ECSU-Net: An Embedded Clustering Sliced U-Net Coupled With Fusing Strategy for Efficient Intervertebral Disc Segmentation and Classification

Anam Nazir, Muhammad Nadeem Cheema, Bin Sheng, *Member, IEEE*, Ping Li, *Member, IEEE*, Huating Li, Guangtao Xue, *Member, IEEE*, Jing Qin, *Member, IEEE*, Jinman Kim, *Member, IEEE*, and David Dagan Feng, *Life Fellow, IEEE*

***Abstract*— Automatic vertebra segmentation from computed tomography (CT) image is the very first and a decisive stage in vertebra analysis for computer-based spinal diagnosis and therapy support system. However, automatic segmentation of vertebra remains challenging due to several reasons, including anatomic complexity of spine, unclear boundaries of the vertebrae associated with spongy and soft bones. Based on 2D U-Net, we have proposed an Embedded Clustering Sliced U-Net (ECSU-Net). ECSU-Net comprises of three modules named segmentation, intervertebral disc extraction (IDE) and fusion. The segmentation module follows an instance embedding clustering approach, where our three sliced sub-nets use axis of CT images to generate a coarse 2D segmentation along with embedding space with the same size of the input slices. Our IDE module is designed to classify vertebra and find the inter-space between two slices of segmented spine. Our fusion module takes the coarse segmentation (2D) and outputs the refined 3D results of vertebra. A novel adaptive discriminative loss (ADL) function is introduced to train the embedding space for clustering. In the fusion strategy, three modules are integrated via a learnable weight control component, which adaptively sets their contribution. We have evaluated classical and deep learning methods on Spineweb dataset-2. ECSU-Net has provided comparable performance to previous neural network based algorithms achieving the best segmentation dice score of 95.60% and classification accuracy of 96.20%, while taking less time and computation resources.**

***Index Terms*— Vertebra segmentation, vertebra classification, computed tomography (CT) images, image fusion, computer-based therapy support system, 2-dimensional U-Net.**

Manuscript received March 31, 2021; revised September 16, 2021; accepted December 9, 2021. Date of publication December 24, 2021; date of current version January 4, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62077037 and Grant 61872241, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102, in part by the Science and Technology Commission of Shanghai Municipality under Grant 18410750700 and Grant 17411952600, and in part by the Project of Shanghai Municipal Health Commission under Grant 2018ZHYL0230. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiangqian Wu. *(Corresponding authors: Bin Sheng; Huating Li.)*

Anam Nazir, Bin Sheng, and Guangtao Xue are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shengbin@sjtu.edu.cn).

Muhammad Nadeem Cheema is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 43600, Pakistan (e-mail: itscheema786@yahoo.com).

Ping Li is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: p.li@polyu.edu.hk).

Huating Li is with the Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China (e-mail: huarting99@sjtu.edu.cn).

Jing Qin is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong (e-mail: harry.qin@polyu.edu.hk).

Jinman Kim and David Dagan Feng are with the Biomedical and Multimedia Information Technology Research Group, School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: dagan.feng@sydney.edu.au).

Digital Object Identifier 10.1109/TIP.2021.3136619

I. Introduction

LOWER back pain (LBP) caused by spinal disorders is reported as a frequent cause for clinical visits [1], [2]. The automatic image segmentation [3]–[5] of the spine obtained from a computed tomography (CT) image is significant in diagnosing spine disorders such as fracture detection, intervertebral disc pathology, and spinal trauma patients [6]. A computer-based spinal therapy support system uses CT images of the patients to extract relevant information by segmentation of the vertebra and create a 3D model [7]. Using these models the surgeon can more thoroughly assess the situation by exploring views of patient's spinal anatomy from different angles and depth [8]–[10]. The bone structures have high contrast in today's medical imaging modalities such as CT scans which provide high resolutions images with numerous quantities. The vertebrae segmentation is still considered a challenging task due to many difficulties like the unclear boundaries of the vertebrae associated to spongy and soft bones, the abnormal spine curves and complex structure of the vertebra [11]–[13]. Moreover, such a large quantity of high resolution images cannot be examined manually. Automated image segmentation could increase precision by eliminating the subjectivity of the clinician [14].

Various techniques have been proposed to overcome the abovementioned shortcomings related to vertebra segmentation in last decade [15]. Traditional segmentation algorithms fully exploit the density, gradient and local similarity information, together with clustering, histogram or graph theory algorithms. Recent spine segmentation research can be categorized into two main approaches: free estimation methods and trainable methods. Free estimation methods do not require an explicit

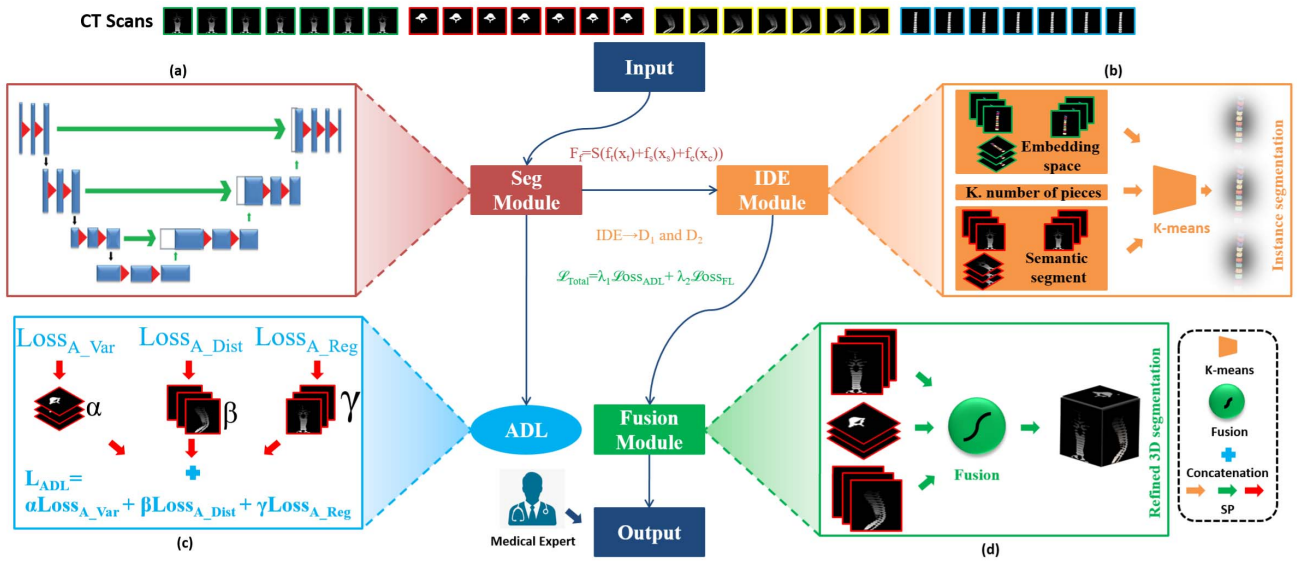

Fig. 1. The overall work flow of proposed approach. (a)-(d) are three modules and calculated loss of our proposed ECSU-Net i.e a) is the segmentation module, b) is the intervertebral disc extraction (IDE) module, c) is the adaptive discriminative loss (ADL), and d) is the fusion module.

model for segmentation and include the followings: classical region growing, watershed, active contours, and graph-cut methods [16], [17]. However, the trainable methods impose a central assumption that the structures of interest organs have repetitive geometry. Therefore, we can utilize the repetitive geometry into a probabilistic representation aimed toward explaining the variation in the shape of the organ and utilize the extracted information for segmentation task. All of these methods require expert human intervention and the manual adjustment of the parameter settings at various distinct steps [8].

Other techniques such as [18] has been proposed which utilized watershed segmentation and directed graph search to locate the vertebral body surfaces whereas identification and segmentation of the vertebrae were not carried out. A 3D deformable model through mesh adaptation was used by [15] for vertebra segmentation but its performance depends on tremendous parameter setup. Chen *et al.* [19] localized and segmented 3D intervertebral discs (IVDs) from MR images. Their approach accomplished disc localization task by estimating the image displacement and classification which were done using a data-driven approach. In another approach [20], a 3D U-Net based segmentation of IVDs was proposed from multi-modality MRI images, where the centers of intervertebral discs were first localized from all the spine samples and then training of network was carried out using the cropped small volumes centered at the localized intervertebral discs. A recent method named S^3egANet [21] proposed to solve the high variety and variability of complex 3D spinal structures through a multi-modality autoencoder along with a cross-modality voxel fusion module to incorporate comprehensive spatial information from multi-modality MRI images and achieved state-of-the-art performance. In the SpineParseNet [22] approach, the authors achieved automated spine parsing for volumetric MR image using a 3D graph convolutional segmentation network for coarse segmentation and a 2D residual U-Net for 2D segmentation refinement. For detecting and labeling vertebrae shapes, Glocker *et al.* [23] implemented a trained model with supervised classification forest; however, this method needed an appropriate feature and require a prior knowledge of the spine shape, therefore, considered as impractical for general and varying image data.

Automatic spine segmentation has challenges associated with it. Some of these challenges are attributed to the anatomic complexity of the spine (33 vertebrae, 23 intervertebral discs, spinal cord, and connecting ribs, etc.), image noise (all real-life data and CT images contain noise), low intensity (in spongy bones and softer bones) [24], and high complexity due to components like 3D convolution or high resolutional 3D images. The majority of spine CT scans of the chest, abdomen or neck cover only part of the spine. Segmentation and classification should therefore not rely on the visibility of certain vertebrae from a single view/axis [25]. A generic vertebra segmentation algorithm therefore needs to be robust with respect to different image resolutions and different coverages of the spine with respect to various axis such as transverse, sagittal and coronal views to provide 3D model of segmented vertebras.

Inspired by the remarkable performance of [26], which shows another direction to tackle instance segmentation by employing clustering on the achieved output, we propose a new 3D segmentation network named an Embedded Clustering Sliced U-Net (ECSU-Net), which can alleviate the intricacy issues related to high resolutional 3D images at input and model complexity problems related to anatomy of vertebra as shown in Fig. 1. ECSU-Net comprises of three modules i.e segmentation, intervertebral disc extraction and fusion modules [27]. In the first module, we exploit 2D U-Net by introducing three sliced sub-nets to process transverse, sagittal, and coronal axis of input image to provide scalability by processing the high resolution 3D CT scans to three subnets in the same way as to process 2D images. In the second module, we introduce an intervertebral disc extraction (IDE) module

to classify and tag the segmented vertebras for measuring distances of adjacent vertebras. For the third module, we make use of fusion concept in the form of concatenating strategy to combine results obtained from the three segmentation subnets (transverse, sagittal, and coronal slices) for refined 3D segmentation results [28].

ECSU-Net improved the general 2D U-Net architecture to incorporate the high resolution 3D input images in a similar way as to process 2D images providing scalability as a novel feature [29]. Our hybrid strategy based approach can be useful for many other similar problems that should provide efficient yet scalable solutions in 3D medical image processing where the target objects usually have to process the high resolution input images. The proposed segmentation network provided comparable performance to previous neural network based algorithms on Spineweb dataset-2 [11], while taking less time and computation resources. Clinically, the proposed model can benefit many automatic spine analysis problems that use high resolution 3D CT scans. The ECSU-Net concurrently performs segmentation of a vertebra as well as classification that whether the vertebra is correctly segmented in the image. Following are the key contributions of this study.

- A novel 3D ECSU-Net as an instance embedding clustering approach is proposed by jointly learning from a hybrid strategy of three modules i.e. segmentation, intervertebral disc extraction and fusion for effective and efficient coarse-to-fine vertebra segmentation and classification.
- The key feature of segmentation module is the improvement in general 2D U-Net to provide scalability by introducing three sliced sub-nets to process high resolution 3D CT scans at input in the same way as to process 2D images.
- The (IDE) module finely classify and tag the segmented intervertebral disc to measure distances of adjacent vertebras to help surgeons for computer-based spinal diagnosis and therapy support system.
- We introduce a fusion strategy, which combines the contribution of features from sliced input data with a learnable weight control module for leveraging the advantages of different axis to preserve spatial details for accurate vertebra classification.

We quantified the similarity in terms of success and failure rates for vertebra segmentation and intervertebral disc classification using average precision and average recall metrics. In according to the above three advantages, our method can compute very efficiently for finer classification of vertebra's structures which can create improved segmentation results.

II. Method

Inspired by [6], we propose a three-stage segmentation and classification model based on 2D U-Net [30], our segmentation model is 3D and comparatively low in cost. Different from [6], our model exploits all the transverse, sagittal and coronal slices of CT images and can be trained with insufficient data and few computational resources (see Fig. 2 and Algorithm 1).

Algorithm 1 ECSU-Net Based Intervertebral Disc Segmentation

Input: 3D_Vertebra_CT-Data
Output: Segmented_Intervertebral_Discs
Initialize_Weights(f_t, f_s, f_c);
Split3Slices$\big(Transverse(f_t), Sagittal(f_s), Coronal(f_c)\big)$;
for $Seg\_Module_{i \leftarrow n}$ && $IDE\_Module_{i \leftarrow n}$ && $Fusion\_Module_{i \leftarrow n}$ **do**
Seg_Module $\leftarrow$ Apply_uNet(f_t, f_s, f_c);
Seg_Loss=$\alpha.L_{VC} + \beta.L_{DT} + \gamma.L_{RG}$;
$\mathscr{L}_{ADL} = \alpha.Loss_{A\_Var} + \beta.Loss_{A\_Dist} + \gamma.Loss_{A\_Reg}$
IDE_Module $\leftarrow IDE$ (Seg_Module)
K-means (IDE Seg_Module);
Fusion_Module$\leftarrow$ $\big($IDE Seg_Module, $(f_t, f_s, f_c)\big)$;
$d_1, d_2 \leftarrow$Calculate_Distance (IDE);
$\mathscr{L}oss_{FL} = \Big\{ ||\sigma - \sigma^v||^2, v = 0, \sigma \leq \sigma^v, or v \epsilon \{1, 2, 3\}$
end for
repeat
Cal_Total_Loss$\leftarrow$ $(\mathscr{L}_{\lambda_1 \mathscr{L}oss_{ADL} + \lambda_2 \mathscr{L}oss_{FL}})$;
Move to next slice;
until Refined_3D_Seg(Segmented_Vertebra+Total_Loss);

A. U-Net

U-Net [30] is a light-weight segmentation network and is especially popular among medical image analysis tasks. The architecture of U-Net is given by [30], consisting of 4 down-sampling modules, one bottom module and 4 upsampling modules. The down-sampling module includes two 3×3 convolutions, a 2×2 down pooling and the upsampling module includes two 3×3 convolutions and a 2×2 up pooling. The down-sampling module's outputs are concatenated with up-sampling modules with same feature map size. And the bottom module only consists of two 3×3 convolutions. Since U-Net is a fully convolutional network, it can handle images with any sizes as long as the image size keeps unchanged after pooling and up-pooling process, i.e. lengths of both side are multiples of 16.

B. Segmentation Module

Segmentation module is applied on three subnets for which the provided input is divided into three branches of transverse, sagittal and coronal slices with identical structures. Segmenting and classifying vertebras by focusing on only one dimension is not always accurate. Experienced radiologists usually compare the current vertebra with its adjacent vertebrae, utilizing the continuity of the spine in its 3D form, i.e. how a normal vertebra in a 3D environment is similar in the shape and signal intensity to its adjacent normal vertebrae. Imitating the radiologists, the segmentation module is designed to take three views of vertebra as input for concrete accuracy and efficacy, i.e transverse, sagittal and coronal views (denote as $x_t; x_s; x_c$ respectively), and leverage a three-slice architecture. We denote the three branches of input with their corresponding three feature as $(x_t; x_s; x_c)$: $(f_t; f_s; f_c)$ respectively. The features obtained by three sliced branches

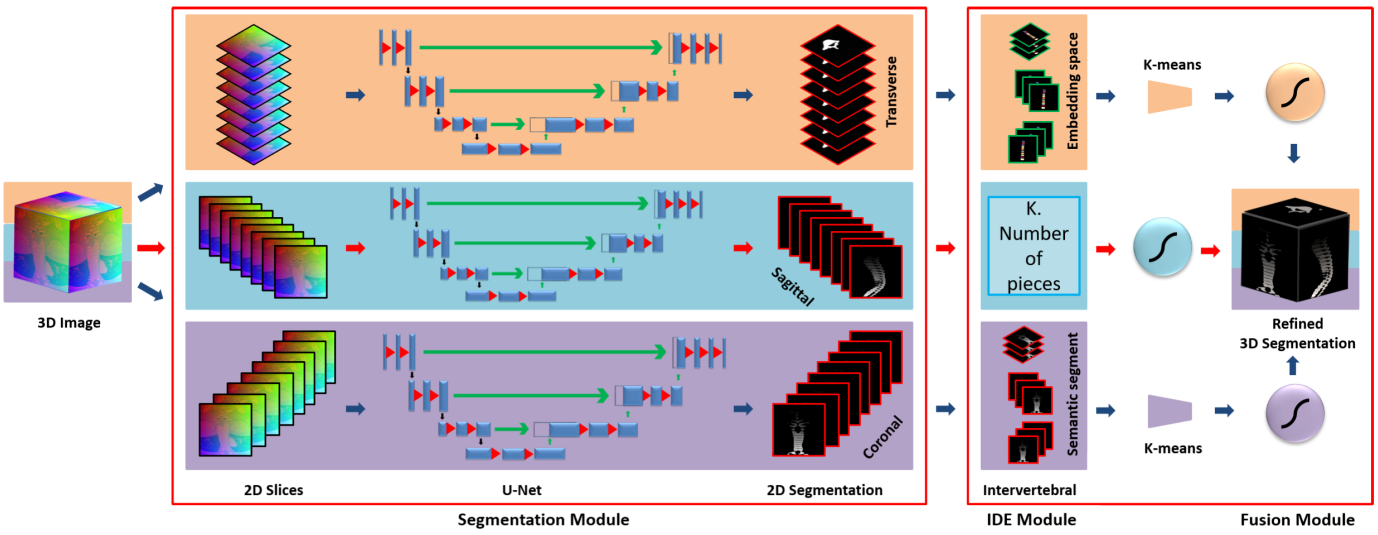

Fig. 2. The ESCU-Net, working design along with detailed visualization of three modules i.e. segmentation, intervertebral disc extraction (IDE) and fusion module.

are input to the rest convolution layers of the stream S after fusing them by pixel-wise addition. The final output features of this stream F_f is expressed as:

$$F_f = S\big(f_t(x_t) + f_s(x_s) + f_c(x_c)\big) \tag{1}$$

The output of the network also includes a number k which represents how main pieces of spine are there in this CT image. The clustering is applied to the outputted embedding space, where we use the K-means algorithm to split the whole 512×512 pixels into $k+1$ categories.

Segmentation network is composed of three independent sub-nets, which are based on 2D U-Net [30]. Compared to 3D convolution based networks, the 2D model is significantly more efficient, smaller in size and easier to train. Since typical CT scan images have high resolution, hence implementing three 2D network require reasonably less computational power than a 3D convolution model. Besides computational consideration, 2D slices based method can fully exploit the insufficient data. Lack of training data is a common issue in biomedical image analysis and it is more severe in case of 3D models, as the annotation of 3D images is expensive and laborious. Our 2D approach can partly alleviate this issue by splitting 3D voxels into multiple 2D images (see Fig. 3).

The input channels (W) after spatial embedding of pixel I_{ij} in a image with size (H, P) can be formalized as: $W_{ij}^{(1)} = I_{ij}$, $W_{ij}^{(2)} = \frac{i}{H}$, and $W_{ij}^{(3)} = \frac{j}{P}$. The sagittal and coronal networks output 20-channel softmax results which stands for the probability for a pixel to be a part of background or one of the 19 vertebrae. Each of three sub-nets is trained on all the transverse, sagittal and coronal slices of CT scans and produces a rough 3D segmentation of the whole image by stacking all the 2D outputs.

1) Adaptive Discriminative Loss (ADL): Our novel idea of ADL found best suited for instance-level segmentation tasks which is inspired by [31]. The core idea of our designed loss is differentiated from the previous losses is in having a special

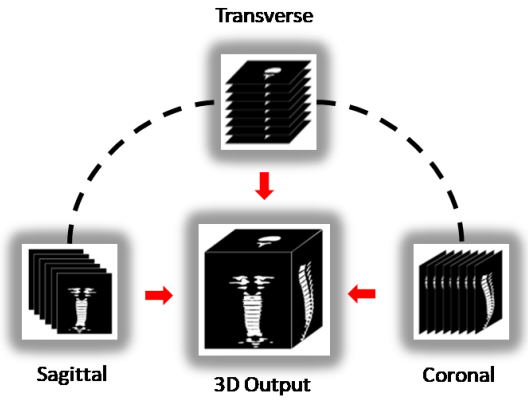

Fig. 3. Illustration to show the transverse, sagittal, and coronal slices of input image to provide scalability for processing the high resolution 3D CT scans.

design to handle losses of three subnets with respect to three axis of input CT image.

$$Loss_{DL} = L = \alpha.L_{VC} + \beta.L_{DT} + \gamma.L_{RG} \tag{2}$$

Above is the general form of discriminative loss function. where the LVC denotes variance loss ($Loss_{A\_Var}$), LDT is representing distance loss ($Loss_{A\_Dist}$) and LRG is showing the regression loss ($Loss_{A\_Reg}$). For generating embedding space of each slice, we have designed a differentiable slice-wise pixel embedding function that maps each pixel in an input CT image to a point in n (for our case $n = 3$) dimensional feature space.

The perception behind our adaptive loss function is that embedding with the same label (same slice) should end up close together, while embedding with a different label (different slices) should end up far apart. We now formulate our discriminative loss in terms of repelling and attracting forces between and within clusters. The ADL divides loss item into three terms: (1) Adaptive variance (A_Var): regraded as the attracting force to draw embedding towards the mean value of

the cluster for transverse, sagittal and coronal slices. (2) Adaptive distance (A_Dist): regarded as the repelling force that pushes the cluster away from each other for transverse, sagittal and coronal slices (3) Adaptive regularization (A_Reg): a small attracting force to draw all clusters embedding to the center of the same embedding cluster of transverse, sagittal and coronal slices.

For our setup, a cluster is defined as a group of pixel embedding sharing the same label with the same slice, e.g. pixels belonging to the same instance and slice. We use the following definitions: C is the number of clusters in the gold standard ground truth, N_e is the number of elements in cluster c, x_i (x_t, x_s, x_c) is an embedding for three slices transverse, sagittal and coronal, μ_e $(\mu_{et}, \mu_{es}, \mu_{ec})$ is the mean embedding of cluster e (the cluster center) for the three slices (e_t, e_s, e_c). The term $||.||$ is representing three distances D_1, D_2 and D_3 for three losses $Loss_{A\_Var}$, $Loss_{A\_Dist}$, $Loss_{A\_Reg}$, and $[x]_+ = max(0; x)$ denotes the hinge. ϵ_v and ϵ_d are respectively the margins for the variance and distance loss. Now the ADL can be written as follows:

$$Loss_{A\_Var} = \frac{1}{C}\sum_{e=1}^{C}\frac{1}{N_e}\Pi_{i=1}^{N_e}\left[||\mu_e - (x_T, x_S, x_C)|| - \epsilon_v\right]^2 + \tag{3}$$

$$Loss_{A\_Dist} = \frac{1}{C(C-1)}\sum_{eq1}^{C}\Pi_{e_S=1}^{C}\Pi_{e_C=1}^{C} \left[2\epsilon_v - ||\mu_{eT} - \mu_{eS} - \mu_{eC}||\right]^2 + \tag{4}$$

$$\therefore eq1 = e_T = 1_{e_T \neq e_S, e_C \neq e_T, e_S \neq e_C}\alpha$$

$$Loss_{A\_Reg} = \frac{1}{C}\sum_{e=1}^{C}||\mu_{e_{(T,S,C)}}|| \tag{5}$$

$$\mathscr{L}_{ADL} = \alpha.Loss_{A\_Var} + \beta.Loss_{A\_Dist} + \gamma.Loss_{A\_Reg} \tag{6}$$

In our case we set $\alpha = \beta = 1$ and $\gamma = 0.0001$. The loss is minimized by stochastic gradient descent. Thus, the classification of vertebrae in the output of the same class are clustered together while those of different classes are repelled apart.

C. Intervertebral Disc Extraction (IDE) Module

After acquiring the embedding space of the input RGB image and get each instance segment of three corresponding slices with the K-means algorithm. we present a searching line algorithm along the y-axis of the embedding space in order to find out the inter-space between two-pieces of spine. Basically, what we done is to find each pixels along each y-axis to see whether its upper and lower area are belong to different instance segment while its own location is in the background. We find this simple algorithm works well when the quality of the outputted embedding space is good enough. We have shown the efficacy of presented idea in the results section with qualitative analysis. However, vertebra classification and tagging require very different features, and both tasks are characterized by high intra-class and high inter-class similarity between the adjacent vertebras. For resolving this issue we have presented a novel concept of ADL for enabling the pixels belong to the same object as close to each other as possible (D_1) on the embedding space while for the pixels belong to different objects, the distance should be as large as possible (D_2). Our IDE module learns to identify vertebra through comparing and contrasting interspace distance between adjacent vertebra using intra class and inter-class distance to deliver fine-grained tagging results as shown in Fig. 4.

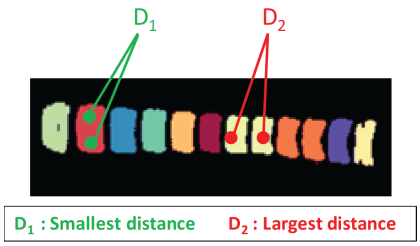

Fig. 4. Intervertebral distance calculation and tagging concept. D1 is the distance between pixel on the same segment of a spine whereas D2 represents distance on the different segments.

D. Fusion Module

Since three sub-nets in segmentation networks may produce different results on some pixels which is hard for them to identify individually, the best idea is to take all their results into consideration and fuse the outputs into one. As, splitting data into 2D slices may omit the local information of reduced dimension, here our fusion strategy can significantly recover the skipped information. Our fusion module takes input in the form of coarse segmentation results from IDE module and outputs the refined 3D segmentation [32]. Degradation in our case is of two types. First one is additive noise due to image denoising and second one is effects of convolution process in the form of blind convolution. To remove denoising effects we have applied mean fusion which is simply averaging the obtained image over multiple realizations with respect to time.

In the fusion module, we have introduce a weight control term which generates adaptive weight h to integrate and control features from three slices f_t, f_s, f_c. An obvious strategy for ignoring incompletely visible vertebrae in the segmentation process would be to train the network only with examples of fully visible vertebrae. However, the proposed scheme requires segmentation of vertebral slices with respect to three axis. If a slice in one plane is incompletely visible, there is possibility of its complete visibility in another axis. Hence inspired by the [25], we therefore choose to incorporate a classification component in our fusion network that classifies each segmented vertebra as correctly segmented or incorrectly segmented. The output comprises of single value in $[0, 1]$, which specifies the possibility that the vertebra is completely visible in the input image with respect to all axis. The obtained segmented images consists of true/segmented part and false/degradation part, which we have restored by fusion strategies. A function ϕ_h is employed to transform fused feature of f_t, f_s, and f_c to weight h_t, h_s, h_c with 3D as follows:

$$h = \phi_h(f_t, f_s, f_c) \tag{7}$$

The fused output f_o is obtained by concatenating $h_t \times f_t, h_s \times f_s$ and $h_c \times f_c$. Total weight loss (product) of fused module can be expressed as:

$$\mathscr{L}oss_{FL} = \Pi\left\{||\sigma - \sigma^{\upsilon}||^2, \upsilon = 0, \sigma \leq \sigma^{\upsilon},\ or\ \upsilon\epsilon\{1,2,3\}, \sigma \geq \frac{1}{\sigma^{\upsilon}}, otherwise\right\} \quad (8)$$

where σ is the weight ratio $\frac{h(t_i,s_i,c_i)}{f(t_i,s_i,c_i)}$ we set a parameter $\sigma^{\upsilon}(\sigma^{\upsilon} > 3)$, which controls the bound of σ. Here we define the total loss of proposed method as:

$$\mathscr{L}_{Total} = \lambda_1 \mathscr{L}oss_{ADL} + \lambda_2 \mathscr{L}oss_{FL} \quad (9)$$

1) Mean Fusion: Let s_a, s_s, s_c denote the 2D segmentation scores on axial, sagittal, and coronal slices of one voxel. Mean fusion is simply averaging the three scores:

$$S_{mean} = \frac{S_a + S_s + S_c}{3} \quad (10)$$

If the mean fusion is implemented on classification results, only sagittal and coronal results are used:

$$S_{mean} = \frac{S_s + S_c}{2} \quad (11)$$

The final results O are:

$$O_{mean} = arg_j max S_{mean,j} \quad (12)$$

Blind deconvolution is an ill posed problem for one single image, which can be solved by accruing multiple acquisitions of the same object. To remove degradation due to blind convolution, for an input image I, with W_i channel in three dimensional sagittal, coronal and transverse plane (x_s, x_t, x_c), the noise n added by blind deconvolution D is defined in terms of:

$$\int [I \times W_i](x_s, x_t, x_c) + n_i(x_s, x_t, x_c) = D_i(x_s, x_t, x_c) \quad (13)$$

we have regularized by point spread function (PSF) regularization which is defined in our case for multiple channels D_1 to D_n as:

$$D_1 = \int I \times W_1, D_2 = I \times W_2, \ldots, D_n = I \times C_n \quad (14)$$

hence we have regularized the terms as:

$$I \times W_1 \times W_2 - W_2 \times W_1 \times I \quad (15)$$

$$D_1 \times W_2 = W_1 \times D_2 \quad (16)$$

Now, after regularization R we have obtained

$$R(W_a) = \frac{1}{2} \sum_{1 \leq a,b \leq i} ||D_a \times W_b - D_b \times W_a||^2 \quad (17)$$

for an additive constrain of $0 \leq W_a(x_s, x_t, x_c) \leq 1 \forall (x_s, x_t, x_c, a)$.

After regularization, vote fusion is implemented for segmentation results with binarization, i.e. it is applied on the binary outputs rather than score so it can be seen as the plural vote result of three input channels.

$$O_{\cdot} = arg_j max S_{\cdot,j} \quad (18)$$

$$O_{1_{vote}} = \begin{cases} 1, & \text{If } o_a + o_s + o_c \geq 1 \\ 0, & otherwise \end{cases} \quad (19)$$

$$O_{2_{vote}} = \begin{cases} 1, & \text{If } o_a + o_s + o_c \geq 2 \\ 0, & otherwise \end{cases} \quad (20)$$

III. Evaluations and Discussions

A. Dataset Details

We evaluate various vertebra segmentation methods, including classical and deep-learning method, on Spineweb dataset-2 [33], which contains spine CT scans of 10 young adults with annotation in .nii format. The CT scans are 3D 12-bit (ranging from 0-4095) gray images. We downsample the data to 8-bit by dividing the pixel values by 16. We also converted the .nii format files to hdf5 files for higher reading speed. We sliced the CT images on three directions, which include over 5000 2D slices for each direction. We split the data set into three sets: train, valid and test. The first 9 scans were used for the training and validation phase and the left one were used for testing. We randomly sampled 10% of slices of the first 9 scans as validation set and 90% as training set. Our training set contains about 4000 slices, validation set contains about 500 slices and test set contains about 500 slices, which slightly vary among different slicing directions. In order to keep the output size unchanged during pooling and up-pooling, we crop the image size to its nearest multiple of 16.

B. Evaluation Parameters

The manually labeled images from the Spineweb dataset-2 [33] of each CT scan are used as ground truth and the results of ECSU-Net segmentation are converted into binary images with the same voxel resolution and image dimensions as the ground truth image. In this study for description, the ECSU-Net segmentation result is indicated by US and the ground truth by GT. We have utilized three types of metrics for quantitative evaluation including similarity metrics (dice coefficient, intersection over union), distance based metrics (average symmetric surface distance), and classical measurements (accuracy, precision, and recall).

1) Dice Coefficient (DC): DC quantifies the degree of the spatial overlap between two binary images to be compared. DC values range between 0-1, where 0 means no overlap and 1 means perfect agreement in compared images. For this research, the DC values are calculated using $\frac{2|US \cap GT|}{|US|+|GT|}$.

2) Intersection Over Union (IoU): We have utilized IoU as a measure to calculate extent of overlap found between ground truth (GT) image and the segmented image (US) obtained from our method.

3) Average Symmetric Surface Distance (ASSD): ASSD determines the measure of border voxels US and GT images.

4) Classical Metrics: We make use of the confusion matrix to achieve classic measurements by calculating four variables: true negative (TN), false negative (FN), true positive (TP), and false positive (FP).

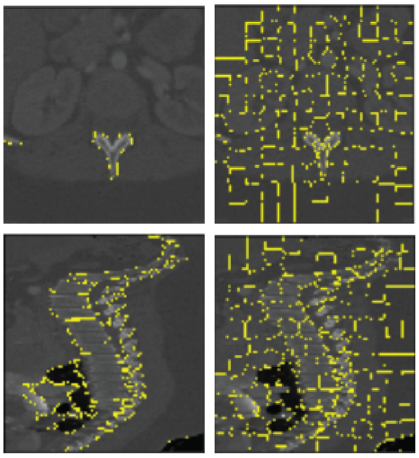

Fig. 5. Left: Felzenszwalb. Right: SLIC. The yellow lines indicates the segmented boundaries.

a) TN: Measures the pixels which are not classified as the vertebra in the ground truth as well as ECSU-Net (correctly detected as background).

b) FN: Pixels classified as the vertebra in the GT image, but not classified as vertebra by the proposed method (falsely detected as background).

c) TP: For this study, TP is a measure of pixels correctly segmented as the vertebra in the ground truth and ECSU-Net segmented images (correctly segmented).

d) FP: FP is the measure of pixels which are not classified as the vertebra in the ground truth, but are classified as the vertebra by ECSU-Net(falsely segmented)

Using these four measures, we have defined precision as $\frac{TP}{TP+FP}$, recall as a measure of true positive rate using $\frac{TP}{TP+FN}$, and accuracy of segmentation as $\frac{TP+TN}{TP+TN+FP+FN}$. Also, we take average of abovementioned metrics.

C. Evaluation Details of Classical Methods

Before presenting results obtained from our model, we have shown the segmentation results obtained from several popular segmentation models, which are detailed below.

- **Felzenszwalb algorithm [34].** The algorithm was proposed by Pedro F. Felzenszwalb in 2004, introducing a graph-based method of image segmentation. It presented a predicate to define the image region's border, and used this predicate to greedily decide how to perform the image segmentation.
- **SLIC [35].** It is an algorithm used to generate superpixels given an image with K-means clustering. The main idea of super-pixel is to abstract the information contained in the image, reducing the computational complexity, obtaining a more robust and generalized representation.

Here we illustrate some sample images segmented by these two algorithms in Fig. 5. Both algorithms tends to over segment the image. SLIC's [35] segmentation presents regions that are more compact, while Felzenszwalb gives bigger segmented regions which are slightly irregular. Moreover, different views also result in varied performance. To solve the over segmentation problem, we used gray scale threshold to classify whether a segmented region belongs to a vertebra.

TABLE I

RESULTS WITH CLASSICAL SEGMENTATION METHODS WITH TRANSVERSE AND SAGITTAL AXIS

Method	Dice (Transverse)	Dice (Sagittal)
Felzenszwalb [34]	0.5863	0.7168
SLIC [35]	0.7165	0.6624

TABLE II

RESULTS FOR FUSED CLASSICAL METHODS A_1 AND A_2 DENOTES FELZ AND SLIC RESPECTIVELY ALONG WITH DICE

Ist Axis	2nd Axis	A1	A2	Dice
Transverse	Sagittal	SLIC	Felz	0.7420
Transverse	Sagittal	SLIC	SLIC	0.7067
Transverse	Sagittal	Felz	SLIC	0.4791
Transverse	Sagittal	Felz	Felz	0.4867
Sagittal	Transverse	Felz	SLIC	0.7193
Sagittal	Transverse	SLIC	Felz	0.5868
Sagittal	Transverse	SLIC	SLIC	0.6944
Sagittal	Transverse	Felz	Felz	0.5640

Considering the two algorithms are designed for 2D images, we also evaluated them along two different axis: transverse and sagittal, results in terms of DC are shown in Table I. As shown, SLIC [35] significantly outperformed Felzenszwalb algorithm [34] along transverse, while Felzenszwalb [34] presented better performance along sagittal. The different performance inspired us to combine their advantages by fusing the results of Felzenszwalb and Huttenlocher [34] and SLIC [35] on the two different axis. The fusing strategy can be stated as follows:

- Perform Felzenszwalb and Huttenlocher [34] or SLIC [35] along the first axis, using gray threshold δ to classify.
- Perform Felzenszwalb and Huttenlocher [34] or SLIC [35] along the second axis. For a region, if it satisfies the threshold requirement, and more than max γ of its pixels is predicted as vertebra by the first axis, label this region as vertebra. If less than min γ of its pixels is predicted as vertebra by the first axis, label this region as background.

In our experiment, we set maximum upper limit to 90, and minimum lower limit to 50, resulting in best performance. The evaluation results are shown in Table II. The best performance is visualized in Fig. 6. As shown, the fusing strategy successfully improved the segmentation performance. Inspired by the different performance along different axis, we introduced a fusing strategy in our proposed model.

D. Quantitative Analysis

1) Quantitative Analysis w.r.t Success and Failure Rates: To show the mean average accuracy of ECSU-Net from a different point of view with respect to transverse, sagittal, and coronal slices of input image, we have carried out experiments to demonstrate the success and failure rates. Following four metrics are used: success rate, failure rate, partial success rate,

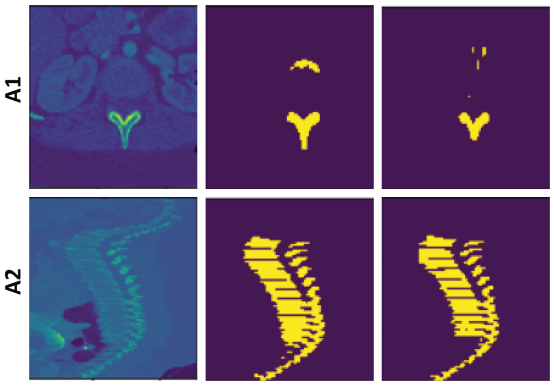

Fig. 6. Left: Original image. Middle: Ground truth. Right: Predicted result. The vertebra is highlighted with yellow. A1 is the algorithm composed on the first axis, and A2 is the algorithm composed on the second axis.

TABLE III

QUANTITATIVE RESULTS IN TERMS OF SUCCESS AND FAILURE RATES ALONG WITH PARTIALLY SUCCESS AND FAILURE RATES (S.R, F.R, P.SR, P.FR RESPECTIVELY). R(S:F) STANDS FOR RATIO OF SUCCESS AND FAILURE RATES OF CLASSIFIED AND TAGGED VERTEBRA

Class	S.R	F.R	P.SR	P.FR	R (S:F) (%)
Class A (Cervical)	61.518	3.083	3.166	31.251	61:3
Class B (Thoracic)	70.916	6.253	5.124	17.666	70:8
Class C (Lumbar)	80.398	2.975	2.953	12.67	82:2

and partial failure rate. The results and statistical analysis of these four metrics are described in Table III. For our case, the success rate is a percentage measured as a total number of successfully detected pixels which are correctly segmented as the vertebra in the ground truth and ECSU-Net segmented images (correctly segmented). Failure rate is a percentage calculated as a number measured pixels, which are not classified as the vertebra in the ground truth as well as ECSU-Net. Partially success rate is metric to show a total number of partially successful measured pixels, which are not classified as the vertebra in the ground truth, but are classified as the vertebra by ECSU-Net (falsely detected as vertebra). Partially failure rate is the number of partially failed pixels classified as the vertebra in the GT image, but not classified as vertebra by the ECSU-Net (falsely detected as background).

2) Training Details and Evaluation of ECSU-Net: In this subsection, we have presented the implementation and evaluation details of ECSU-Net. We trained the three segmentation sub-nets individually with Adam optimizer having learning rate 5×10^{-4}, momentum parameters 0.5 and 0.999. Since the input images are different in size, the batch size is set to 1. All the networks are trained for 10 epochs with Tesla T4 GPU on Google Colab platform. The training of transverse slices took about 40 minutes per epoch and coronal and sagittal slices took 30 minutes per epoch.

a) Individual sub-nets performance: First, we evaluated three 2D U-Net individually. Since the classification outputs of 20-channel coronal and sagittal networks can be easily transformed into 2-channel segmentation results by regarding all 19 vertebrae classes into one class. Besides, in order to overcome the influence of class imbalance (pixels in background class are much more than others), we calculated class-wise mean accuracy, which is the average accuracy on each class. The evaluation results on three sub-nets on test set are shown in Table IV and some inference results are visualized in Fig. 7. Where axis 0 represents segmentation results of transverse plane via first sub-net, axis 1 is showing result obtained from coronal plane with second sub-net while the axis 2 is depicting results achieved via sagittal plane from the third sub-net. With single 2D U-Net, the segmentation results can be quite accurate. It can be seen that the 2D U-Nets with different slicing directions significantly outperform traditional methods. The best segmentation direction is transverse whose dice score can reach 94.61%, because the slices on transverse plane have least noises and are regular in shape. It is hard to find the difference between ground truth and model prediction with human eyes. The best classification direction comes to sagittal network whose mean accuracy is 95.20% and its segmentation results are close to the transverse one.

TABLE IV

TEST SEGMENTATION RESULTS OF OUR PROPOSED ECSU-NET WITH DIFFERENT FUSION STRATEGIES

Method	Accuracy (%)	Precision (%)	Recall (%)	IoU (%)	DC (%)
Transverse only	99.55	93.28	95.98	89.77	94.61
Coronal only	99.34	85.81	98.13	84.43	91.56
Sagittal only	99.53	89.68	98.93	88.82	94.08
1 vote fusion	99.62	96.72	94.47	91.54	95.58
2 vote fusion	99.63	91.95	99.19	91.26	95.43
Mean fusion	99.64	92.07	99.41	91.57	95.60

TABLE V

TEST CLASSIFICATION RESULTS OF ECSU-NET

Method	Accuracy (%)	Mean Accuracy (%)
Transverse only	-	-
Coronal only	96.34	92.53
Sagittal only	91.20	95.03
Mean fusion	93.20	96.34

b) Fusing strategy: We evaluated the fusion strategies in terms of accuracy, precision, recall, IoU and DC parameters. The results are shown in Table IV and Table V, some of the fusion methods achieves higher accuracy while some others have negative effect. For segmentation, the best fusion approach is simple mean fusion, which achieves the best segmentation results among all experiments we have conducted. The other fusion strategies for segmentation also improve the dice score by almost over 1%. For classification, surprisingly, the accuracy of mean fusion outputs is better than single coronal network but worse than sagittal network. Perhaps weighted mean fusion would outperform the current fusion method.

E. Qualitative Results

1) Tagging of Classified Vertebras With Respect to Measured Distance: We have evaluated ECSU-Net by drawing precision-recall curves to qualitatively evaluate its

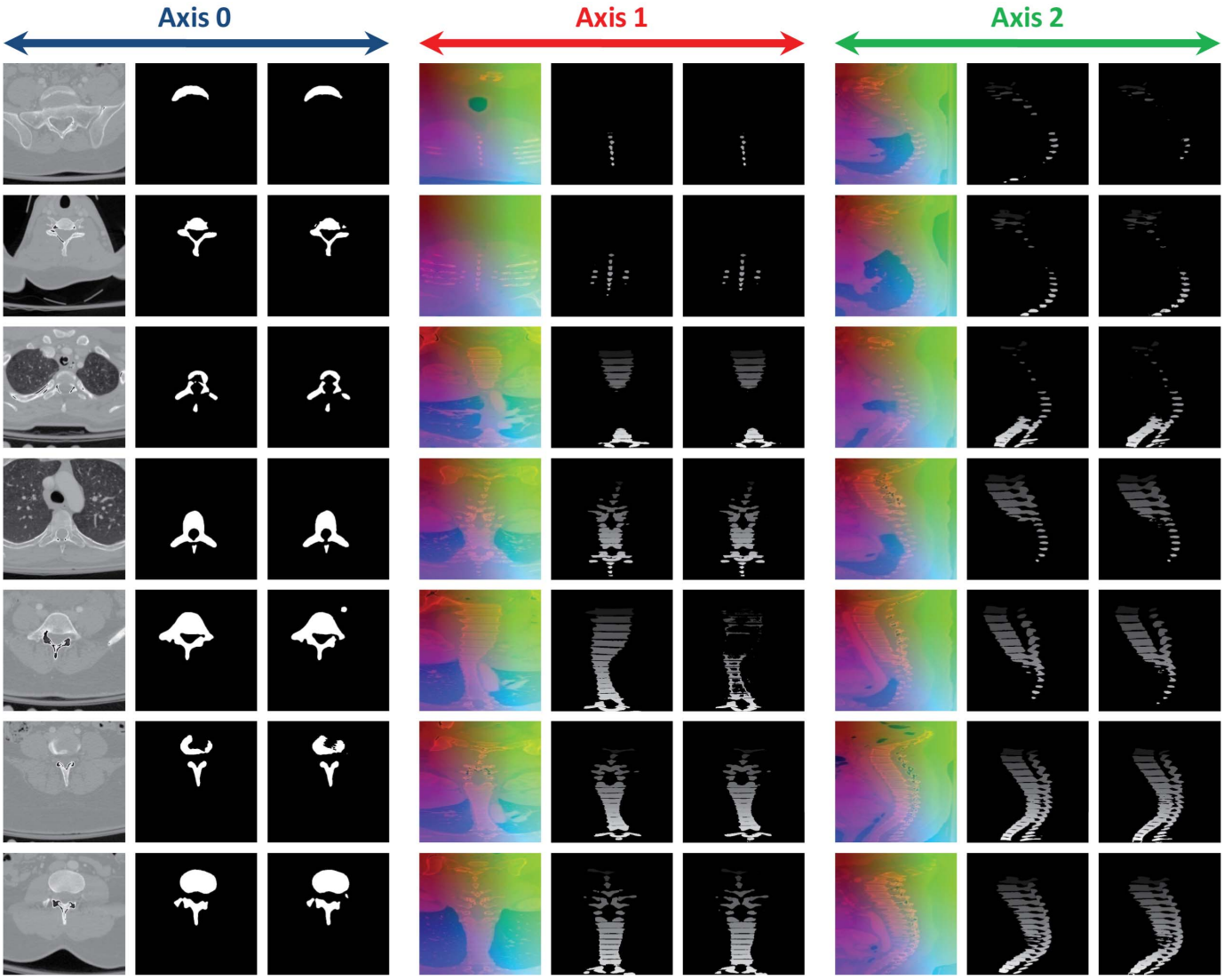

Fig. 7. Visualization of 2D U-Net results: Axis 0 represents segmentation results of transverse plane, axis 1 is showing result obtained from coronal plane while the axis 2 is depicting results achieved using sagittal plane.

performance for tagging the classified vertebra with the measured distances. Our experimental setup draws precision-recall curves for true positive rate and positive predictive value (TPR, PPV) with respect to varying views of slices i.e. transverse, sagittal and coronal slices at specific threshold of IoU value. For analyzing effects of measured distance on tagging the classified vertebra with respect to specific precision-recall values, we have used IoU as a measure to calculate how much overlap found between two regions of adjacent vertebras. IoU permits us to detect the measured distance overlapping to the two adjacent vertebras. The true-positive rate is known as recall while the PPV is termed as precision.

In this work, precision is defined as the ratio of truly tagged vertebra to the total number of vertebra classified by ECSU-Net. Precision values closer to 1.0 means that positive tagging by proposed approach are in fact correct tagging as illustrated in Fig. 8. We have utilized recall for measuring false negative rate. The recall is defined in our scenario as the ratio of truly tagged vertebra to the total number of actual vertebra in an input image as shown in Fig. 9. Recall values near to 1.0 represent almost all vertebra in input CT image are positively tagged by ECSU-Net. We have defined three threshold levels for IoU as follows:

- IoU (%) ($\text{Thershold}_1 = 89.77$): The presented value of IoU shows least overlapping distance (improved precision recall values) between two adjacent vertebras for transverse slice
- IoU (%) ($\text{Thershold}_2 = 84.43$): The presented value of IoU shows least overlapping distance between two adjacent vertebras for coronal slice which means we have obtained improved precision recall values.
- IoU (%) ($\text{Thershold}_3 = 88.82$): The presented value of IoU shows least overlapping distance between two adjacent vertebras with improved precision and recall values for sagittal slice.

F. Ablation Study

We further investigate the individual contribution of the three key architectures in ECSU-Net, i.e. the segmentation,

TABLE VI

ABLATION STUDY RESULTS FOR PROPOSED METHODOLOGY, WHERE SEG.M REPRESENTS SEGMENTATION MODULE, IDEM STANDS FOR INTERVERTEBRAL DISC EXTRACTION MODULE, F. M IS SHOWING FUSION MODULE, T.L REPRESENTS TRIPLET LOSS AND ADL IS SHOWING THE ADAPTIVE DISCRIMINATIVE LOSS

Method	Avg. Accuracy(%)	meanAvg. Accuracy (%)	Avg. Precision (%)	Avg. Recall (%)
Seg.M with MaskRCNN (V_1)	83.43	82.44	85.43	83.49
Seg.M with SegNet (V_2)	84.37	83.44	86.38	85.25
Seg.M with U-Net (V_3)	87.44	86.98	85.18	84.29
Seg.M (V_3) + IDEM + F.M + T.L	86.93	85.43	89.69	88.98
Seg.M (V_3) + IDEM + F.M (without sliced input)	95.43	91.28	92.34	94.34
Seg.M (V_3) + IDEM + F.M + ADL (with sliced input)	99.64	98.95	96.07	99.41

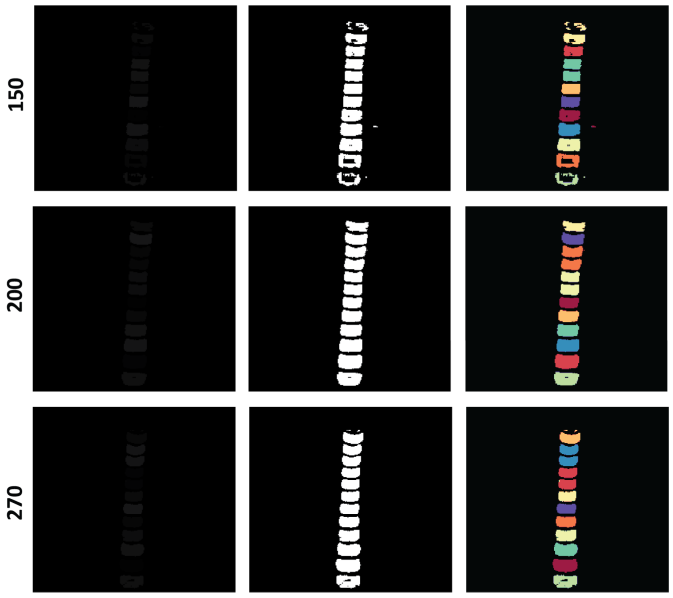

Fig. 8. Qualitative experiment for showing the segmentation and classification performance with different number of training samples We choose the training set with 150, 200 and 270 samples from top to bottom. Column 1 is showing the CT image, column 2 is the coarse segmentation result while column 3 is the representation of refined segmentation and classified vertebra.

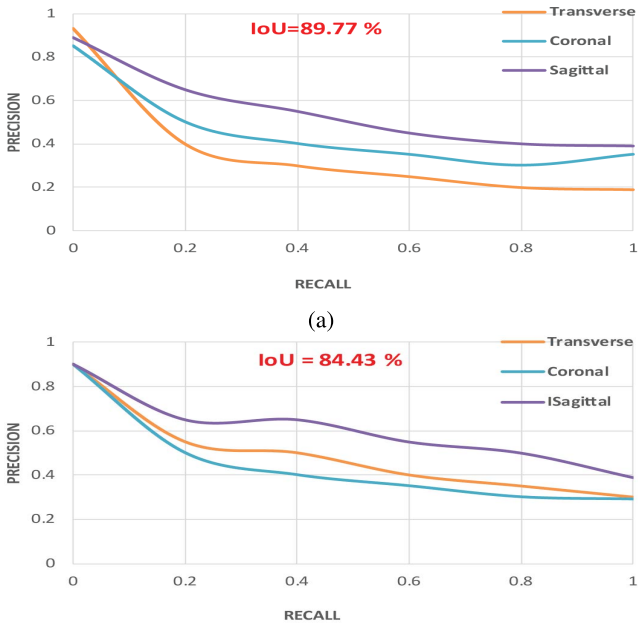

Fig. 9. Receiver operator curves (ROC) for various levels of threshold metric (IoU) with respect to precision and recall values.

intervertebral disc extraction and fusion modules, via the ablation study.

1) Discussion on Various Combination of the Proposed Method: We compare the performance of the segmentation module with and without sliced CT images at input. Another aspect of comparison is carried out by utilizing Mask RCNN [36] and SegNet [37] as an encoder-decoder along with U-Net. Moreover, we have added triplet loss and AD loss along with ECSU-Net whole stream to analyze its performance. The performance of sliced input is similar to without sliced input, but sliced-input architecture provides the possibility to optimize each sliced individually thus making the training process efficient thus easing the trade-off relationship. Compared to without sliced input, the avg. precision and avg. recall of sliced input with AD loss increase up to 4% (92.34% vs 96.07%) and 5% (94.34% vs 99.41%) respectively as shown in Table VI.

We notice that adding triplet loss damages the performance of the ECSU-Net. Triplet loss can enhance the feature extraction ability of ECSU-Net but weaken its training efficiency. Experiments prove our hypothesis that decoupling the three kinds of slices for input data enhance each feature extraction with respect to three dimensions, which as a result can improve the segmentation and classification performance of vertebra. Comparing the performance of proposed model with Mask RCNN [36] and SegNet [37] as an encoder network, we can see form the Table VII that our method achieves higher avg. accuracy and mean avg. accuracy by up to 3-4%(83.43% and 84.37% vs 87.44%) and up to 3-4% (82.44% and 83.44% vs 86.98%) respectively. As, the compared techniques are suited for proposal base segmentation and semantic segmentation which is varied form our goal of instance segmentation. The experimental results reveal that although ECSU-Net has hybrid the interrelated modules, its generalization ability and embedded expert knowledge, significant for learning distinguishable features efficiently from a small amount of training data.

2) Comparison of ADL With Other Losses: In this section, we introduce the choice of our loss function used to train the segmentation network. We tried three different loss functions and finally found the last one achieve the best experiment results.

The first idea is inspired by the triplet loss introduced by [38]. The main idea of triplet loss function is to let the instances within the same category be as close as possible in the outputted embedding space, which just totally match our goal and needs. A common triplet loss function can be written as follow:

$$Triplet(Loss) = max\big(d(a, p) - d(a, n) + margin, 0\big) \tag{21}$$

where a represents the anchors (in our tasks, it is pixel-level embedding vector), p represents a positive image and n represents a negative image, margin is a threshold that defines how far the dissimilarities between the positive and negative samples should be in the embedding space. In our case, to define a proper margin is easier said than done. Although we can turn the scale of the outputted embedding space into a uniform scale, for example from 0 to 1 by using Softmax, since there are multi-class (about 8-10 classes) in one image, the fixed margin is hard to satisfy all pair of classes. Also, running this loss function in this spine segmentation is also not training efficient, since each two different piece of spine need to calculate one loss. Suppose there are $(11 + 1)$ pieces of spine $(+1$ represents the background segment), the overall number of loss we need to calculate and add together is $12 \times 11 = 132$, which might be a huge redundant of training efficiency cost. We tested this loss function with a margin set as 0:5 and add a Softmax layer as the outputted layer and found it works fine but inefficient with respect to training performance.

Since, the intuition of our idea is to let the pixel embedding of the same segments as close as possible, which can be regarded as the distance between the mean value of the two segments as small as possible and the pixel embedding of the different segments as far as possible. Hence, we follow this simple idea and get our second loss function, which can be shown as follow:

$$Loos_2 = a \times dis_{sim} + \frac{b \times 1}{dis_{diff}} \tag{22}$$

where a and b are two weight factors used to balance the two parts of this equation. To improve the training efficiency of this loss function is in each training iteration, we opt to sample a certain number of points to calculate this loss, for example 500 in our experiments settings. We found this loss function not only achieve better training efficiency compared to original triplet loss in our cases but also a better way to avoid overfitting, since in each training iteration, only a part of pixels are taken into consideration. The experiment result of this loss function is much better in both accuracy and efficiency compared to previous triplet loss. However, the problem of this loss function is the embedding of the background is quite noisy, which might affect the clustering result. Table VII is showing the statistical results of ADL with other two losses.

3) Performance With Different Number of Training Samples: In this section, our experiment focus on the segmentation performance with different number of training samples and try to more vividly represent the relationship between the number of samples in the training set and the performance of the final classification. We choose the training set with 150, 200 and 270 samples and train the model independently. We test them on the same validation set and the result is shown in Fig. 8. The visual results shows that with the number of samples in training set increasing, the performance of our model is getting better.

TABLE VII
VARIANTS OF PROPOSED METHODOLOGY ALONG WITH LOSS FUNCTIONS

Method + Loss Function	Avg. Auc (%)	mAvg. Auc (%)	Avg. precision (%)	Avg. Recall (%)
ECSU-Net + Triplet loss	86.93	85.43	89.09	88.98
ECSU-Net + $Loss_2$	95.43	94.31	93.52	96.11
ECSU-Net + ADL	99.64	98.95	92.07	99.41

TABLE VIII
COMPARISON OF PROPOSED METHOD WITH THREE STATE-OF-ART APPROACHES

Method	Accuracy (%)	ASSD (mm)	DC (%)
Lessmann et al. [25]	96.89±0.05	0.45±0.42	94.89±2.13
Vania et al. [8]	98.01±0.13	0.40±0.31	94.00±0.10
Janssens et al. [24]	97.44±0.22	0.38±0.14	95.27±0.52
Ours	**99.64±0.15**	**0.19±0.29**	**95.60±0.15**

G. Comparison With Deep Learning Methods

We have carried out comparison of ECSU-Net with three recent state-of-art vertebra segmentation approaches in terms of DC, accuracy, and ASSD. Table VIII is describing the obtained results on Spineweb dataset-2 [33]. I-FCNN [25] employed instance memory to analyze image patches for searching as well as segmenting vertebra. In [8] a hybrid combination of CNN and FCN along with class redundancy is utilized to enhance vertebra segmentation accuracy while [24] uses a cascade FCNN having localization and segmentation networks for pixel wise multi-class segmentation.

If we compare the classical segmentation results to the simple pixel-wise threshold result, we will find that the dice scores were actually similar. However, if we visualize them, we can see that classical segmentation results look much better than pixel-wise threshold result. This is because classical segmentation algorithms generated more structured regions as compared to thresholding, suffering less from outliers and noises. Moreover, fusion helps by introducing some 3D structure information into the segmentation. Based on the similarity metrics DC provided in Table VIII, ECSU-Net obtained better results than the classic segmentation methods. For vertebra segmentation, ECSU-Net compares favorably with state-of-art methods as we have achieved the DC value up to 95.60±0.15% with an ASSD value of 0.19±0.29 *mm* while taking less time and computational resources because of 2D subnets usage. Vertebrae were classified as correctly or incorrectly segmented with an accuracy of 99.64%. We showed improvement and

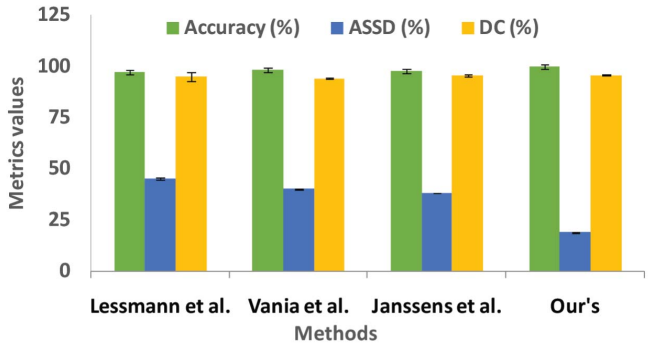

Fig. 10. Graphical comparison of proposed method with Lessmann *et al.* [25],Vania *et al.* [8], and Janssens *et al.* [24].

TABLE IX
COMPARISON OF PROPOSED APPROACH WITH THREE RECENT INTERVERTEBRAL DISC SEGMENTATION APPROACHES

Approach	Dice (%)	ASSD (mm)
Chen et al. [39]	91.60±0.99	0.26±0.45
Payer et al. [40]	91.08±0.90	0.29±0.35
Sekuboyina et al. [41]	88.85±0.20	0.35±0.19
ECSU-Net (ours)	95.60±0.15	0.19±0.29

achieved good results in terms of DC and ASSD as compared to approaches presented in [25], and [8].

The method in [25] has utilized a prior knowledge that the vertebrae are always located next to each other, which may perform substantially better than a regular multiclass FCN with DC score up to 94.9±2.1%. However, segmentation performance in such cases may often susceptible to cascading failure. Although authors in [25] have proposed refinement of the labeling through a maximum likelihood approach, still it can suffers from a similar weakness. Similarly, the approach presented in [8] utilizes class redundancy as a soft constraint to greatly improve the segmentation results with DC up to 94.00±0.1. Redundancy in one case may provide improved segmentation results but at the same time it may cause computational burden which leads to impracticability in clinical routines. The performance of ECSU-Net is comparable to [24] which have utilized pixel-wise multi-class segmentation to map a cropped lumber region volumetric data to its volume-wise labels. However, we have greatly improved in terms of DC score up to 95.60±0.15% with an ASSD value of 0.19±0.29 *mm* with an additional contribution of classifying the segmented vertebras (see Fig. 10). As, our concept of fusion aids in achieving better accuracy by regenerating some ignored 3D structural information during segmentation process.

Payer *et al.* [40] proposed 3D U-Net based multi-staged, patch-wise approach where individual vertebrae are localized and identified with the SpatialConfiguration-Net. Each vertebra is then independently segmented as a binary segmentation with a post processing technique for localization stage's output. Chen *et al.* [39] introduced a multi-staged, patch-based 3D U-Net approach which coarsely localizes the spine along with a U-Net to perform binary segmentation. At the end of the process, a 3D ResNet-model identifies the vertebral class. Using a generative adversarial network Sekuboyina *et al.* [41] proposed Btrfly-Net for labelling sagittal and coronal maximum intensity projections of the spine with the reinforcement of prior learnt. Table IX is showing the comparison of ECSU-Net with above-mentioned approaches.

IV. CONCLUSION

Accurate spinal segmentation and classification from CT images is important for an early diagnosis of spinal disorders, surgical planning and locating spinal pathologies like degenerative disorders, trauma, and fractures. We have presented a way to enhance vertebra segmentation by introducing a three stage network named ECSU-Net, which alleviates intricacy of high resolution input 3D data and anatomical complexity related to the vertebral model. The ECSU-Net concurrently performs segmentation of a vertebra as well as intervertebral disc classification independently of the input image resolutions or spine coverages. Our method efficiently segment the vertebra and provide convincing results in terms of dice score up to 95.60% and classification accuracy of 96.20%. We have provided experimental validation for our claims as well as visualization results to show the practicability of proposed method in clinical routine.

REFERENCES

[1] J. Ajgl and O. Straka, "Covariance intersection in track-to-track fusion: Comparison of fusion configurations," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1127–1136, Mar. 2018.
[2] I. N. Bankman, *Handbook of Medical Image Processing and Analysis*, 2nd ed. Burlington, MA, USA: Academic, 2008.
[3] A. Nazir *et al.*, "OFF-eNET: An optimally fused fully end-to-end network for automatic dense volumetric 3D intracranial blood vessels segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 7192–7202, 2020.
[4] Q. Yu, Y. Shi, J. Sun, Y. Gao, J. Zhu, and Y. Dai, "Crossbar-net: A novel convolutional neural network for kidney tumor segmentation in CT images," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4060–4074, Aug. 2019.
[5] L. Ngo, J. Cha, and J.-H. Han, "Deep neural network regression for automated retinal layer segmentation in optical coherence tomography images," *IEEE Trans. Image Process.*, vol. 29, pp. 303–312, 2020.
[6] A. Sekuboyina, A. Valentinitsch, J. S. Kirschke, and B. H. Menze, "A localisation-segmentation approach for multi-label annotation of lumbar vertebrae using deep nets," *CoRR*, vol. abs/1703.04347, pp. 1–10, Dec. 2017.
[7] R. D. Labati, A. Genovese, E. Munoz, V. Piuri, and F. Scotti, "3-D granulometry using image processing," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1251–1264, Mar. 2019.
[8] M. Vania, D. Mureja, and D. Lee, "Automatic spine segmentation from CT images using convolutional neural network via redundant generation of class labels," *J. Comput. Des. Eng.*, vol. 6, no. 2, pp. 224–232, 2019.
[9] Q. Huang, J. Lan, and X. Li, "Robotic arm based automatic ultrasound scanning for three-dimensional imaging," *IEEE Trans. Ind. Informat.*, vol. 15, no. 2, pp. 1173–1182, Feb. 2019.
[10] F. Meng, H. Li, Q. Wu, B. Luo, and K. N. Ngan, "Weakly supervised part proposal segmentation from multiple images," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 4019–4031, Aug. 2017.
[11] J. H. Siewerdsen *et al.*, "Automatic analysis of global spinal alignment from spine CT images," in *Proc. Med. Imag.*, Mar. 2019, pp. 15–22, doi: 10.1117/12.2513975.
[12] A. B. Oktay and Y. S. Akgul, "Simultaneous localization of lumbar vertebrae and intervertebral discs with SVM-based MRF," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 9, pp. 2375–2383, Sep. 2013.
[13] D. Sierra-Sosa *et al.*, "Scalable healthcare assessment for diabetic patients using deep learning on multiple GPUs," *IEEE Trans. Ind. Informat.*, vol. 15, no. 10, pp. 5682–5689, Oct. 2019.

[14] H. Fehri, A. Gooya, Y. Lu, E. Meijering, S. A. Johnston, and A. F. Frangi, "Bayesian polytrees with learned deep features for multi-class cell segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3246–3260, Jul. 2019.
[15] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz, "Automated model-based vertebra detection, identification, and segmentation in CT images," *Med. Image Anal.*, vol. 13, no. 3, pp. 471–482, Jun. 2009.
[16] D. Li, W. Zhong, K. M. Deh, T. D. Nguyen, M. R. Prince, Y. Wang, and P. Spincemaille, "Discontinuity preserving liver MR registration with three-dimensional active contour motion segmentation," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 7, pp. 1884–1897, Jul. 2019.
[17] D. E. Hyde, J. Peters, and S. K. Warfield, "Multi-resolution graph based volumetric cortical basis functions from local anatomic features," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 12, pp. 3381–3392, Dec. 2019.
[18] J. Yao, S. D. O'Connor, and R. M. Summers, "Automated spinal column extraction and partitioning," in *Proc. IEEE Int. Symp. Biomed. Imag.*, May 2006, pp. 390–393.
[19] C. Chen *et al.*, "Localization and segmentation of 3D intervertebral discs in MR images by data driven estimation," *IEEE Trans. Med. Imag.*, vol. 34, no. 8, pp. 1719–1729, Aug. 2015.
[20] C. Wang, Y. Guo, W. Chen, and Z. Yu, "Fully automatic intervertebral disc segmentation using multimodal 3D U-Net," in *Proc. IEEE 43rd Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, Jul. 2019, pp. 730–739.
[21] T. Li, B. Wei, J. Cong, X. Li, and S. Li, "S3egANet: 3D spinal structures segmentation via adversarial nets," *IEEE Access*, vol. 8, pp. 1892–1901, 2020.
[22] S. Pang *et al.*, "SpineParseNet: Spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 262–273, Jan. 2021.
[23] B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. Criminisi, "Vertebrae localization in pathological spine ct via dense classification from sparse annotations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2013, pp. 262–270.
[24] R. Janssens, G. Zeng, and G. Zheng, "Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional networks," in *Proc. IEEE Int. Symp. Biomed. Imag.*, Jan. 2018, pp. 893–897.
[25] N. Lessmann, B. V. Ginneken, P. A. D. Jong, and I. Išgum, "Iterative fully convolutional neural networks for automatic vertebra segmentation and identification," *Med. Image Anal.*, vol. 53, pp. 142–155, Apr. 2018.
[26] B. De Brabandere, D. Neven, and L. Van Gool, "Semantic instance segmentation with a discriminative loss function," 2017, *arXiv:1708.02551*.
[27] L. Khelifi and M. Mignotte, "A multi-objective decision making approach for solving the image segmentation fusion problem," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3831–3845, Aug. 2017.
[28] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2020.
[29] P. Wang and X. Bai, "Thermal infrared pedestrian segmentation based on conditional GAN," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6007–6021, Dec. 2019.
[30] 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.*, 2015, pp. 234–241.
[31] Z. Han, P. Wang, and Q. Ye, "Adaptive discriminative deep correlation filter for visual object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 155–166, Jan. 2018.
[32] L. Jing, Y. Chen, and Y. Tian, "Coarse-to-fine semantic segmentation from image-level labels," *IEEE Trans. Image Process.*, vol. 29, pp. 225–236, 2020.
[33] J. Yao, J. E. Burns, H. Munoz, and R. M. Summers, "Detection of vertebral body fractures based on cortical shell unwrapping," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2012, pp. 509–516.
[34] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.
[35] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2011.
[36] S. Badhe, V. Singh, J. Li, and P. Lakhani, "Automated segmentation of vertebrae on lateral chest radiography using deep learning," 2020, *arXiv:2001.01277*.
[37] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2015.
[38] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2015, pp. 815–823.
[39] D. Chen, Y. Bai, W. Zhao, S. Ament, J. M. Gregoire, and C. P. Gomes, "Deep reasoning networks for unsupervised pattern de-mixing with constraint reasoning," in *Proc. Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 1500–1509.
[40] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Coarse to fine vertebrae localization and segmentation with SpatialConfiguration-net and U-Net," in *Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2020, pp. 124–133.
[41] A. Sekuboyina, M. Rempfler, A. Valentinitsch, B. H. Menze, and J. S. Kirschke, "Labeling vertebrae with two-dimensional reformations of multidetector CT images: An adversarial approach for incorporating prior knowledge of spine anatomy," *Radiol., Artif. Intell.*, vol. 2, no. 2, Mar. 2020, Art. no. e190074.

Anam Nazir received the M.Sc. degree in computer science from COMSATS University Islamabad, Islamabad, Pakistan, in 2015. She is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. She is also a Lecturer with the Department of Computer Science, COMSATS University Islamabad, Wah Campus, Wah Cantt, Pakistan. Her current research interests include medical image processing and deep learning.

Muhammad Nadeem Cheema received the M.Sc. degree in computer science from COMSATS University Islamabad, Attock Campus, Attock, Pakistan, in 2015. He is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He is also a Lecturer with the Department of Computer Science, COMSATS University Islamabad, Attock Campus. His current research interests include computer vision, medical image analysis, and deep learning.

Bin Sheng (Member, IEEE) received the B.A. degree in English and the B.Eng. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2004, the M.Sc. degree in software engineering from the University of Macau, Taipa, Macau, in 2007, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2011. He is currently a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include virtual reality and computer graphics. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2013. He is currently a Research Assistant Professor with The Hong Kong Polytechnic University, Hong Kong. He has one image/video processing national invention patent and has excellent research project reported worldwide by *ACM TechNews*. His current research interests include image/video stylization, artistic rendering and synthesis, and creative media.

Huating Li received the Ph.D. degree in medicine from Shanghai Jiao Tong University, Shanghai, China, and the Pennington Biomedical Research Center, Baton Rouge, LA, USA, in 2011. She is currently an Associate Professor with the Shanghai Jiao Tong University Affiliated Sixth People's Hospital and the Shanghai Diabetes Institute. Her current research interests include the role of cytokines in the development of fatty liver disease, diabetes, and other obesity-related diseases.

Guangtao Xue (Member, IEEE) received the Ph.D. degree in computer software and theory from Shanghai Jiao Tong University, Shanghai, China, in 2004. He is currently a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. His current research interests include mobile and wireless computing, big data, social networks, distributed computing, and wireless sensor networks.

Jing Qin (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2009. He is currently an Associate Professor with the School of Nursing, The Hong Kong Polytechnic University, Hong Kong. His current research interests include virtual reality for healthcare, medical imaging, deep learning, and health informatics.

Jinman Kim (Member, IEEE) received the B.S. (Hons.) and Ph.D. degrees in computer science from The University of Sydney, Sydney, Australia, in 2001 and 2006, respectively. Since 2006, he has been a Research Associate with the leading teaching hospital, the Royal Prince Alfred. From 2008 to 2012, he was an ARC Postdoctoral Research Fellow, one year leave from 2009 to 2010 to join the MIRALab Research Group, Geneva, Switzerland, as a Marie Curie Senior Research Fellow. Since 2013, he has been with the School of Computer Science, The University of Sydney, where he was a Senior Lecturer, and became an Associate Professor in 2016. His current research interests include medical image analysis and visualization, and computer aided diagnosis.

David Dagan Feng (Life Fellow, IEEE) received the M.Eng. degree in electrical engineering and computer science (EECS) from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.Sc. degree in biocybernetics and the Ph.D. degree in computer science from the University of California, Los Angeles (UCLA), Los Angeles, CA, USA, in 1985 and 1988, respectively, where he received the Crump Prize for Excellence in Medical Engineering. He is currently the Head of the School of Computer Science, the Director of the Biomedical and Multimedia Information Technology Research Group, and the Research Director of the Institute of Biomedical Engineering and Technology, The University of Sydney, Sydney, Australia. He has published over 700 scholarly research papers, pioneered several new research directions, and made a number of landmark contributions in his field. More importantly, however, is that many of his research results have been translated into solutions to real-life problems and have made tremendous improvements to the quality of life for those concerned. He is a fellow of the Australian Academy of Technological Sciences and Engineering. He has served as the Chair for the International Federation of Automatic Control (IFAC) Technical Committee on Biological and Medical Systems, has organized/chaired over 100 major international conferences/symposia/workshops, and has been invited to give over 100 keynote presentations in 23 countries and regions.