BB-UNet: U-Net With Bounding Box Prior

Rosana El Jurdi, Caroline Petitjean, Paul Honeine, and Fahed Abdallah

***Abstract*—Medical image segmentation is the process of anatomically isolating organs for analysis and treatment. Leading works within this domain emerged with the well-known U-Net. Despite its success, recent works have shown the limitations of U-Net to conduct segmentation given image particularities such as noise, corruption or lack of contrast. Prior knowledge integration allows to overcome segmentation ambiguities. This paper introduces BB-UNet (Bounding Box U-Net), a deep learning model that integrates location as well as shape prior onto model training. The proposed model is inspired by U-Net and incorporates priors through a novel convolutional layer introduced at the level of skip connections. The proposed architecture helps in presenting attention kernels onto the neural training in order to guide the model on where to look for the organs. Moreover, it fine-tunes the encoder layers based on positional constraints. The proposed model is exploited within two main paradigms: as a solo model given a fully supervised framework and as an ancillary model, in a weakly supervised setting. In the current experiments, manual bounding boxes are fed at inference and as such BB-Unet is exploited in a semi-automatic setting; however, BB-Unet has the potential of being part of a fully automated process, if it relies on a preliminary step of object detection. To validate the performance of the proposed model, experiments are conducted on two public datasets: the SegTHOR dataset which focuses on the segmentation of thoracic organs at risk in computed tomography (CT) images, and the Cardiac dataset which is a mono-modal MRI dataset released as part of the Decathlon challenge and dedicated to segmentation of the left atrium. Results show that the proposed method outperforms state-of-the-art methods in fully supervised learning frameworks and registers relevant results given the weakly supervised domain.**

***Index Terms*—U-Net, shape prior, location prior, attention maps, weakly supervised segmentation, deep learning.**

I. Introduction

EVER since machine learning emerged as a leading tool for technological development, major breakthroughs have been achieved in various domains such as pattern recognition, natural language processing, classification and image segmentation. Semantic segmentation in image processing is the process of making per-pixel predictions with regards to every pixel in an image, through deriving meaningful segments, contour regions and boundaries. Since the process involves indicating not only what is present in an image but also where, semantic segmentation considers a trade-off between contextual and spatial understanding. First approaches to segmentation with deep learning emerged with the Fully Convolutional Networks (FCNs) [1]. FCNs are structures derived from typical deep models such as VGG16, AlexNet or GoogLeNet by removing the corresponding classification layers, replacing their fully connected layers with convolutional ones and adding an up-sampling layer that is dedicated to transforming coarse outputs into dense predictions. Despite their good performance, FCNs fail to consider global and spatial information, and often result in fuzzy coarse-grained predictions [2]. A pioneering approach is the U-Net model [3], especially popular in medical imaging. U-Net has a symmetric encoder/decoder structure with skip connections. The encoder part is a contracting path composed of stacked convolutional and max pooling layers, whereas the decoder part is an expanding path composed of de-convolutional or bilinear upsampling layers. Layers within the encoder are dedicated to capturing contextual information in order to detect objects/classes present in an image. The decoder layers, on the other hand, help precise localization of patterns including contours and boundaries. As an image moves further into the contracting layers, it decreases in size but increases in depth of its learnt contextual features. In contrast, the decoder layers increase its input size but decrease its depth, thus retaining the model's localization ability. To make use of both contextual and positional features, skip connections between the downsampling (encoder) and upsampling (decoder) paths are utilized. Skip connections concatenate symmetrical features from opposing convolution and de-convolution layers. Through end-to-end training, the U-Net takes on as input an image of any size and produces a segmentation map of similar dimensions. Thus, due to these enhanced properties, U-Net gained a high level of success and has been applied in various segmentation tasks [4], [5].

Taking advantage of U-Net's success, multiple variants emerged in order to increase model performance given different tasks [6], [7]. Despite good performance, such networks often require large amounts of annotated training data, which is not easy to obtain given particular domains such as the medical one. Rather, unannotated or partially labeled data are more easily available or less computationally expensive to obtain. For this reason, recent approaches within the machine learning domain aim to make use of these "not-so-accurate" labels in order to derive proper segmentation masks, thus embracing the weakly supervised learning paradigm.

Manuscript received December 10, 2019; revised April 16, 2020 and June 4, 2020; accepted June 4, 2020. Date of publication June 10, 2020; date of current version September 24, 2020. This work was supported in part by the DAISI project, in part by the European Union with the European Regional Development Fund (ERDF), and in part by the Normandy Region. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Vishal Monga. *(Corresponding author: Rosana El Jurdi.)*

Rosana El Jurdi is with the Normandie Université, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France, and also with the Lebanese University, Beirut, Lebanon (e-mail: rosana.el-jurdi@univ-rouen.fr).

Caroline Petitjean and Paul Honeine are with the Normandie Université, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France (e-mail: caroline.petitjean@univ-rouen.fr; paul.honeine@univ-rouen.fr).

Fahed Abdallah is with the Lebanese University, Beirut, Lebanon, and also with the University of Technology of Troyes, 10300 Troyes, France (e-mail: fahed.abdallah76@gmail.com).

Digital Object Identifier 10.1109/JSTSP.2020.3001502

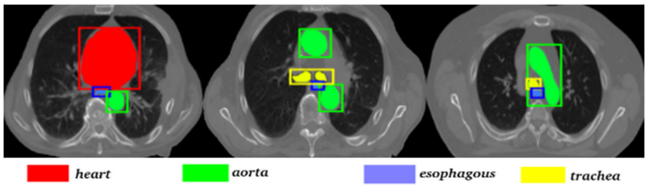

Fig. 1. CT images from the SegTHOR dataset with manual segmentation and bounding boxes overlaid on multiple organs.

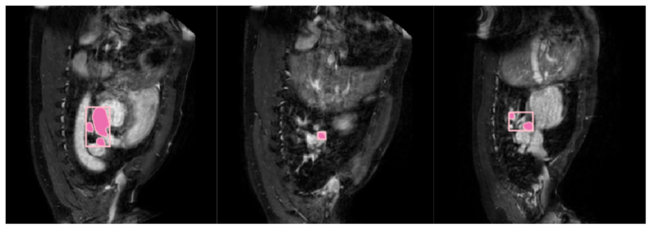

Fig. 2. MR images from the Cardiac dataset with manual segmentation and bounding boxes segmentation overlaid on the left atrium.

Within this framework, dataset labels may be incomplete, inexact or inaccurate. Weak labels can come in different forms such as bounding boxes encompassing the understudied organ [8], [9], image tags [10], seeds generated from object center of mass [11], randomly or by erosion [12]. The objective of weakly supervised segmentation models is to make use of these coarse-grained annotations to derive proper and accurate predictions at the pixel level. Weakly supervised image segmentation can be conducted in two different ways: (i) a two-step iterative approach where initial label estimates are generated from weak labels in the first step, and fine-tuned through a deep learning model in the second step [8], [9]; and (ii) through direct modification of the network (e.g. insertion of customized segmentation layers) to take into account weak labels [10], [13].

In medical imaging, organ segmentation comes with particular challenges, such as low contrast and high noise levels. Given this, recent works aim to exploit anatomical priors with regards to organ shape and position [2], [11], [14].

Generally, anatomical priors refer to expert knowledge and domain expertise that capture spatial as well as location guidelines with respect to the understudied organs. The basic motivation behind the prior-based approaches is that a convolutional neural network (CNN) might suffer from difficulties in differentiating two distinct objects that are consistently in two specific parts of the scan, if they have the same intensity and context.

In this paper, we propose a new model inspired by U-Net that integrates prior information in-between local and global features. We call the proposed model BB-UNet (Bounding Box U-Net) since it uses weak labels and bounding filters to guide the training process onto convergence. The proposed model allows to take advantage of positional and shape features as means of guiding the neural network to find consistent organ contours. We exploit this model within two strategies: one, a fully supervised semantic segmentation strategy that utilizes prior information to overcome noise and low contrast; two, a weakly supervised strategy where training of BB-UNet is initially conducted on a very tiny sample of the dataset. The learnt weights are then used to generate initial label estimates for a much larger weakly supervised dataset.

The proposed model is validated on two segmentation problems. First, we consider a multi-label segmentation problem in computed tomography (CT) imaging. Experiments are conducted on the SegTHOR dataset, which consists of CT images of patients suffering from non-small cell lung cancer. The four organs of interest, which are the heart, aorta, esophagus and trachea, have variable shapes and share same gray-scale intensity values with neighboring tissues as shown in Fig. 1. As a result, it is hard to identify the organs from neighboring tissues or properly separate them. Secondly, a single-label segmentation problem having a multi-component complex organ in magnetic resonance imaging (MRI) scans. Experiments are conducted on the Cardiac dataset, which consists of MR images covering the entire atrium (see Fig. 2). The understudied organ within this dataset is characterized with large variability. Aside from the large range of organ size that varies across slices, the organ is also characterized with multiple components within the same slice that are in close proximity of each other and of different sizes.

The main contributions of this paper are as follows:

1) we propose a novel deep learning model that integrates location and shape constraints into the network architecture in order to overcome segmentation ambiguities;
2) we show that the proposed approach achieves excellent performance in comparison with the state of the art when trained in a fully supervised framework;
3) we incorporate the proposed novel model in a semi-weakly supervised framework where only bounding box tags are present and achieve comparable results with respect to the fully supervised models;
4) we shed light on the role of embedding prior knowledge onto model training relative to data augmentation and post-processing alternatives.

The rest of the paper is organized as follows. Section II provides a brief overview of the state of the art in fully as well as weakly supervised learning for image segmentation. Section III elaborates on the proposed BB-UNet model as well as the multiple frameworks and paradigms explored. Section IV presents the datasets as well as the evaluation metrics and experiment setting. Section V reveals model performance within experiments where all labels are present. Section VI evaluates the robustness of BB-UNet performances across bounding box size variations and dataset distributions. Section VII elaborates on BB-UNet performance using the provided prior information as weak labels. Finally, Section VIII concludes with future works and perspectives.

II. Related Work

A. Segmentation Under Full Supervision With Prior Knowledge

Primary work on segmentation within deep learning emerged with the fully convolutional neural networks (FCNs) [1]. Since then, FCNs and its variants were applied to many segmentation tasks including non-medical [15], [16] and medical ones [17],

[18]. In 2015, U-Net architecture emerged as a powerful structure not only for medical image segmentation but also for natural image segmentation, regression, face alignment and recognition [3]. The main reason that U-Net has registered such a success is its ability to process the image as a whole, thus incorporating global features rather than just local ones. This is mainly due to its symmetrical properties and equivalent distribution of convolutional and de-convolutional layers, as well as skip connections.

Despite this breakthrough, U-Net and its variants are limited in incorporating domain expertise such as location information and explicit anatomical priors [14]. For example, a U-Net may have some difficulties in differentiating two distinct organs that are consistently lying in two specific parts of the scan but are characterized with same intensity values [14]. For this reason, U-Net performance can be further enhanced given methods that exploit prior knowledge such as shape or position.

Segmentation approaches based on CNN integrate prior information either through topological [19] or shape-based loss functions [20] or through adding regularization techniques that conforms predicted shapes with a set of allowed ones [2]. For instance, in [2], a non-linear shape regularization model is trained jointly along U-Net. The main function of their adjoint network is to learn projections of arbitrary shapes onto a manifold space. It then incorporates a loss function that updates the segmentation network (U-Net) parameters based on the regularized predicted segments, the rough predicted segments as well as the ground-truth labels. The authors of [21] adopt a similar regularization approach to that in [2]. However, they target the decoder layer with their U-Net-like structure and train the up-sampling layers through super resolution ground-truth maps.

Another approach to incorporate shape prior is through the use of loss functions that update model training. Both [22] and [2] adopt this approach in order to update their U-Net model parameters. However, whereas the former aims to minimize the Euclidean distance between the predicted and the ground-truth shape, the latter aims, through the loss function, to drive U-Net predictions to be as close as possible to the shape manifold representing allowed shapes while still preserving the variations between the actual ground truth shapes and the learnt shape space. Similar to [2], the authors in [23] demonstrate a manifold of permissible nuclei shapes prepared by a domain expert and incorporate this prior information in the form of a regularizing term that encourages detection inside nuclei boundary while simultaneously penalizing false positives.

Methods in [24] and [11] extend upon U-Net by proposing a novel structure that learns good features for predicting proper segmentation masks of their understudied organ by properly computing organ center of mass from intermediate U-Net-like layers. In [24], a regression model is introduced at the bottleneck level of its U-Net-like structure in order to extract the center of mass corresponding to their understudied organ. The extracted feature map is then merged with that of the decoder layer, then segmentation maps are derived. To avoid anatomically impossible shapes, the authors of [24] extend upon their work to further estimate a probability distribution from the training data with regards to the occurrence probability of the understudied organ. This predefined shape prior is further concatenated with the center of mass feature map and the decoder output [11].

Whereas U-Net was firstly dedicated to medical images, multiple U-Net variants emerged in order to increase model performance in applications that are non-related to the medical field [16], [25]. However, whether for medical or non-medical purposes, such networks usually require large amounts of annotated training data in order to gain their generalization ability, which is not often available. For this reason, prior knowledge such as bounding box or image tags can be considered as a case of weak labels, thus casting the medical image segmentation problem onto the weakly supervised learning domain. In the following, a review of some recently proposed work that take into account these weak labels is presented.

B. Segmentation Under Weak Supervision

Among weakly supervised segmentation methods, one can identify two main approaches: those based on a two-step iterative process that mimics full supervision, and those based on classification model training with modified upper layers [10], [12]. Despite their different concepts, the two approaches use weak labels to derive accurate segmentation maps.

Weakly supervised segmentation through a two-step iterative process that mimics full supervision is a common approach that synthesizes full pixel-level labeled training masks from the available weak labels. Typically, such proposal based techniques iterate two steps: label estimate generation (the proposals) and fully supervised CNN training. Weak labels may be of different types: bounding box labels, image labels, or a mixture of bounding box and image labels.

Image Labels: The EM-Adapt model in [13] implements this two-step iterative process in an Expectation-Maximization (EM) framework where pixel-level annotations are considered as latent variables to be estimated from known image-level labels (E-step). The method then updates the neural network parameters through stochastic gradient descent and a probability distribution that incorporates a pixel distribution and an adaptive bias (M-step). As explained in [12], the EM-Adapt method is generally limited when it comes to leveraging the full power of weak labels. The authors further explain that this problem is generally non-convex and requires Lagrangian dual optimizations which is computationally very expensive. Their proposed method finds a way around the dual Lagrangian optimization by integrating the constraint at network output level. Instead of the EM approach, the authors cast the segmentation task onto a constraint optimization problem where the parameters of the CNN network are found given particular constraint Q with respect to the weak labels.

Bounding Box + Image Labels: In [26], the EM-Adapt method is extended through proposing a good initialization approach of the EM algorithm, with the goal of avoiding local maxima. Thus, instead of an initialization based on a classification task as is usually done, method in [26] focuses on exploiting a combination of saliency and attention maps to kick-start the algorithm.

Bounding Box Labels: Both BoxSup [9] and Simple Does It (SDI) [8] use an iterative training approach to gradually improve generated label estimates. However, whereas SDI exploits a GrabCut-like algorithm [27] for the initial label estimate generation, BoxSup exploits an unsupervised region proposal method such as Multiscale Combinatorial Grouping (MCG) [28]. Moreover, whereas BoxSup modifies the training procedure in order to denoise intermediate outputs, SDI leaves the training algorithm unmodified and focuses on externally denoising input labels through exploiting prior knowledge.

III. Proposed BB-UNet Structure

In this section, we present the proposed BB-UNet model. We first elaborate on its different building blocks and then clarify the different prior information used. In Section III-E, we focus on the BB-UNet principle and training strategies. Finally, a comparison is presented in Section III-F between the proposed BB-UNet and similar state-of-the-art models.

A. BB-UNet Architecture

In the proposed architecture, we extend U-Net to include not just global and local features, but also ones related to position as well as shape priors. As previously stated, U-Net is a symmetric encoder/decoder structure with equivalent distribution of convolutional and de-convolutional layers. In order to make use of local and global information, U-Net utilizes skip connections that concatenate down-sampling features from the contracting path with up-sampling ones from the expanding path. Our main contribution within the BB-UNet structure lies at skip connection levels. Thus, instead of directly concatenating the features from both paths as in U-Net, a third component layer that takes into consideration shape and location information is introduced. This layer is called the BB-ConV layer as shown in Fig. 3. A BB-ConV layer is composed of a 2D-max pooling layer followed by two consecutive convolutional ones. The input to this layer is a bounding map (bounding filter) representing a coarse-grained area where the organs are supposedly located. In summary, the BB-ConV layer takes as input a bounding filter and outputs a feature map that allows the network to enhance its estimation to where an organ can be. This learnt feature map can be considered as a per-pixel weighting factor, enhancing discriminative features over non-significant ones. Overall, BB-UNet is taking two inputs, the CT-image and the bounding filter. Since the filters are inserted at the link between the contextual and location information, we are able to adapt what the model learns, focusing on the attention areas that we are yielding, i.e., enhancing features detection within particular sections of the image.

B. BB-UNet Main Principle

The proposed BB-UNet considers two inputs, the CT image and the bounding filter. Whereas the CT image is fed to the encoder layers in the contracting path for contextual feature extraction as is done within a regular U-Net, the bounding filter is fed independently to the BB-ConV layer for shape and location feature extraction. Within each skip connection, the intersection between the unpooled map from a level contracting layer and the location feature map from the BB-ConV layer is then obtained, and further concatenated with the features from the up-sampling layers. The bounding filter provided to the BB-ConV layer is a binary map indicating the attention area corresponding to the position of the organ(s) under consideration. For single-organ segmentation, a single channel indicating the possible area where the organ may be located is provided to the BB-ConV layer. For the multi-organ segmentation, filters of the different organs are independently fed to the BB-ConV layer in the form of a multi-channel binary map. This multi-channel map is then convolved within the BB-ConV layers for feature extraction. The output of BB-UNet is a segmentation mask derived from learnt relations between the bounding filters as well as the image.

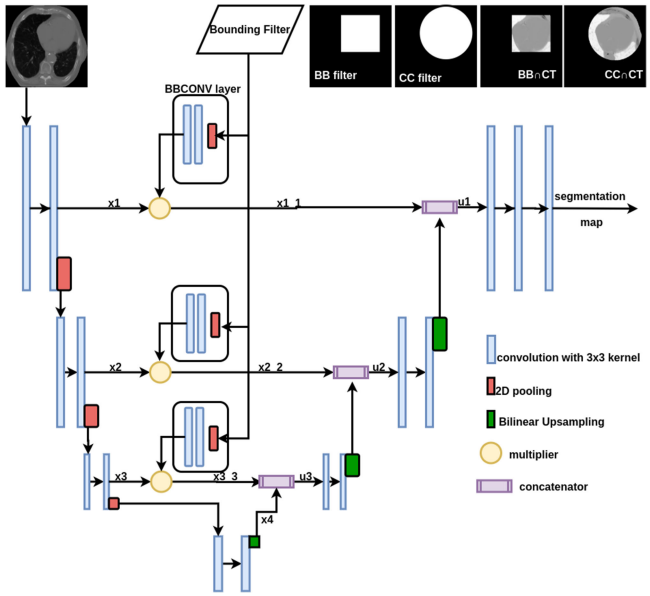

Fig. 3. BB-UNet structure with a bounding filter inserted at the BB-ConV layers. Four possible bounding filters are illustrated: BB: Bounding Box filter, CC: circular filters, BB∩CT: intersection between bounding box and image, CC∩CT: intersection between circular filter and image.

C. Generating Bounding Boxes

Bounding boxes can be generated through 2D or 3D approaches. In 2D, bounding boxes can be generated either through region proposal approaches [29] or through regression/classification based approaches [30]. 3D-bounding box generation can be done by training an end-to-end convolutional network for 3D-object extraction [31], or by extrapolation from 2D bounding box generation techniques as surveyed in [32]. In our implementation, we have considered the case where bounding boxes were generated automatically from the ground truth and defined as the smallest bounding box encompassing the understudied organ with interval ϵ. Despite the fact that we have used manually-obtained bounding boxes for both our training and inference tasks, automatically obtaining coarse grained bounding areas could be considered given current object

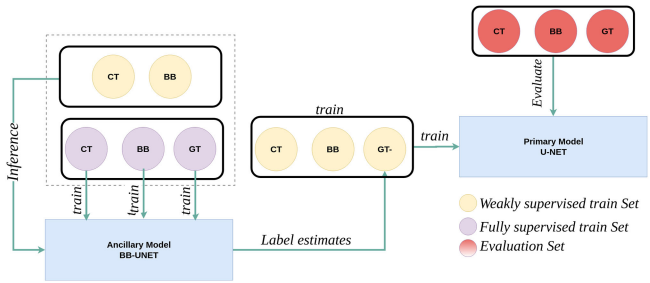

Fig. 4. Weakly supervised segmentation process with BB-UNet. BB: Bounding Box, CT: CT image. GT: Ground Truth segmentations (full annotation).

detection advances in medical imaging [33]. In this case, BB-UNet performance will depend on the shift between detected vs ground-truth bounding box distributions. Ideally in the automatic framework, detected bounding boxes would have to be used for training and testing.

D. Prior Information

We have designed several types of bounding filters. We have considered rectangular bounding boxes denoted as **BB** for bounding box filter.

Noting that the objects of interest (organs) do not have corners or edges, we also considered a circular filter, denoted as **CC**. CC filters are the smallest circles encompassing the bounding boxes. Since organs may share similar intensity values, filters that include the intersection between both the bounding mask as well as the CT image are investigated. Henceforth, the **BB∩CT** and **CC∩CT** filters are introduced in Fig. 3. The proposed model is trained in such a way that it extracts features specific only to the organ understudy with respect to the other relative organs or tissues. This can include shape, size, or organ smoothness.

E. Supervision Strategies

In order to study how well the proposed model performs as a standalone structure, the BB-UNet is firstly trained within a fully supervised framework. Labels used within this framework are the ground truth segmentations provided for each organ respectively. Given this framework, separate channels are fed to the BB-ConV layer relative to each organ independently. In the case of multi-component organs, a unified bounding box encompassing all organ elements is taken into consideration. With respect to BB-UNet output, we consider five classes thus distinguishing between the 4 organs as well as the background. The BB-UNet may also be implemented within a weakly supervised framework. We adopt the iterative method as done in our previous work [34]; however, instead of using the GrabCut algorithm to generate initial label estimates, we focus on training the BB-UNet on a very tiny sample of the training set (ancillary set) in a fully supervised manner. We then used the BB-UNet weights in order to derive proper label estimates for a much larger weakly supervised dataset - the Primary training set. A U-Net model is then trained on the label estimates provided for the Primary-train set. The process is described in Fig. 4.

F. Comparison to State-of-the-Art Models

Fully Supervised Framework: Models most relevant to our work are those developed in [2] and [35]. The SR-UNet in [2] introduces regularization factors by jointly adding an external network to the U-Net model. The main objective of this network is to take into consideration the incomplete, over- or under-segmented shape masks provided by the U-Net and map it to a manifold of training shapes. Despite the importance of regularization, the addition of an adjoint complex model while training will increase model complexity and thus affect model performance. For this reason, we aim to add the regularization structure within the U-Net level, thus fine-tuning the encoder spatial considerations intern without the need for a manifold space.

The Attention-UNet in [35] is most similar to ours in terms of adding attention blocks at the skip-connection level. Thus, both BB-UNet and Attention-UNet focus on imposing convolutional filters midway between encoder and decoder paths. In doing so, both models are able to distinguish between relevant and irrelevant features while training. However, the means to which each model obtains the constraints to these attention maps differ considerably. Whereas [35] aims at exploiting coarse-grained features obtained from U-Net bottleneck as input constraints to the convolutional layers at skip connection levels, our model imposes external activation inputs based on prior knowledge of the dataset. One can think of both models as functioning in different directions within the skip connections. Whereas we impose external activation inputs to guide the network on to where to look and move downwards through the network until bottleneck, [35] exploits inputs provided by the bottleneck output and moving upward through the skip connections.

Weakly Supervised Framework: Models most relevant to our work are those developed in [36] and [8]. The authors of [36] elaborate on a simple to complex (STC) framework, where an initial deep CNN is learnt on simple images and their corresponding saliency maps. An enhanced CNN is learnt on the output of the initial ancillary model as well as the image label. Our work shares similarity with STC in that both methods train a primary U-Net based on the predictions of an ancillary model, which is the BB-UNet in our work. However, the BB-UNet does not use saliency maps or simple data, rather our intuition lies in the idea of developing a robust model that can make full use of the information given a tiny subset of dataset that is fully supervised, so as to derive good initial label estimates for the much larger weakly supervised framework. Whereas STC's main contribution is the use of simple images to infer labels for a much larger weakly supervised dataset, our contribution is to make use of a very small amount of data in order to perform the aforementioned task.

As in SDI [8], we aim to generate good initial label estimates within just one generation step. However, SDI uses $M \cap G+$, which is an intensity-based estimator representing the intersection between multi-scale combinatorial grouping segment proposals [28] and GrabCut [27], as an initial label estimator. We have shown in our previous work [34] that the use of intensity-based algorithms such as $M \cap G$ for example, does not provide

TABLE I
SLICE AND SIZE DISTRIBUTION OF THE SEGTHOR DATASET

	Train	Train Ancillary	Train Primary	Validation	Evaluation	Organ Size (pixels)		
Patients	36	6	30	4	20	Average	Min	Max
Heart	1444	219	1225	155	726	9574	245	23588
Aorta	3363	554	2809	391	1824	1023	81	6336
Trachea	1767	293	1474	220	953	340	72	1244
Esophagus	3510	565	2945	410	1862	226	60	2528
Total	4153	699	3454	497	2281			

proper label representatives. Instead, we aim to generate suitable initial label segments by exploiting a network (ancillary model) trained on a very tiny sample of a fully supervised dataset.

IV. EXPERIMENTAL SETTING

SegTHOR dataset: This dataset consists of 60 CT scans of patients characterized with non-small cell lung cancer and referred for radiotherapy. The dataset was acquired at the cancer center Centre Henri Becquerel in Rouen, France. Organs at risk in CT images including the heart, trachea, aorta and esophagus, were manually segmented by an expert radiotherapist. CT images are 512×512 voxels, and number of slices ranging from 150 to 284 per patient. The dataset was released publicly in a competition conducted at the IEEE International Symposium for Biomedical Imaging 2019.[1]

In these images, some organs share similar gray-scale intensity values with each other as well as neighboring tissues, which makes the segmentation particularly challenging. This phenomenon renders common intensity-based thresholding methods useless and justifies the need for learning-based techniques for segmentation. Moreover, target organs are very close. As a result, bounding filters suffer from a high overlap between neighboring organs (see Fig. 1), which may enhance organ imbalance thus affecting model performance, as is later shown.

Following the SegTHOR challenge, the training considered 40 patients subdivided into a training set of 36 patients and a validation set of 4 patients. The test set includes the remaining 20 patients. Given a weakly supervised framework, further divisions are made as stated in VII. The patient and slice distributions are shown in Table I. The average, minimum and maximum organ size is also provided, over the entire set of patients.

Taking a closer look at the organ slice frequencies in Table I, one can notice that the slice organ distribution suffers from high class imbalance. Thus, the heart as well as the trachea have a small number of slices ($\approx$1000 slices) compared to the aorta and esophagus ($\approx$3000 slices). Taking a closer look at the organ size, the trachea and esophagus are the smallest in size (200 to 340 pixels in average) relative to the heart and aorta ($\approx$10000 pixels in average).

Using body contours provided for each CT slice, images were cropped and resized to a resolution of 512×512 pixels. Image intensities were bound to values between -1000 and 3000, normalized by subtraction of the mean and division by standard deviation at image level. Slices were filtered to keep only images with at least one organ present.

[1]The SegTHOR dataset is available at https://competitions.codalab.org/competitions/21145.

Cardiac Dataset: The Cardiac dataset is part of the Decathlon medical image segmentation challenge [39]. It consists of 20 mono-modal MRI scans covering the entire atrium which were segmented through an automated tool followed by manual correction. MR images are 320×320 voxels and a number of slices ranging between 54 and 76 slice per patient. The dataset was split into 10 patients (670 ± 5 slices) for training, 4 patients (66 ± 1 slices) for validation and 6 patients (416 slices) for testing. The dataset is characterized, as stated in the challenge, by being small, 1351 slices in total, with large variability. Thus, the atrium has a large size range that varies from 3 to 1921 pixels with up to 3 connected components.

Aside from the low contrast that these images are characterized with, the segmentation process is particularly challenging due to many factors, among which the high organ size imbalance over the different slices. Thus, whereas some slices have a considerably large atrium, others contain a segment which is very small. Moreover, the atrium is generally composed of multiple components within the bounding box. Therefore, the model must learn to distinguish between the different parts of the atrium present in the same bounding box.

Model Training and Architecture: The BB-UNet has two main components: the Base U-Net model and the BB-ConV layer. The U-Net implementation in this work is the one provided by a PyTorch implementation of the original U-Net [40]. Feature dimensions extend till 256 feature maps within the bottleneck which is composed of 2 simple convolutional layers. The BB-ConV is composed of a 2D-pooling layer followed by two consecutive convolutions with batch normalization (momentum $= 0.1$) and dropout (factor $= 0.4$). Bounding boxes relative to different organs are fed independently through a multi-channel input onto the BB-ConV layer. Moreover, to overcome the size imbalance between organs versus the background, we distinguish between classes corresponding to the understudied organs (4 organs for SegTHOR, 1 organ for Cardiac) and the background class. To guide the training, a loss approximation of the Dice similarity factor as elaborated in [6] was adopted. Moreover, we have used the Adam optimizer with an initial learning rate of 10^{-3} and a cosign annealing scheduler. Network diagram is presented in Fig. 3.

V. FULLY SUPERVISED SEGMENTATION EXPERIMENTS

In this section, we present results for both SegTHOR and Cardiac datasets. SegTHOR results are compared relative to 3 fully supervised segmentation models: the original U-Net with and without data augmentation [3], and VB-Net, the winner of the ISBI SegTHOR challenge [37]. With regards to Cardiac

TABLE II
AVERAGE DICE RATIO FOR FULLY SUPERVISED MULTI-ORGAN SEGMENTATION. FIRST ROWS REPRESENT NON-FILTERED MODEL PERFORMANCE AND LAST ROWS SHOWS RESULTS AFTER POST PROCESSING USING BOUNDING BOXES

	Heart	Aorta	Trachea	Esophagus
State of the Art				
VB-Net [37]	0.94	0.93	0.91	0.84
U-Net (+ data augmentation) [38]	0.93	0.92	0.86	0.81
U-Net (+ background)	50.37 ± 15.28	85.172 ± 2.21	82.48 ± 5.27	76.56 ± 0.10
U-Net (+ background) + Post.	96.89 ± 1.38	93.20 ± 0.62	97.59 ± 0.24	84.761 ± 1.09
Proposed Models				
BB-UNet-BB	**98.32 ± 0.29**	**96.02 ± 0.46**	**97.82 ± 0.10**	**91.56 ± 0.12**
BB-UNet-BB + Post.	98.59 ± 0.10	96.02 ± 0.46	97.82 ± 0.10	91.56 ± 0.12
BB-UNet-BB ∩ CT	97.57 ± 1.52	95.95 ± 0.31	**97.82 ± 0.34**	**91.74 ± 0.17**
BB-UNet-BB ∩ CT + Post.	98.63 ± 0.10	95.95 ± 0.31	97.97 ± 0.30	91.74 ± 0.17
BB-UNet-CC	93.07 ± 5.8	95.54 ± 0.10	93.526 ± 0.17	90.01 ± 0.28
BB-UNet-CC + Post.	98.07 ± 0.27	95.77 ± 0.23	97.74 ± 0.25	90.29 ± 0.31
BB-UNet-CC ∩CT	82.89 ± 4.8	95.30 ± 0.25	93.41 ± 0.24	89.79 ± 0.34
BB-UNet-CC ∩ CT + Post.	98.1 ± 0.06	95.54 ± 0.26	96.65 ± 0.14	90.07 ± 0.37

dataset, we compare to regular U-Net. In addition, we compare for both datasets U-Net performance after post-processing by filtering predicted segments with bounding box prior (U-Net + Post.). The proposed models are evaluated with the Dice similarity index, defined as twice the intersection divided by the union between ground-truth and predicted segments [41].

A. *SegTHOR Multi-Organ Segmentation*

From the results in Table II, we observe that the proposed model outperforms regular U-Net (3 rd row) by about 15% on the esophagus and trachea, 11% for the aorta, and by a large margin on the heart. This indicates the ability of the proposed model to learn discriminative features both specific for the organs at hand and also relative to their location. With respect to previous leading work that utilize data augmentation (1st and 2nd row), BB-UNet admits comparable results with respect to both VB-Net, the ISBI challenge winner as well as well as U-Net with data augmentation. In fact, model performance between them vary mildly with BB-UNet taking the lead for the esophagus, the trachea, and the aorta given all proposed experiments and in 3 out of 4 experiments in the case of the heart. These results shed light on the role of prior embedded structures in obscuring the need for data augmentation.

In comparison to the state-of-the-art models, bounding boxes within U-Net are required in both training and inference phase. Since we used bounding box prior at test time, it is logical to impose such prior on the reference model (U-Net) and compare the obtained results relative to ours. Filtering the segmentations using the bounding boxes within U-Net, we obtain the post-processed results in row "U-Net (+background) + Post" from Table II. Comparing these results with ours, we realize that indeed the proposed models provide comparable results ($2 \sim 3\%$ higher) than the post-processed U-Net segmentations, while outperforming the "U-Net + Post" in the case of the esophagus. This indicates the importance of properly integrating prior knowledge onto the model structure while training.

Comparing the different proposed structures in Table I, we gather that the BB-UNet with bounding box filters, regardless of whether it is solely the bounding box or the bounding box intersected with the CT image, perform better than circular ones contrary to the binary case. This is due to the fact that circular filters have larger shared areas generated due to the closeness of organs with respect to each other. Thus, attention areas are more overlapping in the case of circular filters than that of bounding boxes. The variation between model performance given bounding box filters (BB and BB∩CT) relative to circular filters (CC and CC∩CT) is significant to note since it opens up the discussion of the dependency of models' performance relative to the approximate area where the organ is estimated to exist in. For example, given segmentation of the aorta which according to Table I is the second largest organ (1023 pixels) and the second most common organ present in slice distribution (3363 slices), the variation in the approximated prior area has no effect on model performance which is stable at ≈0.96.

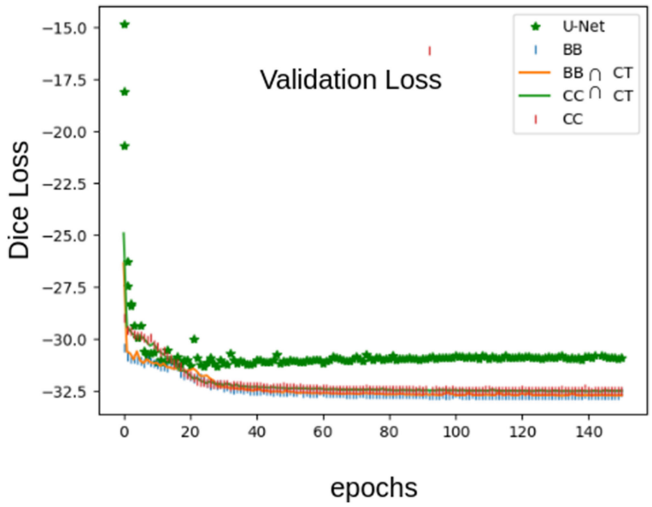

Fig. 5. Evolution Curve of the proposed models as well as U-Net for the validation set. Same Legend as Fig. 3.

Taking a closer look at the evolution of the validation Dice losses relative to the number of epochs, we derive the role of prior with respect to model convergence Fig. 5. It is evident that BB-UNet models tend to converge onto lower losses than regular U-Net during validation, which seems to be losing its generalization ability or sustaining its limited performance with epoch evolution.

TABLE III
AVERAGE DICE ACCURACY FOR FULLY SUPERVISED ORGAN SEGMENTATION OF THE ATRIUM

	Av. Dice (%)	Av. Hausdorff (mm)
State of the Art		
U-Net + (background)	77.95±1.38	2.51±0.11
U-Net + (background) + Post	82.59 ± 0.61	2.33 ± 0.04
Proposed Models		
BB-UNet-BB	89.94 ± 1.54	2.17 ± 0.13
BB-UNet-BB +Post	89.936 ± 1.256	
BB-UNet-BB∩CT	91.83 ± 0.22	2.05 ± 0.014
BB-UNet-BB∩CT +Post	89.403 ± 0.781	
BB-UNet-CC	88.97 ± 0.35	2.10 ± 0.01
BB-UNet-CC +Post	89.592 ± 0.176	
BB-UNet-CC∩CT	88.82 ± 0.26	2.10 ± 0.02
BB-UNet-CC∩CT +Post	89.262 ± 0.281	

B. Cardiac Multi-Component Organ Segmentation

Results with respect to the Atrium are benchmarked in Table III. A closer look at Table III, we realize that the utilization of BB-UNet has direct effect on segmentation quality. Thus, BB-UNet in all its modalities outperform regular U-Net by over 12% with $BB \cap CT$ model registering highest Dice accuracy scores relative to its peers. The significance of BB-UNet is relatively evident when comparing with respect to U-Net Post-processing results (2nd row). Thus, even when we integrate bounding box filters at inference time onto U-Net predictions, BB-UNet still has a leading increase in about 6% in Dice accuracy. This is also verified by the Hausdorff distances of U-Net and U-Net with post-processing relative to the proposed models. Thus, BB-UNet models register a decrease in Hausdorff distances relative to regular U-Net by about 13.54% in worst case scenarios (BB-UNet-BB: 2.51 $\rightarrow$ 2.17) and by about 18% in best case scenarios (BB-UNet-BB $\cap$ CT: 2.51 $\rightarrow$ 2.05). After Post-Processing with bounding boxes onto U-Net segments, the proposed models register a decrease in Hausdorff distance by about 7% in worst case scenarios (BB-UNet-BB: 2.33 $\rightarrow$ 2.17) and about 12% in best case scenarios BB-UNet-BB: 2.33 $\rightarrow$ 2.05. The above results, re-signifies on the importance of the BB-ConV layer as well as the importance of integrating prior embedded structures onto segmentation problems.

Comparing internally the proposed models, we realize that despite the fact that the change is slight, the utilization of intersection filters with bounding boxes (BB-UNet-$BB \cap CT$) is slightly more beneficial. This is explained by the fact that with the atrium organ, segments are often rather composed of multiple components that vary in size and position. Since the bounding box utilized is a unified box that includes all the organ components, the utilization of the intersection filters allows the model to distinguish between the different components while still posing attention on the particular area where the components can be.

VI. ROBUSTNESS ASSESSMENT OF THE PROPOSED MODEL

From the previous results, we have shown that the utilization of BB-ConV layer at the level of the skip connections allows the network to learn intrinsic properties relative to the organs under-study. In this section, we validate the robustness of the BB-UNet performance through two steps. Firstly, we study the invariance of BB-UNet performance given its different modalities when imposing Bounding Box Filtering. Secondly, we conduct a sensitivity analysis with regards to the effect of varying the bounding box size on model performances.

A. Post-Processing Comparison

In this experiment, we have applied the same post-processing step that we did with U-Net in the previous section onto the different BB-UNet modalities. The main objective was to determine whether the BB-ConV layer eliminated the need for post-processing. Results are benchmarked relative to both SegTHOR and Cardiac in Tables II and III under the name (Model +Post.). An ideal case would be a zero gap between the BB-UNet model performance vs BB-UNet + Post. This would mean that imposing bounding box post-processing will not affect the BB-UNet performances across its different modalities. From Table III, we gather that post-processing indeed resulted in little to no variation in Dice accuracy for Cardiac. This means that even given the variation of the shape/type of filter utilized, the BB-UNet still maintained its agreeable performance and does not require Post-processing. With regards to SegTHOR (see Table II), the same conclusion can be drawn when comparing relative to the Aorta and the esophagus. Thus for both organs, the gap between BB-UNet vs BB-UNet + Post. is almost null. This is not the case for the heart and trachea. Thus, a considerable gap (BB-UNET-$CC \cap CT$: 16%, BB-UNET-CC: 6%) is registered relative to the heart and about 3 to 4% for both models relative to the Trachea. Incidentally, going back to Table I that presents the slice distribution per organ and dataset size, the heart as well as the trachea are the smallest in slice size. On the other hand, the esophagus as well as the aorta are the richest ones. This provides us with intuition regarding the impact of BB-ConV relative to dataset size.

B. Sensitivity Analysis

One way to validate the proposed model is through determining the effect of bounding box variation on model performance. To do so, we vary the size of the bounding box with respect to its initial size which is the smallest bounding box that encompasses the organ. We increase the boundaries of each side of the bounding box by xpx pixels denoted by $+xpx$ in Table IV. Table IV shows that given small variations in bounding boxes (increasing bounding box size by 20% heart, 12% aorta, 20% trachea and 27% esophagus) did not reveal any significant change in model performance. In fact, these small variations may have slightly improved the already present accuracies. Conducting further variations for up to 50% of initial bounding box size resulted in slight variation in model performance while still outperforming the U-Net results by a considerable margin.

VII. TOWARDS WEAKLY SUPERVISED SEGMENTATION

In this section, the overall proposed pipeline is presented for weakly supervised training using BB-UNet, in a single organ setting (the heart). For this framework, the training set was further divided into a much smaller training set called train-ancillary,

TABLE IV
EFFECT OF BOUNDING BOX SIZE VARIATION ON DICE ACCURACY. EACH SIDE OF THE BOUNDING BOX IS INCREASED BY l PIXELS, AS SHOWN IN COLUMN **BB. VAR**. THE NEW BOUNDING BOX AREA IS $\times h$ GREATER THAN THE INITIAL BOUNDING BOX AS INDICATED IN **BB. AREA INC.** THE RESULTING NEW DICE ACCURACY IN% IS IN COLUMN **DA**. COMPARISON TO UNET WITHOUT POST-PROCESSING, COLUMN **UNET**

Organ	UNet	BB. Var	BB. Area Inc.	DA	BB-UNet
Heart	0.66	$bb + 50px$	$\times$ 4.11	0.72	0.98
		$bb + 10px$	$\times$ 1.44	0.94	
		$bb + 5px$	$\times$ 1.20	0.98	
Aorta	0.95	$bb + 10px$	$\times$ 2.67	0.92	0.957
		$bb + 5px$	$\times$ 1.72	0.95	
		$bb + 1px$	$\times$ 1.12	0.97	
Trachea	0.88	$bb + 10px$	$\times$ 4.05	0.96	0.98
		$bb + 5$ px	$\times$ 2.27	0.98	
		$bb + 1px$	$\times$ 1.21	0.98	
Esophagus	0.76	$bb + 10px$	$\times$ 5.29	0.60	0.92
		$bb + 5px$	$\times$ 2.73	0.79	
		$bb + 1px$	$\times$ 1.27	0.93	

TABLE V
AVERAGE DICE RESULTS (IN%) FOR WEAKLY SUPERVISED SINGLE-ORGAN (HEART) SEGMENTATION

	mean	std	max	min
Baselines				
U-Net with Circular Labels	54.43	22.46	77.91	0.0
U-Net with GrabCut Labels [34]	64.37	32.67	92.93	0.0
State of the art				
Simple Does It with GrabCut [8]	84.67	3.94	90.57	67.86
EM-Adapt without CRF [13]	25.97	9.61	50.72	0.3
Proposed models				
BB-UNet-BB	83.19	13.32	96.36	9.09
BB-UNet-BB$\cap$CT	84.47	5.97	95.96	60.62
BB-UNet-CC	91.69	11.27	98.77	23.82
BB-UNet-CC$\cap$CT	86.79	15.03	98.54	5.12
Full supervision baseline				
BB-UNet Full Supervision	95.29	3.51	98.55	75.65
U-Net Full Supervision [40]	91.53	11.12	98.79	10.67

consisting of 6 patients (200 slices); and a primary training set called train-primary, containing 30 patients (1244 slices). While the ancillary set has the full labels, we consider only bounding box labels in the primary set (Table I). The overall process is shown in Fig. 4.

Baselines: A naive circular baseline is established as a starting point for our implementations. Given an image and a bounding box, a circular shape encompassed within the bounding box is considered as the label estimate to our model. We also compared with our previous work on SegTHOR where we adopted a similar two-step iterative process, but generated initial label estimates using GrabCut algorithms. Moreover, comparison is done with other common weakly supervised state-of-the-art methods, such as Simple Does It (SDI) [8], and EM-Adapt [13].

As an upper baseline, we resort to the fully supervised setting where all labels as well as the bounding boxes are present. We study two fully supervised scenarios: one, a regular U-Net is trained using fully annotated segmentation maps; two a fully supervised framework enhanced with prior bounding box knowledge, where we train the BB-UNet model with circular filters.

Experimental setup: For weakly supervised segmentation processes, we adopt a similar approach to the one we implemented within [34]. We elaborated on a two-step iterative process where we generated initial label estimates basing on GrabCut-like algorithms. We then fine-tuned using regular U-Net training. In this current work, we replace the GrabCut label estimator by the ancillary BB-UNet model. We train the BB-UNet on a tiny fully supervised sample of the dataset – the ancillary training set, and then use the learnt weights in order to infer label estimates with regards to the much larger weakly supervised dataset – the primary training set (see Fig. 4).

Results analysis: As one can see from Table V, the proposed models outperform state of the art by a considerable margin, with BB-UNet with circular filters taking the lead by an increase in performance of 6% with respect to our previous work [34] and other state of the art methods. This leads us to believe that the proposed model is a viable solution when compared to the fully supervised framework. In this way one can avoid expensively annotating large datasets by making use of only a small partition of full annotation to conduct training.

VIII. CONCLUSION

In this paper, we have proposed a new model, the BB-UNet model, that is inspired by U-Net and that integrates shape and location prior by incorporating bounding areas as filters within the middle of skip connections. The proposed model outperforms the state of the art in both multi-organ and multi-component segmentation settings. We further implemented this BB-UNet within a weakly supervised framework. Promising results indicate the relevance of the proposed method relative to its peers within the state of the art.

Given the fully supervised domain, future steps are to be taken in order to relieve the BB-Unet dependency on bounding areas at inference. This can be done through addressing the feature distribution shift resulting from the augmented BB-ConV layer. Moreover, diagnostic as well as interventional imagery often consist of 3D images. Hence, exploration of the possibility of developing a BB-VNet that can perform 3D segmentation is also an aim that we hope to achieve.

Future work for the weakly supervised approach includes developing training methods suitable for weakly supervised learning using only the BB-UNet model and independent of ancillary vs primary training. This may be done through training the BB-UNet within an unsupervised framework or through an Expectation maximization setting. Moreover, a thorough study should be carried out in order to find suitable loss functions that infer relations between the bounding boxes and the corresponding label segments.

ACKNOWLEDGMENT

The authors would like to acknowledge the CNRS-Lebanon and AUF for granting a doctoral fellowship to R. El Jurdi, as well as the ANR (Project APi, grant ANR-18-CE23-0014) and the CRIANN for providing computational resources.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[2] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2017, pp. 203–211.

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

[4] A. Ghosh, M. Ehrlich, S. Shah, L. Davis, and R. Chellappa, "Stacked u-nets for ground material segmentation in remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. Workshops*, Jun. 2018, pp. 252–2524.

[5] T. Sun, Z. Chen, W. Yang, and Y. Wang, "Stacked u-nets with multi-output for road extraction," in *Proc. Comput. Vision Pattern Recognit. Workshops*, Jun. 2018, pp. 187–1874.

[6] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Int. Conf. 3D Vision*, Oct. 2016, pp. 565–571.

[7] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "NAS-Unet: Neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, pp. 44 247–44 257, 2019.

[8] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Los Alamitos, CA, USA, Jul. 2017, pp. 1665–1674.

[9] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vision*, Los Alamitos, CA, USA, Dec. 2015, pp. 1635–1643.

[10] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jul. 2017, pp. 5957–5966.

[11] C. Zotti, Z. Luo, A. Lalande, and P. Jodoin, "Convolutional neural network with shape prior applied to cardiac MRI segmentation," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 1119–1128, May 2019.

[12] D. Pathak, P. Krähenbühl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. Int. Conf. Comput. Vision*, Dec. 2015, pp. 1796–1804.

[13] G. Papandreou, L. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vision*, Dec. 2015, pp. 1742–1750.

[14] A. V. Dalca, J. Guttag, and M. R. Sabuncu, "Anatomical priors in convolutional networks for unsupervised biomedical segmentation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, Jun. 2018, pp. 9290–9299.

[15] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vision*, Oct. 2016, pp. 239–248.

[16] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 encoder pre-trained on imagenet for image segmentation," 2018, *arXiv:1801.05746*.

[17] X. Zhou, T. Ito, R. Takayama, S. Wang, T. Hara, and H. Fujita, "Three-dimensional CT image segmentation by combining 2D fully convolutional network with 3D majority voting," in *Proc. Deep Learn. Data Labeling Med. Appl.*, 2016, pp. 111–120.

[18] X. Li *et al.*, "Fully convolutional networks for ultrasound image segmentation of thyroid nodules," in *Proc. IEEE 4th Int. Conf. Data Sci. Syst.*, Jun. 2018, pp. 886–890.

[19] A. BenTaieb and G. Hamarneh, "Topology aware fully convolutional networks for histology gland segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2016, pp. 460–468.

[20] Z. Yan, X. Yang, and K.-T. T. Cheng, "A deep model with shape-preserving loss for gland instance segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2018, pp. 138–146.

[21] O. Oktay *et al.*, "Anatomically constrained neural networks: Application to cardiac image enhancement and segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 384–395, Feb. 2018.

[22] A. Arif, S. M. M. Rahman, K. Knapp, and G. Slabaugh, "Shape-aware deep convolutional neural network for vertebrae segmentation," in *Proc. Comput. Methods Clin. Appl. Musculoskeletal Imag.*, 2018, pp. 12–24.

[23] M. Tofighi, T. Guo, J. K. P. Vanamala, and V. Monga, "Deep networks with shape priors for nucleus detection," in *Proc. 25th IEEE Int. Conf. Image Process.*, Oct. 2018, pp. 719–723.

[24] C. Zotti, Z. Luo, O. Humbert, A. Lalande, and P. Jodoin, "GridNet with automatic shape prior registration for automatic MRI cardiac segmentation," in *Proc. Statistical Atlases Comput. Models Heart STACOM, Held Conjunction* , Quebec City, Canada, 2017, vol. 10663 pp. 73–81.

[25] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th ISMIR Conf.*, Paris, France, Sep. 2018, pp. 334–340.

[26] Q. Hou, D. Massiceti, P. K. Dokania, Y. Wei, M.-M. Cheng, and P. H. S. Torr, "Bottom-up top-down cues for weakly-supervised semantic segmentation," in *Proc. Int. Workshop Energy Minimization Methods Comput. Vision Pattern Recognit.*, 2016, pp. 263–277.

[27] C. Rother, V. Kolmogorov, and A. Blake, ""GrabCut": Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[28] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multi-scale combinatorial grouping for image segmentation and object proposal generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 128–140, Jan. 2017.

[29] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, Los Alamitos, CA, USA, Oct. 2017, pp. 2980–2988.

[30] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.

[31] W. Lu, Y. Zhou, G. Wan, S. Hou, and S. Song, "L3-Net: Towards learning based lidar localization for autonomous driving," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 6382–6391.

[32] L. Jiao *et al.*, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128 837–128 868, 2019.

[33] T. Araújo, G. Aresta, A. Galdran, P. Costa, A. M. Mendonça, and A. Campilho, "Uolo—Automatic object detection and segmentation in biomedical images," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 165–173.

[34] R. El Jurdi, C. Petitjean, P. Honeine, and F. Abdallah, "Organ segmentation in CT images with weak annotations: A preliminary study," in *Proc. 27th GRETSI Symp. Signal Image Process.*, Lille, France, Aug. 2019, pp. 1–4.

[35] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," in *Proc. 1st Conf. Med. Imag. Deep Learn. (MIDL)*, Amsterdam, The Netherlands, 2018.

[36] Y. Wei *et al.*, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2017.

[37] M. Han *et al.*, "Segmentation of CT thoracic organs by multi-resolution VB-Nets," in *Proc. Challenge Segmentation THoracic Organs Risk Images*, CEUR Workshop Proceedings, C. Petitjean, S. Ruan, Z. Lambert, and B. Dubray, Eds., vol. 2349. CEUR-WS.org, 2019. [Online]. Available: http://ceur-ws.org/Vol-2349

[38] Z. Lambert, C. Petitjean, B. Dubray, and S. Ruan, "SegTHOR: Segmentation of thoracic organs at risk in CT images," 2019, *arXiv:1912.05950*.

[39] A. Simpson *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019, *arXiv:1902.09063*.

[40] C. S. Perone, Cclauss, E. Saravia, P. L. Ballester, and Mohit Tare, "perone/medicaltorch: Release v0.2," to be published, doi: 10.5281/zenodo.1495335.

[41] W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1451–1461, Nov. 2006.