

Liver Segmentation in Abdominal CT Images via Auto-Context Neural Network and Self-Supervised Contour Attention

Minyoung Chung, Jingyu Lee, Jeongjin Lee*, and Yeong-Gil Shin

Abstract—Accurate image segmentation of the liver is a challenging problem owing to its large shape variability and unclear boundaries. Although the applications of fully convolutional neural networks (CNNs) have shown groundbreaking results, limited studies have focused on the performance of generalization. In this study, we introduce a CNN for liver segmentation on abdominal computed tomography (CT) images that shows high generalization performance and accuracy. To improve the generalization performance, we initially propose an auto-context algorithm in a single CNN. The proposed auto-context neural network exploits an effective high-level residual estimation to obtain the shape prior. Identical dual paths are effectively trained to represent mutual complementary features for an accurate posterior analysis of a liver. Further, we extend our network by employing a self-supervised contour scheme. We trained sparse contour features by penalizing the ground-truth contour to focus more contour attentions on the failures. The experimental results show that the proposed network results in better accuracy when compared to the state-of-the-art networks by reducing 10.31% of the Hausdorff distance. We used 180 abdominal CT images for training and validation. Two-fold cross-validation is presented for a comparison with the state-of-the-art neural networks. Novel multiple N-fold cross-validations are conducted to verify the performance of generalization. The proposed network showed the best generalization performance among the networks. Additionally, we present a series of ablation experiments that comprehensively support the importance of the underlying concepts.

Index Terms—Auto-context neural network, contour attention network, high-level residual shape prior estimation, liver segmentation.

I. INTRODUCTION

MEDICAL image segmentation is an essential prerequisite for clinical applications of a computer-aided diagnosis system, such as volume measurement, treatment planning, and further virtual or augmented surgeries [1], [2]. Among the organs, the liver is a highly in demand as liver diseases are among the primary increasing causes of death worldwide [3]. For accurate surgical planning, such as liver transplantation and resection, volumetric information of the liver is critically required. However, manual or semi-automatic image segmentation of the liver is an impractical task owing to its large shape variability and unclear boundaries. Unlike

other organs, ambiguous boundaries with the heart, stomach, pancreas, and the occurrence of fat result in difficulty in the image segmentation of the liver. Furthermore, manual segmentation is error-prone, which implies that there exists a severe inter- and intra-observer variability in the results.

A vast body of literature on automatic liver segmentation has been previously presented. Many classical methods, before the era of deep learning, employed image- or shape-based approaches [4]–[8]. Among them, an active contour model (ACM) was a popular approach, which regards the segmentation task as a contour delineation [4], [5]. The ACM approach attempts to design an objective energy functional that drives the contour to propagate toward the target object by numerical optimization techniques [9]–[11]. However, the stopping criteria of ACM primarily rely on the local intensity distribution, which easily breaks down owing to the large variance in foreground intensity distribution and unclear boundaries of the liver. Conversely, shape-based methods, such as an active/statistical shape model, were developed to overcome such difficulties [6], [7], [12]–[14]. Shape-based methods are regarded as more successful approaches than simple intensity-based methods owing to the embedded shape priors. However, the shape-based methods also suffer from limited prior information (i.e., lack of liver database) as it is difficult to embed all inter-patient organ shapes. Moreover, fine registration is still challenging owing to irregular boundaries.

With the advent of deep learning, convolutional neural networks (CNNs) have been showing promising results over the conventional methods for the medical image segmentation task [15]–[25]. However, the performance of generalization was not addressed, which is the most important feature in the actual deployment of CNNs for medical image segmentation tasks. Many studies were conducted to obtain a high generalization performance of neural networks, such as weight decay, drop out [26], transfer learning [27], data augmentation [28], domain adaptation [29], [30], and regularization of loss functions [31]. However, these global, systematic techniques demonstrate limitations in adapting to other fields that have severe data deficiency and intrinsic class imbalance (e.g., rare cases of anomalies and phases in medical images). Thus, a domain-specific generalization technique is highly required, especially in the field of medical image analysis. Additionally, it is worth knowing that the image segmentation problem can be resolved by delineating the accurate boundaries of an object in the image, such as in ACM approaches [4], [5], [9]–[11]. However, research focusing on the implantation of

Asterisk indicates corresponding author.

M. Chung, J. Lee, and Y.-G. Shin are with the Department of Computer Science and Engineering, Seoul National University, Korea (e-mail: chungmy@snu.ac.kr).

*J. Lee is with the Department of Computer Science and Engineering, Soong-sil University, Korea (e-mail: profjjlee@naver.com).

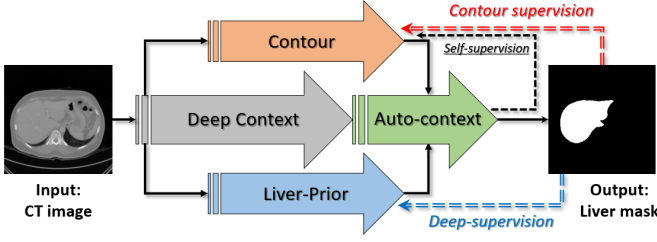


Fig. 1: Overall architecture of the proposed neural network.

a contour scheme to modern end-to-end CNN frameworks is significantly limited. Oppose to a previous study [19], it is difficult to simply supervise a CNN to delineate the full ground-truth contour in a multi-task framework because many ambiguous boundaries exist on the liver border. In this study, we propose a novel CNN architecture to address the aforementioned issues regarding generalization performance and contour scheme implantation.

The base architecture of our network is an auto-context algorithm [32]. We employed the auto-context algorithm [32] to a single neural network by using a liver-prior branch (Fig. 1). The liver-prior branch is deeply supervised to generate the probability of a liver foreground. The prior is then fused with deep contexts for the final auto-context layers. In addition to the auto-context structure, we added another branch, which is also deeply supervised to delineate the contour of a liver. Instead of training the explicit ground-truth contour, we trained sparse contours by a self-supervising method that acts as an implicit contour attention. The self-supervision is obtained by the final prediction of the network, which penalizes the ground-truth contour image based on confidence. The primary underlying principle of the proposed architecture is that the accurate segmentation of a liver can be achieved by a robust shape prior and an accurate contour delineation. This work is an extension of our previous work on a contour embedded network (CENet) [33], which employed the self-supervised contour embedding. The proposed network in this study automated the previous self-supervision of the contour by removing categorical classification loss that was formed by a heuristic threshold value [33]. We referred to our proposed network as automated auto-context CENet (AutoCENet). The network also reduced a large number of parameters based on the compact formulation of the auto-context algorithm.

The remainder of this paper is organized as follows. In Section II, several CNN models, auto-context algorithms, and contour embedding mechanisms are reviewed. The proposed method is described in Section III. The experimental results, discussion, and conclusion are presented in Sections IV, V, and VI, respectively.

II. RELATED WORKS

A. CNNs for Medical Image Segmentation

Since a fully convolutional network (FCN) was introduced, several CNN architectures have been developed for medical image segmentation tasks. To extract 3D anatomical contexts, a 3D U-net [16] was presented by replacing all the

2D convolutional operators in the original U-net with their 3D counterparts. The U-net architecture employs contracting and expanding paths together with skip connections, which combines both low- and high-level features [15]. In [17], a full 3D CNN-based U-net-like architecture was presented to segment volumetric medical images using dice coefficient that tackles the class imbalance problem. The dice loss presented in [17] intrinsically overcame the class imbalance problem by avoiding a strong bias toward background learning. A deep contour-aware network was developed to depict clear contours by designing a multi-task framework [19]. A voxelwise residual network (VoxResNet) [18] performed brain tissue segmentation by employing voxelwise residual connections. Additionally, the authors employed an auto-context algorithm to further refine the voxelwise prediction results [18]. A deep supervision mechanism [34] was employed to supervise multiple intermediate layers, which enhanced the discriminability of the low-level features [22]. The authors argued that when more discriminable low-level features are extracted, a more discriminative final classification can be obtained, which results in the improvement of the generalization performance [22]. A densely connected convolutional architecture [35] was employed by designing a similar architecture as a V-net for the task of multiorgan segmentation [24]. The singularity of the network was the introduction of a trainable grid that learns the shape prior [24]. More recently, the attention mechanism was successfully employed in the 3D U-net architecture to boost the performance of the network in [25]. The authors hierarchically applied the attention gate module to disambiguate task-irrelevant feature contexts in the intermediate layers (AGU-net) [25].

B. Auto-Context Algorithm

The auto-context mechanism fuses implicit shape information and low-level appearance features to perform image segmentation [32]. The posterior distribution of the given segmentation problem is learned with marginal distribution (i.e., classified probability map), which is further combined to learn the final classifiers. The posterior marginal distribution is learned through image patches by calculating the following distribution [32]:

$$p(y_i|\mathbf{x}) = \int p(y_i, \mathbf{y}_{-i}|\mathbf{x}) d\mathbf{y}_{-i}, \quad (1)$$

where \mathbf{x} , \mathbf{y} present a given image and ground-truth label vector, respectively, and \mathbf{y}_{-i} is a marginal set, $\{\mathbf{y} - y_i\}$. We have omitted patch representation for simplicity. Traditional feature extractors (e.g., Haar [36], histogram of oriented gradients [37]) and classifiers (e.g., probabilistic boosting tree [38]) were used for patch-wise prediction for the calculation (1). The algorithm iteratively solves the posterior probability with the previous marginal distribution:

$$p^{(t)}(y_i|\mathbf{x}, \tilde{\mathbf{p}}^{(t-1)}) \longrightarrow p(y_i|\mathbf{x}), \quad (2)$$

where $\tilde{\mathbf{p}}^{(t-1)}$ is a posterior marginal for each pixel i learned according to (1). It was proven by the authors that the algorithm asymptotically converges to $p(y_i|\mathbf{x})$ with a discrete,

iterative process. In contrast to the original paper [32], we used the term “*context*” in this paper as a feature used in the second classifier (i.e., not shape information).

C. Self-Supervised Contour Embedding

The contour features were successfully embedded in the network in [33]. The authors deeply supervised the contour extraction layer by a dynamic modification of the ground-truth contour for each iteration, as presented below:

$$\tilde{\Gamma}_c = \Gamma_c \otimes (\tilde{\mathbf{y}}_p), \quad (3)$$

where \otimes is an element-wise multiplication operator, Γ_c is the ground-truth contour image, and $\tilde{\mathbf{y}}_p$ is a binary image with respect to the threshold value p :

$$\tilde{\mathbf{y}}_p(x) = \begin{cases} 1, & \text{if } \tilde{\mathbf{y}}(x) < p, \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where $\tilde{\mathbf{y}}$ is the output probability prediction of the proposed network for a given iteration. That is, the ground-truth contours were automatically erased if the network successfully delineated the corresponding labels at the output. A manually self-supervised contour embedding mechanism was established by explicit attention to the misclassified contour region. The training of the contour feature was performed by a cross-entropy classification loss for each voxel.

III. METHODOLOGY

The proposed network architecture is composed of three primary branches: liver-prior, context, and contour, i.e., the blue, gray, and orange dotted boxes in Fig. 2, respectively. The liver-prior network is deeply supervised to estimate the ground-truth liver in a lower resolution. The trained posterior (i.e., the output of the liver-prior network) is used as a prior in the remaining auto-context network. Deep features that are trained by the context network are concatenated to the prior for the final auto-context fusion. In addition, the contour attention branch is also deeply self-supervised with a penalized ground-truth contour regarding the output of the network. There are two different non-linear modules in our network: skip-attention block (Fig. 3a) and V-transition layer (Fig. 3b). Each module is comprised of depth-wise separable convolutions, batch normalization [39], rectified linear unit (ReLU) nonlinear activation function [40], skip connection, and channel-wise attention. Details of the architecture are described in the following subsections.

A. Common Feature Extraction

The skip-attention block (Fig. 3a) is first used to extract common features (i.e., shared features in the following layers). Subsequently, the features are fed to the liver-prior, context, and contour sub-networks (Fig. 2). The skip-attention block is composed of non-linear transformation series: separable convolutions, batch normalization, and ReLU non-linear activation function (Fig. 3a). These transformations are skip connected for feature reuse. We introduced depth-wise separable convolutions [41] in the skip-attention block rather

than bottleneck [42] or compression [35] layers for more efficient use of parameters. The attention mechanism is applied to the final output to employ channel-wise attention, similar to [43]. Unlike [43], for simplicity, we directly applied a trainable channel-wise attention vector that is multiplied for each channel.

B. Liver-Prior Inference and Auto-Context Algorithm

The base architecture of the proposed network is the auto-context algorithm. Instead of stacking deep neural layers, our proposed network uses multiple shallow stacks of layers (Fig. 2). The liver-prior and context layers are composed of V-transition layers, which are small V-net-like modules that include down and up transitions together with skip connections (Fig. 3b). The channel-wise attention is applied to the features in the lower resolution. The two identical shape transitions are used in the liver-prior block to subtract each output prediction at a higher level (blue dotted box in Fig. 2). The output is deeply supervised with the ground-truth label image. The dual-passing architecture effectively learns mutually complementary features for the accurate inference of the liver posterior. The objective function for deep supervision of liver-prior can be defined as follows

$$L_p = \mathcal{D}((V_p^0(S(\mathbf{x})) - V_p^1(S(\mathbf{x}))), \mathbf{y}_{\text{dl}}), \quad (5)$$

where $\mathbf{x}, S, \mathbf{y}_{\text{dl}}$ denotes input image, skip-attention block, and the ground-truth liver label at down-scaled resolution, respectively. \mathcal{D} denotes the soft dice loss [17] and V_p^i indicates the i^{th} V-transition in the liver-prior sub-network. Finally, the output feature map is concatenated to the context features (i.e., output of the context sub-network; $V_c(S(\mathbf{x}))$) and is passed through an auto-context sub-network (V_a) for the final refinement:

$$L_f = \mathcal{D}\left(V_a\left(\left[V_c(S(\mathbf{x})), R_p(S(\mathbf{x})), V_c(S(\mathbf{x}))\right]\right), \mathbf{y}_1\right), \quad (6)$$

where R_p denotes the residual output of the liver-prior sub-network, V_c indicates the contour V-transition described in the subsequent subsection, and \mathbf{y}_1 denotes the ground-truth liver label.

The V-transition architecture is visualized in Fig. 3b. The down-transition process down-samples the feature map by a factor of two for each dimension through 2^3 convolutions with stride = 2. Conversely, an up-transition process restores the dimensions through a de-convolution (i.e., transposed convolution). We designed a skip connection and channel-wise attention in the lower dimension. By contracting and expanding paths, the V-transition layer can extract more multiscaled features (i.e., higher receptive field). A 1^3 convolution is applied to the final concatenated features for further propagation. The number of output channels is as illustrated in Fig. 2.

C. Understanding the Network

Let vectors $\mathbf{x} = \{x_i \in \mathbb{R}, i \in \mathbb{R}^3\}$ and $\mathbf{y} = \{y_i \in \{0, 1\}, i \in \mathbb{R}^3\}$ represent the input image and ground-truth label, respectively. The objective of the given segmentation problem is to determine the optimal solution for modeling

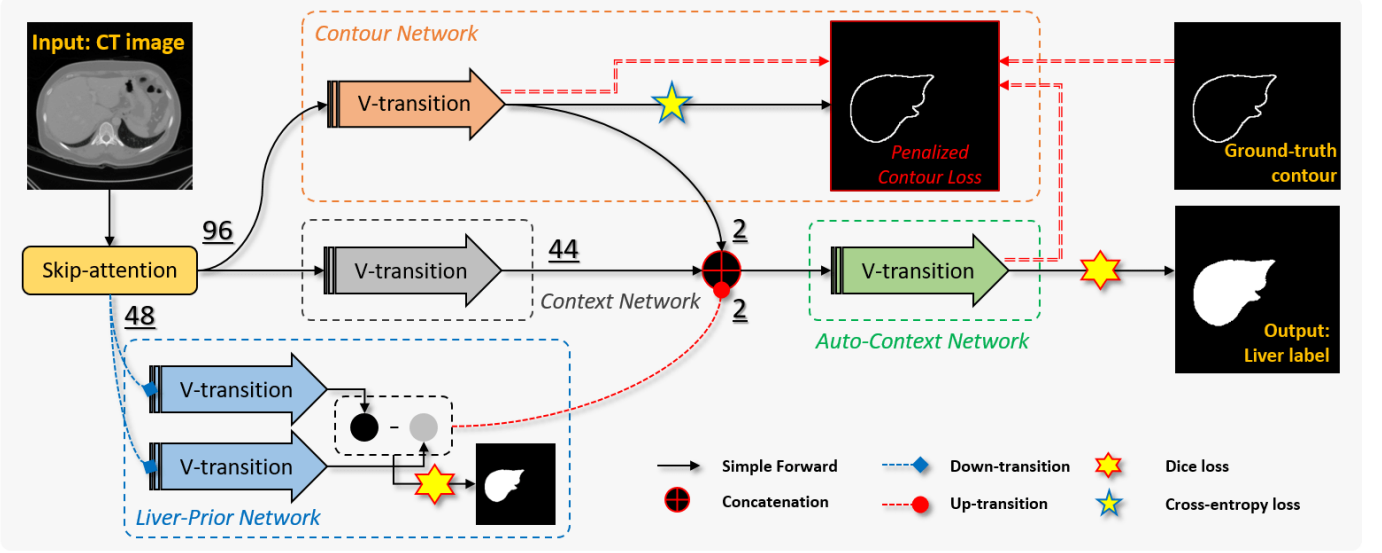


Fig. 2: Proposed 3D network architecture. Stacked V-transitions form a base module with multiple skip connections. The red (i.e., circled arrow) and blue (i.e., squared) arrows indicate up- and down-transition layers, respectively. The red and blue dotted boxes represent the contour and liver-prior transitions, respectively. The two transitions are deeply supervised by the penalized contour and ground-truth liver foreground. The penalized contour loss is formulated by employing the final output, output of the contour network, and the ground-truth contour labels (double lines with red arrows). All the images are displayed in 2D for simplicity. This image is best viewed in color.

a conditional probability distribution, $p(\mathbf{y}|\mathbf{x})$, by determining the maximum a posterior, as presented below:

$$\theta^* = \arg \max_{\theta} p(\mathbf{y}|\mathbf{x}; \theta) = \arg \max_{\theta} p(\mathbf{x}|\mathbf{y}; \theta)p(\mathbf{y}), \quad (7)$$

where θ is a parameter set for classifiers. However, it is significantly difficult to model the likelihood (i.e., $p(\mathbf{x}|\mathbf{y}; \theta)$) and prior (i.e., $p(\mathbf{y})$); moreover, solving the decomposed posterior with a generative approach easily yields inaccurate results, primarily owing to the difficulty in likelihood and prior estimations. Our proposed network iteratively solves the posterior directly using the auto-context method [32]. In auto-context, the previous classification map is used as a shape feature (i.e., the term “context” is used in the original paper) for additional classification. Setting t as a discrete time value, the auto-context is formulated as

$$p^{(t)}(\mathbf{y}|\mathbf{x}, p^{(t-1)}(\mathbf{y}|\mathbf{x}; \theta_{(t-1)}); \theta_t) \longrightarrow p(\mathbf{y}|\mathbf{x}; \theta^*). \quad (8)$$

Unlike the previous approaches [32], [44], we combined the shape-feature extraction procedure with a single-passing neural network. The output of our proposed network for time t can be formulated as

$$p^{(t)}(\mathbf{y}|\mathbf{x}, \tilde{p}^{(t)}(\mathbf{y}|\mathbf{x}; \theta_t); \theta_t), \quad (9)$$

where \tilde{p} is a probability map of shape-residual sub-network (i.e., the blue dotted box in Fig. 2). Applying deep supervision (i.e., auxiliary classifiers), we could obtain a single-passing neural network embedded with a previous posterior. Thus, we avoided using separated classifiers and storing previous classification maps.

D. Self-Supervising Contour Attention

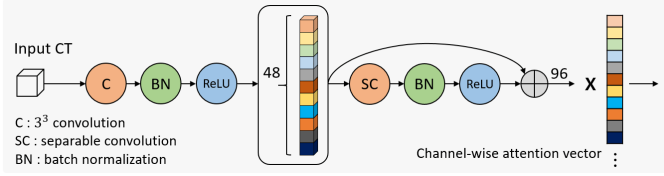
From the base architecture of the aforementioned auto-context framework, we extended our network with an explicit focus on contour features. The primary differences from the original work [33] are the automatization of the training procedure and removal of the categorical classification loss. Unlike defining the threshold and manipulating it manually during the iteration, we employed a penalized contour soft loss with respect to the output predictions of the network. We first calculated the contour weighting map that has larger values for the misclassified contour as follows:

$$\widehat{\Gamma}_c = \Gamma_c \otimes \tilde{\mathbf{y}}_1^{-1}, \quad (10)$$

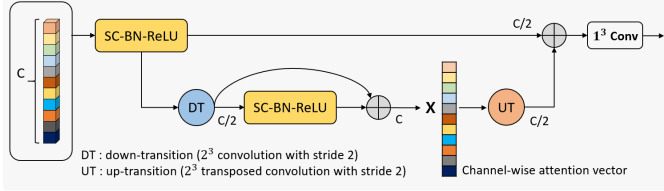
where Γ_c , \otimes , and $\tilde{\mathbf{y}}_1^{-1}$ indicate the ground-truth contour image, element-wise multiplication operator, and the final inverse liver prediction, respectively. The ground-truth contour image contains a value of 1 for the contour and 0 elsewhere. For the inverse prediction, we applied $\tilde{\mathbf{y}}_{1i} = 1 - \tilde{\mathbf{y}}_{li}$ for every i^{th} voxel, where $\tilde{\mathbf{y}}_1$ is the final output prediction of a foreground liver after softmax operation. Finally, we applied a penalized contour loss as follows:

$$L_c = - \sum_{i \in \Omega} (w_0(1 - \Gamma_{c,i})\log(1 - \tilde{\mathbf{y}}_{c,i}) + w_1\Gamma_{c,i}\widehat{\Gamma}_{c,i}\log(\tilde{\mathbf{y}}_{c,i})), \quad (11)$$

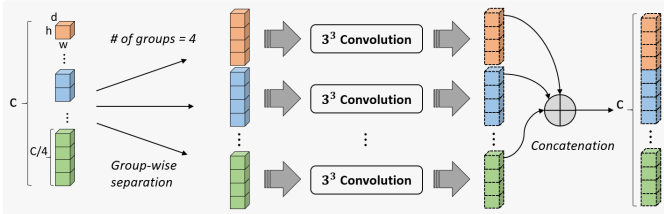
where $\tilde{\mathbf{y}}_c$ is the output prediction of the contour after softmax operation, w_c denotes class-specific weights for class c , and Ω indicates the dimensions of the image (i.e., $\Omega \in \mathbb{R}^3$). Consequently, the contour loss includes sparse contour attention based on the final output (10), which is employed to penalize the confident output of the network at each iteration. The



(a) Skip-attention layer (common feature extraction module). The intermediate features are skip-connected by concatenation. A trainable channel-wise attention vector is employed for the final output features. The number of output features is 96.



(b) V-transition layer. A structured non-linearity module (i.e., SC-BN-ReLU) composed of a series of separable convolutions (SCs), batch normalizations (BNs), and rectified linear units (ReLUs). A multiscaled feature analysis is applied by down-transition through 2^3 convolutions with stride = 2 and up-transition through 2^3 transposed convolutions with stride = 2. A skip-connection and channel-wise attention vector are employed in the lower resolution similar to the skip-attention block. The final output is obtained by a 1^3 convolution applied to the concatenated features.



(c) Depth-wise separable convolutions. The input channels are separated by groups and are convolved separately. The final output is a concatenation of all groups. The number of groups is four in the proposed network.

Fig. 3: Proposed non-linear layers employed in the proposed network: (a) skip-attention, (b) V-transition layers, and (c) depth-wise separable convolutions.

difference between the proposed loss function and the focal loss [45] is that the proposed self-supervision is intended to penalize the confident output regarding the final liver prediction rather than the confidence of the contour itself.

E. Learning the Network

The task of the given learning system is to maximize the posterior, $p(y|x)$. To effectively model the probability distribution, we attempted to train our network model to map the segmentation function $\phi(x) : x \rightarrow \{0, 1\}$ by minimizing the following loss function:

$$L = L_f + \alpha L_p + \beta L_c + \gamma \|W\|_2^2, \quad (12)$$

where L_f , L_p , and L_c indicate objective functions defined at the final output (6), shape prior (5), and contour (11) layers, respectively. W is a whole set of network parameters. α , β , and γ are weighting parameters. The output of the network is obtained by applying softmax to the final output feature maps.

“Xavier” initialization [46] is used for initializing all the weights of the proposed network. While training the network,

we fixed the loss parameters as $\alpha = \beta = 1$ and $\gamma = 0.1$ in (12). We used the rectified Adam optimizer [47] with a batch size of 4 and learning rate of 0.001. We decayed the learning rate by multiplying 0.5 for every 10 epochs. We trained the network for 100 epochs using an Intel i9-7900X desktop system with 3.30 GHz processors, 128 GB of memory, and Nvidia Titan RTX (24 GB) GPU machine. We implemented the network using the PyTorch framework. It took 2h to complete all the training procedures.

F. Data Preparation and Augmentation

We acquired 180 subjects in total: 90 subjects from a publicly available dataset¹ in [24], 20 subjects from MICCAI-SLiver07 dataset [8], 20 subjects from 3Dircadb², 20 subjects from CHAOS challenge³, and additional 30 annotated subjects with the help of clinical experts in the field. In the dataset, the slice thickness ranged from 0.5–5.0mm and pixel sizes ranged from 0.6 – 1.0mm.

The whole dataset was separated into three sets: training, validation, and testing. We first randomly shuffled the dataset and separated 80 images for testing. The remaining 100 images were used for training based on a two-fold cross-validation (i.e., 50 training images and 50 validation images). We resampled all abdominal CT images into $256 \times 256 \times 64$. We pre-processed the image using fixed windowing values: level=10 and width=700 (i.e., we clipped the intensity values under -340 and over 360). After re-scaling, we normalized the input images into the range $[0, 1]$ for each voxel. On-the-fly random affine deformations were subsequently applied to the dataset for each iteration with an 80% probability.

IV. EXPERIMENTS

In our experiments, we evaluated the performance in terms of accuracy and generalization of our proposed network by comparing these results with those of the other state-of-the-art FCN-based models. We used 3D U-net [16], V-net [17], deeply supervised network (DSN) [22], VoxResNet [18], DenseVNet [24], AGU-net [25], CENet [33], and our proposed network, AutoCENet for the performance evaluation.

A. Evaluation Metrics

The segmentation results were evaluated using the F1 score, precision, sensitivity, Hausdorff distance (HD), and average symmetric surface distance (ASSD). The F1 score is defined as follows:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}. \quad (13)$$

Precision and sensitivity are defined by $P = \frac{TP}{TP+FP}$ and $S = \frac{TP}{TP+FN}$, where TP, FN, and FP are the numbers of true positive, false negative, and false positive voxels, respectively. The F1 score is equivalent to the dice coefficient [17]. The surface distance metrics were evaluated on a surface basis:

¹<https://doi.org/10.5281/zenodo.1169361>

²<https://www.ircad.fr/research/3dircadb>

³<https://doi.org/10.5281/zenodo.3367758>

TABLE I
Accuracy evaluation of the proposed network and the state-of-the-arts

Methods	DSC	Precision	Sensitivity	HD [mm]	95% HD [mm]	ASSD [mm]
3D U-net [16]	0.95 ± 0.01	0.94 ± 0.02	0.96 ± 0.02	45.20 ± 31.93	7.77 ± 12.71	1.33 ± 0.91
V-net [17]	0.95 ± 0.02	0.94 ± 0.02	0.95 ± 0.03	26.52 ± 19.05	5.38 ± 3.94	1.20 ± 0.65
DSN [22]	0.92 ± 0.02	0.88 ± 0.04	0.97 ± 0.01	28.63 ± 23.85	7.40 ± 9.33	1.77 ± 1.05
VoxResNet [18]	0.95 ± 0.01	0.95 ± 0.02	0.95 ± 0.02	18.67 ± 11.15	4.99 ± 5.89	1.11 ± 0.49
DenseVNet [24]	0.83 ± 0.05	0.75 ± 0.09	0.94 ± 0.03	37.19 ± 14.52	16.54 ± 8.47	3.98 ± 1.69
AGU-net [25]	0.95 ± 0.01	0.94 ± 0.03	0.96 ± 0.01	31.57 ± 22.22	8.56 ± 13.52	1.34 ± 1.07
CENet [33]	0.95 ± 0.01	0.95 ± 0.02	0.96 ± 0.01	16.68 ± 8.87	3.55 ± 1.36	0.94 ± 0.38
AutoCENet	0.96 ± 0.01	0.95 ± 0.02	0.97 ± 0.01	14.96 ± 4.25	2.92 ± 1.12	0.82 ± 0.32

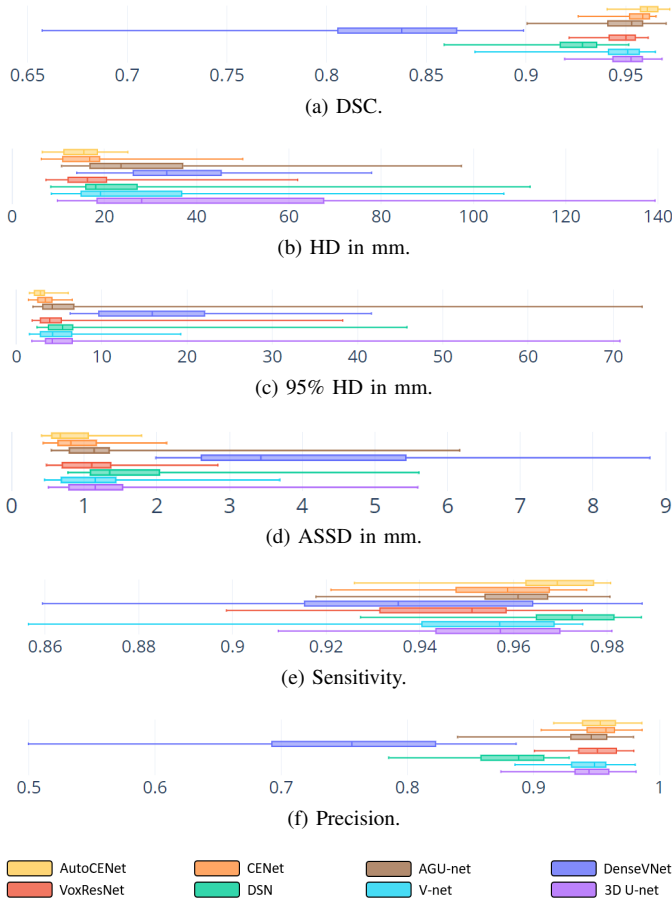


Fig. 4: Box plots of the evaluation metrics for state-of-the-art networks.

HD, 95% HD [33], and ASSD [8]. We applied 95% of voxels for HD to exclude the 5% of outlying voxels. The 95% HD is a better-generalized evaluation of distance because there exist ground-truth variations on a portal vein region. Two-fold cross-validation was used to obtain the quantitative results listed in Table I.

B. Comparison

Table I shows the quantitative results of liver segmentation. The results show that our proposed AutoCENet along with

CENet [33] outperformed other state-of-the-art networks; moreover, our network showed better accuracy while using much fewer parameters than CENet. Table I lists that AutoCENet reduced ASSD by 12.77% when compared to CENet. The lowest precision and sensitivity were presented by DenseVNet [24]. DenseVNet failed to segment the liver accurately because of two significant reasons: 1) the resolution of the network was too low and 2) shape prior was not robust. The excessively coarse dimensions of the network suffer from inaccurate segmentation in the original image resolution. Furthermore, 12^3 resolution of shape prior is too small; moreover, the training images must be accurately and manually cropped for the robustness of the shape prior. There is no specific metric presented in the original paper [24] to crop the testing images automatically. The DSN [22] showed high ASSD because the network was inferred from low resolution. The up-sampling process from $40 \times 40 \times 18$ demonstrated limitations in accurately delineating objects in the original resolution. The results indicate that multiple deep supervisions in DSN enforced the lower-level intermediate features to be discriminative, which resulted in degradation of the overall performance. The AGU-net also presented many false positives as opposed to the architectural design principle proposed in the original paper [25]. The spatial attention-gated units in AGU-net [25] failed to suppress irrelevant background regions as suggested. Conversely, VoxResNet [18] showed the second minimum distance errors. The results of VoxResNet indicates that the auto-context algorithm successfully suppressed false positive responses. The box plots of the results listed in Table I are illustrated in Fig. 4.

C. Ablation Study

We extended our experiments to verify the architectural components of the proposed network. We first validated the auto-context framework that does not exploit contour features (i.e., without contour loss, L_c in (12); AutoCENet). From the base auto-context framework, four additional ablations were studied: without channel-wise attention (AutoNet-att), without the auto-context part (i.e., AutoNet-A), without high-level residual inference (i.e., AutoNet-R), and without both auto-context and high-level residual inference (i.e., AutoNet-AR). In the case of AutoNet-A, we removed the deep supervision for

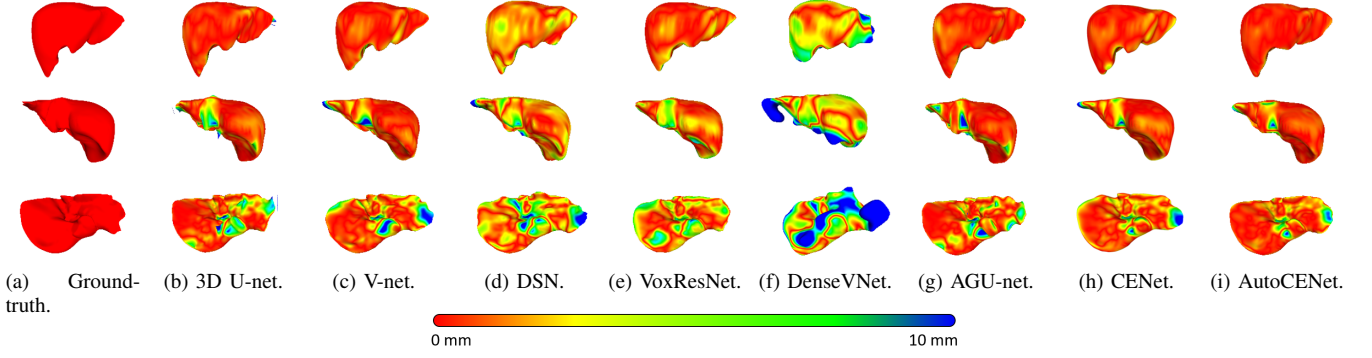


Fig. 5: Visualizations of the test results for state-of-the-art networks. The surface color is visualized based on the distance to the ground-truth surface.

TABLE II
Performance of AutoCENet and its ablations

Methods	DSC	HD [mm]	ASSD [mm]
AutoCENet	0.96 ± 0.01	14.96 ± 4.25	0.82 ± 0.32
AutoNet	0.95 ± 0.01	20.18 ± 8.79	1.04 ± 0.42
AutoNet-att	0.95 ± 0.01	25.73 ± 17.06	1.10 ± 0.53
AutoNet-A	0.95 ± 0.01	33.25 ± 22.80	1.34 ± 0.71
AutoNet-R	0.95 ± 0.01	37.99 ± 25.09	1.23 ± 0.57
AutoNet-AR	0.94 ± 0.01	38.88 ± 28.81	1.32 ± 0.61
AutoCENet+FC	0.95 ± 0.01	27.56 ± 20.34	1.20 ± 0.56
AutoCENet+MC	0.95 ± 0.01	24.20 ± 15.26	1.14 ± 0.56

the liver-prior network (Fig. 2). For AutoNet-R, the high-level residual connection was modified to a sequential connection of V-transitions with the number of intermediate features equal to 48. AutoNet-AR employed both the modifications corresponding to AutoNet-A and AutoNet-R. The results (Table II) showed that the accuracy of all the ablations was lower than the original AutoCENet. In AutoNet ablations, a significant increase pertaining to distance metrics was observed when the auto-context algorithm or residual shape prior were not employed. The results indicate that the auto-context framework and residual shape prior estimation jointly performed an important role in the final accuracy. The results of the liver-prior network with and without the residual inferences showed that the high-level residual connection boosted the performance of the liver-prior network. Sample visualizations of the liver priors are presented in the following subsection.

To verify the proposed self-supervised contour attention loss, we additionally experimented with two different contour losses: full-contour supervision of the ground-truth contour (AutoCENet+FC) and manual self-supervision, which was previously proposed in [33]. The former full supervision was conducted without the penalization term presented in (11). The latter self-supervision was conducted by employing modified contour supervision, as presented in (4) [33]. All the contour variants showed lower accuracy when compared to

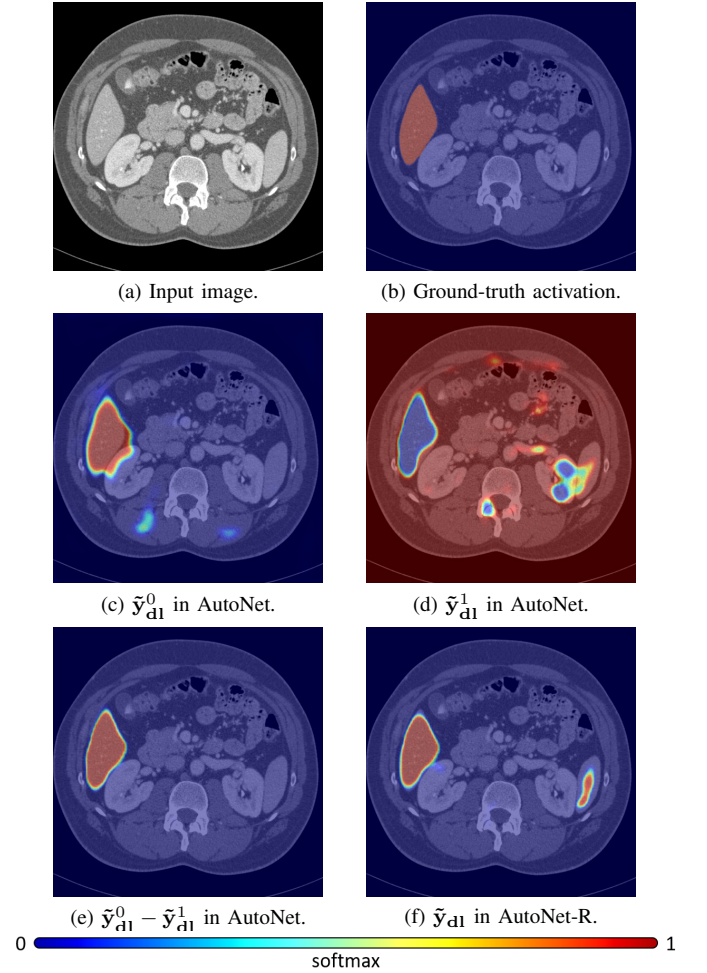


Fig. 6: Liver prior estimations of AutoNet and AutoNet-R.

the original network. The performance of the AutoCENet+FC was more inferior than that of the AutoNet (Table II) in terms of distance measures, indicating that enforcing the network to learn the full ground-truth contour image degrades the performance. Sample visualizations for the fully supervised and self-supervised contour feature maps are illustrated in the following subsection.

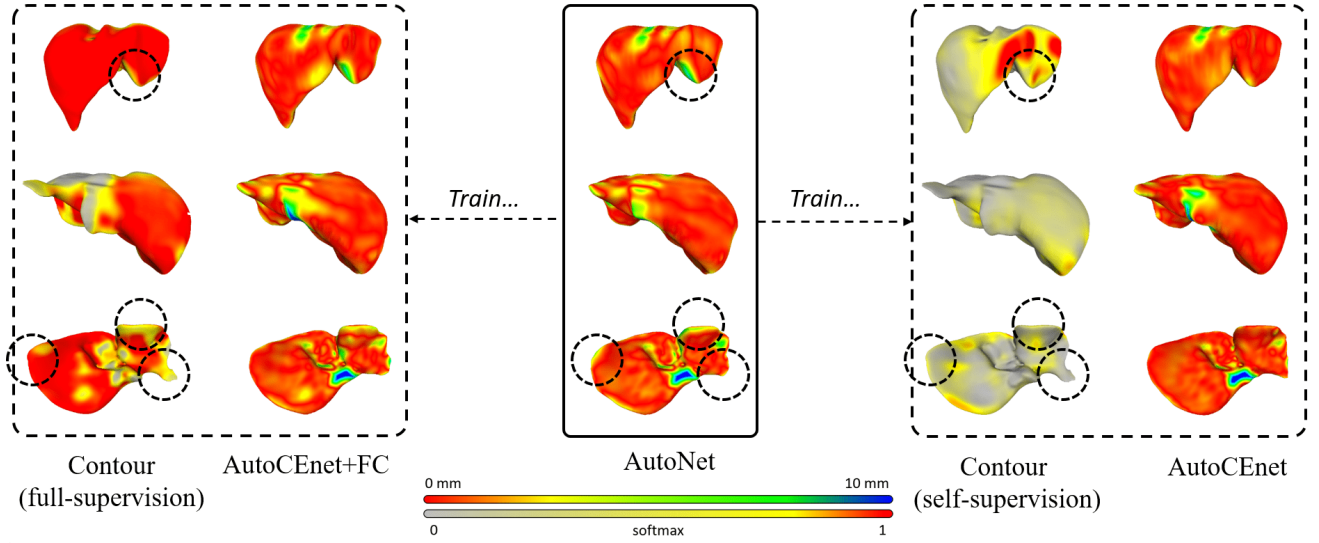


Fig. 7: Visualizations of contour feature map and the final outputs after full training of AutoNet: full-contour supervision (left; AutoCENet+FC) and self-supervision (right; AutoCENet). The self-supervised contour feature map is sparser than that of the full supervision and is implicitly utilized as a strong contour attention. The ground-truth surface is used for visualizing the distribution of the contour feature. The softmax value is normalized into the range [0-1].

D. Liver-Prior and Contour Feature Analysis

Figure 6 shows the liver shape priors that were estimated with and without the proposed residual connection. The predicted probabilities clearly show the effectiveness of the high-level residual connection in shape prior estimation. The posterior of the liver from AutoNet-R (Fig. 6f) shows significant false positive responses when compared to the version with a residual connection (Fig. 6e). The two high-level predictions, i.e., Figs. 6c and 6d, were used as mutual complements to derive accurate liver prediction. The results indicate that the high-level residual inference shows an effective method to estimate accurate prior of a liver region without implementing a more complex and deep architecture of neural layers.

The contour feature map of a fully supervised network (i.e., using ground-truth contour supervision; AutoCENet+FC) was activated within overall contour regions (left box in Fig. 7). The figure illustrates that even with fully supervised training, the network failed to extract full-contour features accurately. Conversely, in the self-supervised network, the contour feature map was activated in the sparse regions (right box in Fig. 7). The sparse contour feature map acted as an implicit attention such that the network can concentrate more on the accurate delineation of boundary regions. By employing the self-supervised contour learning, the network demonstrated an improvement in the final segmentation. Figure 7 illustrates the final output prediction of AutoNet and the following two networks: AutoCENet and AutoCENet+FC. The self-supervised contour responses did not correspond to the initial, weak contours from AutoNet (i.e., the initial sparse contour supervision starts from the weak parts of AutoNet results). A strong indication is that the self-supervised contour feature guides the network to better delineate object contours rather than learning the misclassified counterparts, as illustrated in Fig. 7. That is, the response of the contour feature successively

changes pertaining to the current output prediction, which acts as implicit attention for the network. Note that the contour features are not complementary features that are to be merged for the final output prediction.

E. Multiple N-Fold Validation

Previous research has thoroughly investigated neural networks in an architectural perspective and verified their performances within individual metrics. However, limited academic research has been conducted to show the performance of generalization. To evaluate the performance of generalization, N-fold cross-validations were demonstrated for the presented networks. Figure 8 illustrates the dice loss for the test images (i.e., 80 images) by training the network using 10%, 30%, 50%, 70%, and 90% of training images out of the 100 images. The N-fold experiments approximately proxy the real-life deep learning problem and show an extremely generalized regularization analysis.

The overall test errors increased in a smaller percentage of training images. The proposed AutoCENet showed the best performance of generalization. AutoCENet did not over-fit the training images when compared to the other networks. The VoxResNet [18] was the second-best out of other state-of-the-art networks. The fair performance of VoxResNet was obtained owing to its auto-context algorithm. The severe errors in DenseVNet [24] were caused by weak representative shape prior, as discussed in the aforementioned evaluations.

The ablation networks of AutoCENet showed comparable performances to the other state-of-the-art networks (Fig. 8b). Among AutoNet variations, AutoNet-R was the worst-performing network indicating that residual shape prior estimation performs an important role in an auto-context algorithm. In the cases of contour variants, full supervision of

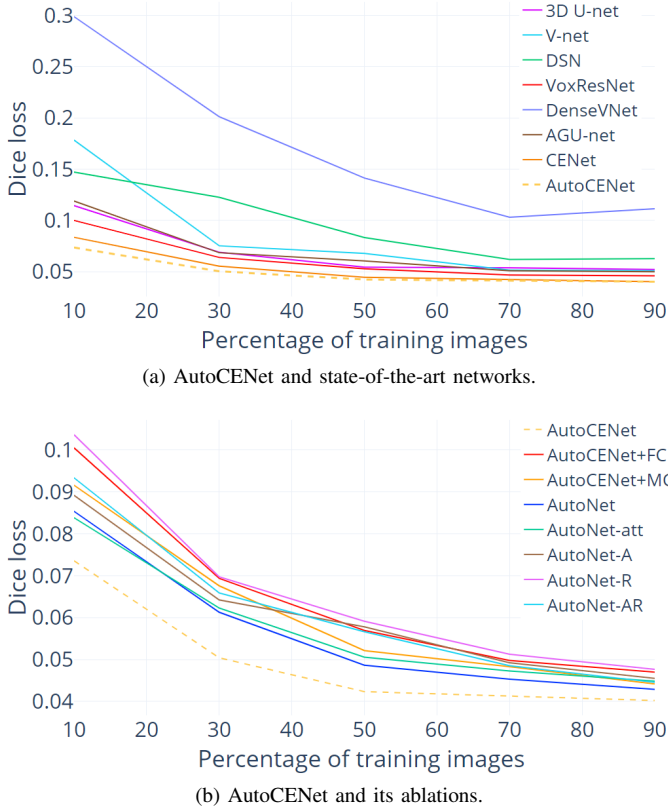


Fig. 8: N-fold cross-validation study of (a) the state-of-the-art networks and (b) the proposed network and its variants. The networks were cross validated by using 10%, 30%, 50%, 70%, and 90% of the images for training out of a total of 100 images. The errors were calculated based on the 80 test images with dice loss.

contour (i.e., AutoCENet+FC) showed the worst performance (Fig. 8b).

V. DISCUSSION

In recent years, the employment of shape priors or neural networks has been the most promising method for the accurate segmentation of a liver. The proposed network avoided using the shape priors because the performance can be highly dependent on the trained shape variations. If the training set is insufficient, the algorithm easily breaks down owing to the quality of the trained prior. Our proposed auto-context algorithm introduced a high-level residual shape prior estimation process that robustly acquired the liver posterior. The embedded liver probability map acted as a post-inference prior, which can be further used for the final accurate classification in an auto-context framework. Consequently, a single-passing auto-context neural network was established without separate classification series, as presented in [32], [44]. The primary underlying principle of the basic auto-context architecture is that the performance of generalization can be achieved by a robust estimation of the overall shape of a liver. In that perspective, high-level residual shape estimation in a lower resolution can successfully achieve the desired task. The architecture suggests that deepening or widening the neural

network is not the only way for complex tasks. Stacking layers sequentially results in difficulty in using parameters effectively and further degrades the regularization of the network. The study of ablation for residual connection demonstrated that the proposed method, which was designed as a task-dependent curriculum, significantly outperformed a simple sequential architecture.

The attention mechanism has demonstrated increasing applicability as a dominant method for modern neural networks. However, the attention mechanism demonstrates a limitation as it is a data-driven algorithm, which indicates that the performance completely relies on the training data distribution. A simple adaptation of the attention mechanism cannot improve the baseline network without explicit guidance. The experimental results showed that the self-attention mechanism presented in AGU-net [25] did not show significant improvement when compared to the basic 3D U-net [16]. It is significantly difficult to create a neural network that focuses more attention on certain features that are useful for the final output. In this study, a self-supervising contour delineation was applied to the intermediate layer that is intended to implicitly guide the network, rather than giving explicit attention, to focus more attention on weak boundary regions that the network has failed to accurately delineate. The self-supervising mechanism was successfully embedded in the network and it improved the final accuracy without any extra false positives.

VI. CONCLUSION

The accurate segmentation of a liver is still a challenging task. Although deep learning demonstrates increasing applications, the lack of annotated medical image data results in difficulty in successfully deploying CNNs in the clinics. Therefore, improving generalization performance is one of the most important tasks for utilizing CNN. In this study, a CNN for liver segmentation was proposed to minimize generalization errors based on the human-designed curriculum (i.e., auto-context). The proposed method minimized the error between training and test images more than any other modern neural networks. In addition, the contour scheme was successfully employed in the network by introducing a self-supervising metric. Instead of exploiting the entire ground-truth contour or self-attention, sparse contours were trained explicitly so that the network can focus on its failures. Based on the experimental results, it was identified that the proposed method performed a significant role in improving accuracy. The newly presented multiple N-fold cross-validation studies also demonstrated the practical applicability of the networks in actual clinics.

REFERENCES

- [1] B. Van Ginneken, C. M. Schaefer-Prokop, and M. Prokop, "Computer-aided diagnosis: how to move from the laboratory to the clinic," *Radiology*, vol. 261, no. 3, pp. 719–732, 2011.
- [2] R. D. Howe and Y. Matsuoka, "Robotics for surgery," *Annual review of biomedical engineering*, vol. 1, no. 1, pp. 211–240, 1999.
- [3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

- [4] K. Suzuki, R. Kohlbrenner, M. L. Epstein, A. M. Obajuluwa, J. Xu, and M. Hori, "Computer-aided measurement of liver volumes in ct by means of geodesic active contour segmentation coupled with level-set algorithms," *Medical physics*, vol. 37, no. 5, pp. 2159–2166, 2010.
- [5] J. Lee, N. Kim, H. Lee, J. B. Seo, H. J. Won, Y. M. Shin, Y. G. Shin, and S.-H. Kim, "Efficient liver segmentation using a level-set method with optimal detection of the initial liver boundary from level-set speed images," *Computer Methods and Programs iBiomedicine*, vol. 88, no. 1, pp. 26–38, 2007.
- [6] X. Zhang, J. Tian, K. Deng, Y. Wu, and X. Li, "Automatic liver segmentation using a statistical shape model with optimal surface detection," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2622–2626, 2010.
- [7] H. Ling, S. K. Zhou, Y. Zheng, B. Georgescu, M. Suehling, and D. Comaniciu, "Hierarchical, learning-based automatic liver segmentation," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.
- [8] T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes *et al.*, "Comparison and evaluation of methods for liver segmentation from ct datasets," *IEEE transactions on medical imaging*, vol. 28, no. 8, pp. 1251–1265, 2009.
- [9] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," in *Proceedings of IEEE international conference on computer vision*. IEEE, 1995, pp. 694–699.
- [10] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001.
- [11] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations," *Journal of computational physics*, vol. 79, no. 1, pp. 12–49, 1988.
- [12] T. Heimann, H. Meinzer, and I. Wolf, "A statistical deformable model for the segmentation of liver ct volumes using extended training data," *Proc. MICCAI Work*, pp. 161–166, 2007.
- [13] A. Wimmer, G. Soza, and J. Hornegger, "A generic probabilistic active shape model for organ segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2009, pp. 26–33.
- [14] E. van Rikxoort, Y. Arzhaeva, and B. van Ginneken, "Automatic segmentation of the liver in computed tomography scans with voxel classification and atlas matching," in *Proceedings of the MICCAI Workshop*, vol. 3. Citeseer, 2007, pp. 101–108.
- [15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [16] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432.
- [17] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 565–571.
- [18] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images," *NeuroImage*, 2017.
- [19] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, "Dcan: Deep contour-aware networks for object instance segmentation from histology images," *Medical image analysis*, vol. 36, pp. 135–146, 2017.
- [20] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.
- [21] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.
- [22] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3d deeply supervised network for automated segmentation of volumetric medical images," *Medical image analysis*, vol. 41, pp. 40–54, 2017.
- [23] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. de Marvao, T. Dawes, D. P. O'Regan *et al.*, "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37, no. 2, pp. 384–395, 2018.
- [24] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Automatic multi-organ segmentation on abdominal ct with dense v-networks," *IEEE Transactions on Medical Imaging*, 2018.
- [25] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.
- [26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [27] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [28] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.
- [29] G. Cai, Y. Wang, L. He, and M. Zhou, "Unsupervised domain adaptation with adversarial residual transform networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.
- [30] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," *arXiv preprint arXiv:1502.02791*, 2015.
- [31] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017.
- [32] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3d brain image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1744–1757, 2010.
- [33] M. Chung, J. Lee, M. Lee, J. Lee, and Y.-G. Shin, "Deeply self-supervised contour embedded neural network applied to liver segmentation," *arXiv preprint arXiv:1808.00739*, 2018.
- [34] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.
- [35] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, no. 2, 2017, p. 3.
- [36] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.
- [38] Z. Tu, "Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1589–1596.
- [39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [40] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.
- [41] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint*, pp. 1610–02357, 2017.
- [42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [44] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging," *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2319–2330, 2017.
- [45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [46] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.
- [47] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.