Body Part Regression With Self-Supervision

Yucheng Tang, Riqiang Gao, Shizhong Han, Yunqiang Chen, Dashan Gao, Vishwesh Nath, Camilo Bermudez, Michael R. Savona, Shunxing Bao, Ilwoo Lyu, Yuankai Huo, *Member, IEEE*, and Bennett A. Landman, *Senior Member, IEEE*

***Abstract*— Body part regression is a promising new technique that enables content navigation through self-supervised learning. Using this technique, the global quantitative spatial location for each axial view slice is obtained from computed tomography (CT). However, it is challenging to define a unified global coordinate system for body CT scans due to the large variabilities in image resolution, contrasts, sequences, and patient anatomy. Therefore, the widely used supervised learning approach cannot be easily deployed. To address these concerns, we propose an annotation-free method named blind-unsupervised-supervision network (BUSN). The contributions of the work are in four folds: (1) 1030 multi-center CT scans are used in developing BUSN without any manual annotation. (2) the proposed BUSN corrects the predictions from unsupervised learning and uses the corrected results as the new supervision; (3) to improve the consistency of predictions, we propose a novel neighbor message passing (NMP) scheme that is integrated with BUSN as a statistical learning based correction; and (4) we introduce a new pre-processing pipeline with inclusion of the BUSN, which is validated on 3D multi-organ segmentation. The proposed method is trained on 1,030 whole body CT scans (230,650 slices) from five datasets, as well as an independent external validation cohort with 100 scans. From the body part regression results, the proposed BUSN achieved significantly higher median R-squared score (= 0.9089) than the state-of-the-art unsupervised method (= 0.7153). When introducing BUSN as a preprocessing stage in volumetric segmentation, the proposed pre-processing pipeline using BUSN approach increases the total mean Dice score of the 3D abdominal multi-organ segmentation from 0.7991 to 0.8145.**

***Index Terms*— Body part regression, self-supervised learning, robust regression, organ navigation, multi-organ segmentation.**

Manuscript received November 4, 2020; revised January 13, 2021; accepted February 3, 2021. Date of publication February 9, 2021; date of current version April 30, 2021. This work was supported in part by the Vanderbilt-12Sigma Research Grant, in part by NSF CAREER 1452485, in part by NIH grants, and in part by NIH 1R01EB017230 (Landman). *(Corresponding authors: Yucheng Tang; Yuankai Huo.)*

Yucheng Tang and Yuankai Huo are with the Department of Electrical Engineering, Vanderbilt University, Nashville, TN 37235 USA (e-mail: yucheng.tang@vanderbilt.edu; yuankai.huo@vanderbilt.edu).

Riqiang Gao, Vishwesh Nath, Shunxing Bao, Ilwoo Lyu, and Bennett A. Landman are with the Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235 USA.

Shizhong Han, Yunqiang Chen, and Dashan Gao are with 12 Sigma Technologies, San Diego, CA 92130 USA.

Camilo Bermudez is with the Department of Biomedical Engineering, Vanderbilt University, Nashville, TN 37235 USA.

Michael R. Savona is with the Department of Medicine and Program in Cancer Biology, Vanderbilt University Medical Center, Nashville, TN 37235 USA.

Digital Object Identifier 10.1109/TMI.2021.3058281

I. Introduction

CLINICALLY acquired CT scans can exhibit large variations in field of view, especially for coverage of the chest, abdomen or pelvis (Figure 1). Without pre-processing, such scans are difficult to use for medical image analysis due to lack of spatial consistency. Using pre-processing to remove inconsistency in field of view helps to localize anatomical regions in each body part and enable more precise image registration and machine learning. Recently, Zhang *et al.* [1] suggested that image quality, quantitative analysis, and anatomical structure localization could be combined to estimate spatial consistency across the human body in CT. The potential applications include content navigation [2], lesion detection [3], classification [4] and segmentation [5], [6], which universally benefit from accurate quantitative assessment of body parts in regions of 1) shoulder and lung, 2) abdomen and 3) pelvis.

To achieve slice-wise tissue navigation and to quantify body consistency, the body part regression technique was proposed, which estimated a uniform spatial location (i.e., global position scores in Figure 2) of axial slices for a particular subject [1]. Initially, body part regression was formed as a supervised learning task using deep learning [1]. However, intensive manual annotation is required to prepare the large-scale training cohort. To alleviate the manual efforts, Yan *et al.* [7] proposed an unsupervised regression network (URN) to perform body part regression in an annotation-free manner. The method required the creation of distance metrics, varied combinations of scalars, isotropic variance, and slice thickness, which resulted in difficulties in creating consistent performance for capturing continuities between slices in medical image volumes.

The spatial location score achieved from the URN [7] can be inaccurate (Figure 2). A natural solution to improve the annotation-free method is to provide extra labeled images in a semi-supervised learning manner. However, manual annotation is resource intensive and typically not desired for accurate assessment due to complexity of patient and scanner protocols [8]. Moreover, tissues and slice thickness typically varies across different sessions and studies [9]. Herein, tissue and organ analysis are challenging problems given inter-subject variance of patient bodies and complicated 3-D volume relationships among anatomies. Holger *et al.* [10] performs selective and iterative approaches for different levels of estimation in field of views based on hierarchical architecture [10]. Han *et al.* proposed pyramid attention [11] solution for

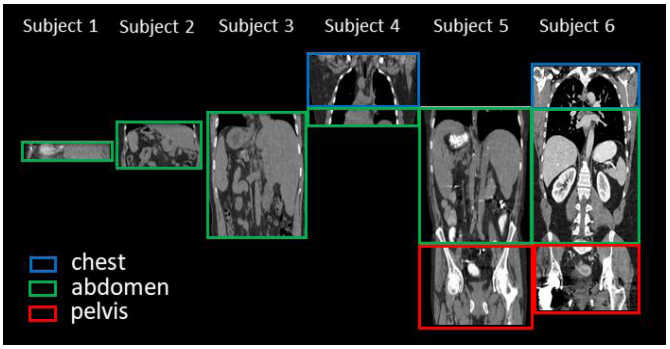

Fig. 1. The de-identified data retrieved from clinical scans under IRB approval exhibited large variations in field of view due to types of scanner, protocols of study or anatomy variance in patients.

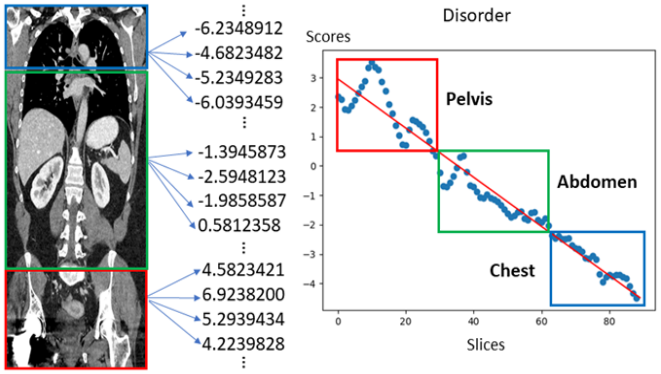

Fig. 2. Slice disorder problem in three regions with the unsupervised regression network (URN). The left panel indicates the global location scores along slices indices. The body part regression values (blue dots) are inconsistent in the right panel compared with an ideal linear relationship (red line).

accurate view selection. Herein, a further integration with body part slice selection and field of view selection is required.

Challenges and limitations with annotation-free method restrict the generalizability and robustness of models. Therefore, to leverage the current framework, we propose a self-supervised approach named blind-unsupervised-supervision network (BUSN) using robust regression [12] and uses the corrected predictions to provide extra supervision. Our contributions are in four folds: (1) an self-supervised solution to boost current body part regression are designed using 1030 multi-center CT scans without using manual labels; (2) a novel unsupervised-supervision method is introduced to achieve robust body part regression; (3) we propose a neighbor message passing correction method to further improve the BUSN results by modeling the spatial relationships between axial slices; and (4) a preprocessing pipeline is proposed using BUSN to normalize the CT volumes spatially, which is evaluated by organ navigation and 3D multi-organ segmentation on normalized scans.

Herein, 1030 whole body CT scans without manual annotation are used to train and evaluate the proposed method. First, regression scores are evaluated with R-squared metric. For body part regression results, the proposed method achieved superior performance compared with baseline methods. Second, organ-wise navigation is performed according to slice by slice scores. This experiment shows that the robust

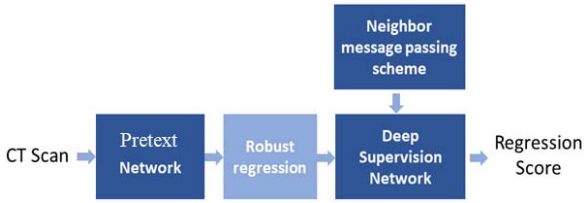

Fig. 3. Body part location inferences is performed in a self-supervised manner without manual annotation. 3D CT scans are fed into the automatic framework, and then into a cascaded unsupervised network. A self-supervision network is performed in 2D. The outputs are position scores.

body part regression is able to be used for localization of anatomies. Last, as an example of application, we trained a 3D multi-organ segmentation model for evaluating performance on abdomen CT scans using BUSN as a preprocessing stage.

II. Method

The proposed BUSN consists of four major portions: (1) an self-supervised learning network, (2) robust regression refinements, (3) a supervised learning network, and (4) neighbor message passing scheme. Note that the proposed unsupervised-supervision method does not require any new manual labels, which is trained from scratch (i.e., pre-trained networks are not used for initialization). In this study, we first trained an unsupervised model with input of only CT slices. Then we performed robust regression for modifying the labels. Last, we trained a network in the paradigm of fully-supervised learning with the annotation-free label. For testing and inference, the end-to-end model trained in the context of supervised learning is used. Figure 3 presents the overall flowchart.

A. Self-Supervision in Body Part Regression

The self-supervised learning paradigm in the task of body part regression. The self-supervised learning is used by the studies that robust modification of first step prediction can be used for supervised learning. Based on the definition, the pretext task assumes two characteristics of deep neural networks. First, the scanning procedures of medical images are serialized which the order of slices are consistent and present natural linearity. Second, fully-supervised training that learns the explicitly-provided prior labels is more accurate and stable than unsupervised learning [13].

B. Blind-Unsupervised-Supervision Network (BUSN)

1) Unsupervised-Regression Network (URN): Convolutional neural network (CNN) methods [14] provide efficient models for vision-learning tasks by employing weight connections among pixels through multiple layers. Recently, Yan *et al.* [7] proposed a self-supervised method that trains a CNN model under constrains of unsupervised learning in body-part regression problem. The unsupervised regression network (URN) consists of convolutional, ReLU, and pooling layers, which is pre-trained by the ImageNet dataset [15]. At the bottom level, a global average pooling layer is added to transfer the feature maps to a single value, followed by a fully connected layer. The unsupervised learning part is illustrated in the left panel of Figure 4.

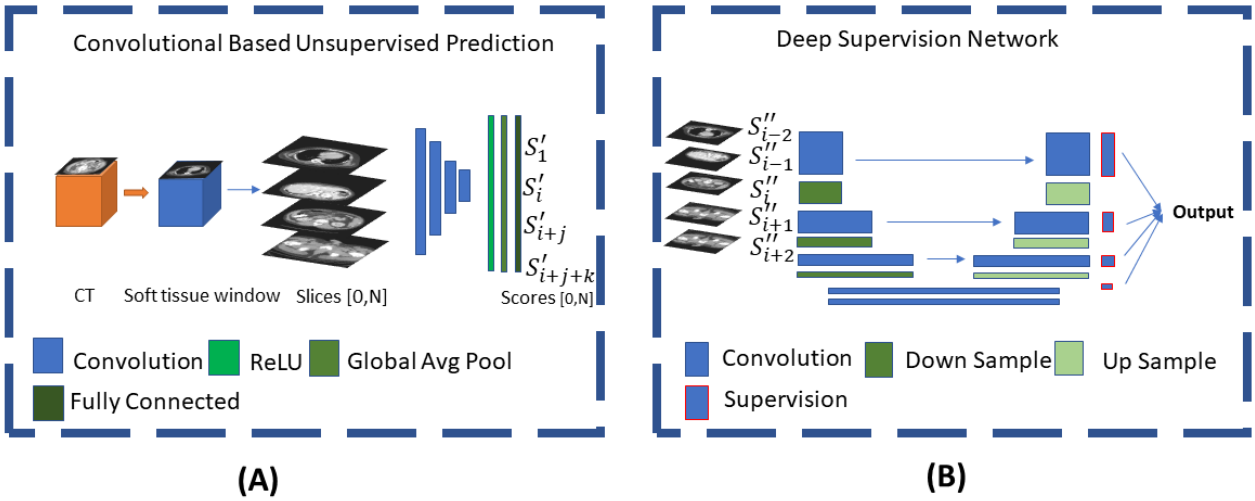

Fig. 4. Proposed deep blind unsupervised-supervision network (BUSN). Panel (A) is the unsupervised network with robust regression refinement. Panel (B) is the deep supervised network using the refined prediction scores. After training, only the right panel (B) is required to perform body part regression.

The encoder part of the unsupervised learning is learnt from the relative locations and distances between slices. Let L_a denotes the loss given sequential slices.

$$L_a = -\sum_{i=1}^{m-1} log(S\,(f\,(i+1) - f\,(i))) \tag{1}$$

where S is a sigmoid function. $f(i)$ is the predicted score of the i^{th} slice. Let L_b be the correlation between slices.

$$L_b = \sum_{i=1}^{m-2} |f\,(i+2) - 2f\,(i+1) + f(i)| \tag{2}$$

The loss function is given by: $L_U = w_a L_a + w_b L_b$. The weights w_a and w_b were empirically set as 1 and 10 according to [7]. L_b indicates the numeric difference between two slice which are proportional to physical distance of two images.

Equation 1 keeps the qualitative order of the regressed slice scores. The value variance between two slice scores are close to the physical distance between the two images. Since we used the sets of neighboring equidistant slices (e.g., slices j, j+k, j+2k, ...), the slice scores should be equidistant as well. In the experiment, the order loss (Eq. 1) and distance loss (Eq. 2) collaborate to constrain each slice score f(i) towards the direction relative to other neighboring slices.

Under the defined loss function (Eq. 1, Eq. 2), the URN output scores range in -15 and 15. The learned regression scores and patient anatomical body parts correspond well (-15: upper chest, -5: upper liver, 0: lower abdomen, 5: lower pelvis). URN is also robust to the varying position, size and imaging variance.

2) Robust Regression Refinement: Robust regression [16] refinement is introduced to further correct the inconsistent prediction values (in Figure 2) by hypothesizing that the distribution of the body part regression scores follow a linear distribution. We adopted the random sample consensus (RANSAC) [16] algorithm. As shown in Figure 7, the RANSAC robust regression help removed the 207 outlier predictions (yellow points) in a volume, and forms the corrected pseudo-labels following the evenly distributed scores. The RANSAC is an iterative approach to evaluate parameters from discrete observed data contains inliers and outliers when outliers are presented to be no influence on the values of the evaluation. RANSAC estimates parameters with high degree of accuracy even with large number of outliers. The linear distributed pseudo labels are used for the second stage training. Unlike many robust estimation approaches such as statistics of least-median and M-estimators squares [17] prevailed in image processing, RANSAC was created by resampling technique that presents candidates by minimum number of observations. The aim of the approach is to model the hidden linear trend from heterogeneous input data using robust linear regression (i.e., resilient to outliers). In our context, the robust regression is deployed to correct the discontinuity of the predicted scores as the training labels for the unsupervised learning (Figure 4a). The robust regression is critical as it helps achieve the expected linearity of score across slice indices.

3) Deep Supervision Network: The section of deep supervision network uses a symmetric convolutional architecture [18]. As shown in Figure 4b, the end-to-end network is defined with multichannel inputs, where each channel is corresponding to a score number. According to the continuous property of slices, the supervision task is formulated as multichannel pixel-wised image to score regression.

The deep supervision network [19] introduced in the BUSN method (Figure 4b) is a standard encoder-decoder CNN. As shown in Figure 4. The end-to-end network is defined with multichannel inputs (2.5D) to leverage the performance from using a single input slice (2D) every time. The U-Net [18] is used as the backbone. Different from the standard U-Net, we concatenate deep supervision (Figure 4b) from different levels to integrate the deep features from coarse to fine. A dense layer is used to convert the long dimensional 1D features from deep supervision to regress a single location score [5], [20]. During the training, we employ the L1 loss for each channel, which is more resilient to outliers compared with L2 loss. To leverage the knowledge of each level and

to alleviate the outliers, the deep supervision scheme acquires weighted sum of losses from each level. The final supervision loss L_S is given by

$$L_S = \sum_{s=1}^{N} w_s L_s \tag{3}$$

where L_S is the loss function of level s, N is the total number of levels, w_s are weight parameters of level s. The weight of final output is set as 1, while the weights of sub-labels are set as 0.1 and 0.01 [19]. N is empirically set to four to present four transpose layers in Figure 4b.

C. Neighbor Message Passing

Yang *et al.* [21] proposed a message passing scheme between landmark probability maps. The neighbor information enhancement was introduced in the context of nearby slices. Inspired by such strategy, we propose a neighbor refinement method to refine the outputs from the deep network by quantifying the relation between nearby slices to enhance the probability map of center slice.

Instead of using only a single slice for prediction, five consecutive slices are used to perform the neighbor message passing. Using the five slices, we predict the slice score of the middle slice as the output, while considering the context information from upper and lower slices.

We denote one probability map P for the center of neighboring slices (nodes), we consider slice $i-1$ and $i+1$ as the first-neighboring slices, while $i-2$ and $i+2$ as the second-neighboring slices. The overall probability distribution is formed between neighboring slices to exchange information and optimization. To express the spatial connections among slices, for each slice i, we denote each node in the graph represent the center slice ω_i. The updated probability map:

$$P(\omega_i|\mathrm{i}) = \frac{1}{|N|}\sum_{j\in N} P\left(\omega_j|j\right) * k(\omega_i|\omega_j) + P(\omega_i|\mathrm{i}) \tag{4}$$

N is the normalization term, we use the uniform factor which equals to the number of neighboring slices. The message passing is conducted by $P\left(\omega_j|j\right) * k(\omega_i|\omega_j)$, $*$ is the convolution operation. $k(\omega_i|\omega_j)$ is the convolution kernel derived from the distribution of annotation-free labels. The multidimensional convolution enables the shifting of neighboring probability maps. $P\left(\omega_j|j\right) * k(\omega_i|\omega_j)$ plays as a strong prior for $P(\omega_i|\mathrm{i})$. The predicted score of each slice can simply be determined by the corresponding probability map followed by global average pooling and linear layer.

Several recent works have studied neighbor message passing (NMP) concept for detection tasks [21]. In our framework, the NMP is used to enhance the probability map on the center slice.

III. Experiment

We evaluated the proposed BUSN method with three experiments. First, the inter-slice consistency of body part regression was assessed directly by calculating the R-squared on a large-scale cohort. Second, the efficacy of BUSN at organ-wise navigation was examined. Third, we apply BUSN as a preprocessing stage within a multi-organ segmentation pipeline.

TABLE I
DATASETS

Dataset	Website	Num
Decathlon pancreas	http://medicaldecathlon.com	421
Decathlon spleen	http://medicaldecathlon.com	61
Decathlon hepatic	http://medicaldecathlon.com	265
LiTS liver	https://competitions.codalab.org/competitions/17094	201
BTCV abdomen	https://zenodo.org/record/1169361	100
TCIA pancreas	https://wiki.cancerimagingarchive.net/display	82

Summary of datasets, BTCV dataset is used for external validation, organ navigation and multi-organ segmentation.

A. Datasets

1) Body Part Regression: The method is evaluated using a large scale of 1030 whole body CT scans from multi-center datasets (Table 1) to compensate the insufficient of views in lung and pelvis. Five multi-center datasets are used only for training, while the sixth, 100 scans of (BTCV) [22] are used for external validation. A total of 230,650 2D slices are obtained from the 3D CT scans in Table 1. The mean and variance on number of slices per scan are 224 and 35. The validation set includes 100 3D scans. All datasets were accessed in de-identified form with institutional review board approval. The in-plane pixel dimension of the volumes varied from 0.4 to 1.2 mm. Each volume was preprocessed by thresholding the soft tissue window (HU from -275 to 275) before being fed into the method. The slice thickness varies from 0.1 to 6 mm.

2) Head Part Regression: We collected de-identified brain MRI images under institutional review board for approval of head part regression. 5111 multi-site T1w MRI scans from nine different projects are used to obtain the large-scale training data, the data collection and preprocessing are discussed in [23]. The testing cohort consisted of 45 T1-weighted (T1w) MRI scans from Open Access Series on Imaging Studies (OASIS) dataset [24] with 1 mm isotropic spatial resolution.

3) Organ Navigation: We used the independent 100 whole abdominal CT volumes from BTCV as the external evaluation. We used all 100 research-controlled cases for evaluation. The in-plane resolution ranges from 0.59×0.59 to 0.98×0.98 mm^2.

4) Multi-Organ Segmentation: We used the same external cohorts of BTCV 100 scans, each with all 12 labeled organs, in the multi-organ segmentation task. We integrate the body part regression method as a preprocessing step, where each slice is assigned a slice score. For each scan, the axial slices with score between -6 and 5 are kept in the final 3D volume. The axial slices outside the range will be cropped out. While the zero-padding is applied to fill the volume if the score range is not be able to cover -6 to 5. Last, all scans are resampled to a unified dimension of [168,168,64], with the resolution of $2 \times 2 \times 6$ mm for training a 3D segmentation network [25].

B. Platform

The experiments are performed using NVIDIA Titan X GPU 12G memory and CUDA 9.0. Training, validation and testing

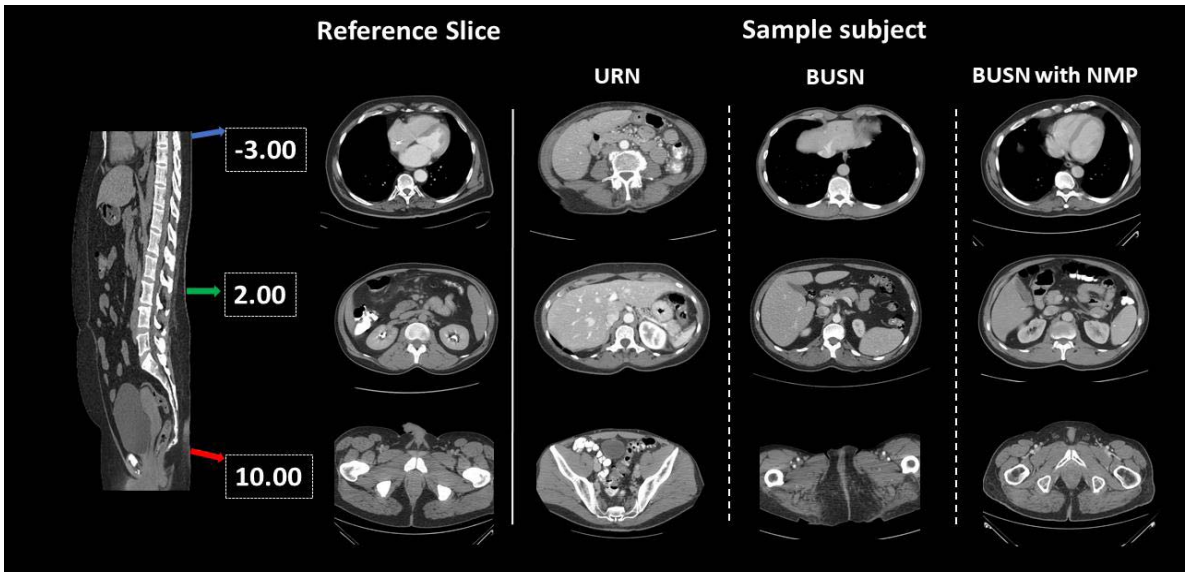

Fig. 5. The three rows show slices in chest, abdomen and pelvis regions in the same subject, under same regression score (−3, 2 and 10) with four columns (URN, BUSN, BUSN with Neighbor Message Passing and ground truth). The slice predicted by BUSN with NMP is closer to the reference slice.

are executed on a Linux workstation with Intel Xeon CPU, 32GB of RAM. The code of all experiments including baseline methods are implemented in python 3.6 with anaconda3. Networks and frameworks are established in Pytorch 1.0.

C. Experiment Design

1) Body Part Regression: To evaluate the accuracy of regression scores, we apply BUSN method to predict a score corresponding to each 2D slice. Slices are soft-tissue windowed and fed into the unsupervised network (figure 4a). A rough score is predicted according to slice thickness and continuity nature. Then, robust regression refinement is implemented for fixing incorrect scores. Finally, an accurate score is predicted from an encoder-decoder structure with deep supervision. In this section, the deep supervision network is composed of multi-level convolution layers, which are deployed in a symmetric scheme to enable efficient inference. ReLU and max-pooling layers are implemented in the encoder part of the network. Pooling layers helped to enlarge the receptive field of neurons where more contextual information is considered in layers. The decoder part consists of convolution, upsampling and ReLU layers. The convolution filters are set to $3 \times 3 \times 3$, while the maxpooling kernel is $2 \times 2 \times 2$. The stride is 1 and downsample/upsample factor is set to 2 in each dimension. We used the Adam optimization with learning rate of 0.0001. The batch size is set to 4. All weights are trained from scratch with random initialization. The results are evaluated with R-squared measurement, relative to true anatomical position. The unsupervised BPR method pipeline was employed as state-of-the-art performance for body part regression. To assess ablation performance, the self-supervision method was performed to the target CT scans with robust correction and the BUSN with neighbor message passing scheme (NMP). All experiments are implemented with same data configuration.

2) Organ Navigation: We evaluated the aforementioned three body part regression models on the withheld 100 CT scans

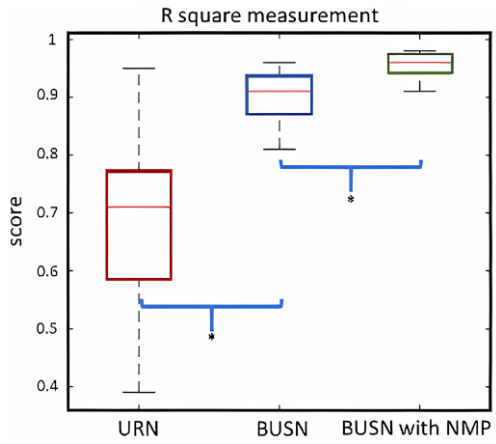

Fig. 6. The boxplot show R-squared results of the body part regression methods are presented. "∗" indicates statistically significant ($p < 0.01$ from paired t-test).

(all organs are labeled). The range of slide scores for each organ from the ground truth labels are used to present the absolute spatial location. When accumulating such scores for all 100 scans, the range of the scores are summarized as a density map (Figure 8). In Figure 8, the range and density of the predicted scores are from the URN, BUSN, and BUSN+NMP are presented in different colors. Paired t-test is used between averaged boundaries scores in Figure 8, '∗' indicates statistically significant improvement.

3) Multi-Organ Segmentation: The effectiveness of the BUSN is evaluated by using it as a preprocessing stage in a segmentation pipeline. The aim of the preprocessing stage is to reduce the spatial variations between scans. Ideally, the similar chunks of body scan can be achieved with the same score range (e.g., −6 to 5) [7]. The training was performed as a multi-channel, multi-class manner. Multi-source Dice loss [26] was employed as the loss to balance the heterogeneous size of all 12 organs. This approach normalizes voxels from

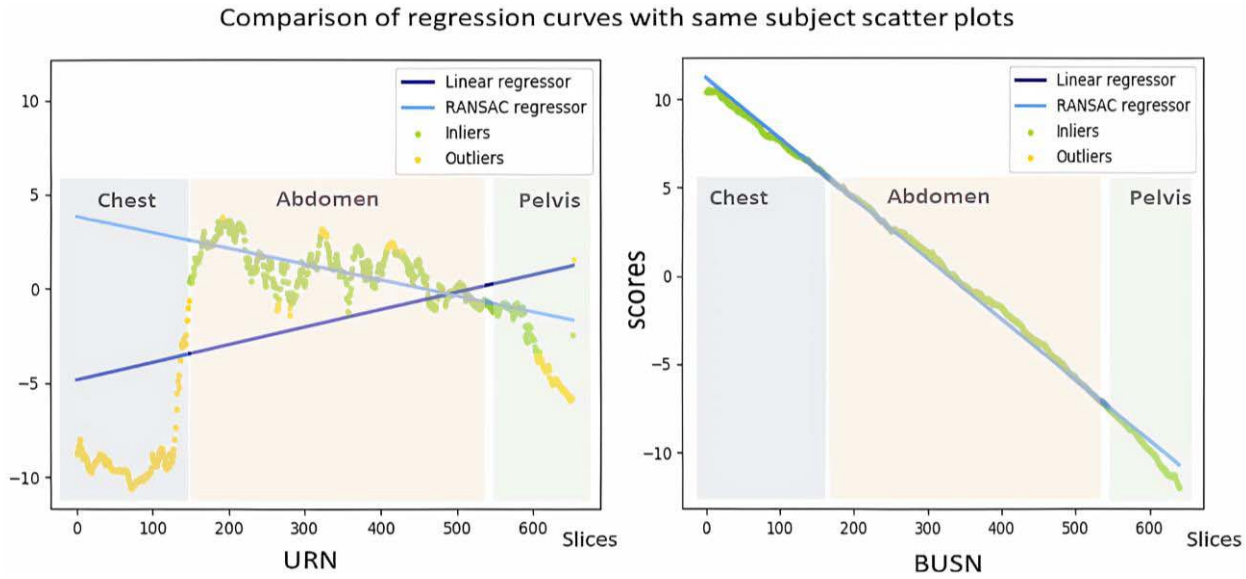

Fig. 7. A representative subject was evaluated with URN (left) and BUSN (right). Green scatters are inliers of influence to the regression, yellow scatters are outliers of no influence to the distributed data. Darker blue line indicates the normal linear regression on scatters points, lighter blue line is the RANSAC regressor result according to inliers. Left panel presents the single URN regression with amounts of outliers result in failure of linearity nature in chest and pelvis regions. Right panel shows the testing result of BUSN method, the distributed scores follows good linearity in chest, abdomen and pelvis regions in CT scan. In summary, BUSN takes advantage of self-supervised network, which presents better continuity in regression result among neighbor slices and shows scatter plots without number of outliers.

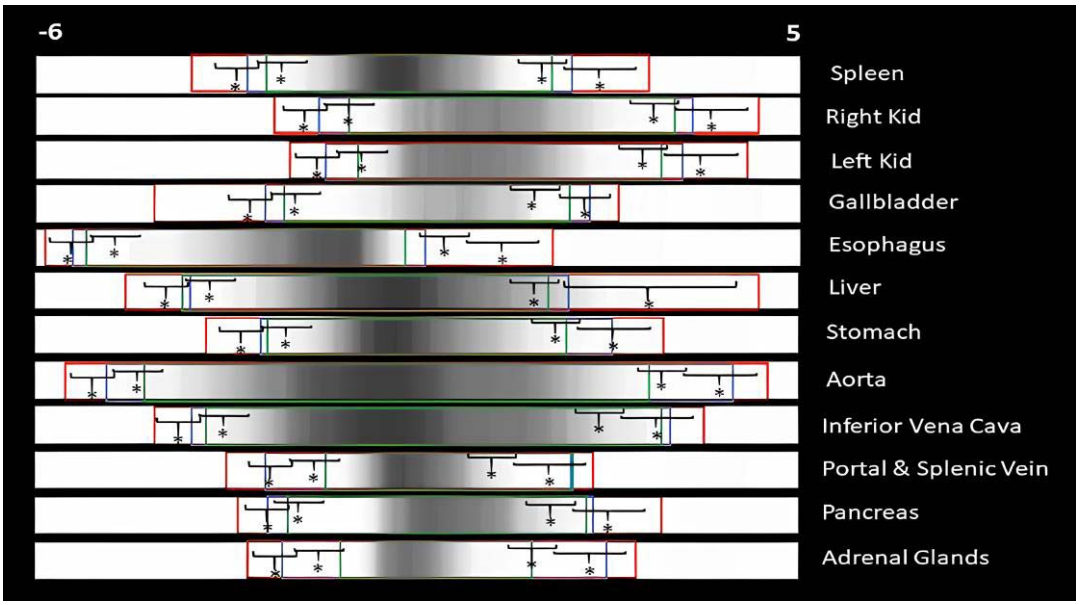

Fig. 8. Organ navigation task and organ-wise body part regression analysis: Density maps represents the distribution of each organ in whole-body CT scan. The red box range represent the URN method, while the blue box is the BUSN-plain method and the green box shows the result in BUSN with neighbor analysis. "$*$" indicates statistically significant (p-value < 0.01 from paired t-test).

prediction, which are not 'activated' in probability maps. During training, all weights are trained from scratch with random initialization. The Adam optimizer is used with a learning rate of $0.0001(\beta_1 = 0.5$ and $\beta_2 = 0.999)$, where the learning rate is decayed by a factor of 10 every 10 epochs.

4) Head Part Regression: We conduct the experiment by evaluation on MRI slices of head, to show the robustness and sensitivity of our method. Similar to body part regression experiments, we implement the baseline and our methods on brain MRI images with same parameters.

IV. Results

A. Body Part Regression

Figure 5 presents qualitative results of body part regression on a randomly selected scan. The URN indicates a slice from lung area to abdomen, a kidney area slice to liver region or an upper pelvis slice to lower pelvis. BUSN and NMP helped fix the ordering problem from the URN model. Figure 6 compares the R-squared error of all methods including URN, BUSN and BUSN with neighbor message passing. Paired

TABLE II
ORGAN DICE RESULT

Organ	NO BR (Baseline)	URN (Ke et al.)		BUSN (Ours)		BUSN + NMP (Ours)
1.Spleen	92.82 ± 2.13	94.57 ± 2.24	*	94.97 ± 2.05	*	**95.61 ± 2.01**
2.Right Kid	89.96 ± 2.54	90.41 ± 2.63	*	92.32 ± 2.43	*	**93.21 ± 2.17**
3.Left Kid	88.93 ± 2.01	90.44 ± 3.21	*	91.26 ± 2.51	*	**92.35 ± 2.12**
4.Gallbladder	53.94 ± 20.12	54.01 ± 21.14	*	54.73 ± 22.31	*	**55.92 ± 18.94**
5.Esophagus	74.81 ± 5.47	75.77 ± 7.53	*	76.78 ± 6.36	*	**76.98 ± 6.02**
6.Liver	94.58 ± 1.94	95.36 ± 2.15	*	95.73 ± 2.86	*	**96.01 ± 1.46**
7.Stomach	82.98 ± 4.21	84.01 ± 6.01	*	84.72 ± 4.12	*	**85.47 ± 3.75**
8.Aorta	90.63 ± 3.45	91.16 ± 3.92	*	91.74 ± 3.05	*	**91.95 ± 2.33**
9.Inferior vena cava	80.65 ± 3.97	81.46 ± 4.62	*	82.86 ± 3.76	*	**82.99 ± 2.19**
10.Potal&splenic vein	62.17 ± 13.48	63.79 ± 14.27	*	69.01 ± 9.57	*	**71.62 ± 9.47**
11Pancreas	67.65 ± 10.27	71.71 ± 10.85	*	73.06 ± 9.48	*	**74.21 ± 9.01**
12Adrenal glands	63.21 ± 11.36	64.02 ± 13.10	*	64.89 ± 10.84	*	**65.17 ± 8.23**

Multi-organ segmentation results with 3D U-Net and different preprocessing strategies are presented with average Dice coefficients. The best performance results marked as bold. BR means bodypart regression. "*" indicates statistically significant (p-value < 0.01 paired t-test) between left and right mean DSC.

TABLE III
SEGMENTATION PERFORMANCE COMPARISON IN TERMS OF MEAN DICE SCORES USING 3D NNUNET

Methods	spleen	R Kid	L Kid	Gall	Eso	liver	Sto	aorta	IVC	Vein	Pan	AG	All
No BR	.942	.901	.894	.557	.752	.953	.831	.908	.808	.623	.693	.650	.793
URN	.948	.915	.908	.571	.760	.961	.855	.911	.819	.640	.730	.672	.808
BUSN	.956.	.927	.928	.594	.773	.967	.868	.917	.830	.701	.758	.695	.826
BUSN+NMP	**.960**	**.934**	**.931**	**.602**	**.781**	**.968**	**.882**	**.919**	**.835**	**.712**	**.769**	**.698**	**.833**

R Kid: right kidney, L Kid: left kidney, Gall: gallbladder, Eso: esophagus, Sto: stomach, IVC: inferior vena cava, Vein: portal and splenic veins, Pan: pancreas, AG: adrenal gland.

t-test is presented as the statistical analysis. From the results, BUSN presents better linearity to unsupervised approach with significant improvement. The NMP framework shows more stable performance compare to BUSN ($p < 0.01$). Figure 7 compares the quantitative results of the body part regression using conventional linear regression and the robust regression of the same cohort. The pure BUSN network fixes the disordering problem in chest, abdomen and pelvis regions using self-supervision.

B. Organ Navigation

Across individuals, the body part regression scores should help localize organs. We performed organ-wise comparisons across 100 individuals (Figure 8). The horizontal range indicates the distribution of each organ over this cohort. For example, in the URN pipeline, the mean of top border score (left in figure 8) of spleen is -4.0217, mean of bottom border is 4.2924, while the means of BUSN is -3.7872 and 3.0158 respectively. A larger range indicates the larger uncertainty in organ localization. The green box shows a more precise evaluation after implemented the proposed BUSN method, which exclude outliers on the boundary of organs. The upper bound of aorta and lower bound of left, right kidney is defined as the abdomen. We empirically selected the scalar -6 to 5 as the range of abdomen region according to the corresponding ground truth label of all of training subjects (e.g. the top most slice that contains esophagus or liver label is score around $-5.5, -6$ is selected to ensure slices above the top most labeled slice).

C. Multi-Organ Segmentation

Table 2 shows the mean Dice similarity coefficient (DSC) and standard deviation of multi-organ segmentation. We compared four methods: (1) without using body part regression, (2) using URN, (3) using the proposed BUSN method, and (4) the proposed BUSN with NMP. Without body part regression, the performance is inferior compared with the results with body part regression method. The average DSC of BUSN with NMP is 0.8145 against URN (0.7991) The BUSN with neighbor message passing performs achieve the generally highest DSCs of organs with smaller variances. The p-value with paired t-test between No BR and URN is 0.00097, 0.015 between URN and BUSN, and 0.0017 between BUSN and BUSN with neighbor message passing.

D. Regression on Head Navigation

Figure 9 shows the regression performance in terms of R-squared measurements on head part regression. We observe

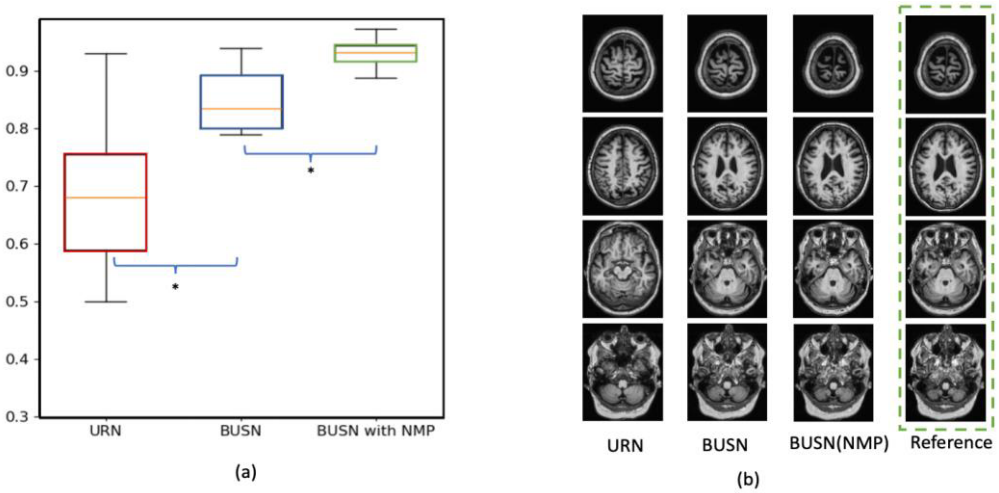

Fig. 9. (a) R-squared measurements of head part regression. "$*$" indicates statistically significant ($p < 0.01$ from paired t-test). (b) The four rows show slices in different brainn regions in the same representative subject, under same regression score (-10, 0 and 10) with four columns (URN, BUSN, BUSN (NMP) and the ground truth reference slice).

that brain images show less variations among subjects than whole-body images. The baseline URN achieves higher mean R-squared score 0.7136 than 0.6741 in whole-body images. From the results, BUSN presents better linearity to URN with significant improvement ($p < 0.01$).

E. Ablation Study of Our Methods

To evaluate the ablative effectiveness of deep supervision. We conduct an ablative experiment to compares the performance with and without deep supervision. Concretely, the method achieves the higher average R-squared scores 0.9584 ± 0.0215 than 0.9398 ± 0.0264 (without deep supervision). In addition, we formulated an external comparison on removing skip connection, to evaluate the efficacy of capturing multiple layers' feature. The result is 0.9584 ± 0.0215 (with skip-connection) against 0.9264 ± 0.0304 (without skip-connection). In addition, for comparing the effect of the introduced self-supervision and robust regression. We conducted an extra validation on using only linear model (RANSAC) and compares with and without further supervisions. Our method achieves the higher average R-squared scores 0.9238 ± 0.0253 than 0.8910 ± 0.0276 (significant improvement with $p < 0.01$ on paired t-test). Next, we compared experiments on using L2 norm error function instead of L1 loss. The R-squared measurements show 0.9584 ± 0.0215 in L1 compared to 0.9483 ± 0.0281. The further discussion on L1 vs L2 norm could be find in [27]. Empirically, larger size of input slices which means higher resolution, can preserve more spatial context in the training. In our experiment, we used the original image pixel-dimension, which presents slightly higher performance compared to down-sampled images. As more CT slices per batch introduced, the inter-slice relationship could be better regulated.

Comparison with state-of-the-art

We also train 3D nnUNet [28] for the multi-organ segmentation networks, which is known or the state-of-the-art segmentation network. The evaluation is conducted with and without BUSN. The results are shown in Table 3. Compare to 3D UNet, nnUNet performed consistently superior DSC scores. The bold values indicated the effectiveness of our BUSN pre-processing step.

V. Discussion and Conclusion

3D medical images are intrinsically spatialized, and the location of anatomies/organs are relatively structured. The goal in this study is to predict a continuous, uniformly distributed score for each axial CT slice along with body coordinate values. The predicted coordinate scores should be linearly dependent to increasing slice indices (e.g., larger scores match lower abdomen region). However, the variation of imaging process, position, size lead to morphology difference in human bodies. Herein, we take the full image of each slice, and intends to preserve the spatial context. The pixel-wised feature maps provide strong prior information, then we use the global average pooling and linear layer to obtain the final regression score. Overall, the numeric difference of the predicted slice scores can be approximately corresponded to the spatial context.

In this paper, we propose an unsupervised-supervision learning body part regression framework in a self-supervised paradigm that achieves superior performance without using manual annotations. Our method outperforms the state-of-the-art methods in terms of the body part regression accuracy. The integration of robust regression analysis leads to the pseudo-ground truth data that are exploited in the context of supervised networks. Furthermore, the part regression enables the superior content navigation (Figure 8) and volumetric segmentation (Table 2). In the segmentation task, we compared our method with the current state-of-the-art performance in the challenge dataset [29]. The averaged DSC score in [29] is 0.832 compared to 0.8179 (ours BUSN). We achieved lower but comparable DSC scores with single 3D UNet model without extra data.

One limitation of the deep learning-based segmentation is the generalizability across different reconstructed scans. In this work, all experiments and results are performed on axial

aligned images. So, the model cannot be applied to coronal view and sagittal view slides. Next, -6 to 5 are empirically set to as the range when normalizing different scans as preprocessing according to the experiments in Figure 8. However, body part regression might not be the optimal choice when applying the modal to the CT including the extremities. A detailed comparison of body part regression to detection, such as multi-atlas labeling is expected. In the future, to further improve segmentation result, it is worthy to investigate cropping volumes for each organ specifically and fitting cropped regions into segmentation models. In the segmentation task, the body part regression preprocessing takes 30 seconds on average per case, which indicates reasonable time efficiency on implementation.

In this study, the self-supervised BURN method outperforms the baseline methods. In the future, the usage of pseudo-ground truth data can be employed in the context of supervised learning methods. Herein, we could take advantage of the stability in supervised learning as well as the unsupervised nature of entire framework. Additionally, robust statistics would benefit for many approaches as quality assurance (QA) is a promising avenue to regularize information and knowledge. Therefore, the proposed self-boosted networks might be an opportunity for broader tasks with consistent performance in regression, segmentation, or classification. Investigation into the boosted approach could provide valuable improvements without extra manual efforts.

ACKNOWLEDGMENT

This study was in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN, USA. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU. The imaging dataset(s) used for the analysis described were obtained from ImageVU, a research resource supported by the VICTR CTSA award (ULTR000445 from NCATS/NIH). Clinical trial was supported by TG Therapeutics. The authors also would like to thank all of the participants that made this work possible. Additionally, they would like to thank data sources including XNAT, NITRC, ABIDE, ADHD, BLSA, EBRL, NDAR, NKI, and OASIS.

REFERENCES

[1] P. Zhang, F. Wang, and Y. Zheng, "Self supervised deep representation learning for fine-grained body part recognition," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 578–582.

[2] X. Xu, F. Zhou, B. Liu, D. Fu, and X. Bai, "Efficient multiple organ localization in CT image using 3D region proposal network," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1885–1898, Aug. 2019.

[3] 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," *J. Med. Imag.*, vol. 5, no. 3, Jul. 2018, Art. no. 036501.

[4] Y. Li and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network," *Sensors*, vol. 18, no. 2, p. 556, Feb. 2018.

[5] Q. Dou *et al.*, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016.

[6] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[7] K. Yan, L. Lu, and R. M. Summers, "Unsupervised body part regression via spatially self-ordering convolutional neural networks," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1022–1025.

[8] E. Kulama, "Scanning protocols for multislice CT scanners," *Brit. J. Radiol.*, vol. 77, no. 1, pp. S2–S9, Dec. 2004.

[9] K. J. Strauss, "Developing patient-specific dose protocols for a CT scanner and exam using diagnostic reference levels," *Pediatric Radiol.*, vol. 44, no. S3, pp. 479–488, Oct. 2014.

[10] H. R. Roth, A. Farag, L. Lu, E. B. Turkbey, and R. M. Summers, "Deep convolutional networks for pancreas segmentation in CT imaging," *Proc. SPIE*, vol. 9413, Mar. 2015, Art. no. 94131G.

[11] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*. [Online]. Available: http://arxiv.org/abs/1805.10180

[12] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. Hoboken, NJ, USA: Wiley, 2005.

[13] J. Buhmann and H. Kuhnel, "Unsupervised and supervised data clustering with competitive neural networks," in *Proc. Int. Joint Conf. Neural Netw. IJCNN*, Jun. 1992, pp. 796–801.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[16] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[17] P. J. Rousseeuw, "Least median of squares regression," *J. Amer. Statist. Assoc.*, vol. 79, no. 388, pp. 871–880, 1984.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015.

[19] Y. Liu and M. S. Lew, "Learning relaxed deep supervision for better edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 231–240.

[20] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[21] D. Yang *et al.*, "Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2017.

[22] Z. Xu *et al.*, "Efficient multi-atlas abdominal segmentation on clinically acquired CT with SIMPLE context learning," *Med. Image Anal.*, vol. 24, no. 1, pp. 18–27, 2015.

[23] Y. Huo *et al.*, "3D whole brain segmentation using spatially localized atlas network tiles," *NeuroImage*, vol. 194, pp. 105–119, Jul. 2019.

[24] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *J. Cognit. Neurosci.*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007.

[25] Ö. Çiçek *et al.*, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016.

[26] Y. Tang *et al.*, "Improving splenomegaly segmentation by learning from heterogeneous multi-source labels," *Proc. SPIE*, vol. 10949, Mar. 2019, Art. no. 1094908.

[27] J. Li, E. Elhamifar, I.-J. Wang, and R. Vidal, "Consensus with robustness to outliers via distributed optimization," in *Proc. 49th IEEE Conf. Decis. Control (CDC)*, Dec. 2010, pp. 2111–2117.

[28] F. Isensee *et al.*, "NnU-Net: Self-adapting framework for U-Net-based medical image segmentation," 2018, *arXiv:1809.10486*. [Online]. Available: http://arxiv.org/abs/1809.10486

[29] Y. Zhou *et al.*, "Prior-aware neural network for partially-supervised multi-organ segmentation," 2019, *arXiv:1904.06346*. [Online]. Available: http://arxiv.org/abs/1904.06346