation has become a staple in computer-aided medical technologies. Body part recognition is universally important across medical imaging applications and is increasingly important given the recent rise in machine-learned disease detection and diagnosis systems. Traditional methods use expert human knowledge and labour to construct image classification features and ground-truth labels for images used in supervised learning applications, which create a natural bottleneck for applying deeper, more powerful machine learning models like Convolutional Neural Networks (CNN). However, recent studies have proposed unsupervised learning strategies that could mitigate the demand for manually labelled data, opting for a more automated and generalizable approach to segmenting anatomical features. In this paper, I seek to explore recent unsupervised learning techniques in image segmentation and apply them to anatomical segmentation in the medical imaging space. The objective is to evaluate an unsupervised deep learning model on its ability to produce accurate segmentation of organs and internal structures into distinct categories from CT scans.

Related Work:

CNN-based image segmentation has been gaining attention in literature, often using object detectors or user inputs to determine parameters for segmentation [1]. Weakly supervised learning approaches using bounding boxes or image-level class labels for training have been widely used as a compromise to the manually labelled data bottleneck; such approaches have been successfully applied to lesion segmentation, maintaining comparable performance to fully supervised counterparts [2]. Unsupervised deep learning methods using CNN architecture mainly focus on learning high-level feature representations, while applications to pixel-level annotations in image segmentation have only been recently investigated [3]. While some proposals can train deep CNN filters using standard backpropagation without ground truth labels, use case and performance optimization can be achieved using a semi-supervised approach under medical imaging modalities. Combined with methods like unsupervised body part regression using spatially self-ordering CNNs, a robust system can be created to generate 3D human bodies for patient-specific diagnostic systems [4].

Approach:

The proposed approach is based on Dr. Kanezaki's paper on unsupervised image segmentation by back propagation [5]. It applies a linear classifier on each pixel of an image into q classes and normalized to [0, 1]. We compute a p-dimensional feature map $\{x_n\}$ from $\{v_n\}$ through M convolutional components consisting of 2D convolution, ReLU activation function, and a batch normalization function corresponding to N pixels of an input image. We set p 3x3 filters for all M components. Next, we obtain a response map $\{y_n = W_c x_n + b_c\}_{n=1}^{N}$ by applying a linear classifier, where $W_c \in \ R^{q*p}$ and $b_c \in \mathrm{R}^{\mathrm{q}}$, then normalizing the response map to $\{y_n'\}$ such that $\{y'_n\}_{n=1}^{N}$has zero mean and unit variance using batch normalization (whitening). Finally, we obtain the cluster label c_n for each pixel by selecting the dimension that has the maximum y_n' value using argmax classification.

1) Constraint on feature similarity:

The first consideration is that pixels with similar characteristics should be assigned the same label. As mentioned, cluster labels $\{c_n\}$ (obtained via argmax of normalized prediction map $\{y_n'\}$) are used as pseudo targets in cross entropy loss between $\{y_n'\}$ and $\{c_n\}$:

$$L_{sim}(\{y_n', c_n\}) = \sum_{n=1}^{N} \sum_{i=1}^{q} -\delta(i - c_n) \ln y_{n,i}'$$

Where:

$$\delta(t) = \begin{cases} 1 & if\ t = 0 \\ 0 & otherwise \end{cases}$$

The training objective of this loss is to strengthen the similarity of alike features when classifying image pixels, as feature vectors of pixels within the same cluster should be similar to one another and pixels of differing classification should have different feature vectors. By minimizing this loss, the network weights should better extract representative feature vectors for its classification of similar pixels.

2) Constraint on spatial continuity:

The second consideration is that spatially continuous pixels should be assigned the same label. To represent this constraint, we consider the L1-norm of horizontal and vertical differences in $\{y_n'\}$:

$$L_{con}(\{y_n'\}) = \sum_{w=1}^{W-1} \sum_{h=1}^{H-1} \left\| y_{w+1,h}' - y_{w,h}' \right\|_1 + \left\| y_{w,h+1}' - y_{w,h}' \right\|_1$$

Where W, H represent the width and height of the input image, and $y_{w,h}'$ represents the pixel value at (w,h) in response map $\{y_n'\}$. By minimizing this loss, the excessive consideration of labels with poor spatial continuity (due to complex patterns or inconsequential texture artifacts) can be reduced.

The combined loss function is:

$$L = \theta L_{sim} + \mu L_{con}$$

With θ, μ constants pertaining to the strength at which each constraint contributes to the overall loss and therefore the impact of each constraint on network training.

3) Constraint on number of labels

To determine how many segmentations should be generated in an image, the model must classify pixels into an arbitrary number q' of clusters where $(1 \leq q' \leq q)$. To prevent undersegmentation, we specify a preference for a large q' by using an *intra-axis* normalization process for the response map $\{y_n\}$ before assigning cluster labels via argmax. We use batch normalization $y'_{n,i} = \frac{y_{n,i}-\mu_i}{\sqrt{\sigma_i^2+\varepsilon}}$, where μ_i and σ_i respectively denote mean and standard deviation, and ε is a constant added to the variance for numerical stability. This gives each $y'_{n,q'}$ an equal chance to be the maximum value of y'_n across the axes of the cluster feature space. While this does not guarantee that every cluster label q' achieves the maximum value for any pixel, many cluster labels will. Thus, this batch normalization process gives the proposed system a preference for a large q'.

The network self-trains by backpropagation in two parts: prediction of cluster labels with fixed network parameters (forward process), and training of network parameters with the fixed predicted cluster labels (backward, based on gradient descent). Like in supervised learning, we calculate a softmax loss between the network responses $\{y'_n\}$ and the refined cluster labels $\{c'_n\}$. Then, we backpropagate error signals to update the parameters of convolutional filters $\{W_m, b_m\}_{m=1}^{M}$ and classifier $\{W_c, b_c\}$. In the paper, they use stochastic gradient descent with moment for updating the parameters. The forward-backward process is iterated T times to obtain the final prediction of cluster labels $\{c_n\}$. In contrast to supervised learning, the proposed CNN has a batch normalization layer between the final convolution layer and the argmax classification layer required to obtain reasonable labels $\{c_n\}$, resulting in multiple solutions of $\{c_n\}$ with different network parameters that achieve near zero loss. Dr. Kanezaki empirically determined that the learning rate of 0.1 and momentum of 0.9 yielded the best results [3].

Other models:

Superpixel model [3]:

An alternative method of addressing spatial continuity of pixel clustering ensure clustering of pixels is by creating superpixel clusters. To achieve this, we extract *K* superpixels $\{S_k\}_{k=1}^{K}$ (K=10,000) from the input image, where S_K denotes a set of pixel indices belonging to the *k*th superpixel. Then, we force all the pixels in each superpixel to have the same cluster label by selecting the most frequent cluster label c_{max} (letting $|c_n|_{n \in \mathrm{S_k}}$ be the number of pixels in S_k that belong to the c_nth cluster). This groups large numbers of spatially close pixels together and assigns them the same label, thereby forcing spatial continuity of cluster labels at each forward propagation iteration.

No batch model:

To evaluate the efficacy of response map normalization as part of the constraints on the number of labels, the proposed model was modified to have this final batch normalization step removed. In this model, each cluster label prediction will have no constraint to force them to all have equal chance to be the maximum value across the feature space, thereby providing no guarantees that each cluster label q will achieve the maximum value for any pixel. This may greatly reduce the number of labels considered in the final $\{c_n\}$ and be vulnerable to a trivial q=1 solution. Since the minimum number of labels is set to 3, the expectation is that many segmentations will result in only 3 labels.

Baseline model:

K-means clustering was used as a pseudo baseline. While it is somewhat ill-suited for this problem formulation, given that k-means requires that the number of labels is given before training, it is a common standard within the field of image segmentation. The number of labels (k) for the k-means model was set to 3 and 15 to capture a representative “under-segmentation” and “over-segmentation” performance threshold during model evaluation.

Model evaluation:

Given the nature of the problem formulation, model evaluation must be done using images with ground truth segmentations. Without access to ground truth segmentations for CT scans, the BSD500 dataset was used instead [6]. The test subset of images from BSD500 comprised of 200 images with several ground truth segmentations, roughly differentiated by the granularity and detail of segmentations (course segmentations had around 3 cluster labels while fine segmentations had upwards of 15). For each image in the dataset, the segmentation made by each model (proposed, superpixel, no batch, and k-means baseline) was evaluated against all ground truth segmentations using IoU (intersection over union) [7], choosing the best IoU score from the set of ground truth comparisons. The hyperparameters used in this test for the proposed, superpixel, and k-means models were chosen from the best experimental results from Dr. Kanezaki’s runs on the same data, with no batch inheriting the proposed model’s hyperparameters. K-means was split into two rounds, one with number of clusters set to 3 and another set to 15. Finally, a box and whisker plot was created to compare the IoU scores from each model [Fig. 1]. This test is mainly to evaluate the general applicability of models for the task of unsupervised image segmentation. When applied to the CT image modality, a different style of test must be performed to find appropriate hyperparameters and evaluate the result of the proposed model on the task of unsupervised segmentation of CT scans.

Contrast-enhanced CT scans from the DeepLesion database were used to find the best hyperparameters and evaluate the proposed model within the domain of CT scans [8]. First, the best number of convolutional components (M), feature similarity strength (θ), and spatial continuity strength (μ) was found by testing various values for each (M = [5, 6, 7, 8, 9, 10, 13], θ = [1, 1.5, 2, 2.5, 3, 4, 5], μ = [1, 1.5, 2, 2.5, 3, 4, 5]). Then, the best segmentation was chosen via visual inspection under the criteria imposed on the model design, namely visual coherence of segments (constraint on feature similarity, constraint on spatial continuity) and appropriateness of pixel groupings as contiguous segments (constraint on number of labels) [Fig. 2]. The hyperparameters from the best segmentation was used to evaluate a spatially correlated series of CT scans to evaluate the consistency of segmentation [Fig. 3].

Results:

The IoU model evaluation using the BSD500 dataset illustrated that the proposed model performs slightly better than the superpixel, no batch, and k-means (k = 3) models in the task of general unsupervised image segmentation [Fig. 1]. While k-means (k = 15) technically performed the best out of all models, this is likely a result of bias towards fine-grain segmentation ground truths in the dataset. As mentioned, k-means models require prior specification of the number of labels, so the results of k-means (k = 15) should be considered under scrutiny. However, k-means (k = 3) and no batch models are comparable in that they both resulted in 3 cluster labels. The no batch model converged to the minimum number of labels, 3, for each image, simulating the end result of specifying a target number of labels conducive to setting the number of labels in k-means. Furthermore, the no batch model still performed better than k-means (k = 3), showing the effectiveness of the convolutional architecture as a whole above the k-means baseline.

Despite the proposed model's success in general unsupervised image segmentation, the CT scan segmentation results were quite poor. It was difficult to ascertain the best quality segmentation of the contrast-enhanced CT image, as repeated trials under the same hyperparameters often result in different segmentations and segmentation quality. The best hyperparameters within the CT modality was determined to be 9 convolutional layers, feature similarity strength = 1.5, and spatial continuity strength = 1 [Fig. 2], values which were within the range of tested hyperparameter values. The majority of trials were able to isolate the lungs, but under the best hyperparameters the proposed model was moderately successful in separating muscle tissue from fat tissue and appropriately classified the air within the esophagus (the teal circle in the center of the image) as part of the non-body elements like the CT bed and clothes. However, the overall quality of this segmentation was poor – some examples include clustering soft organs with the bone, non-existent segmentation within contiguous segmented body parts, and poor segmentation continuity (solid shapes like the bones are jagged or incomplete). None of the parameters tested were able to segment the internal structures within the lungs.

The inconsistency of this model is demonstrated by the segmentations of the spatially correlated CT images. Despite the images being very similar, the model did not produce consistent segmentations across the series. Some were exceptionally poor, while others were better, but had similar issues as the previous test using one CT image. While it was able to accommodate for the introduction of internal body objects, namely the spherical object in the bottom row [Fig. 3], the overall segmentation quality and consistency has much to be desired.

Concluding remarks and extensions:

The proposed model had better results than the other models in general image segmentation, but fails to produce acceptable segmentation for CT scans. This highlights the inefficacy of unsupervised image segmentation using such a model alone without domain knowledge specific to CT imaging. Given the success of weakly-supervised medical image segmentation for tumor detection, the inclusion of ground truth segmentations could greatly improve CT segmentation techniques if applied to this problem formulation (under the domain of holistic classification anatomy). Additionally, other sources of domain knowledge could enhance the quality and specificity of segmentations, such as incorporating body position in the network or image preprocessing to remove the CT bed and clothes. While this negates the time and labour benefits of purely unsupervised image segmentation to CT imaging, this represents the inherent trade-off between supervised and unsupervised machine learning: accuracy and consistency of model results against the effort and expertise required to build it. Given the immense importance of domain knowledge and risk within medicine, extensions to this paper will most likely incorporate some supervised elements and domain knowledge to improve upon the weak performance of fully unsupervised segmentation.

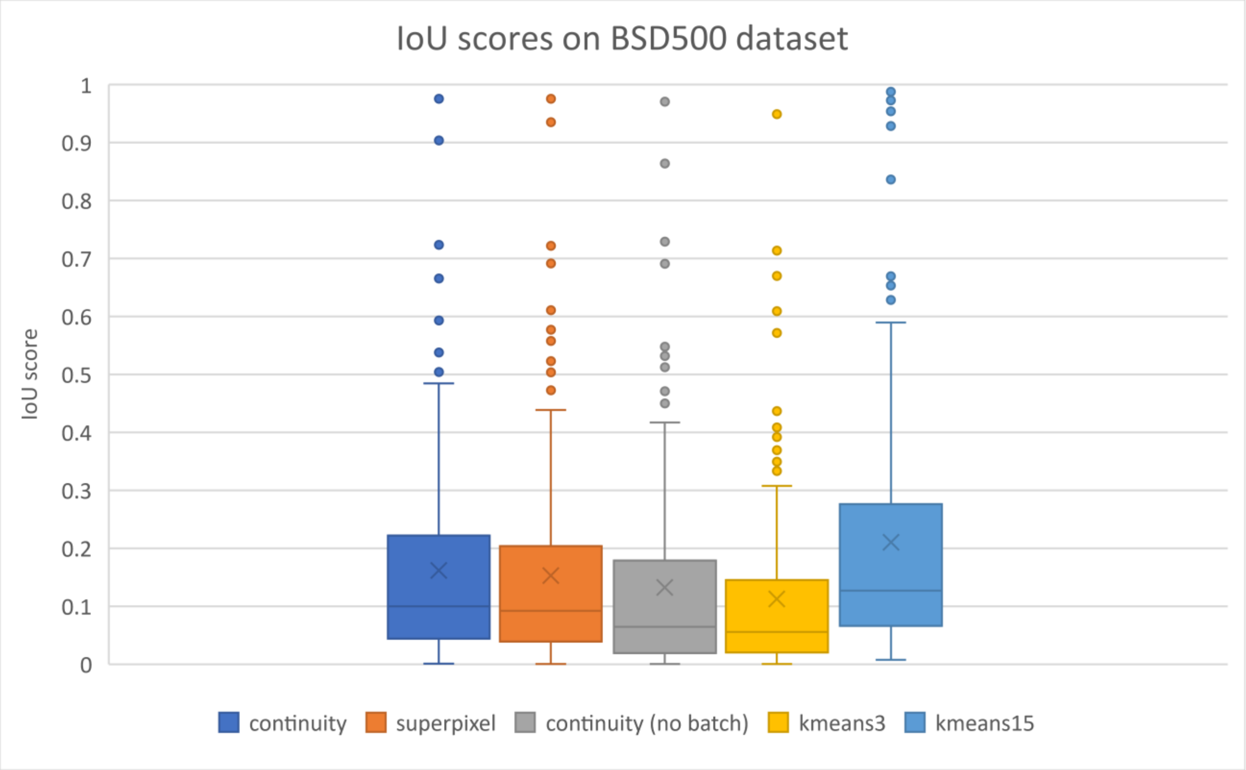


Fig 1: Box and Whisker plot of IoU scores for models' segmentations on the BSD500 test data. There were 200 images and their ground truth labels of various granularity (such as course or fine segmentation) provided for each image in the dataset. Image segmentation performed by each model was evaluated against all ground truth labels for a particular image, keeping the best IoU score as the evaluation. Each column represents the model's IoU performance over all images in the dataset, with "x" denoting the mean IoU score over all images; the center line denoting the median.

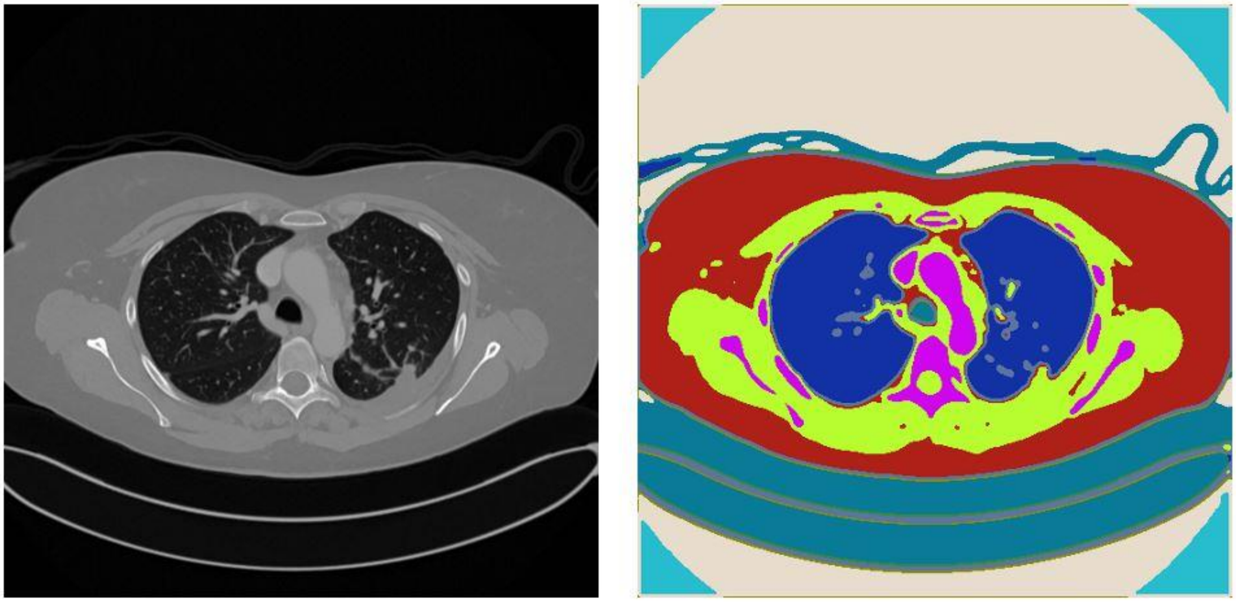


Fig 2: Result of proposed model in segmentation of contrast-enhanced CT image under the best performing hyperparameters (Number of convolutional layers = 9, feature similarity factor = 1.5, spatial continuity factor = 1). Supplemental figures 1, 2, and images in the code base show samples of segmentation results under different parameters.

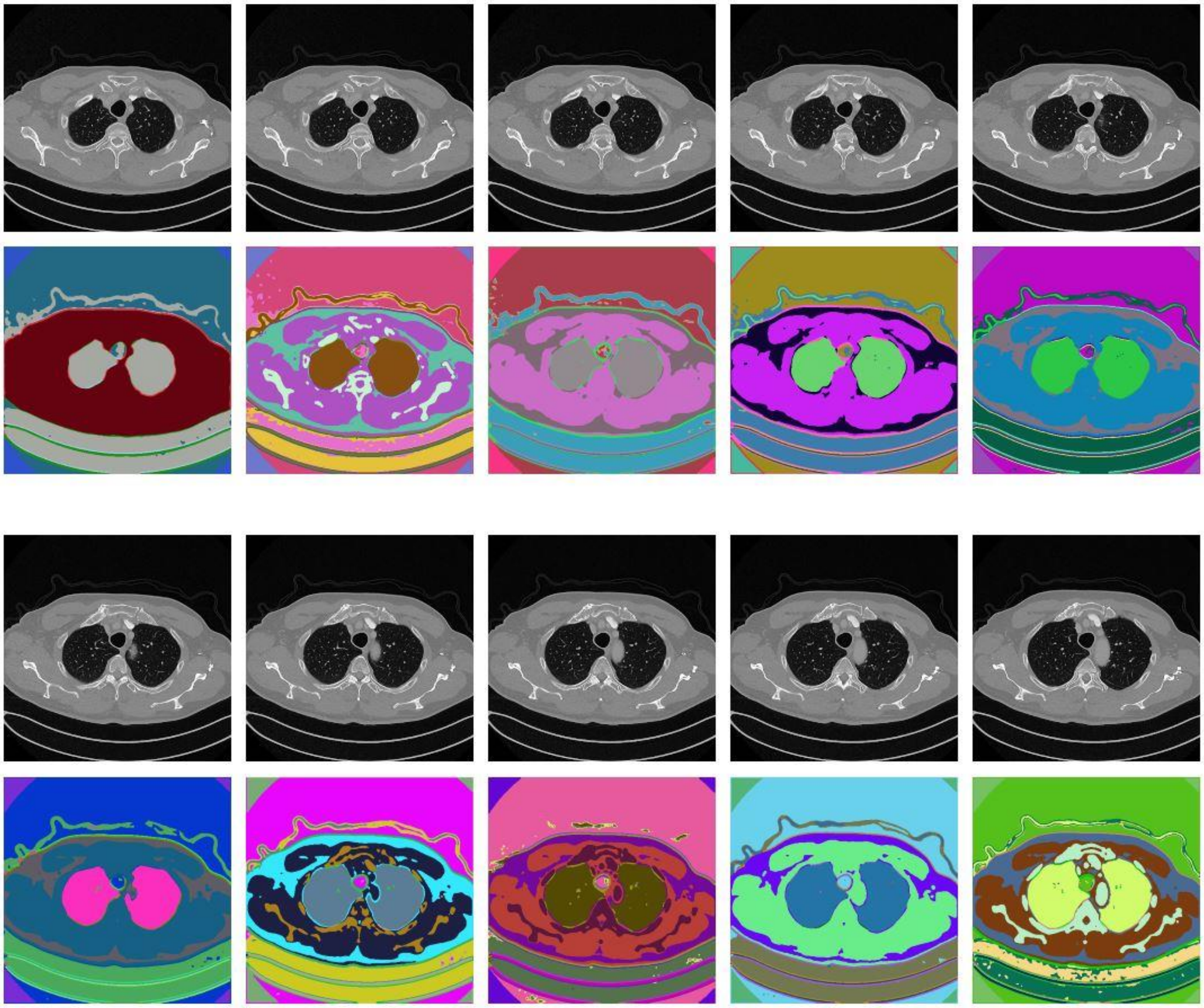


Fig 3: Image segmentations by proposed model on a series of spatially correlated, contrast-enhanced CT images. Top row (left-to-right) and third row (left-to-right) are CT scans along the axial line of one patient; images directly underneath those images are the segmentations performed by the proposed model (M = 9, θ = 1.5, μ = 1). Different segments are shown in different colors, with colors being randomized for each image.

References

[1] J. Tighe and S. Lazebnik, "Finding Things: Image Parsing with Regions and Per-Exemplar Detectors," 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3001-3008, doi: 10.1109/CVPR.2013.386.

[2] J. Cai, Y. Tang, L. Lu, A. P. Harrison, K. Yan, J. Xiao, L. Yang, and R. M. Summers, “Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3D mask generation from 2D RECIST,” Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, pp. 396–404, 2018.

[3] W. Kim, A. Kanezaki and M. Tanaka, "Unsupervised Learning of Image Segmentation Based on Differentiable Feature Clustering," in IEEE Transactions on Image Processing, vol. 29, pp. 8055-8068, 2020, doi: 10.1109/TIP.2020.3011269.

[4] K. Yan, L. Lu and R. M. Summers, "Unsupervised body part regression via spatially self-ordering convolutional neural networks," *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 1022-1025, doi: 10.1109/ISBI.2018.8363745.

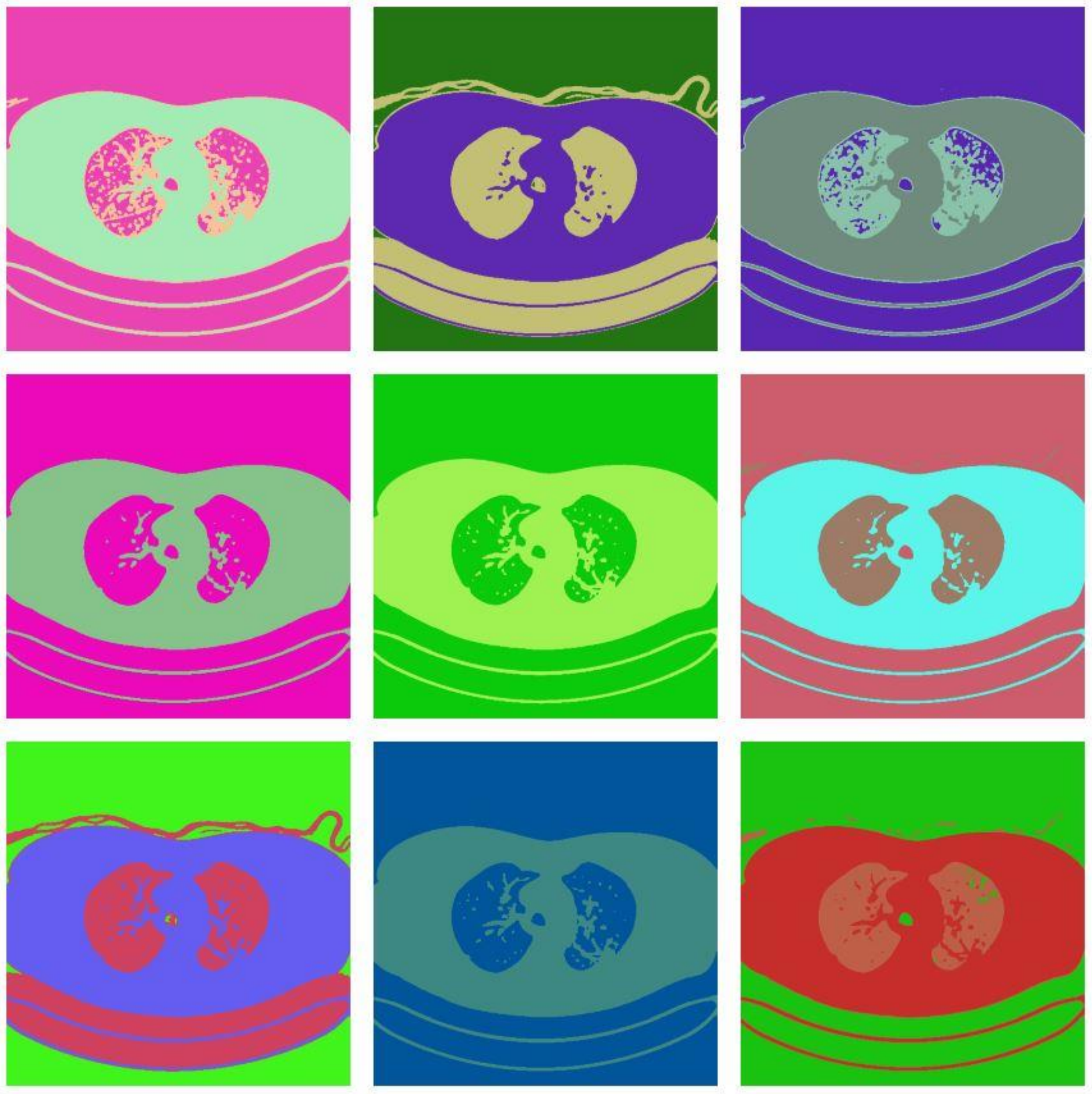
[5] A. Kanezaki, "Unsupervised Image Segmentation by Backpropagation," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 1543-1547, doi: 10.1109/ICASSP.2018.8462533.

[6] P. Arbeláez, M. Maire, C. Fowlkes and J. Malik, "Contour Detection and Hierarchical Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 5, pp. 898-916, May 2011, doi: 10.1109/TPAMI.2010.161.

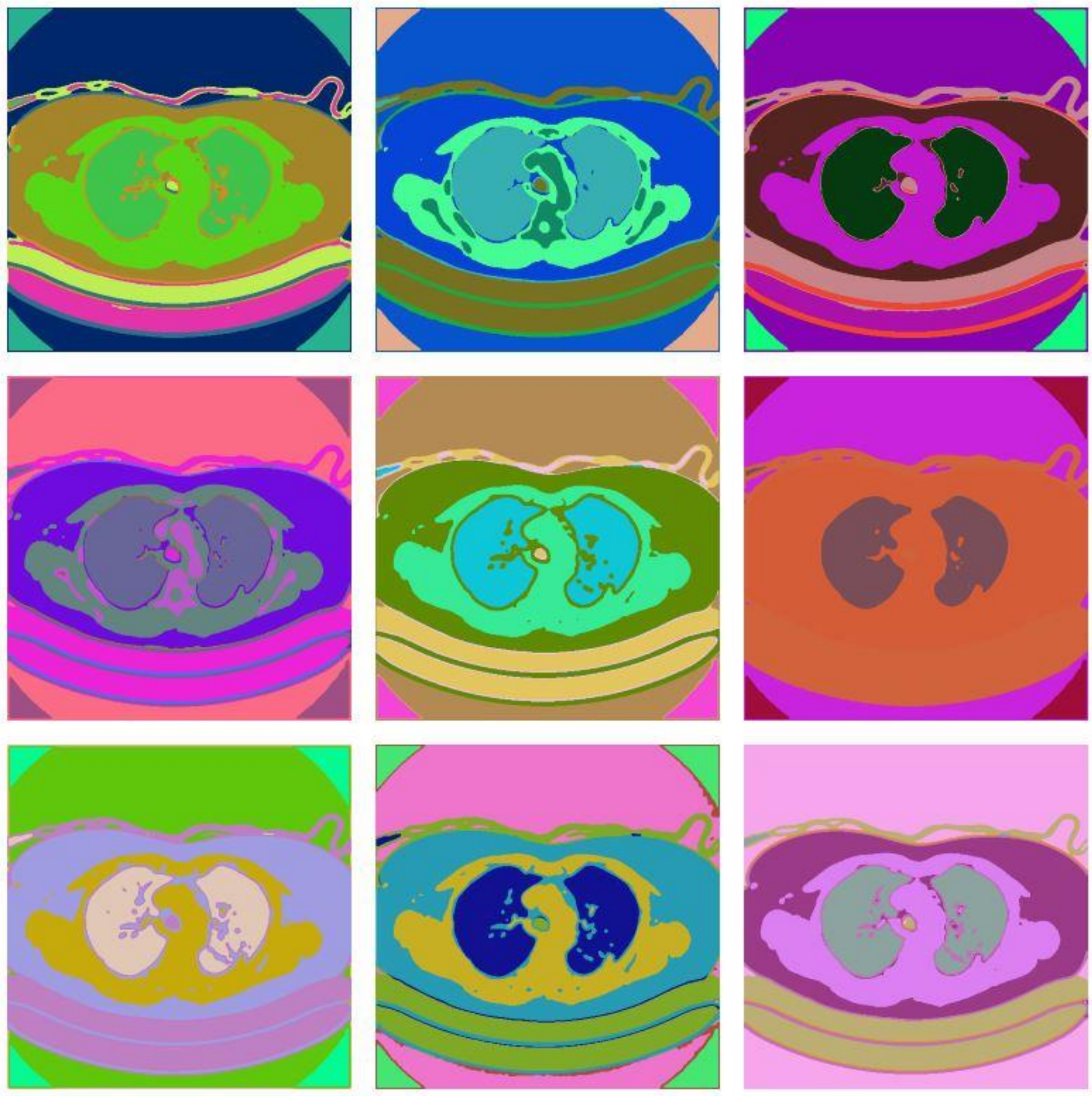
[7] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, “Optimizing the DICE score and Jaccard Index for Medical Image Segmentation: Theory and Practice,” Lecture Notes in Computer Science, pp. 92–100, Oct. 2019.

[8] K. Yan, X. Wang, L. Lu, and R. M. Summers, “Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning,” 2018 Journal of Medical Imaging, vol. 5, no. 03, pp. 1-11, doi: 10.1117/1.JMI.5.3.036501.

Supplemental figures:



Supplemental figure 1: Parameter selection test illustrating segmentation results of proposed model using 5 convolutional components. From left to right, spatial continuity μ = 1, 2, 5 respectively. From top to bottom, feature similarity θ = 1, 2, 5, respectively



Supplemental figure 2: Parameter selection test illustrating segmentation results of proposed model using 13 convolutional components. From left to right, spatial continuity μ = 1, 2, 5 respectively. From top to bottom, feature similarity θ = 1, 2, 5, respectively